# PUBLIC POLICY 639 NOTES
# Quantitative methods for program evaluation

Xiaoyang Ye
*University of Michigan*

December 25, 2017

*"Daunt us from our true delight"* (W. H. Auden, 1936)

# PubPol 639 Section #1
# Causal inference for public policy

Xiaoyang Ye
*University of Michigan*

September 8, 2017

**OBJECTIVES OF SECTION 1**

1. Introduction to the section
2. Getting started with `Stata`
3. Review of stat using `Stata`

## 1 Introduction to section (welcome!)

- **Purpose:** Our weekly review section is meant to supplement the lectures and readings, with an emphasis on applying the empirical analytical skills necessary in the causal inference for public policy (and `Stata`) and ensuring you can correctly interpret and communicate findings in a professional setting.
- **What we expect of you:** Brian and I share a teaching philosophy of high expectations with high support. We expect you to come to lecture and section having completed the readings and spent some time pondering their implications. We expect you to spend time struggling and muddling through material in an attempt to make sense of it before coming to us with questions; this is something you'll have to do as independent public policy researchers. That said, do not suffer endlessly. We are here to help.
- **What you can expect from me:** You can expect careful preparation (e.g. handouts, section files) and thoughtful instruction in section. You can expect prompt and constructive responses to your emails. In addition, you can expect copious feedback on assignments. Whenever you have questions and comments, please let me know.
- **What you should expect of one another:** I expect we cultivate an environment of mutual respect and support. I hope that you will help one other through the rough patches, encourage one another throughout the course of your work, and be open to learning from one another. I aim to teach the section "at the mean," so if you are a `Stata` pro and find yourself with free time in the lab, please help someone who is struggling.
- **What we will do in section:** Section meets weekly on Fridays 2:30-4pm (notice two reschedules in November). By 2:35pm (not the Michigan time), I expect you are in your seats, have loaded Stata, and navigated to the appropriate working directory. Please be respectful of your peers and minimize tardiness.

- **You are very welcome to my office hours:** My office hours are Mondays 11am-2pm (please register at *Google Calendar*), but my schedule is pretty flexible and I can meet with you at other times as well (my office is 5204 Weill Hall). **Virtual office hours** will be held on evenings before assignment dues and quizzes (announcements will be sent in advance).

- **Email me** whenever you have questions at *yxy@umich.edu* with the subject starting with "639 - ".

- **Additional note (what I have learned from Professor John DiNardo):** You may keep a self-created "notebook" for the class. The idea is that ultimately this notebook might prove useful to you when you are called to teach a class such as this, or as a reference for your own research. This notebook should contain, *inter alia*:
  - A Table of Contents and a bibliography
  - Summaries of lecture notes
  - All problem sets and problem set answers (including do and log files).
  - Very short summaries of assigned readings with critical commentary
  - Documented `Stata` (or other software) code and output when appropriate.
  - etc.

## 2 Non-Bayesian empirical-minded `Stata` runner

This class introduces students to the use and interpretation of multiple regression analysis and program evaluation (aka "applied econometrics"). In the review section, we will cover the following topics:

1. Review of `Stata` and stat
2. Randomization
3. Bivariate regression
4. Multiple regression and OVB
5. Non-linearity
6. Bivariate dependent variables
7. DID/Panel Data/CITS/Event study
8. RCT
9. RD

(1) Below is a list of books are useful as theoretical (and empirical) reference.

  - Wooldridge (2010), *Econometric Analysis of Cross Section and Panel Data*

- *Econometrics* by Bruce Hansen, you can download the 2017 version here.
- Hayashi (2000), *Econometrics*

Here is also a list of more empirical books.

- Angrist and Pischke (2008), *Most Harmless Econometrics*
- Murnane and Willett (2010), *Methods Matter*
- Cameron and Trivedi (2009), *Microeconometrics Using Stata*

(2) More books on causal inference.

- Manski (2009), *Identification for Prediction and Decision*
- Imbens and Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*

(3) This class is not about *ex ante* evaluation ("structural").

- DiNardo and Lee (2011), *Program evaluation and research designs*
- Heckman and Vytlacil (2007a), *Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation*
- Heckman and Vytlacil (2007b), *Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments*
- Abbring and Heckman (2007), *Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation*

(4) This class is not about Bayesian econometrics. The key difference between Bayesians and non-Bayesians (frequenists) is about what "probability" is.

- Poirier (1995), *Intermediate Statistics and Econometrics*
- Greenberg (2012), *Introduction to Bayesian Econometrics*.

(5) This class is not about big data and machine learning.

- Athey and Imbens (2017b) provide a nice up-to-date review

# 3 Lecture notes #1

## 3.1 Problems with simple comparison

- *Overstate* the return to education?
  - Omitting other factors (e.g., ability, family background)
  - General equilibrium effects (may understate)
- *Understate* the return to education?

    - Other outcomes (e.g., health, marriage, job occupation)
    - Age/experience effects (Mincerian equation)
- Other issues
    - Sampling (random sample? sample size? representative sample?)
    - Measurement of outcome (e.g., total income vs. wages; timing) and education (e.g. levels of education vs. schooling years; school quality)
    - Costs?

### 3.2 Why do schooling levels differ between peoples?

- Observables
    - Race, ethnicity, gender, family/school/community factors, etc.
    - For causal inference, they often serve as the "control variables"
- Unobservables and endogenous
    - Ability, preference, risk aversion, etc.
    - They are the threats of causal inference
- Unobservables and exogenous
    - Randomized controlled trial is the gold standard
    - They are the identification instruments in quasi-experimental designs
    - E.g., labor market condition, distance from home to school/college, admission and financial aid cutoff, policy and program (what we are supposed to evaluate in PubPol 639 and in your future career)

## 4 Lecture notes #2

### 4.1 Four key questions in causal inference from MHE

1. What is the **causal** relationship of interest?
    - Cause and consequence
2. What is the ideal experiment?
    - Random assignment of the treatment
3. What is the identification strategy?
    - Use observational data to approximate the ideal experiment
4. What is the mode of statistical inference?
    - Use sample data to infer the population
    - Regressions are widely used

### 4.2 Impacts of pollutions

- Chay and Greenstone (2005), "Does air quality matter? Evidence from the housing market"
  - The Clean Air Act legislation imposes strict regulations on polluters in "nonattainment" counties, which are defined by concentrations of TSPs (total suspended particulates) that exceed a federally set ceiling.
  - TSPs nonattainment status is associated with large reductions in TSPs pollution and increases in county-level housing prices. [instrumental variable]
  - The elasticity of housing values with respect to particulates concentrations ranges from **-0.20 to -0.35**.
  - These estimates of the average marginal willingness to pay for clean air are robust to quasi-experimental regression discontinuity and matching specification tests. Further, they are far less sensitive to model specification than cross-sectional and fixed-effects estimates, which occasionally have the "perverse" sign. [RD, matching, fixed effects]
  - The marginal benefit of reductions of TSPs is lower in communities with relatively high pollution levels, which is consistent with preference-based sorting.
  - Overall, the improvements in air quality induced by the mid-1970s TSPs nonattainment designation are associated with a $45 billion aggregate increase in housing values in nonattainment counties between 1970 and 1980.
- Chay and Greenstone (2003), "The Impact of Air Pollution on Infant Mortality: Evidence from Geographic Variation in Pollution Shocks Induced by a Recession"
  - The 1981–1982 recession induced substantial variation across sites in air pollution reductions. This is used to estimate the impact of total suspended particulates (TSPs) on infant mortality. [Natural experiment]
  - A 1-percent reduction in TSPs results in **a 0.35 percent decline** in the infant mortality rate at the county level, implying that 2500 fewer infants died from 1980–1982 than would have in the absence of the TSPs reductions.
  - Most of these effects are driven by fewer deaths occurring within one month of birth, suggesting that fetal exposure is a potential pathophysiologic mechanism.
  - The analysis also reveals nonlinear effects of TSPs pollution and greater sensitivity of black infant mortality at the county level. Importantly, the estimates are stable across a variety of specifications.
- Greenstone and Gallagher (2008), "Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program"
  - This paper uses the housing market to develop estimates of the local welfare impacts of Superfund-sponsored cleanups of hazardous waste sites.
  - If consumers value the cleanups, then the hedonic model predicts that they will lead to increases in local housing prices and new home construction, as well as the migration of individuals that place a high value on environmental quality to the areas near the improved sites.

- We compare housing market outcomes in the areas surrounding the first 400 hazardous waste sites chosen for Superfund cleanups to the areas surrounding the 290 sites that narrowly missed qualifying for these cleanups. [DID]
  - We find that Superfund cleanups are associated with **economically small and statistically insignificant** changes in residential property values, property rental rates, housing supply, total population, and types of individuals living near the sites.
  - These findings are robust to a series of specification checks, including the application of a regression discontinuity design based on knowledge of the selection rule. [RD]
  - Overall, the preferred estimates suggest that the local benefits of Superfund cleanups are small and appear to be substantially lower than the $43 million mean cost of Superfund cleanups.

- Currie and Walker (2011), "Traffic congestion and infant health: Evidence from E-ZPass"
- Sullivan (2017), "The True Cost of Air Pollution: Evidence from the Housing Market"
- Isen, Rossin-Slater, and Walker (2017), "The long-term consequences of the Clean Air Act of 1970"

# 5 `Stata` tip #1: Free `Stata` at UMich

## 5.1 AFS space

- On UMich SITES or VirtualSites machines, your AFS workspace should be M:\ drive.
- On your own computer, you can access your AFS space at: http://mfile.umich.edu
- Creating lab folders
  - On the lab computer, double-click on the **My AFS Space** icon
  - Double-click on Private
  - Create a new folder **PubPol639** and make a directory of sub-folders

## 5.2 Where is the `Stata`?

- You may not have to buy and install Stata on your personal computer
  - You can use PuTTY to connect to UMich's server, see steps here
  - You can transfer files to your AFS Space using WinSCP (windows) or FileZilla (Mac), see steps here
  - I have worked with it in more than 20 airports around the world...

# 6 `Stata` tip #2: Send an email using `Stata`

- `mail`: Module to send emails from Mac/Linux/Unix
- `emailme`: Module to send emails from Windows systems
- `email`: Module to send emails via Python

***Stata codes***

```
* Send an email using Mac's underlying sendmail program to send you emails

// Syntax
// mail From: x@X; To: y@Y; Subject #; # [using filename] [, option]

// Example
mail From: yxy@umich.edu; To: yxy@umich.edu; Subject: 799 Lab 1; ///
Lab 1 Do File has finished.

* It is useful when you run long do files remotely
```

# PubPol 639 Section #2
# Selection bias is everywhere! RCT eliminates it?

Xiaoyang Ye
*University of Michigan*

September 15, 2017

**OBJECTIVES OF SECTION 2**

1. A little bit more statistics
2. Observables and unobservables in the potential outcome framework
3. RCT
4. Lecture notes: Bertrand and Duflo (2017), "Field experiments on discrimination"
5. `Stata` tip: `missings`

## 1 Those statistics are confusing us

### 1.1 Confidence interval for means

- The sample mean is *a point estimate* of the population mean, but it does not reveal the uncertainty associated with the estimate - how far the sample mean may be from the population mean, or what the range is which the population mean is estimated to lie
- Interpretations
  - **in terms of repeated samples**: Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that contain the true population mean would tend toward 95% (or other proportions)
  - **in terms of a single sample**: There is a 95% probability that the calculated confidence interval from some future experiment contains the true value of the population mean
- Misunderstandings
  - A 95% confidence interval does not mean that ~~for a given realized interval there is a 95% probability/level of confidence that the population mean lies within the interval~~
    - Once an experiment is done and an interval calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability - probability refers the future
  - A 95% confidence interval does not mean that ~~95% of the sample data lie within the interval~~

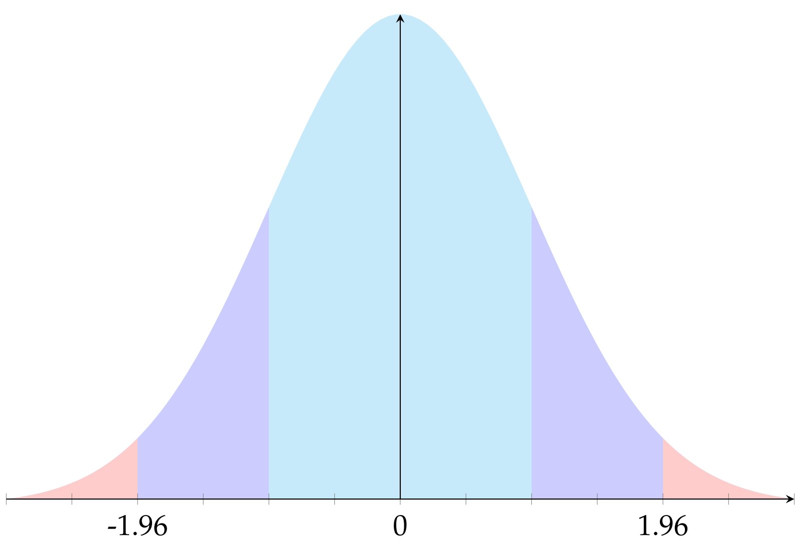
- A particular confidence interval of 95% calculated from one sample does not mean that ~~there is a 95% probability of a sample parameter from a repeat of the experiment falling within this interval~~

- Relationship with one-sample t test ($H_o : \mu = 0$)
  - The estimated confidence interval contains 0 $\Rightarrow$ cannot reject the null hypothesis

## 1.2 (In)equal variances in two-sample t test ($H_0 : \mu_1 = \mu_2$)

- The two-sample t test tests whether there is statistically significant difference between the **population means** in two unrelated groups
- Key assumption: equal variances in the population of the two groups
  - If your variances are unequal, this can affect the Type I error rate
  - Levene's Test of Equality of Variances (`sdtest`)
  - Equal variance t test - `ttest variable, by(group)`
  - Unequal variance t test (Welch's t test) - `ttest variable, by(group) unequal`

## 1.3 Two-tailed t test vs. one-tailed test



Normal distribution

- Three alternative hypotheses in t test (`ttest` provides results for all)
  - t statistic is symmetric and bell-shaped, like normal distribution
  - Two-tailed
    * If you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction
    * .025 is in each tail of the distribution of your test statistic

    * When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions
  - One-tailed
    * If you are using a significance level of .05, a one-tailed test allots all of your alpha to testing the statistical significance in the one direction of interest
    * .05 is in one tail of the distribution of your test statistic
    * When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction

- When to use one-tailed test
  - One-tailed test provides more power to detect an effect
  - You may be tempted to use a one-tailed test whenever you have a hypothesis about the direction of an effect
  - You fail to test for the possibility of the other direction
  - If you consider the consequences of missing an effect in the untested direction and conclude that they are negligible and in no way irresponsible or unethical, then you can proceed with a one-tailed test

- When not to use one-tailed test
  - Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate
  - Choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was.

- Two-tailed p-value is twice the one-tailed p-value
  - t statistic is the same

Reference: UCLA IDRE

### 1.4 Significance level

- Reject the null: P-value$\leqq a$
- $a$ (or, alpha) is the significance level. We primarily use 0.05
- A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference
  - In randomization, we may still have about 5% cases/variables where the treatment and the control are unbalanced

# 2 Causal inference is the core of quantitative analysis in program evaluation

- Four key questions in Angrist and Pischke (2014)
  1. What is the causal relationship of interest?
     - Cause vs. consequence
  2. What is the ideal experiment?
     - What is the treatment?
  3. What is the identification strategy?
     - "Causal research designs"
  4. What is the mode of statistical inference?
     - Statistical techniques: e.g., t test, regression

## 2.1 Potential outcomes

- Causality is defined by potential outcomes, not by realized (observed) outcomes
- *Counterfactual*

$$Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0) \tag{1}$$

  - $T_i$: a dummy variable indicating whether individual $i$ receives treatment ($T_i = 1$) or not ($T_i = 0$);
  - $Y_i(1)$: the outcome of individual $i$ if she receives treatment;
  - $Y_i(0)$: the outcome of individual $i$ if she does not receive treatment.
- A valid causality question must involve well-defined causes (treatments, manipulations), and the counterfactuals should be unambiguously defined.
- For any vaguely-defined problem, It is always good to consider whether there is a treatment/manipulation/thought experiment that could produce meaningful counterfactuals.

## 2.2 Fundamental problem of causal inference by Holland (1986)

- *Individual treatment effect*

$$\tau_i = Y_i(1) - Y_i(0) \tag{2}$$

- **The problem**: We can only observe one of the two potential outcomes
  - *Missing data problem*: Any statistical method dealing with treatment effects necessarily imputes the counterfactual part of the data.
  - Extensions: **Fisher's sharp null hypothesis** and **Permutation Test**

**Average treatment effects**

- The **average treatment effect (ATE)** is defined as

$$\tau_{\texttt{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)]. \tag{3}$$

  - Extensions: Quantile treatment effects

- The **average treatment effect on the treated (ATT)** is defined as

$$\tau_{\texttt{ATT}} = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1]. \tag{4}$$

- The **average treatment effect on the untreated (ATUT)** is defined as

$$\tau_{\texttt{ATU}} = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 0]. \tag{5}$$

### 2.3 Treatment Assignment Mechanism

- Estimation of causal effects starts with the assignment mechanism.
- In general, the **assignment mechanism** could be denoted by the following mapping:

$$(i, \mathbf{X}_i, \mathbf{Y}_i, u_i) \ \rightarrow T_i, \tag{6}$$

  where $\mathbf{X}_i$ includes individual characteristics, $\mathbf{Y}_i = (Y_i(0), Y_i(1))$ are potential outcomes and $u_i$ is come orthogonal error term which allows for randomization.

- Since $T_i$ is a binary random variable, the assignment mechanism could be represented by the **propensity score function**:

$$p(i, \mathbf{X}_i, \mathbf{Y}_i) = \mathbf{P}\left[T_i = 1 \mid i, \mathbf{X}_i, \mathbf{Y}_i\right] \tag{7}$$

- Here are come important constraints about the assignment mechanism:
  - **Individualistic Assignment**:

$$p(i, \mathbf{X}_i, \mathbf{Y}_i) = p(\mathbf{X}_i, \mathbf{Y}_i) = \mathbf{P}\left[T_i = 1 \mid \mathbf{X}_i, \mathbf{Y}_i\right] \tag{8}$$

  - **Probabilistic Assignment** (overlapping):

$$0 < p(i, \mathbf{X}_i, \mathbf{Y}_i) < 1 \tag{9}$$

  - **Unconfoundedness**:

$$p(i, \mathbf{X}_i, \mathbf{Y}_i) = p(\mathbf{X}_i) = \mathbf{P}\left[T_i = 1 \mid \mathbf{X}_i\right]. \tag{10}$$

    * Conditional on covariates (both observed and unobserved), the assignment mechanism is similar to a randomized experiment

* **Omitted variable bias**: In many cases the covariates $\mathbf{X}_i$ are not rich enough, meaning that the assignment mechanism could still be related to the potential outcomes, so that the estimated treatment effects could be biased

### 2.4 Selection Bias

- **Prima facie causal effect** by Holland (1986):

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

  - Without further assumptions, however, the selection bias could be arbitrary large/small which makes the above quantity uninformative about the average treatment effect
  - "We don't know what we don't know"

- Rubin causal model is a special case of **switching regression / general Roy model**:

$$\begin{aligned}\text{Potential Outcomes:}\quad & Y_i(0)=\mathbf{X}_i\beta(0)+u_i(0)\\ & Y_i(1)=\mathbf{X}_i\beta(1)+u_i(1)\\ \text{Selection/Assignment Mechanism:}\quad & \mathbf{1}_{\{T_i=1\}}=F(\mathbf{X}_i\gamma)+\epsilon_i\end{aligned}\tag{11}$$

- The identification is:

$$\mathbf{X}_i\perp(u_i(0),u_i(1),\epsilon_i).$$

- The unconfoundedness assumption is simply:

$$\text{Unconfounded Assignment:}\qquad (u_i(0),u_i(1))\perp\epsilon_i,\tag{12}$$

  which has the intuitive meaning that after controlling the covariates, the errors in determining the potential outcomes are orthogonal to the errors in the selection equation/assignment mechanism

- If the above does not hold, then the potential outcomes will be correlated with the selection equation/assignment mechanism even after partialling out the covariates, hence selection bias will affect estimating the average treatment effect

### 2.5 Causal Research Designs

- This section summarizes the excellent chapter of DiNardo and Lee (2011).

1. By knowledge of **Assignment Mechanism**
   - Random assignment (RCT)
     - Gold standard
   - Regression discontinuity (RD)

    - A close "cousin" of RCT

2. By **Self-Selection**
    - Difference-in-differences (DID)
        - Influence of "other factors" fixed
    - Selection on unobservables and instrumental variables (IV)
        - Conditional on covariates, instrument "as good as randomly assigned"
        - Instrument is uncorrelated with outcomes
        - Another structural approach: Heckman selection model
    - Selection on observables and matching (Matching)
        - Conditional on covariates, treatment "as good as randomly assigned"
        - Including Blinder/Oaxaca decomposition, matching, propensity score matching, re-weighting

- If you are interested in knowing how these research designs have been used to study the core question in the economics of education: **the causal effects of schooling on wages**, please read Chapter 6 in Angrist and Pischke (2014)

# 3 The RCT "revolution"

## 3.1 A brief history

- Randomized Controlled Trial
- Banerjee and Duflo (2017)
    - Fisher (1925) proposed RCT to answer causal questions
    - The first large-scale social experiment was the New Jersey Income Maintenance Experiment, which was initiated in 1968 and tested the impact of income transfers and tax rates on labor supply
    - In mid 1990s, RCT extended to developing countries
    - The last 15 years have seen an explosion in the number, scope, quality, and creativity of field experiments.
- How RCT impacts the way we think about the world: the value of better human capital, reforming education, the design of redistributive program, the design of incentives of public officials, access of financial products, the demand for insurance and other prophylactic products, preferences and preference change, the role of community, and getting people to vote.
- "Handbook of Economic Field Experiments" is strongly recommended for those who are interested in RCTs.

### 3.2 Randomization of the assignment

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

- To eliminate the selection bias: $\mathbf{E}[Y_i(0)|T_i = 1] - \mathbf{E}[Y_i(0)|T_i = 0] = 0$
- Treatment-control balance tests (t test for mean comparison, regression)
  - Potential outcomes are unobservable
  - (Some) covariates are observable
    - Expect that we fail to reject the null
    - (Note: that if we test many baseline characteristics we will reject 5% of the time by chance)
- If the randomization works, any difference in outcome can be attributed to the random assignment

### 3.3 Assess the economic significance of the treatment effect

- **Effect size**: standard deviation (of the population) change
- **Percent change** of the benchmark/control group mean
- **Back of envelop calculation**: total # changes in the real world by the treatment; admittedly uncertain and require a lot of assumptions

### 3.4 RCT is perfect, but the world is not

- Threats that could make the result you are finding not be the causal estimate of interest, examples include
  - Differential attrition
  - Switching between treatment and control
  - Variety of treatment status
  - Hawthorne effects (responses to being treated)
  - John Henry effects (responses to being not treated - the tendency for people based in a control group to perceive themselves at a disadvantage to the experimental group and work harder in order to overcome the perceived deficiency)

### 3.5 More econometrics on RCT

- Athey and Imbens (2017a)
  - Internal validity vs. external validity
  - Complete vs. stratified vs. paired vs. cluster randomization

- Noncompliance (Intent-to-treat, treatment effect on the treated, local average treatment effect, bounds)
- Heterogeneous treatment effect

# 4 Lecture notes #3

## 4.1 Bertrand and Duflo (2017), "Field experiments on discrimination"

- Discrimination
  - "members of a minority group (women, Blacks, Muslims, immigrants, etc.) are treated differentially (less favorably) than members of a majority group with otherwise identical characteristics in similar circumstances. "
  - taste-based discrimination
    - "for the context of the labor market, some employers have a distaste for hiring members of the minority group. They may indulge this distaste by refusing to hire, say, Blacks or, if they do hire them, paying them less than other employees for the same level of productivity.
    - If the fraction of discriminating employers in the economy is sufficiently large, a wage differential will emerge in equilibrium between otherwise identically productive minority and majority employees and this wage differential will be a reflection of the distaste parameter of the marginal employer for minority workers"
  - statistical discrimination
    - "the differential treatment of members of the minority group is due to imperfect information, and discrimination is the result of a signal extraction problem.
    - As a profit-maximizing prospective employer tries to infer the characteristics of a person that are relevant to the market transaction they are considering to complete with that person, they use all the information available to them.
    - When the person-specific information is limited, group specific membership may provide additional valuable information about expected productivity"
  - "While taste-based discrimination is clearly inefficient (simply consider how it constrains the allocation of talent), statistical discrimination is theoretically efficient and hence more easily defendable in ethical terms under the utilitarian argument"
  - Micro-foundations from psychologists
    - including personality development, socialization, social cognition, evolutionary psychology and neuroscience
    - favoritism for in-group members and negativity towards outgroup members
    - the existence of unconscious, unintentional forms of bias
- Experimental methods
  - Audit studies send out individuals who are matched in all observable characteristics except for the one in question (race, criminal record, etc.) to apply for jobs or make purchases, then researchers analyze the responses.

  - Correspondence studies – which represent by far the largest share of field experiments on discrimination – do the same but control for more variables by creating fictitious applicants (often for jobs or apartments) who correspond via mail.

- Costs and benefits of diversity
- Programs to reduce discrimination

### 4.2 `Stata` tip #3: Manage missing values using `missings`

- A more common used command is `missing`, for example, `drop if missing(var1)`
- `missings`: A set of utility commands of managing missing values (., .a-.z in numeric variables, "" in string variables)
- Stata allows us to code different types of numeric missing values. It has 27 numeric missing categories. ".a" to ".z" and "."

***Stata codes***

```
* (1) Install "missings" package
ssc install missings

* (2) Example data set from Stata
webuse nlswork, clear

* (3) "missngs report" issues a report on the number of missing values
missings report
// report for all variables
missings report year
// report for year variable
missings report, min(1000)
// report for variables with at least 1000 missing cases

* (4) "missings list" lists observations with missing values
missings list, min(5)
// list observations with at least 5 missing variables

* (5) "missings table" tabulates observations by the number of missing values
missings table
by race: missings table
// by race groups

* (6) "missings tag" generates a variable containing # of missing values
missings tag, generate(nmissing)

* (7) "missings dropvars" drops any variables which are missing on all values
```

```
gen frog = .
gen toad = .a
gen newt = ""

missings dropvars frog toad newt, force sysmiss
missings dropvars toad, force sysmiss
// sysmiss applies to . missing only (not .a-.z in numeric and "" in string)

* (8) "missings dropobs" drops any observations which are missing on all values
set obs 30000
missings dropobs, drop
```

# PubPol 639 Section #3
# "Instead of complaining about it , I'm just gonna go in everyday and give it my all."

Xiaoyang Ye
*University of Michigan*

September 22, 2017

**OBJECTIVES OF SECTION 3**

1. RCT again
2. Lecture notes: Impacts of class size on test scores
3. `Stata` tip: `outreg2`

## 1 RCT again

### 1.1 Three steps in RCT evaluation

- *Prima facie causal effect* by Holland(1986)

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

1. **Balance check** - no selection bias: $\mathbf{E}[Y_i(0)|T_i = 1] - \mathbf{E}[Y_i(0)|T_i = 0] = 0$
2. **Raw effect estimation** - compare treatment/control: $\mathbf{E}[Y_i(1)|T_i = 1] - \mathbf{E}[Y_i(0)|T_i = 0]$
3. Infer the causal effect given evidence of no selection bias: $\tau_{\text{ATT}} = \mathbf{E}[Y_i(1)|T_i = 1] - \mathbf{E}[Y_i(0)|T_i = 1]$

### 1.2 Internal and external validity

- **Internal validity**
  - "Causality"
  - No selection bias of the specific sample in study
    - Attrition? Compliance? Hawthorne/John Henry effect?
  - The primary concern in causal inference

- **External validity**
  - "Generalizability"
  - Whether the balance check holds for other samples or settings
    * $\mathbf{E}[Y_i(0)|T_i = 1] - \mathbf{E}[Y_i(0)|T_i = 0] =?$
  - Whether the treatment effect is the same for other samples or settings
    * $\mathbf{E}[Y_i(1)|T_i = 1] - \mathbf{E}[Y_i(0)|T_i =? = 0] =?$

### 1.3 Estimation in RCT evaluation

- **Two sample t test**?
  - Only works with two groups

$$Y = \beta_0 + \beta_1 * X_{(one\ group\ dummy)} + \varepsilon$$

- **ANOVA** (analysis of variance)?
  - The t-test is a two-sample case of ANOVA
  - Square of the t-test statistic equals the F statistic in ANOVA

$$Y = \beta_0 + \beta_1 * X_{(one\ or\ more\ group\ dummies)} + \varepsilon$$
$$Y = \beta_0 + \beta_1 * X_{(group\ 1\ dummy)} + \beta_1 * X_{(group\ 2\ dummy)} + \varepsilon$$

- **Regression**?
  - ANOVA is a specific regression model where there are only dummy variables in $X$

$$Y = \beta_0 + \beta_1 * X_{(dummies\ or\ contininous)} + \varepsilon$$

## 2 Lecture notes #4

### 2.1 Impacts of class size on test scores

- Card and Krueger (1992)
- Angrist and Lavy (1999)
- Hoxby (2000)

- Krueger and Whitmore (2001)
- Chetty et al. (2010)
- Dynarski, Hyman, and Schanzenbach (2013)
- Fredriksson, Öckert, and Oosterbeek (2013)
- Battistin, Angrist, and Vuri (2016)
- Bettinger et al. (2017)
- Angrist et al. (2017)

### 2.2 `Stata` tip #4: `outreg2`

- `Outreg2` provides a fast and easy way to produce an illustrative table of regression outputs.
- It can produce various types of output files: dta, word, excel, tex, text
- A full description is in `Stata`: "`help outreg2`"
- A very good example from Oscar Torres-Reyna at Princeton: https://www.princeton.edu/ otorres/Outreg2.pdf

# PubPol 639 Section #4
# The world is linear.

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**OBJECTIVES OF SECTION 4**

1. Introduction to OLS
2. Lecture notes: Effect size
3. `Stata` tip: recode, codebook tab

## 1 OLS regression

OLS is the dominant method used in empirical quantitative analysis. The OLS estimator is BLUE (best linear unbiased estimator) under certain assumptions.

### 1.1 Linear regression with a single regressor

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

- $Y_i$ is the *dependent variable*, *regressand* or *left-hand variable*
- $X_i$ is the *independent variable*, *regressor*, or *right-hand variable*
- $\beta_0$ is the intercept of the population regression line ($\beta_0 + \beta_1 X$)
- $\beta_1$ is the slope
- $\mu_i$ is the error term

#### OLS estimator

The population regression line is **unknown**. We use data from a sample to estimate the parameters.
Ordinary least squares (OLS) estimators ($\hat{\beta_0}$, $\hat{\beta_1}$) minimize the sum of squared errors:

$$min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$$

- The estimators are random variables with a probability distribution (sampling distribution)
- The sample regression line: $\hat{\beta_0} + \hat{\beta_1}X$
- The predicted outcome: $\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1}X_i$
- The residual: $\hat{\mu_i} = Y_i - \hat{Y_i}$

**Interpretation**

***On average, each one unit increase in $x$ is associated with $\hat{\beta_1}$ unit change in $Y$ (p=**).***

**Assumptions**

1. The conditional distribution of $\mu_i$ given $X_i$ has a mean of zero: $\mathbf{E}(\mu_i \mid X_i) = 0$
   - The other factors in $\mu_i$ are not correlated with $X_i$
   - This implies $corr(X_i, \mu_i) = 0$ (this implication does not go the other way)
   - A randomized controlled experiment ensures that $X$ is distributed independently of all other factors
   - In observational data, $X$ is randomly (by natural experiment) or **as if** randomly (e.g., instrumental variables) assigned is the key identification condition for causal inference
2. $(X_i, Y_i)$ for $i = 1, ..., n$ are independent and identically distributed (**i.i.d**) draws from their joint distribution
3. Large outliers are unlikely: $X_i$ and $Y_i$ have nonzero finite fourth moments

If these assumptions hold, the OLS estimators are unbiased, consistent, and normally distributed when the sample is large.

**Model fit**

$$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$ESS = \sum_{i=1}^{n}(\hat{Y_i} - \bar{Y})^2$$
$$SSR = \sum_{i=1}^{n}(Y_i - \hat{Y_i})^2$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

$$Standard\ error\ of\ regression = \sqrt{\frac{SSR}{n-2}}$$

- $R^2$ does not imply the regression is either “good” or “bad”, it only indicates that other important factors also impact the outcome.

**Hypothesis testing**

$$t = \hat{\beta}_1 / SE(\hat{\beta}_1)$$

## 1.2 Linear regression with multiple regressors

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \mu_i$$

- $\beta_1$ is the effect on $Y$ of a unit change in $X1$, holding $X2$ constant (or, controlling for $X2$)
- $\beta_0$ is the expected value of $Y$ when all $X$s equal 0

**Adjusted $R^2$**

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

### 1.2.1 Multicollinearity and the dummy variable trap

- One regressor is an exact linear function of the other regressors
- The key problem: **you are no longer able to "control for all other factors”**

***The dummy variable trap***

```
Binary variables for urbanicity: urban, suburban, rural

If you include all these three variables in the regression along with a
constant term, the perfect multicollinearity problem arises

urban + suburban + rural = 1 for every observation
```

## 2 Lecture notes #5: Effect size

## 3 `Stata` tip #5: recode, codebook

# PubPol 639 Section #5
# The world is conditionally linear.


Xiaoyang Ye
*University of Michigan*


September 5, 2018

**OBJECTIVES OF SECTION 5**

1. Conditional expectation function
2. More on multiple regression
3. Lecture notes: What are we weighting for?
4. `Stata` tip: Working with categorical variables

## 1 Conditional expectation function, based on Angrist and Pischke (2014)

Conditional expectation of outcome $Y_i$ given $X_i$ equals the particular value $x$:

$$\mathbf{E}[Y_i|X_i = x]$$

$\mathbf{E}[Y_i|X_i = x]$ is one point in the range of the *Conditional Expectation Function, CEF*: $\mathbf{E}[Y_i|X_i]$, which includes all the different averages of $Y_i$ as the conditioning variable $X_i$ changes.

### 1.1 Regression and the CEF

The CEF of $Y_i$ is a linear function of the conditioning variable $X_i$:

$$\mathbf{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

Given the linearity assumption, the coefficient on $X_i$ in the CEF will be equal to the coefficient in the regression:

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

$$\beta_1 = \mathbf{E}[Y_i|X_i] - \mathbf{E}[Y_i|X_i - 1] = \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

If the CEF is linear, regression finds it; if not linear, regression finds a good approximation to it.

### 1.2 Zero conditional mean assumption

$\mathbf{E}(\mu_i|X_i) = 0$ is the key assumption to obtain unbiased OLS estimator. When we omit a variable in the regression that is correlated with both $Y_i$ and $X_i$, this assumption violates as $\mathbf{E}(\mu_i|X_i) \neq 0$.

## 2 More on multiple regression

### 2.1 Why we need multiple regression?

1. Omitted variables bias for causal inference - more likely the zero conditional mean assumption holds
2. Reduce standard errors (increase statistical power) - we can explain more of the variation of outcome, thus better predictions
3. Non-linear relationship (e.g., quadratic) - more general functional forms
4. Interested in the relationship holing all other factors constant - *this is how we interpret the beta coefficients*
   - "controlling for"
   - "statistically accounting for"
   - "all else equal"
   - "*ceteris peribus*"

### 2.2 Interpreting $R^2$

The goal of regression is not to maximize $R^2$ or $\bar{R}^2$ but to estimate the causal effect. Although a low $R^2$ indicates that we have not accounted for many other important factors that affect the outcome, if these factors in the error term $\mu_i$ are not correlated with the key variable of interest $X_i$, we can still estimate the causal effect.

A low $R^2$ imply that the error variance is large, we may have a large standard error for statistical precision.

## 3 Lecture notes #6: What are we weighting for?

## 4 `Stata` tip #6: Working with categorical variables

- Full content at https://www.stata.com/manuals13/u25.pdf
- Three types of variables
  - Continuous (in numbers)
  - Categorical (in numbers of strings)
  - Indicator (a special case of categorical with only two groups)

# PubPol 639 Section #6
# To control or not to control for "other factors?"

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**OBJECTIVES OF SECTION 6**

1. OVB and over-control in multiple regression
2. Lecture notes: Non-parametric regression
3. `Stata` tip: In a list/range

## 1 OVB and over-control in multiple regression

## 2 Lecture notes #7: Non-parametric regression

## 3 `Stata` tip #7: In a list/range

# PubPol 639 Section #7
# What if the world is not linear?

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**OBJECTIVES OF SECTION 7**

1. Quadratic terms in linear regression
2. Lecture notes: Returns to education and human capital theory
3. `Stata` tip: forvalues/foreach

## 1 Quadratic terms in linear regression

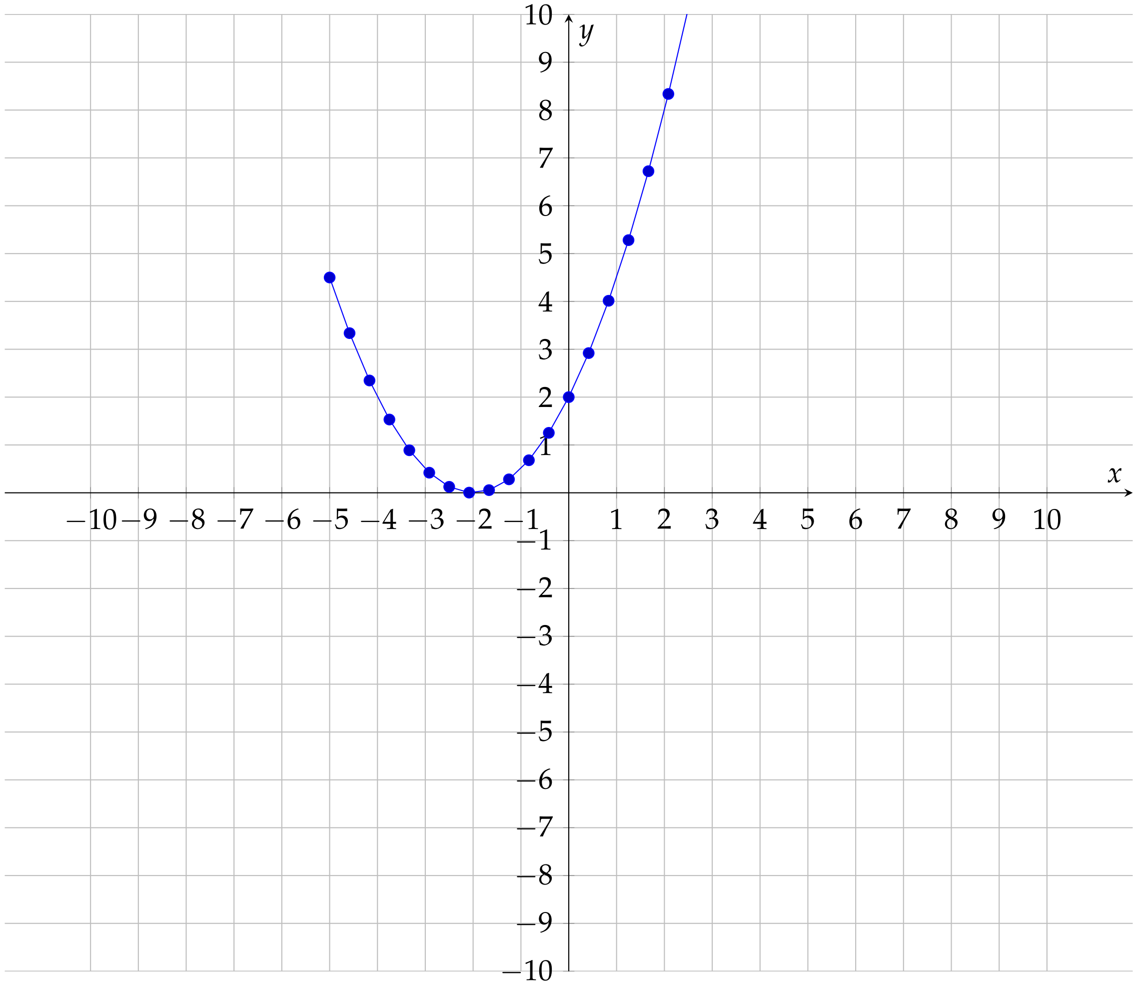
### 1.1 Quadratic equation

In middle school algebra, quadratic equation is:

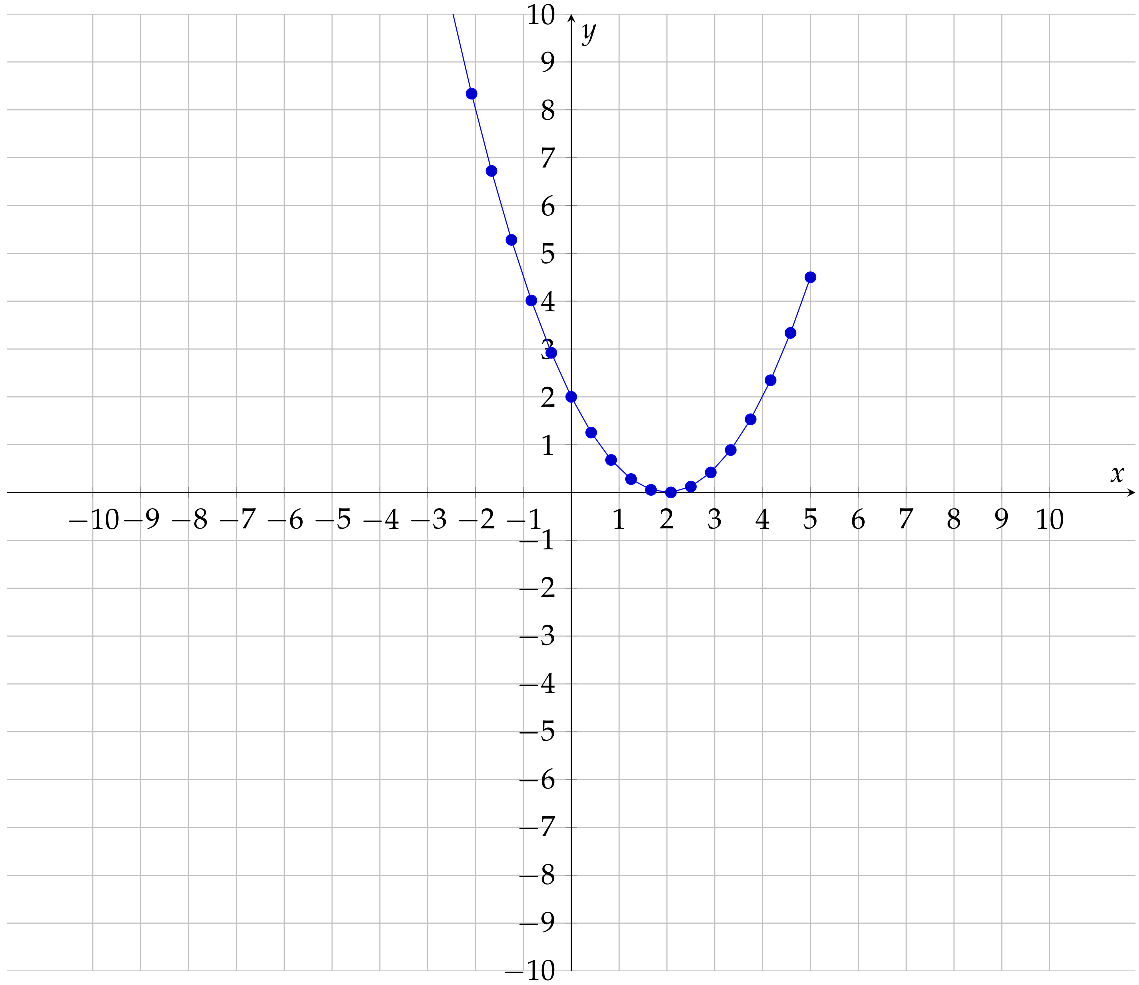
$$Y = a + bX + cX^2$$

- Quadratic equations have between one and three terms, one of which always incorporates $X^2$ (a and b can be 0). c determines whether the associated quadratic line is convex or concave, $X = \frac{-b}{2c}$ determines the position of maximum (concave when c<0) or minimum (convex when c>0) value of $Y$.
- When graphed, quadratic equations produce a U-shaped curve known as a parabola. The line of symmetry is an imaginary line which runs down the center of this parabola and cuts it into two equal halves. This line is commonly referred to as the axis of symmetry. It can be found quite quickly by using a simple algebraic formula.
- The symmetric proposition means that from $-\infty$ to $\infty$ in X, we always have one half of the symmetry line is decreasing and the other half is increasing (the linear line , in contrast, is always increasing or decreasing). When interpreting the changing relationship, always define the range of your data.

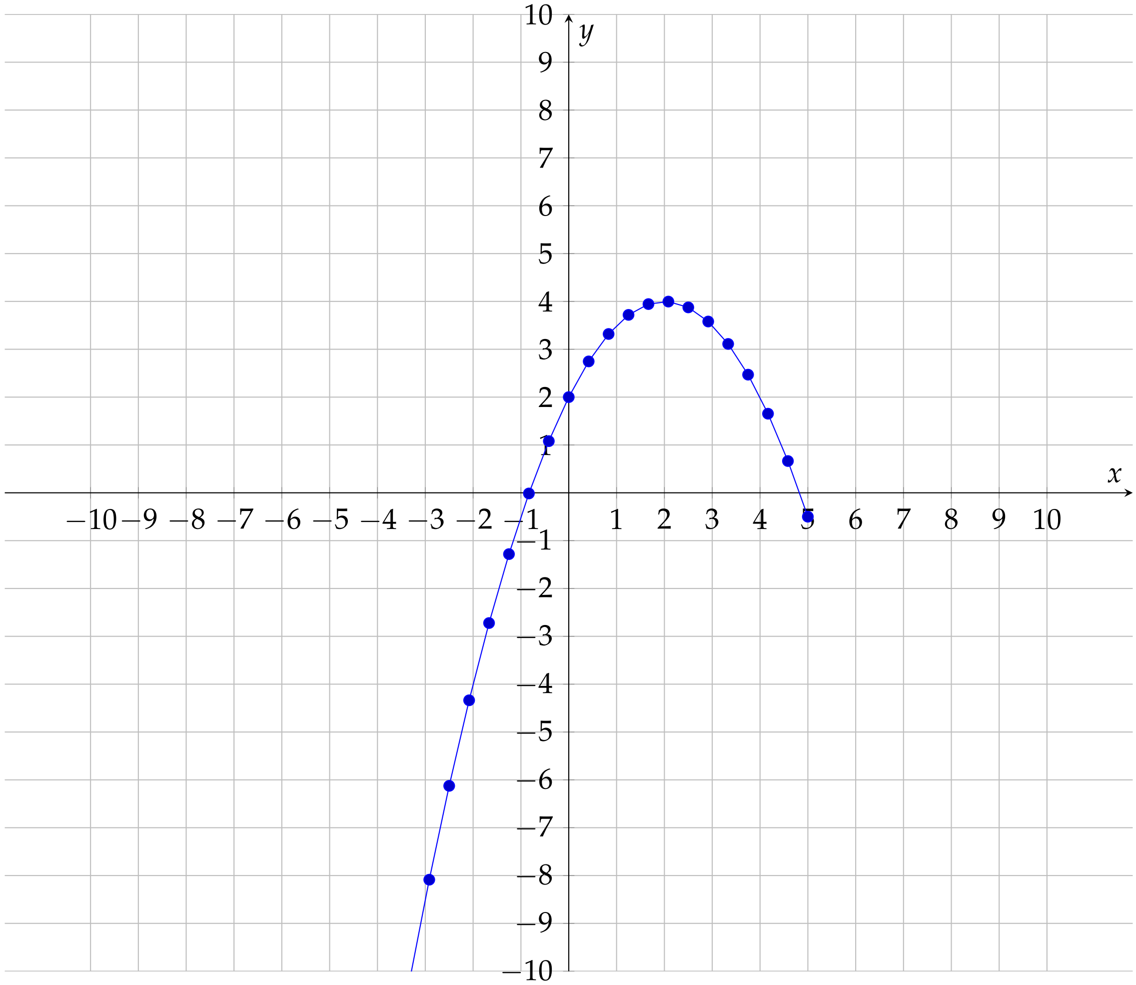
### 1.1.1 Sample figure: $0.5X^2 + 2X + 2$, symmetry line is $X = -2$



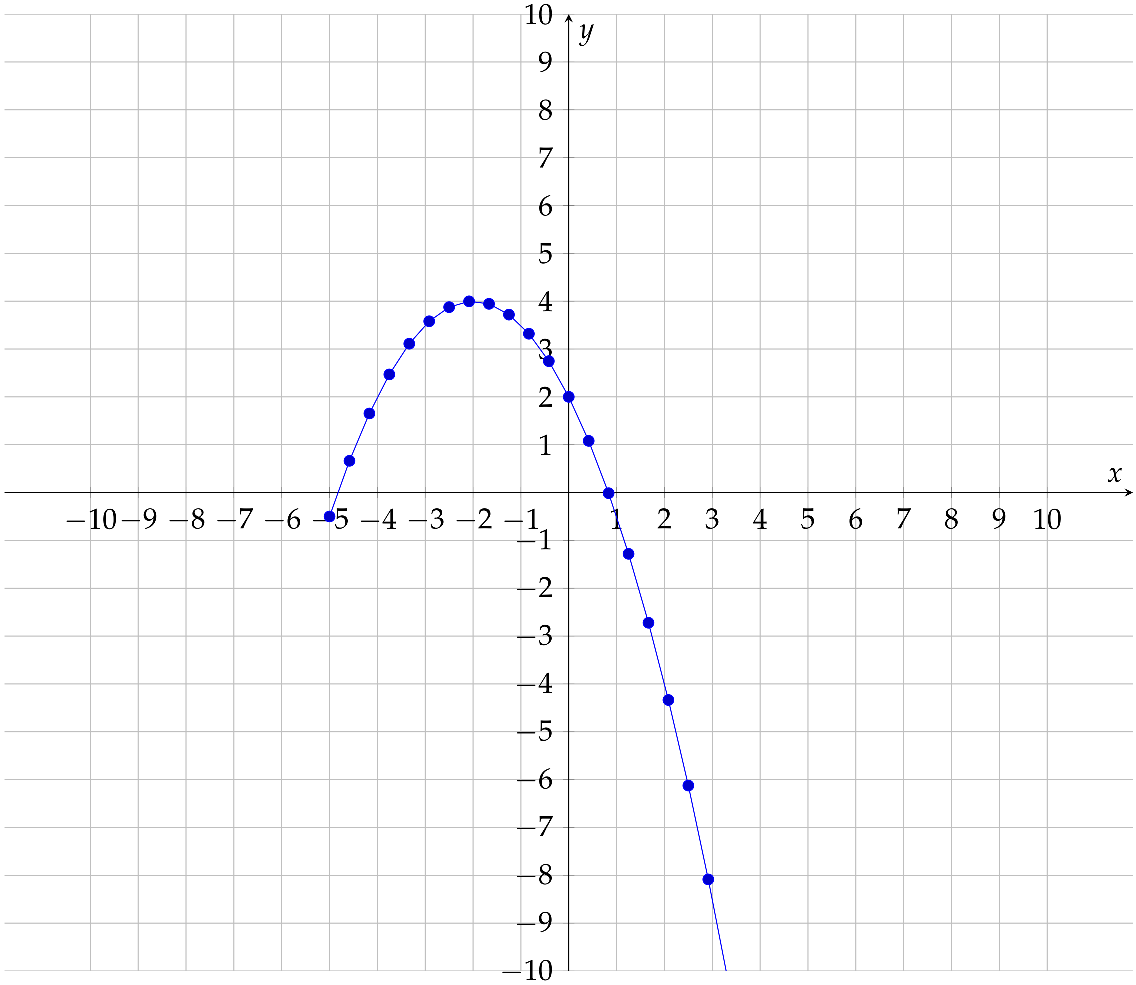
#### 1.1.2 Sample figure: $0.5X^2 - 2X + 2$, symmetry line is $X = 2$



### 1.1.3 Sample figure: -$0.5X^2 + 2X + 2$, symmetry line is $X = 2$



### 1.1.4 Sample figure: $-0.5X^2 - 2X + 2$, symmetry line is $X = -2$




## 1.2 Quadratic term in linear regression

### 1.2.1 Mincer Earnings Equation

$$log\ Y_i = \beta_0 + \beta_1 Schooling_i + \beta_2 Experience_i + \beta_3 Experience_i^2 + \mu_i$$

- $\beta_1$: returns to education
- $\beta_2$ and $\beta_3$: returns to job experience (concave relationship)

### 1.2.2 Regression results

Here are the results from our Section 7 example. Using data from Carneiro, Heckman, and Vytlacil (2011), we estimate the Mincer Earnings Model. The `state` variable is different from the AER paper that we define completing college (`school == 4`) as the state variable.

- $\hat{\beta}_2$ is positive and $\hat{\beta}_3$ us negative, which is consistent with the Mincer Equation prediction

```
. reg wage state exp expsq, robust

Linear regression                               Number of obs     =      1,747
                                                F(3, 1743)        =     110.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1682
                                                Root MSE          =     .45545

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       state |   .5108677   .0281861    18.12   0.000     .4555855    .5661498
         exp |   .0454967   .0118544     3.84   0.000     .0222462    .0687471
       expsq |  -.0008483   .0006899    -1.23   0.219    -.0022014    .0005049
       _cons |   1.925447   .0508128    37.89   0.000     1.825786    2.025107
------------------------------------------------------------------------------

. test exp expsq

 ( 1)  exp = 0
 ( 2)  expsq = 0

       F(  2,  1743) =   43.70
            Prob > F =    0.0000
```

### 1.3 Joint significance of the linear and quadratic terms

In assessing whether it is appropriate to include the quadratic term in the OLS regression, we use F-test to test the joint significance of the estimated parameters of the two variables.

- The **null hypothesis** is that both of the estimated coefficients of the linear term (`exp`) and the quadratic term (`expsq`) are not statistically significant from zero.
- The F-test results indicate that we can reject the null hypothesis ($p<0.05$). The linear term and the quadratic term are jointly statistically significant. Thus, it is appropriate to use the quadratic functional form.
- *Note:* as the quadratic term faces the multicollinearity problem, the single t test is not sufficient to determine whether we should include the quadratic term.

### 1.4 Interpretation using predicted outcomes

Given the non-linear relationship between experience and wage, and the symmetric proposition of quadratic equation, we are no longer able to produce one-parameter change (either increasing or decreasing). It may be true that in some range of X, the relationship is increasing; while in the other range of X, the relationship changes to be decreasing.

We then calculate the difference/change in outcome ($\delta X$) to interpret the results.

We can use the approximate/derivative way or the exact way. Recall the partial relationship (holding education equal):

$$Y = \beta_2 Experience + \beta_3 Experience^2$$

### 1.4.1 The derivative formula

The derivative formula is:

$$\partial Y = \beta_2 \partial Experience + 2 * \beta_3 Experience \partial Experience$$

When we change $Experience$ from $X_1$ to $X_2$, we calculate that $Experience$ is the mid point of the the two values $= \frac{X1+X2}{2}$, and $\partial Experience$ is $(X2 - X1)$

$$\begin{aligned}
\partial Y &= \beta_2 * (X2 - X1) + 2 * \beta_3 * \frac{X2 + X1}{2} * (X2 - X1) \\
&= \beta_2 * (X2 - X1) + \beta_3 * (X2 + X1) * (X2 - X1)
\end{aligned}$$

### 1.4.2 The exact formula

The exact formula is:

$$\begin{aligned}
Y1 &= \beta_2 * X1 + \beta_3 * X1^2 \\
Y2 &= \beta_2 * X2 + \beta_3 * X2^2
\end{aligned}$$

Then the change in predicted outcome is (remember $a^2 - b^2 = (a + b) * (a - b)$):

$$\begin{aligned}
\partial Y = Y2 - Y1 &= (\beta_2 * X2 + \beta_3 * X2^2) - (\beta_2 * X1 + \beta_3 * X1^2) \\
&= \beta_2 * (X2 - X1) + \beta_3 * (X2^2 - X1^2) \\
&= \beta_2 * (X2 - X1) + \beta_3 * (X2 + X1) * (X2 - X1)
\end{aligned}$$

### 1.4.3 Examples

- Change in hourly wage from 0 to 1 year of experience
  - $0.0455 * (1 - 0) + (-0.0008) * (1 + 0) * (1 - 0) = 0.0447$
- Change in hourly wage from 40 to 41 year of experience
  - $0.0455 * (41 - 40) + (-0.0008) * (41 + 40) * (41 - 40) = -0.0193$
  - We know from the regression results that average white males reach their top earnings at $Experience = 28$. After that, each one additional year is associated with decreasing earnings.

- Change in hourly wage from 10 to 20year of experience
  - $0.0455 * (20 - 10) + (-0.0008) * (20 + 10) * (20 - 10) = 0.215$

# 2 Lecture notes #8: Returns to education and human capital theory

# 3 `Stata` tip #8: Forvalues/foreach

# PubPol 639 Section #8
# Causal effects using interaction terms

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**OBJECTIVES OF SECTION 8**

1. Interaction term: CITS
2. Lecture notes: Review of non-linear relationship in linear regression
3. `Stata` tip: CITS graph after regression

## 1 Interaction term: CITS

## 2 Lecture notes #9: Review of non-linear relationship in linear regression

- Note 0: This note serves as an outline of the materials that would be included in quiz 2. You should read and review (1) lecture slides; (2) section slides and notes; (3) SW textbook; (4) Assignment 3 questions and solutions. If you have done all of them and still have questions, email me or come to meet with me on Monday morning. Have fun!

1. Interpret regression results (magnitude, effect size, statistical significance)
   - If X is continuous variable
   - If X is binary/categorical variable variable
2. OVB
   - Draw the OVB diagram
   - Explain the direction of OVB (first explain signs in beta2 and gamma1)
3. Quadratic term
   - Interpret the relationship between X and Y
   - Predicted outcome when changing from X1 to X2, holding all else equal
   - $\hat{Y} = \beta_1 * (X2 - X1) + \beta_2 * (X2 + X1) * (X2 - X1)$
4. Interaction term
   - Interpretation of the main effect and the interaction effect
   - Predicted outcome (writing down the equation helps you) and draw the predicted lines
5. Log
   - **level-log**: each 1 percent increase in X is associated with 0.01*$\beta$ change (be specific: increase or decrease) in Y, holding all else equal.
   - **log-level**: each 1 unit increase in X is associated with 100*$\beta$ percent change (be specific: increase or decrease) in Y, holding all else equal.
   - **log-log**: each 1 percent increase in X is associated with $\beta$ percent change (be specific: increase or decrease) in Y, holing all else equal.
   - Note 1: If $beta > 0.1$ in log-level, we should calculate the exact percent change: (exp($\beta$)-1)*100
   - Note 2: $log(1) = 0$
6. *Relevant Practice Quiz questions*
   - Fall 2014 Quiz 2: (1), (2), (3), (4); Fall 2014 Quiz 3: (1), (2), (3), (4), (5), (6), (7); Winter 2015 Quiz 2: (1), (2), (3); Winter 2015 Quiz 3: (1), (2), (4)
   - Note 3: Some answers are too long, you may want think about shortening them
   - Note 4: The practice quizzes do not cover all the materials that may be included in our quiz 2

# 3 `Stata` tip #9: CITS graph after regression

# PubPol 639 Section #9
# When the outcome is binary

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**OBJECTIVES OF SECTION 9**

1. Binary response models
2. `Stata` tip: Simulation of logit/probit

# 1 Binary response models: A summary

## 1.1 Set-up

In real life, we often make yes/no choices that usually take two values for the outcome (1 and 0). Such variables are called *dummy* or *dichotomous* or *binary response* variables.

- To be, or not to be?
- Love me, or her?
- To enroll at U Michigan, or not?

If the number of your options exceeds two, we will be using other multi-responses models, which are based on the binary response models.

## 1.2 Probability

The expected value of a dummy variable $y_i \in \{1, 0\}$ is the probability of positive outcome (taking the value 1):

$$E(y_i) = 1 * P(y_i = 1) + 0 * P(y_i = 0) = P(y_i = 1) \tag{13}$$

## 1.3 Linear Probability Model

$$y = x\beta + \varepsilon, E(\varepsilon|x) = 0; P(y_i = 1) = x\beta \tag{14}$$

**Problems with LPM: Estimates and statistical inferences**

- The error term $\varepsilon$ is heteroscedastic

$$Var(\varepsilon|x) = P(y = 1)(1 - P(y = 1)) = x\beta(1 - x\beta) \tag{15}$$

  - Can be addressed using WLS
- The error term is not normally distributed (t-test is not guaranteed)
- Predicted $\hat{y}$ can lie outside $[0, 1]$ that does not represent a probability
- Incorrect linearity assumption
  - Estimates will be highly sensitive to the range of data observed in the sample
  - May understate/overstate the magnitude of the true effects
  - As the underlying estimation function is similar $P(y = 1) = \mathbf{F}(x\beta)$, LPM gives the correct sign of the effect

**Mostly harmless econometrics: Probit better than LPM?**

- Practical advantages of LPM
  - Easier to calculate
  - The parameters are directly interpretable
  - Fixed effects and instrumental variables estimators can easily be implemented
    * Note: Adding fixed effects as dummy variables in the probit or logit model will yield biased estimates (when within group N is small)
    * The statistical problem is that, as the number of groups tends to infinity, the number of estimated parameters increases at the same rate. The estimates are not consistent.
    * See more in Greene (2004), The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal*, *7*(1), 98-119.
- Comments on model choice from Jorn-Steffen Pischke:

```
We care about the marginal effects. The LPM will do a pretty good job estimating
those. If the CEF (conditional expectation function) is linear, as it is for a
saturated model, regression gives the CEF - even for LPM. If the CEF is
non-linear, regression approximates the CEF. Usually it does it pretty well.
Obviously, the LPM won't give the true marginal effects from the right nonlinear
model. But then, the same is true for the "wrong" nonlinear model!

The fact that we have a probit, a logit, and the LPM is just a statement to the
fact that we don't know what the "right" model is. Hence, there is a lot to be
said for sticking to a linear regression function as compared to a fairly
arbitrary choice of a non-linear one! Nonlinearity per se is a red herring.
```

## 1.4 Logit and Probit

- Binary response models directly describe the probability $P(y_i = 1)$ of the dependent binary variable $y_i$, using an index function (latent variable):

$$P(y = 1|x) = P(y* > 0|x) = P(x\beta + \varepsilon > 0) = \mathbf{F}(x\beta) \tag{16}$$

- The index function maps the single index into dichotomous choice $[0, 1]$

$$\mathbf{F}(-\infty) = 0, \mathbf{F}(\infty) = 1 \tag{17}$$

  - $\mathbf{F}(x\beta)$ is NOT a linear of $\beta$
- Logit and Probit models are almost identical and the model choice is usually arbitrary
  - Logit: $P(y = 1|x) = \mathbf{F}(x\beta) = \frac{1}{1+\exp^{-x\beta}}$
  - Probit: $P(y = 1|x) = \mathbf{F}(x\beta) = \Phi(x\beta) = \int_{-\infty}^{x\beta} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}t^2}\, dt$

### 1.5 Estimation using Maximum Likelihood

- The likelihood function is:

$$\mathcal{L} = \prod_{i=1}^{N} P_i^{y_i}(1-P_i)^{1-y_i} = \prod_{i=1}^{N} \mathbf{F}(x\beta)^{y_i}(1-\mathbf{F}(x\beta))^{1-y_i} \tag{18}$$

- The corresponding log likelihood function is:

$$log\mathcal{L} = \sum_{i=1}^{N}[(1-y_i)log(1-\mathbf{F}(x\beta)) + y_i log\mathbf{F}(x\beta)] \tag{19}$$

- No analytical solution to the FOCs and numerical optimization routines are used (iterations)

### 1.6 Marginal effects

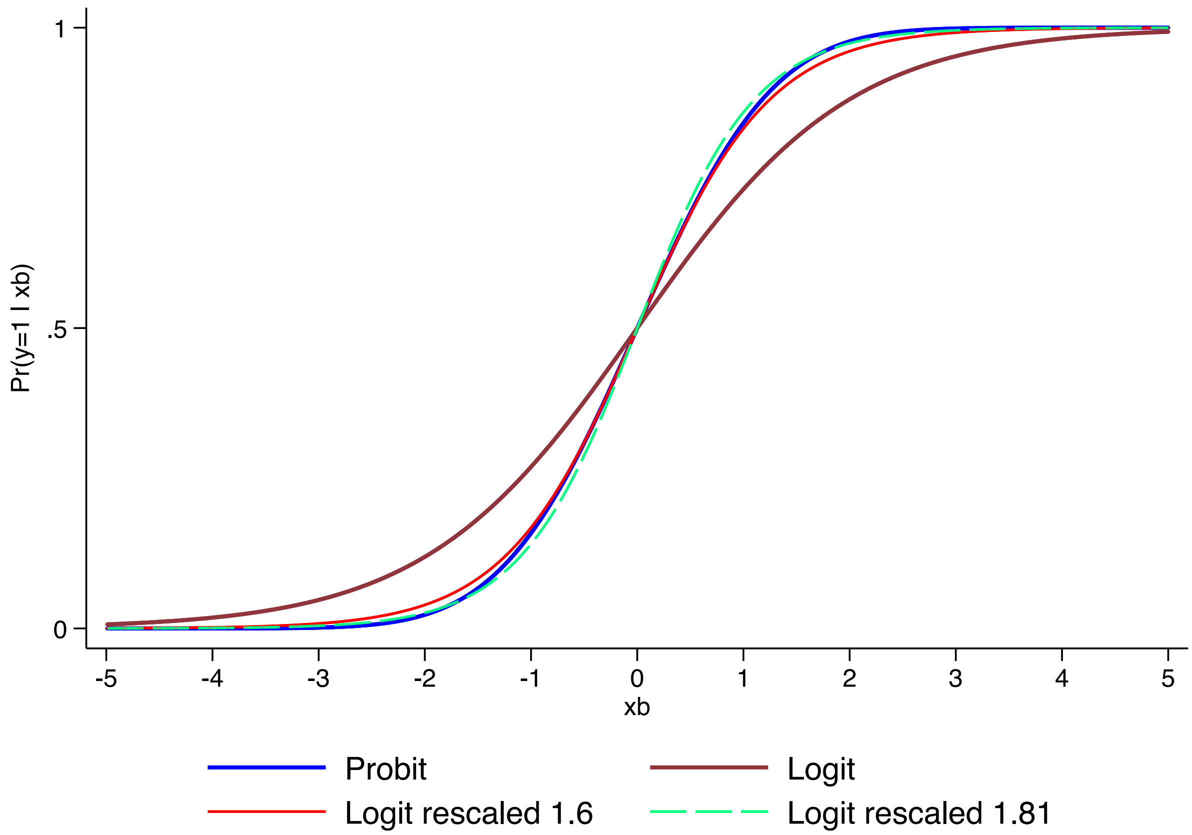
- LPM
  - Raw coefficients ($\beta$) are average marginal effects
  - Important to report robust s.e. due to heteroscedasticity
- Logit and Probit
  - Raw coefficients are NOT marginal effects: changes in predicted outcomes (see below) given a marginal change in $x$
    - Logit: log odds (log of odds ratio)
    - Probit: z-score
  - Marginal effect of $x_{ik} = \frac{\partial P(y_i=1|x_i)}{\partial x_{ik}}$: changes in predicted probabilities given a marginal change in $x$
    - Logit: $\frac{exp(x\beta)}{(1+exp(x\beta))^2}\beta_k$
    - Probit: $\phi(x_i\beta)\beta_k$
  - Marginal effects depend on all $x_ik$ that any individual has a different marginal effect
  - Different ways to summarize the marginal effects
    - average marginal effect
    - marginal effect at mean/median
    - marginal effect at any values
  - `Stata` command: `margins`

### 1.7 More topics

- `cloglog`, assuming the error term is log-log distributed (rather than logit/probit)
- `scobit`, relaxing the assumption that the marginal effect is greatest when $P(y=1) = 0.5$
- `hetprobit`, heteroscedasticity in probit model

- `ivprobit`, endogenous regressors
- `biprobit`, seemingly unrelated regression
- `xtlogit` or `xtprobit`, panel data

## 2 `Stata` tip #10: Simulating and comparing logit and probit models




***Stata codes***

```
* Compare Logit and Probit models
// global date "10_12_2016"

// set observations
clear
set obs 1000

// create a sequence of numbers (-5, 5)
gen xb = _n*0.01 - 5
sum xb

// create logit and probit
gen logit = 1/(1+exp(-xb))
sum logit

gen probit = normal(xb)
sum probit
```

```
// create scaled logit
gen logit_scale1 = 1/(1+exp(-xb*1.6))
gen logit_scale2 = 1/(1+exp(-xb*1.81))

// graph
twoway  (line probit xb, lwid(medthick)) lcolor(blue)   ///
(line logit xb, lwid(medthick))   ///
(line logit_scale1 xb, lcolor(red))    ///
(line logit_scale2 xb, lcolor(mint)   ///
lpattern(longdash)  ///
xtitle("xb", size(small))  ///
ytitle("Pr(y=1 | xb)", size(small))  ///
ylabel(0(.5)1, nogrid labs(small) angle(h))    ///
xlabel(-5(1)5, angle(0) labs(small))  ///
legend(order(1 2 3 4 ) label(1 "Probit") label(2 "Logit")   ///
label(3 "Logit rescaled 1.6") label(4 "Logit rescaled 1.81")   ///
cols(2) region(color(none)) margin(zero))   ///
graphregion(color(white)) bgcolor(none))

graph save "logit_probit_${date}.gph", replace
graph export "logit_probit_${date}.pdf", replace
erase "logit_probit_${date}.gph"
```

# PubPol 639 Section #10
# From estimation to identification

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**OBJECTIVES OF SECTION 10**

1. Casual questions in MHE
2. Difference-in-differences
3. Lecture notes: Review of estimation and DID
4. `Stata` tip: `reghdfe`

## 1 Causal questions in (Angrist and Pischke, 2008)

## 2 Difference in differences using natural experiments

# 3 Lecture notes #10: Review of estimation and DID

- Note 0: This note serves as an outline of the materials that would be included in quiz 3. You should read and review (1) lecture slides, (2) section slides and notes, (3) SW textbook, (4) Assignment 4 questions and solutions. If you have done all of them and still have questions, email me or come to meet with me. Have fun!

## 3.1 Review of generalized linear regressions

### 3.1.1 Interpret regression results (coefficient, economic/policy significance, statistical significance)

- Recent examples in quiz 2 and assignment 4

***If X is a continuous variable variable***

- The "formula"
  - **On average, each one unit increase in X is statistically (in)significantly associated with beta unit change in Y, holding all else equal.**
- Make real world phrases
  1. one unit (X)
  2. X
  3. (in)significantly
  4. beta
  5. unit (Y)
  6. Y
  7. holding all else equal - omit this in the bivariate regression
- Assess the economic significance (may use one of the three ways)
  1. Effect size
  2. Percent change relative to the baseline mean
  3. Compare with another (well-known) variable
- one unit of X
  - Obvious in most cases
    - e.g., 1 dollar increase in school spending
  - 100 percentage point in percent measure
    - e.g., if % free lunch is measured from 0-1
  - 100 percent in log(X)
    - e.g., if X variable is log(spending)
- unit of Y

- Obvious in most cases
- percentage point in dummy Y variable (**Linear probability model**)
  * e.g., Y is a dummy variable indicating college graduation
  * interpret "beta unit" as beta*100 percentage point
- 100 percent in log(Y)

- beta
  - Many times, 100 percentage point (or percent) seems too large
  - We use 1 pp (0.01 unit) or 10 pp (0.1 unit) as the marginal change
  - The corresponding change in Y is also 0.01/0.1 unit

***If X is a binary/categorical variable variable***

- The "formula"
  - **On average, group 1 (*X=1 group*) has statistically (in)significant beta unit higher/lower in Y than group 0 (*X=0 group*), holing all else equal.**

- Real world phrases
  1. group 1
  2. (in)significantly
  3. beta
  4. unit (Y)
  5. higher/lower (or other comparison terms)
  6. Y
  7. group 0
     - For more than two groups, be careful of the reference group

#### 3.1.2 Non-linear relationship in linear regressions

- Quadratic term & interactions
  - Quadratic term can be seen as a special case of interaction term
    * In interaction term regression, we have two linear terms and one interaction term
    * In quadratic term regression, we have only one linear term (if you want, you can write two linear terms for the same X variable) and one interaction/quadratic term
  - Write down the regression model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2$$

- Total effect of $X_1$ on $Y$

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

***Examples***

- Example 1: Quadratic

$$Wage = \alpha Edu + \beta_1 Experience + \beta_2 Experience^2 + \varepsilon$$

  - Total effect of `Experience`: $\frac{\Delta Y}{\Delta Experience} = \beta_1 + 2\beta_2 Experince$
  - The "effect"/association/slope for a worker with 30 years experience: $\beta_1 + 2\beta_2 * 30$
  - Suppose $\beta_1 > 0$ and $\beta_2 > 0$, the "effect" is smaller (the slope is flatter) than that for a worker with 35 years experience $\beta_1 + 2\beta_2 * 35$

- Example 2: Continuous*continuous

$$Wage = \beta_1 Edu + \beta_2 Experience + \beta_3 Edu * Experience + \varepsilon$$

  - Total effect of `Edu`: $\frac{\Delta Y}{\Delta Edu} = \beta_1 + \beta_3 Experince$
  - The "effect"/association/slope of education varies along with one's working experience (if there is no interaction term, the "effect" is constant: $\beta_1$)
  - **Interpretation of** $\beta_3$: On average, each one year increase in job experience is statistically (in)significantly associated with $\beta_3$ unit change in the association between education and wage, holding all else equal.

- Logarithm
  - **level-log**: each 1 percent increase in X is associated with 0.01*$\beta$ change (be specific: increase or decrease) in Y, holding all else equal.
  - **log-level**: each 1 unit increase in X is associated with 100*$\beta$ percent change (be specific: increase or decrease) in Y, holding all else equal.
  - **log-log**: each 1 percent increase in X is associated with $\beta$ percent change (be specific: increase or decrease) in Y, holing all else equal.
    - Note 1: If $beta > 0.1$ in log-level, we should calculate the exact percent change: (exp($\beta$)-1)*100
    - Note 2: $log(1) = 0$
    - Note 3: For binary X variable, NEVER use "association" interpretation, use group difference.

#### 3.1.3 Probit and logit models

- The idea
  - Associated change in the probability of Y given 1 unit/marginal change in X

* $x_{ik} = \frac{\partial P(y_i=1|x_i)}{\partial x_{ik}}$

    * Continuous X: first derivative
    * Binary X: group difference

* LPM
  - $P(y = 1|x) = y = x\beta$
  - Raw coefficients ($\beta$) are average marginal effects

* Logit and Probit
  - $P(y = 1|x) = \mathbf{F}y = \mathbf{F}(x\beta)$
  - Raw coefficients are NOT marginal effects
    * Logit: $\frac{exp(x\beta)}{(1+exp(x\beta))^2}\beta_k$
    * Probit: $\phi(x_i\beta)\beta_k$
  - Marginal effects depend on all $x_ik$ that any individual has a different marginal effect
  - Different ways to summarize the marginal effects
    * average marginal effect
    * marginal effect at mean/median
    * marginal effect at any values

## 3.2 DID (and panel data)

### 3.2.1 Omitted variable bias in panel data

* No secrets in panel data
  - Two additional groups of control variables
    * **panel fixed effects**: all variables change between panels (e.g., states, schools, individuals), but do not change over time within each panel
    * **time fixed effects**: all variables change over time *commonly* for all panels
  - Use the "OVB" triangle to assess the direction of the bias
  - Even when we control for panel and time fixed effects, there may still be panel-time variations
    * DID estimator ($policy * post$)
      · *policy* is the panel difference (between treatment and control groups)
      · *post* is the time difference (between pre and post)

### 3.2.2 DID estimates

- Similar to the interaction term in the interaction term regression

$$Y = \beta_0 + \beta_1 Policy + \beta_2 Post + \beta_3 Policy * Post$$

- Estimating the policy effect
  - You can not simply compare $Y\{Policy = 1, Post = 1\}$ with $Y\{Policy = 0, Post = 1\}$, because the two groups may be very different
    - Panel fixed effects help
  - You can not simply $Y\{Policy = 1, Post = 1\}$ with $Y\{Policy = 1, Post = 0\}$, because there may be many things changing between the two periods
    - Time fixed effects help

$$\begin{aligned} DID &= (Y\{Policy = 1, Post = 1\} - Y\{Policy = 1, Post = 0\}) \\ &\quad - (Y\{Policy = 0, Post = 1\} - Y\{Policy = 0, Post = 0\}) \\ &= (\beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_1)) - (\beta_0 + \beta_2 - (\beta_0)) \\ &= (\beta_2 + \beta_3) - \beta_2 \\ &= \beta_3 \end{aligned}$$

- **Key assumption**: common trends between the two groups
  - After controlling for panel/time fixed effects, the remaining potential OVB is panel-time varying variables
  - In that case, we don't know $\beta_3$ is **solely** from the policy or from other omitted panel-time varying variables
    - Then the control group can't be a counterfactual of the treatment group
  - Common trends in the pre-policy periods between the two groups are often used to validate the no-OVB claim
- Bias when using a wrong control group
  - $bias = Slope_{Policy} - Slope_{Control}$ in pre-policy periods
    - What is the slope?
    - Graph helps you.
  - If $bias = 0$, common trends!

$$\begin{aligned} DID &= (Y\{Policy = 1, Post = 1\} - Y\{Policy = 1, Post = 0\}) \\ &\quad + bias\ term\ in\ the\ pre - policy\ periods \\ &\quad - (Y\{Policy = 0, Post = 1\} - Y\{Policy = 0, Post = 0\}) \\ &= (\beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_1)) + bias - (\beta_0 + \beta_2 - (\beta_0)) \\ &= (\beta_2 + \beta_3) + bias - \beta_2 \\ &= \beta_3 + bias \end{aligned}$$

## 3.3 Practice questions: Winter 2017, Quiz 5

### 3.3.1 Part 1

You want to know whether greater access to charter schools is associated with higher test scores for $8^{th}$ graders. You collect panel data on average test scores, charter school access, and a few other variables for each state from 2000 to 2015.

The table below presents (hypothetical) regression coefficients (and standard errors) from models that regress the *standardized test score* of $8^{th}$ graders on the *fraction of students in the state with access to charter schools* and the *state unemployment rate*. All three of these variables vary over time for each state. The indicated columns include year and/or state fixed effects.

| Dept variable: Standardized test score | | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Fraction with access to charter school (mean = 0.25) | -0.100<br>(0.042) | -0.080<br>(0.032) | 0.010<br>(0.023) |
| | | | |
| State unemployment rate (mean = 4.5) | -0.030<br>(0.010) | -0.030<br>(0.010) | -0.030<br>(0.010) |
| | | | |
| Year FE | No | Yes | Yes |
| | | | |
| State FE | No | No | Yes |
| | | | |
| Constant | 0.010<br>(0.020) | 0.010<br>(0.020) | 0.010<br>(0.020) |
| | | | |
| Sample | 2000-2015 | 2000-2015 | 2000-2015 |
| | | | |
| Observations | 800 | 800 | 800 |
| | | | |

Notes: Standard errors in parentheses, clustered by state.

- Consider several "facts" about charter schools:
  - **Fact A.** States that experienced the fastest growth in the number of charter schools from 2000 to 2015 also experienced the greatest increase in the unemployment rate during that time period.
  - **Fact B.** The number of charter schools (and thus charter access) increased over time across all states from 2000 to 2015, while at the same time aggregate achievement scores of US 8th graders declined.

- **Fact C.** The states that typically have many charter schools (and greater charter access) also tend to have lower levels of parental education.

- Questions

  3 Which of these "facts" (if any) would cause estimates in (1) to suffer from omitted variable bias?

  4 Which of these "facts" (if any) would cause estimates in (2) to suffer from omitted variable bias?

  5 Which of these "facts" (if any) would cause estimates in (3) to suffer from omitted variable bias?

***Solutions***

- Step 1
  - Understand the regressions in columns (1)-(3)
  - Write down the regression equations

$$(1) Y = \beta_0 + \beta_1 * CharterAccess + \beta_2 * Unemployment + \varepsilon$$

$$(2) Y = \beta_0 + \beta_1 * CharterAccess + \beta_2 * Unemployment + YearFE + \varepsilon$$

$$(3) Y = \beta_0 + \beta_1 * CharterAccess + \beta_2 * Unemployment + YearFE + StateFE + \varepsilon$$

- Step 2
  - Use the OVB triangle
    - *In (1), possible omitted variables:* time-state variations, time FE, state FE
    - *In (2), possible omitted variables:* time-state variations, state FE
    - *In (3), possible omitted variables:* time-state variations

- Step 3
  - Determine the variable type in each question
    - *Fact A:* time-state variations (varies over time within state)
    - *Fact B:* time FE (varies over time for all states)
    - *Fact C:* state FE (varies between states, but remains the same for the same state)

- Step 4
  - Answers
    - *3:* B & C
    - *4:* C
    - *5:* None.

- In this question, be careful that, although unemployment rate varies over time within states and generally could be a source of OVB in questions 3-5, **but the regressions already control for unemployment**.
- *If the regressions do not control for unemployment, then the answers should be: ABC, AC, A.*

### 3.3.2 Part 2

The state of Alabama recently passed legislation permitting charter schools to operate starting in 2017. The table below shows the 8th grade test scores in Alabama and two other states from 2014 to 2017.

| 8th grade test scores in 3 states (FAKE DATA) | | | |
|---|---|---|---|
| | Alabama (Treatment group) | Mississippi (Control group 1) | Georgia (Control group 2) |
| 2014 (before) | 20 | 23 | 20 |
| 2015 (before) | 19 | 22 | 18 |
| 2016 (before) | 18 | 21 | 16 |
| 2017 (after) | 23 | 26 | 20 |
| | | | |

- Questions

  8 Using data from 2016 and 2017 for Alabama and Mississippi, construct a DID estimate of the effect of permitting charter schools on test scores

***Solutions***

- Step 1
  - Write down the DID regression equation (interactions of two dummy variables)

$$(1) Y = \beta_0 + \beta_1 * Treatment(Alabama) + \beta_2 * After + \beta_3 * Treatment * After + \varepsilon$$

- Step 2
  - Compute predicted outcome ($\hat{Y}$) for the four groups
    1. $\hat{Y}(AL, 2016) = \beta_0 + \beta_1 = 18$
    2. $\hat{Y}(AL, 2017) = \beta_0 + \beta_1 + \beta_2 + \beta_3 = 23$
    3. $\hat{Y}(MS, 2016) = \beta_0 = 21$

4. $\hat{Y}(MS, 2017) = \beta_0 + \beta_2 = 26$

- Step 3
  - Solve the equations to get $\beta_3$ (the DID estimate)
    - from 1 and 2, we get $\beta_2 + \beta_3$ = 23-18 = 5
    - from 3 and 4, we get $\beta_2$ = 26-21 = 5
    - Finally, we get $\beta_3$ = 5 - 5 =0
  - Step 4 (forget about steps 1-3)

$$\begin{aligned} DID(effect) &= (\hat{Y}(AL, 2017) - \hat{Y}(AL, 2016)) - (\hat{Y}(MS, 2017) - \hat{Y}(MS, 2016)) \\ &= (23 - 18) - (26 - 21) \\ &= 5 - 5 \\ &= 0 \end{aligned}$$

- Questions

9 If you had used Georgia as the control group instead of Mississippi, would your estimate likely over-state (positive OVB) or under-state (negative OVB) the true effect of permitting charter schools? Explain.

10 Is Mississippi a good control?

***<u>Solutions</u>***

- Step 1
  - Note that: the key assumption of DID is **common trends**
  - To evaluate whether a control group is good or not (then bias), you need to test whether the control group has the same pre-policy trend/slope with the treatment group
- Step 2
  - Compute the pre-policy trends
  - For simplicity, you can also just draw some draft graphs
  - Below is a simple way to calculate the slope (may be wrong iif the slope is not linear, then go back to the graphic way)
    - AL slope = $\frac{18-20}{2016-2014}$ = -1
    - MS slope = $\frac{21-23}{2016-2014}$ = -1
    - GA slope = $\frac{16-20}{2016-2014}$ = -2
- Step 3
  - According to Xiaoyang's review notes, $bias = Slope_{Policy} - Slope_{Control}$
    - Between AL and MS: bias = -1 - (-1) = 0, so MS is a good control group
    - Between AL and GA: bias = -1 - (-2) = 1, so these is an upward (overstated, positive) bias

# 4 `Stata` tip #11: CITS graph after regression

# PubPol 639 Section #11
# Quasi-experimental designs


Xiaoyang Ye
*University of Michigan*


September 5, 2018

**OBJECTIVES OF SECTION 11**

1. Difference-in-differences
2. Panel data
3. RD
4. IV and treatment effects
5. Other topics: Matching, synthetic control

## 1 Difference in differences using natural experiments

## 2 Panel data

## 3 Regression discontinuity

## 4 Instrumental variables

## 5 Other topics

# PubPol 639 Section #12
# Program evaluation toolkit 1

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**OBJECTIVES OF SECTION 12**

1. Reading a quantitative & causal inference study
2. Your `Stata` skills

## 1 Reading a quantitative & causal inference study

Adapted from Angrist and Pischke (2014) and http://users.nber.org/ nikolovp/studentresources/re-freport.pdf

### 1.1 Overview

1. What is the causal question?

- Effect of X on Y
  - This should be clearly stated in Abstract/Introduction/Conclusion, many times also in the Title
  - The authors may use proxy variables for X, but the research question is about X

2. What are the key findings?

- What is the estimated effect of X on Y?
  - This should also be clearly stated in Abstract/Introduction/Conclusion
  - Do the authors care about the direction of the effect/relationship, or the magnitude (statistical, economic) as well?
  - Holding *what* constant?

## 1.2 Identification strategy

1. What is the *ideal* experiment?

- The ideal experiment is to randomly assign X (the treatment status) to different samples and then to estimate the causal effect by comparing Y (the outcome) between the treatment group (samples with X) and the control group (samples without X) after a certain period.
  - In many cases, such ideal experiments may not be feasible - or, hypothetical (shortcomings of RCT: e.g., cost, ethical issues)
  - Random sample $\neq$ random assignment
  - "Fundamentally unidentified questions" cannot be answered by experiments. For example, we cannot manipulate one's race to estimate the causal effect of race; but we can manipulate how people *believe* you to be black or white.

2. Does the paper have a clear identification strategy?

- Check how the authors make their argument - this should also be clearly stated in the Method section
  - RCT, yes
  - Quasi-experimental, yes if correctly specified
  - Nothing "(quasi-)experimental", probably not

3. If not, how could omitted variable bias affect the results?

- Check how the authors make their argument.
  - Use the OVB triangle
  - Omitted variables that are correlated with both X and Y

4. If so, what is the identification strategy?

- The most popular designs (that you have learned in PP639)
  - By knowledge of assignment mechanism
    * RCT
    * RD
  - By selection
    * DID
    * Panel data (extensions: ITS, CITS, event study)

        * IV
        * *Other designs: matching, synthetic control, machine learning methods for causal inference*
- One study may combine different identification strategies (with different assumptions)
    - For example, we may use DID, event study, and panel data (recent example: Jackson, Johnson, and Persico (2015))
- Identification $\neq$ estimation
    - Estimation methods (that you have learned): t-test, OLS, OLS with non-linear relationships (quadratic, interaction, log), binary dependent variable models (LPM, Logit, Probit)
    - When asked about empirical approaches, **think about both estimation and identification**

### 1.3 Data

1. What data do the authors use to answer the research question?

- Data source
    - This should also be clearly stated in the Data section

2. What is the unit of observation?

- What does each one observation represent?
    - Example in PP639: one person, one school, one school in a year (school-year observation), county-year observation

3. What is the sample size?

- Check the main tables (summary statistics or the main regression table)

4. How do the authors deal with missing data?

- Does missing data cause non-random sampling problem?

5. Measures of Y and X variables?

- Advantages and disadvantages of using particular measures/proxies
- Survey data vs. administrative data (vs. "big data")
  - Survey data may be detailed, but sampling and sample size may be problems
  - Administrative data are more accurate and may cover the whole population, but they may lack of details (or lack of particular measures) and still have non/mis-reporting issues

### 1.4 Empirical analysis

1. What is the empirical model?

- Many papers should report the main regression model, like this:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it}$$

  - $Y_{it}$ represents the outcome for a unit $i$ in time $t$, $X_{it}$ are the covariates, $\varepsilon_{it}$ is the residual
  - You may want to write down all the covariates, like this:

$$Y_{it} = \beta_0 + \beta_1 * Black_{it} + \beta_2 * Female_{it} + \varepsilon_{it}$$

  - No subscripts $it$ for the parameters ($\beta_0$, $\beta_1$) because parameters are the same for every observation
  - Don't forget about the constant term and fixed effects (if applicable)
- Sample regression model or the prediction model is based on the estimates

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 * Black_{it} + \hat{\beta}_2 * Female_{it}$$

2. Is the interpretation of the coefficients of interest (and the effect size) clear and correct?

- Interpretation will be a separate part in the next section

### 1.5 Threats to validity

1. What are the most important assumptions embedded in the empirical strategy?

- Why do the treatment status vary?

  - By assignment mechanism (e.g., random assignment by researchers, policies, or the nature)?
  - By selection (e.g., some people are more likely to choose the treatment due to either observed in the data or unobserved reasons)?

2. Any internal validity threats? How do they affect the results?

- Threats to internal validity compromise our confidence in saying that a causal relationship exists between the independent and dependent variables
- The biggest threat: OVB - selection bias
  - By assignment mechanism (e.g., random assignment by researchers, policies, or the nature)?
  - By selection (e.g., some people are more likely to choose the treatment due to either observed in the data or unobserved reasons)?
- Other issues: measurement error, instrumentation, John Henry effect

3. What are the robustness checks?

- Empirical studies should have tons of robustness checks
- Check how the authors discuss about these checks and then decide whether you believe the main conclusion still holds

4. Any external validity concerns?

- Threats to external validity compromise our confidence in stating whether the study's results are applicable to other groups
  - Think about whether we care about the direction of the relationship or the specific magnitude
- Population validity: Representative sample? Generalizable?
- Ecological validity: Interactive effect? Reactive effect?

### 1.6 Conclusion

1. What are the policy implications?
2. Any related future research directions?

## 2 Your `Stata` skills do not help your finals, but your future life

- See "Refresh your `Stata` memories in 46 minutes.pdf"

# PubPol 639 Final Review
# Program evaluation toolkit 2

Xiaoyang Ye
*University of Michigan*

September 5, 2018

**Conducting a quantitative & causal inference study**

- Estimation
  - What are we regressing for?
  - The world is often non-linear
- Identification
  - OVB
  - Experimental design
  - Quasi-experimental designs

## 3 Estimation

### 3.1 What are we regressing for?

#### 3.1.1 OLS regression model

- Population regression model

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

- Population regression line

  - A linear line without the residual term

$$Y_i = \beta_0 + \beta_1 X_i$$

- Sample regression line

  - A linear line with estimated parameters from a specific sample
  - A new sample will generate different estimates
  - Predicted outcome: $\hat{Y}_i$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

### 3.1.2 Interpretations

***If X is a continuous variable variable***

- The "formula"
  - **On average, each one unit increase in X is statistically (in)significantly associated with beta unit change in Y, holding all else equal.**
- Make real world phrases
  1. one unit (X)
  2. X
  3. (in)significantly
  4. beta
  5. unit (Y)
  6. Y
  7. holding all else equal - omit this in the bivariate regression
- one unit of X
  - Obvious in most cases
    - e.g., 1 dollar increase in school spending
  - 100 percentage point in percent measure
    - e.g., if % free lunch is measured from 0-1
  - 100 percent in log(X)
    - e.g., if X variable is log(spending)
- unit of Y
  - Obvious in most cases
  - percentage point in dummy Y variable (**Linear probability model**)
    - e.g., Y is a dummy variable indicating college graduation
    - interpret "beta unit" as beta*100 percentage point

- 100 percent in log(Y)
    - interpret "beta unit" as beta*100 percent

- beta
  - Many times, 100 percentage point (or percent) seems too large
  - We use 1 pp (0.01 unit) or 10 pp (0.1 unit) as the marginal change
  - The corresponding change in Y is also 0.01/0.1 unit

***If X is a binary/categorical variable variable***

- The "formula"
  - **On average, group 1 (*X=1 group*) has statistically (in)significant beta unit higher/lower/larger/smaller in Y than group 0 (*X=0 group*), holing all else equal.**

- Real world phrases
  1. group 1, group 0
     - For more than two groups, be careful of the reference group
  2. (in)significantly
  3. beta
  4. unit (Y)
  5. higher/lower (or other comparison terms)
  6. Y
  7. holding all else equal - omit this in the bivariate regression

### 3.1.3 More on interpretations

1. Economic significance (whether the estimated coefficient is *large*?)

   - Effect size
   - Percent change relative to the baseline mean
   - Compare with another (well-known) variable

2. Compare coefficients (absolute value)

   - Same Y, different Xs
     - Partial standardized coefficients: $\beta \times sd(X)$

- Different Ys, Same X
  - Effect size: $\frac{beta}{sd(Y)}$
- Different Ys, different Xs
  - Standardized coefficients: $\frac{\beta \times sd(X}{sd(Y)})$

3. Heteroskedasticity and multicollinearity

- The assumption of homoscedasticity (meaning "same variance") is central to linear regression models
- Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable
- Heteroscedasticity will underestimate the standard error, thus overestimate the t stat (and then p-value and statistical significance)
- Solution: robust standard error, clustered standard error

4. Multicollinearity

- Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated
- Problems
  - The partial regression coefficient due to multicollinearity may not be estimated precisely.
  - The standard errors are likely to be high.
  - Multicollinearity results in a change in the signs as well as in the magnitudes of the partial regression coefficients from one sample to another sample.
  - Multicollinearity makes it tedious to assess the relative importance of the independent variables in explaining the variation caused by the dependent variable.
- Solution: drop one or more of the variables

5. R-squared (and adjusted R-squared)

- R-squared

$$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$
$$SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- $R^2$ does not imply the regression is either "good" or "bad", it only indicates that other important factors also impact the outcome
- We can't say one model is better than the other because the former has larger R-squared
- Model choice should be based whether one model (1) identifies causal effects (no OVB), (2) best fits the data (e.g., linear or non-linear relationship)

• Adjusted R-squared

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

- Adjusted R-squared $\leqq$ R-squared

#### 3.1.4 Prediction

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Predicted outcome for each observation *i*: $\hat{Y}_i$
- Plug in X values of observation *i*
- Don't forget about the constant term
- Be careful of the unit of X values
  - In the example of regressing school average math achievement on school average free lunch student share
  - "% free lunch" is measured from 0 to 1 (100 percent)
  - We need to plug in **0.3** rather than *30* for a school with 30% free lunch students

## 3.2 The world is often non-linear

### 3.2.1 Quadratic term

$$Y = a + bX + cX^2$$

- Quadratic equations have between one and three terms, one of which always incorporates $X^2$ (a and b can be 0). c determines whether the associated quadratic line is convex or concave, $X = \frac{-b}{2c}$ determines the position of maximum (concave when c<0) or minimum (convex when c>0) value of $Y$.
- When graphed, quadratic equations produce a U-shaped curve known as a parabola. The line of symmetry is an imaginary line which runs down the center of this parabola and cuts it into two equal halves. This line is commonly referred to as the axis of symmetry. It can be found quite quickly by using a simple algebraic formula ($\frac{-b}{2c}$).
- The symmetric proposition means that from $-\infty$ to $\infty$ in X, we always have one half of the symmetry line is decreasing and the other half is increasing (the linear line , in contrast, is always increasing or decreasing). When interpreting the changing relationship, always define the range of your data.
- Predicted outcome when changing from X1 to X2, holding all else equal
  - $\Delta\hat{Y} = \beta_1 * (X2 - X1) + \beta_2 * (X2 + X1) * (X2 - X1)$

### 3.2.2 Interaction term

- Write down the regression model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2$$

- Total effect of $X_1$ on $Y$

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

- Example 1: Quadratic

$$Wage = \alpha Edu + \beta_1 Experience + \beta_2 Experience^2 + \varepsilon$$

  - Total effect of `Experience`: $\frac{\Delta Y}{\Delta Experience} = \beta_1 + 2\beta_2 Experince$
  - The "effect"/association/slope for a worker with 30 years experience: $\beta_1 + 2\beta_2 * 30$
  - Suppose $\beta_1 > 0$ and $\beta_2 > 0$, the "effect" is smaller (the slope is flatter) than that for a worker with 35 years experience $\beta_1 + 2\beta_2 * 35$

- Example 2: Continuous*continuous

$$Wage = \beta_1 Edu + \beta_2 Experience + \beta_3 Edu * Experience + \varepsilon$$

  - Total effect of `Edu`: $\frac{\Delta Y}{\Delta Edu} = \beta_1 + \beta_3 Experince$
  - The "effect"/association/slope of education varies along with one's working experience (if there is no interaction term, the "effect" is constant: $\beta_1$)
  - **Interpretation of $\beta_3$**: On average, each one year increase in job experience is statistically (in)significantly associated with $\beta_3$ unit change in the association between education and wage, holding all else equal.

#### 3.2.3 Log

- **level-log**: each 1 percent increase in X is associated with 0.01*$\beta$ change (be specific: increase or decrease) in Y, holding all else equal.
  - diminishing returns to X
- **log-level**: each 1 unit increase in X is associated with 100*$\beta$ percent change (be specific: increase or decrease) in Y, holding all else equal.
  - increasing returns to X
- **log-log**: each 1 percent increase in X is associated with $\beta$ percent change (be specific: increase or decrease) in Y, holing all else equal.
  - Constant elasticity

- Note 1: If $beta > 0.1$ in log-level, we should calculate the exact percent change: (exp($\beta$)-1)*100
- Note 2: $log(1) = 0$

### 3.3 Binary outcome

- Binary response models directly describe the probability $P(y_i = 1)$ of the dependent binary variable $y_i$, using an index function (latent variable):

$$P(y = 1|x) = P(y* > 0|x) = P(x\beta + \varepsilon > 0) = \mathbf{F}(x\beta)$$

- The index function maps the single index into dichotomous choice $[0,1]$

$$\mathbf{F}(-\infty) = 0, \mathbf{F}(\infty) = 1$$

  - $\mathbf{F}(x\beta)$ is NOT a linear of $\beta$

- Logit and Probit models are almost identical and the model choice is usually arbitrary
  - LPM: $P(y=1|x) = \mathbf{F}(x\beta) = x\beta + \varepsilon$
  - Logit: $P(y=1|x) = \mathbf{F}(x\beta) = \frac{1}{1+\exp^{-x\beta}}$
  - Probit: $P(y=1|x) = \mathbf{F}(x\beta) = \Phi(x\beta) = \int_{-\infty}^{x\beta} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}t^2} dt$

- Model choice: LPM vs. Logit/Probit
  - Problems with LPM: Estimates and statistical inferences
    1. The error term $\varepsilon$ is heteroscedastic
    2. The error term is not normally distributed (t-test is not guaranteed)
    3. Predicted $\hat{y}$ can lie outside $[0,1]$ that does not represent a probability
    4. Incorrect linearity assumption
       * Estimates will be highly sensitive to the range of data observed in the sample
       * May understate/overstate the magnitude of the true effects
       * As the underlying estimation function is similar $P(y=1) = \mathbf{F}(x\beta)$, LPM gives the correct sign of the effect - it's OK if you only care about the sign of the effect
  - Practical advantages of LPM
    * Easier to calculate
    * The parameters are directly interpretable

- Marginal effects in logit/probit
  - Coefficients are NOT marginal effects
  - Marginal effects are non-linear functions of all the estimated parameters and explanatory variables
  - Even if a coefficient is "statistically significant," it doesn't guarantee that the marginal effect associated with that coefficient is "statistically significant."

# 4 Identification

## 4.1 The key problem: OVB

### 4.1.1 Identifying causal effects by eliminating all the omitted variable biases

- OVB mechanism
  - The true model

$$Y = \beta_1 X + \beta_2 Z + \mu$$

  - The biased model (omitting $Z$)

$$Y = \gamma X + \varepsilon$$

  - Correlation (not causality) of $X$ and $Z$: variation in $Z$ is NOT determined by $X$

$$Z = \alpha X + \phi$$

  - Rewrite the true model

$$\begin{aligned} Y &= \beta_1 X + \beta_2 Z + \mu \\ &= \beta_1 X + \beta_2(\alpha X + \phi) + \mu \\ &= (\beta_1 + \beta_2 * \alpha)X + \mu \end{aligned}$$

- Biased estimate [$r$] = True estimate [$\beta_1$] + bias [$\beta_2 * \alpha$]
  - bias $< 0$, [$r$] $<$ [$\beta_1$], underestimate
  - bias $> 0$, [$r$] $>$ [$\beta_1$], overestimate
- How to address OVB?
  - Observables (in the data): include in the regression
  - Unobservables but within unit/time: include fixed effects in the regression
  - Unobservables but can be addressed by (quasi-)experimental designs
  - Remaining unobservables (e.g., unit-time varying): Only God can help!

### 4.1.2 Over-controlling

- Over-controlling mechanism

- The biased model (incorrectly including $Z$)

$$Y = \beta_1 X + \beta_2 Z + \mu$$

- The true model

$$Y = \gamma X + \varepsilon$$

- Causality of $X$ and $Z$: variation in $Z$ is determined by $X$

$$Z = \alpha X + \phi$$

- Rewrite the true model

$$\begin{aligned} Y &= \gamma X + \varepsilon \\ &= (\beta_1 + \beta_2 * \alpha) X + \mu \end{aligned}$$

- Biased estimate [$\beta_1$] = True estimate [$\gamma$] - bias [$\beta_2 * \alpha$]
  - bias $< 0$, [$\beta_1$] $>$ [$r$] , overestimate
  - bias $> 0$, [$\beta_1$] $<$ [$r$] , underestimate
- If Z variable is one mechanism factor on the causal path between X and Y, we can't control for Z in the regression
  - For example, when estimating the effects of charter school on student achievement, we can't control for school spending if we believe spending is one factor how charter schools operate differently from traditional public schools

## 4.2 Experimental design: RCT

- See a recent summary in Athey and Imbens (2017a)

## 4.3 Quasi-experimental designs

### 4.3.1 Difference-in-differences

- Similar to the interaction term in the interaction term regression

$$Y = \beta_0 + \beta_1 Policy + \beta_2 Post + \beta_3 Policy * Post + \varepsilon$$

- Estimating the policy effect

- You can not simply compare $Y\{Policy = 1, Post = 1\}$ with $Y\{Policy = 0, Post = 1\}$, because the two groups may be very different
  - Panel fixed effects help
- You can not simply $Y\{Policy = 1, Post = 1\}$ with $Y\{Policy = 1, Post = 0\}$, because there may be many things changing between the two periods
  - Time fixed effects help

$$
\begin{aligned}
DID &= (Y\{Policy = 1, Post = 1\} - Y\{Policy = 1, Post = 0\}) \\
&\quad - (Y\{Policy = 0, Post = 1\} - Y\{Policy = 0, Post = 0\}) \\
&= (\beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_1)) - (\beta_0 + \beta_2 - (\beta_0)) \\
&= (\beta_2 + \beta_3) - \beta_2 \\
&= \beta_3
\end{aligned}
$$

- **Key assumption**: common trends between the two groups
  - After controlling for panel/time fixed effects, the remaining potential OVB is panel-time varying variables
  - In that case, we don't know $\beta_3$ is **solely** from the policy or from other omitted panel-time varying variables
    - Then the control group can't be a counterfactual of the treatment group
  - Common trends in the pre-policy periods between the two groups are often used to validate the no-OVB claim
- Bias when using a wrong control group
  - $bias = Slope_{Policy} - Slope_{Control}$ in pre-policy periods
    - What is the slope?
    - Graph helps you.
  - If $bias = 0$, common trends!

$$
\begin{aligned}
DID &= (Y\{Policy = 1, Post = 1\} - Y\{Policy = 1, Post = 0\}) \\
&\quad + bias\ term\ in\ the\ pre - policy\ periods \\
&\quad - (Y\{Policy = 0, Post = 1\} - Y\{Policy = 0, Post = 0\}) \\
&= (\beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_1)) + bias - (\beta_0 + \beta_2 - (\beta_0)) \\
&= (\beta_2 + \beta_3) + bias - \beta_2 \\
&= \beta_3 + bias
\end{aligned}
$$

### 4.3.2 CITS

- DID is a simplification of CITS; CITS is a more rigorous design in theory
  - DID evaluates the impact of a program by looking at whether the treatment group deviates from its baseline **mean** by a greater amount than the comparison group (common trends)
  - CITS evaluates program impacts by looking at whether the treatment group deviates from its baseline **trend** by a greater amount than the comparison group
- If no common trends, we can't use DID, but may use CITS
  - Relative "Year" = 0 in 2009; =1 in 2011; =2 in 2013; =3 in 2015; =4 in 2017

$$\begin{aligned} Y = \beta_0 &+ \beta_1 Policy + \beta_2 Post + \beta_3 Policy * Post \\ &+ \beta_4 Year + \beta_5 Policy * Year + \beta_6 Post * Year \\ &+ \beta_7 Policy * Post * Year \\ &+ \varepsilon \end{aligned}$$

  - Estimated policy effects

$$\Delta\hat{Y} = \beta_3 + \beta_7 Policy * Post * Year$$

***Example***

*As you probably know, the City of Flint (in Genesee County) switched to a polluted/contaminated water source in April 2014. Since then, the Michigan Department of Health and Human Services has reported an increase of Legionnaires' disease (LD) in Genesee County. The table below shows the number of LD cases reported in 3 Michigan counties in 2009, 2011, 2013, and 2015 (FAKE DATA). Estimate the policy effects using CITS.*

| | Number of Legionnaires Disease (LD) cases in 2 counties | |
|---|---|---|
| | Genesee (Treatment) | Livingston (Control) |
| 2009 (before) | 38 | 48 |
| 2011 (before) | 40 | 49 |
| 2013 (before) | 42 | 50 |
| 2015 (after) | 50 | 53 |
| 2017 (after) | 55 | 55 |

***Solution***

1. Pre-policy trends

   - Treatment: $(42-38)/(2-0)=2$
   - Control: $(50-48)/(2-0)=1$

2. We can't use DID due to non-common trends between the two groups; we then use CITS

3. For CITS graph (graph is very helpful), we identify 8 data points/equations

   (a) Slope (coefficients on "Year") for treatment in pre-policy years

   $$\begin{aligned} Slope &= \beta_4 + \beta_5 * Policy(=1) + \beta_6 * Post(=0) + \beta_7 * Policy(=1) * Post(=0) \\ &= \beta_4 + \beta_5 \\ &= (42-38)/2 = 2 \end{aligned}$$

   (b) Slope (coefficients on "Year") for control in pre-policy years

   $$\begin{aligned} Slope &= \beta_4 + \beta_5 * Policy(=0) + \beta_6 * Post(=0) + \beta_7 * Policy(=0) * Post(=0) \\ &= \beta_4 \\ &= (50-48)/2 = 1 \end{aligned}$$

   (c) Slope (coefficients on "Year") for treatment in post-policy years

   $$\begin{aligned} Slope &= \beta_4 + \beta_5 * Policy(=1) + \beta_6 * Post(=1) + \beta_7 * Policy(=1) * Post(=1) \\ &= \beta_4 + \beta_5 + \beta_6 + \beta_7 \\ &= (55-50)/1 = 5 \end{aligned}$$

   (d) Slope (coefficients on "Year") for control in post-policy years

   $$\begin{aligned} Slope &= \beta_4 + \beta_5 * Policy(=0) + \beta_6 * Post(=1) + \beta_7 * Policy(=0) * Post(=1) \\ &= \beta_4 + \beta_6 \\ &= (55-53)/1 = 2 \end{aligned}$$

   (e) Intercept for treatment in year=0 (first year in pre-policy)

   $$\begin{aligned} Intercept &= \beta_0 + \beta_1 * Policy(=1) + \beta_2 * Post(=0) + \beta_3 * Policy(=1) * Post(=0) \\ &= \beta_0 + \beta_1 \\ &= 38 \end{aligned}$$

(f) Intercept for control in year=0 (first year in pre-policy)

$$\begin{aligned}
Intercept &= \beta_0 + \beta_1 * Policy(=0) + \beta_2 * Post(=0) + \beta_3 * Policy(=1) * Post(=0) \\
&= \beta_0 \\
&= 48
\end{aligned}$$

(g) Intercept for treatment in year=2015 (first year in post-policy)

$$\begin{aligned}
Intercept &= \beta_0 + \beta_1 * Policy(=1) + \beta_2 * Post(=1) + \beta_3 * Policy(=1) * Post(=1) \\
&+ \beta_4 * 3 + \beta_5 * Policy(=1) * 3 \\
&+ \beta_6 * Post(=0) * 3 \\
&+ \beta_7 * Policy(=1) * Post(=0) * 3 \\
&= \beta_0 + \beta_1 + \beta_2 + \beta_3 + (\beta_4 + \beta_5) * 3 \\
&= 50
\end{aligned}$$

(h) Intercept for control in year=2015 (first year in post-policy)

$$\begin{aligned}
Intercept &= \beta_0 + \beta_1 * Policy(=0) + \beta_2 * Post(=1) + \beta_3 * Policy(=1) * Post(=0) \\
&+ \beta_4 * 3 + \beta_5 * Policy(=0) * 3 \\
&+ \beta_6 * Post(=0) * 3 \\
&+ \beta_7 * Policy(=0) * Post(=0) * 3 \\
&= \beta_0 + \beta_2 + (\beta_4) * 3 \\
&= 53
\end{aligned}$$

4. Solve the above 8 equations

- $\beta_7 = (5-2) - (2-1) = 2$

$$\begin{aligned}
&\beta_4 + \beta_5 &&=2 \\
&\beta_4 &&=1 \\
&\beta_4 + \beta_5 + \beta_6 + \beta_7 &&=5 \\
&\beta_4 + \beta_6 &&=2
\end{aligned}$$

  – Short solution:

$$\begin{aligned}
beta_7 = &(Slope_{treatment,post} - Slope_{treatment,pre}) \\
&- (Slope_{control,post} - Slope_{control,pre})
\end{aligned}$$

- $\beta_3 = (50 - 38 - 3 * 2) - (53 - 48 - 3 * 1) = 4$

$$
\begin{aligned}
\beta_0 + \beta_1 &= 38 \\
\beta_0 &= 48 \\
\beta_0 + \beta_1 + \beta_2 + \beta_3 + (\beta_4 + \beta_5) * 3 &= 50 \\
\beta_0 + \beta_2 + (\beta_4) * 3 &= 52
\end{aligned}
$$

– Short solution:

$$
\begin{aligned}
beta_3 = &(Y_{treatment,post} - Y_{treatment,pre} + Slope_{treatment,pre} \times Year) \\
&- (Y_{control,post} - Y_{control,pre} + Slope_{control,pre} \times Year)
\end{aligned}
$$

5. Evaluate the effects over years: $\beta_3 + \beta_7 * Post\_Year$
   - In 2015: $4 + 2 * 0 = 5$
   - In 2017: $4 + 2 * 1 = 6$

#### 4.3.3 RD

- Regression Discontinuity is a highly rigorous method for evaluating social programs
- Candidates are selected for treatment (or not) based on whether their "score" on a numeric rating exceeds a designated threshold or cut-point
  – Candidates scoring above or below a certain threshold are selected for inclusion in the treatment group
  – Candidates on the other side of the threshold constitute a comparison group
- **Assumption**: candidates are barely below or above the cutoff are similar in both observables and unobservables
  – In a RCT, assignment is based on a "coin flip" In an RD design, assignment is based on whether individuals are above or below a known cut-off on a measurable criterion
- Steps
  – Collecting data on (1) the running variable X, (2) the cutoff score, (3) the treatment group, (4) outcomes, and (5) covariates
  – *First stage:* Show (and estimate) discontinuity in treatment based on X
  – *Continuity assumption:* Show (and estimate) continuity in covariates based on X
  – *Second stage:* Show (and estimate) change in outcome based on X

#### 4.3.4 Other topics: panel data, event study, IV

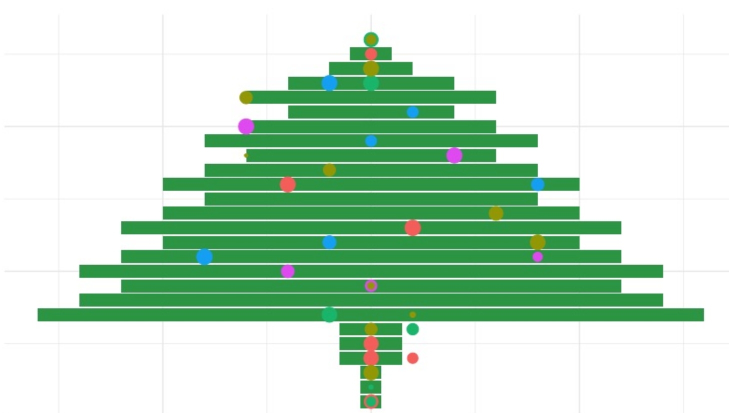
- Learn by doing in your future career
- Help is always available

### 4.4 Applying research to real world policy: External validity

- Threats to external validity compromise our confidence in stating whether the study's results are applicable to other groups
  - Think about whether we care about the direction of the relationship or the specific magnitude
- Population validity: Representative sample? Generalizable?
- Ecological validity: Interactive effect? Reactive effect?

## Thank you and happy new new 2018!

- Keep in touch ☺
  - Email, Facebook, Twitter
  - Website: www-personal.umich.edu~yxy/



Xiaoyang wishes you a Happy New Year 2018

## References


Abbring, Jaap H and James J Heckman. 2007. "Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation." *Handbook of econometrics* 6:5145–5303.

Angrist, Joshua D and Victor Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics* 114 (2):533–575.

Angrist, Joshua D, Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2017. "Maimonides Rule Redux." Tech. rep., National Bureau of Economic Research.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

———. 2014. *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Athey, Susan and Guido W Imbens. 2017a. "The econometrics of randomized experiments." *Handbook of Economic Field Experiments* 1:73–140.

———. 2017b. "The state of applied econometrics: Causality and policy evaluation." *The Journal of Economic Perspectives* 31 (2):3–32.

Banerjee, Abhijit Vinayak and Esther Duflo. 2017. "An Introduction to the "Handbook of Field Experiments"." *Handbook of Economic Field Experiments* 1:1–24.

Battistin, E, JD Angrist, and Daniela Vuri. 2016. "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno." *American Economic Journal: Applied Economics* .

Bertrand, Marianne and Esther Duflo. 2017. "Field experiments on discrimination." *Handbook of Economic Field Experiments* 1:309–393.

Bettinger, Eric, Christopher Doss, Susanna Loeb, Aaron Rogers, and Eric Taylor. 2017. "The effects of class size in online college courses: Experimental evidence." *Economics of Education Review* 58:68–85.

Cameron, Adrian Colin and Pravin K Trivedi. 2009. *Microeconometrics using Stata*, vol. 5. Stata press College Station, TX.

Card, David and Alan B Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *The Journal of Political Economy* 100 (1):1–40.

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil. 2011. "Estimating marginal returns to education." *The American economic review* 101 (6):2754–2781.

Chay, Kenneth Y and Michael Greenstone. 2003. "The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession." *The Quarterly Journal of Economics* 118 (3):1121–1167.

———. 2005. "Does air quality matter? Evidence from the housing market." *Journal of Political Economy* 113 (2):376–424.

Chetty, R, JN Friedman, N Hilger, E Saez, DW Schanzenbach, and D Yagan. 2010. "How does your kindergarten classroom affect your earnings? Evidence from Project Star." *The Quarterly Journal of Economics* 126 (4):1593–1660.

Currie, Janet and Reed Walker. 2011. "Traffic congestion and infant health: Evidence from E-ZPass." *American Economic Journal: Applied Economics* 3 (1):65–90.

DiNardo, John and David S Lee. 2011. "Program evaluation and research designs." *Handbook of labor economics* 4:463–536.

Dynarski, Susan, Joshua Hyman, and Diane Whitmore Schanzenbach. 2013. "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion." *Journal of Policy Analysis and Management* 32 (4):692–717.

Fisher, Ronald Aylmer. 1925. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2013. "Long-Term Effects of Class Size." *The Quarterly Journal of Economics* 128 (1):249–285.

Greenberg, Edward. 2012. *Introduction to Bayesian Econometrics*. Cambridge University Press.

Greene, William. 2004. "The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects." *The Econometrics Journal* 7 (1):98–119.

Greenstone, Michael and Justin Gallagher. 2008. "Does hazardous waste matter? Evidence from the housing market and the superfund program." *The Quarterly Journal of Economics* 123 (3):951–1003.

Hayashi, Fumio. 2000. *Econometrics*. Princeton University Press.

Heckman, James J and Edward J Vytlacil. 2007a. "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation." *Handbook of econometrics* 6:4779–4874.

———. 2007b. "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments." *Handbook of econometrics* 6:4875–5143.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81 (396):945–960.

Hoxby, Caroline M. 2000. "The effects of class size on student achievement: New evidence from population variation." *The Quarterly Journal of Economics* 115 (4):1239–1285.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Isen, Adam, Maya Rossin-Slater, and W Reed Walker. 2017. "Every breath you take - Every dollar you'll make: The long-term consequences of the Clean Air Act of 1970." *Journal of Political Economy* 125 (3):848–902.

Jackson, C Kirabo, Rucker C Johnson, and Claudia Persico. 2015. “The effects of school spending on educational and economic outcomes: Evidence from school finance reforms.” *The Quarterly Journal of Economics* 131 (1):157–218.

Krueger, Alan B and Diane M Whitmore. 2001. “The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR.” *The Economic Journal* 111 (468):1–28.

Manski, Charles F. 2009. *Identification for prediction and decision*. Harvard University Press.

Murnane, Richard J and John B Willett. 2010. *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Poirier, Dale J. 1995. *Intermediate statistics and econometrics: A Comparative approach*. Mit Press.

Sullivan, Daniel M. 2017. “The True Cost of Air Pollution: Evidence from the Housing Market.” Tech. rep., Working Paper.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. MIT press.