Review of Lecture 1 on Two Phase Design

Definition. Assume for each subject i, $1 \leq i \leq n$, we observe $(X_i, R_i, Y_i R_i)$, where

1. $(X_1, Y_1), \ldots, (X_n, Y_n) \sim$ i.i.d. F. X_i is the covariate, and Y_i is the outcome (Y_i exists for everyone; some of the Y_i's are not observed by the investigator);

2. R_i is a binary variable (response indicator): Y_i is observed if $R_i = 1$; otherwise Y_i is missing.

3. (Ignorability) $R_i \perp Y_i \mid X_i$: R_i is independent of Y_i given X_i.

Example. (Simplified double sampling design, extended)

In a study we follow n patients (consider them to be i.i.d.) to observe their Y_i = time to death, but each patient is lost at some time X_i. At the end of the study, we are able to find out some Y_i's with extra cost. Consider the following 3 strategies of deciding whose Y_i to find out:

1. For each patient i, we flip a coin with success probability 0.5 and record $R_i = 1$ if the coin shows head, $R_i = 0$ if the coin shows tail. Then, we observe those Y_i with $R_i = 1$.

2. For each patient i, we flip a coin with success probability $p_i \propto X_i$, i.e. *the probability of showing head is larger if the patient dropped out later in the study.* We record $R_i = 1$ if the coin shows head, $R_i = 0$ if the coin shows tail. Then, we observe those Y_i with $R_i = 1$.

3. Constantine (an independent investigator, i.e. he's not part of "we") knows every patient's Y_i. He flips coin for each patient i with success probability $p_i \propto Y_i$, i.e. *the probability of showing head is larger if the patient lived longer.* He records $R_i = 1$ if the coin shows head, $R_i = 0$ if the coin shows tail. Then, he provides the recorded R_i to us, and we observe those Y_i with $R_i = 1$.

Question: Which of the above three strategies meet(s) the ignorability assumption ($R_i \perp Y_i \mid X_i$)?

Types of missing data

1. **Missing Completely At Random (MCAR)**: $R \perp (Y, X)$
   Data are missing independently of both observed and unobserved data.

2. **Missing At Random (MAR)**: $R \perp Y \mid X$
   Given the observed data, data are missing independently of unobserved data.

3. **Not Missing At Random (NMAR)**: $R \not\perp Y \mid X$
   Missing observations related to values of unobserved data.

Reference:

- Section 1.3, Statistical Analysis with Missing Data (2014), John Wiley & Sons, R. Little & D. Rubin
- Inference and Missing Data (1976), Biometrika, D. Rubin

Question from last lecture:

1. Do we need $R_i \perp Y_i \mid X_i$ in order for $\hat{\tau}_1$ to be consistent?

   Recall: If we have a *correctly specified* regression model $\hat{E}(Y \mid X)$ for $E(Y \mid X)$, then (under some convergence condition about $\hat{E}(Y \mid X) \to E(Y \mid X)$)

   $$\hat{\tau}_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y \mid X_i)$$

   is a consistent estimator for τ.

2. How to check $R_i \perp Y_i \mid X_i$ assumption?

   Sensitivity Analysis:

   1) Assuming $R_i \perp Y_i \mid X_i$, we have some inference (e.g. point estimate and confidence interval for estimating $E(Y)$).

   2) Relax the assumption, and see how the inference change. E.g. assume some correlation ρ between R_i and Y_i given X_i, and the point estimate and CI will become a function on ρ.

   3) If the inference changes a lot, then "the result is sensitive to violation of the ignorability assumption".

3.2 Efficient Influence Function for τ in Two Phase Design

Theorem 3. *Consider an estimand τ. Let F_0 be the true distribution, and $\tau_0 := \tau(F_0)$ be the true value of the estimand. Without additional assumptions, the influence function of the best estimator for τ_0 has the form*

$$\begin{aligned} \varphi^{eff}(z) &= L_{F_0}(\tau; \delta_z) \\ &= \lim_{\epsilon \to 0} \frac{\tau\left\{(1-\epsilon)F_0 + \epsilon\delta_z\right\} - \tau(F_0)}{\epsilon}, \end{aligned}$$

where δ_z is a point mass on z. φ^{eff} is called the (nonparametric) Efficient Influence Function for the estimand τ.

Remark. "Without additional assumptions" means we allow the distribution of (X, R, Y) to take any form. For example, we don't require $E(Y \mid X)$ to be a linear model; we don't require $P(R = 1 \mid X)$ to be a logistic regression.

Intuitively, the "best" estimator means the estimator with the smallest variance among all consistent and asymptotically normal estimators. Rigorously, the "best" estimator means the asymptotically most efficient estimator among all regular asymptotically linear (RAL) estimators.

Corollary 1. *Assume F_0 is the true distribution for (X, R, YR) in the Two Phase Design. For the estimand $\tau = E_X\left\{E\left(Y \mid X, R = 1\right)\right\}$, its Efficient Influence Function is*

$$\varphi^{eff}(z) = -\tau_0 + E_0(Y \mid R = 1, X = x) + \frac{[y - E_0(Y \mid R = 1, X = x)] \times r}{e_0(x)},$$

where $z = (x, r, yr)$ is a data point, $\tau_0 = \tau(F_0)$ is the true parameter value, $E_0(Y \mid R = 1, X = x)$ is the true regression model (i.e. $E(Y \mid R = 1, X = x)$ under F_0), and $e_0(x) = P_{F_0}(R = 1 \mid X = x)$ is the true propensity score.

Remark. Corollary 1 doesn't require ignorability assumption, because the

estimand here is $E_X\left\{E\left(Y \mid X, R=1\right)\right\}$. Ignorability assumption is required for it to equal $E(Y)$.

Proof. (Sketch)

By the above theorem, we have

$$\varphi^{\text{eff}}\left(z\right) = \lim_{\epsilon \to 0} \frac{\tau\left\{(1-\epsilon)F_0 + \epsilon\delta_z\right\} - \tau(F_0)}{\epsilon}. \tag{1}$$

To simplify notation, define

$$F_\epsilon := (1-\epsilon)F_0 + \epsilon\delta_z,$$

and

$$\begin{aligned} \tau_0 &= E_{F_0}\left\{E_{F_0}\left(Y \mid R=1, X\right)\right\} \\ &=: E_0^1 E_0^2 (Y \mid R=1, X), \\ \tau_\epsilon := \tau(F_\epsilon) &= E_{F_\epsilon}\left\{E_{F_\epsilon}\left(Y \mid R=1, X\right)\right\} \\ &=: E_\epsilon^1 E_\epsilon^2 (Y \mid R=1, X). \end{aligned}$$

With the above notation, (1) (without the limit) becomes:

$$\begin{aligned} \frac{\tau\left\{(1-\epsilon)F_0 + \epsilon\delta_z\right\} - \tau(F)}{\epsilon} &= \frac{\tau(F_\epsilon) - \tau(F_0)}{\epsilon} \\ &= \frac{1}{\epsilon}\left\{E_\epsilon^1 E_\epsilon^2 (YR \mid R=1, X) - E_0^1 E_0^2 (YR \mid R=1, X)\right\} \\ \text{(omit the terms in } E) &= \frac{1}{\epsilon}\left\{E_\epsilon^1 E_\epsilon^2 - E_0^1 E_0^2\right\} \\ \text{(add and subtract)} &= \frac{1}{\epsilon}\left\{E_\epsilon^1 E_\epsilon^2 - E_0^1 E_\epsilon^2 + E_0^1 E_\epsilon^2 - E_0^1 E_0^2\right\} \\ \text{(linearity of expectation)} &= \frac{1}{\epsilon}\left(E_\epsilon^1 - E_0^1\right) E_\epsilon^2 + \frac{1}{\epsilon} E_0^1 \left(E_\epsilon^2 - E_0^2\right) \\ \text{(add and subtract)} &= \frac{1}{\epsilon}\left(E_\epsilon^1 - E_0^1\right)\left(E_\epsilon^2 - E_0^2\right) + \frac{1}{\epsilon}\left(E_\epsilon^1 - E_0^1\right) E_0^2 + \frac{1}{\epsilon} E_0^1 \left(E_\epsilon^2 - E_0^2\right) \\ &= \text{I} + \text{II} + \text{III}. \end{aligned}$$

Claim 1: $\lim_{\epsilon \to 0} \mathrm{I} = 0$.

Intuitively, $\frac{1}{\epsilon}\left(E_\epsilon^1 - E_0^1\right)\left(E_\epsilon^2 - E_0^2\right) = \frac{1}{\epsilon} O(\epsilon^2) = O(\epsilon)$.

Claim 2: $\lim_{\epsilon \to 0} \mathrm{II} = -\tau_0 + E_0\left(Y \mid R = 1, X = x\right)$.

Proof in the case when X is discrete taking value in $\mathcal{X}$, with probability mass function $p_0(x)$ under F_0:

$$\begin{aligned}
\frac{1}{\epsilon}\left(E_\epsilon^1 - E_0^1\right) E_0^2 &= \frac{1}{\epsilon} \sum_{x' \in \mathcal{X}} \left[(1-\epsilon) p_0(x') + \epsilon 1(x' = x) - p_0(x')\right] E_0(Y \mid R = 1, X = x') \\
&= \sum_{x' \in \mathcal{X}} \left[-p_0(x') + 1(x' = x)\right] E_0(Y \mid R = 1, X = x') \\
&= -\sum_{x' \in \mathcal{X}} p_0(x') E_0(Y \mid R = 1, X = x') + \sum_{x' \in \mathcal{X}} 1(x' = x) E_0(Y \mid R = 1, X = x') \\
&= -\tau_0 + E_0(Y \mid R = 1, X = x).
\end{aligned}$$

Claim 3: $\lim_{\epsilon \to 0} \mathrm{III} = \frac{[y - E_0(Y|R=1, X=x)] \times r}{e_0(x)}$.

Proof in the case when X is discrete taking value in $\mathcal{X}$, with probability mass function $p_0(x)$ under F_0:

For a random draw from $F_\epsilon = (1-\epsilon)F_0 + \epsilon \delta_z$, define

$$\delta = \begin{cases} 0 & \text{if the random draw is from } F_0 \\ 1 & \text{if the random draw is from } \delta_z \end{cases}$$

then $\delta \sim \text{Bernoulli}(\epsilon)$.

For any $x' \in \mathcal{X}$,

$$\begin{aligned}
E_\epsilon(Y \mid R = 1, X = x') &= P_\epsilon(\delta = 1 \mid R = 1, X = x') \times \underbrace{E_\epsilon(Y \mid R = 1, X = x', \delta = 1)}_{= y, \text{ if } x = x', r = 1; \text{ otherwise undefined}} \\
&\quad + P_\epsilon(\delta = 0 \mid R = 1, X = x') \times \underbrace{E_\epsilon(Y \mid R = 1, X = x', \delta = 0)}_{= E_0^2(Y|R=1, X=x')},
\end{aligned}$$

where

$$\begin{aligned}
& & P_\epsilon(\delta = 1 \mid R = 1, X = x') \\
\text{(Bayes' rule)} & = & \frac{P_\epsilon(\delta = 1, R = 1, X = x')}{P_\epsilon(R = 1, X = x')} \\
& = & \frac{P_\epsilon(\delta = 1) \times P_\epsilon(R = 1, X = x' \mid \delta = 1)}{P_\epsilon(R = 1, X = x')} \\
& = & \frac{\epsilon 1(x' = x, r = 1)}{\epsilon 1(x' = x, r = 1) + (1 - \epsilon) p_0(x') e_0(x')} \\
& =: & \xi,
\end{aligned}$$

so

$$E_\epsilon(Y \mid R = 1, X = x') = \xi y + (1 - \xi) E_0(Y \mid R = 1, X = x')$$

(if $x \neq x'$ or $r \neq 1$, $E_\epsilon(Y \mid R = 1, X = x', \delta = 1)$ is undefined but $\xi = 0$, so the above equation still holds.)

Therefore

$$\begin{aligned}
\text{III} &= \frac{1}{\epsilon} E_0^1 \left(E_\epsilon^2 - E_0^2 \right) \\
&= \frac{1}{\epsilon} \sum_{x' \in \mathcal{X}} p_0(x') \left\{ \xi y + (1 - \xi) E_0(Y \mid R = 1, X = x') - E_0(Y \mid R = 1, X = x') \right\} \\
&= \frac{1}{\epsilon} \sum_{x' \in \mathcal{X}} p_0(x') \xi \left\{ y - E_0(Y \mid R = 1, X = x') \right\} \\
&= \sum_{x' \in \mathcal{X}} p_0(x') \frac{1(x' = x, r = 1) \left\{ y - E_0(Y \mid R = 1, X = x') \right\}}{\epsilon 1(x' = x, r = 1) + (1 - \epsilon) P_0(X = x') e_0(x')},
\end{aligned}$$

so

$$\begin{aligned}
\lim_{\epsilon \to 0} \text{III} &= \sum_{x' \in \mathcal{X}} p_0(x') \frac{1(x' = x, r = 1) \left\{ y - E_0(Y \mid R = 1, X = x') \right\}}{p_0(x') e_0(x')} \\
&= \frac{[y - E_0(Y \mid R = 1, X = x)] \times r}{e_0(x)}.
\end{aligned}$$

□

Corollary 2. *Assuming ignorability ($R \perp Y \mid X$), under the setup of Corollary 1, the Efficient Influence Function for $\tau = E(Y) = E_X\{E(Y \mid R = 1, X)\}$ is*

$$\varphi^{eff}(z) = -\tau_0 + E_0(Y \mid X = x) + \frac{[y - E_0(Y \mid X = x)] \times r}{e_0(x)}.$$

3.3 Doubly-Robust Estimator

3.3.1 Construction of the Estimator

Now we have the Efficient Influence Function for τ in the Two Phase Design, let's construct an estimator based on the EIF.

Theorem. *(Unbiased Estimating Equation, revisit)*

Suppose $Z_1, \ldots, Z_n$ are i.i.d. from some distribution F_θ, with $\theta \in \Theta \subset \mathbb{R}^k$. Suppose $g(z, \theta)$ is a k-dimensional vector-valued function satisfying

$$E_{F_\theta}\{g(Z, \theta)\} = 0, \qquad \text{for all } \theta \in \Theta.$$

Then under some regularity conditions, the estimator $\hat{\theta}$ that solves

$$\sum_{i=1}^{n} g(Z_i, \hat{\theta}) = 0$$

is a consistent estimator for θ, i.e.

$$\hat{\theta} \xrightarrow{P} \theta.$$

Denote $Z = (X, R, YR)$, assume it comes from some distribution F_τ with

$E(Y) = \tau$. Denote $\varphi^{\text{eff}}(z) \equiv \varphi^{\text{eff}}(z, \tau)$, and we have

$$\begin{aligned} E_{F_\tau}\left[\varphi^{\text{eff}}(Z, \tau)\right] &= E\left[-\tau + E(Y \mid X) + \frac{[Y - E(Y \mid X)] \times R}{e(X)}\right] \\ &= -\tau + E_X\left(E(Y \mid X)\right) + E_X\left\{\frac{[Y - E(Y \mid X)] \times R}{e(X)}\right\} \\ &= -\tau + E(Y) + E_X\left\{\underbrace{E\left(\frac{[Y - E(Y \mid X)] \times R}{e(X)} \mid X\right)}_{=0 \text{ (ignorability)}}\right\} \\ &= 0. \end{aligned}$$

Therefore, we can construct an estimator $\hat{\tau}_3$ that solves:

$$\sum_{i=1}^{n} \varphi^{\text{eff}}(Z_i, \hat{\tau}_3) = 0,$$

i.e. $\hat{\tau}_3$ is the solution to

$$\sum_{i=1}^{n}\left\{-\tau + E(Y \mid X_i) + \frac{[Y_i - E(Y \mid X_i)] \times R_i}{e(X_i)}\right\} = 0. \tag{2}$$

However, we don't know $E(Y \mid X)$ or $e(X)$ because we are not making any assumptions about the distribution of (X, R, Y) except for ignorability. Assume $\hat{E}(Y \mid X)$ and $\hat{e}(x)$ are working models (i.e. fitted models, but not necessarily correct models) for $E(Y \mid X)$ and $e(x)$, respectively. Plug them into (2), and it motivates the follwing estimator: $\hat{\tau}^{\text{DR}}$ is the solution to

$$\sum_{i=1}^{n}\left\{-\tau + \hat{E}(Y \mid X_i) + \frac{\left[Y_i - \hat{E}(Y \mid X_i)\right] \times R_i}{\hat{e}(X_i)}\right\} = 0, \tag{3}$$

i.e.

$$\hat{\tau}^{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{E}(Y \mid X_i) + \frac{\left[Y_i - \hat{E}(Y \mid X_i)\right] \times R_i}{\hat{e}(X_i)}\right\}.$$

3.3.2 Double Robustness

$\hat{\tau}^{\mathrm{DR}}$ is constructed by solving the estimating equation (3). By Theorem 2 (unbiased estimating equation), $\hat{\tau}^{\mathrm{DR}}$ will be consistent if the estimating equation is unbiased.

1) If $\hat{E}(Y \mid X) = E(Y \mid X)$:

$$\begin{aligned}
& E\left\{-\tau + \hat{E}(Y \mid X) + \frac{\left[Y - \hat{E}(Y \mid X)\right] \times R}{\hat{e}(X)}\right\} \\
= \;& -\tau + E\left\{E(Y \mid X)\right\} + E\left\{\frac{[Y_i - E(Y \mid X)] \times R}{\hat{e}(X)}\right\} \\
= \;& -\tau + \tau + E\left\{E\left(\frac{[Y - E(Y \mid X)] \times R}{\hat{e}(X)} \mid X\right)\right\} \\
\text{(by ignorability)} \quad = \;& E\left\{\frac{1}{\hat{e}(X)}\underbrace{E\left(Y - E(Y \mid X) \mid X\right)}_{=0} E(R \mid X)\right\} \\
= \;& 0.
\end{aligned}$$

2) If $\hat{e}(x) = e(x)$:

$$
\begin{aligned}
& E\left\{-\tau + \hat{E}(Y \mid X) + \frac{\left[Y - \hat{E}(Y \mid X)\right] \times R}{\hat{e}(X)}\right\} \\
= \; & -\tau + E\left\{\hat{E}(Y \mid X)\right\} + E\left\{\frac{\left[Y - \hat{E}(Y \mid X)\right] \times R}{e(X)}\right\} \\
= \; & -\tau + E\left\{\hat{E}(Y \mid X)\right\} + E\left\{E\left(\frac{\left[Y - \hat{E}(Y \mid X)\right] \times R}{e(X)} \mid X\right)\right\} \\
\text{(by ignorability)} \; = \; & -\tau + E\left\{\hat{E}(Y \mid X)\right\} + E\left\{\frac{1}{e(X)} E\left(\left[Y - \hat{E}(Y \mid X)\right] \mid X\right) \underbrace{E(R \mid X)}_{=e(X)}\right\} \\
= \; & -\tau + E\left\{\hat{E}(Y \mid X)\right\} + E\left[E(Y \mid X)\right] - E\left\{\hat{E}(Y \mid X)\right\} \\
= \; & 0.
\end{aligned}
$$

Corollary 3. *If either the regression model $\hat{E}(Y \mid X)$ or the propensity score model $\hat{e}(x)$ is correct, $\hat{\tau}^{DR}$ is consistent for τ. If both models are correct, $\hat{\tau}^{DR}$ is asymptotically the most efficient.*

4 Deductive Derivation of the Doubly-Robust Estimator

4.1 Review of the Above Estimation Procedure

1. Derive EIF (using Gâteaux derivative or some other approach):

$$
\begin{aligned}
\varphi^{\text{eff}}(z, \tau, F) &= \lim_{\epsilon \to 0} \frac{\tau\left\{(1-\epsilon)F + \epsilon\delta_z\right\} - \tau(F)}{\epsilon} \\
&= -\tau(F) + E(Y \mid X = x) + \frac{[y - E(Y \mid X = x)] \times r}{e(x)}
\end{aligned}
$$

2. Plug EIF into estimating equation:

$$\sum_{i=1}^{n} \left\{ -\tau + E(Y \mid X_i) + \frac{[Y_i - E(Y \mid X_i)] \times R_i}{e(X_i)} \right\} = 0$$

3. Plug in working model for the unknown part in the estimating equation, and solve for $\hat{\tau}$:

$$\sum_{i=1}^{n} \left\{ -\tau + \hat{E}(Y \mid X_i) + \frac{\left[Y_i - \hat{E}(Y \mid X_i)\right] \times R_i}{\hat{e}(X_i)} \right\} = 0$$

Motivation: EIF has been derived in some problems, but the derivation is complicated. In many problems, the form of EIF is still unknown (e.g. median(Y) in Two Phase Design). Can we automate the procedure?

4.2 Deductive Derivation

1. Specify a working distribution F_w as distribution for (X, R, YR):

   (a) $p_w(x)$: empirical distribution on the observed $X_1, \ldots, X_n$

   (b) $e_w(x) = P_w(R = 1 \mid X = x)$: logistic regression fit

   (c) $E_w(Y \mid R = 1, X = x)$: linear regression fit

   $$E_w(Y \mid R = 1, X = x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

2. Relax F_w by 1 degree of freedom, to get $F_w(\eta)$ for some tuning parameter η:

   Keep $p_w(x)$ and $e_w(x)$ as fitted, add δ to the intercept of the linear regression fit:

   $$E_w(Y \mid R = 1, X = x; \eta) = \eta + \hat{\beta}_0 + \hat{\beta}_1 x.$$

3. Have a program to compute $\tau\left(F_w(\eta)\right)$ and $\tau\left(F_{w(Z_i,\epsilon)}(\eta)\right)$, where $F_{w(Z_i,\epsilon)}(\eta)$ is the ϵ-comtamination of $F_w(\eta)$ by data point Z_i:

$$F_{w(Z_i,\epsilon)}(\eta) = (1-\epsilon)F_w(\eta) + \epsilon\delta_{Z_i}.$$

4. Numerically solve for $\hat{\eta}$ that makes the following estimating equation zero, for a pre-specified tiny ϵ:

$$\sum_{i=1}^{n} \frac{\tau\left(F_{w(Z_i,\epsilon)}(\eta)\right) - \tau\left(F_w(\eta)\right)}{\epsilon} = 0.$$

5. Output the (numerical) estimate $\tau\left(F_w(\hat{\eta})\right)$.

Reference:

- Deductive Derivation and Turing-Computerization of Semiparametric Efficient Estimation (2015), Biometrics, Frangakis et. al.