Entropy balanced causal effects

A central result for balanced covariate strategies to identify causal effects involves determining the conditional expectation of observed outcomes $Y = DY_1 + (1-D)\,Y_0$ (with binary treatment $D = 0, 1$) given the covariates X via iterated expectations.

1 Propensity-score weighting

Below we describe this evaluation where the propensity score $p(X) \equiv \Pr(D = 1 \mid X)$ is the basis for balancing covariates then connect to entropy balancing.

$$\begin{aligned}
E\left[\frac{DY}{p(X)} \mid X\right] &= E\left[\frac{DY_1}{p(X)} \mid X\right] \\
&= \sum_{D=0}^{1} \int \frac{DY_1}{p(X)} f(DY_1 \mid X)\, dY_1 \\
&= \sum_{D=0}^{1} \int \frac{DY_1}{p(X)} f(Y_1 \mid D, X) \Pr(D \mid X)\, dY_1 \\
&= \int \frac{(0)\,Y_1}{p(X)} f(Y_1 \mid D, X) \Pr(D = 0 \mid X)\, dY_1 \\
&\quad + \int \frac{(1)\,Y_1}{p(X)} f(Y_1 \mid D, X) \Pr(D = 1 \mid X)\, dY_1 \\
&= \frac{\Pr(D = 1 \mid X)}{p(X)} \int (1)\,Y_1 f(Y_1 \mid X)\, dY_1 \\
&= E[Y_1 \mid X]
\end{aligned}$$

and

$$\begin{aligned}
E\left[\frac{(1-D)\,Y}{1-p(X)} \mid X\right] &= E\left[\frac{(1-D)\,Y_0}{1-p(X)} \mid X\right] \\
&= \sum_{D=0}^{1} \int \frac{(1-D)\,Y_0}{1-p(X)} f(DY_0 \mid X)\, dY_0 \\
&= \sum_{D=0}^{1} \int \frac{(1-D)\,Y_0}{1-p(X)} f(Y_0 \mid D, X) \Pr(D \mid X)\, dY_0 \\
&= \int \frac{(1)\,Y_0}{1-p(X)} f(Y_0 \mid D, X) \Pr(D = 0 \mid X)\, dY_0 \\
&\quad + \int \frac{(0)\,Y_0}{1-p(X)} f(Y_0 \mid D, X) \Pr(D = 1 \mid X)\, dY_0 \\
&= \frac{\Pr(D = 0 \mid X)}{1-p(X)} \int (1)\,Y_0 f(Y_0 \mid X)\, dY_0 \\
&= E[Y_0 \mid X]
\end{aligned}$$

Ignorability (conditional mean independence of potential outcomes, Y_1 and Y_0, with treatment, D) implies

$$E[Y_0 \mid X] = E[Y_0 \mid X, D=1] = E[Y_0 \mid X, D=0]$$

and

$$E[Y_1 \mid X] = E[Y_1 \mid X, D=1] = E[Y_1 \mid X, D=0]$$

Thus, average treatment effects conditional on X can be identified via

$$\begin{aligned} ATE(X) &= E\left[\frac{DY}{p(X)} \mid X\right] - E\left[\frac{(1-D)Y}{1-p(X)} \mid X\right] \\ &= E\left[\frac{(D-p(X))Y}{p(X)(1-p(X))} \mid X\right] \end{aligned}$$

$$\begin{aligned} ATT(X) &= \frac{E\left[\frac{DY}{p(X)} \mid X, D=1\right] - E\left[\frac{(1-D)Y}{1-p(X)} \mid X, D=1\right]}{\Pr(D=1 \mid X)} \\ &= E\left[\frac{(D-p(X))Y}{(1-p(X))} \mid X\right] / \Pr(D=1 \mid X) \end{aligned}$$

and

$$\begin{aligned} ATUT(X) &= \frac{E\left[\frac{DY}{p(X)} \mid X, D=0\right] - E\left[\frac{(1-D)Y}{1-p(X)} \mid X, D=0\right]}{\Pr(D=0 \mid X)} \\ &= E\left[\frac{(D-p(X))Y}{p(X)} \mid X\right] / \Pr(D=0 \mid X) \end{aligned}$$

Iterated expectations combined with full common support and ignorability implies

$$\begin{aligned} E_X[E[Y_0 \mid X]] &= E[Y_0] \\ &= E[Y_0 \mid D=1] = E[Y_0 \mid D=0] \end{aligned}$$

and

$$\begin{aligned} E_X[E[Y_1 \mid X]] &= E[Y_1] \\ &= E[Y_1 \mid D=1] = E[Y_1 \mid D=0] \end{aligned}$$

Hence, unconditional average treatment effects can be identified as

$$\begin{aligned} ATE &= E_X[ATE(X)] \\ &= E_X\left[E\left[\frac{(D-p(X))Y}{p(X)(1-p(X))} \mid X\right]\right] \\ &= E\left[\frac{(D-p(X))Y}{p(X)(1-p(X))}\right] \end{aligned}$$

$$\begin{aligned} ATT &= E_X\left[ATT\left(X\right)\right] \\ &= E_X\left[E\left[\frac{\left(D-p\left(X\right)\right)Y}{\left(1-p\left(X\right)\right)} \mid X\right] / \Pr\left(D=1 \mid X\right)\right] \\ &= E\left[\frac{\left(D-p\left(X\right)\right)Y}{\left(1-p\left(X\right)\right)}\right] / \Pr\left(D=1\right) \end{aligned}$$

and

$$\begin{aligned} ATUT &= E_X\left[ATUT\left(X\right)\right] \\ &= E_X\left[E\left[\frac{\left(D-p\left(X\right)\right)Y}{p\left(X\right)} \mid X\right] / \Pr\left(D=0 \mid X\right)\right] \\ &= E\left[\frac{\left(D-p\left(X\right)\right)Y}{p\left(X\right)}\right] / \Pr\left(D=0\right) \end{aligned}$$

Hahn [1998] demonstrates that knowledge of the propensity score does not improve efficiency for estimating ATE but it does for estimating subpopulation averages, ATT or $ATUT$. Much of the attention to weighting strategies for identifying causal effects has consequently focused on ATT. Hirano, Imbens, and Ritter [2003] propose a weighted average treatment effect

$$E\left[g\left(X\right)\left(\frac{DY}{p\left(X\right)}-\frac{\left(1-D\right)Y}{1-p\left(X\right)}\right)\right] / E\left[g\left(X\right)\right]$$

If the weight, $g\left(X\right)$, is the propensity score, $p\left(X\right)$, then the weighted average treatment effect equals the average treatment effect on the treated, ATT.

$$E\left[p\left(X\right)\left(\frac{DY}{p\left(X\right)}-\frac{\left(1-D\right)Y}{1-p\left(X\right)}\right)\right] / E\left[p\left(X\right)\right] = E\left[\frac{\left(D-p\left(X\right)\right)Y}{\left(1-p\left(X\right)\right)}\right] / \Pr\left(D=1\right)$$

Surprisingly, Hirano, Imbens, and Ritter [2003] show weighting by their nonparametric estimator of the propensity score is more efficient than weighting by the true propensity score.

2 Entropy balancing

In practice, it can be difficult to achieve covariate balance with propensity score weighting. Several studies show that including the propensity score can increase selection bias when covariate balance is poor (Drake [1993], Smith and Todd [2001], and Diamond and Sekhon [2006]). Entropy balancing attempts to address the concern from another angle. Weights on the control sample are entropy balanced to the treatment sample.

Hainmueller's [2012] entropy balanced ATT estimator is similar Hirano, Imbens, and Ritter's [2003] weighted average treatment effect estimator for ATT where entropy weighting is employed rather than the propensity score. $E\left[Y_1 \mid D=1\right]$ can be estimated as

$$\frac{\sum\limits_i D_i Y_i}{\sum\limits_i D_i}$$

while the counterfactual $E[Y_0 \mid D=1]$ can be estimated by

$$\frac{\sum_i (1-D_i) w_i Y_i}{\sum_i (1-D_i) w_i}$$

where w_i is a weight for each untreated (control) individual chosen to minimize a distance metric, $h(\cdot)$, as discussed below. Hence, the ATT estimator is

$$ATT = \frac{\sum_i D_i Y_i}{\sum_i D_i} - \frac{\sum_i (1-D_i) w_i Y_i}{\sum_i (1-D_i) w_i}$$

2.1 Optimization

The program to balance the covariates of the untreated to the treated is

$$\begin{aligned} \min_{w_i \geq 0} H(w) = & \sum_i (1-D_i) h(w_i) \\ s.t. & \sum_i (1-D_i) w_i c_{ri}(X_i) = m_r \\ & \sum_i (1-D_i) w_i = 1 \end{aligned}$$

where $h(w_i)$ is typically $w_i \log \frac{w_i}{q_i}$ with base measure q_i (often set to $\frac{1}{n_0}$, n_0 is the control subsample size), and m_r refers to $r = 1, \ldots, R$ moment restrictions drawn from the treatment group while $c_{ri}(X_i) = X_i^r$ or $c_{ri}(X_i) = (X_i - \mu)^r$ for the control group.

Formulating the Lagrangian

$$\begin{aligned} \mathcal{L} = & \sum_i (1-D_i) w_i \log \frac{w_i}{q_i} + \sum_{r=1}^{R} \lambda_r \left(\sum_i (1-D_i) w_i c_{ri}(X_i) - m_r \right) \\ & + (\lambda_0 - 1) \left(\sum_i (1-D_i) w_i - 1 \right) \end{aligned}$$

and solving the first order condition yields the solution for the weights

$$w_i^* = \frac{q_i \exp\left[-\sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \right]}{\sum_i (1-D_i) q_i \exp\left[-\sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \right]}$$

where the denominator is a normalizing constant

$$Z \equiv \sum_i (1-D_i) q_i \exp\left[-\sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \right]$$

The multipliers solve the moment constraints (provided there is a solution).

Additionally, substitution of w_i^* into the Lagrangian produces

$$\begin{aligned}\mathcal{L}\left(w_i^*\right) &= -\log \sum_i \left(1 - D_i\right) q_i \exp\left[-\sum_{r=1}^{R} \lambda_r c_{ri}\left(X_i\right)\right] - \sum_{r=1}^{R} \lambda_r m_r \\ &= -\log Z - \sum_{r=1}^{R} \lambda_r m_r\end{aligned}$$

In other words, if a solution exists the above primal program can be translated as a simpler, unconstrained dual program[1]

$$\max_{\lambda} -\log Z - \sum_{r=1}^{R} \lambda_r m_r$$

or

$$\min_{\lambda} \log Z + \sum_{r=1}^{R} \lambda_r m_r$$

As with other balancing (weighting) approaches, trimming may be advised. Further, trimming may be needed to satisfy the moment constraints.

This approach is similar to Jaynes' [2003] maximum entropy probability assignment in which moment conditions quantify one's background knowledge and lead to probability assignment. For instance, knowledge of the mean with continuous support bounded below leads to assigning an exponential probability distribution (possibly, with displaced support). Also, knowledge of the first and second moments with continuous support yields a normal probability distribution assignment.

However, irrespective of the extent of covariate balance, identification rests on ignorability as well as common support (balanced covariates). Conditional mean independence or strong ignorability (conditional stochastic independence) of potential outcomes, Y_0 and Y_1, and treatment, D, is a thought experiment that the observable data cannot support or refute due to the counterfactual nature of the question posed.

3 Examples

Next, we explore some simple examples where the focus is identification of ATT. We consider entropy balanced weighting (EBW), propensity score weighting (PSW), propensity score matching (PSM), and exogenous dummy variable regression, $E\left[Y_1 \mid D = 1\right] - E\left[Y_0 \mid D = 0\right]$ (DVR). DVR is analogous to entropy balancing where equal weights are applied to the untreated subsample. Entropy balancing is based on the first moment (the mean) of X (the set of covariates) unless otherwise indicated.

[1] If the primal program is infeasible then the dual program is unbounded.

Example 1 (homogeneous outcome, balanced covariates) *Suppose the data generating process (DGP) is as follows.*

Y_1	Y_0	TE	Y	X_1	X_2	X_3	D	$p(X)$	w
26	25	1	26	1	5	2	1	0.5	
25	24	1	25	2	4	2	1	0.5	
22	21	1	22	3	1	3	1	0.5	
19	18	1	19	4	2	1	1	0.5	
26	25	1	25	1	5	2	0	0.5	0.25
25	24	1	24	2	4	2	0	0.5	0.25
22	21	1	21	3	1	3	0	0.5	0.25
19	18	1	18	4	2	1	0	0.5	0.25

where w refers to the weights identified by entropy balancing for the $D = 0$ (untreated or control) subpopulation.

ATT	1
$ATUT$	1
ATE	1

This DGP exhibits ignorable treatment, homogeneous outcome, balanced covariates, and full support. Therefore, all methods identify ATT.

estimator for ATT	*quantity identified*
EBW	1
PSW	1
PSM	1
DVR	1
EW	1

Example 2 (homogeneous outcome, unbalanced coveriates) *Suppose the data generating process (DGP) is as follows.*

Y_1	Y_0	TE	Y	X_1	X_2	X_3	D	$p(X)$	w
22	21	1	22	3	1	3	1	0.667	
25	24	1	25	2	4	2	1	0.5	
22	21	1	22	3	1	3	1	0.667	
19	18	1	19	4	2	1	1	0.5	
26	25	1	25	1	5	2	0	0	0
25	24	1	24	2	4	2	0	0.5	0.25
22	21	1	21	3	1	3	0	0.667	0.5
19	18	1	17	4	2	1	0	0.5	0.25

where w refers to the weights identified by entropy balancing for the $D = 0$ (untreated or control) subpopulation.

ATT	1
$ATUT$	1
ATE	1

This DGP exhibits homogeneous outcome but endogenous treatment, unbalanced covariates, and limited common support. All methods except DVR identify ATT.

estimator for ATT	*quantity identified*
EBW	1
PSW	1
PSM	1
DVR	0

Example 3 (homogeneous outcome, no common support) *Suppose the data generating process (DGP) is as follows.*

Y_1	Y_0	TE	Y	X_1	X_2	X_3	D	$p(X)$	w
26	25	1	26	1	5	2	1	1	
25	24	1	25	2	4	2	1	1	
22	21	1	22	3	1	3	1	1	
19	18	1	19	4	2	1	1	1	
14	13	1	25	1	1	2	0	0	0.25
19	18	1	24	2	2	2	0	0	0.25
34	33	1	21	3	5	3	0	0	0.25
25	24	1	18	4	4	1	0	0	0.25

where w refers to the weights identified by entropy balancing for the $D = 0$ (untreated or control) subpopulation.

ATT	1
ATUT	1
ATE	1

This DGP exhibits homogeneous outcome but unbalanced covariates and no common support. Therefore, the data suggest no reason for any method to identify ATT. However, in this knife-edge case entropy balancing (as well as DVR) is effective.[2]

estimator for ATT	*quantity identified*
EBW	1
PSW	*NA*
PSM	*NA*
DVR	1

Since the propensity score is not bounded away from zero and one, both PSW and PSM are inapplicable in this setting.

Example 4 (heterogeneous outcome, balanced covariates) *Suppose the*

[2] Typically, in a case lacking support like this there is no feasible solution to the entropy-balancing primal (and the dual program is unbounded).

data generating process (DGP) is as follows.

Y_1	Y_0	TE	Y	X_1	X_2	X_3	D	$p(X)$	w
25.50	22	3.50	25.50	1	5	2	1	0.5	
25.84	21	4.84	25.84	2	4	2	1	0.5	
21.96	15	6.96	21.96	3	1	3	1	0.5	
18.82	13	5.82	18.82	4	2	1	1	0.5	
26.50	22	4.50	22	1	5	2	0	0.5	0.25
24.16	21	3.16	21	2	4	2	0	0.5	0.25
22.04	15	7.04	15	3	1	3	0	0.5	0.25
19.18	13	6.18	13	4	2	1	0	0.5	0.25

where w refers to the weights identified by entropy balancing for the $D = 0$ (untreated or control) subpopulation.

ATT	5.28
$ATUT$	5.22
ATE	5.25

Although outcome is heterogeneous, this DGP exhibits ignorable treatment, balanced covariates, and full support. Therefore, all methods identify ATT.[3]

estimator for ATT	*quantity identified*
EBW	5.28
PSW	5.28
PSM	5.28
DVR	5.28

Example 5 (heterogeneous outcome, unbalanced covariates) *Suppose the data generating process (DGP) is as follows.*

Y_1	Y_0	TE	Y	X_1	X_2	X_3	D	$p(X)$	w
21.58	14.16	7.42	21.58	3	1	3	1	0.667	
24.06	20	4.06	24.06	2	4	2	1	0.333	
21.78	15.84	5.94	21.78	3	1	3	1	0.667	
18.34	16	2.34	18.34	4	2	1	1	0.5	
25.52	19.98	5.54	19.98	2	4	2	0	0.333	0.125
25.41	20.02	5.40	20.02	2	4	2	0	0.333	0.125
22.63	15	7.63	15	3	1	3	0	0.667	0.5
19.66	16	3.66	16	4	2	1	0	0.5	0.25

where w refers to the weights identified by entropy balancing for the $D = 0$ (untreated or control) subpopulation.

ATT	4.94
$ATUT$	5.56
ATE	5.25

[3] The results from dummy variable regression are often interpreted as ATE, which is erroneous in this case.

This DGP exhibits ignorable treatment (conditional mean independence for outcome without treatment is satisfied), heterogeneous outcome, unbalanced covariates, and complete support. Since treatment is not exogenous, DVR fails but the other estimators identify ATT.

estimator for ATT	*quantity identified*
EBW	4.94
PSW	4.94
PSM	4.94
DVR	3.69

Example 6 (heterogeneous nonignorable outcome, balanced covariates) *Suppose the data generating process (DGP) is as follows.*

Y_1	Y_0	TE	Y	X_1	X_2	X_3	D	$p(X)$	w
25.50	22.50	3	25.50	1	5	2	1	0.5	
25.84	19.16	6.68	25.84	2	4	2	1	0.5	
21.96	15.04	6.92	21.96	3	1	3	1	0.5	
18.82	13.18	5.64	18.82	4	2	1	1	0.5	
26.50	21.50	5	21.50	1	5	2	0	0.5	0.25
24.16	20.84	3.32	20.84	2	4	2	0	0.5	0.25
22.04	14.96	7.08	14.96	3	1	3	0	0.5	0.25
19.18	12.82	6.34	12.82	4	2	1	0	0.5	0.25

where w refers to the weights identified by entropy balancing for the $D = 0$ (untreated or control) subpopulation.

ATT	5.56
$ATUT$	5.44
ATE	5.5

This DGP exhibits balanced covariates, and full support. Outcome is nonignorable (as conditional mean independence fails for both potential outcome with treatment and potential outcome without treatment) and heterogeneous. Surprisingly, the bias offsets such that ATT estimators identify ATE.

estimator for ATT	*quantity identified*
EBW	5.5
PSW	5.5
PSM	5.5
DVR	5.5

In the next example, we expand entropy balancing to include the first moments, the second moments (including cross terms) along with first moments, and the fourth moments along with the second (including cross terms) and first moments. Entropy weights associated with the first moment balancing are denoted w_1, while weights based on the second moment balancing are denoted w_2, and weights for the fourth moment balancing are denoted w_4.

Example 7 (heterogeneous outcome, expanded entropy balancing) *Suppose the data generating process (DGP) is as follows.*

Y_1	Y_0	TE	Y	X_1	X_2	X_3	D	$p(X)$	w_1	w_2	w_4
10.43	4.43	6	10.43	1	1.134	0.293	1	0.5			
11.21	7.21	4	11.21	2	1.567	0.646	1	0.4			
12	8	4	12	3	2	1	1	0.667			
13.79	3.79	10	13.79	4	2.433	1.354	1	0.4			
16.57	7..57	9	16.57	5	2.866	1.707	1	0.5			
14	9	5	14	3	2	1	1	0.667			
14	5	9	14	3	2	1	1	0.667			
9.43	−0.57	10	9.43	1	1.134	0.293	1	0.5			
10.21	5.21	5	10.21	2	1.567	0.646	1	0.4			
12	6	6	12	3	2	1	1	0.667			
14.79	7.79	7	14.79	4	2.433	1.354	1	0.4			
17.57	9.57	8	17.57	5	2.866	1.707	1	0.5			
7.43	4.43	3	4.43	1	1.134	0.293	0	0.5	0.0833	0.0712	0.0833
9.21	7.21	2	7.21	2	1.567	0.646	0	0.4	0.0833	0.0878	0.0556
11	8	3	8	3	2	1	0	0.667	0.0833	0.0942	0.1667
12.79	4.79	8	4.79	4	2.433	1.354	0	0.4	0.0833	0.0878	0.0556
14.57	7.57	7	7.57	5	2.866	1.707	0	0.5	0.0833	0.0712	0.0833
9.21	7.21	2	7.21	2	1.567	0.646	0	0.4	0.0833	0.0878	0.0556
12.79	4.79	8	4.79	4	2.433	1.354	0	0.4	0.0833	0.0878	0.0556
7.43	−0.57	8	−0.57	1	1.134	0.293	0	0.5	0.0833	0.0712	0.0833
9.21	4.21	5	4.21	2	1.567	0.646	0	0.4	0.0833	0.0878	0.0556
11	6	5	6	3	2	1	0	0.667	0.0833	0.0942	0.1667
12.79	7.79	5	7.79	4	2.433	1.354	0	0.4	0.0833	0.0878	0.0556
14.57	9.57	5	9.57	5	2.866	1.707	0	0.5	0.0833	0.0712	0.0833

where w_1, w_2, w_4 refers to the weights identified by entropy balancing for the $D = 0$ (untreated or control) subpopulation based on the first moment, second moment, and fourth moment, respectively.

ATT	6.9167
$ATUT$	5.0833
ATE	6

This DGP exhibits unbalanced covariates, and full support. Outcome is ignorable (as conditional mean independence is satisfied for potential outcome without treatment) and heterogeneous. Only entropy balancing with first, second (including cross terms), and fourth moment identifies ATT, while both propensity score approaches identify ATT. It is not surprising that EBW (w_1) fails as the means of X in the two subpopulations are the same but the covariates are unbalanced. Adding the second moment moves entropy balancing closer to ATT but it still

falls short.

estimator for ATT	*quantity identified*
$EBW\,(w_1)$	7.0833
$EBW\,(w_2)$	7.025
$EBW\,(w_4)$	6.9167
PSW	6.9167
PSM	6.9167
DVR	7.0833

As identification conditions for EBW, PSW, and PSM are the same, their asymptotic (or identification) behavior is largely the same (with the exception of knife-edge cases). Hence, design choice depends on matching their finite sample properties to the setting at hand where we estimate the propensity score as well as compromise on the number of moments on which to entropy balance the covariates. Hainmueller's [2012] simulations suggest promise for entropy balancing. Nevertheless, the choice is likely context specific and, in part, dependent on the analyst's background knowledge and causal thought experiment.