Inequality in Health

Tutorial 1

Nikolaos Prodromidis, M.Sc.

nikolaos.prodromidis@uni-due.de

Winter Term 2022/2023

Types of Endogeneity Problems

If we are interested in estimating causal effects, three types of endogeneity problems might arise:

1. Omitted variable bias (OVB)
2. Measurement error
3. Simultaneity bias

Omitted Variable Bias

We are interested in the causal effect of X on Y and estimate

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{1}$$

However, the true data generating process is given by

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 W + \eta. \tag{2}$$

Omitted Variable Bias

In case (i) $\alpha_2 \neq 0$ and (ii) X and W are correlated, e.g. via the relationship

$$W = \gamma_0 + \gamma_1 X + \zeta, \tag{3}$$

regressing Y on X only is not able to recover the true effect α_1 as W enters the error term ($\varepsilon = \eta + \alpha_2 W$) and thus the OLS assumption $\mathbb{E}[\varepsilon \mid X] = 0$ is violated.

Omitted Variable Bias

Inserting (3) into (2) gives

$$Y = \underbrace{(\alpha_0 + \alpha_2\gamma_0)}_{\beta_0} + \underbrace{(\alpha_1 + \alpha_2\gamma_1)}_{\beta_1} X + \underbrace{(\eta + \alpha_2\zeta)}_{\varepsilon}. \tag{4}$$

Thus, $\beta_1 = (\alpha_1 + \alpha_2\gamma_1) \neq \alpha_1$.

Measurement Error

Assume the outcome Y is generated by

$$Y = \beta_0 + \beta_1 X^* + \varepsilon \tag{5}$$

but we only observe

$$X = X^* + \eta \tag{6}$$

with $\mathbb{E}[\eta] = 0$ and η uncorrelated with X^*.

Measurement Error

Inserting (6) into (5) gives

$$\begin{aligned} Y &= \beta_0 + \beta_1 (X - \eta) + \varepsilon \\ &= \beta_0 + \beta_1 X + \underbrace{(\varepsilon - \beta_1 \eta)}_{\zeta}. \end{aligned} \tag{7}$$

As both X and ζ are functions of η, $\mathbb{E}[\zeta \mid X] = 0$ is violated and

$$\operatorname{plim} \widehat{\beta}_1 \neq \beta_1. \tag{8}$$

Simultaneity Bias

Assume that the relationship between Y and X is given by the following equations:

$$Y = \alpha_0 + \alpha_1 X + \varepsilon \tag{9}$$
$$X = \beta_0 + \beta_1 Y + \eta \tag{10}$$

If we are interested in the effect of X on Y (9), the OLS estimate $\widehat{\alpha}_1$ will not be unbiased.

Simultaneity Bias

To make this clear, we insert (9) into (10) and obtain

$$X = (\beta_0 + \beta_1 \alpha_0) + \beta_1 \alpha_1 X + (\eta + \beta_1 \varepsilon)\,. \tag{11}$$

Even without solving for X we see that X is a function of ε and thus the assumption $\mathbb{E}\left[\varepsilon \mid X\right] = 0$ is violated.

Endogeneity Problems: Examples

What is the relationship between education health?
People with more education appear to be healthier. The two variables are highly associated.
But is this relationship causal?

Instrumental Variables: Education and Health

Imagine: We know that there is a variable Z that is not affected by health or education. We also know that this variable is not correlated with any omitted variables.
This Z variable only affects education. This "movement" in Z can be used to mitigate endogeneity concerns.

- IV: Uses variation in our treatment variable (e.g education) that is not endogenous.

Think it as an event that impacts our health but all the effects are going through education.

IV Assumptions

Instrumental variable (IV) techniques are frequently applied to deal with endogeneity problems. Using a suitable instrument Z allows us to estimate causal effects if endogeneity problems are present. Two assumptions must hold for a variable Z to be a suitable instrument:

1. Exclusion restriction
2. Existence of (strong) first stage

IV Assumptions – Exclusion Restriction

Assume we are interested in the causal effect of X on Y in the *structural equation*

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{12}$$

but X is correlated with the error term ε. The exclusion restriction states that the instrument Z must not be correlated with ε which is equivalent to assuming that Z does not have a direct effect on Y.

IV Assumptions – Existence of First Stage

Further, Z must be correlated with X such that in the *first stage equation*

$$X = \pi_0 + \pi_1 Z + \eta \tag{13}$$

the coefficient $\pi_1 \neq 0$.

IV Estimation

IV estimation exploits the fact that only a part of X is correlated with the error term ε. That means if we could exclude the variation in X which causes the endogeneity problem, we could estimate causal effects using the "good" (i.e., exogenous) variation in X.

So in order to find suitable instruments, we must ask (i) which factors determine X, and (ii) which parts of X lead to the endogeneity problem.

IV Estimation

By inserting the first stage (13) into the structural equation (12), we obtain

$$Y = \underbrace{(\beta_0 + \beta_1 \pi_0)}_{\gamma_0} + \underbrace{\beta_1 \pi_1}_{\gamma_1} Z + \underbrace{(\varepsilon + \beta_1 \eta)}_{\zeta}. \tag{14}$$

This relationship is also called *reduced form*.

- Does reduce form show a causal effect?

IV Estimation

As we know that the reduced form estimate γ_1 is equal to the product of the causal effect of interest β_1 and the first stage estimate π_1, we can obtain an estimate of β_1 as

$$\widehat{\beta}_1 = \frac{\widehat{\gamma}_1}{\widehat{\pi}_1}. \tag{15}$$

IV Estimators

- Indirect Least Squares (ILS):

$$\widehat{\beta}_1^{ILS} = \frac{\widehat{\gamma}_1}{\widehat{\pi}_1}$$

- IV estimator (IV):

$$\widehat{\beta}_1^{IV} = \left(Z'X\right)^{-1} Z'Y$$

- Two-Stage Least Squares (2SLS):

$$\widehat{\beta}_1^{2SLS} = \left(\widehat{X}'\widehat{X}\right)^{-1} \widehat{X}'Y$$

- Generalized Method of Moments (GMM):

$$\widehat{\beta}_1^{GMM} = \left(X'P_Z X\right)^{-1} X'P_Z Y$$

- Wald estimator (Wald):

$$\widehat{\beta}_1^{Wald} = \frac{\mathbb{E}\left[Y \mid Z = 1\right] - \mathbb{E}\left[Y \mid Z = 0\right]}{\mathbb{E}\left[X \mid Z = 1\right] - \mathbb{E}\left[X \mid Z = 0\right]}$$

Weak Instruments

Weak instruments are instruments which hardly explain any variation in X. Weak instruments might lead to heavily biased estimates:

$$\mathbb{E}\left[\widehat{\beta}_1^{2SLS} - \beta_1\right] \approx \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2}\left[\frac{1}{F+1}\right] \tag{16}$$

where F denotes the F-Statistic for the test $\pi_1 = 0$. Thus, if F is small, the bias becomes large. In practice, we should only consider instruments with $F > 10$.

LATE

Standard IV estimation assumes homogeneous causal effects. However, not all individuals need to be affected by the instrument in the same way:

$$\begin{aligned} X_i &= X_{0i} + \left(X_{1i} - X_{0i}\right) Z_i \\ &= \pi_0 + \pi_{1i} Z_i + \eta_i \end{aligned}$$

LATE – Compliance

	$X_{0i} = 1$	$X_{0i} = 0$
$X_{1i} = 1$	Always-Takers	Compliers
$X_{1i} = 0$	Defiers	Never-Takers

LATE – Assumptions

- Independence: instrument is as good as randomly assigned, independent of potential outcomes/treatment assignment.

→ First stage and reduced form coefficients have a causal interpretation!

- Exclusion: instrument affects outcome only via X.
- Monotonicity: $\pi_{1i} \geq 0 \ \forall i$ or $\pi_{1i} \leq 0 \ \forall i$

→ Defiers do not exist!

⇒ LATE estimates the causal effect for the subgroup of compliers.

Causal effect of education on health: Lleras-Muney, A. (2005)

- Idea: Use changes on laws that affect education as instruments. If those laws forced children to stay more at school and if education improves health then the affected individuals from the laws should be healthier.
- Historical data: State-level changes on laws 1910s-1930s
- Instrument: Laws (regarding compulsory education) when an individual was 14 years old.

Results

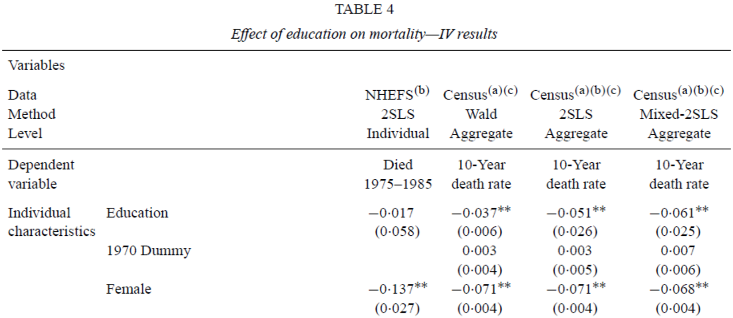


TABLE 4

Effect of education on mortality—IV results

Variables					
Data		NHEFS[b]	Census[a][c]	Census[a][b][c]	Census[a][b][c]
Method		2SLS	Wald	2SLS	Mixed-2SLS
Level		Individual	Aggregate	Aggregate	Aggregate
Dependent variable		Died 1975–1985	10-Year death rate	10-Year death rate	10-Year death rate
Individual characteristics	Education	−0·017	−0·037**	−0·051**	−0·061**
		(0·058)	(0·006)	(0·026)	(0·025)
	1970 Dummy		0·003	0·003	0·007
			(0·004)	(0·005)	(0·006)
	Female	−0·137**	−0·071**	−0·071**	−0·068**
		(0·027)	(0·004)	(0·004)	(0·004)

Identification strategy discussion

Are changes in laws really exogenous? Critique:
Mazumder, B., 2008. Does education improve health? A reexamination of the evidence from compulsory schooling laws. Economic Perspectives, 32(2).

Overview

- One of the most frequently applied approaches for causal analysis in microeconomic research (especially health, labor, and education economics) and other social sciences (sociology, political science)
- Easily implemented: OLS
- But: strong assumptions, usually not testable

History

- First application (probably): English physician John Snow (1855) studied cholera epidemics in London
- Study: show that cholera is transmitted by contaminated drinking water (not "bad air")
- Setting:
  - 1849: two water companies, both obtained water from the dirty part of the Thames
  - 1853: one water company moved to an area with cleaner water
- Result: death rates fell sharply in districts supplied by the company that moved in comparison to the other districts

Basic Idea

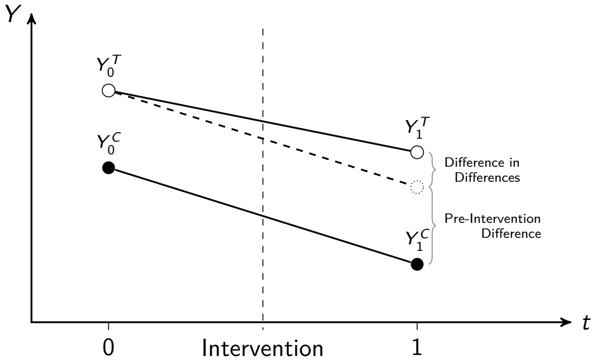
- Post-treatment differences between treatment and control group outcomes might be due to
  1) Treatment
  2) Unobserved/unobservable time-invariant heterogeneity
- In RCTs time-invariant heterogeneity should be zero but this is unlikely in most quasi-experimental settings.

⇒ Comparison of average post-treatment outcomes $\neq$ treatment effect

$$\mathbb{E}\left[\overline{Y}_1^T - \overline{Y}_1^C\right] \neq \tau$$

Basic Idea

⇒ Idea: use pre-treatment difference to decompose post-treatment difference into treatment effect and heterogeneity.



Implementation

- Simplest case: two groups, two periods

$$\hat{\tau}_{DID} = \overline{Y}_1^T - \overline{Y}_0^T - \left[\overline{Y}_1^C - \overline{Y}_0^C\right]$$

- Better: OLS regression

$$Y_{it} = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot D_i + \tau_{DID} \cdot tD_i + \mathbf{X}_{it}\beta + \varepsilon_{it}$$

  where **X** is a set of covariates.

- Advantages:
  - Possible to control for covariates
  - Inference becomes very easy
  - Various extensions possible (additional time periods, groups, continuous treatments, etc.)

Implementation

Generalization:

$$Y_{it} = \mu_i + \lambda_t + \tau_{DID} \cdot \mathbb{1}\left(t \geq T_0\right) \cdot D_i + \mathbf{X}_{it}\beta + \varepsilon_{it}$$

where

μ_i set of group dummies equal to 1 in case of unit i

λ_t set of time dummies equal to 1 in case of period t

T_0 treatment period

Assumptions

1) SUTVA - Stable unit treatment value assumption:
   - Observational units are either affected or unaffected by the treatment
   - Violated if treatment has an impact on untreated units (e.g. via equilibrium effects)
2) EXOG - Exogeneity:
   - Included covariates must be exogenous, i.e., unaffected by the treatment

   ⇒ Use covariates measured prior to treatment

Assumptions

3) NEPT - No effect prior to treatment:
   - Anticipation effects not allowed
4) COSU - Common support:
   - Observations with characteristics x exist in all four sub-samples
   - Violated if control group characteristics differ from treatment group (e.g., control group: individuals aged 25-50, treatment group: individuals aged 65-75)
5) CT - Common trend:
   - Treatment and control group follow a common trend in absence of treatment
   - Allows to use control group as counterfactual
   - Plausibility test: event study graphs (see next slide)

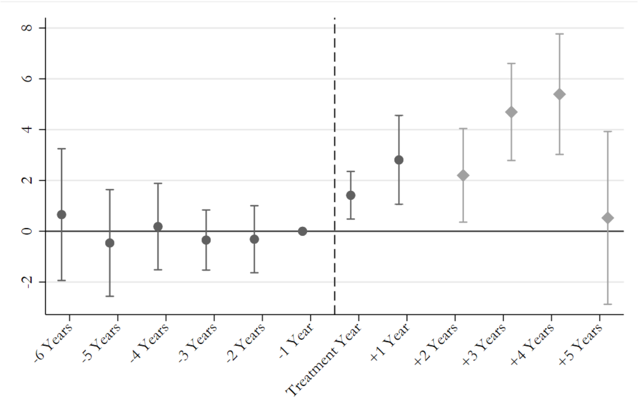
Excursion: Event Study Graphs

- CT not directly testable but plausibility can be evaluated
- Regress Y on lags and leads of treatment variable (several pre-treatment periods needed):

$$\begin{aligned} Y_{it} = &\beta_0 + \ldots + \tau_{-3}\mathbb{1}\left(t=-3\right) + \tau_{-2}\mathbb{1}\left(t=-2\right) \\ &+ \tau_1\mathbb{1}\left(t=1\right) + \tau_2\mathbb{1}\left(t=2\right) + \ldots + \beta_1 \cdot D_i + \epsilon_{it} \end{aligned}$$

- If pre-period coefficients insignificant → CT plausible

Excursion: Event Study Graphs



Introduction

- In many applications, treatment is determined by a deterministic rule, i.e., some continuous variable crossing a threshold.
- Examples:
  - Political parties winning an election with $>$50% of votes
  - Scholarships for students above a certain test score
  - Retirement eligibility above age 65
- Regression Discontinuity Designs (RDD) exploit knowledge of the treatment assignment process for causal inference.

General Requirements

- Requirements:
  - Score (e.g. vote share, test score, age)
  - Treatment (e.g. winning an election, receiving scholarship, receiving old-age pension)
  - Threshold (e.g. 50%, 90%, 65 years)
- Distinction between two scenarios:
  - Perfect compliance: score determines treatment → sharp RD
  - Imperfect compliance: score determines probability of treatment with discontinuity around threshold → fuzzy RD

Sharp RD Treatment Assignment

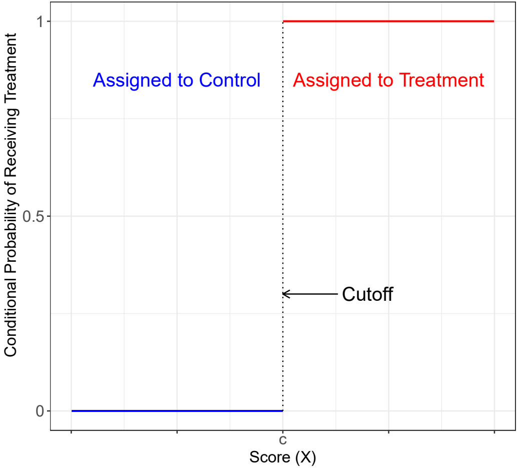


Figure: Source: Cattaneo et al. (2019)

Frameworks

- Two alternative frameworks with different causal parameters, different identification assumptions, and different estimation and innference methods:
  - **Continuity-based approach**: requires continuity of potential outcomes around threshold
  - **Local randomisation approach**: assumes random assignment in neighborhood around threshold
- Commonality: only untreated units below and only treated units above threshold

Continuity-Based Approach

- Only one potential outcome is observed on each side of the threshold.

⇒ We do not observe treated and untreated units with the same value of the score.

- However, at the threshold we “almost" observe both potential outcomes.
- Assumption: potential outcomes are continuous around the threshold.

Potential Outcomes

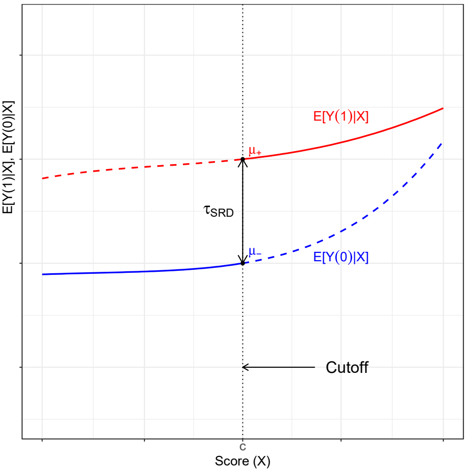


Figure: Source: Cattaneo et al. (2019)

RD Treatment Effects

- Consequence: Causal effects are only available for units with score equal to cutoff.
- The resulting parameter τ_{SRD} can be interpreted as local average treatment effect on the treated.
- Formally, the effect is defined as:

$$\tau_{SRD} = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right] \tag{17}$$

where X_i is the score and c is the threshold.

- Hahn et al. (2001) show that

$$\mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right] = \lim_{x \downarrow c} \mathbb{E}\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y_i \mid X_i = x\right] \tag{18}$$

⇒ If potential outcomes are continuous functions at c, the difference between limits of observed outcomes is the ATET at the cutoff.

Estimation

- Traditional setting: OLS regression

$$
\begin{aligned}
\mathbb{E}\left[Y_{0i}\right] &= \alpha + f\left(x_i\right) + \varepsilon_{0i} \\
\mathbb{E}\left[Y_{1i}\right] &= \alpha + \tau + f\left(x_i\right) + \varepsilon_{1i}
\end{aligned}
$$

- Specification allows for flexible functions of the score.
- It does not use observations from one group to estimate parameters for the other group.

Disadvantages of Parametric Approach

- RDD requires good approximations at boundary points but linear regression focuses on good overall fit

→ Poor approximation at boundary points.

- Results are often sensitive to the order of polynomial.
- Inference is often misleading: over-rejection.

⇒ Conclusion: Rely on local polynomial estimation!

Local Polynomial Regression

- Local polynomial methods only use observations close to the threshold, separately for treatment and control.
- Formally, observations in the interval $[c - h; c + h]$ are included where h is called **bandwidth**.
- To ensure that observations close to the threshold receive more weight, a weighting scheme determined by a kernel function $K(\cdot)$ is adopted.
- Kernel functions need to satisfy three criteria:
  1. $\int_{-\inf}^{\inf} K(u) = 1$
  2. $K(u) = K(-u)$
  3. $K(u) \geq 0$

Introduction

- Previously, we focused on the basic case: sharp RD under the continuity-bases approach
- However, there is another framework for analysing sharp RD designs, the local randomisation approach.
- The main idea of this approach is that, if the treatment can be considered as good as randomly assigned in a neighbourhood around the cutoff, the setting can be analysed similar as in a randomised experiment.

Local Randomisation RDD

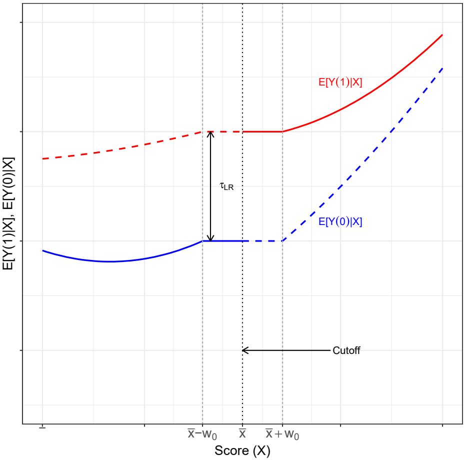


Figure: Source: Cattaneo et al. (2018)

Treatment Effect

- The local randomisation approach identifies τ_{SRD}^{LR} defined as

$$\tau_{SRD}^{LR} = \frac{1}{n_{W_0}} \left(\sum_{i:X_i \in W_0} Y_i(1) - \sum_{i:X_i \in W_0} Y_i(0) \right), \qquad (19)$$

  where n_{W_0} denotes the number of observations within W_0.

- The corresponding estimator is

$$\hat{\tau}_{SRD}^{LR} = \frac{1}{n_{W_{0,+}}} \sum_{i:X_i \in W_{0,+}} Y_i(1) - \frac{1}{n_{W_{0,-}}} \sum_{i:X_i \in W_{0,-}} Y_i(0). \qquad (20)$$

- Difference between τ_{SRD} and τ_{SRD}^{LR}:
  - τ_{SRD}: causal effect at a single point
  - τ_{SRD}^{LR}: causal within an interval

⇒ For $w \to 0$, both coefficients become more conceptually similar.

Introduction

- Methods based on **unconfoundedness** (UC) are very common in the programme evaluation literature.
- The UC assumption requires that *conditional on observed covariates*, there are no **unobserved factors** associated with both assignment and potential outcomes.
- This is a **strong** assumption, but we often have datasets which have been designed to make it plausible.
- The logic is similar to the ideas underlying standard **multiple regression analysis**.
- UC implies we have a rich set of predictors of the treatment indicator – so that adjusting for these lead to valid estimates of causal effects.

Introduction II

- Combined with a **linearity assumption**, UC justifies linear regression.
- If the linear approximation is *not* accurate, regression estimates of average treatment effects can be **severely biased**.
- As a result of this sensitivity, more sophisticated approaches have been developed.
- Some of these rely on the **propensity score** (the conditional probability of receiving treatment).

A Simple Parametric Model

- Define the following models for potential outcomes:

$$\mu_0 (x) = \mathbb{E}\left[Y_i^0 | X_i = x\right]$$
$$\mu_1 (x) = \mathbb{E}\left[Y_i^1 | X_i = x\right]$$

- By definition, the average treatment effect conditional on $X = x$ is $\tau (x) = \mu_1 (x) - \mu_0 (x)$.
- Under UC, we can estimate $\mu_1 (x)$ using the treated subsample and $\mu_0 (x)$ using the untreated subsample to get:

$$\hat{\tau}_{reg} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\mu}_1 (x) - \hat{\mu}_0 (x)\right)$$

A Simple Parametric Model II

- In the simplest case, we assume a **linear** specification:

$$\mu_0 (x) = \alpha_0 + \beta_0' (x - \psi_X)$$
$$\mu_1 (x) = \alpha_1 + \beta_1' (x - \psi_X)$$

where ψ_X is the population mean.

- Naturally, we replace it with the sample average.
- Then

$$\hat{\tau}_{reg} = \hat{\alpha}_1 - \hat{\alpha}_0,$$

which can be obtained from the coefficient of D in a regression of Y_i on $\left(1, D_i, X_i, D_i \cdot \left(X_i - \bar{X}\right)\right)$

Problems of Linear Regression

- In the simple linear case, the treatment effect estimator can be expressed as

$$\hat{\tau}_{reg} = \bar{Y}_1 - \bar{Y}_0 - \left(\frac{N_0 \hat{\beta}_1}{N_0 + N_1} + \frac{N_1 \hat{\beta}_0}{N_0 + N_1} \right)' \left(\bar{X}_1 - \bar{X}_0 \right)$$

- Thus, the average difference in **outcomes** is adjusted by the difference in **covariates**.
- If averages of covariates are very different, this adjustment is **large**.
- In this case, results can be sensitive to minor changes in specification.
- Unless the linear approximation is *globally* accurate, this may lead to **severe biases**.
- If averages of covariates are very different, collinear with treatment – thus exacerbating misspecification problems.

Matching Estimators

- Matching estimators impute the missing potential outcomes using outcomes of a few **neighbours** in the opposite group.
- Given the matching metric, the researcher only has to choose the **number of matches**.
- Using a single match leads to the most credible inference with the least **bias**, at the cost of lower **precision**.
- Matching estimators are normally used when
  1. The interest is in the ATT and
  2. There is a large pool of controls.

The Propensity Score

- Under *UC*, independence holds also after conditioning only on the propensity score (PS): $e\left(x\right) = \Pr\left(D_i = 1 | X_i = x\right)$

$$\begin{aligned} D_i \perp\!\!\!\perp \left(Y_i^1, Y_i^0\right) \Big| \, X_i \\ \Rightarrow D_i \perp\!\!\!\perp \left(Y_i^1, Y_i^0\right) \Big| \, e\left(X_i\right) \end{aligned}$$

- Hence, within a subpopulation with the same PS, outcomes are independent of the treatment indicator.
- Thus, it is sufficient to adjust only for differences in PS between treated and control units.

The Propensity Score

- Suppose we estimate the PS using a logit model and MLE.

$$e(x) = \frac{\exp\left(h(x)'\gamma\right)}{1 + \exp\left(h(x)'\gamma\right)}$$

  for some selected vector of functions $h(x)$.
- Our estimated PS $\hat{e}(x)$ will thus be based on $\hat{\gamma}_{MLE}$.
- But we need to choose the vector of functions $h(x)$.

Nearest Neighbour Matching

- Nearest neighbour matching with M neighbours: impute outcomes by

$$\hat{Y}_i^d = \begin{cases} Y_i & \text{if } D_i = d \\ \frac{1}{M} \sum_{j \in J_M(i)} Y_j & \text{if } D_i \neq d \end{cases}$$

  and estimate $\hat{\tau}_{match} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{Y}_i^1 - \hat{Y}_i^0 \right)$.

- This matching algorithm entails tradeoffs between **variance** and **bias**:
  - Matching with **replacement**: if there are PS regions with *many* treated per control. Increases average **match quality** but also the **variance**.
  - **Oversampling** (many neighbours per treated) reduces variance but increases bias (poorer matches used).