2. Preliminaries: Basic Causal Inference

Statistical model

In general: Statistical inference concerns relationships among variables in a population

- Z = random vector comprising variables of interest
- *Statistical model:* A class of probability distributions thought to contain the true distribution of Z
- E.g., probability density or mass function

$$p_Z(z;\theta)$$

  indexed by parameter θ (fully parametric model)
- The model represents relationships among elements of Z in the population

Throughout this course: Most of the time, uppercase letters represent random variables/vectors, lowercase letters represent realized values of these (rare exceptions with Greek symbols)

Associational inference

Ubiquitous example: Classical linear regression model

- $Z = (X, Y)$, scalar outcome Y, covariates $X = (X_1, \ldots, X_k)^T$

$$Y = \beta_1 + \beta_2^T X + \epsilon = \beta_1 + \beta_{21} X_1 + \cdots + \beta_{2k} X_k + \epsilon$$

  $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\epsilon \perp\!\!\!\perp X$ ($\perp\!\!\!\perp$ = "independent of")
- Describes the conditional mean of Y given X, $E(Y|X)$, and thus the *association* between Y and X
- Average outcome for individuals with $X = (x_1, \ldots, x_j + 1, \ldots, x_k)^T$ is β_{2j} units greater than that for $X = (x_1, \ldots, x_j, \ldots, x_k)^T$, $\beta_{2j} > 0$
- So if $\beta_{2j} > 0$, larger values of x_j are associated with larger average outcomes in the population
- But cannot infer from the model alone that *intervening* to increase x_j by one unit (if even possible) will *cause* an increase of β_{2j} in average outcome

Causal inference

Goal of researchers:

- Interest is almost always in *causal* rather than associational relationships (whether or not researchers admit it)
- E.g., does administering treatment option A lead to more beneficial outcomes than option B, so that the improvement can be attributed to giving A rather than B?
- Of obvious relevance to the development of dynamic treatment regimes

Can we establish and estimate such causal relationships based on data?

- Different approaches in different disciplines
- Fruitful approach: Use *potential outcomes* (aka *counterfactuals*)
- Neyman (1923), Rubin (1974, 2005), Robins (1986, 1987)

2. Preliminaries: Basic Causal Inference

2.1 Introductory Remarks

2.2 Point Exposure Studies

2.3 Potential Outcomes and Causal Inference

2.4 Estimation of Causal Effects via Outcome Regression

2.5 Review of M-estimation

2.6 Estimation of Causal Effects via the Propensity Score

2.7 Doubly Robust Estimation of Causal Effects

Usual point exposure study

- Sample of n individuals from a population of interest
- Each individual receives (is exposed to) one of several interventions (e.g., treatment options)
- Outcome of interest Y is subsequently ascertained for each
- Intervention received A, individual characteristics (covariates) prior to the intervention X (at *baseline*) recorded for each

Data: $Z_i = (X_i, A_i, Y_i),\ i = 1, \ldots, n$

- Independent and identically distributed (i.i.d.) across i

Simplest case: Two options coded as 0 or 1; set of possible options

$$\mathcal{A} = \{0, 1\}$$

Example

Antihypertensive study: Does an antihypertensive drug (1) reduce systolic blood pressure (SBP) after 6 months relative to no drug (0) in individuals with SBP $>$ 140 mmHg?

- Sample n individuals from this population; some receive 0, some receive 1; measure SBP at entry (baseline) and at 6 months
- *Clinical trial:* Each individual i is randomly assigned to 0 or 1
- *Observational study:* Each individual i is assigned to 0 or 1 at physician's discretion
- Outcome of interest Y_i = SBP at 6 months $-$ SBP at baseline
- $A_i = 0$ or 1, X_i = pre-treatment covariates (age, weight, race, gender, health history, etc)

Example

Goal: Use the data $(X_i, A_i, Y_i),\ i = 1, \ldots, n$, to infer a causal relationship between drug and outcome

- Is reduction in SBP using drug 1 greater than with no drug (0)?
- Usually, more precisely stated as

  If the entire population of interest of individuals with SBP $>$ 140 *mmHg were treated with drug 1, would the average reduction in SBP be greater than that if the entire population were treated with no drug (0)?*

A statistical model

Assumed model:

$$Y \mid A = 0 \sim \mathcal{N}(\mu_0, \sigma^2), \qquad Y \mid A = 1 \sim \mathcal{N}(\mu_1, \sigma^2),$$

$$\sigma^2 > 0, \quad \theta = (\mu_0, \mu_1, \sigma^2)$$

- Difference in average outcome (change in SBP) among individuals observed to receive drug (1) and those who did not (0)

$$\delta = E(Y|A = 1) - E(Y|A = 0) = \mu_1 - \mu_0 \tag{2.1}$$

- Reflects the the association between Y, observed outcome, and A, treatment received

A statistical model

- *Observational study:* δ does not necessarily reflect the causal relationship of interest because individuals who do not receive drug may be inherently different (younger, healthier, smoke less, etc) than those who do
- *Confounder:* A variable related to both the outcome and to which treatment is received; may distort, or confound, the apparent effect of treatment on outcome
- *Clinical trial:* Whether or not an individual receives drug or not is at random, so independent of any individual characteristics
- Thus, it is widely accepted that there are no confounders, so δ reflects purely the causal effect the drug in a clinical trial

These observations can be formalized through the framework of potential outcomes

2. Preliminaries: Basic Causal Inference

Potential outcomes

Philosophically: Causality is a complicated concept

- An intervention may trigger a series of events that ultimately affect outcome
- The point along this sequence we attribute to causality can be difficult to establish
- *Useful simplification:* Potential outcomes

Potential outcomes

In general:

- $\mathcal{A}$ = set of possible treatment options (feasible for all individuals)
- $Y^*(a)$ = outcome that *would be achieved* by a randomly chosen individual in the population if he/she *were to receive* option $a \in \mathcal{A}$ (random variable)
- *Potential outcome* or *counterfactual*
- Hypothetical construct – can conceive of the outcome an individual would have under any treatment option
- Can think of $Y^*(a)$ for any option $a \in \mathcal{A}$ as an inherent characteristic of an individual

Causal treatment effect

Two treatment options: $\mathcal{A} = \{0, 1\}$

Causal treatment effect:

- For any individual, two potential outcomes $Y^*(0)$ and $Y^*(1)$
- Intuitively: If the difference in outcomes an individual would achieve on each treatment $\neq 0$; i.e.,

$$\{Y^*(1) - Y^*(0)\} \neq 0,$$

  this non-zero difference must be *attributable to the treatments*
- *Causal treatment effect*

$$\{Y^*(1) - Y^*(0)\}$$

- Individual-specific

Average causal treatment effect

Challenge: Usually, only one of $Y^*(1)$ or $Y^*(0)$ can be observed for any individual, so cannot obtain the causal treatment effect

Average causal treatment effect:

$$\delta^* = E\{Y^*(1) - Y^*(0)\} = E\{Y^*(1)\} - E\{Y^*(0)\} \qquad (2.2)$$

- δ^* = difference between average outcome that would be achieved if all individuals in the population were to receive option 1 and that if all were to receive option 0
- So (2.2) has a causal interpretation

Can we estimate the average causal treatment effect δ^* using data from a point exposure study?

Formal causal problem

Goal: Estimate δ^* in (2.2) from i.i.d. observed data

$$(X_i, A_i, Y_i), \ i = 1, \ldots, n$$

- I.e., estimate $E\{Y^*(1)\}$ and $E\{Y^*(0)\}$, which are features of the distribution of $Y^*(1)$ and $Y^*(0)$, from the distribution of (X, A, Y)
- Under what conditions can we do this?

A key assumption is required

Stable Unit Treatment Value Assumption (SUTVA)

$$Y_i = Y_i^*(1)A_i + Y_i^*(0)(1 - A_i),\ i = 1, \ldots, n \tag{2.3}$$

- Rubin (1980), aka the *consistency or stability assumption*
- The outcome Y_i observed for individual i, who received treatment A_i, is the same as his potential outcome for that treatment regardless of the conditions under which he received that treatment
- E.g., the outcome i would have if randomized to treatment 1 in a clinical trial is the same as that if she instead received 1 at the discretion of her physician
- Implies *no interference:* Potential outcomes for an individual are unaffected by treatments received or potential outcomes of other individuals
- No interference is often reasonable; an *exception* is when the treatments are vaccines for prevention of an infectious disease

Review of conditional independence

Independence: $Z_1 \perp\!\!\!\perp Z_2$ if

$$p_{Z_1,Z_2}(z_1, z_2) = p_{Z_1}(z_1) p_{Z_2}(z_2) \tag{2.4}$$

$$p_{Z_1|Z_2}(z_1|z_2) = p_{Z_1}(z_1), \text{ if } p_{Z_2}(z_2) > 0 \tag{2.5}$$

$$p_{Z_2|Z_1}(z_2|z_1) = p_{Z_2}(z_2), \text{ if } p_{Z_1}(z_1) > 0 \tag{2.6}$$

for all realizations z_1, z_2

- $p_Z(z)$ is the probability mass function $P(Z = z)$ if Z is discrete or the probability density if Z is continuous
- $p_{Z_1|Z_2}(z_1|z_2)$ is the conditional probability mass function or density of Z_1 given Z_2

Review of conditional independence

Conditional independence: $Z_1 \perp\!\!\!\perp Z_2 | Z_3$ if

$$p_{Z_1,Z_2|Z_3}(z_1, z_2|z_3) = p_{Z_1|Z_3}(z_1|z_3)p_{Z_2|Z_3}(z_2|z_3), \text{ if } p_{Z_3}(z_3) > 0 \quad (2.7)$$

$$p_{Z_1|Z_2,Z_3}(z_1|z_2, z_3) = p_{Z_1|Z_3}(z_1|z_3), \text{ if } p_{Z_2,Z_3}(z_2, z_3) > 0 \quad (2.8)$$

$$p_{Z_2|Z_1,Z_3}(z_2|z_1, z_3) = p_{Z_2|Z_3}(z_2|z_3), \text{ if } p_{Z_1,Z_3}(z_1, z_3) > 0 \quad (2.9)$$

for all realizations z_1, z_2, z_3

- I.e., (2.4)-2.6) hold conditionally at all levels of z_3

We make heavy use of (2.7)-(2.9) throughout the course

Randomized studies

Can show: Under SUTVA (2.3), data from a randomized study (e.g., clinical trial) can be used to estimate the average causal treatment effect δ^* in (2.2)

- As before, randomization ensures treatment assignment is independent of all other factors, including individual characteristics
- Including the outcome an individual *would achieve* under any of the possible treatment options
- That is, for any individual, randomization ensures that

$$\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A \tag{2.10}$$

Randomized studies

Be careful! Do not confuse (2.10) with treatment assignment being independent of *observed outcome*,

$$Y \perp\!\!\!\perp A$$

- By SUTVA, this is equivalent to

$$\{Y^*(1)A + Y^*(0)(1-A)\} \perp\!\!\!\perp A$$

  which clearly is not true
- $Y \perp\!\!\!\perp A$ corresponds to the hypothesis of no treatment effect

Randomized studies

Fundamental result: Under SUTVA (2.3) and (2.10), the average (associational) treatment difference (2.1)

$$\delta = E(Y|A=1) - E(Y|A=0)$$

is the same as the average causal treatment effect (2.2)

$$\delta^* = E\{Y^*(1)\} - E\{Y^*(0)\}$$

Demonstration: Consider $E(Y|A=1)$

$$\begin{aligned} E(Y|A=1) &= E\{Y^*(1)A + Y^*(0)(1-A)|A=1\} \\ &= E\{Y^*(1)|A=1\} = E\{Y^*(1)\} \end{aligned}$$

by SUTVA and then (2.10); similarly $E(Y|A=0) = E\{Y^*(0)\}$. Thus

$$\delta = E(Y|A=1) - E(Y|A=0) = E\{Y^*(1)\} - E\{Y^*(0)\} = \delta^* \quad (2.11)$$

Randomized studies

Implication of (2.11):

- $E(Y|A=1)$ is the average outcome among individuals observed to receive treatment 1
- Can be estimated consistently by the sample average outcome among those receiving treatment 1
- Similarly for $E(Y|A=0)$
- Thus

$$\widehat{\delta} = \overline{Y}_1 - \overline{Y}_0$$

is a consistent estimator for δ, where (sample averages)

$$\overline{Y}_1 = \frac{\sum_{i=1}^{n} A_i Y_i}{\sum_{i=1}^{n} A_i} \quad \text{and} \quad \overline{Y}_0 = \frac{\sum_{i=1}^{n} (1-A_i) Y_i}{\sum_{i=1}^{n} (1-A_i)}$$

- And by (2.11) is a consistent estimator for δ^*

Observational studies

Complication: Individuals receive treatment according to physician discretion or their own choice

- Thus, individuals who receive treatment 1 may have different characteristics from those who receive treatment 0
- So they may be *prognostically different*
- Because $\{Y^*(1), Y^*(0)\}$ reflect prognosis (how an individual would fare on either treatment),

$$\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A$$

  in (2.10) is no longer reasonable
- Thus, the foregoing arguments do not hold, and it is not necessarily the case that $\widehat{\delta}$ consistently estimates δ^*

Observational studies

Hope:

- Suppose individual characteristics (covariates) X^* ascertained prior to treatment can be identified that are associated with both prognosis and treatment selection, i.e., are *confounders*
- Among individuals sharing the same X^*, all factors associated with treatment selection and outcome are taken into account, so that treatment assignment is effectively *at random*
- Formally

$$\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A | X^* \tag{2.12}$$

Difficulty:

- All variables in X^* may not have been captured in the observed data, so that $X^* \not\subset X$
- There may be variables U, *unmeasured confounders*, $U \subset X^*$ but $U \not\subset X$

No unmeasured confounders assumption

Critical assumption: All variables X^* used to make treatment decisions are captured in the data, so that $X^* \subseteq X$, and

$$\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A|X \tag{2.13}$$

- Assumption of *no unmeasured confounders*
- Aka *strong ignorability* assumption (Rubin, 1978)
- *Fundamental difficulty:* It is *impossible to verify* from the observed data that there are no unmeasured confounders and thus that (2.13) holds
- I.e., cannot tell from the data at hand if there are additional variables not recorded in the data that are associated with both prognosis and treatment selection
- Adoption of (2.13) must be justified based on expertise, specific situation, etc

Assumptions

We assume that SUTVA and the no unmeasured confounders (NUC) assumption hold henceforth

- In practice, these must be critically evaluated for relevance on a case by case basis

Under SUTVA (2.3) and NUC (2.13) (and a further assumption):

- The average causal effect δ^* in (2.2) can be identified from the distribution of observed data (X, A, Y)

2. Preliminaries: Basic Causal Inference

Identifiability of δ^*

Fundamental calculation: Consider $E\{Y^*(1)\}$

$$E\{Y^*(1)\} = E[E\{Y^*(1)|X\}] = E[E\{Y^*(1)|X, A=1\}] \quad (2.14)$$
$$= E\{E(Y|X, A=1)\} \quad (2.15)$$

- Second equality in (2.14) follows by NUC
- (2.15) follows by SUTVA
- Outer expectations are wrt marginal distribution of X

$$E[E\{Y^*(1)|X\}] = \int_{\mathcal{X}} E\{Y^*(1)|X=x)\}\, p_X(x)\, d\nu(x)$$

$$E[E\{Y^*(1)|X, A=1\}] = \int_{\mathcal{X}} E(Y|X=x, A=1)\, p_X(x)\, d\nu(x)$$

$$\neq \int_{\mathcal{X}} E(Y|X=x, A=1) p_{X|A}(x|A=1) d\nu(x) = E(Y|A=1)$$

$\nu(\cdot)$ is a dominating measure (Lebesgue, counting measure)

Identifiability of δ^*

Similarly: $E\{Y^*(0)\} = E\{E(Y|X, A = 0)\}$

Result: The average causal treatment effect (2.2)

$$\delta^* = E\{E(Y|X, A = 1)\} - E\{E(Y|X, A = 0)\} \qquad (2.16)$$

- δ^* can be expressed in terms of the observed data (X,A,Y)

Also required: For $E[E\{Y^*(1)|X, A = 1\}]$ and $E\{E(Y|X, A = 1)\}$ (and similarly for $A = 0$) to be well defined, must have

$$p_{X,A}(x, a) = P(X = x, A = a) > 0 \quad a = 0, 1$$

for all $x \in \mathcal{X}$ with $p_X(x) > 0$

- *Convention:* We usually treat random variables as discrete to avoid measure theoretic distractions
- This holds if

$P(A = a|X = x) > 0$ for all x such that $p_X(x) = P(X = x) > 0$ (2.17)

- (2.17) is referred to as the *positivity assumption*

Outcome regression

Implication of (2.16):

$$\delta^* = E\{E(Y|X, A = 1)\} - E\{E(Y|X, A = 0)\}$$

depends on $E(Y \mid X, A)$, the *regression* of observed outcome on covariates and observed treatment received

True regression relationship:

$$E(Y|X = x, A = a) = Q(x, a)$$

- $Q(x, a)$ is the true function of x and a relating observed outcome to covariates and treatment received and thus

$$\delta^* = E\{Y^*(1)\} - E\{Y^*(0)\} = E\{Q(X, 1)\} - E\{Q(X, 0)\}$$

- $Q(x, a)$ ordinarily unknown in practice

Outcome regression

Posit a (parametric) regression model: Assume

$$E(Y|X = x, A = a) = Q(x, a; \beta)$$

- $Q(X, A; \beta)$ is a linear or nonlinear function of x, a, and finite-dimensional parameter β; for example
- Y continuous, linear model (w/ or w/o interaction)

$$Q(x, a; \beta) = \beta_1 + \beta_2 a + \beta_3^T x + \beta_4^T xa \tag{2.18}$$

- Y binary, $\text{logit}(p) = \log\{p/(1-p)\}$

$$\text{logit}\{Q(x, a; \beta\} = \beta_1 + \beta_2 a + \beta_3^T x + \beta_4^T xa$$

- Fit by ordinary least squares (OLS), weighted least squares (WLS), maximum likelihood, etc

Outcome regression

Important: A posited model $Q(x, a; \beta)$ may or may not be *correctly specified*

- The model $Q(x, a; \beta)$ is correctly specified if there exists β_0, referred to as the true value of β, such that $Q(x, a; \beta_0)$ is the true function $Q(x, a)$
- If no such β_0 exists, then the model is not correctly specified

If $Q(x, a)$ were known: Obvious estimator for δ^* based on (X_i, A_i, Y_i), $i = 1, \ldots, n$

$$n^{-1} \sum_{i=1}^{n} \{Q(X_i, 1) - Q(X_i, 0)\}$$

Outcome regression estimator

Assuming $Q(x, a; \beta)$ is correctly specified: Given estimator $\widehat{\beta}$ for β, obvious estimator for δ^* is the *outcome regression estimator*

$$\widehat{\delta}^*_{OR} = n^{-1} \sum_{i=1}^{n} \{Q(X_i, 1; \widehat{\beta}) - Q(X_i, 0; \widehat{\beta})\} \tag{2.19}$$

- Under model (2.18) with no interaction

$$\{Q(x, 1; \beta) - Q(x, 0; \beta)\} = (\beta_1 + \beta_2 + \beta_3^T x) - (\beta_1 + \beta_3^T x) = \beta_2$$

and $\widehat{\delta}^*_{OR} = \widehat{\beta}_2$; similarly for arbitrary function $\phi(x; \beta_1)$

$$Q(x, a; \beta) = \phi(x; \beta_1) + \beta_2 a$$

- In general, substitute fitted model in (2.19)

2. Preliminaries: Basic Causal Inference

Large sample inference

As for any estimator for anything:

- In addition to $\widehat{\delta}^*_{OR}$, require an estimate of the *uncertainty* associated with $\widehat{\delta}^*_{OR}$ as an estimator for δ^* (standard errors, confidence intervals, etc)
- Require the *sampling distribution* of $\widehat{\delta}^*_{OR}$
- Properties of $\widehat{\delta}^*_{OR}$ depend on those of $\widehat{\beta}$
- For all but simplest models, must appeal to *large sample theory approximation*

Generic statistical model

$$Z_1, \ldots, Z_n \quad \text{i.i.d.} \quad Z \sim p_Z(z)$$

Goal: Inference on p-dimensional parameter θ

- θ fully characterizes the distribution of Z, model $p_Z(z; \theta)$
- θ characterizes features of the distribution, e.g., expectation
- Assume the model is correctly specified, where θ_0 is the true value of θ

Statistical problem: Derive an estimator for θ and its large sample properties

- In many common models, natural/popular estimators are *M-estimators*

M-estimator

An M-estimator for θ is the solution (assuming it exists and is well defined) to the $(p \times 1)$ system of *estimating equations*

$$\sum_{i=1}^{n} M(Z_i; \widehat{\theta}) = 0 \tag{2.20}$$

- $M(z;\theta) = \{M_1(z;\theta), \ldots, M_p(z;\theta)\}^T$ is a $(p \times 1)$ *unbiased estimating function* satisfying

$$E_\theta\{M(Z;\theta)\} = 0 \quad \text{for all } \theta$$

- E.g., for model $p_Z(z;\theta)$

$$E_\theta\{M(Z;\theta)\} = \int M(z;\theta)\, p_Z(z;\theta)\, d\nu(z) = 0 \quad \text{for all } \theta$$

- Suppress subscript for evaluation at θ_0 (expectation wrt true distribution of Z)

M-estimator

Maximum likelihood estimator: Fully parametric model $p_Z(z;\theta)$

$$M(z;\theta) = \frac{\partial \log\{p_Z(z;\theta)\}}{\partial\theta}$$

- Right hand side is $(p \times 1)$ vector of derivatives of the logarithm of $p_Z(z;\theta)$ with respect to the elements of θ
- I.e., the score

M-estimator result

Under regularity conditions: $\widehat{\theta}$ satisfying (2.20) is a consistent and asymptotically normal estimator for θ

$$\widehat{\theta} \xrightarrow{p} \theta_0$$

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma) \tag{2.21}$$

- $\xrightarrow{p}$ denotes convergence in probability
- (2.21) is shorthand meaning the left hand side converges in distribution to a $\mathcal{N}(0, \Sigma)$ random vector (covariance matrix Σ)

Standard M-estimator argument

$$0 = \sum_{i=1}^{n} M(Z_i; \theta_0) + \left\{ \sum_{i=1}^{n} \frac{\partial M(Z_i; \theta^*)}{\partial \theta^T} \right\} (\widehat{\theta} - \theta_0)$$

θ^* is a value between $\widehat{\theta}$ and θ_0

$$\frac{\partial M(z; \theta^*)}{\partial \theta^T} = \begin{pmatrix} \frac{\partial M_1(z;\theta)}{\partial \theta_1}, & \cdots, & \frac{\partial M_1(z;\theta)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial M_p(z;\theta)}{\partial \theta_1}, & \cdots, & \frac{\partial M_p(z;\theta)}{\partial \theta_p} \end{pmatrix}_{\theta=\theta^*}$$

Rearrange as

$$\left\{ -n^{-1} \sum_{i=1}^{n} \frac{\partial M(Z_i; \theta^*)}{\partial \theta^T} \right\} n^{1/2} (\widehat{\theta} - \theta_0) = n^{-1/2} \sum_{i=1}^{n} M(Z_i; \theta_0) \quad (2.22)$$

Standard M-estimator argument

By consistency of $\widehat{\theta}$ and because θ^* is between $\widehat{\theta}$ and θ_0, under regularity conditions

$$-n^{-1}\sum_{i=1}^{n}\frac{\partial M(Z_i;\theta^*)}{\partial\theta^T} \xrightarrow{p} -E\left\{\frac{\partial M(Z_i;\theta_0)}{\partial\theta^T}\right\}$$

Assuming $E\{\partial M(Z,\theta^*)/\partial\theta^T\}$ is nonsingular, with increasing probability as $n\to\infty$ so is the left hand side

$$\left\{-n^{-1}\sum_{i=1}^{n}\frac{\partial M(Z_i;\theta^*)}{\partial\theta^T}\right\}^{-1} \xrightarrow{p} \left[-E\left\{\frac{\partial M(Z_i;\theta_0)}{\partial\theta^T}\right\}\right]^{-1}$$

Thus rewrite (2.22) as

$$n^{1/2}(\widehat{\theta}-\theta_0) = \left\{-n^{-1}\sum_{i=1}^{n}\frac{\partial M(Z_i;\theta^*)}{\partial\theta^T}\right\}^{-1} n^{-1/2}\sum_{i=1}^{n}M(Z_i;\theta_0) \quad (2.23)$$

Standard M-estimator argument

By central limit theorem

$$n^{1/2} \sum_{i=1}^{n} M(Z_i; \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big[0, E\{M(Z; \theta_0) M^T(Z; \theta_0)\}\Big]$$

Then by Slutsky's theorem

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma) \tag{2.24}$$

$$\Sigma = \left[E\left\{\frac{\partial M(Z; \theta_0)}{\partial \theta^T}\right\}\right]^{-1} E\{M(Z; \theta_0) M^T(Z; \theta_0)\} \left[E\left\{\frac{\partial M(Z; \theta_0)}{\partial \theta}\right\}\right]^{-1}$$

known as the *sandwich formula*

Standard M-estimator argument

Estimate

$$\left[E\left\{\frac{\partial M(Z_i;\theta_0)}{\partial \theta^T}\right\}\right]^{-1} \quad \text{by} \quad \left[n^{-1}\sum_{i=1}^{n}\frac{\partial M(Z_i;\widehat{\theta})}{\partial \theta^T}\right]^{-1}$$

$$E\{M(Z;\theta_0)M^T(Z;\theta_0)\} \quad \text{by} \quad n^{-1}\sum_{i=1}^{n} M(Z_i;\widehat{\theta})M^T(Z_i;\widehat{\theta})$$

Sandwich variance estimator: Substitute in Σ to obtain $\widehat{\Sigma}$

Approximate sampling distribution for $\widehat{\theta}$: From (2.24)

$$\widehat{\theta} \dot{\sim} \mathcal{N}(\theta_0, n^{-1}\widehat{\Sigma}) \tag{2.25}$$

- "$\dot{\sim}$" means "approximately distributed as"
- Standard errors for elements of $\widehat{\theta}$ are square roots of diagonal elements of $n^{-1}\widehat{\Sigma}$

Illustration: OLS estimator

OLS estimator solves: With model $Q(x, a; \beta)$, β as θ

$$\sum_{i=1}^{n} \frac{\partial\, Q(X_i, A_i; \beta)}{\partial \beta} \{Y_i - Q(X_i, A_i; \beta)\} = 0$$

- Estimating function

$$M(z; \beta) = \frac{\partial\, Q(x, a; \beta)}{\partial \beta} \{Y - Q(x, a; \beta)\}$$

- With correctly specified model is unbiased

$$\begin{aligned} & E_\beta \left[\frac{\partial\, Q(X, A; \beta)}{\partial \beta} \{Y - Q(X, A; \beta)\} \right] \\ & = E_\beta \left(E_\beta \left[\frac{\partial\, Q(X, A; \beta)}{\partial \beta} \{Y - Q(X, A; \beta)\} | X, A \right] \right) \\ & = E_\beta \left[\frac{\partial\, Q(X, A; \beta)}{\partial \beta} \{E_\beta(Y|X, A) - Q(X, A; \beta)\} \right] = 0 \end{aligned}$$

Illustration: OLS estimator

Can show:

$$\left[E\left\{\frac{\partial M(Z;\beta_0)}{\partial \beta^T}\right\}\right]^{-1} = \left[-E\left\{\frac{\partial Q(X,A;\beta_0)}{\partial \beta}\frac{\partial Q(X,A;\beta_0)}{\partial \beta^T}\right\}\right]^{-1}$$

estimated by

$$\left[-n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial Q(X_i,A_i;\widehat{\beta})}{\partial \beta}\frac{\partial Q(X_i,A_i;\widehat{\beta})}{\partial \beta^T}\right\}\right]^{-1}$$

$$E\{M(Z;\beta_0)M^T(Z;\beta_0)\} = E\left\{\text{var}(Y|X,A)\frac{\partial Q(X,A;\beta_0)}{\partial \beta}\frac{\partial Q(X,A;\beta_0)}{\partial \beta^T}\right\}$$

$$\text{var}(Y|X,A) = E[\{Y - Q(X,A;\beta_0)\}^2|X,A]$$

Illustration: OLS estimator

Approximate sampling distribution:

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \qquad \widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, n^{-1}\widehat{\Sigma})$$

- E.g., assuming $\text{var}(Y|X, A) = \sigma^2$, estimated by $\widehat{\sigma}^2$

$$\widehat{\Sigma} = \widehat{\sigma}^2 \left[-n^{-1} \sum_{i=1}^{n} \left\{ \frac{\partial\, Q(X_i, A_i; \widehat{\beta})}{\partial \beta} \frac{\partial\, Q(X_i, A_i; \widehat{\beta})}{\partial \beta^T} \right\} \right]^{-1}$$

- If $\text{var}(Y|X = x, A = a) = V(x, a)$, the optimal estimator for β solves

$$\sum_{i=1}^{n} \frac{\partial\, Q(X_i, A_i; \beta)}{\partial \beta} V^{-1}(X_i, A_i)\{Y_i - Q(X_i, A_i; \beta)\} = 0$$

  and similar results are obtained
- E.g., in generalized linear models, $V(x, a)$ is a known function of $E(Y|X = x, A = a)$

$\widehat{\delta}^*_{OR}$ as an M-estimator

Assume: $Q(x, a; \beta)$ is correctly specified, β estimated by OLS, and solve jointly in δ^* and β the $(p+1 \times 1)$ "stacked" estimating equations

$$\sum_{i=1}^{n} \{Q(X_i, 1; \beta) - Q(X_i, 0; \beta) - \delta^*\} = 0$$

$$\sum_{i=1}^{n} \frac{\partial\, Q(X_i, A_i; \beta)}{\partial \beta} \{Y_i - Q(X_i, A_i; \beta)\} = 0$$

- With $\theta = (\delta^*, \beta^T)^T$, unbiased estimating function

$$M(z; \theta) = \begin{pmatrix} Q(x, 1; \beta) - Q(x, 0; \beta) - \delta^* \\ \dfrac{\partial\, Q(x, a; \beta)}{\partial \beta} \{y - Q(x, a; \beta)\} \end{pmatrix}$$

- β_0 and δ^*_0 are true values of β and δ^*

$$n^{1/2} \begin{pmatrix} \widehat{\delta}^*_{OR} - \delta^*_0 \\ \widehat{\beta} - \beta_0 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$$

$\widehat{\delta}^*_{OR}$ as an M-estimator

Result: With Σ_{11} $(1,1)$ element of Σ

$$\widehat{\delta}^*_{OR} \dot{\sim} \mathcal{N}(0, n^{-1}\widehat{\Sigma}_{11}), \quad \widehat{\Sigma}_{11} = A_n + C_n^T D_n^{-1} B_n D_n^{-1} C_n$$

$$A_n = n^{-1} \sum_{i=1}^{n} \left\{ Q(X_i, 1; \widehat{\beta}) - Q(X_i, 0; \widehat{\beta}) - \widehat{\delta}^* \right\}^2$$

$$B_n = n^{-1} \sum_{i=1}^{n} \left[\{ Y_i - Q(X_i, A_i; \widehat{\beta}) \}^2 \frac{\partial\, Q(X_i, A_i; \widehat{\beta})}{\partial \beta} \frac{\partial\, Q(X_i, A_i; \widehat{\beta})}{\partial \beta^T} \right]$$

$$C_n = n^{-1} \sum_{i=1}^{n} \left\{ \frac{\partial\, Q(X_i, 1; \widehat{\beta})}{\partial \beta} - \frac{\partial\, Q(X_i, 0; \widehat{\beta})}{\partial \beta} \right\}$$

$$D_n = n^{-1} \sum_{i=1}^{n} \left\{ \frac{\partial\, Q(X_i, A_i; \widehat{\beta})}{\partial \beta} \frac{\partial\, Q(X_i, A_i; \widehat{\beta})}{\partial \beta^T} \right\}$$

2. Preliminaries: Basic Causal Inference

The propensity score

Two treatment options: A takes values in $\mathcal{A} = \{0, 1\}$

$$\pi(X) = P(A = 1|X)$$

- Rosenbaum and Rubin (1983)
- Can be generalized to > 2 options (later)

Conditional independence given the propensity score: Under NUC, $\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A|X$

$$\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A|\pi(X) \tag{2.26}$$

- $0 < \pi(x) < 1$ is one-dimensional
- Can show by a conditioning argument that

$$P\{A = 1|Y^*(1), Y^*(0), \pi(X)\} = E\{\mathrm{I}(A = 1)|Y^*(1), Y^*(0), \pi(X)\} = \pi(X)$$

demonstrating (2.26), where $\mathrm{I}(B) = 1$ if B is true, = 0 otherwise

The propensity score

Under SUTVA and NUC:

$$E\{Y^*(1)\} = E\left[E\{Y^*(1)|\pi(X)\}\right] = E\left[E\{Y^*(1)|\pi(X), A=1\}\right]$$
$$= E\left[E\{Y|\pi(X), A=1\}\right]$$

and similarly $E\{Y^*(0)\} = E\left[E\{Y|\pi(X), A=0\}\right]$ and thus

$$\delta^* = E\left[E\{Y \mid \pi(X), A=1\} - E\{Y \mid \pi(X), A=0\}\right] \qquad (2.27)$$

Modeling the propensity score:

- *Randomized study:* $\pi(x)$ is known, often independent of x
- *Observational study:* $\pi(x)$ is unknown and *modeled*
- E.g., logistic regression model

$$\text{logit}\{\pi(x;\gamma)\} = \gamma_1 + \gamma_2^T x \quad \text{or} \quad \pi(x;\gamma) = \frac{\exp(\gamma_1 + \gamma_2^T x)}{1 + \exp(\gamma_1 + \gamma_2^T x)} \qquad (2.28)$$

Maximum likelihood estimator $\widehat{\gamma}$ based on data (X_i, A_i), $i = 1, \ldots, n$

Propensity score stratification

Based on (2.27): Rosenbaum and Rubin (1983)

- Stratify individuals into S groups based on estimated propensities $\pi(X_i;\widehat{\gamma})$, $i = 1,\ldots,n$, by choosing $0 = c_0 < c_1 < \cdots < c_S = 1$ such that individual i belongs to group j if

$$c_{j-1} < \pi(X_i;\widehat{\gamma}) \leq c_j,\ j = 1,\ldots,S$$

- Estimate δ^* by

$$\widehat{\delta}^*_S = \sum_{j=1}^{S} (\overline{Y}_{1j} - \overline{Y}_{0j})\,(n_j/n)$$

  $\overline{Y}_{1j}$ and $\overline{Y}_{0j}$ are the sample average outcomes among individuals in the jth group receiving treatments 1 and 0, n_j = number of individuals in group j

- Suggestion: take $S = 5$ (stratification on quintiles)

Inverse propensity score weighting

More formal basis: Semiparametric theory for missing data (Robins, Rotinitzky, and Zhao, 1994; Tsiatis, 2006)

- If we could observe $\{Y_i^*(1), Y_i^*(0)\}, i = 1, \ldots, n$, obvious estimators for $E\{Y^*(1)\}$ and $E\{Y^*(0)\}$

$$n^{-1} \sum_{i=1}^{n} Y_i^*(1) \quad \text{and} \quad n^{-1} \sum_{i=1}^{n} Y_i^*(0)$$

- But, as in SUTVA, observe $Y_i^*(1)$ only when $A_i = 1$ and $Y_i^*(0)$ only when $A_i = 0$
- "Missing data problem" suggests *inverse probability weighted complete case estimators* originally proposed by Horvitz and Thompson (1952)

Inverse propensity score weighting

Estimation of $E\{Y^*(1)\}$: Weight outcomes for individuals with $A = 1$ by $1/\pi(X)$; represent themselves and others sharing X with $A = 0$

$$n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{\pi(X_i)} \tag{2.29}$$

- E.g., if $\pi(X_i) = 1/3$, i represents himself and 2 others
- Similarly, estimate $E\{Y^*(0)\}$ by

$$n^{-1} \sum_{i=1}^{n} \frac{(1 - A_i) Y_i}{1 - \pi(X_i)} \tag{2.30}$$

Positivity assumption: (2.29) and (2.30) are problematic if $\pi(X) = 0$ or $\pi(X) = 1$, so must have $0 < \pi(X) < 1$ a.s.

- That is, as in (2.17), for $a = 0, 1$,

$$P(A = a|X = x) > 0 \quad \text{for all } x \text{ such that } p_X(x) = P(X = x) > 0$$

Formal justification

(2.29) and (2.30) are unbiased estimators for $E\{Y^*(1)\}$ and $E\{Y^*(0)\}$: Consider (2.29) under the positivity assumption (2.17)

$$E\left\{\frac{AY}{\pi(X)}\right\} = E\left\{\frac{AY^*(1)}{\pi(X)}\right\} = E\left[E\left\{\frac{AY^*(1)}{\pi(X)}\middle| Y^*(1), X\right\}\right]$$
$$= E\left[\frac{E\{A|Y^*(1), X\}Y^*(1)}{\pi(X)}\right] = E\{Y^*(1)\}$$

using SUTVA and, by NUC,

$$E\{A|Y^*(1), X\} = E\{A|X\} = P(A = 1|X) = \pi(X)$$

- Similarly

$$E\left\{\frac{(1-A)Y}{1-\pi(X)}\right\} = E\{Y^*(0)\}$$

Formal justification

Suggests estimators for δ^*:

- If $\pi(x)$ known

$$\widehat{\delta}^*_{IPW} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{\pi(X_i)} - \frac{(1 - A_i) Y_i}{1 - \pi(X_i)} \right\} \tag{2.31}$$

- *Observational studies:* Posit and fit propensity model $\pi(x; \gamma)$ (e.g., logistic)

$$\widehat{\delta}^*_{IPW} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{\pi(X_i; \widehat{\gamma})} - \frac{(1 - A_i) Y_i}{1 - \pi(X_i; \widehat{\gamma})} \right\} \tag{2.32}$$

- Here, assume that the propensity model $\pi(x; \gamma)$ is correctly specified; i.e., there exists γ_0 such that $\pi(x; \gamma_0) = \pi(x)$

Approximate sampling distribution for (2.32)

Stacked M-estimating equations: $\theta = (\delta^*, \gamma^T)^T$

$$\sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{\pi(X_i;\widehat{\gamma})} - \frac{(1-A_i)Y_i}{1-\pi(X_i;\widehat{\gamma})} - \delta^* \right\} = 0$$

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left\{ A_i - \frac{\exp(\gamma_1 + \gamma_2^T X_i)}{1+\exp(\gamma_1 + \gamma_2^T X_i)} \right\} = 0$$

- Can derive approximate sampling distribution for $\widehat{\delta}^*_{IPW}$ using M-estimation theory

Outcome regression vs. inverse propensity weighted estimators

- Both require SUTVA, NUC, and the positivity assumption
- Outcome regression estimator: Assumption on the conditional distribution of Y given X and A; i.e., on $E(Y|X = x, A = a)$; model $Q(x, a; \beta)$ must be correctly specified for $\widehat{\delta}^*_{OR}$ to be a consistent estimator for δ^* (no assumption on $\pi(x)$ required)
- IPW estimator: Assumption on $\pi(x) = P(A = 1|X = x)$; propensity model $\pi(X; \gamma)$ must be correctly specified for $\widehat{\delta}^*_{IPW}$ to be a consistent estimator for δ^* (no assumption on $E(Y|X = x, A = a)$ required)
- Tradeoff – which is harder to model?

Counterintuitive result

With correctly specified model $\pi(x;\gamma)$: Theoretically, $\widehat{\delta}^*_{IPW}$ in (2.32) with γ estimated by $\widehat{\gamma}$ is *more efficient* than (2.31) with γ known

Simple special case, estimation of $E\{Y^*(1)\}$: True $\pi(x) = 1/2$; correctly specified model $\pi(x;\gamma) = \gamma$, $\gamma_0 = 1/2$,

$$\widehat{\gamma} = n^{-1}\sum_{i=1}^{n} A_i$$

Estimators: γ known and γ estimated

$$n^{-1}\sum_{i=1}^{n}\frac{A_i Y_i}{\pi(X;\gamma_0)} = \sum_{i=1}^{n}\frac{A_i Y_i}{(n/2)} = \widehat{\mu}_1$$

$$n^{-1}\sum_{i=1}^{n}\frac{A_i Y_i}{\pi(X;\widehat{\gamma})} = \frac{n^{-1}\sum_{i=1}^{n} A_i Y_i}{n^{-1}\sum_{i=1}^{n} A_i} = \frac{\sum_{i=1}^{n} A_i Y_i}{\sum_{i=1}^{n} A_i} = \overline{Y}_1$$

Counterintuitive result

Can be shown:

$$n^{1/2}(\widehat{\mu}_1 - \mu_1) = 2n^{-1/2}\sum_{i=1}^{n}(A_i Y_i - \mu_1/2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\sigma_1^2 + \mu_1^2)$$

$$n^{1/2}(\overline{Y}_1 - \mu_1) = \left\{ n^{-1}\sum_{i=1}^{n} A_i \right\}^{-1} n^{-1/2}\sum_{i=1}^{n} A_i(Y_i - \mu_1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\sigma_1^2)$$

- $\overline{Y}_1$ with γ estimated is relatively more efficient than $\widehat{\mu}_1$ with γ known
- Similar result for estimation of $E\{Y^*(0)\}$ and δ^*
- This phenomenon persists for much more complicated inverse probability weighted estimators

2. Preliminaries: Basic Causal Inference

2.1 Introductory Remarks
2.2 Point Exposure Studies
2.3 Potential Outcomes and Causal Inference
2.4 Estimation of Causal Effects via Outcome Regression
2.5 Review of M-estimation
2.6 Estimation of Causal Effects via the Propensity Score
2.7 Doubly Robust Estimation of Causal Effects

Augmented inverse probability weighted estimators

Under the assumption that $\pi(x;\gamma)$ is correctly specified: All consistent and asymptotically estimators for δ^* are asymptotically equivalent to an estimator of form

$$\widehat{\delta}^*_{AIPW} = n^{-1}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{\pi(X_i;\widehat{\gamma})} - \frac{(1-A_i)Y_i}{1-\pi(X_i;\widehat{\gamma})} - \{A_i - \pi(X_i;\widehat{\gamma})\}h(X_i)\right] \quad (2.33)$$

- Semiparametric theory (Tsiatis, 2006)
- $h(X)$ is any arbitrary function of X, $\widehat{\delta}^*_{IPW}$ takes $h(X) \equiv 0$
- Can show $E_\gamma[\{A - \pi(X;\gamma_0)\}h(X)] = 0$ for any $h(X)$, so $\widehat{\delta}^*_{AIPW}$ is consistent for δ^*
- The additional term in (2.33) "augments" $\widehat{\delta}^*_{IPW}$ to increase efficiency

Optimal AIPW estimator

Among the class (2.33): The optimal, efficient estimator; i.e., with smallest asymptotic variance, is obtained with

$$h(X) = \frac{E(Y|X, A = 1)}{\pi(X)} + \frac{E(Y|X, A = 0)}{\{1 - \pi(X)\}}$$

- $\pi(x)$ is the true propensity score
- $E(Y|X, A)$ is not known, but can posit a model $Q(x, a; \beta)$
- Given fitted models $\pi(x; \widehat{\gamma})$ and $Q(x, a; \widehat{\beta})$, suggests the estimator

$$\widehat{\delta}^*_{DR} = n^{-1} \sum_{i=1}^{n} \left[\frac{A_i Y_i}{\pi(X_i; \widehat{\gamma})} - \frac{(1 - A_i) Y_i}{1 - \pi(X_i; \widehat{\gamma})} - \frac{\{A_i - \pi(X_i; \widehat{\gamma})\}}{\pi(X_i; \widehat{\gamma})} Q(X_i, 1; \widehat{\beta}) \right.$$
$$\left. - \frac{\{A_i - \pi(X_i; \widehat{\gamma})\}}{1 - \pi(X_i; \widehat{\gamma})} Q(X_i, 0; \widehat{\beta}) \right] \qquad (2.34)$$

Double robustness

Result: $\widehat{\delta}^*_{DR}$ is a consistent estimator for δ^* if only one of the models $\pi(x;\gamma)$ or $Q(x,a;\beta)$ is correctly specified

- $\widehat{\delta}^*_{DR}$ is "robust to" misspecification of one of these models
- "Two tries" to develop a correct model leading to a consistent estimator

Demonstration: Suppose as $n \to \infty$, for some γ^* and β^*

$$\widehat{\gamma} \xrightarrow{p} \gamma^* \quad \text{and} \quad \widehat{\beta} \xrightarrow{p} \beta^*$$

- $\widehat{\gamma}$ and $\widehat{\beta}$ are M-estimators
- If $\pi(x;\gamma)$ is correctly specified, $\gamma^* = \gamma_0$; else, $\widehat{\gamma}$ is a function of the data so has some limit in probability γ^*
- Similarly, if $Q(x,a;\beta)$ is correctly specified, $\beta^* = \beta_0$; else, $\widehat{\beta}$ is a function of the data so has some limit in probability β^*

Double robustness

Thus: $\widehat{\delta}^*_{DR}$ in (2.34) converges in probability to

$$E\left[\frac{AY}{\pi(X;\gamma^*)} - \frac{(1-A)Y}{1-\pi(X;\gamma^*)} - \frac{\{A-\pi(X;\gamma^*)\}}{\pi(X;\gamma^*)}Q(X,1;\beta^*) - \frac{\{A-\pi(X;\gamma^*)\}}{1-\pi(X;\gamma^*)}Q(X,0;\beta^*)\right]$$

which, using SUTVA and algebra, can be rewritten as

$$E\{Y^*(1) - Y^*(0)\}$$

$$+ E\left[\frac{\{A-\pi(X;\gamma^*)\}}{\pi(X;\gamma^*)}\{Y^*(1) - Q(X,1;\beta^*)\}\right] \quad (2.35)$$

$$+ E\left[\frac{\{A-\pi(X;\gamma^*)\}}{1-\pi(X;\gamma^*)}\{Y^*(0) - Q(X,0;\beta^*)\}]\right] \quad (2.36)$$

- Want to show (2.35) and (2.36) = 0

Double robustness

By NUC and (2.7):

$$(2.35) = E\left(E\left[\frac{\{A-\pi(X;\gamma^*)\}}{\pi(X;\gamma^*)}\{Y^*(1)-Q(X,1;\beta^*)\}\middle| X\right]\right)$$

$$= E\left(\frac{E\left[\{A-\pi(X;\gamma^*)\}\middle| X\right]}{\pi(X;\gamma^*)}E[\{Y^*(1)-Q(X,1;\beta^*)\}|X]\right)$$

$$(2.36) = E\left(\frac{E\left[\{A-\pi(X;\gamma^*)\}\middle| X\right]}{1-\pi(X;\gamma^*)}E[\{Y^*(0)-Q(X,0;\beta^*)\}|X]\right)$$

- $\pi(x;\gamma)$ *correct:* $\gamma^* = \gamma_0$, $\pi(X;\gamma^*) = \pi(X;\gamma_0) = \pi(X)$

  $$E[\{A-\pi(X;\gamma^*)\}|X] = E[\{A-\pi(X)\}|X] = E(A|X) - \pi(X) = 0$$

  using $E(A|X) = \pi(X)$, so that (2.35) and (2.36) = 0

Double robustness

By NUC and (2.7):

$$(2.35) = E\left(E\left[\frac{\{A-\pi(X;\gamma^*)\}}{\pi(X;\gamma^*)}\{Y^*(1)-Q(X,1;\beta^*)\}\middle| X\right]\right)$$

$$= E\left(\frac{E\left[\{A-\pi(X;\gamma^*)\}\middle| X\right]}{\pi(X;\gamma^*)}E[\{Y^*(1)-Q(X,1;\beta^*)\}|X]\right)$$

$$(2.36) = E\left(\frac{E\left[\{A-\pi(X;\gamma^*)\}\middle| X\right]}{1-\pi(X;\gamma^*)}E[\{Y^*(0)-Q(X,0;\beta^*)\}|X]\right)$$

- *$Q(x,a;\beta)$ correct:* $\beta^* = \beta_0$, $Q(X,a;\beta^*) = Q(X,a;\beta_0)$
  $= E(Y|X,A=a) = E\{Y^*(a)|X\}$, $a=0,1$

$$E[\{Y^*(a)-Q(X,a;\beta^*)\}|X] = E[\{Y^*(a)-Q(X,a;\beta_0)\}|X] = 0, \quad a=0,1$$

so that (2.35) and (2.36) = 0

Efficient estimator

Result: $\widehat{\delta}^*_{DR}$ is doubly robust

If both propensity and outcome regression models are correctly specified:

- $\widehat{\delta}^*_{DR}$ in (2.34) based on the optimal choice of $h(X)$ achieves the smallest asymptotic variance among all AIPW estimators of the form (2.33)
- This fundamental result follows from semiparametric theory

Randomized study: The propensity $\pi(x)$ is known

- $\widehat{\delta}^*_{DR}$ is a consistent estimator for δ^* whether or not the outcome regression model $Q(x, a; \beta)$ is correctly specified
- The augmentation term yields increased efficiency over $\widehat{\delta}^*_{IPW}$
- $\widehat{\delta}^*_{OR}$ still requires a correct outcome regression model

Asymptotic properties

Using M-estimation theory: $\theta = (\delta, \gamma^T, \beta^T)^T$, and $\widehat{\delta}^*_{DR}$, $\widehat{\gamma}$, and $\widehat{\beta}$ solve "stacked" estimating equations; for example

$$\sum_{i=1}^{n} \left[\frac{A_i Y_i}{\pi(X_i;\gamma)} - \frac{(1-A_i)Y_i}{1-\pi(X_i;\gamma)} - \frac{\{A_i - \pi(X_i;\gamma)\}}{\pi(X;\gamma)} Q(X_i, 1; \beta) \right.$$
$$\left. - \frac{\{A_i - \pi(X_i;\gamma)\}}{1-\pi(X;\gamma)} Q(X_i, 0; \beta) - \delta^* \right] = 0$$

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left\{ A_i - \frac{\exp(\gamma_0 + \gamma_1^T X_i)}{1 + \exp(\gamma_0 + \gamma_1^T X_i)} \right\} = 0$$

$$\sum_{i=1}^{n} \frac{\partial\, Q(X_i, A_i; \beta)}{\partial \beta} \{ Y_i - Q(X_i, A_i; \beta) \} = 0$$

Onward

With this background, we are ready to single decision tackle treatment regimes. . .