

# Two Phase Design and Efficient Influence Function Statistical Theory II

Tianchen Qian

November 3, 2015

## **Outline:**

1. Definition of Two Phase Design and the parameter of interest
2. Estimation with parametric assumption
  - (a) Correctly specified regression model
  - (b) Correctly specified propensity score model
3. Robust estimation motivated by efficient influence function
  - (a) Definition of influence function, Gâteaux derivative
  - (b) Efficient influence function for the parameter of interest in Two Phase Design
  - (c) Doubly-robust estimator
4. Numerical construction of a deductive estimator (research with Dr. Frangakis)

# 1 Definition of the Problem

**Definition 1.** (Two Phase Design)

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

1.  $(X_1, Y_1), \dots, (X_n, Y_n) \sim$  i.i.d.  $F$ .  $X_i$  is the covariate, and  $Y_i$  is the outcome;
2.  $Y_i$  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  $X_i$ .

*Remark.* Why is this called “Two Phase Design”? A picture for illustration.

*Remark.* The goal is to estimate  $\tau := E_F(Y)$ , i.e. the population mean of the outcome. This is a missing data problem. It is also an elementary problem in causal inference.

**Example 1.** (Randomized Clinical Trial)

**Example 2.** (Simplified double sampling design)

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, and we decide “who to find out” based on the dropout time  $X_i$ , i.e.  $R_i$  only depends on  $X_i$ . In this case we have  $R_i \perp Y_i \mid X_i$ .

**Theorem 1.** (*Identifiability*)

*Under the assumptions in Definition 1,*

$$E(Y) = E_X \{E(Y \mid X, R = 1)\}.$$

*Proof.* (Blackboard.)

□

*Remark.* This provides a way to convert the missing data problem ( $E(Y)$  where not every  $Y_i$  is observed) to the observed data ( $E(Y \mid X, R = 1)$  only involves  $Y_i$  with  $R_i = 1$ ).

## 2 Parametric Modeling

### 2.1 With correctly specified regression model

A natural idea is to use  $X_i$  to predict  $\hat{Y}_i$  for those missing  $Y_i$ . To formalize it:

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

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

is a consistent estimator for  $\tau$ .

**Example 3.** (linear regression;  $\hat{\tau}_1$  unbiased when  $\hat{E}(Y | X) = E(Y | X)$ . Blackboard.)

*Remark.*  $\hat{\tau}_1$  may be bad if the regression model  $\hat{E}(Y | X)$  is incorrect.

## 2.2 With correctly specified propensity score model

**Definition 2.** *Propensity score*  $e(x)$  is defined as

$$e(x) := P(R = 1 | X = x).$$

The second idea is to “weight” the data with observed  $Y$ , so that they “represent” the missing  $Y$ . To formalize it:

If we have a *correctly specified* propensity score model  $\hat{e}(x)$  for  $e(x)$ , then (under some convergence condition about  $\hat{e}(x) \rightarrow e(x)$ )

$$\hat{\tau}_2 = \frac{1}{n} \sum_{i=1}^n \frac{Y_i R_i}{\hat{e}(X_i)}$$

is a consistent estimator for  $\tau$ .

**Example 4.** (Intuition on weighting; logistic regression;  $\hat{\tau}_2$  unbiased when  $\hat{e}(x) = e(x)$ . Blackboard.)

*Remark.*  $\hat{\tau}_2$  may be bad if the propensity score model  $\hat{e}(x)$  is incorrect.

## 3 A More Robust Estimator

### 3.1 Preliminaries

**Theorem 2.** (*Unbiased Estimating Equation*)

Suppose  $X_1, \dots, X_n$  are i.i.d. from some distribution  $F_\theta$ , with  $\theta \in \Theta \subset \mathbb{R}^k$ . Suppose  $g(x, \theta)$  is a  $k$ -dimensional vector-valued function, which is continuously differentiable in  $x$  and  $\theta$  and satisfies

$$E_\theta \{g(X, \theta)\} = 0, \quad \text{for all } \theta \in \Theta.$$

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

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

is a consistent estimator for  $\theta$ , i.e.

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

**Example 5.** (Score equation. Blackboard.)

**Definition 3.** Suppose  $Z_1, \dots, Z_n$  are i.i.d. data points from some distribution  $F_\theta$ , with  $\theta \in \Theta \subset \mathbb{R}^k$ . An estimator  $\hat{\theta} = \hat{\theta}(Z_1, \dots, Z_n)$  for  $\theta$  is said to be *asymptotically linear*, if there exists  $\varphi(z)$  such that

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(Z_i) + o_P(1),$$

with  $E[\varphi(Z)] = 0$  and  $E[\varphi(Z)\varphi(Z)']$  finite and non-singular.  $\varphi(z)$  is called the *influence function* for the estimator  $\hat{\theta}$ .

**Example 6.** Consider  $Z_1, \dots, Z_n \sim$  i.i.d.  $N(\theta, 1)$ , and  $\hat{\theta} = \bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$ .

We have

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &= \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n Z_i - \theta \right) \\
&= \sqrt{n} \times \frac{1}{n} \left( \sum_{i=1}^n Z_i - n\theta \right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n (Z_i - \theta).
\end{aligned}$$

So we have  $\varphi(z) = z - \theta$ , and  $E[\varphi(Z)] = E[Z - \theta] = 0$ .

**Example 7.** Consider the Two Phase Design. Here a data point is  $Z_i = (X_i, R_i, Y_i R_i)$ . Estimand is  $\tau := E(Y)$ .

For the first estimator

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

*if regression model is correctly specified and under some convergence condition we have*

$$\begin{aligned}
\sqrt{n}(\hat{\tau}_1 - \tau) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \hat{E}(Y | X_i) - \tau \right) + o_P(1), \\
\text{so } \varphi((x, r, yr), \hat{\tau}_1) &= \hat{E}(Y | X = x) - \tau.
\end{aligned}$$

For the second estimator

$$\hat{\tau}_2 = \frac{1}{n} \sum_{i=1}^n \frac{Y_i R_i}{\hat{e}(X_i)},$$

*if propensity score model is correctly specified and under some convergence*

condition we have

$$\begin{aligned}\sqrt{n}(\hat{\tau}_2 - \tau) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{Y_i R_i}{\hat{e}(X_i)} - \tau \right) + o_P(1), \\ \text{so } \varphi((x, r, yr), \hat{\tau}_2) &= \frac{Y_i R_i}{\hat{e}(X_i)} - \tau.\end{aligned}$$

**Definition 4.** Suppose  $\tau$  is some estimand (a functional defined on a collection of distributions  $\mathcal{F}$ ), and  $F$  and  $G$  are two distributions (such that  $(1 - \epsilon)F + \epsilon G \in \mathcal{F} \forall \epsilon \in [0, 1]$ ). The *Gâteaux derivative* of  $\tau$  at  $F$  in the direction of  $G$  is defined by

$$L_F(\tau; G) = \lim_{\epsilon \rightarrow 0} \frac{\tau\{(1 - \epsilon)F + \epsilon G\} - \tau(F)}{\epsilon}.$$

*Remark.* From a mathematical perspective, the Gâteaux derivative is a generalization of the concept of a directional derivative to functional analysis.

From a statistical perspective, it represents the rate of change in an estimand (statistical functional) at a distribution  $F$ , upon a small amount of contamination by another distribution  $G$ .

**Example 8.**  $\tau = \text{mean}$ ,  $F = N(0, 1)$ ,  $G = N(c, 1)$ .  $\tau(F) = \text{mean}(F) = 0$ . For some  $\epsilon \in (0, 1)$ ,  $(1 - \epsilon)F + \epsilon G$  is a mixture of  $F$  and  $G$ :  $X \sim (1 - \epsilon)F + \epsilon G$  means that

$$\begin{cases} X \sim F & \text{with prob } (1 - \epsilon) \\ X \sim G & \text{with prob } \epsilon \end{cases}$$

So

$$\begin{aligned}\tau\{(1 - \epsilon)F + \epsilon G\} &= E(X) \quad \text{when } X \sim (1 - \epsilon)F + \epsilon G \\ &= (1 - \epsilon) \times 0 + \epsilon \times c \\ &= \epsilon c,\end{aligned}$$



and

$$\begin{aligned}L_F(\tau; G) &= \lim_{\epsilon \rightarrow 0} \frac{\tau\{(1-\epsilon)F + \epsilon G\} - \tau(F)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{\epsilon c - 0}{\epsilon} \\ &= c.\end{aligned}$$

$c$  is the rate of change in the mean, when perturbing  $N(0, 1)$  in the direction of  $N(c, 1)$ .

## 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.  $(X_1, Y_1), \dots, (X_n, Y_n) \sim$  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  $\hat{E}(Y \mid X) \rightarrow E(Y \mid X)$ )

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

is a consistent estimator for  $\tau$ .

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  $\rho$  between  $R_i$  and  $Y_i$  given  $X_i$ , and the point estimate and CI will become a function on  $\rho$ .

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

## 3.2 Efficient Influence Function for $\tau$ in Two Phase Design

**Theorem 3.** Consider an estimand  $\tau$ . 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  $\tau_0$  has the form

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

where  $\delta_z$  is a point mass on  $z$ .  $\varphi^{\text{eff}}$  is called the (nonparametric) Efficient Influence Function for the estimand  $\tau$ .

*Remark.* “Without additional assumptions” means we allow the distribution of  $(X, R, Y)$  to take any form. For example, we don’t require  $E(Y | X)$  to be a linear model; we don’t require  $P(R = 1 | X)$  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  $\tau = E_X \{E(Y | X, R = 1)\}$ , its Efficient Influence Function is

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

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

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

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

*Proof.* (Sketch)

By the above theorem, we have

$$\varphi^{\text{eff}}(z) = \lim_{\epsilon \rightarrow 0} \frac{\tau \{(1 - \epsilon)F_0 + \epsilon\delta_z\} - \tau(F_0)}{\epsilon}. \quad (1)$$

To simplify notation, define

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

and

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

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

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

**Claim 1:**  $\lim_{\epsilon \rightarrow 0} \text{I} = 0$ .

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

**Claim 2:**  $\lim_{\epsilon \rightarrow 0} \text{II} = -\tau_0 + E_0(Y | R = 1, X = x)$ .

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

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

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

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

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

where

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

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

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

so

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

□



**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*

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

### 3.3 Doubly-Robust Estimator

#### 3.3.1 Construction of the Estimator

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

**Theorem.** *(Unbiased Estimating Equation, revisit)*

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

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

*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  $\theta$ , i.e.*

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

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

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

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

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

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

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

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

i.e.

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

### 3.3.2 Double Robustness

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

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

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

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

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

**Corollary 3.** *If either the regression model  $\hat{E}(Y | X)$  or the propensity score model  $\hat{e}(x)$  is correct,  $\hat{\tau}^{DR}$  is consistent for  $\tau$ . 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):

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

2. Plug EIF into estimating equation:

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

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

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

**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  $X_1, \dots, X_n$
- (b)  $e_w(x) = P_w(R = 1 | X = x)$ : logistic regression fit
- (c)  $E_w(Y | R = 1, X = x)$ : linear regression fit

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

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

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

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

3. Have a program to compute  $\tau(F_w(\eta))$  and  $\tau(F_{w(Z_i, \epsilon)}(\eta))$ , where  $F_{w(Z_i, \epsilon)}(\eta)$  is the  $\epsilon$ -contamination 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  $\epsilon$ :

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

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

**Reference:**

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