Corporate Finance

Lecture 2: Review of Econometrics

Tong Li (李通)

School of Economics, Xiamen University
Wang Yanan Institute for Studies in Economics (WISE), Xiamen University

Email: litong@xmu.edu.cn

Announcements

- 1st News Summary will be due today.
- Requirements for the 1st presentation have been uploaded.

Outline for This Lecture

1. Econometrics and Corporate Finance
2. OLS Regressions
3. Correlation vs. Causality
4. Endogeneity and Solutions

What is Econometrics?

- Application of statistical methods to economic data
- Essential for analyzing financial data and testing theories
- Bridges the gap between theory and real-world data

Types of Economic Data in Econometric Analysis

1. **Cross-Sectional Data** (截面数据)
   - Data collected at a single point in time from multiple subjects (e.g., firms, individuals, countries)
   - Example: sales data from 100 firms in 2023

Types of Economic Data in Econometric Analysis

1. **Cross-Sectional Data** (截面数据)
   - Data collected at a single point in time from multiple subjects (e.g., firms, individuals, countries)
   - Example: sales data from 100 firms in 2023

2. **Time Series Data** (时间序列数据)
   - Data collected on a single subject over multiple time periods
   - Example: Annual sales of a firm from 2011 to 2020

Types of Economic Data in Econometric Analysis

1. **Cross-Sectional Data** (截面数据)
   - Data collected at a single point in time from multiple subjects (e.g., firms, individuals, countries)
   - Example: sales data from 100 firms in 2023

2. **Time Series Data** (时间序列数据)
   - Data collected on a single subject over multiple time periods
   - Example: Annual sales of a firm from 2011 to 2020

3. **Panel Data** (面板数据)
   - Data that combines both cross-sectional and time series elements
   - Tracks multiple subjects across time periods
   - Example: Annual sales of 100 firms over 10 years

Types of Economic Data in Econometric Analysis

1. **Cross-Sectional Data** (截面数据)
   - Data collected at a single point in time from multiple subjects (e.g., firms, individuals, countries)
   - Example: sales data from 100 firms in 2023
2. **Time Series Data** (时间序列数据)
   - Data collected on a single subject over multiple time periods
   - Example: Annual sales of a firm from 2011 to 2020
3. **Panel Data** (面板数据)
   - Data that combines both cross-sectional and time series elements
   - Tracks multiple subjects across time periods
   - Example: Annual sales of 100 firms over 10 years
4. **Pooled Cross-Sectional Data** (混合截面数据)
   - Cross-sectional data collected at different time points, but not necessarily following the same subjects
   - Example: Household income surveys from 2010 and 2020, but with different households

Importance of Econometrics in Corporate Finance

- Common Questions:
  - How does corporate capital structure affect firm value?
  - What is the impact of investment decisions on performance?
- Need for Empirical Evidence:
  - Validate theoretical models
  - Support decision-making

Ordinary Least Squares (OLS) Regression

- Basic Form of OLS:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

  - y: Dependent variable
  - x: Independent variables
  - ϵ: Error term

Ordinary Least Squares (OLS) Regression

- Basic Form of OLS:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

  - y: Dependent variable
  - x: Independent variables
  - ϵ: Error term

- Key assumptions of OLS:
  - **Linearity**: Relationship between x and y is linear
  - **Exogeneity**: x is uncorrelated with the error term ϵ
  - **No Omitted Variable Bias**
  - **No Perfect Multicollinearity**: Independent variables are not perfectly correlated
  - **Homoscedasticity**: Constant variance of errors

OLS in Corporate Finance: One Example (Gulen and Ion, RFS 2016)

- Background
  - Businesses often face a significant amount of uncertainty regarding the timing, content, and potential impact of policy decisions.
- Research Question:
  - Does policy-related uncertainty affect corporate investment decisions?

OLS in Corporate Finance: One Example (Gulen and Ion, RFS 2016)

- Background
  - Businesses often face a significant amount of uncertainty regarding the timing, content, and potential impact of policy decisions.
- Research Question:
  - Does policy-related uncertainty affect corporate investment decisions?
- OLS Regression Equation:

$$\frac{CAPX_{i,t+n}}{TA_{i,t+n-1}} = \alpha_i + \beta_1 PU_{i,t} + \beta_2 TQ_{i,t} + \beta_3 \frac{CF_{i,t}}{TA_{i,t-1}} + \beta_4 SG_{i,t} + \delta M_t + QRT_t + \epsilon_{i,t+n}$$

  - $\frac{CAPX_{i,t+n}}{TA_{i,t+n-1}}$: capital expenditure (scaled by total assets) of firm i in year $t+n$;
  - $PU_{i,t}$: policy uncertainty faced by firm i in year t;
  - $TQ_{i,t}$, $\frac{CF_{i,t}}{TA_{i,t-1}}$, and $SG_{i,t}$: firm-level control variables;
  - M_t: macroeconomic factors, e.g., GDP growth;
  - α_i: firm fixed effects; QRT_t: quarter fixed effects

OLS in Corporate Finance: One Example (Gulen and Ion, RFS 2016)

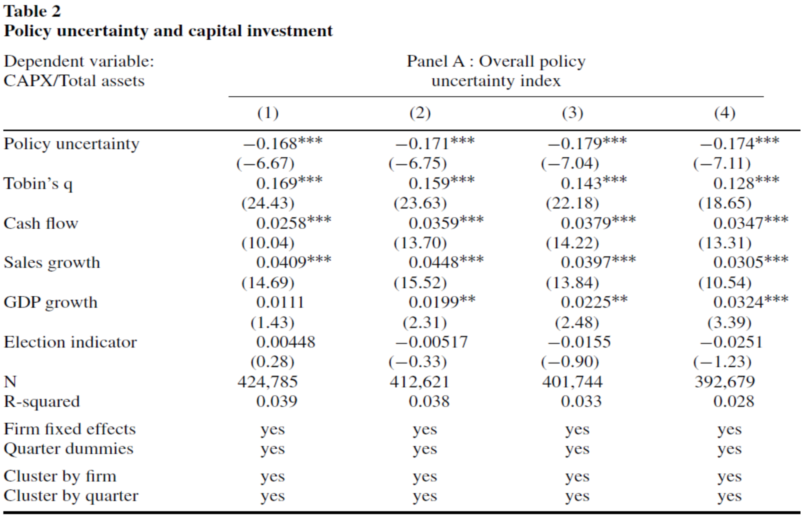


Table 2
Policy uncertainty and capital investment

Dependent variable: CAPX/Total assets	Panel A : Overall policy uncertainty index			
	(1)	(2)	(3)	(4)
Policy uncertainty	−0.168***	−0.171***	−0.179***	−0.174***
	(−6.67)	(−6.75)	(−7.04)	(−7.11)
Tobin's q	0.169***	0.159***	0.143***	0.128***
	(24.43)	(23.63)	(22.18)	(18.65)
Cash flow	0.0258***	0.0359***	0.0379***	0.0347***
	(10.04)	(13.70)	(14.22)	(13.31)
Sales growth	0.0409***	0.0448***	0.0397***	0.0305***
	(14.69)	(15.52)	(13.84)	(10.54)
GDP growth	0.0111	0.0199**	0.0225**	0.0324***
	(1.43)	(2.31)	(2.48)	(3.39)
Election indicator	0.00448	−0.00517	−0.0155	−0.0251
	(0.28)	(−0.33)	(−0.90)	(−1.23)
N	424,785	412,621	401,744	392,679
R-squared	0.039	0.038	0.033	0.028
Firm fixed effects	yes	yes	yes	yes
Quarter dummies	yes	yes	yes	yes
Cluster by firm	yes	yes	yes	yes
Cluster by quarter	yes	yes	yes	yes

Statistical Softwares for Econometric Analysis

- **Stata**
- R
- Python
- SAS
- ...

Correlation vs. Causality

- As researchers, we are interested in making **causal** statements
  - What is the effect of a change in corporate taxes on firms' leverage choices?
  - Does hiring a female CEO lead to better environmental performance of the firm?
- Avoid using "associated" or "correlated" to describe the relation between two variables
- Casual statements are essential for decision making (e.g., policymakers, corporate managers).

What do we mean by causality?

- If our linear model is the following

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

  and we want to infer β_1 as the causal effect of x_1 on y, holding all else equal, then we need to make the following assumptions:

  1. $E(\epsilon) = 0$
  2. $E(\epsilon | x_1, \cdots, x_k) = E(\epsilon)$

- This is “conditional mean independence” (CIM).

- Generally speaking, you need the estimation error to be uncorrelated with all the x.

- CIM is violated whenever an independent variable (x) is correlated with the error term (ϵ).

What is Endogeneity?

- Endogeneity is correlation between the error term (ϵ) and explanatory variables (x) in a regression

$$y = \alpha + \beta x + \epsilon$$

What Causes Endogeneity?

What Causes Endogeneity?

1. **Omitted Variables**: missing factors affecting y

What Causes Endogeneity?

1. **Omitted Variables**: missing factors affecting y

$$wage = \alpha + \beta education + \epsilon$$

   - Hypothesis: $\beta > 0$, More education leads to higher wages.

What Causes Endogeneity?

1. **Omitted Variables**: missing factors affecting y

$$wage = \alpha + \beta education + \epsilon$$

   - Hypothesis: $\beta > 0$, More education leads to higher wages.
   - BUT, people with higher ability (unobserved and omitted) tend to earn more and they likely get more education.

What Causes Endogeneity?

1. **Omitted Variables**: missing factors affecting y

$$wage = \alpha + \beta education + \epsilon$$

   - Hypothesis: $\beta > 0$, More education leads to higher wages.
   - BUT, people with higher ability (unobserved and omitted) tend to earn more and they likely get more education.

2. **Measurement Error**: variables are misspecified or suffer from inaccurate data collection

$$Investment = \alpha + \beta_1 CashFlow + \beta_2 Q + \epsilon$$

   - Hypothesis: $\beta_1 > 0$, more internal funds mitigate financial constraints and lead firms to invest more.

What Causes Endogeneity?

1. **Omitted Variables**: missing factors affecting y

$$wage = \alpha + \beta education + \epsilon$$

   - Hypothesis: $\beta > 0$, More education leads to higher wages.
   - BUT, people with higher ability (unobserved and omitted) tend to earn more and they likely get more education.

2. **Measurement Error**: variables are misspecified or suffer from inaccurate data collection

$$Investment = \alpha + \beta_1 CashFlow + \beta_2 Q + \epsilon$$

   - Hypothesis: $\beta_1 > 0$, more internal funds mitigate financial constraints and lead firms to invest more.
   - BUT, firms with more cash have better investment opportunities (mismeasured by Tobin's Q) and they investment more.

What Causes Endogeneity?

1. **Omitted Variables**: missing factors affecting y

$$wage = \alpha + \beta education + \epsilon$$

   - Hypothesis: $\beta > 0$, More education leads to higher wages.
   - BUT, people with higher ability (unobserved and omitted) tend to earn more and they likely get more education.

2. **Measurement Error**: variables are misspecified or suffer from inaccurate data collection

$$Investment = \alpha + \beta_1 CashFlow + \beta_2 Q + \epsilon$$

   - Hypothesis: $\beta_1 > 0$, more internal funds mitigate financial constraints and lead firms to invest more.
   - BUT, firms with more cash have better investment opportunities (mismeasured by Tobin's Q) and they investment more.

3. **Reverse Causality (Simultaneity)**: y may also affect x

$$Interest\ Rate = \alpha + \beta Loan\ Amount + \epsilon$$

$$Loan\ Amount = \theta + \delta Interest\ Rate + \eta$$

Consequences of Endogeneity

- Statistically speaking, regression parameters cannot be identified.
  - OLS estimates will not be estimates of what we want.
- Practically speaking, it is very hard (impossible) to interpret the results.
  - We cannot rule out alternative explanations for findings.

What to do about Endogeneity?

- Do NOT argue that you can control for *all* of the relevant factors.
  - Adding more control variables in the regression usually does not help.

- Reasonable solution: find some *exogenous* variations in the endogenous variables

- What are exogenous variations?
  - We need the endogenous variable (x) to change for reasons unrelated to the outcome variable (y).

Common Tools to Establish Causality

1. Instrumental Variables
2. Difference-in-Differences (natural/qusi-natural experiments)
3. Regression Discontinuity Design

Instrumental Variables (IV)

Addressing Endogeneity with IV

- Instrumental variable (z):
  - Correlated with endogenous variable (x)
  - Uncorrelated with error term (ϵ)

- Purpose: Isolate the variation in x that is exogenous

Family CEOs and Firm Performance

$$Performance = \alpha + \beta FamilyCEO + \epsilon$$

Family CEOs and Firm Performance

$$Performance = \alpha + \beta FamilyCEO + \epsilon$$

- Endogeneity: the characteristics of the firm and family that cause it to choose a family CEO may also cause the change in performance.

Family CEOs and Firm Performance

$$Performance = \alpha + \beta FamilyCEO + \epsilon$$

- Endogeneity: the characteristics of the firm and family that cause it to choose a family CEO may also cause the change in performance.
- An ideal experiment
  - Take a bunch of firms and randomly assign to a CEO who is either a family member or non-family member.

Family CEOs and Firm Performance

$$Performance = \alpha + \beta FamilyCEO + \epsilon$$

- Endogeneity: the characteristics of the firm and family that cause it to choose a family CEO may also cause the change in performance.
- An ideal experiment
  - Take a bunch of firms and randomly assign to a CEO who is either a family member or non-family member.
- Bennedsen, Nielsen, Perez-Gonzalez, and Wolfenzon (2007, QJE)
  - Male first-child firms are more likely to pass on control to a family CEO.
  - Gender of the departing CEO's first child is unlikely to affect firms' future outcomes.
- Method: using the gender of a departing CEO's firstborn child as an IV for $FamilyCEO$.

Implementing IV Estimation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \epsilon$$

- Suppose x_k is an endogenous variable (i.e., $cov(x_k, \epsilon) \neq 0$), and z is an instrument for x_k.

Implementing IV Estimation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \epsilon$$

- Suppose x_k is an endogenous variable (i.e., $cov(x_k, \epsilon) \neq 0$), and z is an instrument for x_k.
- The IV estimation can be implemented in two stages (2SLS):

Implementing IV Estimation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \epsilon$$

- Suppose x_k is an endogenous variable (i.e., $cov(x_k, \epsilon) \neq 0$), and z is an instrument for x_k.

- The IV estimation can be implemented in two stages (2SLS):
  - **First-stage**: regress x_k on other x's and z; obtain the predicted value of x_k (i.e., $\hat{x}_k$)
    - ★ Estimate $x_k = \theta_0 + \theta_1 x_1 + \cdots + \theta_{k-1} x_{k-1} + \gamma z + \eta$
    - ★ Calculate $\hat{x}_k = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_{k-1} x_{k-1} + \hat{\gamma} z$

Implementing IV Estimation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \epsilon$$

- Suppose x_k is an endogenous variable (i.e., $cov(x_k, \epsilon) \neq 0$), and z is an instrument for x_k.
- The IV estimation can be implemented in two stages (2SLS):
  - **First-stage**: regress x_k on other x's and z; obtain the predicted value of x_k (i.e., $\hat{x}_k$)
    - ★ Estimate $x_k = \theta_0 + \theta_1 x_1 + \cdots + \theta_{k-1} x_{k-1} + \gamma z + \eta$
    - ★ Calculate $\hat{x}_k = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_{k-1} x_{k-1} + \hat{\gamma} z$
  - **Second-stage**: replace x_k with $\hat{x}_k$ in the original model
    - ★ $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k \hat{x}_k + \epsilon$

Implementing IV Estimation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \epsilon$$

- Suppose x_k is an endogenous variable (i.e., $cov(x_k, \epsilon) \neq 0$), and z is an instrument for x_k.
- The IV estimation can be implemented in two stages (2SLS):
  - **First-stage**: regress x_k on other x's and z; obtain the predicted value of x_k (i.e., $\hat{x}_k$)
    - ★ Estimate $x_k = \theta_0 + \theta_1 x_1 + \cdots + \theta_{k-1} x_{k-1} + \gamma z + \eta$
    - ★ Calculate $\hat{x}_k = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_{k-1} x_{k-1} + \hat{\gamma} z$
  - **Second-stage**: replace x_k with $\hat{x}_k$ in the original model
    - ★ $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k \hat{x}_k + \epsilon$
- Why do we need z? Why not just use other x's?

Conditions for a Valid Instrument

- IV can address endogeneity problems only if properly executed.

- You need at least as many instruments as endogenous regressors.
  - You can test the overidentifying restrictions of the model if you have more instruments than endogenous regressors.

Conditions for a Valid Instrument

- IV can address endogeneity problems only if properly executed.

- You need at least as many instruments as endogenous regressors.
  - You can test the overidentifying restrictions of the model if you have more instruments than endogenous regressors.

- Valid IVs must satisfy two conditions:

Conditions for a Valid Instrument

- IV can address endogeneity problems only if properly executed.

- You need at least as many instruments as endogenous regressors.
  - You can test the overidentifying restrictions of the model if you have more instruments than endogenous regressors.

- Valid IVs must satisfy two conditions:
  1. **Relevance condition**
     - ★ Conditionally correlated with the endogenous variable
     - ★ Testable (e.g., first-stage F-statistic in 2SLS)

Conditions for a Valid Instrument

- IV can address endogeneity problems only if properly executed.
- You need at least as many instruments as endogenous regressors.
  - You can test the overidentifying restrictions of the model if you have more instruments than endogenous regressors.
- Valid IVs must satisfy two conditions:
  1. **Relevance condition**
     - ★ Conditionally correlated with the endogenous variable
     - ★ Testable (e.g., first-stage F-statistic in 2SLS)
  2. **Exclusion condition**
     - ★ Uncorrelated with ϵ (i.e., cannot explain y other than through the endogenous variable)
     - ★ NOT testable
     - ★ Must have compelling arguments for instrument validity (and rule out alternative hypotheses with robustness tests)

Practical Issues to Keep in Mind

1. Do NOT estimate the two stages on your own!
   - Let statistical software do it for you. In Stata, this can be done by commends like *IVREG*.
2. Always report the first-stage results.
   - This is an indirect test for the relevance condition (i.e., $\gamma \neq 0$, R^2, F-statistic).

Further IV Examples in Corporate Finance

- Babina, Fedyk, He, and Hodson (2024, JFE)
  - Firms' AI investments and firm growth

- Giroud, Mueller, Stomper, and Westerkamp (2012, RFS)
  - Debt overhang and firm performance

Difference-in-Differences (DID)

Natural Experiments

- Recall: we need exogenous variations in the endogenous variables to address endogeneity.

- In science, researchers can simply design randomized experiments to achieve this.
  - To determine effect of new drug, you randomly give it to certain patients

Natural Experiments

- Recall: we need exogenous variations in the endogenous variables to address endogeneity.

- In science, researchers can simply design randomized experiments to achieve this.
  - To determine effect of new drug, you randomly give it to certain patients

- In social science, we normally cannot run such experiments.
  - E.g., we can't randomly assign a firm's debt-to-asset to determine its effect on investment

Natural Experiments

- Recall: we need exogenous variations in the endogenous variables to address endogeneity.
- In science, researchers can simply design randomized experiments to achieve this.
  - To determine effect of new drug, you randomly give it to certain patients
- In social science, we normally cannot run such experiments.
  - E.g., we can't randomly assign a firm's debt-to-asset to determine its effect on investment
- Our solution is to rely on **natural experiments**, which are basically situations where some "natural" event causes an exogenous change in a variable of interest, x.
  - Natural experiments occur when something happens that affects one group but not another, like a law being passed or a natural event occurring.

Natural Experiments

- Recall: we need exogenous variations in the endogenous variables to address endogeneity.
- In science, researchers can simply design randomized experiments to achieve this.
  - To determine effect of new drug, you randomly give it to certain patients
- In social science, we normally cannot run such experiments.
  - E.g., we can't randomly assign a firm's debt-to-asset to determine its effect on investment
- Our solution is to rely on **natural experiments**, which are basically situations where some "natural" event causes an exogenous change in a variable of interest, x.
  - Natural experiments occur when something happens that affects one group but not another, like a law being passed or a natural event occurring.
  - Example 1: A new minimum wage law is passed in some states but not others.

Natural Experiments

- Recall: we need exogenous variations in the endogenous variables to address endogeneity.

- In science, researchers can simply design randomized experiments to achieve this.
  - To determine effect of new drug, you randomly give it to certain patients

- In social science, we normally cannot run such experiments.
  - E.g., we can't randomly assign a firm's debt-to-asset to determine its effect on investment

- Our solution is to rely on **natural experiments**, which are basically situations where some "natural" event causes an exogenous change in a variable of interest, x.
  - Natural experiments occur when something happens that affects one group but not another, like a law being passed or a natural event occurring.
  - Example 1: A new minimum wage law is passed in some states but not others.
  - Example 2: A wildfire impacts certain companies in one region but not others.

Cross-sectional Difference

- One way to estimate the treatment effect is to compare the “treated” group (those affected by the law or event) and the “control” group (those not affected) **after the treatment**.

- E.g., one could compare corporate outcomes (e.g., profit margin) of companies in states that passed a new law to those in states that didn’t.

- Simple Equation:

$$y = \alpha + \beta \times Treated + \epsilon$$

  - y: the outcome (e.g., profit margin);
  - $Treated$: dummy variable that equals 1 in a state with the law and 0 otherwise.

- Problem with the cross-sectional difference:

  - If you only compare two groups at one point in time (after the treatment), there could be **other differences** between the groups that affect the outcome. For example, firms in different states may have different economic conditions regardless of the new law.

Time-series Difference

- Another way to estimate the treatment effect is to compare the outcome after the treatment with the outcome before the treatment for just the treated group.

- Simple Equation:

$$y = \alpha + \beta \times Post + \epsilon$$

  - y: the outcome (e.g., profit margin);
  - $Post$: dummy variable that equals 1 if the period is after the law adoption, and 0 otherwise.

Time-series Difference

- Another way to estimate the treatment effect is to compare the outcome after the treatment with the outcome before the treatment for just the treated group.

- Simple Equation:

$$y = \alpha + \beta \times Post + \epsilon$$

  - y: the outcome (e.g., profit margin);
  - $Post$: dummy variable that equals 1 if the period is after the law adoption, and 0 otherwise.

- Problem with the time-series difference:
  - If you only compare before and after within the same group (treated group), the outcome might have changed over time even without the treatment due to **other time trends**, like changes in the economy.

Linking Cross-section Difference with Time-series Difference

- The two single difference estimators complement each another.
- The cross-sectional comparison avoids the problem of omitted trends by comparing two groups over the same time period.
- The time series comparison avoids the problem of unobserved differences between two different groups of firms by looking at the same firms before and after the change.

Linking Cross-section Difference with Time-series Difference

- The two single difference estimators complement each another.
- The cross-sectional comparison avoids the problem of omitted trends by comparing two groups over the same time period.
- The time series comparison avoids the problem of unobserved differences between two different groups of firms by looking at the same firms before and after the change.
- The double difference, difference-in-differences (DID), estimator combines these two estimators to take advantage of both estimators' strengths.
  - We compare the **change** in the outcome (before and after the treatment) for the treated group and the control group.
  - This allows us to **cancel out any differences** that would have affected both groups equally (like economic trends) or that exist between the groups even without the treatment.

Linking Cross-section Difference with Time-series Difference

- The two single difference estimators complement each another.

- The cross-sectional comparison avoids the problem of omitted trends by comparing two groups over the same time period.

- The time series comparison avoids the problem of unobserved differences between two different groups of firms by looking at the same firms before and after the change.

- The double difference, difference-in-differences (DID), estimator combines these two estimators to take advantage of both estimators' strengths.
  - We compare the **change** in the outcome (before and after the treatment) for the treated group and the control group.
  - This allows us to **cancel out any differences** that would have affected both groups equally (like economic trends) or that exist between the groups even without the treatment.

- Example: compare the change in average wage **before and after** a minimum wage law in **both treated and control states**.

Difference-in-Differences Estimator

- Intuition: compare **change** in y pre- versus post-treatment for treated group [1st difference] to **change** in y pre- versus post-treatment for control group [2nd difference]

- The difference-in-differences approach can be implemented using

$$y_{i,t} = \alpha + \beta_1 Post_t + \beta_2 Treated_i + \beta_3 (Treated_i \times Post_t) + \epsilon_{i,t}$$

  - $Post_t$ equals 1 if period t occurs after the treatment, and 0 otherwise.
  - $Treated_i$ equals 1 for i is in the treated group, and 0 otherwise.
  - $Treated_i \times Post_t$: the interaction term between $Treated_i$ and $Post_t$.

- Uncover the DID estimator from group means

	Pre-Treatment	Post-Treatment	Difference
	(1)	(2)	(2)-(1)
Treatment Group (a)	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_3$
Control Group (b)	α	$\alpha + \beta_1$	β_1
Difference (a)-(b)	β_2	$\beta_2 + \beta_3$	β_3

One DID Example: Tax and Corporate Investment

- Some cities raise corporate taxes, while others keep them the same.
  - Treated group: Firms in cities that raised taxes.
  - Control group: Firms in cities that didn't raise taxes.
  - Outcome: Firms' investment levels.

- DID Estimator:

$$y = \alpha + \beta_1 \times Post + \beta_2 \times Treated + \beta_3 \times Treated \times Post + \epsilon$$

- $\hat{\beta}_3$ captures the effect of higher corporate taxes on firms' investment, controlling for broader trends in both groups.

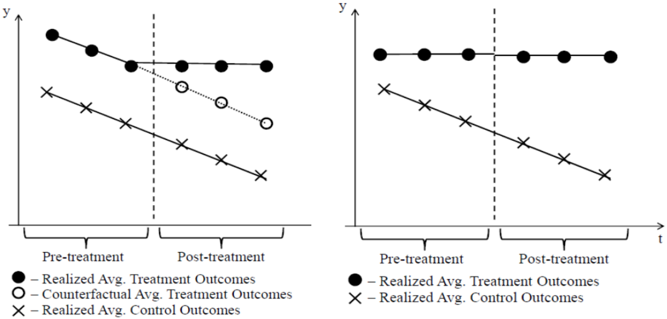
Parallel Trends Assumption

- Key Identification Assumption: in the absence of treatment, the average change in y would have been the same for both the treatment and control groups.
    - a.k.a. the **parallel trends assumption** since it requires that the trend in the outcome variable for both treatment and control groups during the pre-treatment period are similar.

Parallel Trends Assumption

- Key Identification Assumption: in the absence of treatment, the average change in y would have been the same for both the treatment and control groups.
  - a.k.a. the **parallel trends assumption** since it requires that the trend in the outcome variable for both treatment and control groups during the pre-treatment period are similar.

- Visualization of parallel trends assumption:



  - The left panel satisfies the assumption; the right panel violates the assumption.

DID Examples in Corporate Finance

- Derrien and Kecskes (2013, JF)
  - analyst coverage and corporate financial policies
- Mukherjee, Singh, and Žaldokas (2017, JFE)
  - corporate taxes and innovation

Do corporate taxes hinder innovation?

Abhiroop Mukherjee, Manpreet Singh, and Alminas Žaldokas

JFE, 2017

Motivation

- High corporate taxes are often debated:
    - Do they hurt innovation?
    - Or, do they fund public services without affecting competitiveness?

- **Key Question**: Do corporate taxes have real effects on firms' innovation activities?

Motivation

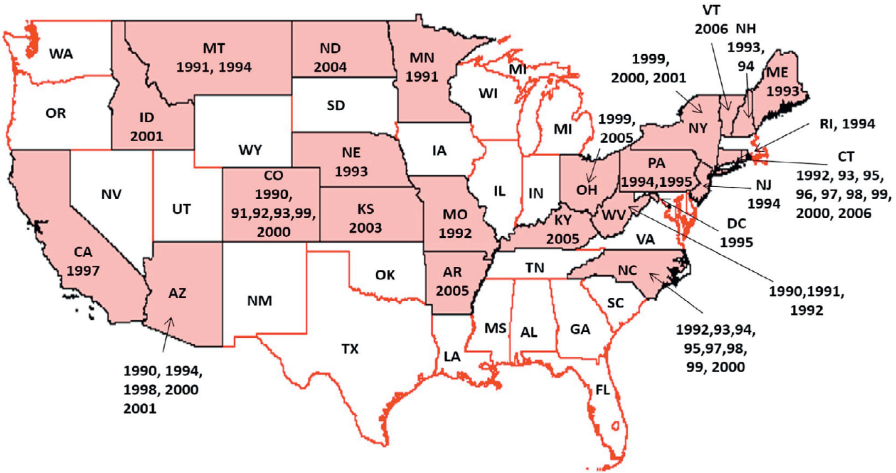
- High corporate taxes are often debated:
  - Do they hurt innovation?
  - Or, do they fund public services without affecting competitiveness?
- **Key Question**: Do corporate taxes have real effects on firms' innovation activities?
- Endogeneity concerns:
  - **Reverse causality:** a state might lower taxes to attract more innovative firms.
  - **Omitted variables**: states with a robust education system might have both lower taxes and higher innovation.

Motivation

- High corporate taxes are often debated:
  - Do they hurt innovation?
  - Or, do they fund public services without affecting competitiveness?
- **Key Question**: Do corporate taxes have real effects on firms' innovation activities?
- Endogeneity concerns:
  - **Reverse causality:** a state might lower taxes to attract more innovative firms.
  - **Omitted variables**: states with a robust education system might have both lower taxes and higher innovation.
- Identification Strategy: difference-in-differences framework using U.S. state-level corporate tax changes as exogenous shocks

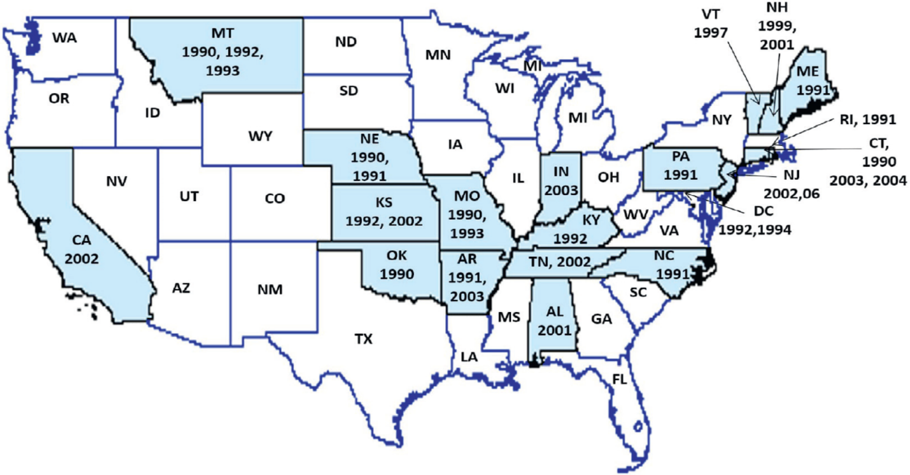
Geography of state corporate income tax changes, 1990 – 2006

1A: Tax decreases



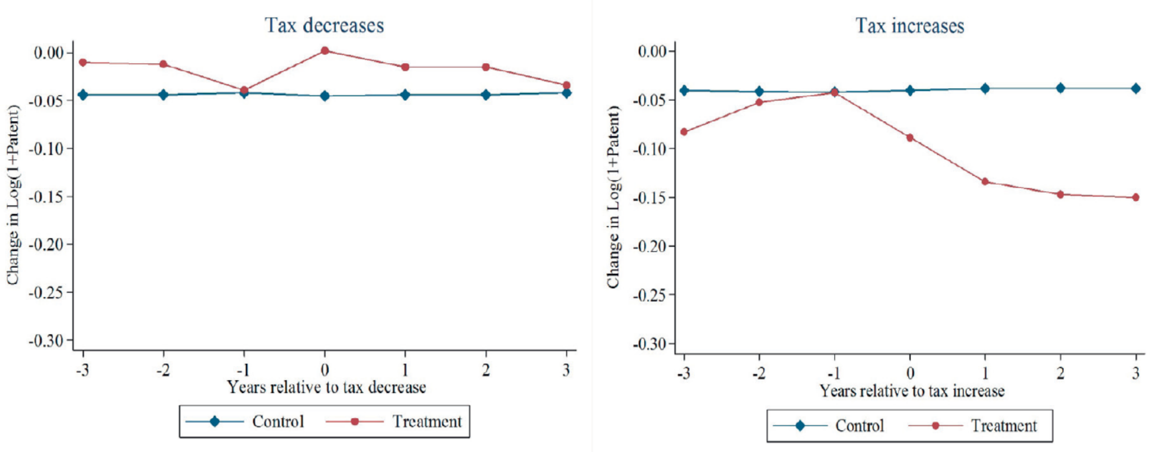
Geography of state corporate income tax changes, 1990 – 2006

1B: Tax increases

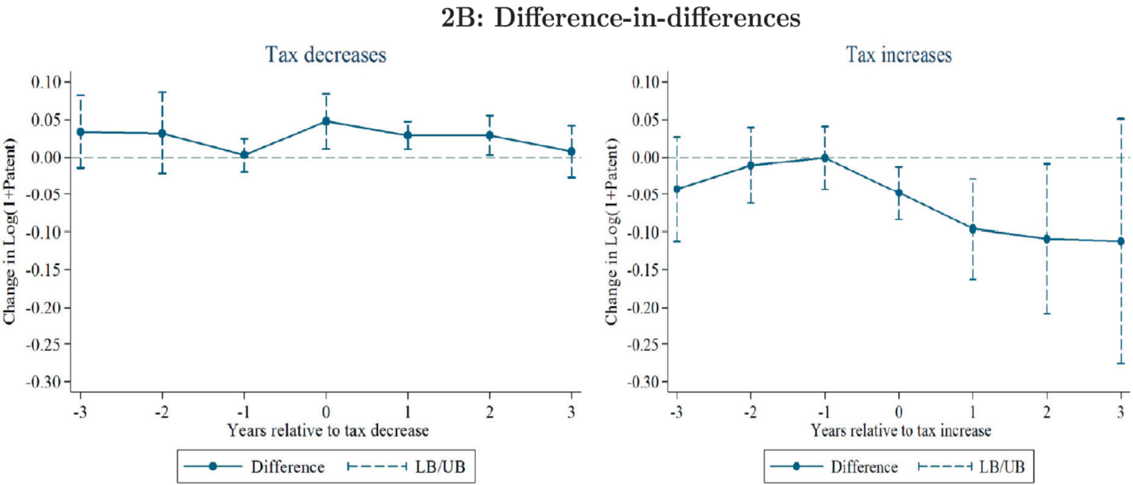


Graphical Plots (DID)

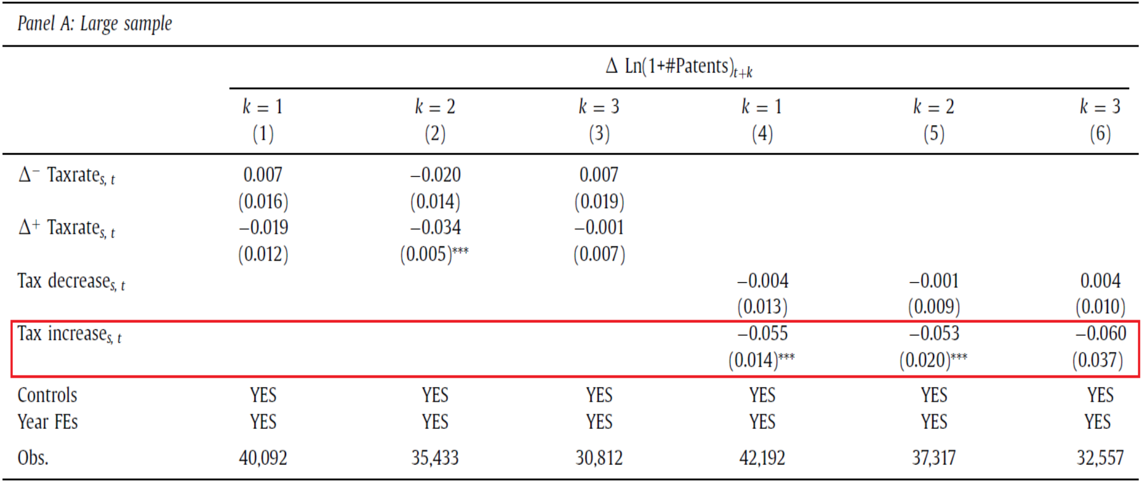
2A: Treatment and control



Graphical Plots (DID)



Baseline Results (DID)



Panel A: Large sample						
	Δ Ln(1+#Patents)$_{t+k}$					
	k = 1	k = 2	k = 3	k = 1	k = 2	k = 3
	(1)	(2)	(3)	(4)	(5)	(6)
Δ^{-} Taxrate$_{s,\ t}$	0.007	−0.020	0.007			
	(0.016)	(0.014)	(0.019)			
Δ^{+} Taxrate$_{s,\ t}$	−0.019	−0.034	−0.001			
	(0.012)	(0.005)***	(0.007)			
Tax decrease$_{s,\ t}$				−0.004	−0.001	0.004
				(0.013)	(0.009)	(0.010)
Tax increase$_{s,\ t}$				−0.055	−0.053	−0.060
				(0.014)***	(0.020)***	(0.037)
Controls	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES
Obs.	40,092	35,433	30,812	42,192	37,317	32,557

Regression Discontinuity Design (RDD)

RDD Intuition

- RDD intuition: observations (e.g., firm, individual, etc.) are "treated" based on known cutoff rule
  - E.g., for some observable variable, x, an observation is treated if $x \geq x'$
  - This cutoff is what creates the discontinuity.
- Researcher is interested in how this treatment affects outcome variable of interest, y.

RDD Intuition

- RDD intuition: observations (e.g., firm, individual, etc.) are “treated” based on known cutoff rule
  - E.g., for some observable variable, x, an observation is treated if $x \geq x'$
  - This cutoff is what creates the discontinuity.
- Researcher is interested in how this treatment affects outcome variable of interest, y.
- Recall: we need exogenous variations in the endogenous variables to address endogeneity.

RDD Intuition

- RDD intuition: observations (e.g., firm, individual, etc.) are "treated" based on known cutoff rule
  - E.g., for some observable variable, x, an observation is treated if $x \geq x'$
  - This cutoff is what creates the discontinuity.
- Researcher is interested in how this treatment affects outcome variable of interest, y.
- Recall: we need exogenous variations in the endogenous variables to address endogeneity.
- Exogenous variation is a consequence of RDD as long as agents are unable to precisely manipulate their x value near the cutoff.
  - No manipulation is an important identification assumption in RDD.

Examples of Discontinuity

- A borrower FICO score $>$ 620 makes securitization of the loan more likely (Keys, Mukherjee, Seru, and Vig, QJE 2010).
- Firms exceeding a certain size threshold suffer from mandatory requirements on the minimum fraction of outside directors (Black and Kim, JFE 2012).
- Firms that are ranked higher than a threshold in a competition can be awarded R&D grants (Howell and Brown, RFS 2023).

RDD Terminology

- The variable that determines treatment assignment, x, is called the “forcing variable”.
  - a.k.a running variable, assignment variable, ratings variable, selection variable
- x' is the known cutoff/threhold
- Estimating local treatment effect: compare outcomes *just* above and below x'
  - $y(0)$ is outcome absent treatment
  - $y(1)$ is outcome with treatment

Two Types of RDD

1. **Sharp RDD**
   - Assignment to treatment only depends on x; i.e., agents are treated with probability 1 if $x \geq x'$ (or $x \leq x'$ depending on the assignment rule).

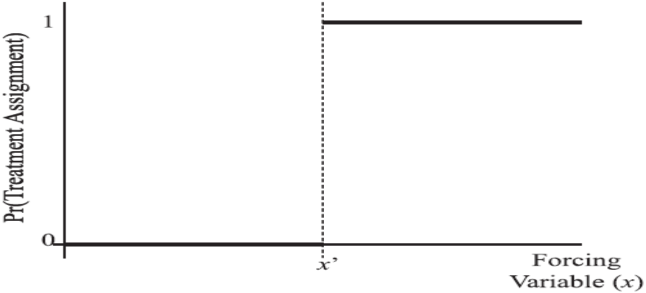
2. **Fuzzy RDD**
   - Having $x \geq x'$ only increases probability of being treated; i.e., other factors (besides x) will influence whether agents are treated or not.

Sharp RDD Assumption #1

- Assignment to treatment occurs through known and deterministic decision rule:

$$d = d(x) = \begin{cases} 1, & \text{if } x \geq x', \\ 0, & \text{otherwise.} \end{cases}$$

    - Treatment assignment is determined only by x.



Probability of Treatment Assignment in Sharp RDD

Sharp RDD Examples

- Students receive national merit scholarship if GPA $> x'$.
  - Scholarships are awarded solely based on a student's previous GPA
  - Thistlethwaite and Campbell (1960) used this to study effects of scholarship on career plans.

Sharp RDD Examples

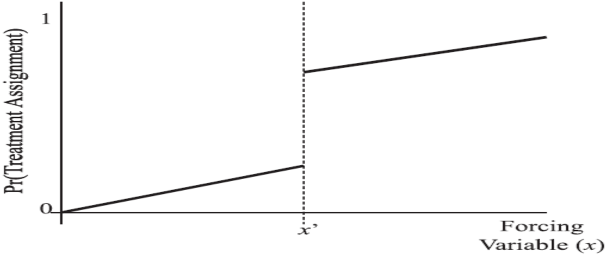
- Students receive national merit scholarship if GPA $> x'$.
  - Scholarships are awarded solely based on a student's previous GPA
  - Thistlethwaite and Campbell (1960) used this to study effects of scholarship on career plans.
- Labor unions are formed if more than 50% of the voters are in favor of unionization.
  - Unionization is determined only by the percentage of votes that support it.
  - Bradley, Kim, and Tian (2017, MS) used this to examine effects of labor unions on corporate innovation.

Fuzzy RDD Assumption #1

- Assignment to treatment occurs in a stochastic manner where the **probability of assignment** (a.k.a. propensity score) has a known discontinuity at x'

$$0 < \lim_{x \downarrow x'} \Pr ob(d = 1|x) - \lim_{x \uparrow x'} \Pr ob(d = 1|x) < 1$$

  - Treatment is determined by x as well as other factors.
  - It can go the other way around.



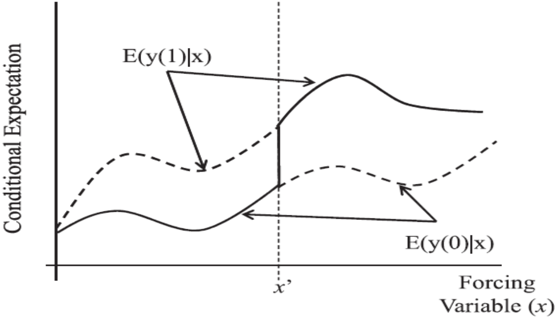
Probability of Treatment Assignment in Fuzzy RDD

Fuzzy RDD Examples

- FICO score $>$ 620 increases likelihood of loan being securitized.
  - Credit score is an important determinant of loan securitization.
  - But extent of loan documentation, lender, etc., will matter as well.
  - Keys, Mukherjee, Seru, and Vig, (2010, QJE) used this setting to examine whether securitization reduces lenders' screening incentives.

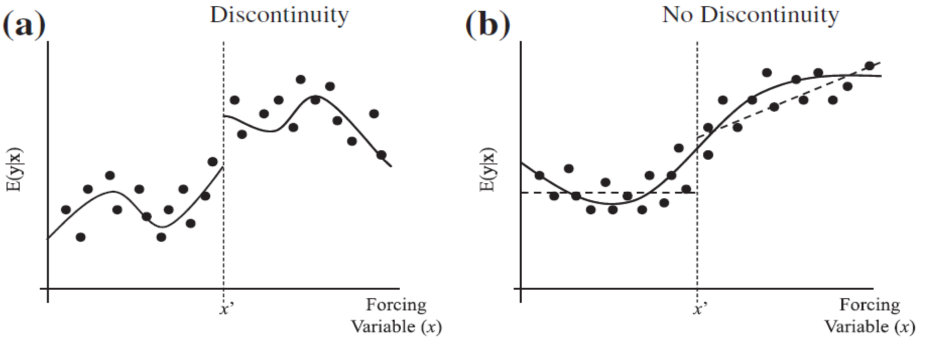
RDD Assumption #2

- Both potential outcomes, $E(y(0)|x)$ and $E(y(1)|x)$, are continuous in x at x'.
  - This means y would be a smooth function around threshold absent treatment.
  - For proper identification, RDD relies on the variation that occurs precisely at the discontinuity.
  - RDD estimators may not capture the true effect if the estimation includes points that are far from the discontinuity.
  - Both types of RDDs share this assumption.



Graphical Analysis

- A scatter plot helps a lot to visually inspect whether a discontinuity exists and whether chosen polynomial order seems to fit the data well.
- In RDD, it is recommended to always present such a graph as a simple "eyeball test".



How to Conduct the Graphical Analysis?

- First, divide x into bins, making sure no bin contains x' as an interior point.
  - E.g., if x ranges between 0 and 10 and treatment occurs for $x \geq x' = 5$, you could construct 10 bins, $[0,1),[1,2),\cdots,[9,10]$.
  - Or, if $x' = 4.5$, you could use something like $[0,0.5),[0.5,1.5),[1.5,2.5)$, etc.
- Second, calculate average y in each bin, and plot this average above the midpoint for each bin.
  - Plotted averages represent a non-parametric estimate of $E[y|x]$.
- Third, estimate your RDD and plot predicted values of y from the estimation, i.e., fit the functions in your RDD estimation.

Remarks on Graphical Analysis

- NEVER use RDD if there is no obvious jump around x'!

- NEVER use RDD if non-parametric plots suggests jumps in y at other points besides x'.
  - It is possible that jump at x' is driven by something else that is unrelated to treatment.

- The choice of optimal # of bins (i.e., bin width) is subjective because of tradeoff between precision and bias.
  - Robustness checks help again.

Estimating Sharp RDD

- There are generally two ways to estimate RDD:
  1. Use all data, but control for effect of x on y in a very general and rigorous way.
  2. Use less rigorous controls for effect of x, but only use data in small window around threshold.

- What are the strength and weakness for each method?
  - The first way uses more data but increases risk of bias that observations further from threshold might vary for other reasons.
  - The second way reduces the bias, but we might get imprecise estimates because the estimate can be very noisy due to the much smaller number of observations.

Method #1: Estimating Sharp RDD, *Using All Data*

- An easier way to estimate the effect is to run the following regression:

$$y_i = \alpha + \beta d_i + f(x_i - x') + d_i \times g(x_i - x') + \epsilon_i$$

  - d_i equals one if $x_i > x'$, and zero otherwise.
  - $\beta = \beta^a - \beta_b$
  - $f(x_i - x')$ and $d_i \times g(x_i - x')$ control for the effect of x on y below and above x', respectively.

- Can we drop $d_i \times g(x_i - x')$?
  - No, unless you assume that the functional form between x and y is same above and below x'.

- What should we use for $f(\cdot)$ and $g(\cdot)$?
  - In practice, a high-order polynomial (e.g., quadratic, cubic, 4th-order) function is used for both $f(\cdot)$ and $g(\cdot)$.
  - No obvious ways to determine which order to use. Just include multiple specifications as robustness checks.

Method # 2: Estimating Sharp RDD, *Using Window*

- The second approach does the same as Method #1, but
  - Restrict analysis to a smaller window around x'
  - Use linear functions as controls

- E.g., estimate the following using data in the window $[x' - \delta, x' + \delta]$ for small $\delta > 0$.

$$y_i = \alpha + \beta d_i + \gamma^b (x - x') + \gamma^a d_i (x - x') + \epsilon_i$$

  - Notice that $f(\cdot)$ and $g(\cdot)$ are linear functions here.

- **Practical Issue**: How to choose the appropriate window width?
  - No right answers. Again, we need to rely on robustness checks using multiple window width.

Estimating Fuzzy RDD

- In fuzzy RDD, comparison of average y immediately above and below threshold (as done in Sharp RDD) won't work.
  - Because not all observations above threshold are treated and not all below are untreated; $x \geq x$. just increases <u>probability</u> of treatment.

Estimating Fuzzy RDD

- In fuzzy RDD, comparison of average y immediately above and below threshold (as done in Sharp RDD) won't work.
  - Because not all observations above threshold are treated and not all below are untreated; $x \geq x.$ just increases <u>probability</u> of treatment.

- What we can do is to **use** $x \geq {'}x.$ **as an IV for treatment**.

Estimating Fuzzy RDD

- In fuzzy RDD, comparison of average y immediately above and below threshold (as done in Sharp RDD) won't work.
    - Because not all observations above threshold are treated and not all below are untreated; $x \geq x.$ just increases probability of treatment.

- What we can do is to **use** $x \geq {}'x.$ **as an IV for treatment**.

- Before talking about the implementation, let's introduce a threshold indicator, T_i.

$$T = T(x) = \begin{cases} 1, & \text{if } x \geq x', \\ 0, & \text{otherwise.} \end{cases}$$

    - Recall d_i is the treatment indicator (i.e., $d_i = 1$ if treated).
    - E.g., $d_i = 1$ if loan is securitized; $T_i = 1$ if FICO score > 620, which increases probability loan is securitized.

Estimating Fuzzy RDD

- The probability of treatment, $Pr(d_i|x_i) = E[d_i|x_i]$, is relevant to whether passing the threshold:
$$E[d_i|x_i] = \delta + \phi T_i$$
- More generally, we can control for the distance between x_i and x' (as we do in sharp RDD):
$$E[d_i|x_i] = \delta + \phi T_i + g(x - x')$$
- In fuzzy RDD, whether an agent is treated is determined by its probability of treatment and some random disturbance, i.e.,
$$d_i = E[d_i|x_i] + \omega_i$$
where ω is a random error independent of x.
- Therefore, a fuzzy RDD can be described by a two equation system:
$$y_i = \alpha + \beta d_i + f(x - x') + \epsilon_i$$
$$d_i = \delta + \phi T_i + g(x - x') + \omega_i$$
- This can be estimated using 2SLS.

Internal Validity Checks

- Here are some tests in RDD to validate its identification assumptions:
  1. Graphic analysis
  2. Robustness tests regarding the chosen polynomial
  3. Robustness tests regarding the chosen bandwidth
  4. Checks for "No Manipulation".
  5. Balance tests
  6. Falsification tests

No Manipulation

- A big concern in RDD is that agents might be able to manipulate their x value around the threshold.

- With manipulation, y might exhibit jump around x' absent treatment because of manipulation.
  - In other words, the RDD Assumption #2 is violated.
  - E.g., in Chava and Roberts (2008, JF), covenant violations are based on financial figures reported by the company, which has a clear incentive to avoid violating a covenant if doing so is costly.

- If agents can't perfectly manipulate it, then there will still be randomness in treatment. RDD still works in such cases.

- To conduct this test, you can plot the histogram or density of the forcing variable.

Example of Histogram Plot

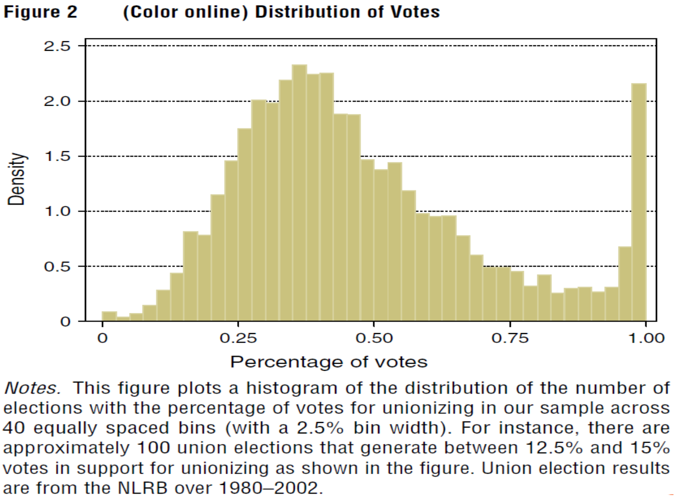


Figure 2 (Color online) Distribution of Votes

Notes. This figure plots a histogram of the distribution of the number of elections with the percentage of votes for unionizing in our sample across 40 equally spaced bins (with a 2.5% bin width). For instance, there are approximately 100 union elections that generate between 12.5% and 15% votes in support for unionizing as shown in the figure. Union election results are from the NLRB over 1980–2002.

Source: Bradley, Kim, and Tian (2017, MS)

Balance Tests

- RDD assumes observations near but on opposite sides of cutoff are comparable.

- How to conduct these tests?
  - Using graphical analysis or RDD, make sure other observable factors that might affect y don't exhibit jump at threshold x'.

- These tests do NOT prove validity of RDD. Why?

Balance Tests

- RDD assumes observations near but on opposite sides of cutoff are comparable.

- How to conduct these tests?
  - Using graphical analysis or RDD, make sure other observable factors that might affect y don't exhibit jump at threshold x'.

- These tests do NOT prove validity of RDD. Why?
  - There could be discontinuity in unobservables! Indeed, there is no way to perfectly prove causality.

Falsification Tests

1. False Threshold
   - Repeat the RDD analysis assuming that the treatment occurs at a false threshold
   - You should expect insignificant results with the fake threshold.

2. False Sample
   - This can be done if threshold x' only existed for certain types of observations (e.g., firms in particular years).
     - ★ E.g., law that created discontinuity was passed in a given year, but didn't exist before that, or maybe the law didn't apply to some firms.
   - Make sure no effect in years where there was no discontinuity or for observations where there isn't supposed to be an effect!

RDD Examples in Corporate Finance

- Chava and Roberts (2008, JF)
  - loan covenant violations → corporate investment

- Bradley, Kim, and Tian (2017, MS)
  - labor unions → corporate innovation

Do Unions Affect Innovation?

Daniel Bradley, Incheol Kim, Xuan Tian

MS, 2017

Motivation

- **Key Question**: Do unions promote or impede firm innovation?
  - Innovation is a crucial driver of economic growth.
  - Unions in the United States are regulated and can be altered by labor laws and regulations over time.
- Two Competing Hypotheses:
  - Employee Protectionism (e.g., higher productivity due to more efforts) ⇒ Higher Innovation.
  - Misaligned Incentives (e.g., talent departure due to lower wage inequality) ⇒ Lower Innovation.
- Identification Challenges:
  - **Omitted Variables**: union election results could be correlated with firm unobservable characteristics that affect firm innovation output.
  - **Reverse Causality**: firms with low innovation potential may be more likely to pass unionization elections (e.g., due to higher job security).

RDD as a Solution

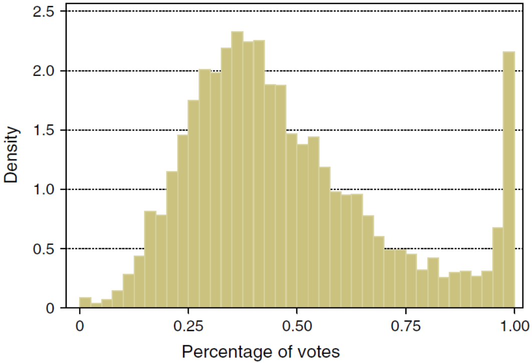
- A regression discontinuity design (RDD) that relies on “locally” exogenous variation in unionization generated by these elections that pass or fail by a small margin of votes.

- This approach compares firms’ innovation output subsequent to union elections that pass to those that do not pass by a small margin.

- Why does this setting address the concerns?
  - For these close-call elections, passing is very close to an independent, random event and therefore is unlikely to be correlated with firm unobservable characteristics.

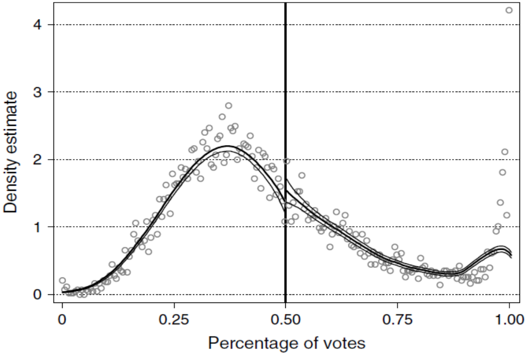
Tests for “No Manipulation”

Figure 2 **(Color online) Distribution of Votes**



Notes. This figure plots a histogram of the distribution of the number of elections with the percentage of votes for unionizing in our sample across 40 equally spaced bins (with a 2.5% bin width). For instance, there are approximately 100 union elections that generate between 12.5% and 15% votes in support for unionizing as shown in the figure. Union election results are from the NLRB over 1980–2002.

Figure 3 **Density of Union Vote Shares**



Notes. This figure plots the density of union vote shares following the procedure in McCrary (2008). The *x* axis is the percentage of votes favoring unionization. The dots depict the density estimate. The solid line represents the fitted density function of the forcing variable (the number of votes) with a 95% confidence interval around the fitted line. Union election results are from the NLRB over 1980–2002.

Balance Test

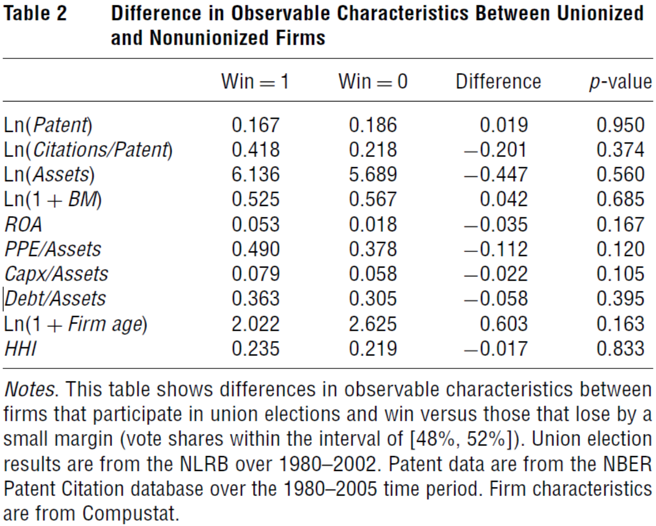
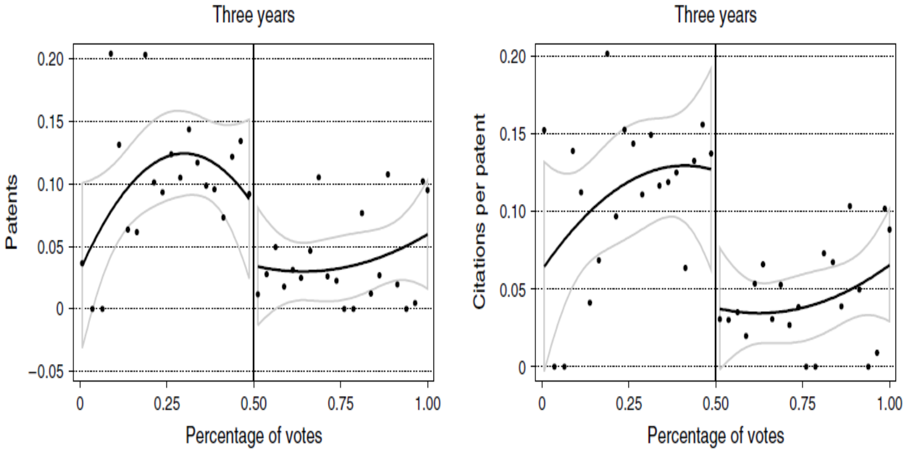


Table 2 Difference in Observable Characteristics Between Unionized and Nonunionized Firms

	Win = 1	Win = 0	Difference	*p*-value
Ln(*Patent*)	0.167	0.186	0.019	0.950
Ln(*Citations/Patent*)	0.418	0.218	−0.201	0.374
Ln(*Assets*)	6.136	5.689	−0.447	0.560
Ln(1 + *BM*)	0.525	0.567	0.042	0.685
ROA	0.053	0.018	−0.035	0.167
PPE/Assets	0.490	0.378	−0.112	0.120
Capx/Assets	0.079	0.058	−0.022	0.105
Debt/Assets	0.363	0.305	−0.058	0.395
Ln(1 + *Firm age*)	2.022	2.625	0.603	0.163
HHI	0.235	0.219	−0.017	0.833

Notes. This table shows differences in observable characteristics between firms that participate in union elections and win versus those that lose by a small margin (vote shares within the interval of [48%, 52%]). Union election results are from the NLRB over 1980–2002. Patent data are from the NBER Patent Citation database over the 1980–2005 time period. Firm characteristics are from Compustat.

Graphical Analysis



RDD Estimation: All Data

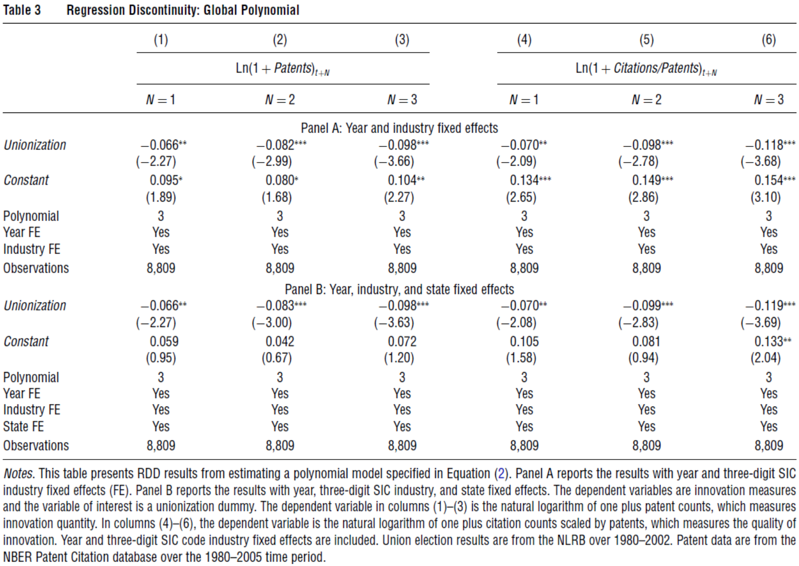


Table 3 Regression Discontinuity: Global Polynomial

	(1)	(2)	(3)	(4)	(5)	(6)
	$\text{Ln}(1+\textit{Patents})_{t+N}$			$\text{Ln}(1+\textit{Citations/Patents})_{t+N}$		
	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$
			Panel A: Year and industry fixed effects			
Unionization	-0.066^{**}	-0.082^{***}	-0.098^{***}	-0.070^{**}	-0.098^{***}	-0.118^{***}
	(−2.27)	(−2.99)	(−3.66)	(−2.09)	(−2.78)	(−3.68)
Constant	0.095^{*}	0.080^{*}	0.104^{**}	0.134^{***}	0.149^{***}	0.154^{***}
	(1.89)	(1.68)	(2.27)	(2.65)	(2.86)	(3.10)
Polynomial	3	3	3	3	3	3
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,809	8,809	8,809	8,809	8,809	8,809
			Panel B: Year, industry, and state fixed effects			
Unionization	-0.066^{**}	-0.083^{***}	-0.098^{***}	-0.070^{**}	-0.099^{***}	-0.119^{***}
	(−2.27)	(−3.00)	(−3.63)	(−2.08)	(−2.83)	(−3.69)
Constant	0.059	0.042	0.072	0.105	0.081	0.133^{**}
	(0.95)	(0.67)	(1.20)	(1.58)	(0.94)	(2.04)
Polynomial	3	3	3	3	3	3
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,809	8,809	8,809	8,809	8,809	8,809

Notes. This table presents RDD results from estimating a polynomial model specified in Equation (2). Panel A reports the results with year and three-digit SIC industry fixed effects (FE). Panel B reports the results with year, three-digit SIC industry, and state fixed effects. The dependent variables are innovation measures and the variable of interest is a unionization dummy. The dependent variable in columns (1)–(3) is the natural logarithm of one plus patent counts, which measures innovation quantity. In columns (4)–(6), the dependent variable is the natural logarithm of one plus citation counts scaled by patents, which measures the quality of innovation. Year and three-digit SIC code industry fixed effects are included. Union election results are from the NLRB over 1980–2002. Patent data are from the NBER Patent Citation database over the 1980–2005 time period.

RDD Estimation: Data in a Small Window

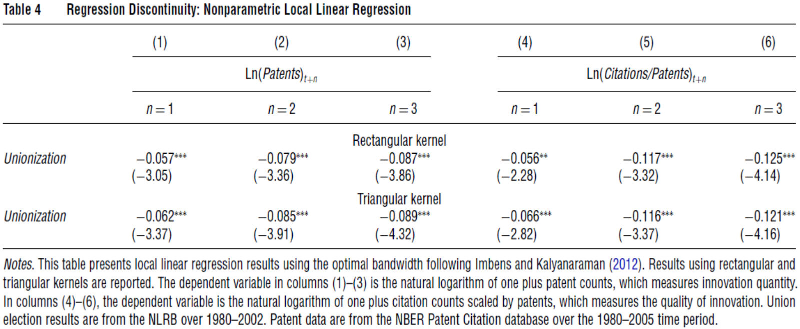
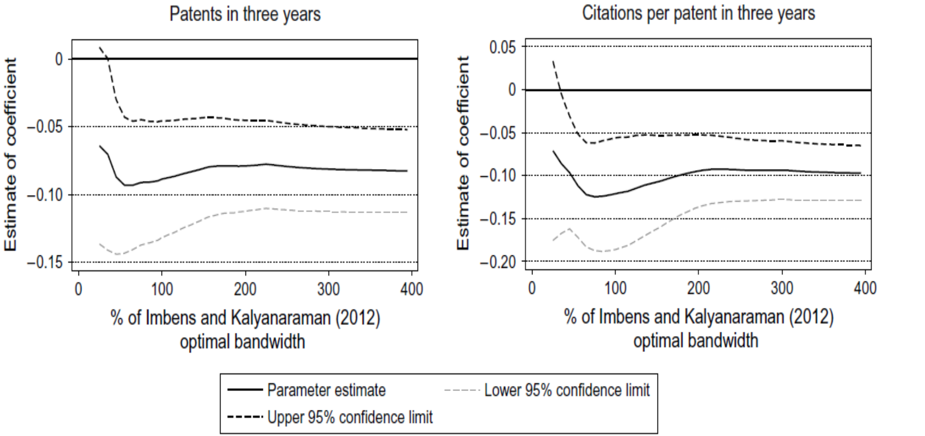


Table 4 Regression Discontinuity: Nonparametric Local Linear Regression

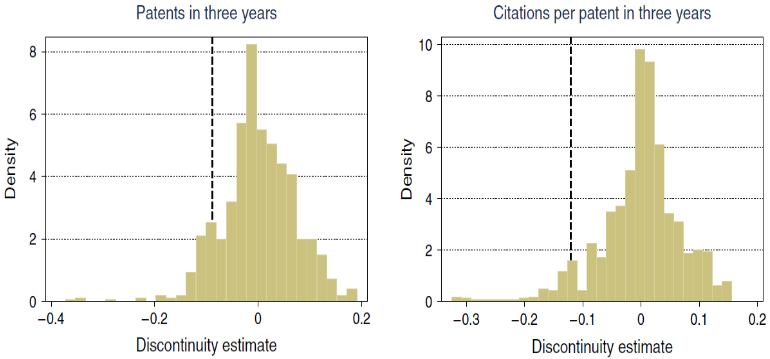
	(1)	(2)	(3)	(4)	(5)	(6)
	$\text{Ln}(\textit{Patents})_{t+n}$			$\text{Ln}(\textit{Citations/Patents})_{t+n}$		
	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
			Rectangular kernel			
Unionization	-0.057^{***}	-0.079^{***}	-0.087^{***}	-0.056^{**}	-0.117^{***}	-0.125^{***}
	(−3.05)	(−3.36)	(−3.86)	(−2.28)	(−3.32)	(−4.14)
			Triangular kernel			
Unionization	-0.062^{***}	-0.085^{***}	-0.089^{***}	-0.066^{***}	-0.116^{***}	-0.121^{***}
	(−3.37)	(−3.91)	(−4.32)	(−2.82)	(−3.37)	(−4.16)

Notes. This table presents local linear regression results using the optimal bandwidth following Imbens and Kalyanaraman (2012). Results using rectangular and triangular kernels are reported. The dependent variable in columns (1)–(3) is the natural logarithm of one plus patent counts, which measures innovation quantity. In columns (4)–(6), the dependent variable is the natural logarithm of one plus citation counts scaled by patents, which measures the quality of innovation. Union election results are from the NLRB over 1980–2002. Patent data are from the NBER Patent Citation database over the 1980–2005 time period.

Robustness with Bin Bandwith



Placebo Test



Notes. This figure plots a histogram of the distribution of the RDD estimates from placebo tests. The *x* axis represents the RDD estimates from a placebo test that artificially assumes an alternative threshold other than 50%. The dashed vertical line represents the RDD estimate at the true 50% threshold. Union election results are from the NLRB over 1980–2002. Patent data are from the NBER Patent Citation database over the 1980–2005 time period.

Further Reading

- Stock & Watson, Introduction to Econometrics
- Angrist & Pischke, Mostly Harmless Econometrics: An Empiricist's Companion
- Rodríguez, Stata Tutorial