ECON 7670: Empirical Methods

Elliott Isaac

Department of Economics
Tulane University

January 26, 2023

Empirical Methods

Introduction

- Many of the things we discuss constitute theory:
  - Utility functions, demand curves, and supply curves are theory
  - Changes in optimal bundles are theory
  - Equilibrium price and quantity changes are theory
- Understanding theory may tell us what to **expect** as the result of a policy change, but it won't tell us how much
  - Theory will tell us that price increases and quantity decreases when you levy a tax on producers
  - But it won't tell you how much price increases or how much quantity decreases without being more specific

Introduction

- The fundamental issue faced by empirical public economists is disentangling correlation from causality
  - **Correlated**: two economic variables are correlated if they move together (either in the same or opposite directions)
  - **Causal**: two economic variables are causally related if the movement of one causes movement of the other

Correlation Versus Causation

Example from Fisher (1976)

There was once a cholera epidemic in Russia. The government, in an effort to stem the disease, sent doctors to the worst-affected areas. The peasants of a particular province observed a very high correlation between the number of doctors in a given area and the incidence of cholera in that area. Relying on this fact, they banded together and murdered their doctors.

Correlation Versus Causation

- The identification problem: given that two variables are correlated, how do you identify whether one variable is causing another?
    - Note: identification is not a statistical property; it is an outcome of your empirical approach
- Example: Harvard freshmen in 1988 who took SAT preparation courses scored, on average, 63 points lower than those who hadn't
- Why?

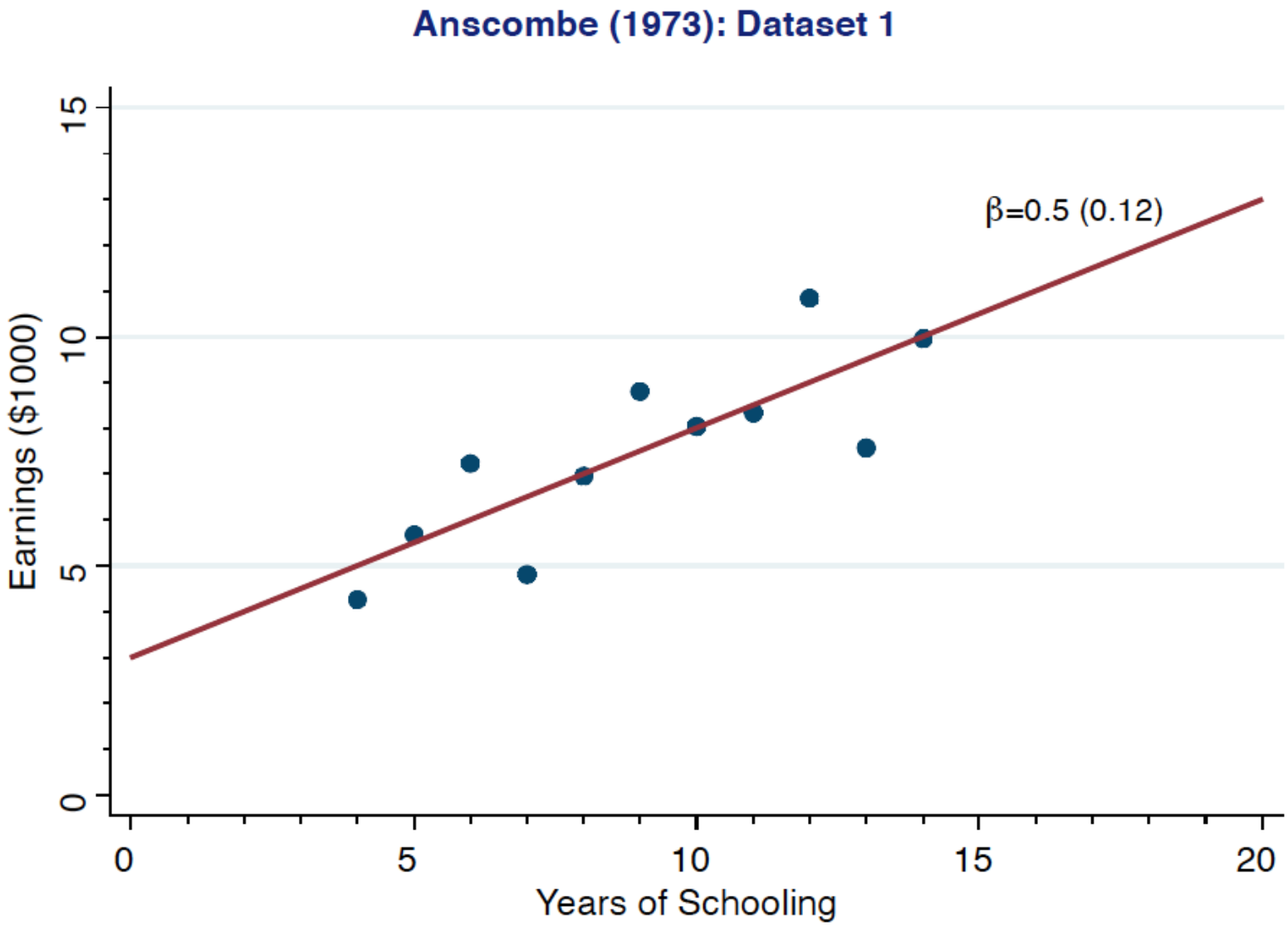
Correlation Versus Causation

- The identification problem: given that two variables are correlated, how do you identify whether one variable is causing another?
    - Note: identification is not a statistical property; it is an outcome of your empirical approach
- Example: Harvard freshmen in 1988 who took SAT preparation courses scored, on average, 63 points lower than those who hadn't
- Why?
- The students who needed the most help were more likely to take the SAT preparation courses
- The course did not **cause** students to do worse; rather, students who were more likely to do worse were the ones who took the course

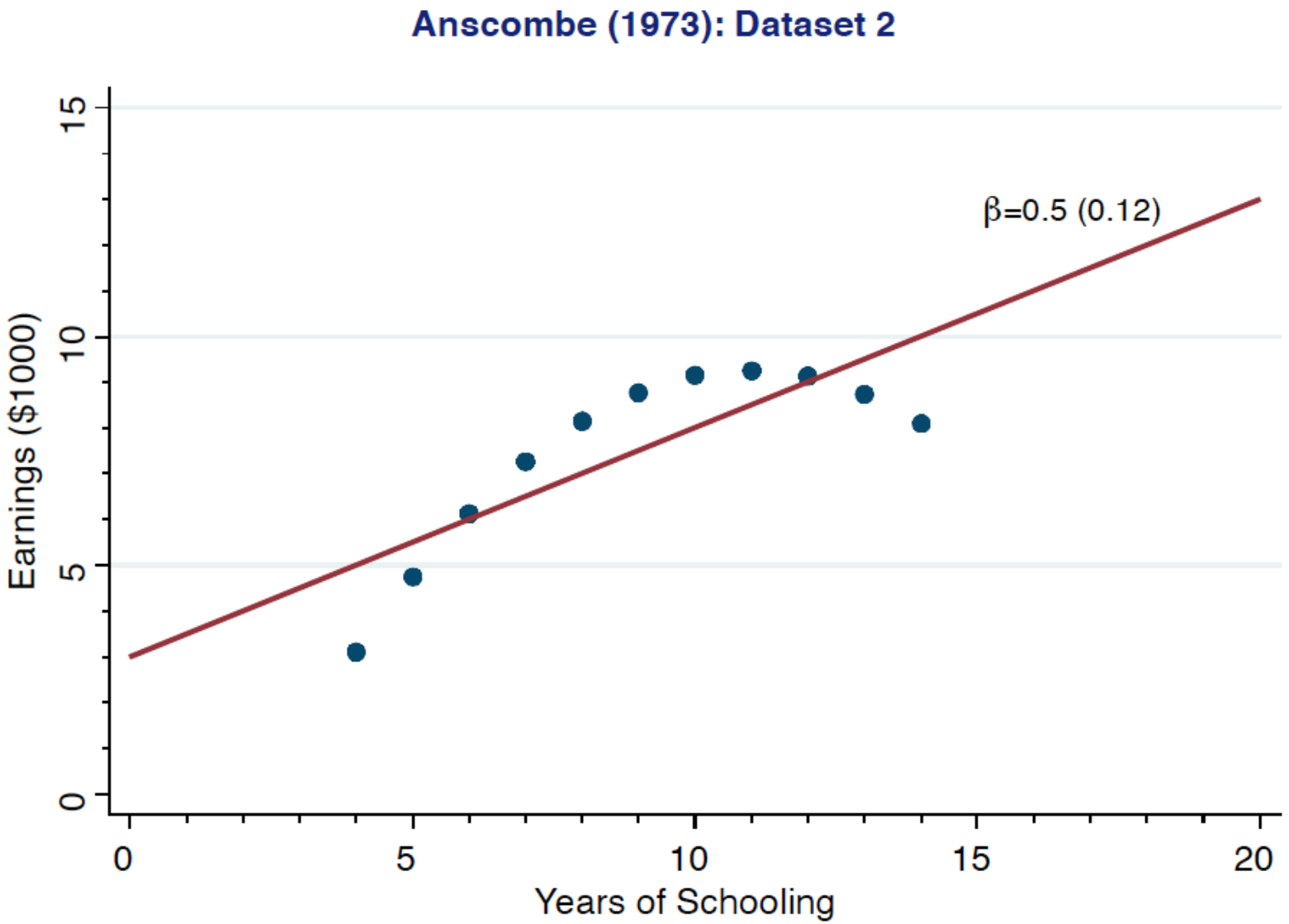
The Identification Problem

- Link to video example
- In above example, we are not able to determine the true relationship between the movements
- In other words, we cannot **identify** the true causal path

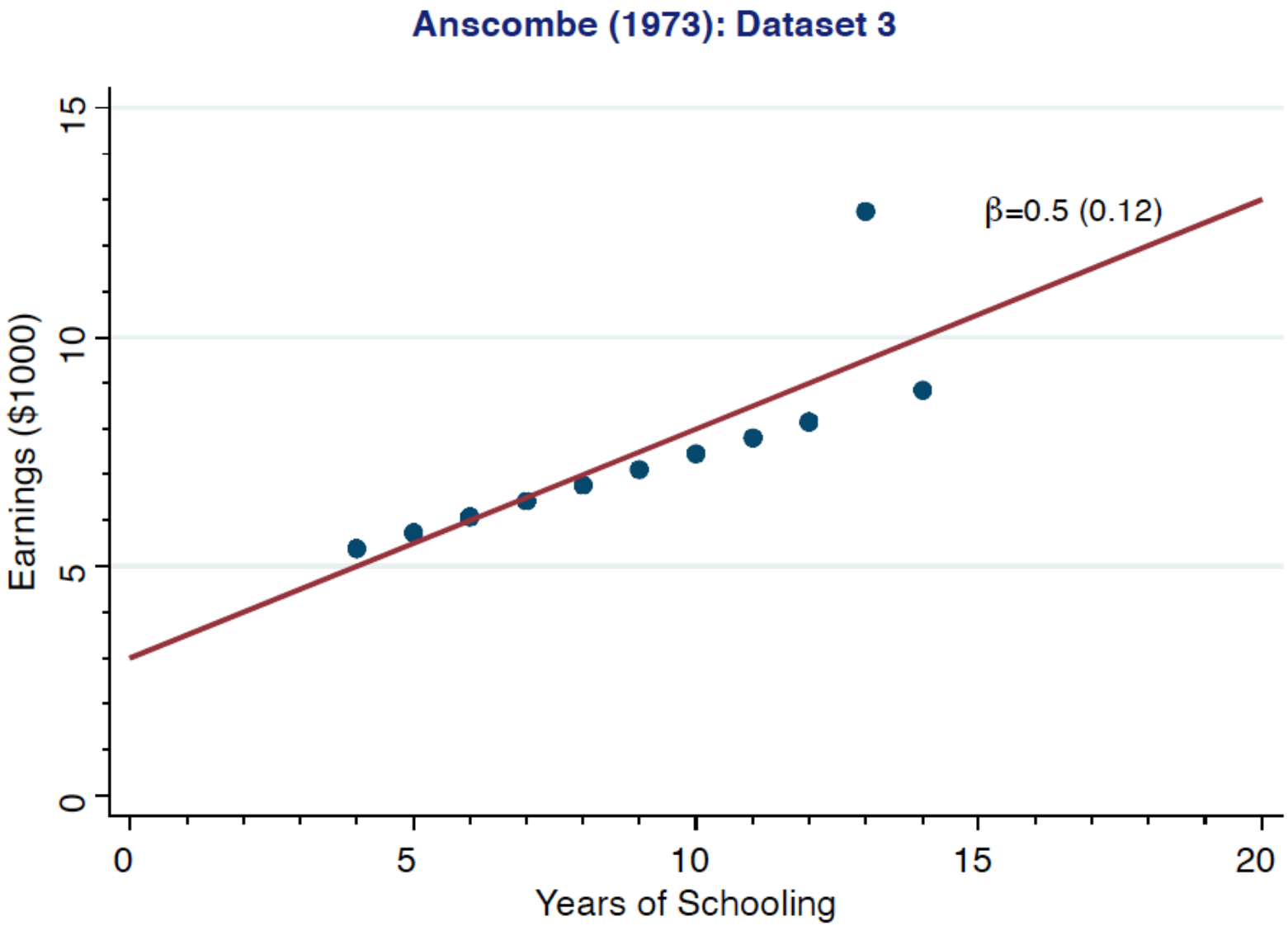
The Identification Problem



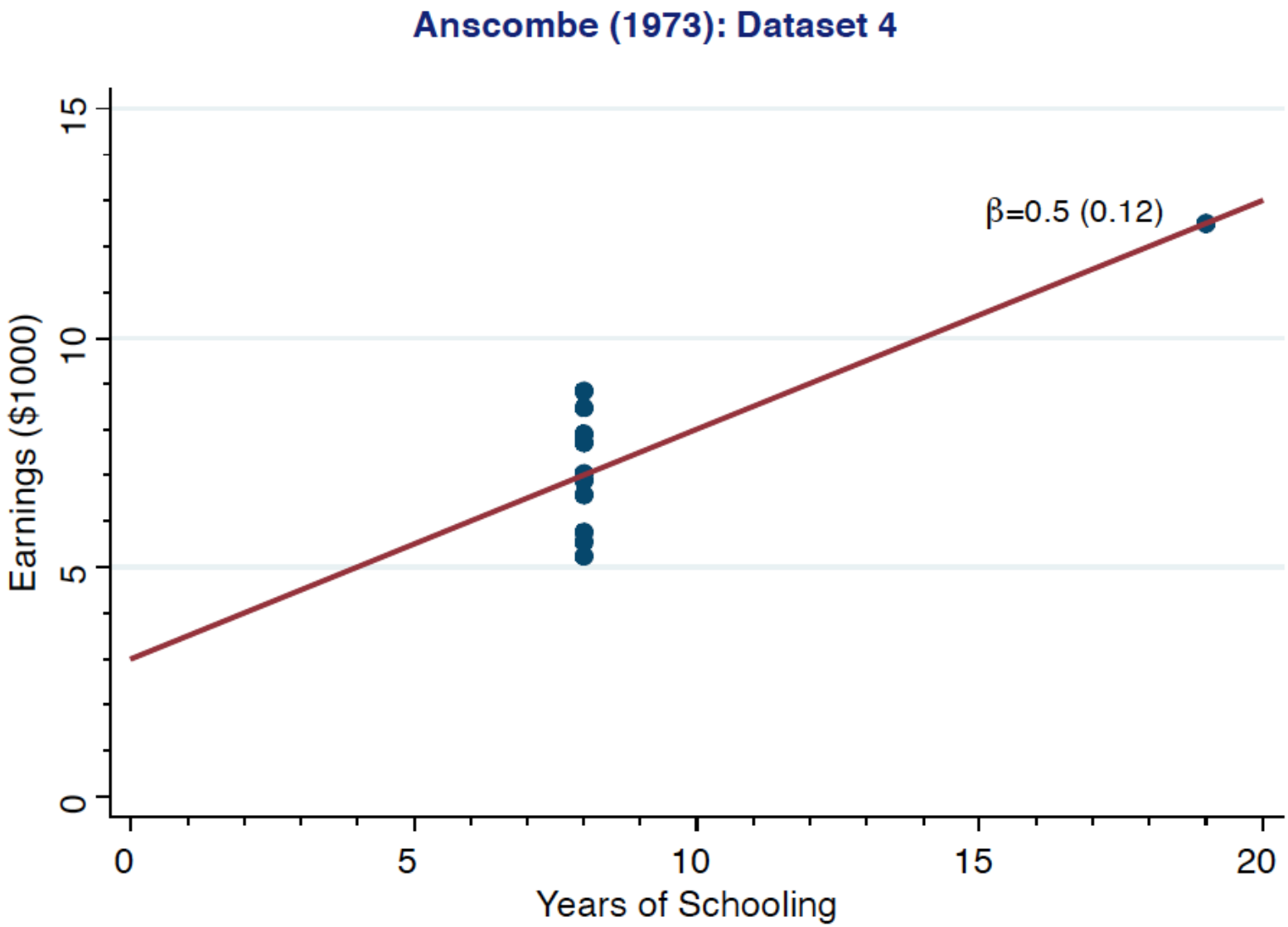
The Identification Problem



The Identification Problem



The Identification Problem



Correlation Versus Causation

- For any correlation between two variables A and B, there are four possible explanations that could result in correlation:
  - A is causing B
  - B is causing A
  - Some third factor is causing both A and B
  - Some or all of the above
- Policy makers often want to know empirical results to predict how government interventions will affect behaviors
- But knowing correlations provides no predictive power
- Prediction requires an understanding of the causal links between factors

Developing Empirical Research Designs

- Consider ideal experimental design first
- Then formulate a feasible design and analyze its flaws relative to the ideal design
- Frontier for empirical papers: tradeoff between quality of research design and importance/novelty

Developing Empirical Research Designs

- Empirical studies often start by formulating clear research designs
- Why develop an explicit design rather than simply use all available variation in, for example, tax rates?
- Consider estimating the effect of a treatment (e.g., a tax) T on outcome y

$$y_i = \alpha + \beta T_i + \varepsilon_i$$

- Treatment is assigned based on a "selection" model

$$T_i = \alpha_T + \beta_T X_i + \eta_i$$

- Treatment may be non-random: $cov(X_i, \varepsilon_i) \neq 0$, $cov(\eta_i, \varepsilon_i) \neq 0$

Developing Empirical Research Designs

- Traditional approach to accounting for confounding factors or selection: control for observables X_i when estimating treatment effect

$$y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i \tag{1}$$

- Can be done using OLS regression, matching, propensity-score reweighting, etc.
- Problem with these approaches: don't know the source of variation in T_i
  - Must be some reason that one person got treated an another did not, even if they are perfectly matched on observables
  - η_i must be correlated with T_i to have variation in $T_i|X_i$
  - But that same unobserved factor could also affect the outcome: no way to know if $cov(\eta_i, \varepsilon_i) = 0$
  - Cannot be sure that estimate of treatment effect (β) is consistent

Developing Empirical Research Designs

- A "research design" is a source of variation in η_i that is credibly unrelated to ε_i
  - Ex.: a reform that affects people above age 65 but not below
  - People at age 64 and 65 are likely to have similar outcomes $\rightarrow cov(\eta_i, \varepsilon_i) = 0$
- General lesson: "controlling" for confounding factors using regression or reweighting will rarely give you convincing estimates
- However, reweighting can be a useful technique to obtain better control groups when paired with a quasi-experimental design

Causal Inference

In an Ideal World...

- In an ideal world, we could:
  - Clone a group of individuals
  - Place the clones in a parallel universe where the only difference is the treatment of interest
  - Observe the differences between the originals and the clones
  - Any differences would be attributed to the treatment of interest because everything else is identical
  - Obviously, this is not possible, but we can come close through randomized trials

Randomized Control Trials

- **Randomized control trial**: the ideal type of experiment designed to test causality, whereby a group of individuals is randomly divided into a treatment group, which receives the treatment of interest, and a control group, which does not
  - **Treatment group**: the set of individuals who are subject to an intervention being studied
  - **Control group**: the set of individuals comparable to the treatment group who are not subject to the intervention being studied
- In essence, trial participants are assigned to treatment or control by the flip of a coin
- Almost any empirical problem we discuss can be thought of as a comparison between treatment and control groups

Bias

- **Bias**: any source of difference between treatment and control groups that is correlated with the treatment but is not due to the treatment
    - Example: Those taking SAT preparation courses may be of lower test-taking ability than those not taking the courses

Bias

- **Bias**: any source of difference between treatment and control groups that is correlated with the treatment but is not due to the treatment
    - Example: Those taking SAT preparation courses may be of lower test-taking ability than those not taking the courses
- This kind of bias does not exist in randomized trials because the only difference between treatment and control groups is the flip of a coin
- The coin flip/treatment is not consistently related to any other characteristic difference between the groups

Issues with Randomized Trials

- Ethical issues
  - Being in the treatment or control group may harm volunteers
- The results are only relevant for the population of volunteers, who may differ in substantial ways from the overall population of interest
- Attrition bias is still a concern
  - **Attrition**: reduction in the size of samples over time, which, if not random, can lead to biased estimates
  - Example: what if everyone who is negatively affected by the treatment leaves the experiment? Your results will show the treatment has a positive effect because all the negative effects are gone
- Randomized trials can be very expensive

Causal Inference with Observational Data

Observational Data

- Randomized trials are not always feasible
- Most often, we have observational data instead
  - **Observational data**: data generated by individual behavior observed in the real world, not in the context of deliberately designed experiments
- We often conceptualize these techniques in terms of **treatment** and **control** groups
  - **Treatment group**: individuals affected by the policy
  - **Control group**: individuals unaffected by the policy
  - Our main concern is to remove any sources of bias between the two groups
    - i.e., remove any differences between the groups that might affect their outcomes **other than the treatment**

Quasi-Experiments

- Quasi-experiments have become an increasingly common middle ground between randomized trials and cross-sectional regression analysis

- **Quasi-experiments**: changes in the economic environment that create nearly identical treatment and control groups for studying the effect of that environmental change, allowing public economists to take advantage of randomization created by external forces

Difference-in-Differences

- Example: Suppose we are interested in estimating the causal effect of TANF generosity on labor supply
  - A large portion of TANF recipients are single mothers
  - We could construct a sample of single mothers from Arkansas and Louisiana for 2007, 2008, and 2009
  - In 2008, Arkansas raised its TANF benefit guarantee by 20% and Louisiana left their program unchanged
  - Single mothers in Arkansas are our **treatment group** and single mothers in Louisiana are our **control group**
  - By comparing single mothers' labor supply between 2007–2009 across Arkansas and Louisiana, we can obtain an estimate of the impact of TANF on labor supply

Difference-in-Differences

TANF Example

In theory, we could estimate an OLS regression (estimation method) of the effect of TANF generosity (explanatory variable) on hours worked (outcome variable) of single mothers in Arkansas only (treatment group). So long as the sample of single mothers remains identical from 2007–2009 and the only change in Arkansas over this time period is the increase in TANF benefit guarantee, then any estimated change in hours worked would reflect only the causal effect of TANF benefits.

- This is an **identification argument**
  - It is a statement that provides a claim and justification for why your regression coefficient is **actually** the true causal effect you are trying to estimate
  - Identification of a causal effect is an argument to be made, not a statistical property
- Whether this is a **valid** identification argument is another question

Difference-in-Differences

TANF Example

In theory, we could estimate an OLS regression (estimation method) of the effect of TANF generosity (explanatory variable) on hours worked (outcome variable) of single mothers in Arkansas only (treatment group). So long as the sample of single mothers remains identical from 2007–2009 and the only change in Arkansas over this time period is the increase in TANF benefit guarantee, then any estimated change in hours worked would reflect only the causal effect of TANF benefits.

- Do you believe this identification argument? Why or why not?

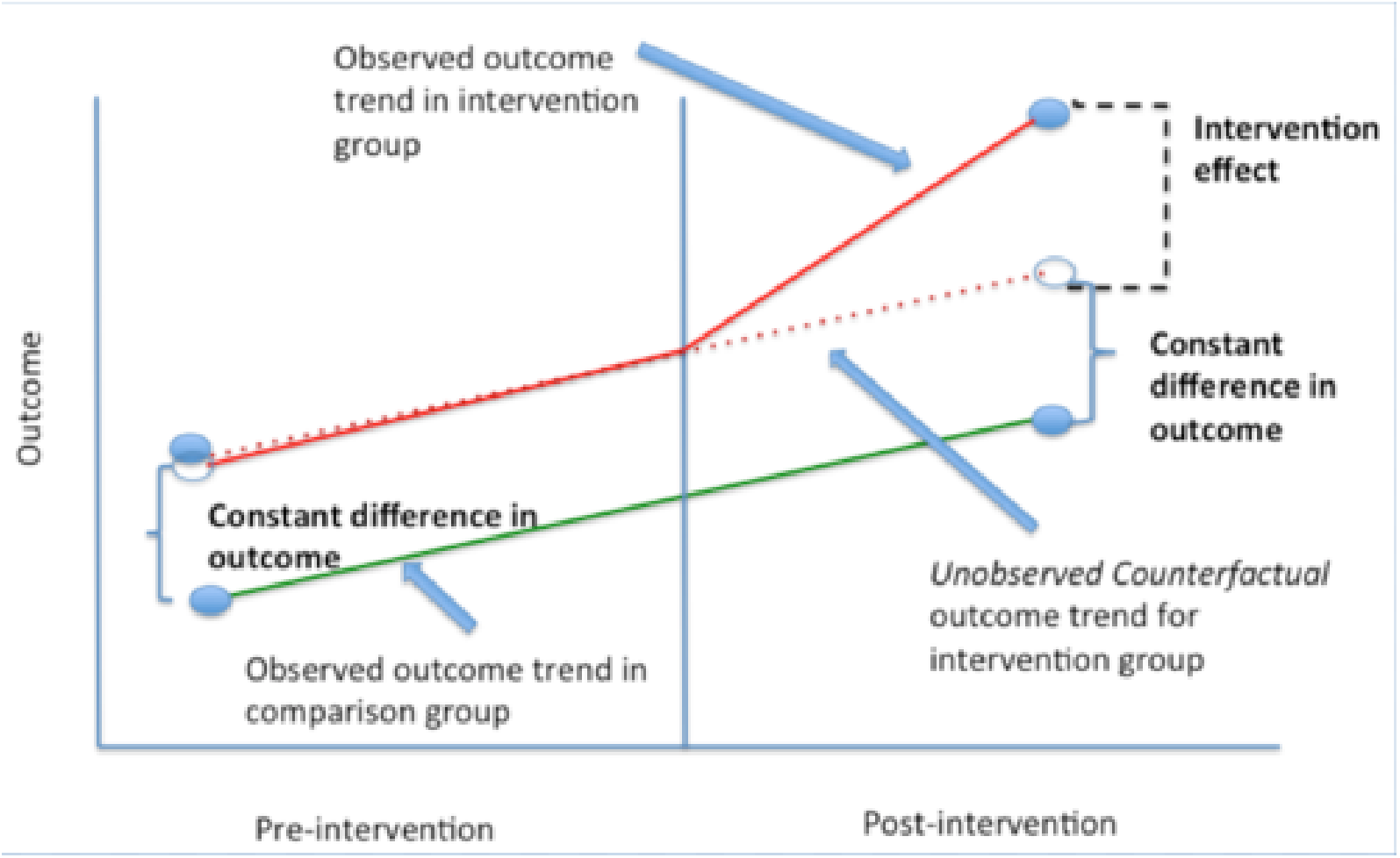
Difference-in-Differences

TANF Example

In theory, we could estimate an OLS regression (estimation method) of the effect of TANF generosity (explanatory variable) on hours worked (outcome variable) of single mothers in Arkansas only (treatment group). So long as the sample of single mothers remains identical from 2007–2009 and the only change in Arkansas over this time period is the increase in TANF benefit guarantee, then any estimated change in hours worked would reflect only the causal effect of TANF benefits.

- Do you believe this identification argument? Why or why not?
- 2007–2009 was during the Great Recession
- It seems likely that single mothers' labor supply changed in Arkansas between 2007–2009 for reasons other than TANF
- So comparing before/after among the treatment group only would result in biased estimates

Difference-in-Differences



Difference-in-Differences

- Single mothers in Arkansas and Louisiana both experienced the Great Recession
- If some reduction in labor supply among single mothers in Arkansas is driven by the Great Recession, then we should see the same reduction in labor supply among single mothers in Louisiana:

Difference-in-Differences

- Single mothers in Arkansas and Louisiana both experienced the Great Recession
- If some reduction in labor supply among single mothers in Arkansas is driven by the Great Recession, then we should see the same reduction in labor supply among single mothers in Louisiana:

$$\begin{aligned} \text{Hours}(AR, 2009) - \text{Hours}(AR, 2007) &= \text{Treatment effect} \\ &\quad + \text{Bias from Great Recession} \\ \text{Hours}(LA, 2009) - \text{Hours}(LA, 2007) &= \text{Bias from Great Recession} \\ \hline \text{Difference} &= \text{Treatment effect} \end{aligned} \tag{2}$$

Difference-in-Differences

- The previous technique is called a difference-in-differences (DD) estimator
- **Difference-in-differences estimator**: the difference between the changes in outcomes for the treatment group that experiences an intervention and the control group that does not

Difference-in-Differences

- The previous technique is called a difference-in-differences (DD) estimator
- **Difference-in-differences estimator**: the difference between the changes in outcomes for the treatment group that experiences an intervention and the control group that does not
- DD estimators combine time series and cross-sectional analyses to address problems with each
  - Time series: DD estimators use policy changes to compare variables over time
  - Cross-sectional: DD estimates use multiple cross-sections over time to control for differences between groups

Difference-in-Differences

- The DD estimator requires a parallel trends assumption for identification
- **Parallel trends assumption**: the differences over time between the treatment and control groups moved parallel to each other before the policy change and would have remained parallel if the treatment had never occurred
- What if the Great Recession affected Arkansas and Louisiana differently?

Difference-in-Differences

$$
\begin{aligned}
\text{Hours}(AR, 2009) - \text{Hours}(AR, 2007) =& \text{Treatment effect} \\
&+ \text{AR bias from Great Recession} \\
\text{Hours}(LA, 2009) - \text{Hours}(LA, 2007) =& \text{LA bias from Great Recession} \\
& \text{---------------------} \\
\text{Difference} =& \text{Treatment effect} + \\
& \text{(AR bias - LA bias)}
\end{aligned}
\tag{3}
$$

- If the trends between the two groups are not parallel, then we end up with a biased result
- Quasi-experimental studies often:
  1. Make an argument that they have removed bias between the groups
  2. Use additional/alternative control groups to confirm the bias has been removed

Estimating Difference-in-Differences

The standard difference-in-differences set-up requires:

- Two periods of data: pre- and post-policy change
- Two groups: treated and untreated
- A variable, *Treat*, which equals 1 if the individual is in the treatment group (regardless of the period)

```
gen Treat = (treated == 1)
```

- A variable, *Post*, which equals 1 for all individuals observed in the post-period
  - `gen Post = (year >= policy_change_year)`

Estimating a Difference-in-Differences

Then estimate a simple OLS regression (using `i.` notation):

```
reg Y i.Treat#i.Post Treat Post X
```

- Or manually create the *Treat* $\times$ *Post* variable and use it directly:

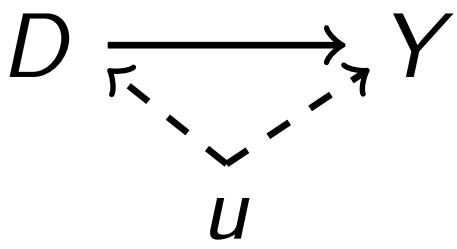
```
gen Treat_Post = Treat*Post
reg Y Treat_Post Treat Post X
```

Estimating Difference-in-Differences

- Difference-in-differences becomes more complicated if:
  - Treated individuals are treated at different times (staggered treatment)
  - All individuals eventually become treated
  - Treatment effect varies by time or sub-groups (heterogeneous treatment)
  - Trends do not appear to be parallel between treatment and control groups
- Wooldridge (2021) is a good starting point
- Callaway and Sant'Anna (2021), Goodman-Bacon (2021), and Sun and Abraham (2021) are now canonical in this literature

Instrumental Variables

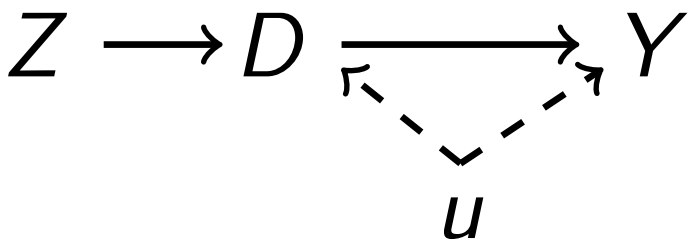
- We are interested in estimating the causal effect of a treatment (D) on the outcome (Y)
- But there are often unobserved factors that affect both whether an individual is treated and the value of their outcome:



- If you run the regression $Y = \beta_0 + \beta_1 D + \varepsilon$, then u will bias your coefficient because it affects both D and Y
- The coefficient you get would not be causal

Instrumental Variables

- An instrument (Z) is a variable that affects Y *only because it affects* D
- In other words:



- Example: Suppose D consists of people making choices
  - Sometimes changes in D affect Y, and other times changes in D reflect changes in Y via u
  - Z induces *some*, but not all, of D to change
  - $\Rightarrow$ Y changes *because* D changes
  - This is the causal effect

Estimating an IV Coefficient

- You can run 2 sets of OLS regressions for the first and second stage, plus the `predict` command (not recommended):

```
reg D Z X
predict D_hat
reg Y D_hat X
```

- Or you can use the user-built `ivreg2` command in Stata to estimate first stage, reduced form, and second stage in a single step (recommended):

```
ivreg2 Y X (D = Z), savefirst saverf
```

- As a check, you should get an identical coefficient through the following process:
  1. Estimate the first stage (OLS): $D = \alpha_0 + \alpha_1 Z + \alpha_2 X + \varepsilon$
  2. Estimate the reduced form (OLS): $Y = \gamma_0 + \gamma_1 Z + \gamma_2 X + \mu$
  3. Divide to get the causal effect: $\beta_1 = \frac{\gamma_1}{\alpha_1}$

Estimating an IV Coefficient

The standard IV set-up requires an instrument, Z, that satisfies:

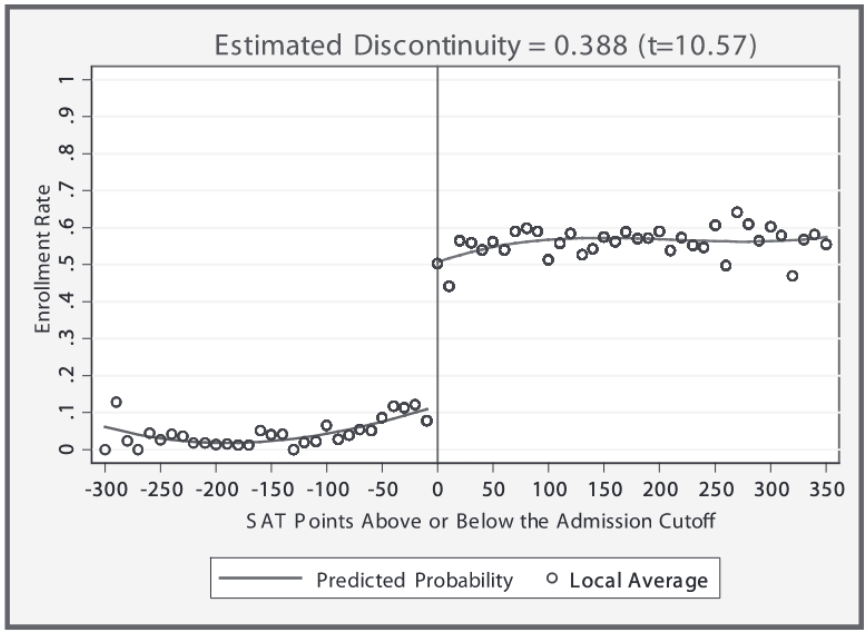
1. Relevance: The instrument has a causal effect on treatment status, D
2. Independence of Y: The instrument is as good as randomly assigned
3. Exclusion restriction: The instrument does not have its own direct effect on Y
4. Monotonicity: All those who are affected by the instrument are affected in the same way

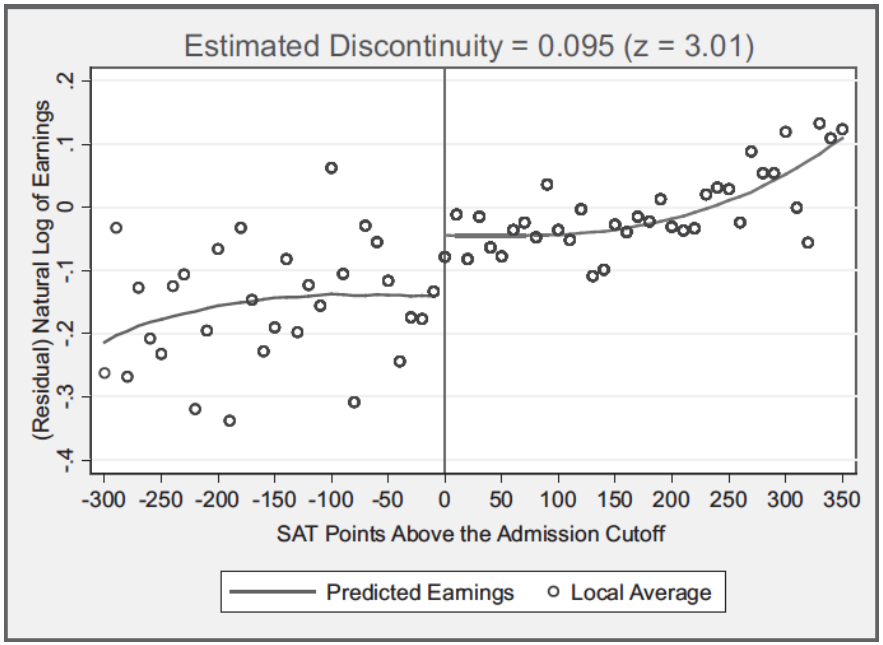
Regression Discontinuity

- Sometimes there is a threshold that determines whether you are treated or not
- Example: Suppose we are interested in estimating the causal effect of attending the state flagship university on earnings after college
  - Note: We are interested in the effect of attending the school, but enrolling at the school is a choice students make
  - The school uses an SAT threshold to determine admission, which affects whether the student can/will enroll
  - If we zoom in close to the threshold then it might be reasonable to assume that small deviations in SAT score, which place a student on one side or the other of the threshold, are random
  - Assuming the side of the threshold a student ends up on is random (conditional on zooming in close to the threshold) allows us to interpret the effect of being above the threshold on earnings as causal (identification)

Regression Discontinuity

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY





- Enrollment discontinuity is the first stage
- Earnings discontinuity is the reduced form

Source: Hoekstra (2009)

Estimating a Regression Discontinuity

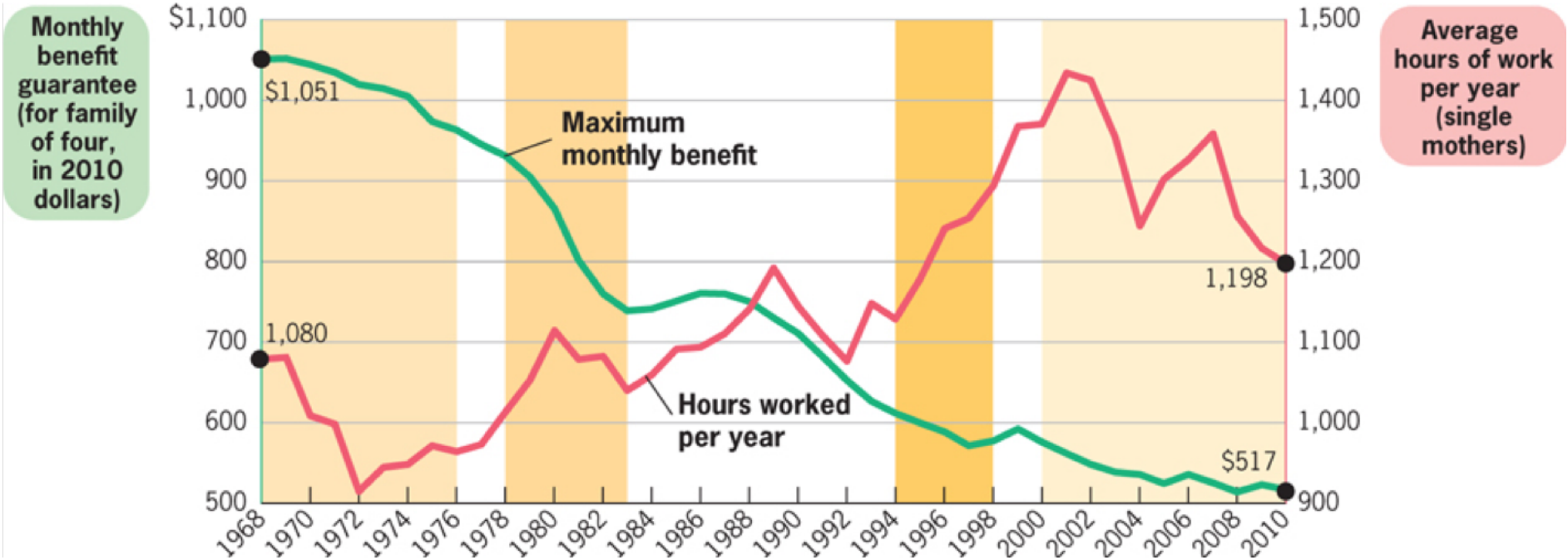
- To estimate an RD you need:
  - A "running variable"
  - A known threshold in the running variable, or a convincing way to find it
  - No manipulation: Units are not able to intentionally move themselves from one side of the threshold to the other
- Calonico, Cattaneo, and Titiunik (2014) and Calonico et al. (2017) are good starting points and include user-written Stata commands: `rdrobust` and `rdplot`

Not-So-Causal-Inference Methods

Time Series Analysis

- **Time series analysis**: analysis of the comovement of two series (variables) over time
- We could gather data over time on the TANF benefit guarantee each year and compare these data to the amount of labor supply delivered by single mothers in the same years
- **Key**: Time series analysis uses variation over time

Time Series Analysis



- The time series shows a negative relationship between benefit guarantees and labor supply

Issues with Time Series Analysis

- There are a number of reasons (other than causation) why single mothers may work more today than in 1968:
  - Changing societal norms
  - Better and more options for child care
  - Declining marriage rates
- Examining subperiods of this time series may reveal different relationships

Time Series Analysis

- Time series analysis is useful when there are clear factors that change
- To claim a causal relationship, researchers must be convinced that there are not other factors at play
- Believe it or not, arguments about causality often come down to researchers trying to convince each other that there are not other confounding factors
- Identification of a causal effect is an argument to be made, not a statistical property

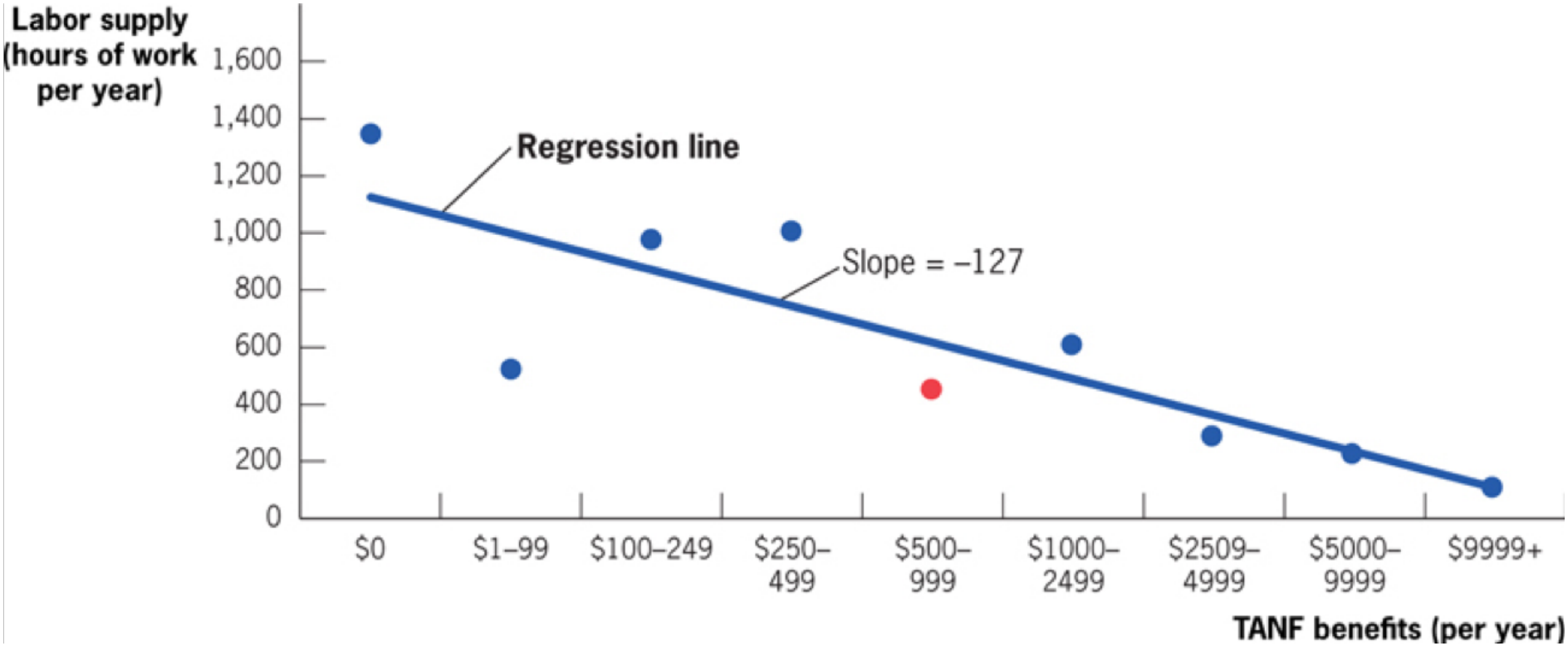
Cross-Sectional Regression Analysis

- **Cross-sectional regression analysis**: statistical analysis of the relationship between two or more variables exhibited by many individuals at one point in time
- In its simplest form, we would gather data for two variables:
  - **Dependent variable**: how many hours each individual works
  - **Independent variable**: the TANF benefit each individual faces
- Regression analysis finds the line that best fits the data, and measures the slope of that line
- **Key**: Cross-sectional regression analysis uses variation over units of observation

Cross-Sectional Regression Analysis

- Example: Current Population Survey (CPS)
  - The CPS collects information each month from individuals in the U.S. on a variety of economic and demographic issues
  - Every March, a special supplement asks respondents about their sources of income and hours of work last year
  - We can take all the single mothers from the CPS to create a dataset of hours worked and TANF benefits

Cross-Sectional Regression Analysis



- The regression line shows the best linear approximation to the relationship between TANF benefits and labor supply for these data
- In words: Each doubling of TANF benefits reduces labor supply by 127 hours

Issues with Cross-Sectional Regression Analysis

- It is possible that there is a third factor causing changes to labor supply and TANF benefits:

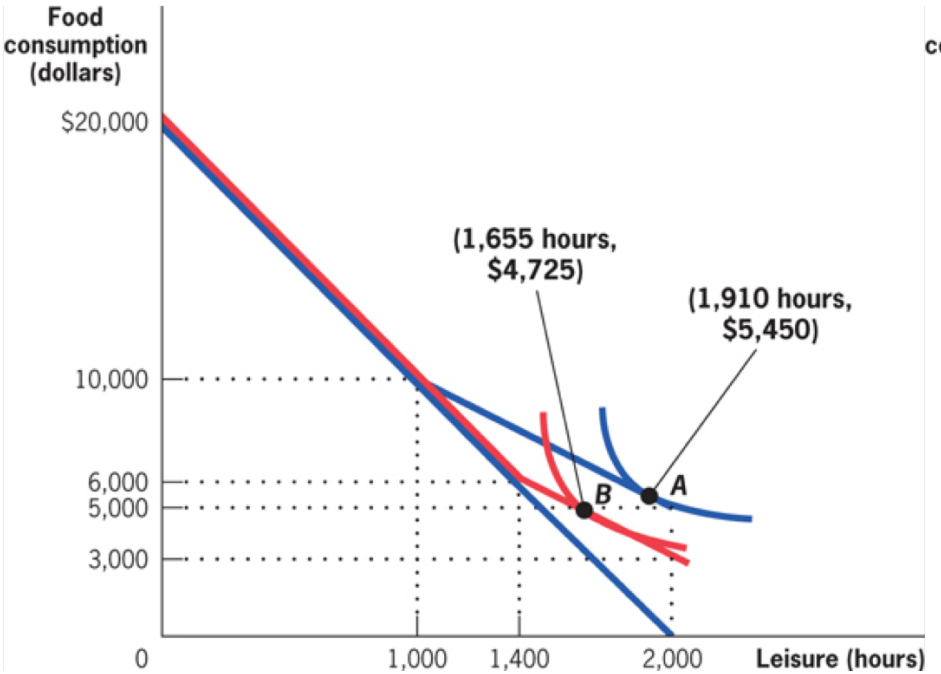
Issues with Cross-Sectional Regression Analysis

- It is possible that there is a third factor causing changes to labor supply and TANF benefits: **preferences for leisure**

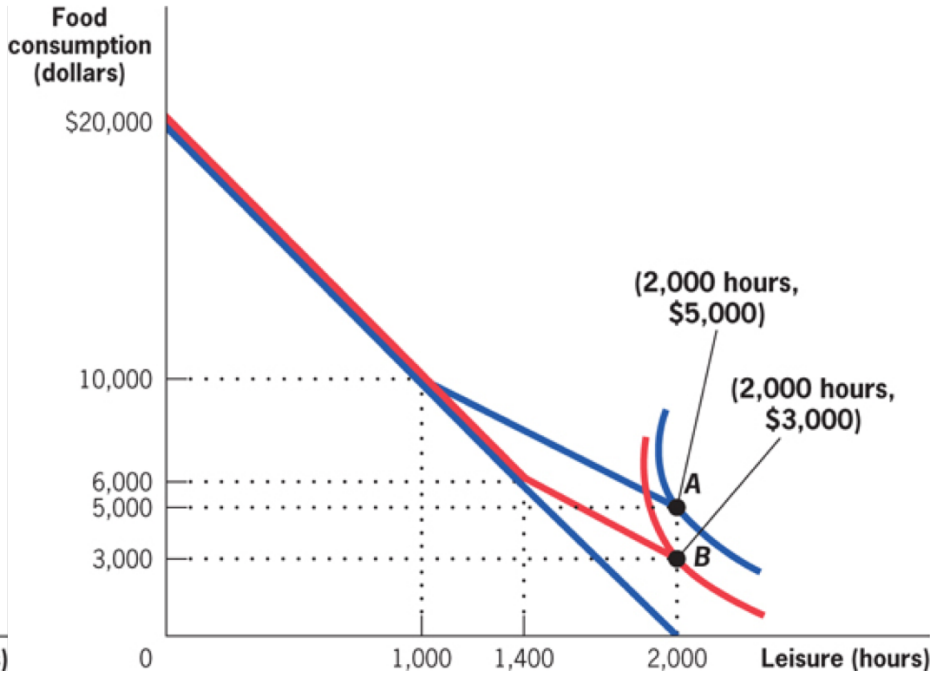
Issues with Cross-Sectional Regression Analysis

- It is possible that there is a third factor causing changes to labor supply and TANF benefits: **preferences for leisure**
- Because TANF benefits decrease as the recipient works more, individuals who take more leisure automatically get higher TANF benefits
- TANF benefits and leisure may be correlated because higher leisure is causing higher TANF benefits, not the other way around

Example: Preferences for Leisure



Sarah



Naomi

- Sarah works for 90 hours, earns \$900, and receives \$4,550 from TANF (blue lines)
- Naomi does not work, earns \$0, and receives \$5,000 from TANF (blue lines)

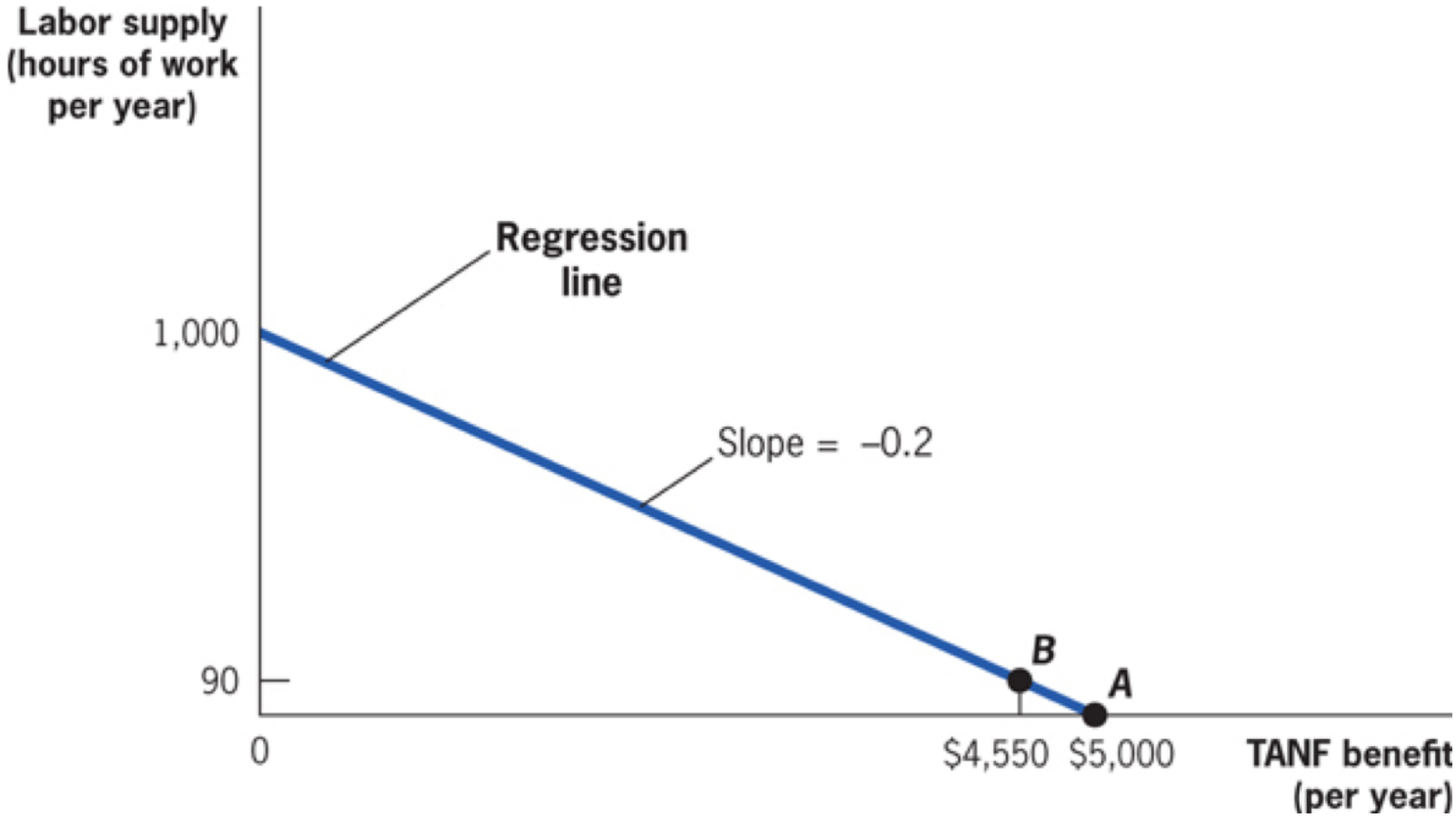
Example: Preferences for Leisure

- A cross-sectional dataset of Sarah and Naomi would look like:

Table:

Individual	**Hours of work**	**TANF benefits**
Sarah	90	4,550
Naomi	0	5,000

Example: Preferences for Leisure



- If we run a regression on data from Sarah and Naomi, we would again conclude that TANF and labor supply are negatively related
- But this relationship is not causal because it is partly driven by preferences

Example: Preferences for Leisure

- What did we learn?
- Differences in preferences for leisure between Sarah and Naomi created differences in labor supply, which created differences in TANF benefits
  - Naomi received higher TANF benefits
  - But part of her labor supply reduction is due to preferences, too

Control Variables

- One advantage of regression analysis is the ability to include control variables
- **Control variables**: variables that are included in cross-sectional regression models to account for differences between treatment and control groups that can lead to bias
- Control variables attempt to control (take into account) other differences across individuals in a sample
- Then any remaining correlation between the dependent & independent variables can be interpreted as a causal effect

Control Variables

- For example:
  - Suppose the CPS included a variable called "tastes for leisure"
  - This variable has two categories: "prefers leisure" and "prefers work"
  - Assume individuals within each category have identical preferences for leisure/work
  - We could divide the sample by preferences for leisure and re-do the analysis within each group
  - Within each group, different preferences for leisure cannot confound the relationship between TANF benefits and labor supply because preferences are identical within each group

Control Variables

- In reality, the available control variables are unlikely to perfectly capture what we want them to
- Example: tastes for leisure are ultimately unmeasurable
- But there are many things we can take into account:
  - Age
  - Education
  - Race
  - Work experience
  - Marital status
- All of the above may influence labor supply, and are important to control for when trying to estimate (for example) the causal effect of TANF on labor supply

References I

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2017. "Rdrobust: Software for Regression-discontinuity Designs." *The Stata Journal* 17 (2): 372–404.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014. "Robust Data-Driven Inference in the Regression-Discontinuity Design." *The Stata Journal* 14 (4): 909–946.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* 225 (2): 254–277.

References II

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199.

Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." *Working Paper.*