Propensity Scores and Covariate Balance

Erica E. M. Moodie

Department of Epidemiology, Biostatistics, & Occupational Health
McGill University
Montreal, QC, Canada

erica.moodie@mcgill.ca

McGill

Session goals

- When and why is a marginal effect desirable?
- How can the parameters of a marginal effect be estimated?
  - Can we use traditional approaches?
  - What are the 'newer' approaches?
- What assumptions do we need, and how can we check them?
- Construction of accessible methods of estimation using familiar statistical tools with tractable statistical properties.
- Alternative uses of the propensity score construction.
- Generalization to the case of non-binary exposures.

Road map

1. Marginal effects in a point-source treatment setting
   - Definition
   - Regression and stratification
2. The propensity score: a method to recover covariate balance
   - Definition
   - PS stratification
   - PS matching
   - PS regression
   - inverse probability of treatment weighting
3. Double robustness
4. Generalizing the propensity score
5. Modelling considerations

Concept: Average Potential Outcomes

The *causal* (unconfounded) *effect* of exposure Z on outcome Y is a measure of how much Y changes as Z is manipulated.

- Here Z is not treated as a random variable, but a manipulable quantity that may influence Y.
- Other variables (confounders), X, may also influence Y.
- $Y(z)$ denotes the outcome if the exposure Z is set equal to z :
  - $Y(z)$ is termed a counterfactual or potential outcome.
- A causal quantity of interest is then

$$\mathbb{E}[Y(z)] = \int y f_{Y(z),X}(y,x)\,\mathrm{d}y\mathrm{d}x$$

that is, an average potential outcome (APO).

Aim

Estimate $\mathbb{E}[Y(z)]$ using a random sample of data

$$(x_i, z_i, y_i), i = 1, \ldots, n$$

for z in some set of values

- $z \in \{0, 1\}$
- $z \in \{0, 1, 2, \ldots, K\}$
- $z \in (a, b)$

The earlier approach to estimation using sample averaging can be adopted to produce an estimator.

Idealized calculation

$$
\begin{aligned}
\mathbb{E}[Y(z)] &= \int y\, f_{Y(z),X}(y,x)\;\mathrm{d}y\mathrm{d}x \\
&= \int y\, f_{Y(z)|X}(y|x) f_X(x)\;\mathrm{d}y\,\mathrm{d}x \\
&= \int y\, f_{Y|Z,X}(y|z,x) f_X(x)\;\mathrm{d}y\,\mathrm{d}x
\end{aligned}
$$

where the final line follows if there is no confounding, and we make certain standard assumptions about the counterfactuals.

The quantity $\mathbb{E}[Y(z)]$ could then be estimated from a hypothetical 'experimental' sample of data by replacing the integral by sample average calculation.

Difference from earlier formulation

Note: An important difference from the earlier formulation of the randomized study calculation is that the variables X are present, and may also influence response.

We still seek a 'marginal' quantity, averaging over the distribution of X, at this stage.

Small problem: assumptions

Often, we do not have access to experimental data. There is no intervention on behalf of the researcher, the data are recorded observationally.

If we could correctly specify the model $f_{Y|Z,X}(y|z,x)$, or at least the conditional expectation

$$\mathbb{E}[Y|Z=z,X=x]$$

then this would not be a problem, as we could simply use the iterated expectation result and estimate

$$\widehat{\mathbb{E}}[Y(z)]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y|Z=z,X=x_i].$$

Assumptions?

What assumptions do we need to get the ‘right’ answer, i.e. an unbiased estimator of the marginal mean $\mathbb{E}[Y(z)]$, via regression when data are obtained observationally?

- Correct model specification (of mean of Y given Z and X)
- No unmeasured confounding → exchangeability
- Independence → no interference
- No extrapolation → positivity
- Well-defined exposure → cannot have multiple versions of treatment

What if we *cannot* satisfy the first assumption?

Unconfounded effect estimation by design

- We can sometimes estimate the APO (or a contrast of APOs, such as the average treatment effect: ATE) by designing a randomized control trial.
- Recall the setting in the case of a binary exposure:
    - obtain a random sample of size n of individuals from the target population, and measure their X values;
    - according to some random assignment procedure, intervene to assign treatment Z to individuals, and measure their outcome Y;
    - the link between X and Z is broken by the random allocation.

Unconfounded effect estimation by design

- Recall that this procedure led to the valid use of the estimators of the ATE based on (1) and (2) from the previous section.
- The important feature of the randomized study is that we have, for confounders X (indeed all predictors)

$$f_{X|Z}(x|1) \equiv f_{X|Z}(x|0) \quad \text{for all } x,$$

or equivalently, in the case of a binary confounder,

$$\Pr[X = 1|Z = 1] = \Pr[X = 1|Z = 0].$$

- The distribution of X is *balanced* across the two exposure groups; this renders direct comparison of the outcomes possible. Probabilistically, X and Z are independent.

Constructing a balanced sample

- In a non-randomized study, there is a possibility that the two exposure groups are *not balanced*

$$f_{X|Z}(x|1) \neq f_{X|Z}(x|0) \quad \text{for some } x,$$

or in the binary case

$$\Pr[X = 1|Z = 1] \neq \Pr[X = 1|Z = 0].$$

- If X influences Y also, then this imbalance renders direct comparison of outcomes in the two groups impossible.

Constructing a balanced sample

- While *global* balance may not be present, it may be that *local* balance, i.e. within certain strata of the sample, may be present.
- That is, for $x \in \mathcal{S}$ say, we might have balance; within $\mathcal{S}$, X is independent of Z.

$$f_{X|Z:\mathcal{S}}(x|1 : x \in \mathcal{S}) = f_{X|Z}(x|0 : x \in \mathcal{S})$$

- Then, for individuals who have X values in $\mathcal{S}$, there is the possibility of direct comparison of the treated and untreated groups.
- We might then restrict attention to causal statements relating to stratum $\mathcal{S}$.

Constructing a balanced sample

- For discrete confounders, we might consider defining strata where the X values are *precisely matched*, and then comparing treated and untreated within those strata.
- Consider matching strata $\mathcal{S}_1, \ldots, \mathcal{S}_K$. We would then be able to compute the ATE by noting that

$$\mathbb{E}[Y(1) - Y(0)] = \sum_{k=1}^{K} \mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k] \Pr[X \in \mathcal{S}_k]$$

  - $\mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k]$ may be estimated nonparametrically from the data by using (1) or (2) for data restricted to have $x \in \mathcal{S}_k$.
  - $\Pr[X \in \mathcal{S}_k]$ may be estimated using the empirical proportion of x that lie in $\mathcal{S}_k$.

Constructing a balanced sample

- For continuous confounders, we might consider the same strategy: consider matching strata $\mathcal{S}_1, \ldots, \mathcal{S}_K$. Then the formula

$$\mathbb{E}[Y(1) - Y(0)] = \sum_{k=1}^{K} \mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k] \Pr[X \in \mathcal{S}_k]$$

  still holds.
- However we must assume a model for how $\mathbb{E}[Y(1) - Y(0)|X \in \mathcal{S}_k]$ varies with x for $x \in \mathcal{S}_k$.
- In both cases, inference is restricted to the set of X space contained in

$$\bigcup_{k=1}^{K} \mathcal{S}_k.$$

Constructing a balanced sample

- In the continuous case, the above calculations depend on the assumption that the treatment effect is similar for x values that lie 'close together' in predictor (confounder) space. However

  I. Unless we can achieve exact matching, then the term 'close together' needs careful consideration.

  II. If X is moderate or high-dimensional, there may be insufficient data to achieve adequate matching to facilitate the estimation of the terms

  $$\mathbb{E}[Y(1) - Y(0)|X \in \mathcal{S}_k];$$

  recall that we need a large enough sample of treated and untreated subjects in stratum $\mathcal{S}_k$.

- Nevertheless, matching in this fashion is an important tool in causal comparison.

Balance via the propensity score

- We now come to the important concept of the propensity score that facilitates causal comparison via a balancing approach.
- Recall: our goal is to mimic the construction of the randomized study that facilitates direct comparison between treated and untreated groups.
- We may not be able to achieve this globally, but possibly can achieve it locally in strata of X space.

- The question is how to define these strata.

Balance via the propensity score

- Recall that in the binary exposure case, balance corresponds to being able to state that within $\mathcal{S}$, X is *independent* of Z.
- This can be achieved if $\mathcal{S}$ is defined in terms of a statistic, $e(X)$ say. That is, we consider the conditional distribution

  $$f_{X|Z,e(X)}(x|z,e)$$

  and attempt to ensure that, given $e(X) = e$, Z is independent of X, so that within strata of $e(X)$, the treated and untreated groups are directly comparable.

Balance via the propensity score

- By Bayes theorem, for $z = 0, 1$, we have that

$$f_{X|Z,e(X)}(x|z,e) = \frac{f_{Z|X,e(X)}(z|x,e) f_{X|e(X)}(x|e)}{f_{Z|e(X)}(z|e)}$$

- Now, as Z is binary, we must be able to write the density in the denominator as

$$f_{Z|e(X)}(z|e) = p(e)^z (1 - p(e))^{1-z} \qquad z \in 0, 1$$

where $p(e)$ is a probability, a function of the fixed value e, and where $0 < p(e) < 1$.

Balance via the propensity score

- Therefore, in order to make the density $f_{X|Z,e(X)}(x|z,e)$ functionally independent of z, and so achieve the independence we seek, we need

$$f_{Z|X,e(X)}(z|x,e) = p(e)^z(1-p(e))^{1-z} \qquad z \in 0,1.$$

- But $e(X)$ is a function of X, so automatically we have that

$$f_{Z|X,e(X)}(z|x,e) \equiv f_{Z|X}(z|x).$$

  Therefore, we require that

$$f_{Z|X}(z|x) = f_{Z|X}(z|x,e) = p(e)^z(1-p(e))^{1-z} \equiv f_{Z|e(X)}(z|e)$$

  for all relevant z, x.

Balance via the propensity score

- This can be achieved by choosing the statistic

  $$e(x) = \Pr_{Z|X}[Z = 1|x]$$

  and setting $p(.)$ to be the identity function, so that

  $$f_{Z|X}(z|x) = e^z(1-e)^{1-z} \quad z = 0, 1.$$

- More generally, choosing $e(x)$ to be some monotone transform of $\Pr_{Z|X}[Z = 1|x]$ would also achieve the same balance.
- The corresponding random variable $e(X)$ defines the strata via which the causal calculation can be considered.

Balance via the propensity score

- The function $e(x)$ defined in this way is the *propensity score*[1]. It has the following important properties
  (i) as seen above, it is a balancing score; conditional on $e(X)$, X and Z are independent.
  (ii) it is a *scalar* quantity, irrespective of the dimension of X.
  (iii) in noting that for balance we require that

$$f_{Z|X}(z|x) \equiv f_{Z|e(X)}(z|e),$$

the above construction demonstrates that if $\widetilde{e}(X)$ is another balancing score, then $e(X)$ is a function of $\widetilde{e}(X)$. That is, $e(X)$ is the 'coarsest' balancing score.

[1] see Rosenbaum & Rubin (1983), Biometrika

Evaluating the propensity score

- To achieve balance we must have

$$e(X) = \Pr[Z = 1|X]$$

correctly specified; that is, for confounders X, we must precisely specify the model $\Pr[Z = 1|X]$.

  - If X comprises entirely discrete components, then we may be able to estimate $\Pr[Z = 1|X]$ entirely nonparametrically, and satisfactorily if the sample size is large enough.
  - If X has continuous components, it is common to use parametric modelling, with

$$e(X; \alpha) = \Pr[Z = 1|X; \alpha].$$

Balance then depends on *correct specification* of this model.

Unconfoundedness given the propensity score

- The assumption of ‘no unmeasured confounders’ amounts to assuming that the potential outcomes are jointly *independent* of exposure assignment given the confounders, that is

$$\{Y(0), Y(1)\} \perp\!\!\!\perp Z \mid X.$$

- With a correctly specified propensity score, we now have that

$$Y(z) \perp\!\!\!\perp Z \mid e(X) \qquad \text{for all } z.$$

Estimation using the propensity score

- We now consider the same stratified estimation strategy as before, but using $e(X)$ instead X to stratify.
- Consider strata $\mathcal{S}_1, \ldots, \mathcal{S}_K$ defined via $e(X)$. In this case, recall that

$$0 < e(X) < 1$$

  so we might consider an equal quantile partition, say using quintiles.
- Then we have

$$\mathbb{E}[Y(1) - Y(0)] = \sum_{k=1}^{K} \mathbb{E}[Y(1) - Y(0) | e(X) \in \mathcal{S}_k] \Pr[e(X) \in \mathcal{S}_k]$$

  still holds approximately if the $\mathcal{S}_k$ are small enough.

Estimation using the propensity score

- This still requires us to be able to estimate

  $$\mathbb{E}[Y(1) - Y(0)|e(X) \in \mathcal{S}_k]$$

  which requires us to have a sufficient number of treated and untreated individuals with $e(X) \in \mathcal{S}_k$ to facilitate the 'direct comparison' within this stratum.
- If the expected responses are constant across the stratum, the formulae (1) and (2) may be used.

Matching

The derivation of the propensity score indicates that it may be used to construct *matched* individuals or groups that can be compared directly.

- if two individuals have precisely the same value of $e(x)$, then they are exactly matched;
- if one of the pair is treated and the other untreated, then their outcomes can be compared directly, as any imbalance between their measured confounder values has been removed by the fact that they are matched on $e(x)$;
- this is conceptually identical to the standard procedure of matching in two-group comparison.

Matching

For an exactly matched pair (i_1, i_0), treated and untreated respectively, the quantity

$$Y_{i_1} - Y_{i_0}$$

is an unbiased estimate of the ATE

$$\mathbb{E}[Y(1) - Y(0)];$$

more typically we might choose m such matched pairs, usually with different $e(x)$ values across pairs, and use the estimate

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{i_1} - Y_{i_0})$$

Matching

Exact matching is difficult to achieve, therefore we more commonly attempt to achieve approximate matching:

- may match one treated to M untreated ($1 : M$ matching)
- caliper matching;
- nearest neighbour/kernel matching;
- matching with replacement.

Most standard software packages have functions that provide automatic matching using a variety of methods.

Propensity Score Regression

Up to this point we have considered using the propensity score for stratification, that is, to produce directly comparable groups of treated and untreated individuals.

Causal comparison can also be carried out using regression techniques: that is, we consider building an estimator of the APO by *regressing* the outcome on a function of the exposure and the propensity score.

Regressing on the propensity score is a means of controlling the confounding.

Propensity Score Regression

If we construct a model

$$\mathbb{E}[Y|Z = z, e(Z, X) = e] = \mu(z, e)$$

then because potential outcomes $Y(z)$ and Z are independent given $e(Z, X)$, we have

$$\mathbb{E}[Y(z)|e(Z, X) = e] = \mathbb{E}[Y|Z = z, e(z, X) = e] = \mu(z, e)$$

and therefore

$$\mathbb{E}[Y(z)] = \mathbb{E}_{e(z,X)}[\mathbb{E}[Y|Z = z, e(z, X)]] = \mathbb{E}_{e(z,X)}[\mu(z, e(z, X))].$$

Propensity Score Regression

That is, to estimate the APO, we might

- fit the propensity score model $e(Z, X)$ to the observed exposure and confounder data by regressing Z on X;
- fit the conditional outcome model $\mu(z, e)$ using the fitted $e(Z, X)$ values, $\widehat{e}(z_i, x_i)$;
- for each z of interest, estimate the APO by

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{\mu}(z, \widehat{e}(z, x_i)).$$

Propensity Score Regression

If the propensity function $e(Z, X) \equiv e(X)$, we proceed similarly, and construct a model

$$\mathbb{E}[Y|Z = z, e(X) = e] = \mu(z, e)$$

then

$$\mathbb{E}[Y(z)|e(X) = e] = \mathbb{E}[Y|Z = z, e(X) = e] = \mu(z, e)$$

and therefore

$$\mathbb{E}[Y(z)] = \mathbb{E}_{e(X)}[\mathbb{E}[Y|Z = z, e(X)]] = \mathbb{E}_{e(X)}[\mu(z, e(X))].$$

Propensity Score Regression

To estimate the APO:

- fit the propensity score model $e(X)$ to the observed exposure and confounder data by regressing Z on X;
- fit the conditional outcome model $\mu(z, e)$ using the fitted $e(X)$ values, $\widehat{e}(x_i)$;
- for each z of interest, estimate the APO by

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(z, \widehat{e}(x_i)).$$

Example: Binary Exposure

We specify

- $e(X;\alpha) = \Pr[Z=1|X,\alpha]$ then regress Z on X to obtain $\widehat{\alpha}$ and fitted values $\widehat{e}(X) \equiv e(X;\widehat{\alpha})$.
- $\mathbb{E}[Y|Z=z, e(X)=e;\beta] = \mu(z,e;\beta)$ and estimate this model by regressing y_i on z_i and $\widehat{e}(x_i)$. For example, we might have that

$$\mathbb{E}[Y|Z=z_i, e(X_i)=e_i;\beta] = \beta_0 + \beta_1 z_i + \beta_2 e_i.$$

This returns $\widehat{\beta}$.

We finally compute the predictions under this model, and average them to obtain the APO estimate

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n}\sum_{i=1}^{n} \mu(z, \widehat{e}(x_i); \widehat{\beta}).$$

Example: Binary Exposure

We then compute the predictions under this model, and average them to obtain the APO estimate

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n}\sum_{i=1}^{n} \mu(z, \widehat{e}(z, x_i); \widehat{\beta}).$$

Note that here the propensity terms that enter into μ are computed at the target z values, and

not the observed exposure values.

Propensity Score Regression

These procedures require us to make two modelling choices:

- the propensity model, $e(Z, X)$ or $e(X)$;
- the outcome mean model $\mu(z, e)$.

Note that *both models must be correctly specified* for consistent inference; however, the resulting estimators often have low variability compared to competing estimators.

Misspecification of the outcome mean model will lead to bias; this model needs to capture the outcome to exposure and propensity function relationship correctly.

Inverse probability weighting

If we could intervene at the population level to set $Z = z$ for all individuals independently of their X value, we might rewrite $\mathbb{E}[Y(z)]$ as

$$\mathbb{E}[Y(z)] = \int y \mathbb{1}_{z}(\mathrm{z}) f_{Y(z),X}(y, x) \, \mathrm{d}y \, \mathrm{dz} \, \mathrm{d}x$$

and take a random sample from the population with density

$$\mathbb{1}_{z}(\mathrm{z}) f_{Y(z),X}(y, x) \equiv \mathbb{1}_{z}(\mathrm{z}) f_{Y|\mathrm{Z},X}(y|\mathrm{z}, x) f_X(x).$$

We could then construct the moment estimate

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n} \sum_{i=1}^{n} y_i$$

as $z_i = z$ for all i.

Average potential outcome: Experimental data

In a randomized (experimental) study, suppose that exposure $Z = z$ is assigned with probability determined by $f_Z(z)$. Then

$$\mathbb{E}[Y(z)] = \frac{\int y \, \mathbb{1}_z(\mathrm{z}) \, f_{Y|\mathrm{Z},X}(y|\mathrm{z}, x) f_X(x) f_Z(\mathrm{z}) \, \mathrm{d}y \, \mathrm{dz} \, \mathrm{d}x}{\int \mathbb{1}_z(\mathrm{z}) f_Z(\mathrm{z}) \, \mathrm{dz}}$$

This suggests the Monte Carlo estimates

$$\widehat{\mathbb{E}}[Y(z)] = \frac{\sum_{i=1}^{n} \mathbb{1}_z(\mathrm{z}_i) y_i}{\sum_{i=1}^{n} \mathbb{1}_z(\mathrm{z}_i)} \qquad \text{or} \qquad \widehat{\mathbb{E}}[Y(z)] = \frac{1}{n f_Z(z)} \sum_{i=1}^{n} \mathbb{1}_z(\mathrm{z}_i) y_i$$

Average potential outcome

Commonly, we want to carry out a comparison of average potential outcomes at different values of z, e.g.:

$$\widehat{\mathbb{E}}[Y(1) - Y(0)] \quad = \quad \frac{\sum_{i=1}^{n} \mathbb{1}_1(z_i) y_i}{\sum_{i=1}^{n} \mathbb{1}_{=1}(z_i)} - \frac{\sum_{i=1}^{n} \mathbb{1}_0(z_1) y_i}{\sum_{i=1}^{n} \mathbb{1}_0(z_i)}$$

or

$$\widehat{\mathbb{E}}[Y(1) - Y(0)] \quad = \quad \frac{1}{n f_Z(1)} \sum_{i=1}^{n} \mathbb{1}_1(z_i) y_i - \frac{1}{n f_Z(0)} \sum_{i=1}^{n} \mathbb{1}_0(z_i) y_i.$$

Average potential outcome: Observational data

Denote by $P_{\mathcal{E}}$ the probability measure for samples drawn under the experimental measure corresponding to the density

$$f^{\mathcal{E}}_{Y|Z,X}(y|z,x) f^{\mathcal{E}}_{X}(x) f^{\mathcal{E}}_{Z}(z)$$

Now consider the case where the data arise from the observational (non-experimental) measure $P_{\mathcal{O}}(\mathrm{d}y, \mathrm{d}z, \mathrm{d}x)$. We have

$$\begin{aligned} \mathbb{E}[Y(z)] &= \frac{1}{f^{\mathcal{E}}_{Z}(z)} \int y \mathbb{1}_{z}(z)\, P_{\mathcal{E}}(\mathrm{d}y, \mathrm{d}z, \mathrm{d}x) \\ &= \frac{1}{f^{\mathcal{E}}_{Z}(z)} \int y \mathbb{1}_{z}(z)\, \underbrace{\frac{P_{\mathcal{E}}(\mathrm{d}y, \mathrm{d}z, \mathrm{d}x)}{P_{\mathcal{O}}(\mathrm{d}y, \mathrm{d}z, \mathrm{d}x)}}_{\textcircled{1}} P_{\mathcal{O}}(\mathrm{d}y, \mathrm{d}z, \mathrm{d}x) \end{aligned}$$

In terms of densities $\textcircled{1}$ becomes

$$\frac{f^{\mathcal{E}}_{Y|Z,X}(y|z,x) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x)}{f^{\mathcal{O}}_{Y|Z,X}(y|z,x) f^{\mathcal{O}}_{Z|X}(z|x) f^{\mathcal{O}}_{X}(x)} = \frac{f^{\mathcal{E}}_{Y|Z,X}(y|z,x)}{f^{\mathcal{O}}_{Y|Z,X}(y|z,x)} \times \frac{f^{\mathcal{E}}_{Z}(z)}{f^{\mathcal{O}}_{Z|X}(z|x)} \times \frac{f^{\mathcal{E}}_{X}(x)}{f^{\mathcal{O}}_{X}(x)}$$

Estimation

This suggests the (nonparametric) estimators

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{z}(Z_i)Y_i}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)} \qquad \text{(IPW0)}$$

which is unbiased, or

$$\widehat{\mathbb{E}}[Y(z)] = \frac{\sum_{i=1}^{n} \frac{\mathbb{1}_{z}(Z_i)Y_i}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}}{\sum_{i=1}^{n} \frac{\mathbb{1}_{z}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}} \qquad \text{(IPW)}$$

which is consistent, each provided $f^{\mathcal{O}}_{Z|X}(.|.)$ correctly specifies the conditional density of Z given X for all (z, x).

The inverse probability weighting constructs a pseudo-population in which there are no imbalances on measured covariates between the exposure groups.

Assumptions

What assumptions do we need to get the ‘right’ answer, i.e. an unbiased estimator of the marginal mean $\mathbb{E}[Y(z)]$, via IPW?

- Correct model specification (of mean of Z given X)
- No unmeasured confounding
- Independence
- No extrapolation
- Well-defined exposure

Parametric modelling: two-stage approach

In the formulation, the nonparametric models

$$f^{\mathcal{O}}_{Z|X}(z|x) \qquad \mu(z,x)$$

are commonly replaced by parametric models

$$f^{\mathcal{O}}_{Z|X}(z|x;\alpha) \qquad \mu(z,x;\beta) = \int y\, f^{\mathcal{O}}_{Y|Z,X}(y|z,x;\beta)\,\mathrm{d}y.$$

Parameters (α, β) are estimated from the observed data by regressing

- Stage I: Z on X using $(z_i, x_i), i = 1, \ldots, n$.
- Stage II: Y on (Z, X) using $(y_i, z_i, x_i), i = 1, \ldots, n$.

and using plug-in version of (IPW).

Double robustness

- The IPW is popular, perhaps unduly so given that it is provably less efficient than PS regression.
- Can we improve upon it?

Double robustness

The IPW can be *augmented*. Note that

$$\mathbb{E}[Y(z)] = \mathbb{E}[Y(z) - \mu(z, X)] + \mathbb{E}[\mu(z, X)]$$

where $\mu(z, x) = \mathbb{E}[Y | Z = z, X = x]$.

This gives the alternate estimator

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}_z(Z_i)(Y_i - \mu(Z_i, X_i))}{f^{\mathcal{O}}_{Z|X}(Z_i | X_i)} + \frac{1}{n} \sum_{i=1}^{n} \mu(z, X_i). \tag{AIPW}$$

Doubly robust IPW

Equation (AIPW) is doubly robust (i.e. consistent even if one of $f^{\mathcal{O}}_{Z|X}(z|x)$ and $\mu(z,x)$ is misspecified).

- If $\mu(z, X_i)$ is correctly specified, then $\mathbb{E}[Y_i - \mu(Z_i, X_i)] \to 0$, and the first term in the augmented estimator disappears (asymptotically), leaving the term $\frac{1}{n}\sum_{i=1}^{n} \mu(z, X_i)$ which is consistent for $\mathbb{E}[Y(z)]$.
- If $f^{\mathcal{O}}_{Z|X}(Z_i|X_i)$ is correctly specified, then $\frac{\mathbb{1}_z(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)} \to 1$, and so
$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_z(Z_i)(-\mu(Z_i, X_i))}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)} + \frac{1}{n}\sum_{i=1}^{n} \mu(z, X_i) \to 0,$$
leaving $\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_z(Z_i)Y_i}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}$, which is again consistent.

Further, $\text{Var}_{\text{AIPW}} \leq \text{Var}_{\text{IPW}}$.

Alternative view of augmentation

Scharfstein et al. (1999), Bang & Robins (2005) write the estimating equation yielding (AIPW) as

$$\sum_{i=1}^{n} \frac{\mathbb{1}_{z}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}(Y_i - \mu(Z_i, X_i)) + \sum_{i=1}^{n} \{\mu(z, X_i) - \mu(z)\} = 0$$

Alternative view of augmentation (cont.)

The first summation is a component of the score obtained when performing OLS regression for Y with mean function

$$\mu(z,x) = \mu_0(z,x) + \epsilon \frac{\mathbb{1}_{z}(z)}{f^{\mathcal{O}}_{Z|X}(z|x)}.$$

and $\mu_0(z,x)$ is a conditional mean model, and ϵ is a regression coefficient associated with the derived predictor

$$\frac{\mathbb{1}_{z}(z)}{f^{\mathcal{O}}_{Z|X}(z|x)}.$$

Alternative view of augmentation

Therefore, estimator (AIPW) can be obtained by regressing Y on (X, Z) for fixed z using the mean specification $\mu(z, x)$, and forming the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mu_0(Z_i, X_i) + \widehat{\epsilon}\frac{\mathbb{1}_z(Z_i)}{f_{Z|X}^{\mathcal{O}}(Z_i|X_i)}\right\}.$$

In a parametric model setting, this becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mu_0(Z_i, X_i; \widehat{\beta}) + \widehat{\epsilon}\frac{\mathbb{1}_z(Z_i)}{f_{Z|X}^{\mathcal{O}}(Z_i|X_i; \widehat{\alpha})}\right\}$$

where α is estimated from Stage (I), and β is estimated along with ϵ in the secondary regression.

Augmentation and contrasts

The equivalent to (AIPW) for estimating the ATE for binary treatment

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

is merely $\widehat{\mathbb{E}}[Y(1)] - \widehat{\mathbb{E}}[Y(0)]$ or

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}_1(Z_i)}{f^{\mathcal{O}}_{Z|X}(1|X_i)} - \frac{\mathbb{1}_0(Z_i)}{f^{\mathcal{O}}_{Z|X}(0|X_i)}\right](Y_i - \mu(Z_i, X_i)) \\ + \frac{1}{n}\sum_{i=1}^{n}\{\mu(1, X_i) - \mu(0, X_i)\}\,.$$

Augmentation and contrasts

Therefore we can repeat the above argument and base the contrast estimator on the regression of Y on (X, Z) using the mean specification

$$\mu(z, x) = \mu_0(z, x) + \epsilon \left[\frac{\mathbb{1}_1(z)}{f^{\mathcal{O}}_{Z|X}(1|x)} - \frac{\mathbb{1}_0(z)}{f^{\mathcal{O}}_{Z|X}(0|x)} \right]$$

Beyond binary exposures

The theory developed above extends beyond the case of binary exposures.

Recall that we require *balance* to proceed with causal comparisons; essentially, with strata defined using X or $e(X)$, the distribution of X should not depend on Z.

We seek a scalar statistic such that, conditional on the value of that statistic, X and Z are independent. In the case of general exposures, we must consider balancing scores that are functions of *both* Z and X.

Beyond binary exposures

For a balancing score $b(Z, X)$, we require that

$$X \perp Z \mid b(Z, X).$$

We denote $B = b(Z, X)$ for convenience.

Consider the conditional distribution $f_{Z|X,B}(z|x, b)$: we wish to demonstrate that

$$f_{Z|X,B}(z|x, b) = f_{Z|B}(z|b) \qquad \text{for all } z, x, b.$$

That is, we require that B completely characterizes the conditional distribution of Z given X.

Beyond binary exposures

This can be achieved by choosing the statistic

$$b(z, x) = f_{Z|X}(z|x)$$

in line with the choice in the binary case.

The balancing score defined in this way is termed the

Generalized Propensity Score

which is a balancing score for general exposures.

Beyond binary exposures

Note, however, that this choice that mimics the binary exposure case is not the only one that we might make. The requirement

$$f_{Z|X,B}(z|x,b) = f_{Z|B}(z|b)$$

for all relevant z, x is met if we define $b(Z,X)$ to be *any* sufficient statistic that characterizes the conditional distribution of Z given X.

It may be possible, for example, to choose functions purely of X.

Beyond binary exposures

Normally distributed exposures

Suppose that continuous valued exposure Z is distributed as

$$Z|X = x \sim \text{Normal}(x\alpha, \sigma^2)$$

for row-vector confounder X. We have that

$$f_{Z|X}(z|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(z - x\alpha)^2\right\}$$

Beyond binary exposures

Normally distributed exposures

We might therefore choose

$$b(Z, X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Z - X\alpha)^2\right\}.$$

However, the linear predictor

$$b(X; \alpha) = X\alpha$$

also characterizes the conditional distribution of Z given X; if we know that $x\alpha = b$, then

$$Z|X = x \equiv Z|B = b \sim \text{Normal}(b, \sigma^2).$$

In both cases, parameters α are to be estimated.

Beyond binary exposures

The generalized propensity score inherits all the properties of the standard propensity score;

- it induces balance;
- if the potential outcomes and exposure are independent given X under the unconfoundeness assumption, they are also independent given $b(Z, X)$.

The generalized propensity score can then be used in regression, weighting, matching or stratification approaches.

Key considerations

In light of the previous discussions, in order to facilitate causal comparisons, there are several key considerations that practitioners must take into account.

1. **The importance of no unmeasured confounding.**

   When considering the study design, it is essential for valid conclusions to have measured and recorded all confounders.

Key considerations (cont.)

2. **Model construction for the outcome regression.**
   - ideally, the model for the expected value of Y given Z and X, $\mu(z, x)$, should be correctly specified, that is, correctly capture the relationship between outcome and the other variables.
   - if this can be done, then no causal adjustments are necessary.
   - conventional model building techniques (variable selection) can be used; this will prioritize predictors of outcome and therefore will select all confounders;
   - however, in finite sample, this method may omit weak confounders that may lead to bias.

Key considerations (cont.)

3. **Model construction for the propensity score.**
   Ideally, the model for the (generalized) propensity score, $e(x)$ or $b(z, x)$, should be correctly specified, that is, correctly capture the relationship between the exposure and the confounders. We focus on

   3.1 identifying the *confounders*;
   3.2 *ignoring* the *instruments*: instruments do not predict the outcome, therefore cannot be a source of bias (unless there is unmeasured confounding) - however they can increase the variability of the resulting propensity score estimators.
   3.3 the need for the specified propensity model to induce *balance*;
   3.4 ensuring *positivity*, so that strata constructed from the propensity score have sufficient data within them to facilitate comparison;
   3.5 effective model selection.

Key considerations (cont.)

Note: Conventional model selection techniques (stepwise selection, selection via information criteria, sparse selection) *should not be used* when constructing the propensity score.

This is because such techniques prioritize the accurate prediction of exposure conditional on the other predictors; however, this is *not* the goal of the analysis.

These techniques may merely select strong instruments and omit strong predictors of outcome that are only weakly associated with exposure.

Key considerations (cont.)

Note: An apparently conservative approach is to build rich (highly parameterized) models for both $\mu(z, x)$ and $e(x)$.

This approach prioritizes *bias elimination* at the cost of *variance inflation* for the resulting estimators.

Note: Statistical approaches to model selection (or ‘causal discovery’) are no substitute for expert, subject-area knowledge relating to the likely data generating mechanisms.

Key points: Summary

- A marginal summary attempts to answer questions relevant to policy makers: *what is the expected outcome, averaged over the covariate distribution in my population?*
- Such questions help to avoid the ‘trap’ of contrasting those who are observed to be treated and untreated, as these may be very different (w.r.t confounding variables) groups of individuals.
- To recover a marginal summary, we need to restore, or create, balance on covariates between the treatment groups.
- We can only restore balance on covariates that we have measured. It is crucial to understand the context of the question to begin to assess whether all confounders have been measured.

Key points: Summary

- The propensity score can be used in as part of an adjustment procedure in various ways that utilize standard statistical tools.
  - regression;
  - weighting;
  - weighted regression.
- The propensity score for binary exposures can be extended to more general settings based on the balancing principle
  - generalized propensity score.
- some consideration of variable or model selection in constructing regression procedures is necessary.

Selected references

- Rosenbaum and Rubin (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* **6**:41–55.
- Rubin (1978) Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6**:34–58.
- Bang H, Robins JM, (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–73.
- Scharfstein, DO, Rotnitzky, A, and Robins, JM, (1999), Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120.

Selected references

- Hirano, K. and Imbens, G, (2004). The propensity score with continuous treatment, in *Applied Bayesian Modelling and Causal Inference from Missing Data Perspectives*, A. Gelman, X.-L. Meng (Eds.), Wiley, New York, 73–84
- Imai, K and van Dyk, DA, (2004). Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association*, 99(467), 854–66
- Imbens, G, (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3), 706–10