ECONOMETRICS II (ECO 2401)

Victor Aguirregabiria

Spring 2018

TOPIC 4: INTRODUCTION TO THE EVALUATION OF TREATMENT EFFECTS

1. Introduction and Notation

2. Randomized treatment

3. Conditional independence

4. Difference-in-Differences (a variant of the CI assumption)

5. Randomized eligibility: LATE

6. Regression Discontinuity Design

7. Roy's Model

1. INTRODUCTION.

• We are interested in estimating the **causal effect** of an explanatory variable D on an outcome variable Y.

• This setting is very general: effect of a drug on cholesterol level; effect of education of labor earnings; effect of price on demand; effect of a wage tax on employment; effect of a competition policy on firms' profits; etc, etc, etc.

• We consider a stylized but very general model where D is binary, $D \in \{0, 1\}$, and Y can be continuous or discrete, e.g., $D =$ University degree, $Y =$ Earnings.

• Following the language in this literature we denote D as the **treatment variable**, and Y is the **outcome variable**.

- $D = 1$ indicates that the subject "receives treatment" or "is in the treatment group", or " is in the experimental group";

- $D = 0$ indicates that the subject "does not receive treatment" or "is in the control group".

Example: A retail chain is interested in estimating the effect on demand of a 20% discount in the price of its key product. The firm decides to implement this price discount in some of its stores (experimental group) and keep the regular price in other stores (control group).

• Let Y_0 and Y_1 be latent variables that represent the outcome variable for an individual without and with treatment, respectively.

• We have an observation of the outcome variable Y per individual. Therefore, we observe:

$$Y = \begin{cases} Y_0 & \text{if} \quad D = 0 \\ Y_1 & \text{if} \quad D = 1 \end{cases} = (1 - D)\ Y_0 + D\ Y_1$$

• The **Treatment Effect** for an individual is:

$$TE \equiv Y_1 - Y_0$$

• Note: Even if we could observe the same individual with and without treatment, it would be at different moments [More on this below].

• Subjects are heterogeneous in multiple dimensions, and **Treatment Effects** can be very heterogeneous across individuals.

- The effect of a medicine drug varies substantially across patients;

- The effect of a university degree on earnings can be very different across individuals;

- The effect of a price reduction on demand can be substantially different across stores of the same chain.

- ...

• Ideally, we would like to estimate the TE of each individual. However, this is not feasible because we observe an individual either with or without treatment but not both.

• Under some conditions / restrictions, we will be able to estimate some features of the distribution of the TEs in the population of interest.

• A commonly used parameter that measures the aggregate effect of a treatment is the **Average Treatment Effect**, ATE.

$$ATE = \mathbb{E}(TE) = \mathbb{E}(Y_1 - Y_0)$$

• An the **Conditional Average Treatment Effect**.

$$ATE(x) = \mathbb{E}(Y_1 - Y_0 \mid X = x)$$

where X is a vector of predetermined attributes of the subject.

• $ATE(x)$ is the ATE for subpopulation of individuals with $X = x$.

Regression-like representation of the model

- Define $\mu_0 \equiv \mathbb{E}(Y_0)$ and $\mu_1 \equiv \mathbb{E}(Y_1)$ such that we can write:

$$\begin{cases} Y_0 &= \mu_0 + U_0 \\ \\ Y_1 &= \mu_1 + U_1 \end{cases}$$

where, by construction, $\mathbb{E}(U_0) = \mathbb{E}(U_1) = 0$.

- Note that, by definition, $ATE = \mu_1 - \mu_0$.

Regression-like representation of the model [2]

• Using these definitions, we have that:

$$\begin{aligned} Y &= (1-D)\ (\mu_0 + U_0) + D\ (\mu_1 + U_1) \\ &= \alpha + \beta\ D + e \end{aligned}$$

where $\alpha = \mu_0$, $\beta = \mu_1 - \mu_0 = ATE$, and

$$e = U_0 + (U_1 - U_0)D$$

• We will show below the Regression-Like representation of the model that includes the X variables.

Estimation of ATE and Endogeneity Problem

• Let $\{y_i, d_i, x_i : i = 1, 2, ..., N\}$ be a random sample of N individuals, some with treatment ($d_i = 1$), and others without treatment ($d_i = 0$).

• The researcher is interested in estimating these data to estimate ATE or/and $ATE(x)$.

• We now present two simple and intuitive estimators of the ATE:

- Difference in means estimator:

- OLS estimator of Y on D.

• We show that they are equivalent and, without further restrictions, they are inconsistent estimators of the ATE.

Estimation of ATE and Endogeneity Problem [2]

Difference in means estimator:

$$\widehat{ATE}_{DM} = \overline{y}_{D=1} - \overline{y}_{D=0}$$

with

$$\overline{y}_{D=1} = \frac{\sum_{i=1}^{N} y_i \; d_i}{\sum_{i=1}^{N} d_i} \quad \text{and} \quad \overline{y}_{D=0} = \frac{\sum_{i=1}^{N} y_i \; (1 - d_i)}{\sum_{i=1}^{N} (1 - d_i)}$$

OLS Estimator:

$$\widehat{ATE}_{OLS} = \widehat{\beta}_{OLS} = \frac{\sum_{i=1}^{N} (y_i - \overline{y}) \; (d_i - \overline{d})}{\sum_{i=1}^{N} (d_i - \overline{d})^2}$$

Equivalence of Difference-in-means and OLS of ATE

$$\widehat{ATE}_{OLS} = \frac{\sum_{i=1}^{N}(y_i - \overline{y})\ (d_i - \overline{d})}{\sum_{i=1}^{N}(d_i - \overline{d})^2}$$

- Note:

$$\begin{aligned}\sum_{i=1}^{N}\left(d_i - \overline{d}\right)^2 &= \sum_{i=1}^{N} d_i^2 - 2\sum_{i=1}^{N} d_i\overline{d} + \sum_{i=1}^{N}\overline{d}^2 \\ &= N\ \overline{d} - 2N\ \overline{d}^2 + N\ \overline{d}^2 \\ &= n\ \overline{d}(1-\overline{d})\end{aligned}$$

- And:

$$\sum_{i=1}^{N}\left(d_i - \overline{d}\right)\ (y_i - \overline{y}) = \sum_{i=1}^{N} d_i y_i - N\ \overline{d}\ \overline{y}$$

Equivalence of Difference-in-means and OLS of ATE [2]

- Therefore:

$$\begin{aligned}
\widehat{ATE}_{OLS} &= \frac{\sum_{i=1}^{N} d_i y_i - N\ \overline{d}\ \overline{y}}{N\ \overline{d}(1-\overline{d})} \\
&= \frac{1}{1-\overline{d}}\left(\frac{\sum_{i=1}^{N} d_i y_i}{\sum_{i=1}^{N} d_i} - \overline{y}\right) \\
&= \frac{1}{1-\overline{d}}\left(\overline{y}_{(D=1)} - \overline{y}\right)
\end{aligned}$$

- Note that:

$$\begin{aligned}
\overline{y} &= N^{-1} \sum_{i=1}^{N} d_i y_i + (1-d_i) y_i \\
&= \overline{d}\ \overline{y}_{(D=1)} + (1-\overline{d})\ \overline{y}_{(D=0)}
\end{aligned}$$

Equivalence of Difference-in-means and OLS of ATE [3]

- Thus,

$$\begin{aligned}\widehat{ATE}_{OLS} &= \frac{1}{1-\overline{d}}\left(\overline{y}_{(D=1)} - \overline{d}\,\overline{y}_{(D=1)} - (1-\overline{d})\,\overline{y}_{(D=0)}\right) \\ &= \overline{y}_{(D=1)} - \overline{y}_{(D=0)}\end{aligned}$$

Inconsistency of DM / OLS Estimators

• Is this estimator consistent? No, without further assumptions.

• It is clear that $\widehat{ATE}_{DM} \to_p \mathbb{E}(Y \mid D = 1) - \mathbb{E}(Y \mid D = 0)$, and if D is NOT independent of Y_0 and Y_1:

$$\begin{aligned} \mathbb{E}(Y \mid D = 1) - \mathbb{E}(Y \mid D = 0) &= \mathbb{E}(Y_1 \mid D = 1) - \mathbb{E}(Y_0 \mid D = 0) \\ &\neq \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = ATE \end{aligned}$$

• In Economics or social sciences, we expect the "choice of treatment" D to be correlated with the "effect of treatment" $Y_1 - Y_0$. Examples.

Inconsistency of DM / OLS Estimators [2]

- In the regression like representation of the model:

$$Y = \alpha + \beta\ D + e$$

where

$$e = U_0 + (U_1 - U_0)D$$

- Such that:

$$\mathbb{E}\,(D\ e) = \mathbb{E}\,(D\ U_0 + D(U_1 - U_0)) = \mathbb{E}\,(D\ U_1) \neq 0$$

2. RANDOMIZED TREATMENT

• Suppose that the treatment dummy D is independent of the latent outcome variables Y_0 and Y_1.

$$D \perp\!\!\!\perp (Y_0, Y_1)$$

where $\perp\!\!\!\perp$ represents "statistical independence".

• Given that $Y = (1 - D)\, Y_0 + D\, Y_1$ and $D \perp\!\!\!\perp (Y_0, Y_1)$:

$$\begin{cases} \mathbb{E}(Y \mid D = 0) = \mathbb{E}(Y_0 \mid D = 0) = \mathbb{E}(Y_0) \\ \\ \mathbb{E}(Y \mid D = 1) = \mathbb{E}(Y_1 \mid D = 1) = \mathbb{E}(Y_1) \end{cases}$$

such that:

$$ATE \equiv \mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y \mid D = 1) - \mathbb{E}(Y \mid D = 0)$$

and ATE is identified from data of $\{Y, D\}$

RANDOMIZED TREATMENT [2]

• We can construct root-N consistent estimators of $\mathbb{E}(Y \mid D = 1)$ and $\widehat{\mathbb{E}}(Y \mid D = 0)$ using:

$$\overline{y}_{D=1} = \frac{\sum_{i=1}^{N} y_i \; d_i}{\sum_{i=1}^{N} d_i} \quad \text{and} \quad \overline{y}_{D=0} = \frac{\sum_{i=1}^{N} y_i \; (1 - d_i)}{\sum_{i=1}^{N} (1 - d_i)}$$

• Then, a root-N consistent estimator of ATE is:

$$\widehat{ATE} = \overline{y}_{D=1} - \overline{y}_{D=0}$$

ENDOGENOUS TREATMENT

• The main concern in this literature is the **endogeneity of treatment**.

$$D \text{ is not independent of } TE = Y_1 - Y_0$$

• The assumption of $D \perp\!\!\!\perp (Y_0, Y_1)$ is equivalent to assume that treatment is perfectly randomized. This assumption is not plausible in most applications unless there is a **randomized experiment and all the individuals comply to their treatment assigment**.

• This may be a realistic condition in some randomized experiments in medical or natural science experiments, or even in lab experiments in experimental economics.

• However, it is quite **unrealistic in social sciences**, even in randomized field experiments in social sciences.

ENDOGENOUS TREATMENT [2]

• In general, treatment D is not independent of the potential outcomes Y_0 and Y_1. Individuals tend to self-select into treatment or not treatment according to their individual-specific benefits of treatment, i.e., according to Y_0 and Y_1.

• In field randomized experiments in social sciences, we typically can **randomize eligibility to treatment but not treatment itself**.

	Treatment	*No Treatment*
Eligible	**Compliers**	Not Compliers
Not Elligible	Not Compliers	**Compliers**

- In general:

  - Some subjects eligible to treatment choose not to take the treatment;

  - Some subjects not eligible decide to take an alternative but similar treatment.

Regression-like representation of the model [2]

***

- The OLS estimator of Y on D in the linear regression $Y = \alpha + \beta\ D + e$ is:

$$\widehat{\beta}_{OLS} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})\ (d_i - \overline{d})}{\sum_{i=1}^{n}(d_i - \overline{d})^2}$$

Regression-like representation of the model [2]

• Is the OLS estimator of β (i.e., the ATE) consistent?

• Consistency of the OLS requires $\mathbb{E}(D\ e) = 0$. Let's see that **this condition holds under randomized treatment.**

• Randomized treatment implies $D \perp\!\!\!\perp (U_0, U_1)$ and therefore $\mathbb{E}(U_0 \mid D) = \mathbb{E}(U_1 \mid D) = 0$.

$$\begin{aligned} \mathbb{E}(D\ e\ ) &= \mathbb{E}(D\ [U_0 + D(U_1 - U_0)]) \\ &= \Pr(D = 1)\ \mathbb{E}(U_1 \mid D = 1) = 0 \end{aligned}$$

Regression-like representation of the model [3]

• Without a randomized experiment, the unobservable component of the potential outcomes, U_0 and U_1, can be correlated with the treatment dummy D and this implies correlation between the error term e and the regressor D.

• The OLS estimator $\widehat{\beta}_{OLS} = \overline{y}_{D=1} - \overline{y}_{D=0}$ will be inconsistent.

3. CONDITIONAL INDEPENDENCE

• A weaker version of the assumption of independence between treatment and potential outcomes is that this independence holds only conditional on a vector of observable individual characteristics (control variables) X.

$$D \perp\!\!\!\perp (Y_0, Y_1) \mid X$$

• Given that $Y = (1 - D)\, Y_0 + D\, Y_1$ and $D \perp\!\!\!\perp (Y_0, Y_1) \mid X$:

$$\begin{cases} \mathbb{E}(Y \mid D = 0,\ X = x) = \mathbb{E}(Y_0 \mid D = 0,\ X = x) = \mathbb{E}(Y_0 \mid X = x) \\ \mathbb{E}(Y \mid D = 1,\ X = x) = \mathbb{E}(Y_1 \mid D = 1,\ X = x) = \mathbb{E}(Y_1 \mid X = x) \end{cases}$$

such that:

$$ATE(x) \equiv \mathbb{E}(Y_1 - Y_0 \mid X = x) = \mathbb{E}(Y \mid D = 1,\ X = x) - \mathbb{E}(Y \mid D = 0,\ X = x$$

and the conditional ATE(x) is identified. Then, we can also identify:

$$ATE = \mathbb{E}_X(\ ATE(x)\ )$$

CONDITIONAL INDEPENDENCE [2]

• **Estimation:** With conditional independence $D \perp\!\!\!\perp (Y_0, Y_1) \mid X$ but without unconditional independence $D \not\perp\!\!\!\perp (Y_0, Y_1)$, the estimator $\widehat{ATE} = \overline{y}_{D=1} - \overline{y}_{D=0}$ of the ATE is inconsistent.

• To estimate consistently the ATE we need to condition on X and estimate first the conditional ATE(x).

• Is X is a vector of discrete random variables (and our sample is relatively large), we can estimate ATE(x) using **frequency estimators** to estimate $\mathbb{E}(Y \mid D = 1,$ and $\mathbb{E}(Y \mid D = 0,\ X = x)$:

$$\widehat{ATE}(x) = \overline{y}_{D=1}(x) - \overline{y}_{D=0}(x)$$

with

$$\begin{aligned}
\overline{y}_{D=1}(x) &= \frac{\sum_{i=1}^{N} y_i \; d_i \; 1\{x_i = x\}}{\sum_{i=1}^{N} d_i \; 1\{x_i = x\}} \\
\overline{y}_{D=0}(x) &= \frac{\sum_{i=1}^{N} y_i \; (1 - d_i) \; 1\{x_i = x\}}{\sum_{i=1}^{N} (1 - d_i) \; 1\{x_i = x\}}
\end{aligned}$$

CONDITIONAL INDEPENDENCE [3]

• If X contains continuous variables (or if our sample is not so large) we can estimate ATE(x) using **Kernel Estimators** to estimate $\mathbb{E}\left(Y \mid D = 1,\ X = x\right)$ and $\mathbb{E}\left(Y \mid D = 0,\ X = x\right)$:

$$\widehat{ATE}(x) = \overline{y}_{D=1}(x) - \overline{y}_{D=0}(x)$$

with

$$\overline{y}_{D=1}(x) = \frac{\sum_{i=1}^{N} y_i\ d_i\ K\left(\dfrac{x_i - x}{b_N}\right)}{\sum_{i=1}^{N} d_i\ K\left(\dfrac{x_i - x}{b_N}\right)}$$

$$\overline{y}_{D=0}(x) = \frac{\sum_{i=1}^{N} y_i\ (1 - d_i)\ K\left(\dfrac{x_i - x}{b_N}\right)}{\sum_{i=1}^{N} (1 - d_i)\ K\left(\dfrac{x_i - x}{b_N}\right)}$$

Regression-like representation under CI

• Define $\mu_0(x) \equiv \mathbb{E}(Y_0 \mid X = x)$ and $\mu_1(x) \equiv \mathbb{E}(Y_1 \mid X = x)$ such that we can write:

$$\begin{cases} Y_0 &= \mu_0(x) + U_0 \\ \\ Y_1 &= \mu_0(x) + U_1 \end{cases}$$

where, by construction, $\mathbb{E}(U_0 \mid X = x) = \mathbb{E}(U_1 \mid X = x) = 0$.

• Note that, by definition, $ATE(x) = \mu_1(x) - \mu_0(x)$.

Regression-like representation under CI [2]

- Taking into account that $Y = (1 - D)Y_0 + D\ Y_1$:

$$\begin{aligned} Y &= (1 - D)\ \ (\mu_0(X) + U_0) + D\ \ (\mu_1(X) + U_1) \\ &= \alpha(X) + \beta(X) * D + e \end{aligned}$$

where:

$$\begin{aligned} \alpha(X) &= \mu_0(X) \\ \beta(X) &= ATE(X) \\ e &= U_0 + D\,(U_1 - U_0) \end{aligned}$$

Regression-like representation under CI [3]

• Under the CI assumption, the OLS estimation of $\beta(x)$ in this regression model provides a consistent estimator of the $ATE(x)$.

$$Y = \alpha(X) + \beta(X) * D + e$$

This is because, under the CI Assumption we have that:

$$\mathbb{E}(e \mid X, D = 0) = \mathbb{E}(U_0|X, D = 0) = \mathbb{E}(U_0) = 0$$

$$\mathbb{E}(e \mid X, D = 1) = \mathbb{E}(U_1|X, D = 1) = \mathbb{E}(U_1) = 0$$

• We can apply (nonparametric) Least Squares to estimate consistently $ATE(X)$.

Regression-like representation under CI [4]

• Suppose that $\alpha(x)$ and $\beta(x)$ are well approximated by a polynomial of order q in x: When x is a scalar:

$$y_i = \left[\alpha_0 + \alpha_1 x_i + ... + \alpha_q x_i^q\right] + \left[\beta_0 + \beta_1 x_i + ... + \beta_q x_i^q\right] d_i + e_i$$

• We can estimate parameters $\alpha's$ and $\beta's$ by OLS and the construct the estimate of the ATE(x):

$$\widehat{ATE}(x) = \widehat{\beta}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + ... + \widehat{\beta}_q x^q$$

Curse of dimensionality in NP estimation of ATE(x)

• The Kernel and Polynomial series estimators of ATE(x) suffer of the well-known curse of dimensionality in NP estimator. The speed of convergence of $\widehat{ATE}(x)$ to the true $ATE(x)$ declines with the number of continuous explanatory variables in the vector X. The estimator can be very imprecise unless we have very large samples. When X is discrete, these estimators have good asymptotic properties, but we still need sufficient observations for each discrete value of x.

• A possible approach is to construct an estimate of the unconditional ATE given the estimates of ATE(x):

$$\widehat{ATE} = \frac{1}{N} \sum_{i=1}^{N} \widehat{ATE}(x_i)$$

Curse of dimensionality in NP estimation of ATE(x) [2]

• This estimator $\widehat{ATE}$ is root-N consistent and asymptotically normal (Newey, ET 1994) despite $\widehat{ATE}(x)$ have lower speed of convergence due to continuous regressors.

• However, in some applications we are interested in the conditional ATE(x).

• Furthermore, even if we are interested only in unconditional ATE, the finite sample properties of the previous estimator $\widehat{ATE}$ are affected by the poor and imprecise estimates $\widehat{ATE}(x_i)$.

• Rosenbaum and Rubin (Biometrika, 1983) provide an interesting and useful approach to deal with this curse of dimensionality in the NP estimation of ATE.

Rosenbaum and Rubin (1983) Matching estimator using the Propensity Score

• Since D is a binary variable, its distribution conditional on $X = x$ is Bernoulli with probability $P(x)$ where:

$$P(x) \equiv \mathsf{Pr}\left(D = 1 \mid X = x\right)$$

In the TE literature, $P(x)$ is denoted the *Propensity Score*.

• Note that $P(x)$ contains all the information in the distribution of D conditional on $X = x$. Therefore, if D is independent of (Y_0, Y_1) conditional on X, then it should be also true that D is independent of (Y_0, Y_1) conditional on $P(X)$.

$$D \perp\!\!\!\perp (Y_0, Y_1) \mid P(X)$$

Matching estimator using the Propensity Score [2]

• Define $\widetilde{\mu}_0(p) \equiv \mathbb{E}(Y_0 \mid P(X) = p)$ and $\widetilde{\mu}_1(p) \equiv \mathbb{E}(Y_1 \mid P(X) = p)$ such that we can write:

$$\begin{cases} Y_0 &= \widetilde{\mu}_0(p) + U_0 \\ \\ Y_1 &= \widetilde{\mu}_1(p) + U_1 \end{cases}$$

where, by construction, $\mathbb{E}(U_0 \mid P(x)) = \mathbb{E}(U_1 \mid P(x)) = 0$.

• Note that, by definition, $ATE(p) = \widetilde{\mu}_1(p) - \widetilde{\mu}_0(p)$.

Matching estimator using the Propensity Score [3]

• The CI assumption implies that $ATE(p)$ is identified as:

$$\begin{aligned} ATE(p) &= \mathbb{E}(Y_1 \mid P(X) = p) - \mathbb{E}(Y_0 \mid P(X) = p) \\ &= \mathbb{E}(Y \mid D = 1,\ P(X) = p) - \mathbb{E}(Y \mid D = 0,\ P(X) = p) \end{aligned}$$

• Based on this insight, Rosenbaum and Rubin proposed the following estimator of the ATE. Let $\widehat{p}_i \equiv \widehat{P}(x_i)$ be a consistent estimator of the propensity score for individual i. Then:

$$\widehat{ATE} = \frac{1}{N} \sum_{i=1}^{N} \widehat{ATE}(\widehat{p}(x_i))$$

where

$$\widehat{ATE}(p) = \overline{y}_{D=1}(p) - \overline{y}_{D=0}(p)$$

Matching estimator using the Propensity Score [4]

with:

$$\overline{y}_{D=1}(p) = \frac{\sum_{i=1}^{N} y_i \; d_i \; K\left(\frac{\widehat{p}_i - p}{b_N}\right)}{\sum_{i=1}^{N} d_i \; K\left(\frac{\widehat{p}_i - p}{b_N}\right)}$$

$$\overline{y}_{D=0}(p) = \frac{\sum_{i=1}^{N} y_i(1 - d_i) \; K\left(\frac{\widehat{p}_i - p}{b_N}\right)}{\sum_{j=1}^{N}(1 - d_i) \; K\left(\frac{\widehat{p}_i - p}{b_N}\right)}$$

and $\widehat{p}_i = \widehat{p}(x_i) = \sum_{j=1}^{N} d_j \; K\left(\frac{x_j - x_i}{b_N}\right) / [\sum_{j=1}^{N} K\left(\frac{x_j - x_i}{b_N}\right)]$.

• Now, the dimension of the conditioning variables in the estimation of the conditional expectations $\overline{y}_{D=1}(p)$ and $\overline{y}_{D=0}(p)$ is 1 (the propensity score)

instead of $\mathbf{dim}(X)$. This improves the asymptotic and finite sample properties of the estimators of ATE.

4. DIFFERENCES-IN-DIFFERENCES (DiD)

- DiD is a particular case of Conditional Indipendence when we have

  (1) Panel Data;

  (2) A particular structure of the Treatment variable D;

  (3) An assumption about the components structure of the unobservables U_0, U_1

- Suppose that we have panel data $\{y_{it}, d_{it}, x_{it}\}$ for $t = 1,...,T$, with $T \geq 2$.

- The treatment dummy $D_{it} \in \{0, 1\}$ has the following structure:

$$D_{it} = T_i \times 1\{t \geq t^*\}$$

- $T_i \in \{0, 1\}$ is the dummy that indicates that an individual i belongs to the experimental group.

- $1\{t \geq t^*\}$ is the dummy that indicates that period t is a period of treatment.

DIFFERENCES-IN-DIFFERENCES [2]

• The model is the same as before but for panel data: $Y_{0,it}$ and $Y_{1,it}$ are the latent variables that represent the outcome variable for an individual without and with treatment, respectively. And have that:

$$\begin{cases} Y_{0,it} &= \mu_0 + U_{0,it} \\ \\ Y_{1,it} &= \mu_1 + U_{1,it} \end{cases}$$

with $\mu_0 \equiv \mathbb{E}(Y_{0,it})$ and $\mu_1 \equiv \mathbb{E}(Y_{1,it})$

• The model is completed with an assumption about the component structure of $U_{0,it}$ and $U_{1,it}$:

$$U_{0,it} = \eta_i + \gamma_{0t} + u_{0it}$$

$$U_{1,it} = \eta_i + \gamma_{1t} + u_{1it}$$

Note $\eta_{0i} = \eta_{1i}$. Key restriction.

DIFFERENCES-IN-DIFFERENCES [3]

- Model:

$$Y_{it} = (1 - D_{it})\, Y_{0,it} + D_{it} Y_{1,it}$$

that we can represent as:

$$Y_{it} = \alpha + \beta\ D_{it} + \left[U_{0,it} + D_{it}\ \left(U_{1,it} - U_{0,it}\right)\right]$$

where $\alpha = \mu_0$, $\beta \equiv \mu_1 - \mu_0 = ATE$, and

$$U_{0,it} + D_{it}\ \left(U_{1,it} - U_{0,it}\right) =$$

$$\eta_i + \gamma_{0t} + u_{0it} + D_{it}\ (\gamma_{1t} - \gamma_{0t} + u_{1it} - u_{0it})$$

DIFFERENCES-IN-DIFFERENCES [4]

• The DiD estimator is simply the OLS estimator in the equation in first-differences when we include time-dummies:

$$\Delta Y_{it} = \beta\ \Delta D_{it} + \widetilde{\gamma}_t\ TD_t + \Delta e_{it}$$

• Note that:

$$\Delta D_{it} = \begin{cases} 0 & \text{for} \quad t < t^* \text{ or } t > t^* \\ \\ T_i & \text{for} \quad t = t^* \end{cases}$$

Therefore, the model has information about β only at $t = t^*$.

$$\Delta Y_{it^*} = \beta\ T_i + \widetilde{\gamma}_{t^*} + \Delta e_{it^*}$$

DIFFERENCES-IN-DIFFERENCES [5]

- And according to the model:

$$\begin{aligned} \Delta e_{it^*} &= u_{0it^*} - u_{0it^*-1} + u_{1it^*} - u_{0it^*-} \\ &= u_{1it^*} - u_{0it^*-1} \end{aligned}$$

- Consistency os the DiD estimator requires:

  T_i is independent of the transitory shocks $u_{1it^*} - u_{0it^*-1}$

  More importantly, the error component restriction: $\eta_{0i} = \eta_{1i}$

5. RANDOMIZED ELIGIBILITY TO TREATMENT

• Let $Z \in \{0, 1\}$ be a random variable that represents whether the individual is eligible to treatment ($Z = 1$) or not ($Z = 0$). This Z variable comes from a randomized experiment.

• In general, in field experiments in the social sciences, $Z \neq D$.

- We observe subjects with $\{z_i = 1$ and $d_i = 0\}$: eligible but not taking the treatment;

- We observe (os suspect) subjects with $\{z_i = 0$ and $d_i = 1\}$: non-eligible but taking a similar / alternative treatment.

• Using Z as a proxy for D generates an inconsistent estimator [See Below]

• However, we show below that, under some additional assumptions, Z can be used as an instrument for D.

• This IV estimator is not a consistent estimator for the ATE for all the population.

• However, thsi IV estimator is a consistent estimator of the ATE for a particular subpopulation of subjects: the compliers.

IV Estimator

- For $z = 0, 1$, let $P(z)$ be the propensity score $P(z) \equiv \mathsf{Pr}(D = 1 \mid Z = z)$.

- Consider the following assumptions on the instrument Z.

  [Independence] Z is independent of potential outcomes (Y_0, Y_1);

  [Relevance] Z is correlated with treatment, i.e., $P(1) > P(0)$.

- Consider the regression-like representation of the model:

$$Y = \alpha + \beta \; D + e$$

- The IV estimator of the ATE is:

$$\widehat{\beta}_{IV} = \left[\sum_{i=1}^{N} (z_i - \overline{z}) \left(d_i - \overline{d} \right) \right]^{-1} \left[\sum_{i=1}^{N} (z_i - \overline{z}) (y_i - \overline{y}) \right]$$

Wald Estimator

• Wald Estimator is defined as:

$$\widehat{\beta}_{Wald} = \frac{\overline{y}_{Z=1} - \overline{y}_{Z=0}}{\overline{d}_{Z=1} - \overline{d}_{Z=0}}$$

where $\overline{y}_{Z=1}$ and $\overline{d}_{Z=1}$ are the sample means of Y and D, respectively, for the subsample of observations with $Z = 1$, and similarly, $\overline{y}_{Z=0}$ and $\overline{d}_{Z=0}$ are the sample means of Y and D for the subsample of observations with $Z = 0$.

• We can show that for this model, **the IV and the Wald estimators are the same**.

$$\begin{aligned}
\widehat{\beta}_{IV} &= \frac{\sum_{i=1}^{N} z_i \left(y_i - \overline{y}\right)}{\sum_{i=1}^{N} z_i \left(d_i - \overline{d}\right)} = \frac{\sum_{i=1}^{N} z_i y_i - N_1 \, \overline{y}}{\sum_{i=1}^{N} z_i d_i - N_1 \, \overline{d}} = \frac{N_1 \left(\overline{y}_1 - \overline{y}\right)}{N_1 \left(\overline{d}_1 - \overline{d}\right)} \\
&= \frac{\frac{N_1 \, N_0}{N} \left(\overline{y}_{Z=1} - \overline{y}_{Z=0}\right)}{\frac{N_1 \, N_0}{N} \left(\overline{d}_{Z=1} - \overline{d}_{Z=0}\right)} = \frac{\overline{y}_{Z=1} - \overline{y}_{Z=0}}{\overline{d}_{Z=1} - \overline{d}_{Z=0}} = \widehat{\beta}_{Wald}
\end{aligned}$$

Inconsistency of IV (Wald) Estimator for ATE

• In general, this IV is NOT a consistent estimator of the ATE.

• Though the instrument Z is independent of U_0 and U_1, it is correlated with the error term $e = U_0 + D(U_1 - U_0)$.

$$\begin{aligned} \mathbb{E}(e \mid Z = 0) &= \mathbb{E}(U_0 + D(U_1 - U_0) \mid Z = 0) \\ &= P(0)\, \mathbb{E}(U_1 - U_0 \mid D = 1) \end{aligned}$$

And,

$$\begin{aligned} \mathbb{E}(e \mid Z = 1) &= \mathbb{E}(U_0 + D(U_1 - U_0) \mid Z = 1) \\ &= P(1)\, \mathbb{E}(U_1 - U_0 \mid D = 1) \end{aligned}$$

• Such that:

$$\mathbb{E}(e \mid Z = 1) - \mathbb{E}(e \mid Z = 0) = [P(1) - P(0)]\ \mathbb{E}(U_1 - U_0 \mid D = 1) \neq 0$$

- For the IV estimator to be consistent, we need $\mathbb{E}(Z\ e) = 0$.

- Note that $\mathbb{E}(Z\ e) = \mathsf{Pr}(Z = 1)\ \mathbb{E}(e \mid Z = 0) + \mathsf{Pr}(Z = 0)\ \mathbb{E}(e \mid Z = 0)$.

- Given that $\mathbb{E}(e \mid Z = 1) - \mathbb{E}(e \mid Z = 0) = [P(1) - P(0)]\ \mathbb{E}(U_1 - U_0 \mid D = 1)$, note that

$$\mathbb{E}(Ze) = [\mathsf{Pr}(Z = 0) + \mathsf{Pr}(Z = 1)\ \ [P(1) - P(0)]]\ \ \mathbb{E}(U_1 - U_0|D = 1)$$

that in general is different to zero.

Inconsistency of IV in Random Coefficients Models with Endogeneity

• More generally, in models with random coefficients and endogenous variables, the error term of the regression model includes interactions between the random coefficient and the endogenous variable.

• In these models, IV estimation does not provide a consistent estimator of the average coefficient.

• Consider the model:

$$Y_i = X_i \, \beta_i + \varepsilon_i \qquad \text{with } \beta_i = \beta + v_i$$

such that

$$Y_i = X_i \, \beta + e_i \qquad \text{with } e_i = \varepsilon_i + X_i v_i$$

where X_i is correlated with v_i, but there is a vector of instruments Z_i that is independent of ε_i and v_i.

• The IV estimator

$$\widehat{\beta}_{IV} = \left(\sum_{i=1}^{N} z_i' x_i \right)^{-1} \left(\sum_{i=1}^{N} z_i' y_i \right)$$

is an asymptotically biased estimator of β.

• The reason is simple: despite Z_i is independent of ε_i and v_i, it is not independent of $e_i = \varepsilon_i + X_i v_i$.

LOCAL AVERAGE TREATMENT EFFECT (LATE)

• Though the IV estimator is an inconsistent estimator of ATE (when we have heterogeneous treatment effects), **under some conditions (Monotonicity)**, the IV is a consistent estimator of the the ATE for a subpopulation of individuals: **the Compliers**.

• To understand some assumptions of the model and some properties of the estimators, it is useful to define the following latent variables:

$D_0 =$ Treatment indicator under the hypothetical case that individual were not eligible, i.e., when $Z = 0$;

$D_1 =$ Treatment indicator under the hypothetical case that individual were eligible, i.e., when $Z = 1$.

- D_0 and D_1 are unobservable. All what we observe is the treatment D.

$$D = (1 - Z)\ D_0 + Z\ D_1$$

LATE [2]

- According to these latent variables, we can define

	$D_0 = 0$	$D_0 = 1$
$D_1 = 0$	Never Takers	Defiers
$D_1 = 1$	**Compliers**	Always Takers

[Assumption: Monotonicity] For every individual, $D_1 \geq D_0$, i.e., there are not defiers.

LATE [3]

• Using the definitions of "individual types" above, the assumption of Monotonicity establishes that **there are not "Defiers"** in the population.

• Under the assumptions of Independence, Relevance, and Monotonicity, the IV estimator converges in probability to the Local Average Treatment Effect parameter defined as

$$LATE \equiv E\left(Y_1 - Y_0 \mid D_1 > D_0\right).$$

• **LATE is the ATE for the subpopulation of Compliers.**

Proof IV is a consistent estimator of LATE

- As we have shown before, the IV and the Wald estimator are the same:

$$\widehat{\beta}_{IV} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{d}_1 - \overline{d}_0}$$

- By the LLN, $\widehat{\beta}$ converges in probability to $\dfrac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$.

- Now, we show that, under the Monotonicity assumption,

$$\frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} = E\left(Y_1 - Y_0 \mid D_1 > D_0\right) = LATE$$

Proof IV is a consistent estimator of LATE [2]

• Note that $Y = Y_0 + D\,(Y_1 - Y_0)$, and $D = D_0 + Z\,(D_1 - D_0)$. Therefore, (by independence of Z with (Y_0, Y_1, D_0, D_1)):

$$\begin{aligned} E(Y|Z=1) &= E(Y_0 + D_1(Y_1 - Y_0) \mid Z=1) \\ &= E(Y_0 + D_1(Y_1 - Y_0)\ ) \end{aligned}$$

• And

$$\begin{aligned} E(Y|Z=0) &= E(Y_0 + D_1(Y_1 - Y_0)\ |Z=0) \\ &= E(Y_0 + D_0(Y_1 - Y_0)\ ) \end{aligned}$$

Proof IV is a consistent estimator of LATE [3]

- Therefore, the numerator of the PLIM of IV is:

$$\begin{aligned} \text{Numerator of PLIM of IV} &= E(Y_0 + D_1(Y_1 - Y_0)\ ) - E(Y_0 + D_0(Y_1 - Y_0)\ ) \\ &= E((D_1 - D_0)\ (Y_1 - Y_0)\ ) \end{aligned}$$

- By the monotonicity assumption, $(D_1 - D_0)$ can be only 0 or 1. Therefore,

$$\text{Numerator of PLIM of IV} = \Pr(D_1 - D_0 > 0)\ E(Y_1 - Y_0 \mid D_1 - D_0 > 0)$$

Proof IV is a consistent estimator of LATE [4]

• Similarly, for the denominator of the PLIM of IV we have that (by independence of Z with (D_0, D_1))

$$\begin{aligned} E(D|Z=1) &= E(D_0 + Z(D_1 - D_0)\,|Z=1) \\ &= E(D_1) \end{aligned}$$

And (by independence of Z with (D_0, D_1))

$$\begin{aligned} E(D|Z=0) &= E(D_0 + Z(D_1 - D_0)\,|Z=0) \\ &= E(D_0) \end{aligned}$$

Proof IV is a consistent estimator of LATE [5]

• The denominator of the PLIM of IV is:

$$\text{Denominator of PLIM of IV} \;=\; E(D_1 - D_0)$$

• Again, by the monotonicity assumption, $(D_1 - D_0)$ can be only 0 or 1, such that $E(D_1 - D_0) = \text{Pr}(D_1 - D_0 > 0)$. Therefore,

$$\begin{aligned}\text{PLIM of IV} \;&=\; \frac{\text{Pr}(D_1 - D_0 > 0)\; E(Y_1 - Y_0 \mid D_1 - D_0 > 0)}{\text{Pr}(D_1 - D_0 > 0)} \\ &=\; E(Y_1 - Y_0 \mid D_1 - D_0 > 0) = LATE\end{aligned}$$

What if Monotonicity does not hold? What is the plim of the IV?

External Validity of LATE

• How different is the LATE to the ATE? Can we apply the LATE (ATE of compliers) to the rest of the population (Always Takers and Never Takers).

• In general, we cannot. However, if the proportion of compliers in the population is large (e.g., $> 80\%$), we can be more confident about the external validity of the LATE. If thsi proportion is small (e.g., $< 20\%$) we should be very cautious.

• Under the Monotonicity assumption, **we can identify the proportion of compliers in the population**.

Identifying the Proportion of Compliers

• Let θ_C, θ_A, θ_N, and θ_D, be the proportion of compliers, always-takers, nerver-takers, and defiers in the population.

• Under Monotonicity, we have that $\theta_D = 0$, such that $\theta_C + \theta_A + \theta_N = 1$.

• We have that:

$$\begin{aligned}
\Pr(D = 1 \mid Z = 0) &= \theta_C \Pr(D = 1 | Z = 0; C) + \theta_A \Pr(D = 1 | Z = 0; A) \\
&+ \theta_N \Pr(D = 1 | Z = 0; N) \\
&= \theta_A
\end{aligned}$$

• Similarly,

$$\begin{aligned}
\Pr(D = 1 \mid Z = 1) &= \theta_C \Pr(D = 1 | Z = 1; C) + \theta_A \Pr(D = 1 | Z = 1; A) \\
&+ \theta_N \Pr(D = 1 | Z = 1; N) \\
&= \theta_C + \theta_A
\end{aligned}$$

- Therefore,

$$\theta_C = \mathsf{Pr}\left(D = 1 \mid Z = 1\right) - \mathsf{Pr}\left(D = 1 \mid Z = 0\right)$$

What if Monotonicity does not hold?

What is the interpretation of $\Pr(D = 1 \mid Z = 1) - \Pr(D = 1 \mid Z = 0)$?

6. REGRESSION DISCONTINUITY (RD)

• Van der Klaauw (2002) uses a RD approach to estimate the effect of financial aid on students' decisions to accept admission to a given college. He exploits discontinuities in an administrative formula that determines aid based on SAT score, GPA, & other components.

• Angrist and Lavy (1999) estimate the effect of class size on student test scores, with identification coming from a rule requiring that one classroom be added in a school whenever average class size exceeds a predetermined threshold. Here class size is a discontinuous (and note: non-monotonic) function of enrollment in the student's school.

• Black (1999) uses a RD approach to estimates parents' willingness to pay for school quality by comparing housing prices near school district boundaries.

• Suppose that the probability of treatment (of $D = 1$) depends on some observable variable X, which is continuous. The variable X need not be independent of Y_0 and Y_1 (of TE).

• Define:

$$P(x) \equiv \mathsf{Pr}\left(D = 1 \mid X = x\right)$$

• The key feature of the RD approach is that $P(x)$ is such that there is a point x_0 in which $P(.)$ is discontinuous. Note that, though this is a necessary condition to apply a RD approach, it is not really an assumption because $P(x)$ is identified at every point in the support of X, so we can check whether this discontinuity exits or not.

• The key identification assumption is that the functions $\mu_0(x) \equiv E(Y_0|X = x)$ and $\mu_1(x) \equiv E(Y_1|X = x)$ are continuous functions of x.

ASSUMPTION RD: The functions $\mu_0(x) \equiv E(Y_0|X = x)$ and $\mu_1(x) \equiv E(Y_1|X = x)$ are continuous at $X = x_0$.

• Under this assumption any observed discontinuity in $E(Y|X = x)$ should be associated with the policy effect.

- Under Assumption RD it is possible to show that:

$$ATE(x_0) = \frac{E(Y|x_0)^+ - E(Y|x_0)^-}{P(x_0)^+ - P(x_0)^-}$$

where:

$$\begin{aligned} E(Y|x_0)^+ &\equiv \lim_{x \to x_0^+} E(Y|X = x_0) \\ E(Y|x_0)^- &\equiv \lim_{x \to x_0^-} E(Y|X = x_0) \\ P(x_0)^+ &\equiv \lim_{x \to x_0^+} P(x_0) \\ P(x_0)^- &\equiv \lim_{x \to x_0^+} P(x_0) \end{aligned}$$

- Note that $ATE(x)$ is identified only at x_0.

7. ROY'S MODEL

• Rational agents self-select in markets, occupations, education levels, etc, that maximize their payoff.

• Roy's (1951) "Thoughts on the Distribution of Earnings," is a seminal paper on this topic. He discusses the optimizing choices of workers selecting between fishing and hunting.

• Workers have skills in each occupation/sector, and they select the sector that gives them the highest expected earnings. **Roy's model is a model of comparative advantage**.

• Since that seminal paper, there has been very substantial amount of methodological and empirical work in Econometrics on the identification and estimation of Roy's model.

ROY'S MODEL

7.1. The Model

7.2. Indetification with (log)Normal distributions of skills

7.3. Nonparametric identification

7.4. Generalized Roy's model

7.1. THE MODEL

• Two occupations [or industries, or countries, etc] indexed by $d \in \{0, 1\}$. A worker is endowed with skills for each occupation (S_0 and S_1).

• Let π_0 and π_1 be the market prices of skills [the same for all workers in the market] in occupations 0 and 1, respectively, such that earnings of a worker in occupation $d \in \{0, 1\}$ are:

$$W_d = \pi_d \, S_d$$

• A worker selects the occupation that maximizes her earnings:

$$W_1 \geq W_0 \quad \Leftrightarrow \quad \text{Worker selects occupation 1}$$

$$W_1 < W_0 \quad \Leftrightarrow \quad \text{Worker selects occupation 0}$$

7.1. THE MODEL [2]

• Define the variables:

$$\begin{aligned} Y_d &\equiv \ln W_d = \ln \pi_d + \ln S_d \quad \text{(i.e., log-earnings in occupation } d\text{)} \\ D &\equiv 1\{ \text{ worker selects occupation 1}\} \end{aligned}$$

• For $d = 0, 1$, define the parameters $\mu_d \equiv \mathbb{E}(Y_d) = \ln \pi_d + \mathbb{E}(\ln S_d)$, and the random variables $U_d \equiv Y_d - \mu_d$.

• The model can be described in terms of the following equations:

$$\begin{cases} Y &= (1 - D)\ Y_0 + D\ Y_1 \\ Y_d &= \mu_d + U_d \qquad \text{for } d = 0, 1 \\ D &= 1\{Y_1 \geq Y_0\} \end{cases}$$

This is the TE model but with the assumation that individuals choose "treatment" to maximze earnings.

7.1. THE MODEL [3]

• Roy's main purpose was to understand the implications of self-selection on the distribution of earnings in different occupations.

• For $d = 0, 1$, define: $\Delta_d \equiv \mathbb{E}(Y_d|D = d) - \mathbb{E}(Y_d)$.

• If $\Delta_d > 0$ we say that there is **positive selection into occupation** d**;** i.e., workers selecting occupation d have on **more skills** in this occupation than the average worker in the population.

• If $\Delta_d < 0$ we say that there is **negative selection into occupation** d**;** i.e., workers selecting occupation d have on **less skills** in this occupation than the average worker in the population.

• What are the predictions of the model about Δ_0 and Δ_1?

7.1. THE MODEL [4]

- Note that:

$$\begin{aligned} D &= 1\{Y_1 \geq Y_0\} \\ &= 1\{U_0 - U_1 \leq \mu_1 - \mu_0\} \\ &= 1\left\{\frac{V}{\sigma_V} \leq \gamma\right\} \end{aligned}$$

where $V \equiv U_0 - U_1$ and $\gamma \equiv \dfrac{\mu_1 - \mu_0}{\sigma_V}$

7.1. THE MODEL [5]

- Under normality of U_0 and U_1:

$$\begin{aligned}
\mathbb{E}\left(Y_1|D=1\right) &= \mu_1 + \mathbb{E}\left(U_1 \mid V \leq \mu_1 - \mu_0\right) \\
&= \mu_1 + \mathbb{E}\left(\frac{\sigma_{1V}}{\sigma_V^2} V \mid V \leq \mu_1 - \mu_0\right) \\
&= \mu_1 + \frac{\sigma_{1V}}{\sigma_V^2}\, \sigma_V\, \mathbb{E}\left(\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} \leq \frac{\mu_1 - \mu_0}{\sigma_V}\right) \\
&= \mu_1 - \frac{\sigma_{1V}}{\sigma_V}\, \frac{\phi\left(\gamma\right)}{\Phi\left(\gamma\right)}
\end{aligned}$$

7.1. THE MODEL [6]

- Similarly,

$$\begin{aligned}
\mathbb{E}\left(Y_0|D=0\right) &= \mu_0 + \mathbb{E}\left(U_0 \mid V > \mu_1 - \mu_0\right) \\
&= \mu_0 + \mathbb{E}\left(\frac{\sigma_{0V}}{\sigma_V^2} V \mid V > \mu_1 - \mu_0\right) \\
&= \mu_0 + \frac{\sigma_{0V}}{\sigma_V^2}\,\sigma_V\,\mathbb{E}\left(\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > \gamma\right) \\
&= \mu_0 + \frac{\sigma_{0V}}{\sigma_V}\,\frac{\phi\left(\gamma\right)}{1-\Phi\left(\gamma\right)}
\end{aligned}$$

7.1. THE MODEL [7]

• Taking into account that $\sigma_{0V} = \sigma_0^2 - \sigma_{01}$ and $\sigma_{1V} = \sigma_{01} - \sigma_1^2$, and defining γ

$$\Delta_0 = \frac{\sigma_0^2 - \sigma_{01}}{\sigma_V} \frac{\phi(\gamma)}{1 - \Phi(\gamma)}$$

$$\Delta_1 = \frac{\sigma_1^2 - \sigma_{01}}{\sigma_V} \frac{\phi(\gamma)}{\Phi(\gamma)}$$

• The signs of Δ_0 and Δ_1 depend on the signs of $[\sigma_0^2 - \sigma_{01}]$ and $\left[\sigma_1^2 - \sigma_{01}\right]$, respectively.

• Note that: $\sigma_V^2 = [\sigma_0^2 - \sigma_{01}] + \left[\sigma_1^2 - \sigma_{01}\right] >$, so at least one of the two terms is positive, and it can be both.

7.1. THE MODEL [8]

	$\sigma_1^2 - \sigma_{01} < 0$	$\sigma_1^2 - \sigma_{01} > 0$
$\sigma_0^2 - \sigma_{01} < 0$	Imposible	Positive selection in 1 Negative selection in 0
$\sigma_0^2 - \sigma_{01} > 0$	Negative selection in 1 Positive selection in 0	Positive selection in 1 Positive selection in 0

7.1. THE MODEL [9]

• **Which type of occupation has positive selection?** The occupation where the distribution of skills is more heterogeneous, more disperse.

• To see this, note that

$$[\sigma_0^2 - \sigma_{01}] = \sigma_0\sigma_1\left[\frac{\sigma_0}{\sigma_1} - \rho\right]$$

$$[\sigma_1^2 - \sigma_{01}] = \sigma_0\sigma_1\left[\frac{\sigma_1}{\sigma_0} - \rho\right]$$

such that the sign of Δ_0 is determined by the sign of $\left[\frac{\sigma_0}{\sigma_1} - \rho\right]$, and the sign of Δ_1 is determined by the sign of $\left[\frac{\sigma_1}{\sigma_0} - \rho\right]$.

- If $\frac{\sigma_1}{\sigma_0} > 1$, then $\Delta_1 > 0$, and if If $\frac{\sigma_0}{\sigma_1} > 1$, then $\Delta_0 > 0$.

7.2. Indetification: Normal distributions

• Suppose that we have cross-sectional data $\{y_i, d_i : i = 1, 2, ..., N\}$. Can we identify the parameters of the Roy's model $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1, \sigma_{01})$?

• Heckman and Honore (ECMA, 1990) show that we normal distributions the parameters are uniquely identified from the following moments in the data:

$\Pr(D = 1)$; $\mathbb{E}(Y|D = 0)$; $\mathbb{E}(Y|D = 1)$; $\mathbb{V}(Y|D = 0)$; and $\mathbb{V}(Y|D = 1)$

• They also show that, without regressors, the model is not identified if we consider a nonparametric specification of the unobservables.

• Then, they present nonparametric identification results when the model includes regressors X.

7.3. Nonparametric Indetification: Exclusion restrictions

• Consider the model with repressors, such that $\mu_d(X) \equiv \mathbb{E}(Y_d|X)$, and assume that U_0 and U_1 are independent of X.

• Suppose that X includes three groups of variables: $X = (Z_0, Z_1, X_c)$ such that:

$$\mu_0(X) = \mu_0(Z_0, X_c)$$

$$\mu_1(X) = \mu_1(Z_1, X_c)$$

• Furtherefore Z_0 and Z_1 have continuous support and $\mu_d(Z_d, X_c)$ is strictly monotonic in Z_d, and

$$\lim_{Z_d \to -\infty} \mu_d(Z_d, X_c) = -\infty$$

Nonparametric Indetification: Exclusion restrictions [2]

- We have that:

$$\mathbb{E}\left(Y \mid X, D=0\right) \;=\; \mu_0(Z_0, X_c) + \mathbb{E}\left(U_0 \mid V > \mu_1(Z_1, X_c) - \mu_0(Z_0, X_c)\right)$$

- Therefore,

$$\begin{aligned}
\lim_{Z_1 \to -\infty} \mathbb{E}\left(Y \mid X, D=0\right) \;=\; & \mu_0(Z_0, X_c) \\
& + \mathbb{E}\left(U_0 \mid V > \lim_{Z_1 \to -\infty} \mu_1(Z_1, X_c) - \mu_0(Z_0, X_c)\right) \\
= \; & \mu_0(Z_0, X_c) + \mathbb{E}\left(U_1 \mid V > -\infty\right) \\
= \; & \mu_0(Z_0, X_c)
\end{aligned}$$

and $\mu_0(Z_0, X_c)$ is identified everywhere.

Nonparametric Indetification: Exclusion restrictions [3]

- Similarly, we have that:

$$\mathbb{E}\left(Y \mid X, D=1\right) \;=\; \mu_1(Z_1, X_c) + \mathbb{E}\left(U_1 \mid V \leq \mu_1(Z_1, X_c) - \mu_0(Z_0, X_c)\right)$$

- Therefore,

$$\begin{aligned}
\lim_{Z_0 \to -\infty} \mathbb{E}\left(Y \mid X, D=1\right) \;=\; & \mu_1(Z_1, X_c) \\
& + \mathbb{E}\left(U_1 \mid V \leq \mu_1(Z_1, X_c) - \lim_{Z_0 \to -\infty} \mu_0(Z_0, X_c)\right) \\
= \; & \mu_1(Z_1, X_c) + \mathbb{E}\left(U_1 \mid V \leq +\infty\right) \\
= \; & \mu_1(Z_1, X_c)
\end{aligned}$$

and $\mu_0(Z_0, X_c)$ is identified everywhere.

Nonparametric Indetification: Exclusion restrictions [4]

• For estimation, we can use nonparametric methods.

• Define the choice probability $P_D(X) = \mathsf{Pr}(D = 1|X)$. The model implies that $P_D(X) = F_V(\mu_1(Z_1, X_c) - \mu_0(Z_0, X_c))$, and if $F_V(.)$ is strictlu increasing:

$$\mu_1(Z_1, X_c) - \mu_0(Z_0, X_c) = F_V^{-1}[P_D(X)]$$

• Note that $\mathbb{E}(U_1 \mid V \leq \mu_1(Z_1, X_c) - \mu_0(Z_0, X_c))$ is a function of $\mu_1(Z_1, X_c) - \mu_0(Z_0, X_c)$ only, and therefore we can represented as a function of $P_D(X)$.

$$\mathbb{E}(U_1 \mid V \leq \mu_1(Z_1, X_c) - \mu_0(Z_0, X_c)) = s_1(P_D(Z_0, Z_1, X_c))$$

Nonparametric Indetification: Exclusion restrictions [5]

- Therefore, we can write

$$\mathbb{E}(Y \mid X, D = 1) \;=\; \mu_1(Z_1, X_c) + s_1(P_D(Z_0, Z_1, X_c))$$

- For the subsample of observations with $d_i = 1$, consider the regression model:

$$\begin{aligned} y_i &= \mu_1(z_{1i}, x_{ci}) + s_1(p_i) + e_i \\ &= h(z_{1i}, x_{ci})'\beta_1 + s_1(p_i) + e_i \end{aligned}$$

- We can use Robinson (1988) or Yatchew (2003) to estimate β_1 in this model.