# Endogeneity in empirical corporate finance

JAN HANOUSEK, CERGE-EI

E-MAIL: JAN.HANOUSEK@CERGE-EI.CZ

ROBUST 2018, January 23, 2018

# Roadmap

1. Basic endogeneity problem
2. Econometric responses
3. General advice
4. Application to capital structure research
5. Conclusions

# 1. Basic endogeneity problem

- The most important and pervasive issue confronting studies in empirical finance is **endogeneity**, which can be loosely defined as a correlation between the explanatory variables and the error term in the regression

- Endogeneity leads to biased and inconsistent parameter estimates that make reliable inference virtually impossible

- In many cases, endogeneity can be severe enough to reverse even qualitative inference

# 1. Basic endogeneity problem

i. Regression framework
   a. Omitted variables
   b. Simultaneity
   c. Measurement error

ii. Potential outcomes and treatment effects
   a. Notation and framework
   b. Link to regression and endogeneity

iii. Identifying and discussing endogeneity problem

# 1.i. Regression framework

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Key assumptions for OLS to produce consistent estimates of parameters:

1. Random sample of observations on $y$ and $(x_1, \ldots, x_k)$
2. Mean zero error term (i.e., $E(u) = 0$)
3. No linear relationships among the explanatory variables (i.e., no perfect collinearity)
4. Error term that is uncorrelated with each explanatory variable (i.e., $cov(x_j, u) = 0 \; j = \overline{1, k}$)

# 1.i. Regression framework

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

For unbiased estimates, one must replace assumption 4. with the following:

4'. Error term with zero mean conditional on the explanatory variables (i.e., $E(u|X) = 0$)

- Assumption 4 (or 4a) should be the focus of most research designs because violation of this assumption is the primary cause of inference problems.
- Yet, *there is no way to empirically test whether a variable is correlated with the regression error term because the error term is unobservable.*

# 1.i.a. Omitted variables

- Omitted variables refer to those variables that should be included in the vector of explanatory variables, but for various reasons are not

- True economic relation: $y = \beta_0 + \beta_1 x_1 + \ \ldots + \beta_k x_k + \gamma w + u$
- Estimable population regression: $y = \beta_0 + \beta_1 x_1 + \ \ldots + \beta_k x_k + v$, where
  $v = \gamma w + u$ – composite error term

- If the omitted variable $w$ is correlated with any of the explanatory variables, then the composite error tem $v$ is correlated with the explanatory variables. In this case, OLS estimation will typically produce inconsistent estimates of all of the elements of $\beta$

# 1.i.a. Omitted variables

- Suppose only one variable, say $x_j$, is correlated with the omitted variable, then

$$plim\ \widehat{\beta_l} = \beta_l \quad l\ \neq j$$

$$plim\ \widehat{\beta_j} = \beta_j + \gamma\ \phi_j,\ \ \phi_j = \frac{cov(x_j, w)}{Var(x_j)} \quad (if\ l =\ j)$$

- The last equation is useful for understanding the direction and potential magnitude of any omitted variables inconsistency: if $\gamma$ and $\phi_j$ have the same sign, then the asymptotic bias is positive, otherwise is negative

# 1.i.b. Simultaneity

- Simultaneity bias occurs when $y$ and one or more of the $x$'s are determined in equilibrium so that it can plausibly be argued either that $x_k$ causes $y$ or that $y$ causes $x_k$

- To illustrate simultaneity bias:

$$y = \beta x + u$$

$$x = \alpha y + v \quad u\ uncorrelaed\ with\ v$$

- Then,

$$\hat{\beta} = \frac{cov(x, y)}{var(x)} = \frac{cov(x, \beta x + u)}{var(x)} = \beta + \frac{cov(x, u)}{var(x)} = \beta + \frac{\alpha(1 - \alpha\beta)var(u)}{\alpha^2 var(u) + var(v)}$$

- This example illustrates the general principle that, unlike omitted variable bias, simultaneity bias is difficult to sign because it depends on the relative magnitudes of different effects, which cannot be known a priori.

# 1.i.c. Measurement error

- Most empirical studies in corporate finance use proxies for unobservable or difficult to quantify variables. Any discrepancy between the true variable of interest and the proxy leads to measurement error

- Measurement error in dependent variable:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + v,$$

where $v = w + u$ – composite error term,

$w \equiv y - y^*$, $y^*$ - unobservable measure, and

$y$ – observable version of or proxy for $y^*$

- The statistical implications of measurement error in the dependent variable are similar to those of an omitted variable

# 1.i.c. Measurement error

- Measurement error in independent variable: $y = \beta_0 + \beta_1 x_1 + \ \ldots + \beta_k x_k + v$, where $v = u \ - \beta_k w$ – composite error term, $w \ \equiv x_k - \ x_k^*$, $x_k^*$ - unobservable measure, and $x_k$ - its observable proxy

- $plim\ \widehat{\beta_k} = \beta_k \left(\frac{\sigma_r^2}{\sigma_r^2 + \sigma_w^2}\right)$ $\qquad plim\ \widehat{\beta_j} = \phi_{yx_j} \ - plim(\widehat{\beta_k})\phi_{xx_j}$

- $\sigma_r^2 \ - error\ variance\ from\ regression\ of\ x_k^*\ on\ (1, x_1, \ldots, x_{k-1})$

- $\phi_{yx_j} - coefficient\ on\ x_j\ in\ projection\ of\ y\ on\ (x_1, \ldots, x_{k-1})$
  $\phi_{xx_j} - coefficient\ on\ x_j\ in\ projection\ of\ x_k\ on\ (x_1, \ldots, x_{k-1})$

- Measurement error in $x_k$ generally produces inconsistent estimates of all of the $\beta_j$, even when the measurement error is uncorrelated with the other explanatory variables

# 1.ii. Potential outcomes and treatment effects

- Many studies in empirical corporate finance compare the outcomes of two or more groups.
- The quantity of interest in each of these studies is the causal effect of a binary variable(s) on the outcome variables.
- This quantity is referred to as a treatment effect, a term derived from the statistical literature on experiments

# 1.ii.a. Notation and framework

$$y = \begin{cases} y(0) & if\ d = 0 \\ y(1) & if\ d = 1 \end{cases} = y(0) + d[y(1) - y(0)],$$

where $y$ – observable outcome variable, $d$ – observable treatment indicator

- The problem of inference in this setting is tantamount to a missing data problem
- To estimate the treatment effect, researchers are forced to estimate

$$E(y|d = 1) - E(y|d = 0)$$

# 1.ii.a. Notation and framework

- To estimate the treatment effect, researchers are forced to estimate

$$E(y|d = 1) - E(y|d = 0)$$

- This can be rewritten by

$$E(y|d = 1) - E(y|d = 0) = \{E[y(1)|d = 1] - E[y(0)|d = 1]\} - \{E[y(0)|d = 1] - E[y(0)|d = 0]\}$$

The first difference on the right-hand side of equation is average treatment effect on the treated. The second difference is the selection bias.

Thus, a simple comparison of treatment and control group averages does not identify a treatment effect.

# 1.ii.b. Link to regression and endogeneity

$$y = \beta_0 + \beta_1 d + u,$$

where $\beta_0 = E[y(0)]$, $\beta_1 = y(1) - y(0)$,

and $u = y(0) - E[y(0)]$

$$E(y|d = 1) - E(y|d = 0) = \beta_1 + [E(y(0)|d = 1) - E(y(0)|d = 0)]$$

- OLS estimation of regression equation will not recover $\beta_1$, rather, the estimate $\widehat{\beta_1}$ will be confounded by the selection bias term, unless treatment assignment is random with respect to the potential outcomes.

# 1.ii.b. Link to regression and endogeneity

$$y = \beta_0 + \beta_1 d + u,$$

where $\beta_0 = E[y(0)]$, $\beta_1 = y(1) - y(0)$, and $u = y(0) - E[y(0)]$

$$E(y|d = 1) - E(y|d = 0) = \beta_1 + [E(y(0)|d = 1) - E(y(0)|d = 0)]$$

- The implication of nonrandom assignment for estimating causal treatment effects is akin to the implications of including an endogenous dummy variable in a linear regression

- The solution is similar: find random variation in the treatment assignment or, equivalently, exogenous variation in the dummy variable

# 1.iii. Identifying and discussing endogeneity problem

- Necessary first step in any empirical corporate finance study focused on disentangling alternative hypotheses or identifying causal effects is identifying the endogeneity problem and its implications for inference
- There are a number of questions that should be answered before putting forth a solution
  - ✓ Specifically, what is the endogenous variable(s)?
  - ✓ Why are they endogenous?
  - ✓ What are the implications for inferences of the endogeneity problems?
  - ✓ What are the alternative hypotheses about which one should be concerned?

# 2. Econometric responses

- Econometric techniques aimed at addressing endogeneity problems can be broadly classified into two categories
  - The first category includes techniques that rely on a clear source of exogenous variation for identifying the coefficients of interests
    - Examples of these techniques include instrumental variables, difference-in-differences estimators, and regression discontinuity design
  - The second category includes techniques that rely on more heavily on modeling assumptions, as opposed to a clear source of exogenous variation
    - Examples of these techniques include panel data methods (e.g., fixed and random effects), matching methods, and measurement error methods

# 2. Econometric responses

i. Panel data techniques

ii. Regression discontinuity design

iii. Matching methods

iv. Instrumental variables

v. Natural experiment/Difference-in-difference

# 2.i. Panel data techniques

- Again, one of the most common causes of endogeneity in empirical corporate finance is omitted variables, and omitted variables are a problem because of the considerable heterogeneity present in many corporate finance settings. Panel data sometimes offer a partial, but by no means complete and costless, solution to this problem

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it},$$

$$u_{it} = c_i + e_{it}, \qquad i = \overline{1, N}, t = \overline{1, T}$$

The term $c_i$ can be interpreted as capturing the aggregate effect of all the unobservable, time-invariant explanatory variables for $y_{it}$

# 2.i. Panel data techniques

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}, \qquad u_{it} = c_i + e_{it}, \qquad i = \overline{1,N}, t = \overline{1,T}$$

- If $c_i$ and $x_{it}$ are correlated, then $c_i$ is referred to as “fixed effect”.
- The possible remedies to the endogeneity problem in this case are
  - Deviations-from-individual-means regression

$$(y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} = \beta_1 \left( x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it} \right) + (e_{it} - \frac{1}{T}\sum_{t=1}^{T} e_{it})$$

  - First differencing $\Delta y_{it} = \beta_1 \Delta x_{it} + \Delta e_{it}$

# 2.i. Panel data techniques

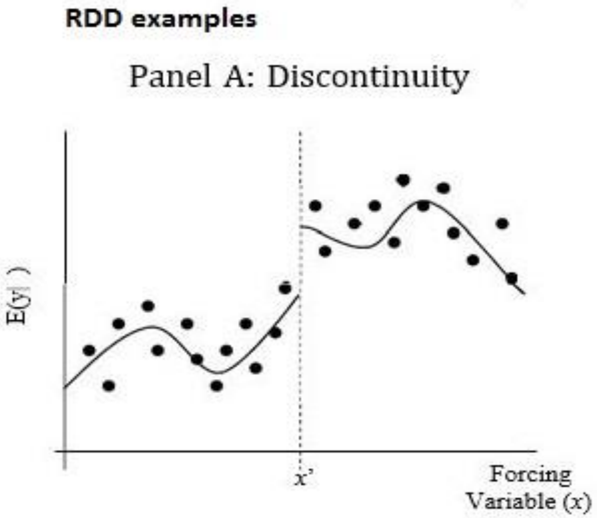
- **Advantages**:
  - Panel data techniques are useful since they at least limit the scope of the endogeneity problem, because you cannot be criticized for omitting some factor that does not vary through the time.
  - You have controlled for that.
  - It is a bit of a solution in that sense, but it can never completely take care of the problem unless the source of the endogeneity is perfectly known

# 2.i. Panel data techniques

- **Disadvantages**:
  - Including fixed effects can exacerbate measurement problems
  - If the research question is inherently aimed at understanding cross-sectional variation in a variable, then the fixed effects defeat this purpose
  - They do nothing to address endogeneity associated with correlation between $x_{it}$ and $c_i$
  - In some instances fixed effects eliminate the most interesting or important variation researchers wish to explain

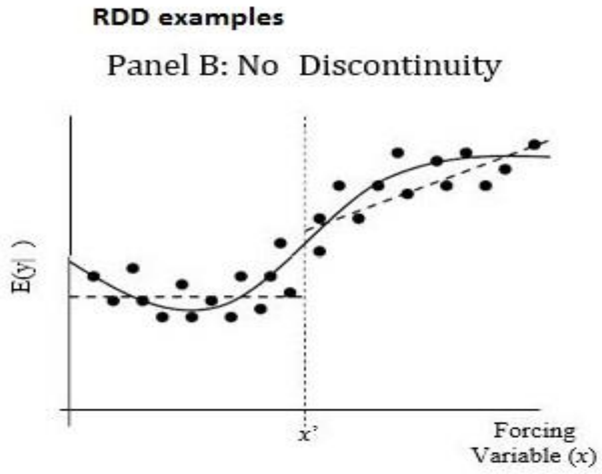
# 2.ii. Regression discontinuity design

- Regression discontinuity design (RDD) is quasi-experimental technique
- The idea is that we have some sort of threshold, and when we cross this threshold we can test if there’s a difference in behavior



$$y = \alpha + \beta d + u$$

$$\beta = \frac{\lim\limits_{x \downarrow x'} E(y|x) - \lim\limits_{x \uparrow x'} E(y|x)}{\lim\limits_{x \downarrow x'} E(d|x) - \lim\limits_{x \uparrow x'} E(d|x)}$$



The most important assumption is local continuity. In other words, the potential outcomes for subjects just below the threshold is similar to those just above the threshold

# 2.ii. Regression discontinuity design

- **Advantages**:
  - Right around that threshold, you can plausibly say that firms on either side of that threshold are roughly the same.
  - Their characteristics are roughly the same.
  - If you can make that claim, this is pretty solid way to try to deal with the endogeneity criticism

# 2.ii. Regression discontinuity design

- **Disadvantages**:
  - You do not come across that many clean threshold types of events that you can really use in the contexts studied in corporate finance
  - Subjects on different sides of threshold, no matter how close, may not be comparable because of sorting (basically, due to manipulation with threshold)
  - There may be situations in which the treatment did not exist or groups for which the treatment does not apply, perhaps because of eligibility considerations. In this case, one can execute the RDD for this era or group in the hopes of showing no estimated treatment effect. This analysis could reinforce the assertion that the estimated effect is not due to a coincidental discontinuity or discontinuity in unobservables

# 2.iii. Matching methods

- Matching methods estimate the counterfactual outcomes of subjects by using the outcomes from a subsample of “similar” subjects from the other group (treatment or control)

- Main assumptions for this method to work $\big(y(0), y(1)\big) \perp d|\, X \quad 0 < \Pr(d = 1|X) < 1$

- The estimated or imputed potential outcomes for observation $i$ are

$$\hat{y}_i(0) = \begin{cases} y_i & if\ d_i = 0 \\ \frac{1}{M} \sum_{\{j\, \in indices\ of\ matches\ to\ i\}} y_j & if\ d_i = 1 \end{cases}$$

$$\hat{y}_i(1) = \begin{cases} \frac{1}{M} \sum_{\{j\, \in indices\ of\ matches\ to\ i\}} y_j & if\ d_i = 0 \\ y_i & if\ d_i = 1 \end{cases}$$

# 2.iii. Matching methods

- The estimated or imputed potential outcomes for observation $i$ are

$$\hat{y}_i(0) = \begin{cases} y_i & if\ d_i = 0 \\ \frac{1}{M} \sum_{\{j\ \in indices\ of\ matches\ to\ i\}} y_j & if\ d_i = 1 \end{cases}$$

$$\hat{y}_i(1) = \begin{cases} \frac{1}{M} \sum_{\{j\ \in indices\ of\ matches\ to\ i\}} y_j & if\ d_i = 0 \\ y_i & if\ d_i = 1 \end{cases}$$

- With estimates of the potential outcomes, the matching estimator of, for instance, is

$$\frac{1}{N} \sum_{i=1}^{N} [\hat{y}_i(1) - \ \hat{y}_i\ (0)]$$

# 2.iii. Matching methods

- **Advantages**:
  - Matching is less parametric that linear regression
  - Matching makes it easy to explicitly see the area of common support
  - Can mitigate asymptotic biases arising from endogeneity or self-selection

- **Disadvantages**:
  - That is great if you can identify what you think are the most plausible set of factors so you can match up firms and do a differences-in-differences sort of approach. But, almost by definition, these cannot be perfect because we are talking about correlated omitted variables that we cannot really identify
  - Many questions should be answered: How many matches should one use for each observation? Match with or without replacement? Which covariates to use?

# 2.iv. Instrumental variables

- Instrumental variables (IV) are a standard way to deal with endogeneity
- An instrument, $z$, is a variable that satisfies two conditions that we refer to as the relevance and exclusion conditions
  - The first condition requires that the partial correlation between the instrument and endogenous variable not be zero

$$x_k = \alpha_0 + \alpha_1 x_1 + \ \dots + \alpha_{k-1} x_{k-1} \ \ + \gamma z + v$$
$$test\colon \gamma = 0$$

  - The second condition is the exclusion condition that requires that $cov(z, u) = 0$

# 2.iv. Instrumental variables

- **Advantages**:
  - Instrumental variable is the common way to deal with endogeneity problem
  - Relevance is pretty easy to demonstrate most of the time with an instrumental variable

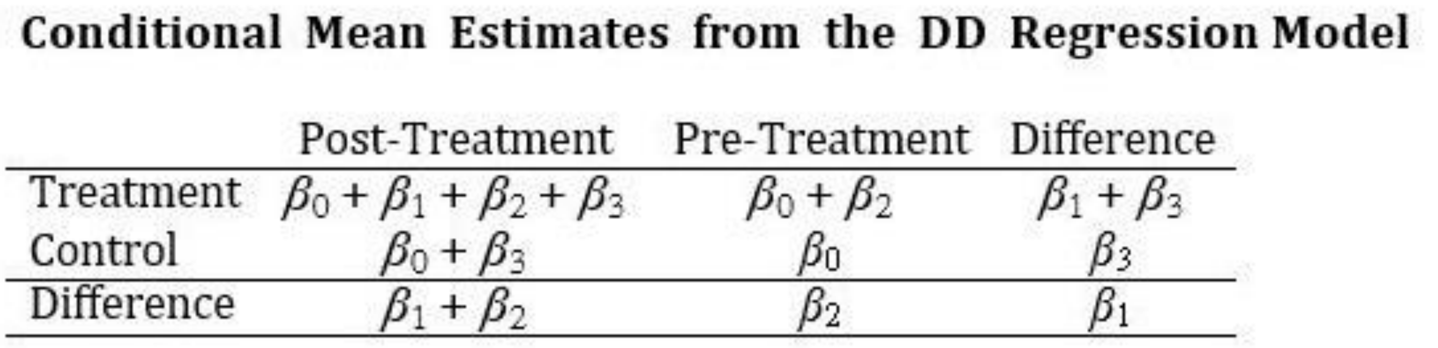
- **Disadvantages**:
  - Unfortunately, it is really easy to criticize instrumental variables on the exclusion part
  - You need to have compelling arguments relying on economic theory and a deep understanding of the relevant institutional details are the most important elements of justifying an instrument's validity
  - It is often the case that in corporate finance more than one regressor is endogenous. In this case, inference about all of the regression coefficients can be compromised if one can find instruments for only a subset of the endogenous variables
  - Faces tradeoff between external and internal validity, like all other strategies

# 2.v. Natural experiment/Difference-in-differences

- Difference-in-difference (DD) estimators are used to recover the treatment effects stemming from sharp changes in the economic environment, government policy, or institutional environment

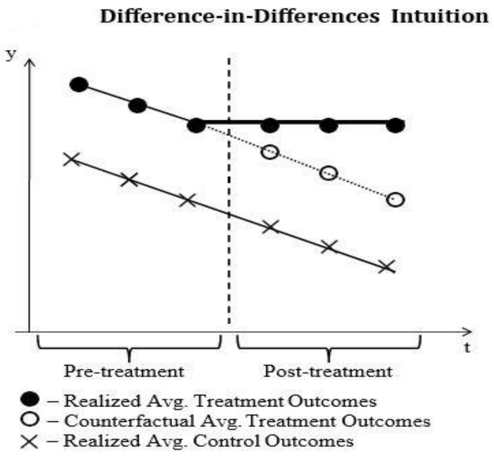
- These estimators usually go hand in hand with the natural or quasi- experiments created by these sharp changes. However, the exogenous variation created by natural experiments is much broader than any one estimation technique

$$y = \beta_0 + \beta_1 d * p + \beta_2 d + \beta_3 p + u$$

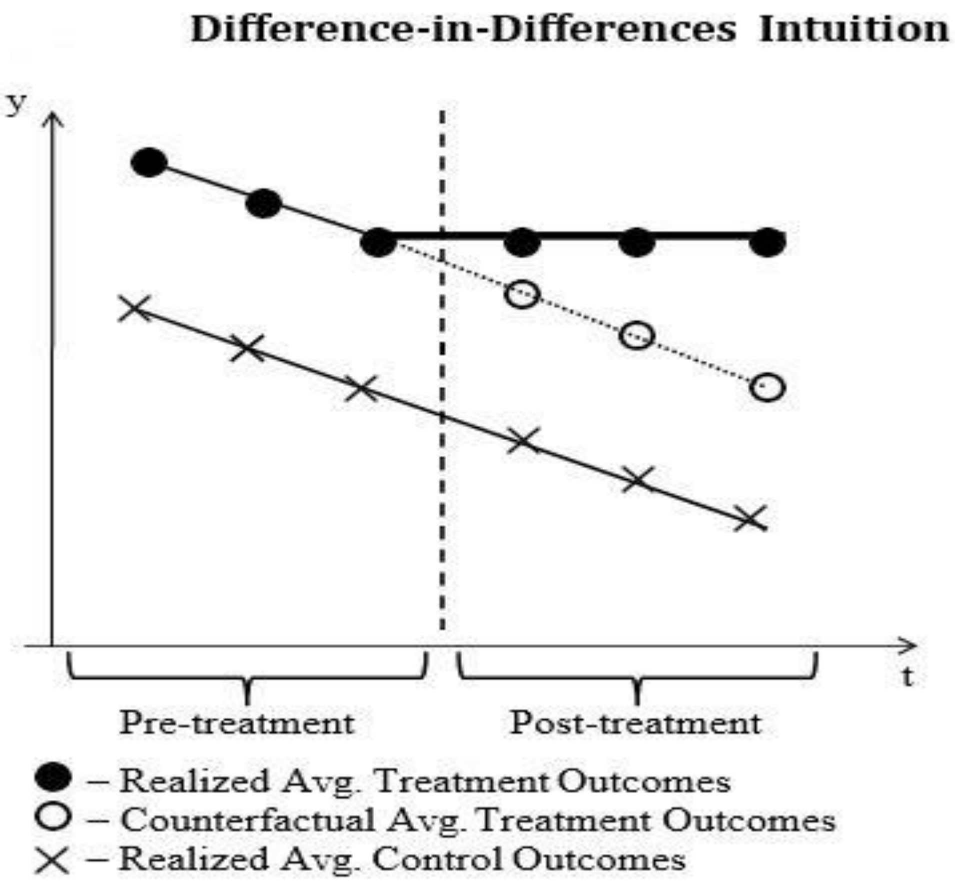


**Conditional Mean Estimates from the DD Regression Model**

| | Post-Treatment | Pre-Treatment | Difference |
|---|---|---|---|
| Treatment | $\beta_0 + \beta_1 + \beta_2 + \beta_3$ | $\beta_0 + \beta_2$ | $\beta_1 + \beta_3$ |
| Control | $\beta_0 + \beta_3$ | $\beta_0$ | $\beta_3$ |
| Difference | $\beta_1 + \beta_2$ | $\beta_2$ | $\beta_1$ |



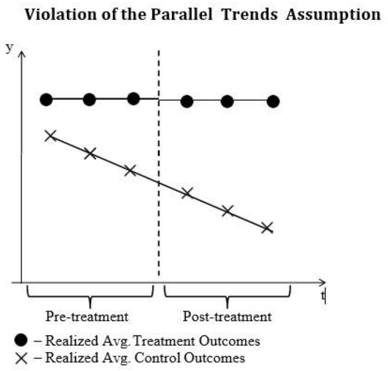
# 2.v. Natural experiment/Difference-in-differences



# 2.v. Natural experiment/Difference-in-differences

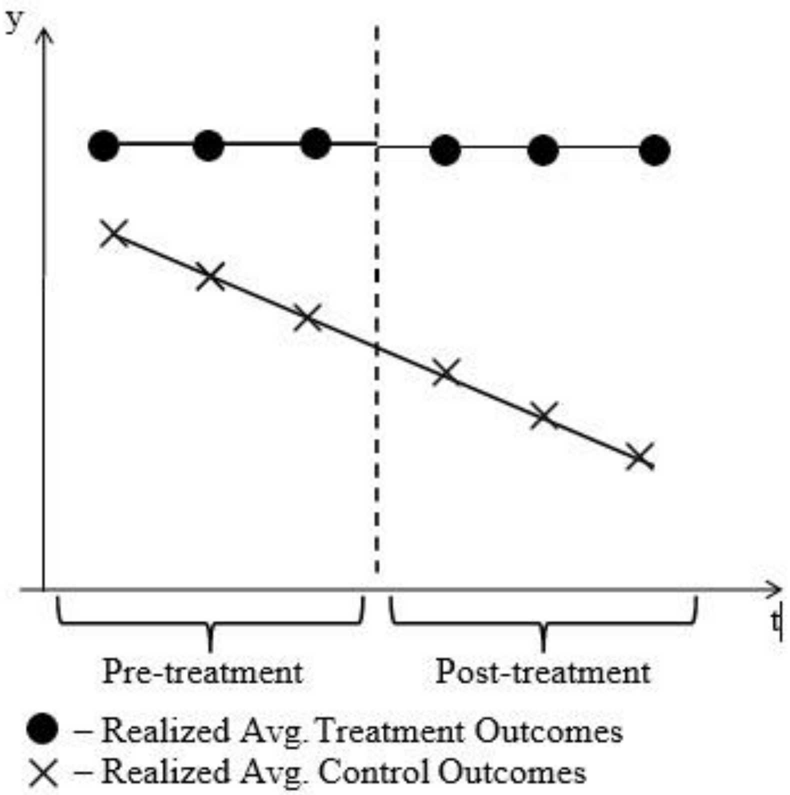
- **Advantages**:
  - What is great about natural experiments is that you can identify some plausibly exogenous event, and if it is truly exogenous, then you can do a diff-in-diff analysis and claim causality in a pretty credible fashion

- **Disadvantages**:
  - What you tend to see a little bit more often now that you saw before is that authors are starting with an experiment rather than the question. That is not quite the way you want to go about doing research
  - Narrows the scope of questions being asked
  - Importance of the parallel trends assumption



# 2.v. Natural experiment/Difference-in-differences

**Violation of the Parallel Trends Assumption**



# 3. General advice

1. Identify an interesting question
2. Develop hypotheses from first principles
3. Identify best experimental design
4. Let the data speak – discuss and explore all potential explanations for basic facts
5. Do not be afraid to point out limitations and caveats

# 4. Application to capital structure research

## Research approaches

i. Classical OLS or panel regressions

ii. Natural experiment

iii. Descriptive data analysis

iv. Longitudinal studies

v. Structural models

vi. Clinical studies

# 4. Application to capital structure research

- **Fundamental question**: What are the key determinants of capital structure decisions?

- **Possible determinants (market frictions)**:
  1. Taxes
  2. Bankruptcy costs
  3. Agency costs
  4. Asymmetric information

- Another whole segment of this literature is just getting at static model versus dynamic models

- Variety of different research approaches give us information in different aspects of the set information that we need to ultimately draw some conclusions about capital structure. However, they all have different limitations

# 4.i. Capital structure research: OLS

- You have got leverage on the left-hand side, and you have got a set of testable determinants on the right-hand side along with other control variables

- The usual estimators are panel regression techniques

- We end up identifying some effects that seem to matter, and we say that leverage is associated with profitability, growth opportunities as measured by market-to-book, and usually also firm size

- But when we think about relating that to capital structure theories, we find ourselves in this trap where it is consistent with multiple theories: static models, tradeoff models, and pecking order models

# 4.ii. Capital structure research: Natural experiment

**Do taxes affect financing decisions?**

- A good example is studies that have looked at staggered changes in the income tax rates across US states or across countries(Heider, & Ljungqvist (2013); Faccio, & Xu (2013))

$$\Delta D_{ijst} = \beta \Delta T_{s,t-1} + \delta \Delta X_{i,t-1} + \theta \Delta Z_{j,t-1} + \epsilon_{ijst}$$

$i = firm$

$j = industry$

$t = year$

$s = state$

- If done in a correct manner, this type of study is really useful for getting at this causal connection between taxes, and capital structure change

# 4.ii. Capital structure research: Natural experiment

**Do taxes affect financing decisions?**

- A good example is studies that have looked at staggered changes in the income tax rates across US states or across countries(Heider, & Ljungqvist (2013); Faccio, & Xu (2013))

$$\Delta D_{ijst} = \beta \Delta T_{s,t-1} + \delta \Delta X_{i,t-1} + \theta \Delta Z_{j,t-1} + \epsilon_{ijst}$$

$$i = firm$$
$$j = industry$$
$$t = year$$
$$s = state$$

- If done in a correct manner, this type of study is really useful for getting at this causal connection between taxes, and capital structure change

- But, in doing this sort of study, you have got to be clear what your goal is. This study is asking: “Do taxes have an impact on capital structure at the margin?” You are able to isolate the casual impact of taxes on leverage decisions. But it is a subset of the information

# 4.iii. Capital structure research: Descriptive data analysis

**Study observed discontinuity in financing behavior** (Denis, & McKeon (2012))

- Questions: Why did they do it? How does leverage subsequently evolve?
- The authors start with the left-hand side variable, try to observe some major discontinuity in the financing behavior of the firm, and then back out what seems to be driving that change in the financing behavior
- Results: Firms borrow primarily to meet investment needs. Subsequent rebalancing to target is neither rapid nor the results of pro-active attempts to return to target
- Implications: Managing toward stationary target is not a first order concern. Observed capital structures appear to be driven more by investment-related capital needs
- But, when we do a study like this, we in no way can claim some causation

# 4.iii. Capital structure research: Descriptive data analysis

**Study within-firm variation leverage ratios** (DeAngelo, & Roll (2014))

- Main findings: Substantial instability in leverage ratios of individual firms. Extended periods of stability are largely limited to low leverage periods. Strong association between departures from leverage stability and company expansion (case-base evidence)
- Implications: Credible theories of capital structure must be able to explain substantial time-series variation. Over a fairly wide range, leverage per se is of second-order importance to firm valuation. Main determinants of observed leverage ratios are factors other than those traditionally associated with leverage targets
- It is very informative to subsequent theory because now any theory that comes along that purports to be modelling how capital structure is actually chosen has to deal with the facts that are out there

# 4.iv. Capital structure research: Longitudinal study

**Study capital structures of US non-financial firms over the last century** (Graham, Leary, & Roberts (2014))

- Questions: How (if at all) have capital structures changed? Do existing models account for these changes? If not, what forces drive variation in financial policy?

- Main findings: Large increases in leverage over time. Cash holdings decline over time concomitantly with secular increase in leverage. Firm characteristics, on average, do not change in a way consistent with greater debt capacity. Negative association between corporate and government leverage (crowding out)

- They did not establish any causation between variables and leverage. Even with their last result with the government crowding out, the most they are saying is that there is some hint of that in the data. But, it is useful in terms of subsequent studies

# 4.iv. Capital structure research: Longitudinal study

**Study capital structure decisions of US firms during the period 1905-1924** (Bargeron, Denis, & Lehn (2014))

- Introduction of corporate and personal taxes
- WWI – large, transitory shock to investment

- Main findings: Little evidence shocks to corporate and individual taxes have a meaningful impact on observed leverage ratios. Strong evidence that changes in leverage are positively related to investment and negatively related to cash flows
- Challenge: Sample period is characterized by other shocks: Panic of 1907; Creation of Federal Reserve – 1913; Post-war depression – 1920-21
- When you are doing any sort of longitudinal study, if you are going to focus in on specific shocks that interest you, you have to be aware of the fact that they are not the only shocks that are going on during this period of time

# 5. Conclusions

1. Identification problems are pervasive in empirical corporate finance
2. Econometrics can help, but can never solve the problem
3. Natural experiments are very useful, but can limit the types of questions that can be studied in a way that is harmful to the pursuit of knowledge
4. Other methodologies can be useful even if they are not able to provide complete identification