5. Observational Studies

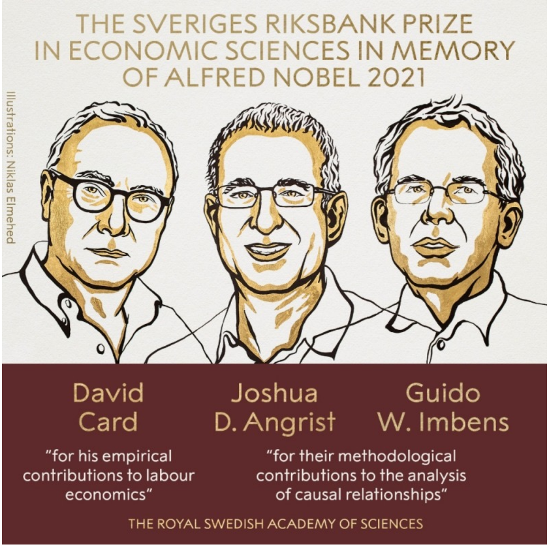
ISS5096 || ECI

Jaewon ("Jay-one") Yoo
National Tsing Hua University

Where are we? Where are we going?

- So far: experiments where design makes things easier.
- Today: what happens when we have observational studies to work with?
  - Begin with **identification**, **selection on observables**, and **DAGs**.
  - Rest of the course will cover different designs for observational studies.
- Q: Why are observational studies in causal inference important? (What are the limitations of RCTs?)

Where are we? Where are we going?



Source: Twitter *@NobelPrize*

Where are we? Where are we going?

Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

ANSWERING CAUSAL QUESTIONS USING OBSERVATIONAL DATA

The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel

"Taken together, therefore, the Laureates' contributions have played a central role in establishing the so-called design-based approach in economics. This approach – aimed at emulating a randomized experiment to answer a causal question ***using observational data*** *– has transformed applied work and improved researchers' ability to answer causal questions of great importance for economic and social policy using observational data." (p.2)*

1/ Identification in Observational Studies

Randomized Experiment Review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ is the probability of treatment assignment.
  - p_i is controlled by & known to researcher in an experiment.

Randomized Experiment Review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ is the probability of treatment assignment.
  - p_i is controlled by & known to researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:
  1. **Positivity**: assignment is probabilistic (and not deterministic):
     $0 < \mathbb{P}[D_i = 1] < 1$

Randomized Experiment Review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ is the probability of treatment assignment.
  - p_i is controlled by & known to researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:
  1. **Positivity**: assignment is probabilistic (and not deterministic): $0 < \mathbb{P}[D_i = 1] < 1$
  2. **Unconfoundedness**: $\mathbb{P}[D_i = 1 | \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1]$
     - Treatment assignment does not depend on any potential outcomes.
     - Sometimes written as $D_i \perp\!\!\!\perp (\mathbf{Y}(1), \mathbf{Y}(0))$.

What is the Selection Problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i is related to POs.

What is the Selection Problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i is related to POs.
- What can we learn about the ATE here? Look at the diff-in-means:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

What is the Selection Problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i is related to POs.
- What can we learn about the ATE here? Look at the diff-in-means:

$$\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \quad \text{(consistency)}
\end{aligned}$$

What is the Selection Problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i is related to POs.
- What can we learn about the ATE here? Look at the diff-in-means:

$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]
\end{aligned}
$$

What is the Selection Problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i is related to POs.
- What can we learn about the ATE here? Look at the diff-in-means:

$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

What is the Selection Problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i is related to POs.
- What can we learn about the ATE here? Look at the diff-in-means:

$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

- Without unconfoundedness: naive diff-in-means = PATT + selection bias
- **Selection bias**: how different the treated and control groups are in terms of their potential outcome under control.

Selection Bias = Unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_{\text{treated}}}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: a combination of two unknown quantities.
  - Can’t distinguish if a diff-in-means is the ATT or selection bias.

Selection Bias = Unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_{\text{treated}}}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: a combination of two unknown quantities.
  - Can't distinguish if a diff-in-means is the ATT or selection bias.

- Example: effect of comment sections on support for online influencers.
  - Naive estimate: influencers do worse without comment sections than with them.
  - $\rightsquigarrow$ negative ATT **OR** positive ATT with large negative selection bias.
  - SB = influencers that disable user comments are worse than those that keep them, even if they posted the same content.

Selection Bias = Unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_{\text{treated}}}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: a combination of two unknown quantities.
    - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of comment sections on support for online influencers.
    - Naive estimate: influencers do worse without comment sections than with them.
    - $\rightsquigarrow$ negative ATT **OR** positive ATT with large negative selection bias.
    - SB = influencers that disable user comments are worse than those that keep them, even if they posted the same content.
- With an unbounded Y_i, we cannot even bound the ATT because, in principle, SB could be anywhere from $-\infty$ to ∞.

Selection Bias = Unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_{\text{treated}}}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: a combination of two unknown quantities.
    - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of comment sections on support for online influencers.
    - Naive estimate: influencers do worse without comment sections than with them.
    - $\rightsquigarrow$ negative ATT **OR** positive ATT with large negative selection bias.
    - SB = influencers that disable user comments are worse than those that keep them, even if they posted the same content.
- With an unbounded Y_i, we cannot even bound the ATT because, in principle, SB could be anywhere from $-\infty$ to ∞.
- We say ATT (as well as ATE) are **unidentified** w/o further assumptions.

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$.
    - $\rightsquigarrow$ We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$.
    - $\rightsquigarrow$ We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ (/(p)saI/) is **identified** if we can write it as function of $\mathbb{P}$.
  - Would we know this quantity if we had access to unlimited data?
  - $\rightsquigarrow$ no worrying about estimation uncertainty here.

What is identification?

- **Identification** connects the counterfactual to the observed.
    - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
    - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
    - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$.
        - $\leadsto$ We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ (/(p)saI/) is **identified** if we can write it as function of $\mathbb{P}$.
    - Would we know this quantity if we had access to unlimited data?
    - $\leadsto$ no worrying about estimation uncertainty here.
- Connecting counterfactuals to the observational requires **assumptions**.
    - **"What is your identification strategy?"** = what are the assumptions that allow you to claim that you've estimated a causal effect?
    - Research design can help justify assumptions (experiments, RDD, etc).
    - Or you will need to justify them through argument.

Identification vs. Estimation

- Identification tells us **what** to estimate, not **how**.
    - If identified, we know our causal parameter is some function of $\mathbb{P}$.

Identification vs. Estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, let's consider the **population** diff-in-means:

$$\mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0]$$

  - But, $\mathbb{P}$ is not directly observable since it's a population distribution!

Identification vs. Estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, let's consider the **population** diff-in-means:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

  - But, $\mathbb{P}$ is not directly observable since it's a population distribution!
- Once identified, we need to actually **estimate** the function of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
  - Now just estimating conditional expectations, etc.
  - $\rightsquigarrow$ **after identification, causal inference part done**
  - Purely a statistical question from here on out.

Identification vs. Estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, let's consider the **population** diff-in-means:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

  - But, $\mathbb{P}$ is not directly observable since it's a population distribution!
- Once identified, we need to actually **estimate** the function of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
  - Now just estimating conditional expectations, etc.
  - $\rightsquigarrow$ **after identification, causal inference part done**
  - Purely a statistical question from here on out.
- Identification comes first, then comes estimation.
  - Without identification, properties of the estimator are unimportant.
  - keep them separate: estimator shouldn't drive identification.

What is Confounding?

- **Confounding**: treatment and potential outcomes are not independent!
  - Due to “common causes” of Y_i and D_i.
  - Main concern in observational studies.

What is Confounding?

- **Confounding**: treatment and potential outcomes are not independent!
  - Due to "common causes" of Y_i and D_i.
  - Main concern in observational studies.
- Pervasive in management/social sciences:
  - Effect of job training program on employment (confounder: motivation)

What is Confounding?

- **Confounding**: treatment and potential outcomes are not independent!
  - Due to “common causes” of Y_i and D_i.
  - Main concern in observational studies.
- Pervasive in management/social sciences:
  - Effect of job training program on employment (confounder: motivation)
  - Effect of college GPAs on salary (confounder: intelligence)

What is Confounding?

- **Confounding**: treatment and potential outcomes are not independent!
  - Due to “common causes” of Y_i and D_i.
  - Main concern in observational studies.
- Pervasive in management/social sciences:
  - Effect of job training program on employment (confounder: motivation)
  - Effect of college GPAs on salary (confounder: intelligence)
  - Effect of income on voting (confounder: age)

What is Confounding?

- **Confounding**: treatment and potential outcomes are not independent!
  - Due to "common causes" of Y_i and D_i.
  - Main concern in observational studies.
- Pervasive in management/social sciences:
  - Effect of job training program on employment (confounder: motivation)
  - Effect of college GPAs on salary (confounder: intelligence)
  - Effect of income on voting (confounder: age)
  - Effect of corporate giants on economic development (confounder: previous economic development)
- Confounding $\rightsquigarrow$ incomplete identification of ATE $\rightsquigarrow$ biased estimators.
- What to do?

2/ Selection on Observables

Observational Studies

- Many different types of identification assumptions we'll cover.

Observational Studies

- Many different types of identification assumptions we'll cover.
- Begin with most common observational assumptions:

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i | \mathbf{X}_i$
    - Also called: conditional unconfoundedness, weak ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeability, etc.
    - $\rightsquigarrow$ Conditional on some covariates, D_i is (effectively) randomly assigned.
2. **Positivity** or **Overlap**: $0 < P[D_i = 1 | \mathbf{X}_i] < 1$
    - Treatment and control are both possible at every value of $\mathbf{X}_i$
    - $\rightsquigarrow$ There are both treated and untreated units for each level of $\mathbf{X}_i$ (i.e., "common support").

Observational Studies

- Many different types of identification assumptions we'll cover.
- Begin with most common observational assumptions:
  1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i | \mathbf{X}_i$
     - Also called: conditional unconfoundedness, weak ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeability, etc.
     - $\rightsquigarrow$ Conditional on some covariates, D_i is (effectively) randomly assigned.
  2. **Positivity** or **Overlap**: $0 < P[D_i = 1 | \mathbf{X}_i] < 1$
     - Treatment and control are both possible at every value of $\mathbf{X}_i$
     - $\rightsquigarrow$ There are both treated and untreated units for each level of $\mathbf{X}_i$ (i.e., "common support").
- We'll take $\mathbf{X}_i$ as a 'given' for now and see later how we might choose it.

Observational Studies

- Many different types of identification assumptions we'll cover.
- Begin with most common observational assumptions:
  1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i | \mathbf{X}_i$
     - Also called: conditional unconfoundedness, weak ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeability, etc.
     - $\rightsquigarrow$ Conditional on some covariates, D_i is (effectively) randomly assigned.
  2. **Positivity** or **Overlap**: $0 < P[D_i = 1 | \mathbf{X}_i] < 1$
     - Treatment and control are both possible at every value of $\mathbf{X}_i$
     - $\rightsquigarrow$ There are both treated and untreated units for each level of $\mathbf{X}_i$ (i.e., "common support").
- We'll take $\mathbf{X}_i$ as a 'given' for now and see later how we might choose it.
- These are assumptions that **can be wrong!**

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) - Y_i(0) | \mathbf{X}_i]]
\end{aligned}
$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1)|\mathbf{X}_i] - \mathbb{E}[Y_i(0)|\mathbf{X}_i]]
\end{aligned}
$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1)|\mathbf{X}_i] - \mathbb{E}[Y_i(0)|\mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1)|D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0)|D_i = 0, \mathbf{X}_i]]
\end{aligned}$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) - Y_i(0) | \mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) | \mathbf{X}_i] - \mathbb{E}[Y_i(0) | \mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) | D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) | D_i = 0, \mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i | D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i | D_i = 0, \mathbf{X}_i]]
\end{aligned}
$$

- Useful to write the treated and control CEFs:

$$
\mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) | \mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) | \mathbf{X}_i = \mathbf{x}]
$$

    - How the mean of the potential outcomes vary with the covariates.

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) - Y_i(0) | \mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) | \mathbf{X}_i] - \mathbb{E}[Y_i(0) | \mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) | D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) | D_i = 0, \mathbf{X}_i]] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i | D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i | D_i = 0, \mathbf{X}_i]]
\end{aligned}
$$

- Useful to write the treated and control CEFs:

$$
\mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) | \mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) | \mathbf{X}_i = \mathbf{x}]
$$

  - How the mean of the potential outcomes vary with the covariates.

- Key part of the identification above:

$$
\underbrace{\mu_1(\mathbf{x})}_{\text{counterfactual}} = \underbrace{\mathbb{E}[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}]}_{\text{observational}}, \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}]
$$

Regression Estimation of the ATE

- Identification done, now turn to estimation!

Regression Estimation of the ATE

- Identification done, now turn to estimation!
- Regression estimators $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_1(\mathbf{x})$.
    - Might be linear or nonlinear models.
    - Safest practice: estimate separate regressions in each treatment group.
    - Sometimes called an **imputation** or **plug-in** estimator.

Regression Estimation of the ATE

- Identification done, now turn to estimation!
- Regression estimators $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_1(\mathbf{x})$.
  - Might be linear or nonlinear models.
  - Safest practice: estimate separate regressions in each treatment group.
  - Sometimes called an **imputation** or **plug-in** estimator.
- Regression/imputation estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

Regression Estimation of the ATE

- Identification done, now turn to estimation!
- Regression estimators $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_1(\mathbf{x})$.
  - Might be linear or nonlinear models.
  - Safest practice: estimate separate regressions in each treatment group.
  - Sometimes called an **imputation** or **plug-in** estimator.
- Regression/imputation estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- Procedure:
  1. Obtain predicted values for all units when $D_i = 1$.
  2. Obtain predicted values for all units when $D_i = 0$.
  3. Take the average difference between these predicted values.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\widehat{\mu}_1(x)$ and $\widehat{\mu}_0(x)$ are from the same OLS model $Y \sim D + X$.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\widehat{\mu}_1(x)$ and $\widehat{\mu}_0(x)$ are from the same OLS model $Y \sim D + X$.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i.
- Fully interacted OLS:
  - $\widehat{\mu}_1(x)$ and $\widehat{\mu}_0(x)$ are from fully interacted OLS with centered covariates.
  - The two CEFs are estimated separately with different slopes for treated and control.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
    - $\widehat{\mu}_1(x)$ and $\widehat{\mu}_0(x)$ are from the same OLS model $Y \sim D + X$.
    - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i.
- Fully interacted OLS:
    - $\widehat{\mu}_1(x)$ and $\widehat{\mu}_0(x)$ are from fully interacted OLS with centered covariates.
    - The two CEFs are estimated separately with different slopes for treated and control.
    - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i.
- These make two very different assumptions about the CEFs!

Variance Estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?

Variance Estimation

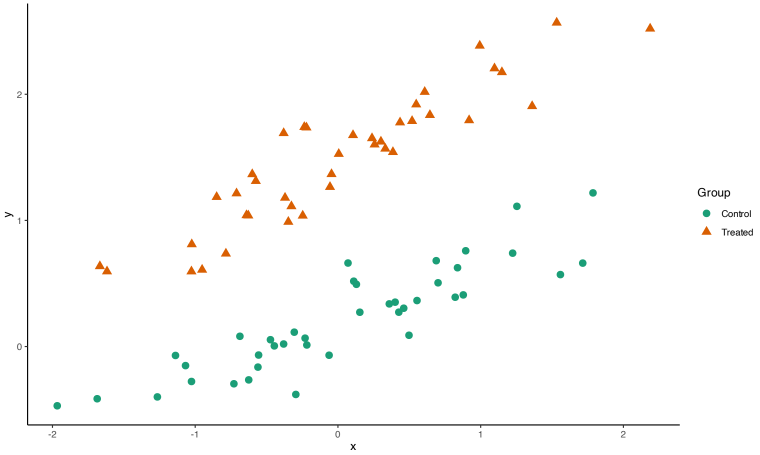
- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!

Variance Estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!
- Computational alternative: **(nonparametric) bootstrap**
    - Randomly resample n rows of the data with replacement.
    - Refit the regressions on the bootstrapped data.
    - Calculate $\widehat{\tau}_{\text{reg}}$ in each bootstrap.
    - Repeat several times and use empirical variance of the bootstraps.

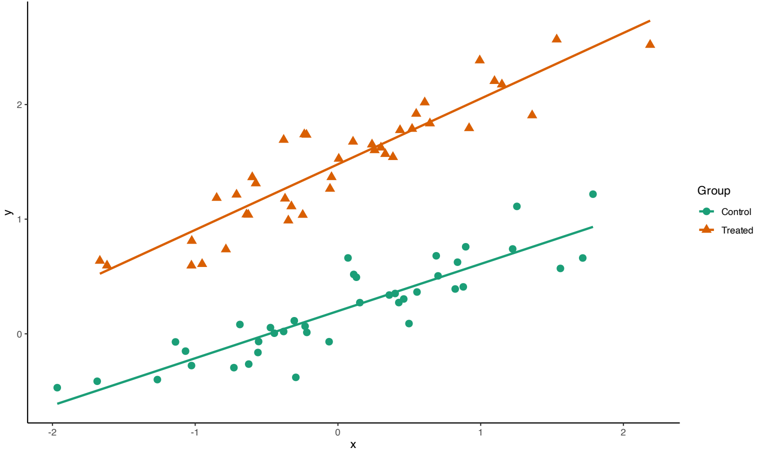
Imputation Estimator Visualization

```
1  > toy_data <- read_csv("https://bit.ly/3vOy2Ao")
```



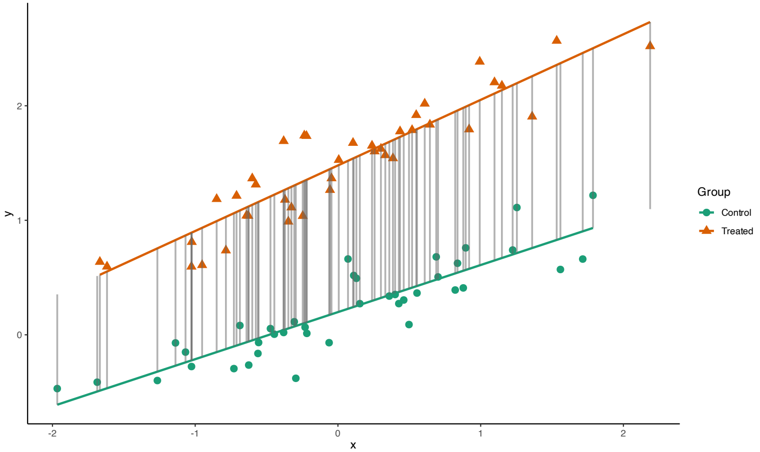
Imputation Estimator Visualization

```
> lm0 <- lm(y~x, data = toy_data, subset = d==0); lm1 <- lm(y~x, data = toy_data, subset = d==1)
```



Imputation Estimator Visualization

```
> mu0.imps = predict(lm0, toy_data); mu1.imps = predict(lm1, toy_data)
> cat("Estimate of ATE:", mean(mu1.imps - mu0.imps))
Estimate of ATE: 1.285176
```



Fully Interacted OLS & Imputation Estimator

- What if $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$ are from fully interacted OLS with centered covariates?
    - Equivalent to running separate models for $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$ (i.e., imputation estimator)

```
> toy_data$x_tilde <- toy_data$x - mean(toy_data$x)
> mod_full <- lm(y ~ d + x_tilde + d * x_tilde, data = toy_data)

> cat("\nEstimate of ATE (Imputation):", mean(mu1.imps - mu0.imps),
      "\nEstimated coefficient on Di from full int.", mod_full$coefficients["d"])

Estimate of ATE (Imputation): 1.285176
Estimated coefficient on Di from full int. 1.285176
```

- $\rightsquigarrow$ Recall: Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
    - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i.
    - Would be the same for uninteracted model, except the variance will be larger (less precision).

Variance Estimation w/ Bootstrap

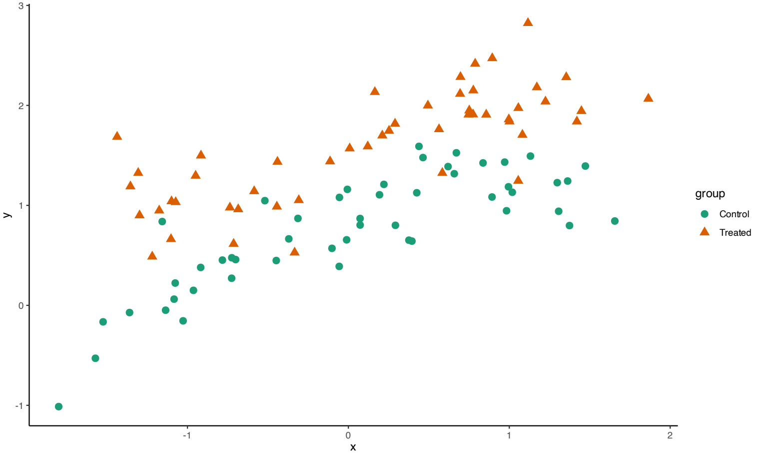
- Again, how do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
  - (Nonparametric) bootstrap: recall source of variance is due to **sampling**
  - Idea: view sample (data) as "population" $\rightsquigarrow$ in-sample "sampling"

```
> set.seed(02138); sims<-500; tau_hat_draws<-rep(NA, sims)
> for (i in 1:sims) { # Repeat the following several times
    # 1. Randomly resample n rows of the data with replacement
    sample_boot <- dplyr::slice_sample(toy_data, n = nrow(toy_data), replace = TRUE)

    # 2. Refit the regressions on the bootstrapped data
    model <- lm(y ~ d + x_tilde + d*x_tilde, data = toy_data)
    dat1 <- sample_boot; dat1$d <- 1
    dat0 <- sample_boot; dat0$d <- 0
    mu1_hat <- predict(model, newdata = dat1)
    mu0_hat <- predict(model, newdata = dat0)

    # 3. Calculate tau_hat in each bootstrap
    tau_hat_draws[i] <- mean(mu1_hat - mu0_hat)
  }

> # 4. Use empirical variance of the bootstraps
> var(tau_hat_draws)
[1] 0.000254049
```

Nonlinear Relationships

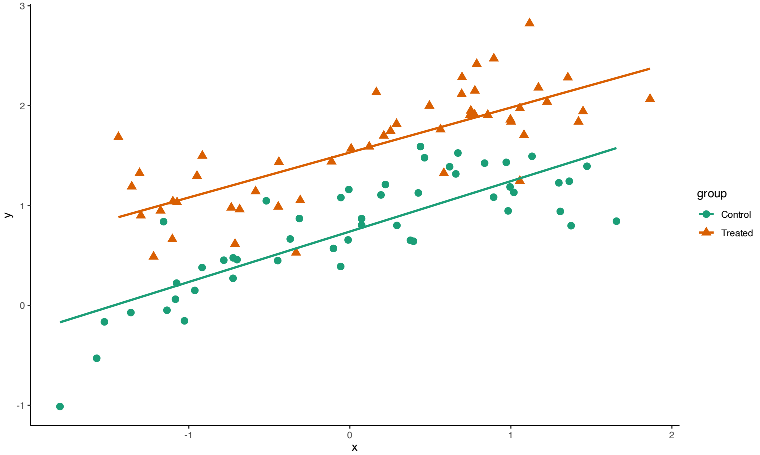
- Same idea but with nonlinear relationship between Y_i and X_i:

```
> toy_data_02 <- read_csv("https://bit.ly/4cOg0Pl") # sim data with nonlinear relationship
```



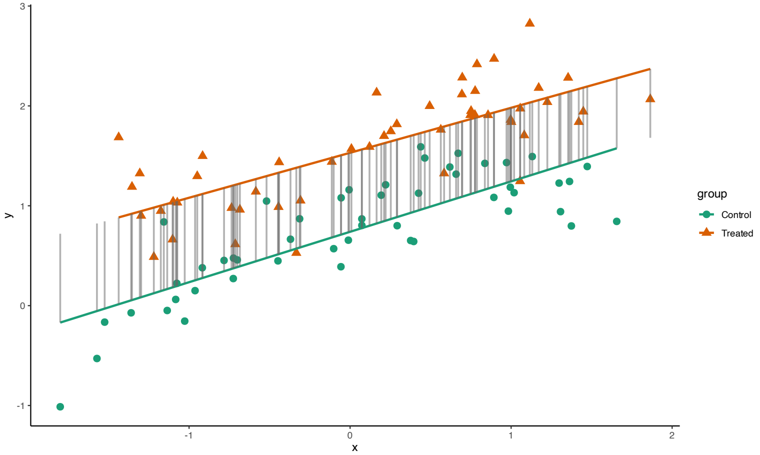
Nonlinear Relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



Nonlinear Relationships

- Same idea but with nonlinear relationship between Y_i and X_i:

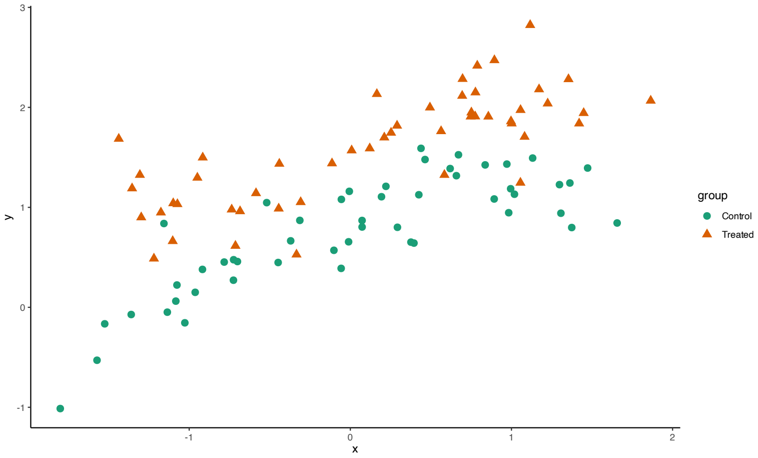


Using Semiparametric Regression

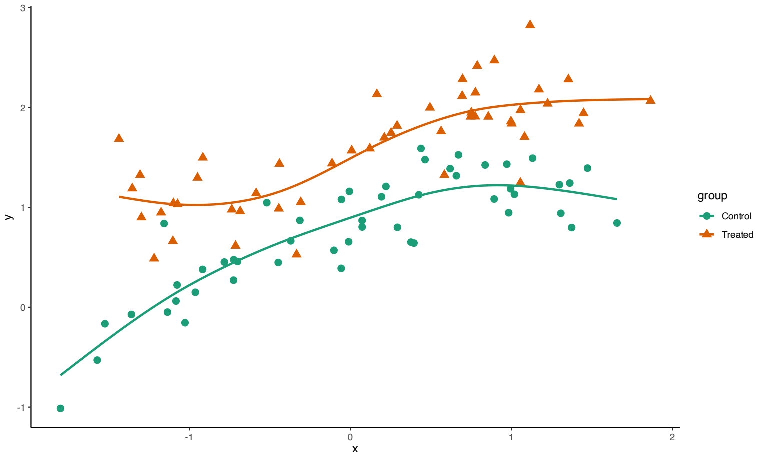
- Here, CEFs are nonlinear, but we don't know their form.
- We can use generalized additive models (GAMs) from the mgcv package for 'flexible' estimation:

```
> library(mgcv)
> gam1 <- gam(y~s(x), data = toy_data_02, subset = group=="Treated")
> gam0 <- gam(y~s(x), data = toy_data_02, subset = d==0); summary(gam0)

Family: gaussian
Link function: identity

Formula:
y ~ s(x)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2167     0.0307   7.059 2.05e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
     edf Ref.df      F p-value
s(x)   1      1 140.9  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.782   Deviance explained = 78.8%
GCV = 0.03969  Scale est. = 0.037705  n = 40
```

Using GAMS



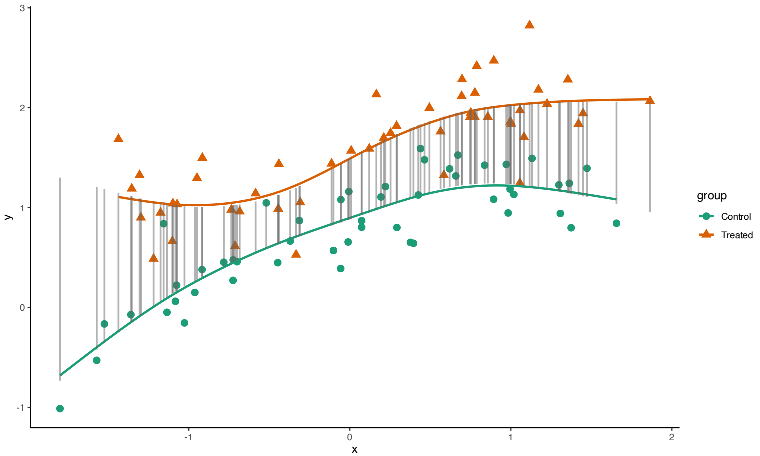
Using GAMS



Using GAMS

- We can estimate $\widehat{\tau}_{\text{reg}}$ using the imputation estimator.

```
> cat("Estimate of ATE (GAM):",mean(predict(gam1) - predict(gam0)))

Estimate of ATE (GAM): 0.8379884
```



Onto the presentations & discussions!

Contact Information:
jaewon.yoo@iss.nthu.edu.tw
https://j1yoo4.github.io/