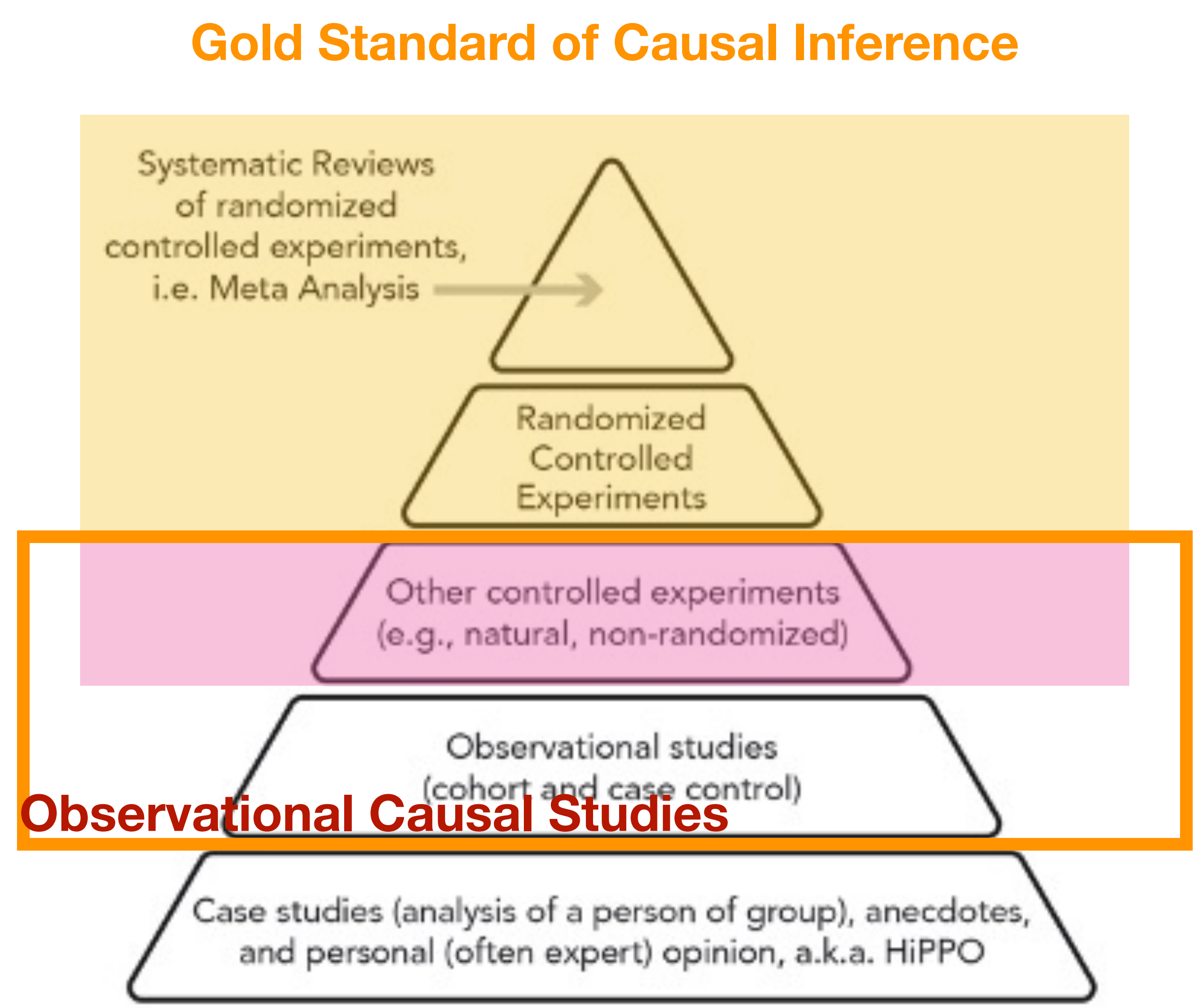
Digital Experimentation Methods

Session 9: Observational Causal Studies

Shan Huang, HKU

Hierarchy of Evidence

- Random Assignments of Variants (Treatments)
    - Randomized Controlled Experiments
    - Multiple randomized controlled experiments
        - Fisher’s Meta-Analysis
- As you go down, the trust level declines
- Other Controlled Experiment : Treatment without Random Assignment
- Observational Studies: No Treatment
- Case Studies: Subjective Evidence



Observational Causal Studies

- More complex in data analysis with low trust in causal effects.
- Harder to scale it up in companies, compared to A/B testing.
- Companies started to invest in observational causal studies about 3 (China) -5(US) years ago.
- Randomized controlled experiments started about 5 (China) - 10 (US) years ago.
- Data-driven (informed) decision makings are the future.
- Only causal effects can inform the choices among different strategies.

When Controlled Experiments Are Not Possible

A. What are the treatment and control groups?

B. Is there a random assignment of the treatment? Why?

- What is the impact of COVID-19 on users' social behavior on Facebook/WeChat?
- What is the impact on product engagement if a user switches their phone from an iPhone to an Android?
- What is the impact of Apple's policy change on WeChat user behaviors?
- What is the impact of Tiktok's new features on Kuaishou's users' behaviors?

When Controlled Experiments Are Not Possible

- When the change to be tested is not under the control of organizations.
  - Third party's decision
  - Competitors' decision
  - Users' decisions
  - Natural disasters
- When establishing a *Control* may incur too large an opportunity cost
  - Experiments can be costly during the rare event
    - A new feature for red pockets during the spring festival
    - Running ads during Super Bowl
    - Measure the long-term treatment effects

Quasi-experiments

- The goal is to measure the causal impact of a change (treatment).
- Compare the outcome of a treated population (treatment) to the outcome for an untreated population (control).

Outcome for treated - Outcome for untreated

= [Outcome for treated - Outcome for treated if not treated] +
[Outcome for treated if not treated - Outcome for untreated]

= Treatment Effects on treated + Selection bias

To minimize it. It's almost impossible to completely remove it without randomization

Quasi-experiments

Outcome for treated - Outcome for untreated

= [Outcome for treated - Outcome for treated if not treated] + [Outcome for treated if not treated - Outcome for untreated]

= Treatment Effects on treated + Selection bias

Minimize the difference between Control and Treatment

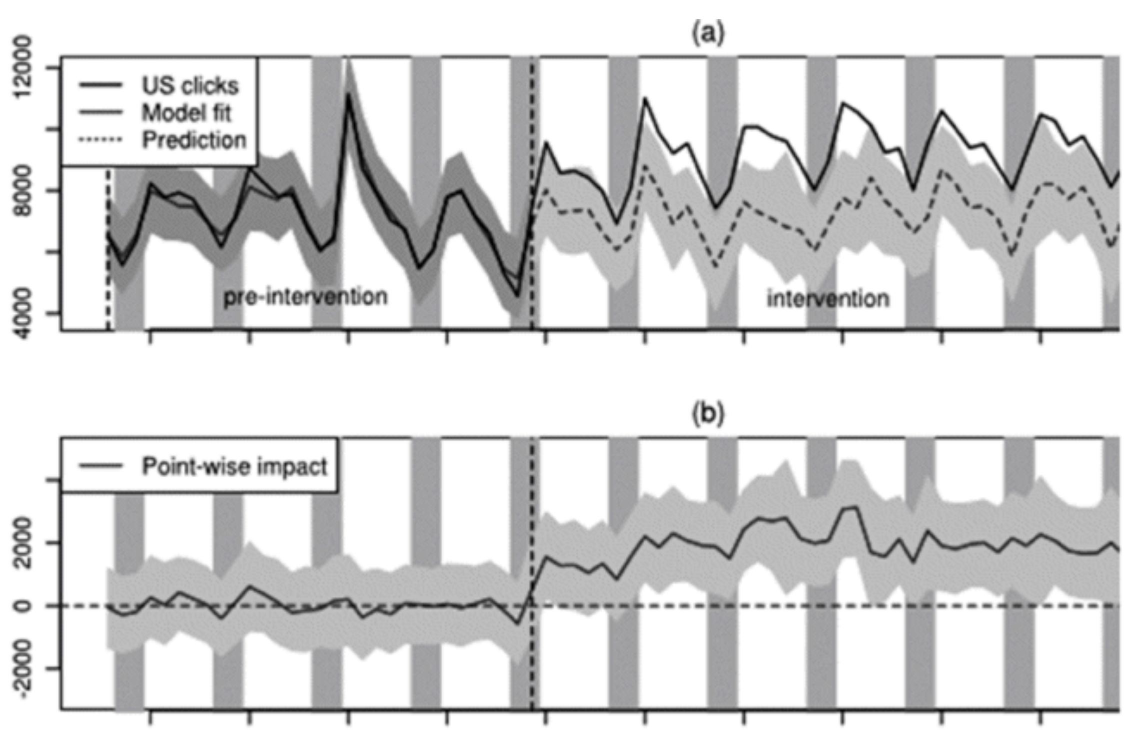
- Challenges are:
  - How to construct Control and Treatment Groups
  - How to model the impact given those Control and Treatment Groups

Interrupted Time Series (ITS)

- A Quasi-Experiment Design: Treatment without a random assignment
- Use the same population for Control and Treatment
- Vary what the population experiences over time.
    - e.g., treatment is the big change to products and cannot be controlled

Y

Δ



[Bayesian structural time series]

1. **Use the data before the Treatment to train the model for prediction.**
2. **- - Counterfactual Y after launching the new feature: Model Predictions**

Interrupted Time Series (ITS)

- The prediction model considers only the information before the treatment
- Confounding Factor
    - The factor unique to post-treatment periods
        - Time effects
        - new changes after launching new features
- How could we improve the design?
    - Switch on & off of the treatment multiple times
        - Average out the confounding effects
    - What are the risks?
        - Hurt user experience

Regression Discontinuity Design

- A methodology to identify the Comparable Treatment and Control Groups by a clear threshold.
- Treatment: Just above the threshold
- Control: Just below the threshold
- Example: Study the effects of university education on Income
  - University admission line is 570
  - Treatment: Just above 570 e.g., [570,575]
  - Control: Just below [565,570)
  - Among the people [565,575], passing the line is likely a random assignment.

Almost the Same expect for taking university

Regression Discontinuity Design

- Goal: Assess the impact of drinking on deaths
- Facts: Americans over 21 can drink legally
- What is the RDD design to answer this question? (Threshold)
  - Threshold: 21 years old
  - Compare the death rate among those just below and above 21 years old.
- What can be the confounders?
  - Other factors that share the same threshold
  - e.g., the legal age of 21 is also for legal gambling

Regression Discontinuity Design

- Carpenter, Christopher, and Carlos Dobkin. "The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age." *American Economic Journal: Applied Economics* 1, no. 1 (2009): 164-82.

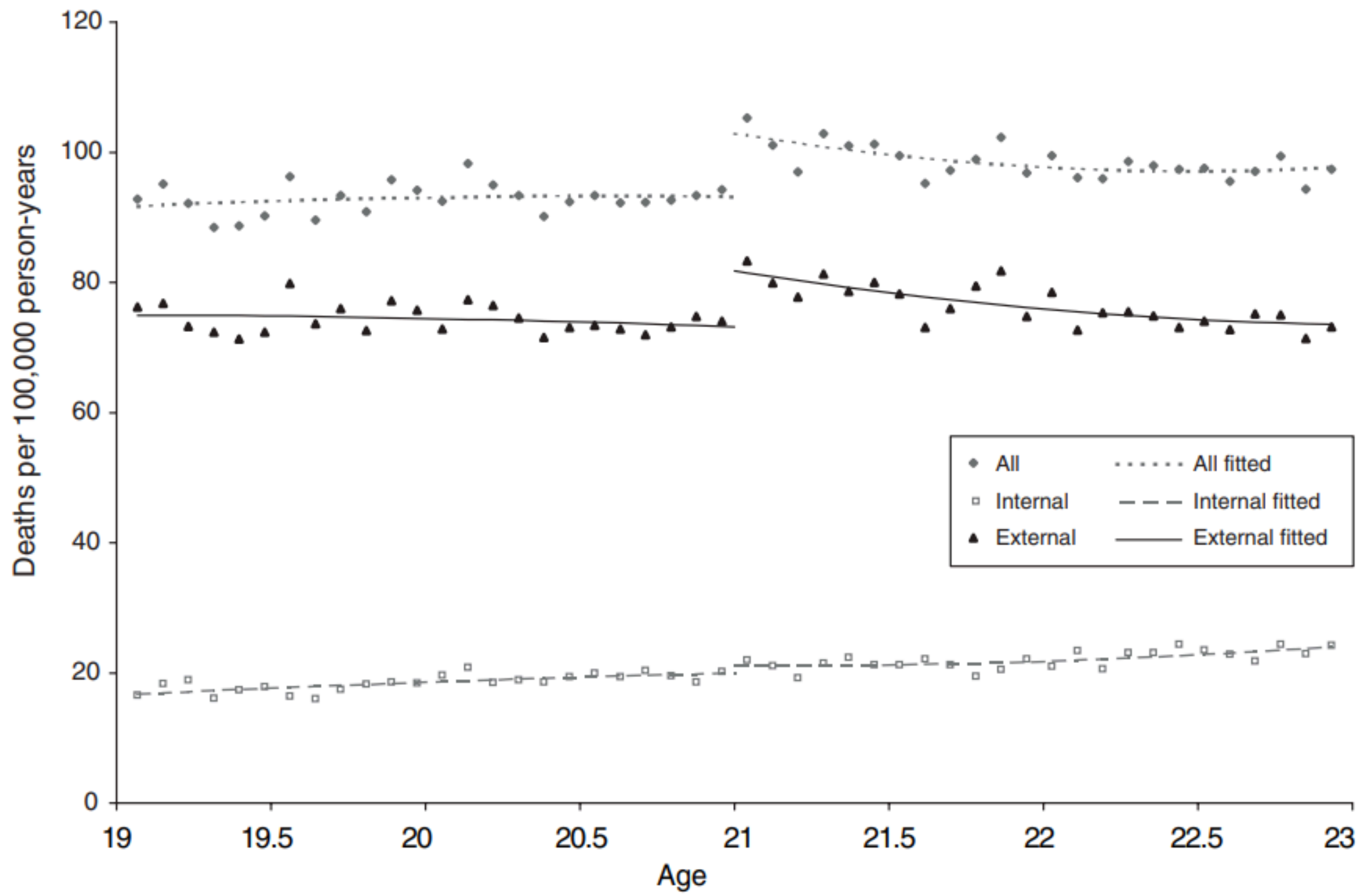


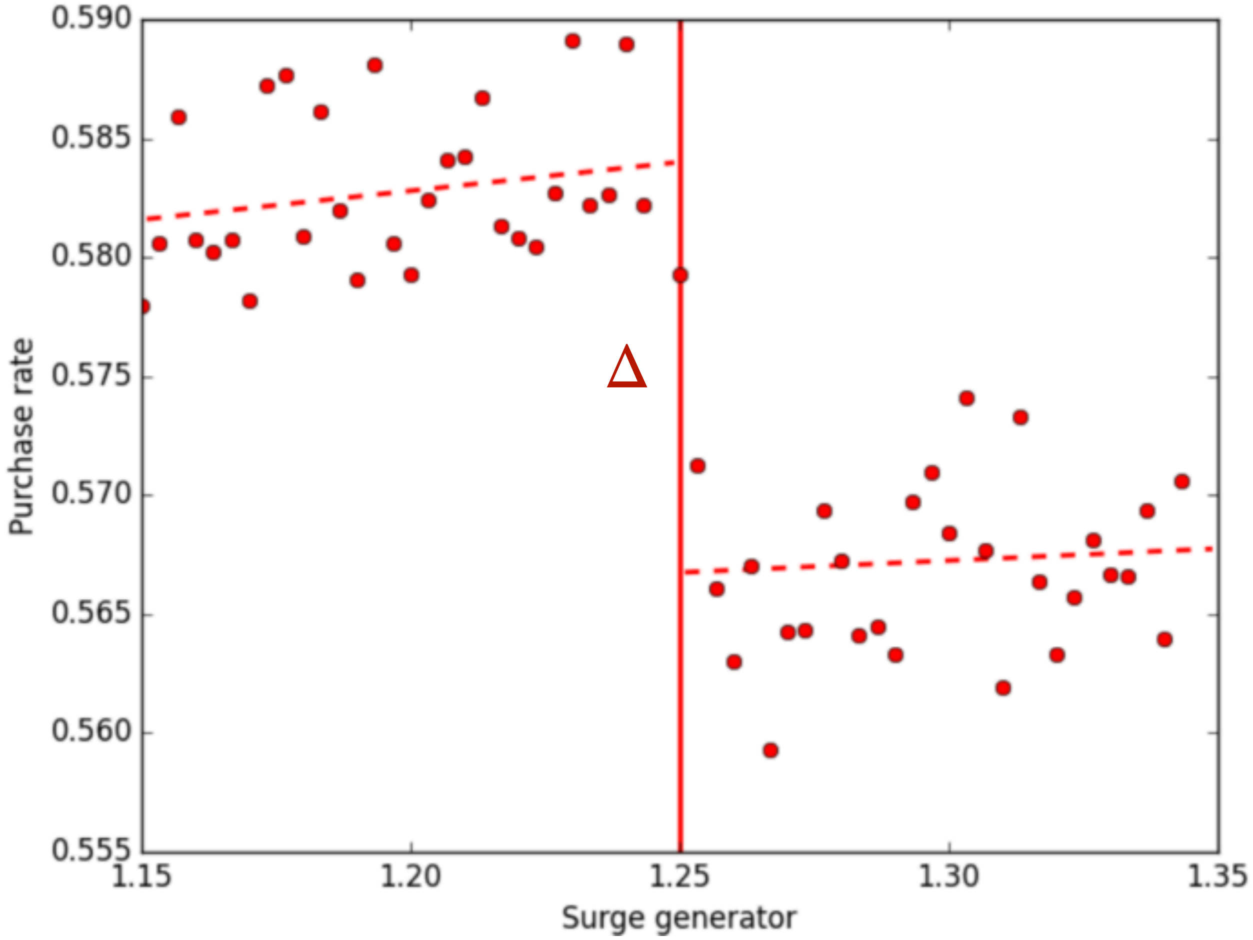
FIGURE 3. AGE PROFILE FOR DEATH RATES

Notes: Deaths from the National Vital Statistics Records. Includes all deaths that occurred in the United States between 1997–2003. The population denominators are derived from the census. See online Appendix C for a list of causes of death.

Regression Discontinuity Design @ Uber

Effects of Surge Pricing on Demand of Uber (purchase rate)

- Sharp Cutoff: surge generator = 1.25
- Assumption: very close to the cut-off point are similar with respect to any relevant confounding variables.
- What is the RDD design?
    - Treatment: Just Above 1.25 and with surge price
    - Control: Just Below 1.25 and without surge price



Regression Discontinuity Design

- A sharp cutoff, e.g.,
  - eBay sellers' score > T can get a badge.
    - Identify the signalling effects of badge on sales
  - Taobao sellers' sales > T are provided with a new tool
    - Identify the effects of new tool on sales
- Assumptions:
  - Users just above and below the cutoff are of no differences on Y except for receiving the treatment or not.
  - "Almost Random" Assignment is based on the cutoff

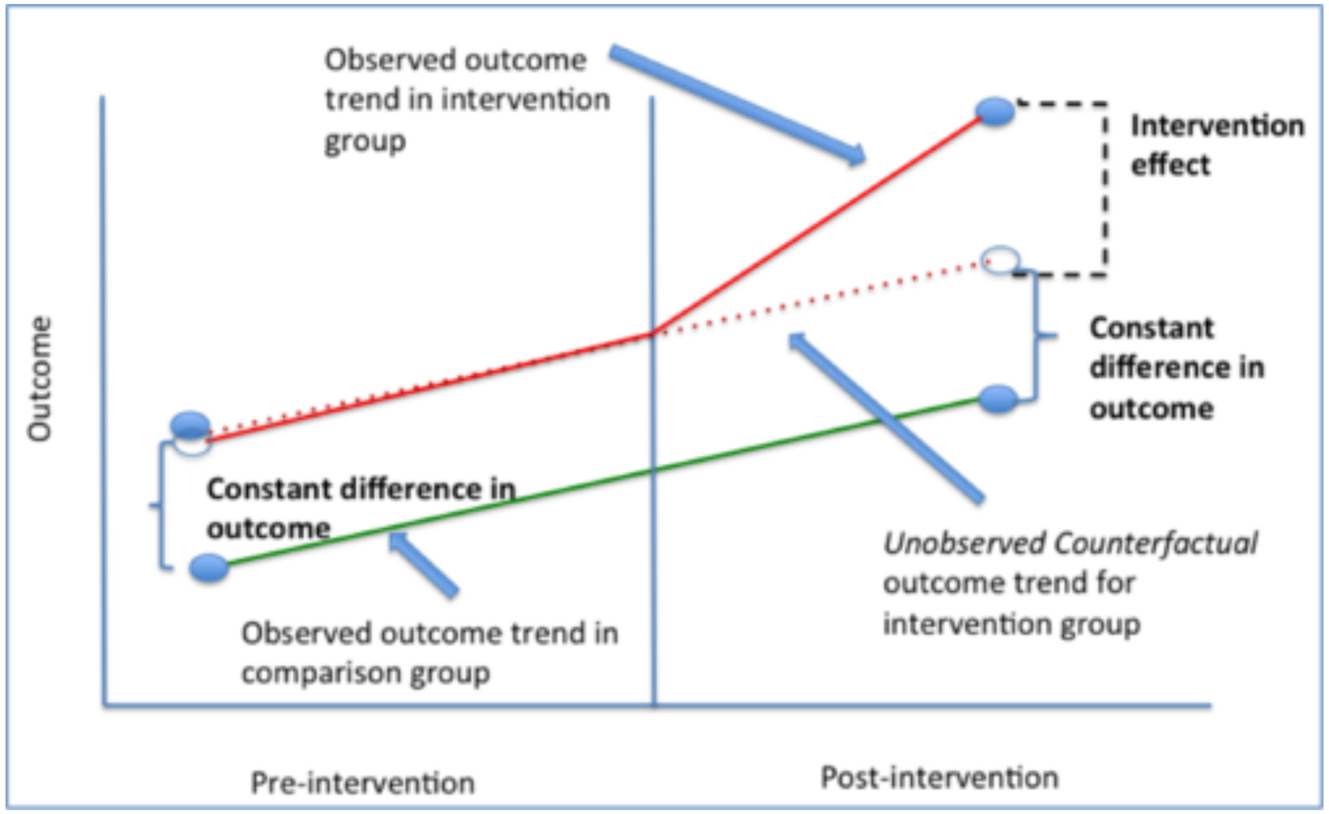
Difference-in-Difference

1. A policy only impacts a subset of users

- e.g., a new feature is first launched on iPhone users
    - However, iPhone users are systematically different from Android users.
    - Find a group of Android users as Control
        - **Check whether the trends on Y (metrics) before the Treatment are the same or not for the users.**

Identify the Control Group

- Use the control group to control for the confounding effects
- Assume the same trends without the treatment after the treatment



Example: DiD @ Seeking Alpha

Chen, Hailiang, Yu Jeffrey Hu, and Shan Huang. "Monetary incentive and stock opinions on social media." *Journal of Management Information Systems* 36.2 (2019): 391-417.

- Seeking Alpha is one of the biggest investment-related social media websites in the U.S.
- In January 2011, SA launched a premium partnership program that enables its contributors to earn $10 per 1,000 page views received by their "premium" articles
  - What is the treatment here?
  - Monetary Incentive for content contribution
- We used DID approach to examine the impact of this policy change.

Example: DiD @ Seeking Alpha

- Treatment: Users who participated in the program
- Control: Users who did not participate in the program
- Assumptions:
  - Control Group's users were not affected by the policy change
  - Control Group's trend on Y is very similar to that of Treatment Group before the policy change

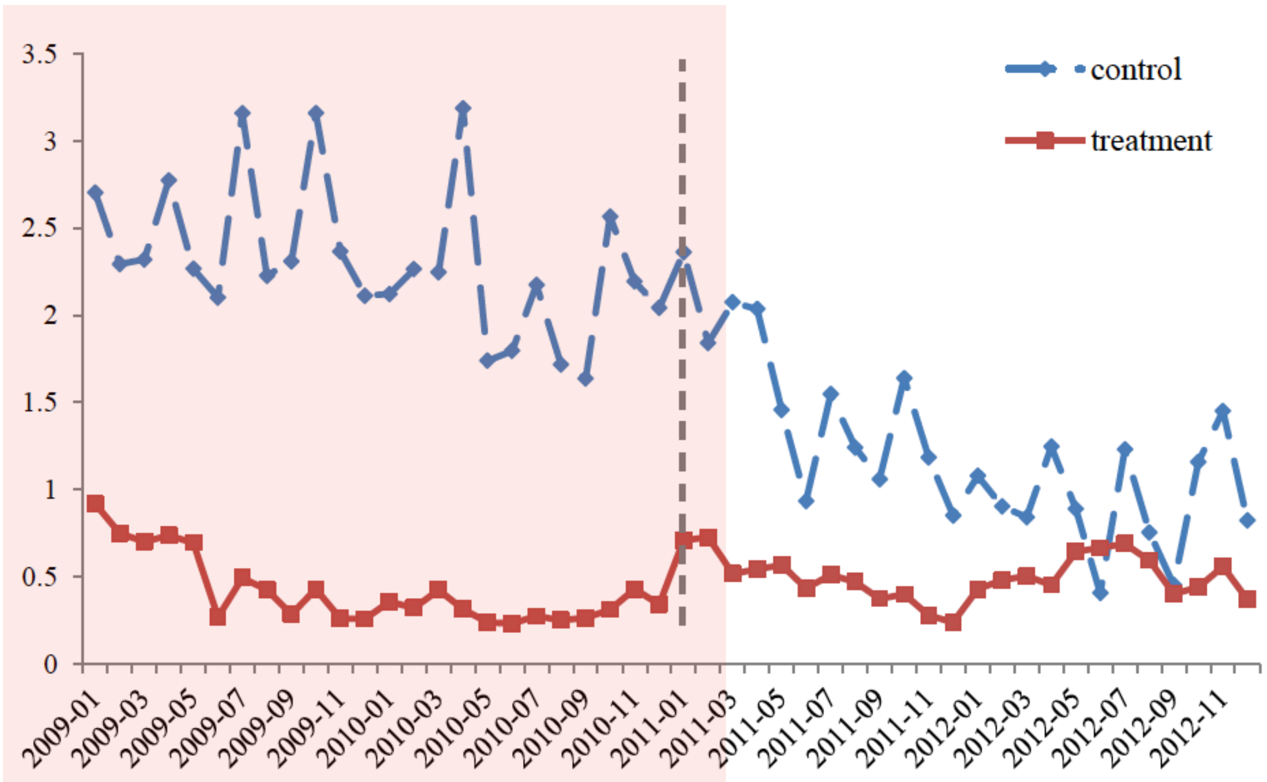


Figure 1. Average number of articles per contributor in each month

Propensity Score Matching

- Construct two comparable groups of units based on observed characteristics.
- Comparable in the sense that:
  - They share the variables that can impact the metrics
  - e.g., OEC is user engagement (# visits /week)
    - Control and Treatment Groups are (almost) the same on the variables that can affect engagement
- Propensity Score Matching (PSM) provides a way to construct two comparable groups.

Propensity Score Matching

- Instead of matching on covariates directly, PSM matches on a single number: the propensity score
  - $p_i = pr(T_i | X_i) = \dfrac{exp[\beta_0 + \beta_1 X + \epsilon_i]}{1 + exp[\beta_0 + \beta_1 X + \epsilon_i]}$
  - $|p_i - p_j| < \sigma$ (a small number)
- (i, j) are (almost) equally likely to be treated but happen to be in control and treatment groups.
  - Users i and j are equally likely to adopt a new feature.
  - User i happens to adopt it, while user j happens not to.
- Find many such pairs and construct Control (j) and Treatment (i) Groups.

Propensity Score Matching

- A useful methodology to construct comparable/matched groups: Treated units vs. Untreated units
  - Android vs. iPhone users
  - Comparable cities
- PSM is one of the most popular matching methods used in the industry.
- Other popular matching methods:
  - Synthetic Control
  - Coarsened Exact Matching (CEM)
- Can we combine PSM and DiD? How?
  - Use PSM to construct the groups with similar trends before the treatment