Introduction to Treatment Effect Models

Yu-Chin Hsu

Institute of Economics
Academia Sinica

Introduction

- Many empirical questions in economics, finance and accounting are interested in the causal effects of programs or policies.

Goal of this talk

- Review treatment effect models.
  - Rubin causal model.
  - Identification and estimation results under unconfoundedness.
- Introduce some estimators that can be used in the research.
- Access the plausibility of the unconfoundedness assumption.
- Estimating Conditional Average Treatment Effects: Abrevaya, Hsu and Lieli (2015, JBES).

References:

- "Recent Developments in the Econometrics of Program Evaluation," by Imbens and Wooldridge (2009, Journal of Economic Literature).
- "Econometric Analysis of Cross Section and Panel Data," by Wooldridge (2010, Chapter 21).
- "Matching Methods in Practice: Three Examples," by Imbens (2015, Journal of Human Resources.)

Rubin Causal Model

- Treatment assignment:

$$D = \begin{cases} 1, & \text{if the individual receives treatment,} \\ 0, & \text{otherwise.} \end{cases}$$

- Potential outcomes:
  - $Y(0)$, the outcome that would be observed if the individual did not receive the treatment.
  - $Y(1)$, the outcome that would be observed if the individual received the treatment.
- Observe:
  - Treatment indicator: D.
  - Outcome of interest: $Y = DY(1) + (1 - D)Y(0)$.
  - A vector of covariates: X.

What is Treatment Effect?

Treatment Effect for the Whole Population:

- Causal effects of programs or policies.
- Interested in the relation between $g(Y(1))$ and $g(Y(0))$, where $g(\cdot)$ is a functional of random variables.
- Average treatment effects: $E[Y(1)] - E[Y(0)]$, where $g(Y) = E[Y]$.
- Quantile treatment effects: $F^{-1}_{Y(1)}(\tau) - F^{-1}_{Y(0)}(\tau)$, where $g(Y) = F^{-1}_{Y}(\tau)$ denotes the τ-th quantile of Y.
- We can consider other inequality measures such as Gini indexes.

What is Treatment Effect? (Cont'd)

Treatment Effect for the Treated Population:

- Causal effects of programs or policies for the treated individuals.
- Average treatment effects of the treated: $E[Y(1)\big|D=1]-E[Y(0)|D=1]$, where $g(Y)=E[Y\big|D=1]$.
- Quantile treatment effects of the treated: $F^{-1}_{Y(1)\big|D=1}(\tau)-F^{-1}_{Y(0)\big|D=1}(\tau)$.
- Treatment effects for the non-treated can be defined similarly.

Main Difficulty in Treatment Effect Model

- Only observe $Y = DY(1) + (1 - D)Y(0)$.
- Never observe both potential outcomes, $Y(1)$ and $Y(0)$.
- We have a missing variable problem.
- Extra conditions are needed for identification.

Two Types of Identifying Conditions:

1. Unconfoundedness assumption: $D \perp (Y(1), Y(0)) \big| X$. Selection-on-observable, ignorability, conditional independence, exogeneity.
   - Without covariates, unconfoundedness assumption reduces to $D \perp (Y(1), Y(0))$ which is the random experiment assumption.
2. Endogenous assignment with a valid binary instrument.

- Propensity score: $p(x) = P(D = 1|X = x)$.
- Overlap Assumption: $0 < \underline{p} \leq p(x) \leq \bar{p} < 1$.
- Overlap Assumption: The supports of $X|D = 1$ and $X|D = 0$ are the same.
- Why these two are equivalent?

Average Treatment Effects

- Average treatment effects (ATE): $\beta = E[Y(1) - Y(0)]$.
- Under unconfoundedness assumption and overlap assumption,

$$\beta = E_X\Big[E[Y|D=1,X] - E[Y|D=0,X]\Big]. \tag{1}$$

Proof

or

$$\beta = E\left[\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\right]. \tag{2}$$

Proof

Parametric Estimators

- Parametric Imputation Estimator based on (1):

$$\hat{\beta}_{imp} = \frac{1}{n} \sum_{i=1}^{n} \rho_1(X_i, \hat{\theta}_1) - \rho_0(X_i, \hat{\theta}_0),$$

  where $\rho_1(X_i, \theta_1)$ and $\rho_0(X_i, \theta_0)$ are parametric models for $\rho_1(x) = E[Y(1)|X = x]$ and $\rho_0(x) = E[Y(0)|X = x]$.
- For a parametric model $\rho_d(x) = \rho_d(x, \theta_1)$, θ_d can be estimated by

$$\hat{\theta}_d = \arg\min_{\theta \in \Theta} D_i^d (1 - D_i)^{1-d} (Y_i - \rho_d(x, \theta))^2.$$

Parametric Estimators (Cont'd)

- Traditionally, we estimate the following model:

$$Y = \alpha + \beta D + \gamma X + \epsilon, \tag{3}$$

  and use an OLS estimator to estimate β.
- This is equivalent to impose the following parametric models on $\rho_1(X)$ and $\rho_0(X)$:

$$\rho_1(X, \theta_1) = \alpha_1 + \gamma_1 X, \qquad \rho_0(X, \theta_0) = \alpha_0 + \gamma_0 X,$$

  with $\theta_1 = (\alpha_1, \gamma_1) = (\alpha + \beta, \gamma)$ and $\theta_0 = (\alpha_0, \gamma_0) = (\alpha, \gamma)$.
- OLS is equivalent to the following:

$$\hat{\theta}_d = (\hat{\alpha}_d, \hat{\gamma}_d) = \arg\min_{a,r} D_i^d (1 - D_i)^{1-d} (Y_i - a - rX_i)^2.$$

  while imposing a constraint: $\gamma_1 = \gamma_0$.
- In this model, $ATE = CATE(x)$ for all x, i.e., treatment effect is homogenous.

Parametric Estimators (Cont'd)

- Or, we estimate the following model:

$$Y = \alpha + \beta D + \gamma X + \eta DX + \epsilon, \tag{4}$$

- This is equivalent to impose the following parametric models on $\rho_1(X)$ and $\rho_0(X)$:

$$\rho_1(X, \theta_1) = \alpha_1 + \gamma_1 X, \qquad \rho_0(X, \theta_0) = \alpha_0 + \gamma_0 X,$$

  with $\theta_1 = (\alpha_1, \gamma_1) = (\alpha + \beta, \gamma + \eta)$ and $\theta_0 = (\alpha_0, \gamma_0) = (\alpha, \gamma)$.
- OLS is equivalent to the following:

$$\hat{\theta}_d = (\hat{\alpha}_d, \hat{\gamma}_d) = \arg\min_{a,r} D_i^d (1 - D_i)^{1-d} (Y_i - a - rX_i)^2.$$

- $CATE(x) = \beta + \eta x$ for all x. Therefore, the coefficient of η is interpret as the marginal effect of X on CATE.
- Treatment heterogeneity over covariate values is allowed.
- $ATE = \beta + \eta E[X]$.

Parametric Estimators (Cont'd)

- Parametric Inverse Probability Weighted Estimator based on (2):

$$\hat{\beta}_{ipw} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{p(X_i, \hat{\gamma})} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i, \hat{\gamma})},$$

  where $p(X_i, \gamma)$ is a parametric model for $p(x) = E[D = 1|X = x]$.
- For a parametric model, say Probit or Logit, $p(x) = p(x, \gamma)$, γ can be estimated by

$$\hat{\gamma} = \arg\min_{r \in \Gamma} D_i \log(p(X_i, r)) + (1 - D_i)\log(1 - p(X_i, r)).$$

Remarks on Parametric Estimators

- Implementation is easy.
- Asymptotics is easier to derive.
- Asymptotic normality follows standard arguments.
- To make inference, bootstrap method works.
- However, these estimators are subject to model misspecification resulting in inconsistent estimator.

Nonparametric IPW Estimator

- Proposed by Hirano, Imbens and Ridder (2003, HIR):

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \hat{p}(X_i)},$$

  where $\hat{p}(x)$ is a non-parametric estimator for $p(x)$.
- Nonparametric Imputation Estimators are available too. The results are similar to nonparametric IPW estimator, but I have been working on nonparametric IPW estimator, so I am more familiar with this method. Hence, the nonparametric imputation estimators will be briefly discussed later.

Nonparametric IPW Estimator (Cont'd)

- Under regularity conditions, $\sqrt{n}(\hat{\beta} - \beta) \sim N(0, \mathcal{V}_\beta)$, where $\mathcal{V}_\beta = E[\phi(Y, D, X)^2]$ with

$$\begin{aligned}\phi(Y, D, X) =& \frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)} - \beta \\ & - \Big(\frac{\rho_1(X)}{p(X)} + \frac{\rho_0(X)}{1-p(X)}\Big)\big(D - p(X)\big).\end{aligned}$$

- $\mathcal{V}_\beta$ is the semiparametric efficiency bound for β as shown by Hahn (1988).

How to nonparametrically estimate $p(x)$

- We suggest the Series logit estimator (SLE) proposed by HIR.
- Let $\phi = (\phi_1, ..., \phi_{d_x})' \in \mathbb{Z}_+^{d_x}$ be a d_x-dimensional vector of non-negative integers, where $\mathbb{Z}_+$ denotes the set of non-negative integers.
- Let $\{\phi(k)\}_{k=1}^{\infty}$ be a sequence including all distinct $\phi \in \mathbb{Z}_+^{d_x}$ such that $|\phi(k)|$ is non-decreasing in k and let $x^\phi = \prod_{j=1}^{d_x} x_j^{\phi_j}$.
- For any integer K, define $R^K(x) = (x^{\phi(1)}, ..., x^{\phi(K)})'$ as a vector of power functions.
- Let $\Lambda(a) = \exp(a)/(1+\exp(a))$ be the logistic cumulative distribution function (CDF).
- The SLE for $p(X_i)$ is defined as $\hat{p}(x) = \Lambda\big(R^K(x)'\hat{\pi}_K\big)$, where

$$\hat{\pi}_K = \arg\max_{\pi_k} \frac{1}{n} \sum_{i=1}^{n} D_i \cdot \log\left(\Lambda\big(R^K(X_i)'\pi_K\big)\right) + (1-D_i) \cdot \log\left(1 - \Lambda\big(R^K(X_i)'\pi_K\big)\right).$$

How to make inference?

- To make inference or construct confidence interval, we need a consistent estimator for $\mathcal{V}_\beta$.
- Let $\hat{\rho}_1(x)$ and $\hat{\rho}_0(x)$ be

$$\hat{\rho}_1(x) = R^K(x) \cdot \Big(\frac{1}{n}\sum_{i=1}^{n} R^K(X_i)' R^K(X_i)\Big)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} R^K(X_i)' \frac{D_i Y_i}{\hat{p}(X_i)},$$

$$\hat{\rho}_0(x) = R^K(x) \cdot \Big(\frac{1}{n}\sum_{i=1}^{n} R^K(X_i)' R^K(X_i)\Big)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} R^K(X_i)' \frac{(1-D_i) Y_i}{1-\hat{p}(X_i)},$$

  where $R^K(X_i)$'s are the same as SLE.
- Then a consistent estimator for $\mathcal{V}_\beta$ is given by

$$\begin{aligned}\widehat{\mathcal{V}}_\beta = & \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} - \hat{\beta} \\ & - \Big(\frac{\hat{\rho}_1(X_i)}{\hat{p}(X_i)} + \frac{\hat{\rho}_0(X_i)}{1-\hat{p}(X_i)}\Big)\big(D_i - \hat{p}(X_i)\big)\Big)^2.\end{aligned}$$

- Alternatively, one can use bootstrap.

Discussions on Nonparametric Estimators

- They are semiparametric efficient.
- They are not subject to model misspecification.
- However, they depend on various nonparametric estimators for conditional mean functions.
- For nonparametric estimations, there are tuning parameters, e.g. number of power series terms or bandwidth, to pick and the results can be sensitive to the choices of the tuning parameters.

Access the Unconfoundedness Assumption

- Unconfoundedness assumption is not testable (without further assumptions).
- However, there are indirect methods that we can use to access the plausibility of it.
- One idea is to estimate the treatment on a pseudo-outcome, a variable known to be unaffected by the treatment, i.e., the treatment effect should be zero.
- If the estimated treatment effect of this variable is close to zero, the unconfoundedness assumption is considered more plausible.
- "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," by Donald, Hsu and Lieli (2014, JBES) propose the first direct test for the unconfoundedness assumption under IV assumption.

Conditional Average Treatment Effects (CATE)

- "Estimating Conditional Average Treatment Effects," by Abrevaya, Hsu and Lieli (2015, JBES) introduce Conditional Average Treatment Effect (CATE) designed to capture the heterogeneity of a treatment effect across subpopulations.
- $CATE(x_1) = E[Y(1) - Y(0)|X_1 = x_1]$, where X_1 is a subset set of covariates X.
  - expected effect of smoking on birthweight as a function of X_1.
  - expected effect of smoking on birthweight for a mother randomly chosen from the subpopulation $X_1 = x_1$.
- Under unconfoundedness assumption,

$$CATE(x_1) = E\Big[\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\Big|X_1 = x_1\Big].$$

Proposed CATE Estimator (semiparametric)

- **Parametric propensity score estimator:** The estimator $\hat{\theta}_n$ of the propensity score model $p(x, \theta)$, $\theta \in \Theta \subset R^d$, $d < \infty$, satisfies $\sup_{x \in \mathcal{X}} |p(x, \hat{\theta}_n) - p(x, \theta)| = O_p(n^{-1/2})$ for any $\theta \in \Theta$.
- Estimator for $CATE(x_1)$ is

$$\widehat{CATE}_\theta(x_1) = \frac{\frac{1}{nh_1^\ell} \sum_{i=1}^n \left(\frac{D_i Y_i}{p(x, \hat{\theta}_n)} - \frac{(1-D_i) Y_i}{1-p(x, \hat{\theta}_n)} \right) K_1\left(\frac{X_{1i} - x_1}{h_1}\right)}{\frac{1}{nh_1^\ell} \sum_{i=1}^n K_1\left(\frac{X_{1i} - x_1}{h_1}\right)},$$

where $K_1(u)$ is a kernel function and h_1 is a bandwidth.

Proposed CATE Estimator (semiparametric) (Cont'd)

- Then

$$\begin{aligned}
&\sqrt{nh_1^\ell}(\widehat{CATE}_\theta(x_1) - CATE(x_1)) \\
=&\frac{1}{\sqrt{nh_1^\ell}} \frac{1}{f_{x_1}(x_1)} \sum_{i=1}^{n} \psi_\theta(X_i, Y_i, D_i) K_1\Big(\frac{X_{1i} - x_1}{h_1}\Big) + o_p(1) \\
\xrightarrow{d}&\mathcal{N}\left(0, \frac{\|K_1\|_2^2 \sigma_{\psi_\theta}^2(x_1)}{f_{x_1}(x_1)}\right),
\end{aligned}$$

where $\psi_\theta(x, y, d) = \frac{d(y-m_1(x))}{p(x)} - \frac{(1-d)(y-m_0(x))}{1-p(x)} - CATE(x_1)$.
$\sigma_{\psi_\theta}^2(x_1) = E[\psi_\theta^2(X, Y, D) | X_1 = x_1]$.

- In the paper, we also propose fully nonparametric estimator for CATE.
- All of the results for CATE extend to the following cases easily:
  - Conditional average treatment effects of the Treated (CATT).
  - Conditional local average treatment effects (CLATE).
  - Conditional average treatment effects of the Treated (CLATT).
- Lee, Okui and Whang (JAE, 2017). "Doubly Robust Uniform Confidence Band for the Conditional Average Treatment Effect Function".
- Fan, Hsu, Lieli and Zhang (JBES, forthcoming). "Estimation of Conditional Average Treatment Effects with High-Dimensional Data".

Application

Effect of a mother's smoking during pregnancy on baby's birthweight.

- Many estimates of the average effect, but
- we don't know very much about how heterogeneous the effect is across relevant subpopulations:
  - (how) does it depend on mother's age, education, family income, etc.
- Data: North Carolina Vital Statistics; all live births between 1988-2002
  - many of the mother's personal characteristics recorded.
  - information on mother's zip code$\Rightarrow$additional covariates.
  - **per capita income** in the mother's zip code serves as a proxy for family income.
- Focus on first time black mothers to save computation time.
  - 157,989 observations.

Is This a Relevant Problem?

Very much so. Low birth weight

- associated with high healthcare costs (direct and later);
- evidence that it adversely affects health, educational, and labor market outcomes later in life;
- potentially contributes to intergenerational persistence of socioeconomic inequality;

$\Rightarrow$ important to understand role of risk factors such as smoking.

Large applied economics literature: Abrevaya and Dahl (2008, JBES), Almond et al. (2005, QJE), Abrevaya (2006, JAE), Walker et al. (2009, SEJ), etc.

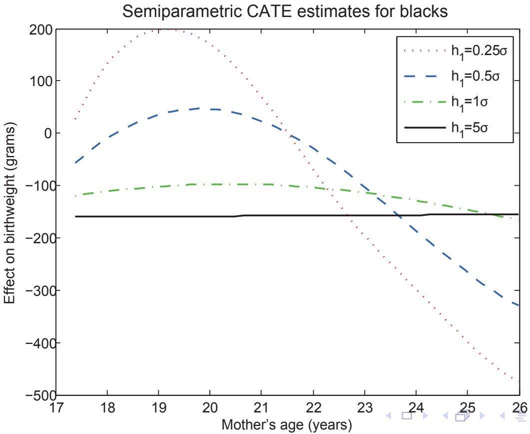
Identification

- Comparing average birthweight across smoking vs. non-smoking mothers likely does not identify causal effect (due to confounding factors).
- Assumption: all relevant confounding factors can be observed.
- Baseline specification:

$$
\begin{aligned}
X = & \text{[baby's gender, } \textbf{mother's age}\text{, marital status, educ,...} \\
& \text{prenatal care, zip location, per capita income]}
\end{aligned}
$$

- Some form of unconfoundedness often used: Almond et al. (2005), da Veiga and Wilder (2008), Walker et al. (2009).

CATE as a function of age: semiparametric results



Extensions

- Donald and Hsu (2014, JoE). "Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models."
- Hsu (2017, Econometrics Journal). "Consistent Tests for Conditional Treatment Effects."
- Hsu, Lai and Lieli (JBES, forthcoming). "Estimation and Inference for Counterfactual Treatment Effects."
- Hsu, Lee, Lai and Liao (2020, work in progress). "Testing Treatment Effect Monotonicity under Unconfoundedness Assumption."
- Treatment effect with High-Dimensional Data.
- Mediation analysis.
- Regression Discontinuity.

Thanks!

Proof of identification of ATE

$$E_x\Big[E\Big[Y|D=1,X\Big]\Big]=E_x\Big[E\Big[Y(1)\Big|D=1,X\Big]\Big]$$
$$=E_x\Big[E\Big[Y(1)\Big|X\Big]\Big]=E[Y(1)].$$

Back

Proof of identification of ATE

$$
\begin{aligned}
&E\left[\frac{DY}{p(X)}\right] = E_x\left[E\left[\frac{DY}{p(X)}\Big| X\right]\right] \\
=&E_x\left[p(X)E\left[\frac{DY}{p(X)}\Big| X, D=1\right] + (1-p(X))E\left[\frac{DY}{p(X)}\Big| X, D=0\right]\right] \\
=&E_x\left[E[Y(1)\Big| X, D=1]\right] = E_x\left[E[Y(1)\Big| X]\right] = E[Y(1)].
\end{aligned}
$$

Back