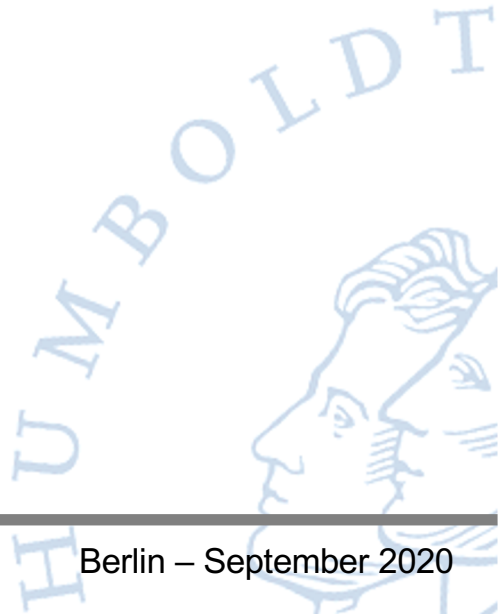
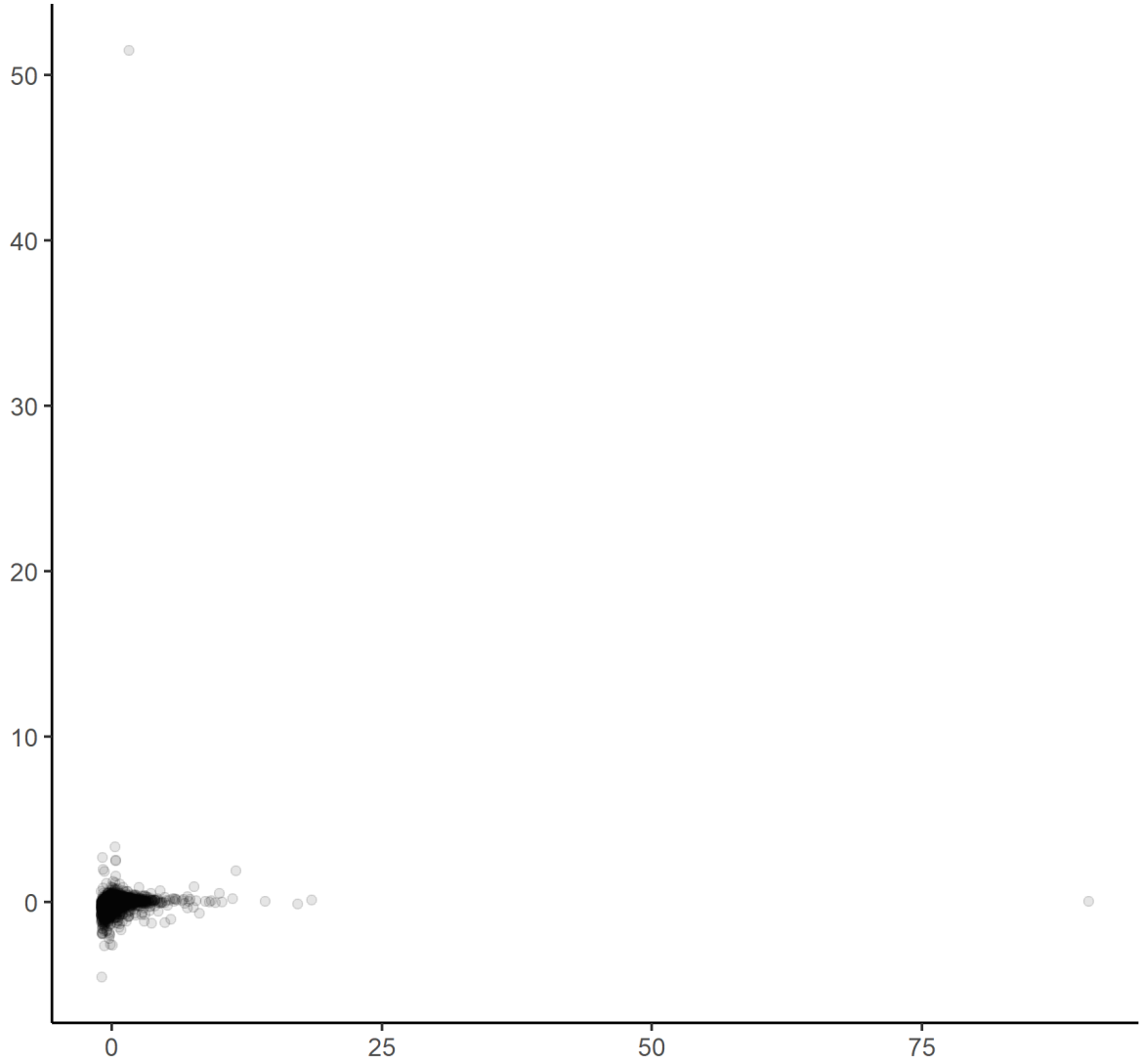
Wednesday, 09.09.2020

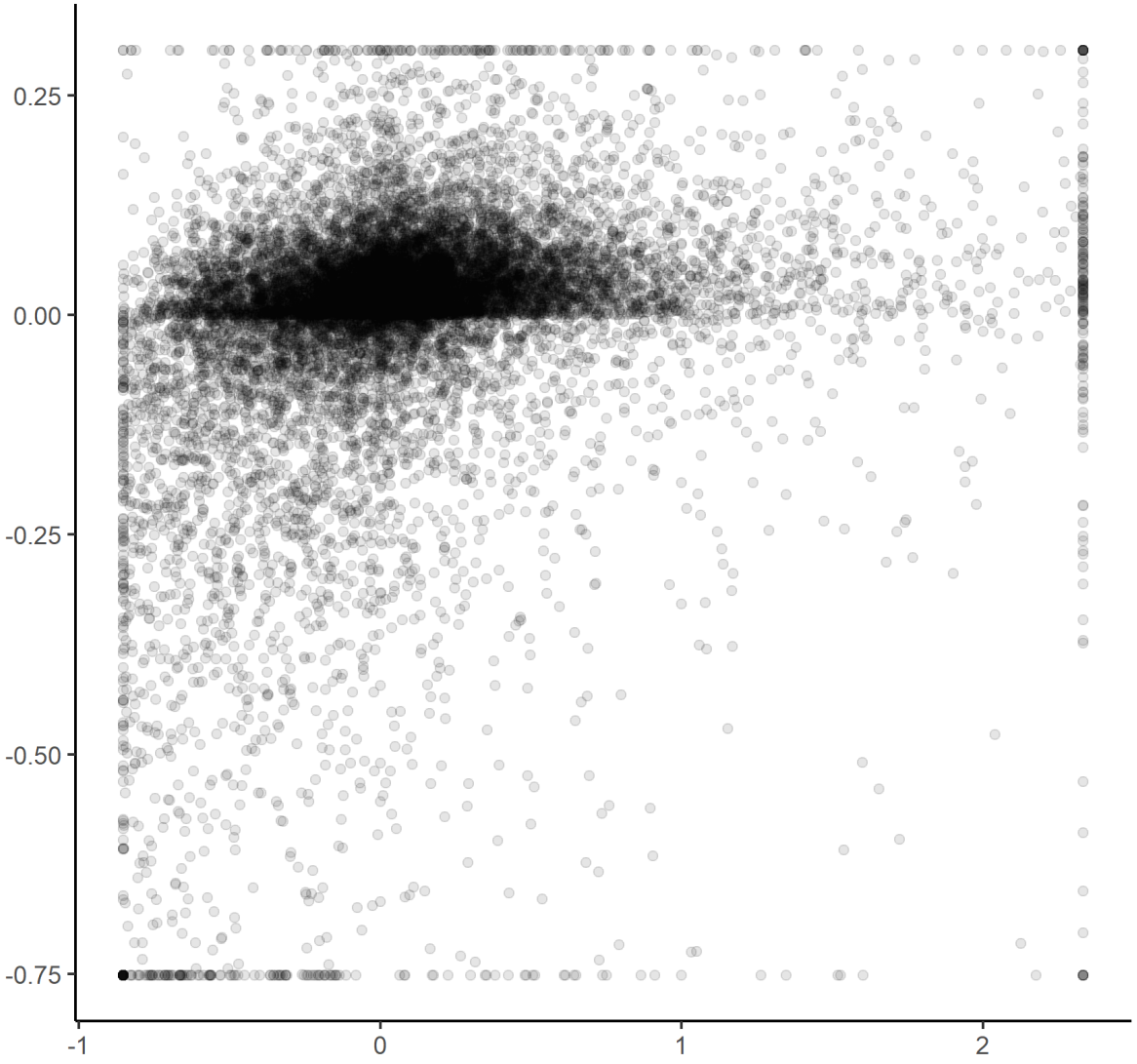
- Paper: Breuer and Windisch (2019)
- Research design
  - Identification strategies
  - From the research question to the research setting
- Paper: Gassen and Muhn (2018)
- Execution
  - Data wrangling
  - Exploration
  - Modelling and testing
- Class project
  - Exploratory data analysis
  - Discussion of potential research questions



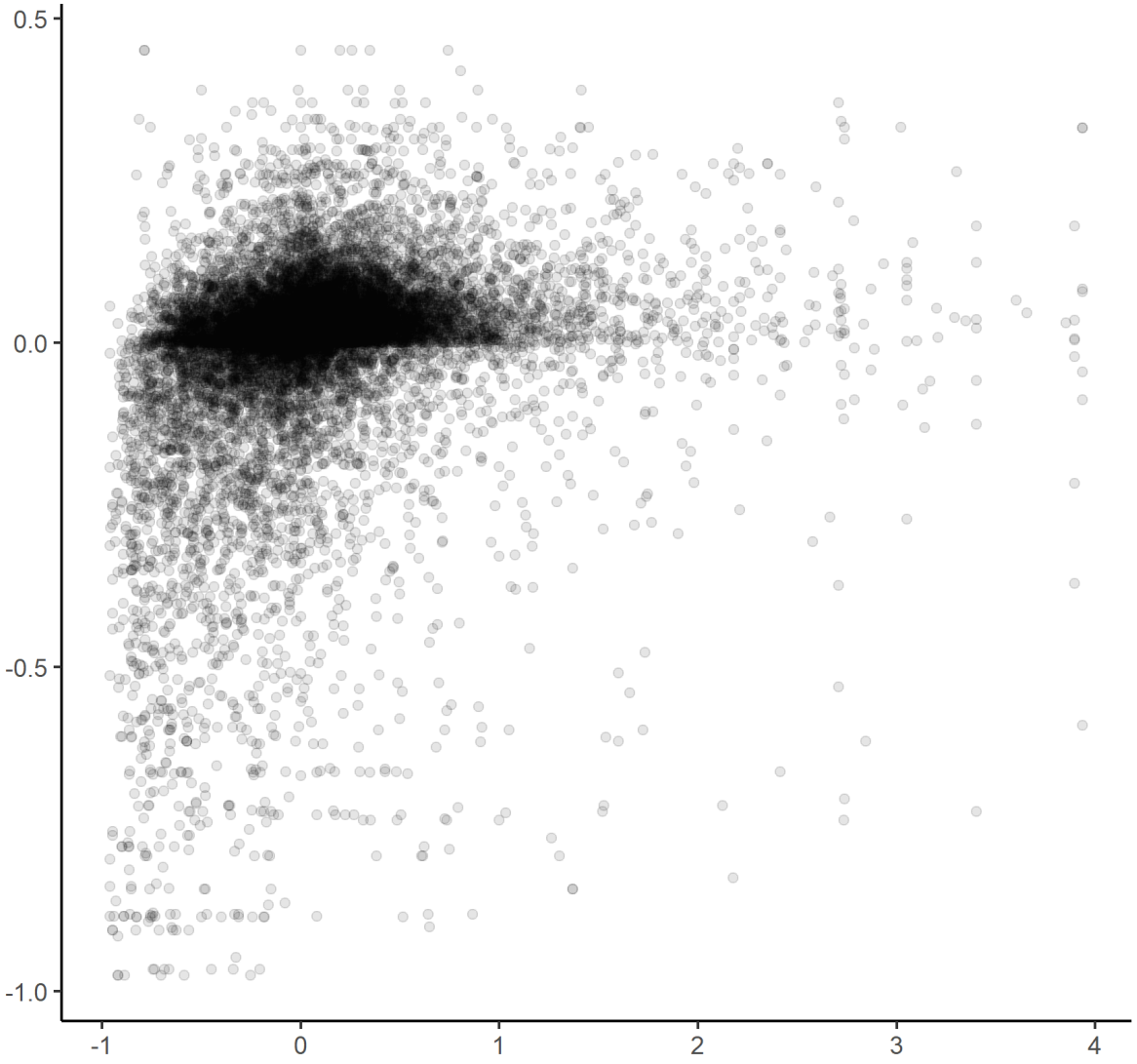
Which association is this?



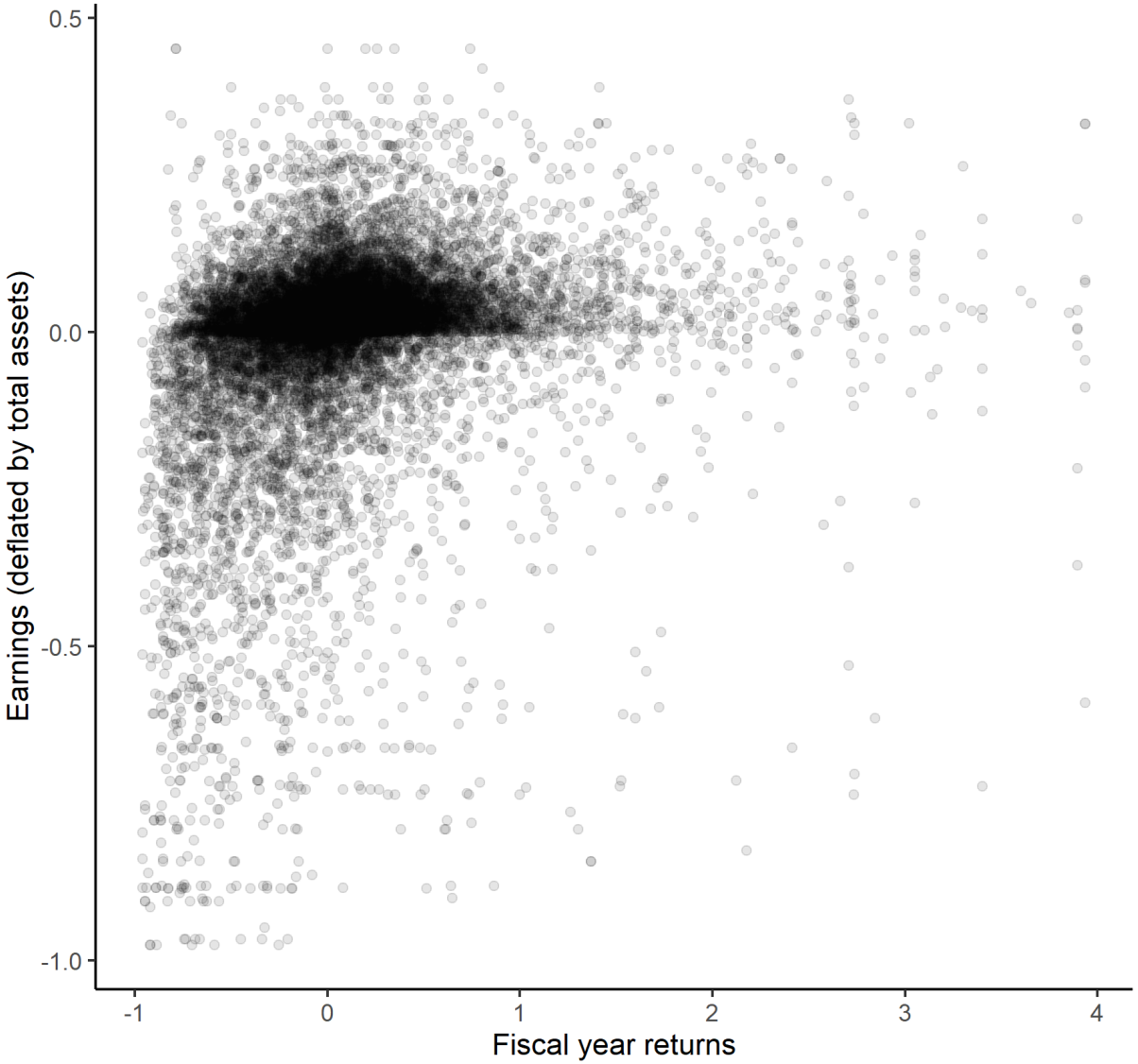
Now better?



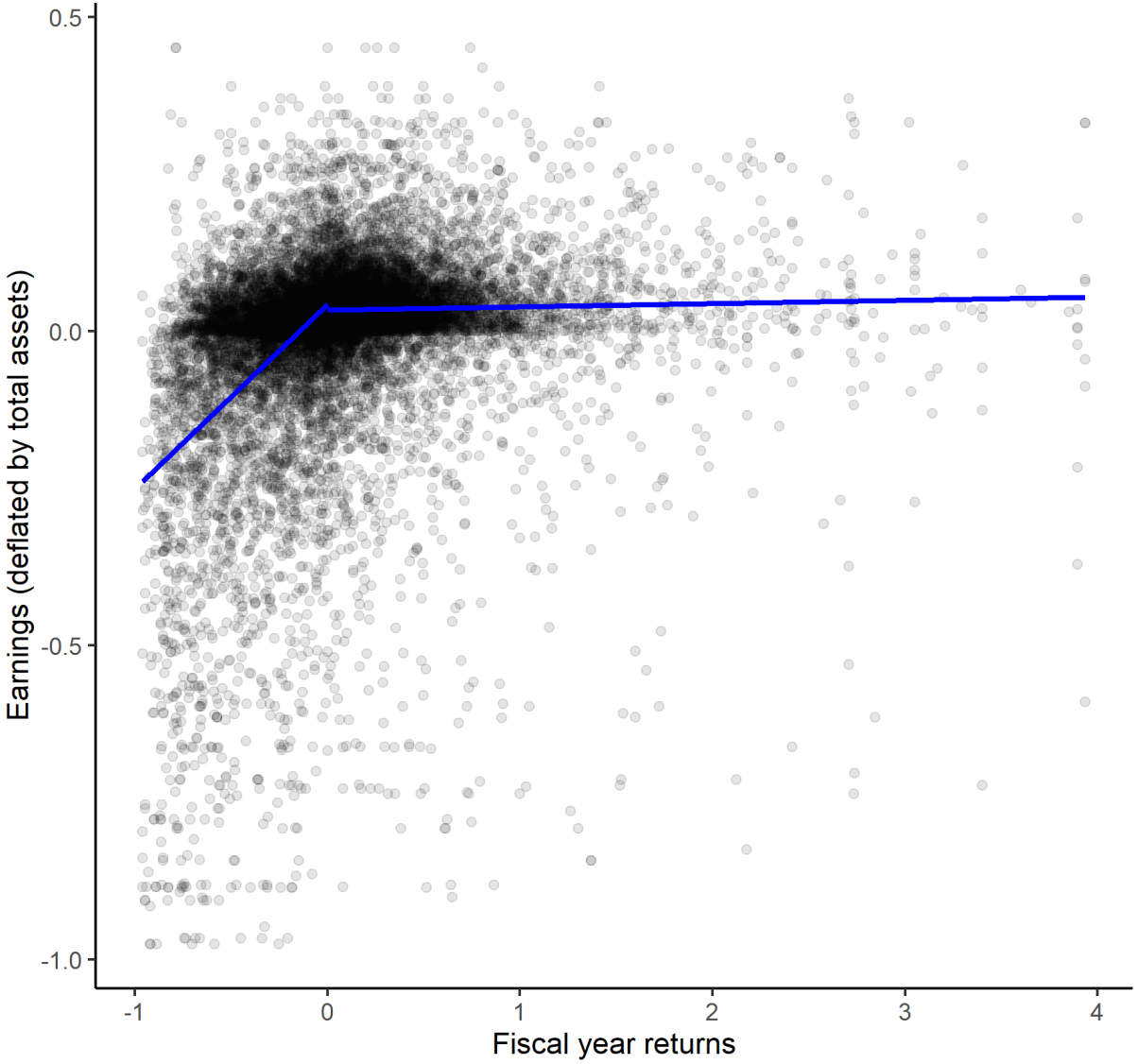
And what about now? What is the difference to the last plot?



Revealed…

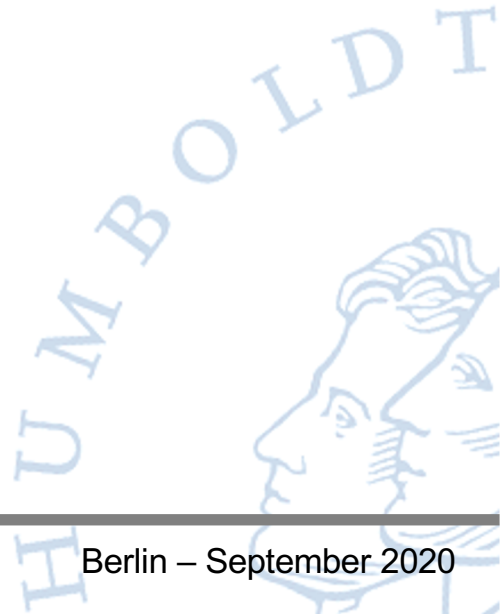

0.5
0.0
-0.5
-1.0
Earnings (deflated by total assets)
-1
0
1
2
3
4
Fiscal year returns

Behold: The Basu conservatism regression (on German data)



Wednesday, 09.09.2020

- Paper: Breuer and Windisch (2019)
- Research design
  - Identification strategies
  - From the research question to the research setting
- Paper: Gassen and Muhn (2018)
- Execution
  - Data wrangling
  - Exploration
  - Modelling and testing
- Class project
  - Exploratory data analysis
  - Discussion of potential research questions



The problem of endogeneity

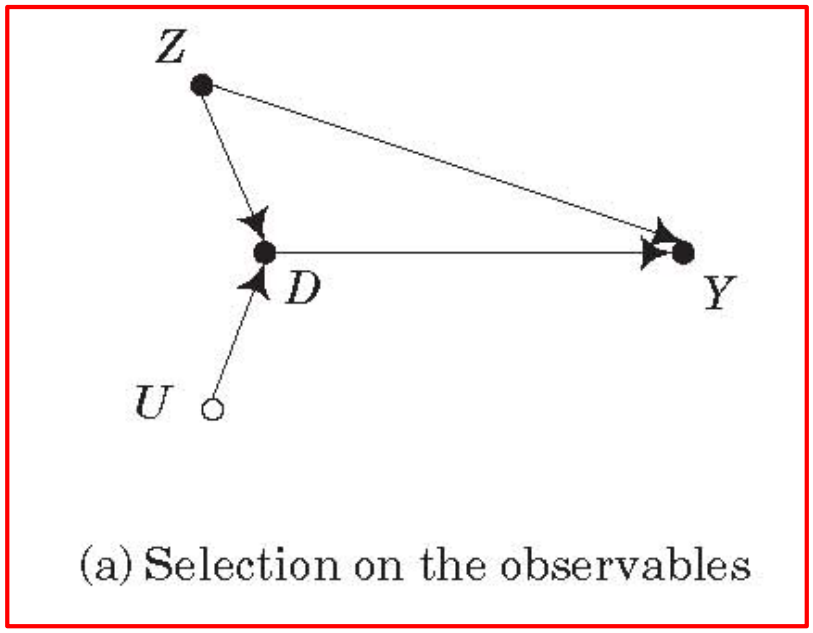
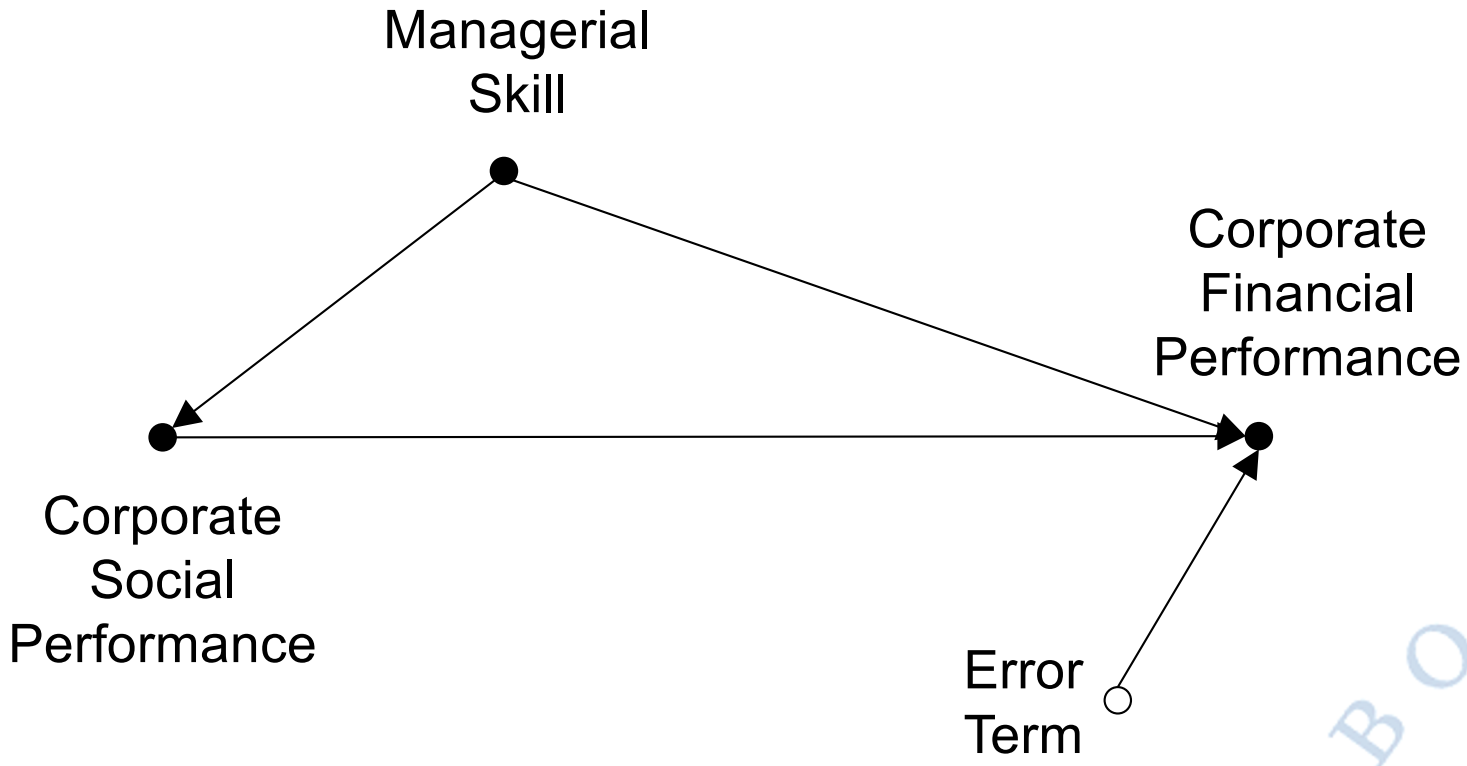


Figure 3.9: Causal diagrams for the terminology from econometric modeling of treatment selection.

Morgan and Winship, 2007, page 81

An artificial CSR example



The approach yields consistent estimates if…

- An regression approach like

$$Y = \alpha + \delta D + \beta X + \varepsilon^{*}$$

- yields are consistent estimate for δ if
  - a fully flexible parametrization of X is used and
  - the causal effect of D does not vary with X and
  - D is uncorrelated to

$$\upsilon^{0} + D\left(\upsilon^{1} - \upsilon^{0}\right)$$

for every unique sub-sample defined by X

$$\upsilon^{0} \equiv Y^{0} - E\left[Y^{0}\right]$$

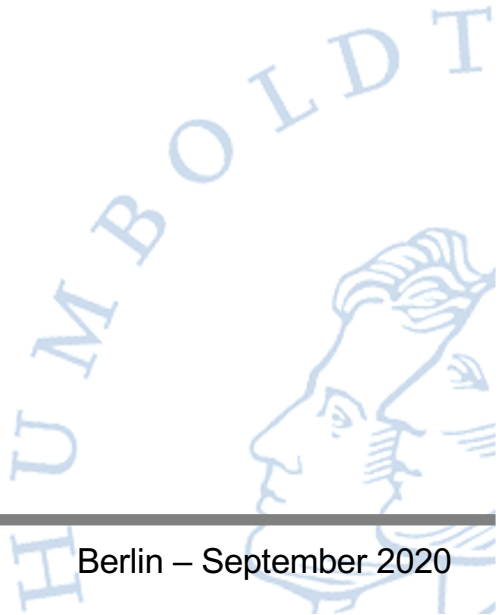
$$\upsilon^{1} \equiv Y^{1} - E\left[Y^{1}\right]$$

- This a sufficient, not a necessary condition

Identification by Stratification: The idea

- Matching is building on the conditioning concept
- The analysis is conditioned on a set of variables (S) so the the following two assumptions hold

$$E\left[Y^1 \middle| D=1, S\right] = E\left[Y^1 \middle| D=0, S\right]$$
$$E\left[Y^0 \middle| D=1, S\right] = E\left[Y^0 \middle| D=0, S\right]$$

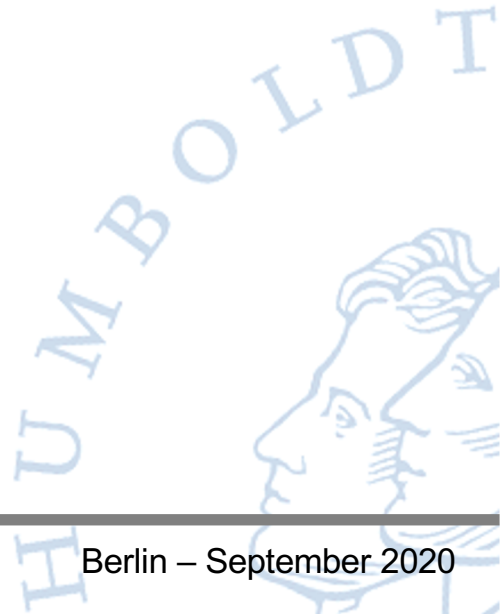


A very common problem

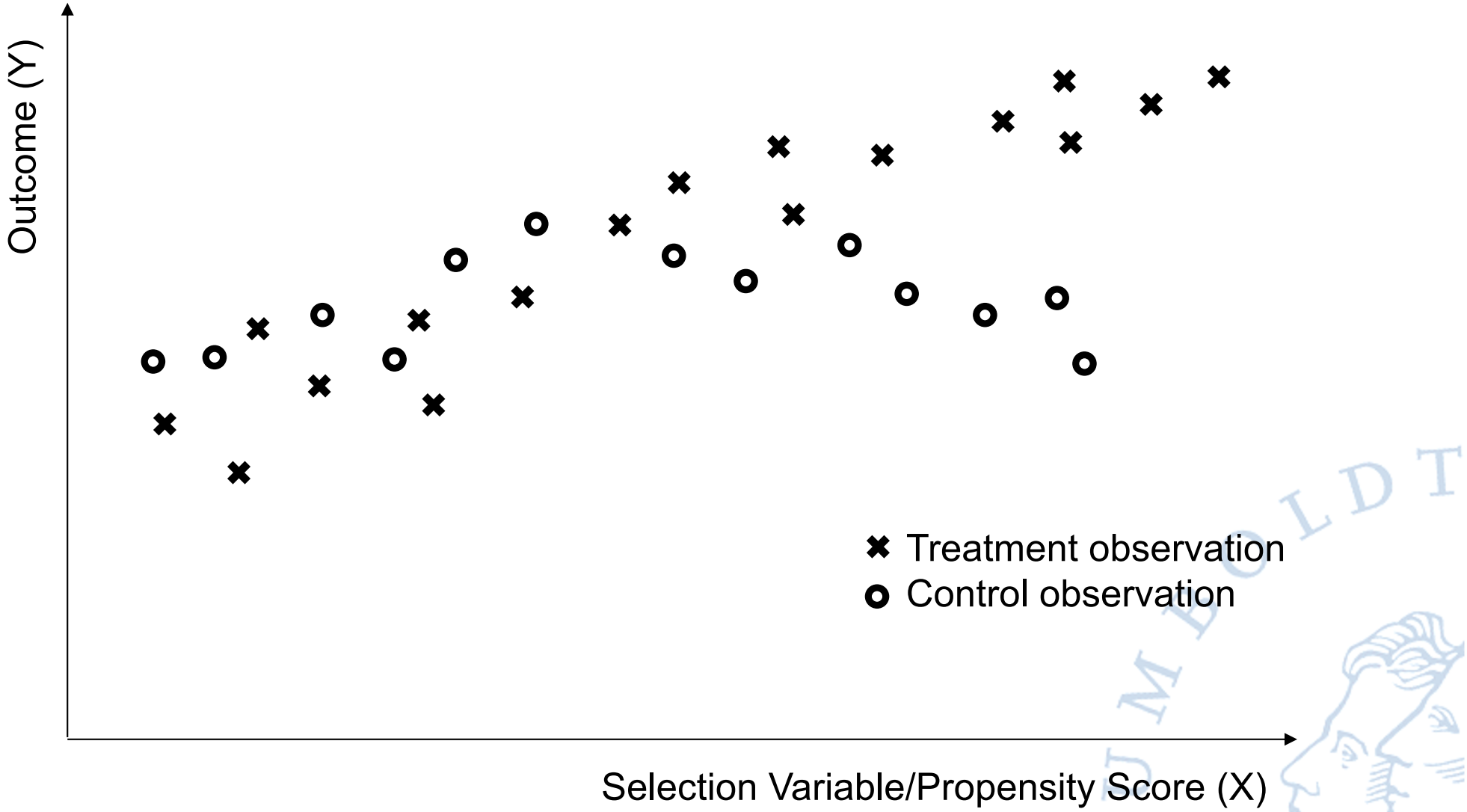
A	B	D	Y
6.3518	10.8536	1	33.954
6.2882	8.7145	1	29.5618
1.5309	7.3589	0	8.1728
2.0941	3.1708	0	3.5488
5.5448	9.5072	0	16.3599
3.0961	6.8125	1	17.6984
5.4246	10.435	1	27.8291
5.9696	9.206	1	27.263
4.9298	9.1949	1	25.0199
4.26	6.3047	0	10.7836
-0.3507	10.1243	0	9.2277
5.5437	13.0252	1	33.3831
1.7486	9.8714	0	13.0454
5.5004	9.3891	1	27.2569

If the data which form S define multiple strata (like A,B to the left), than matching by stratification becomes unfeasible because of sparseness.

Matching by weighting on propensity scores is the remedy!



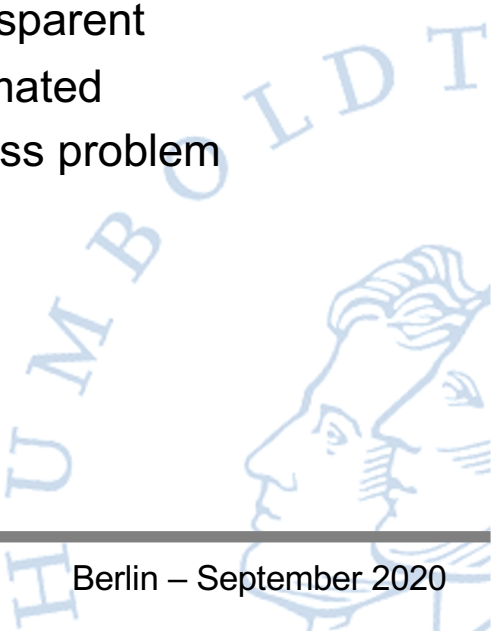
The idea of matching...



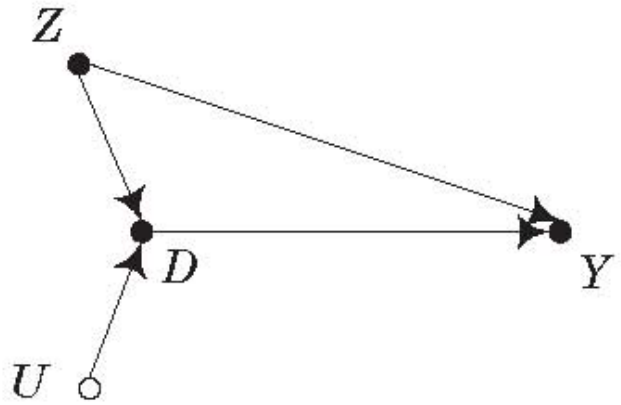
Wrapping things up: Regression and matching compared

- Regression approaches:
  - + Easy to understand for the average reader
  - + Easy to implement
  - + Fully saturated models (if possible) are great to understand the causal effect
  - - Basic TE is not easy to link to ATE, ATT and ATC, weighting is necessary and needs propensity scores (see below)
  - - Areas without common support are easily overlooked

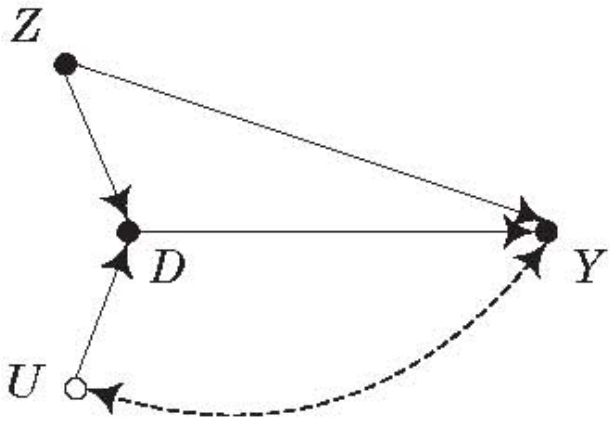
- Matching approaches:
  - + Make the problems of the counterfactual analysis much more transparent
  - + Allow for explicit control of the way counterfactual effects are estimated
  - + Propensity scores are an efficient way to circumvent the sparseness problem
  - - Harder to implement and to communicate
  - - Tough to estimate a good choice model
  - - Standard errors of the second stage are a problem



The problem of endogeneity



(a) Selection on the observables



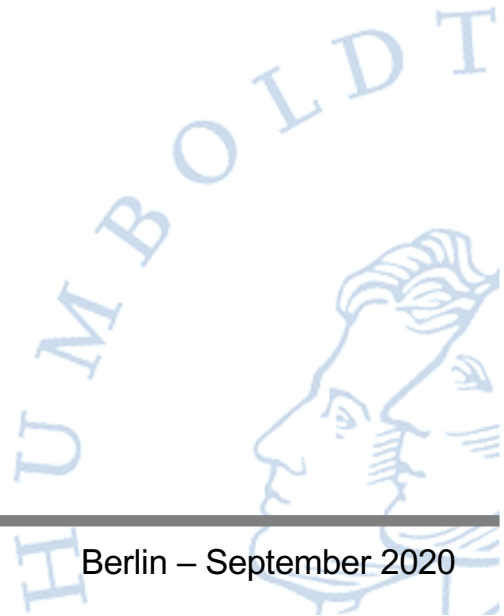
(b) Selection on the unobservables

Figure 3.9: Causal diagrams for the terminology from econometric modeling of treatment selection.

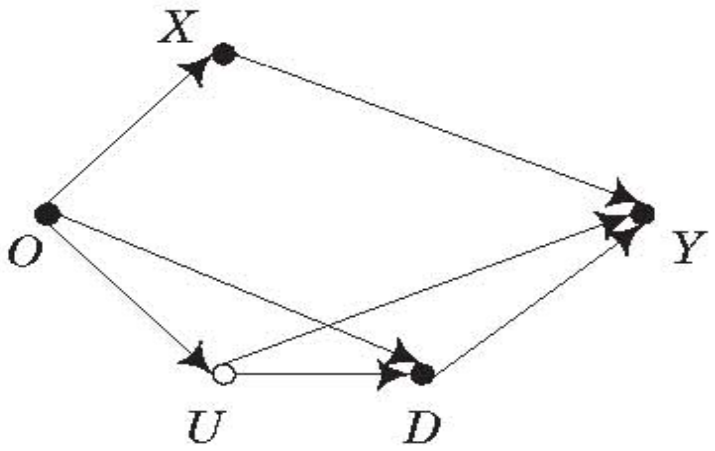
Morgan and Winship, 2007, page 81

Four strategies

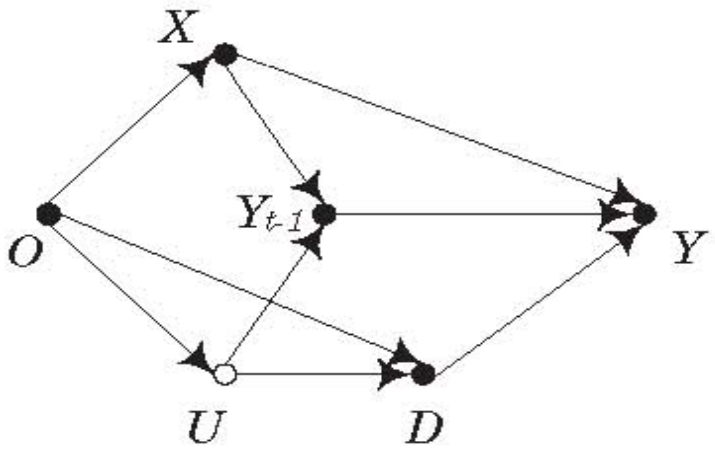
- Using the time structure of the data to close back doors
- Using instrumental variables
- Using mechanisms to open a front door
- Using two-stage approaches to estimate the effect of unobservable variables



Using the time structure of the setting…



(a) The identification puzzle

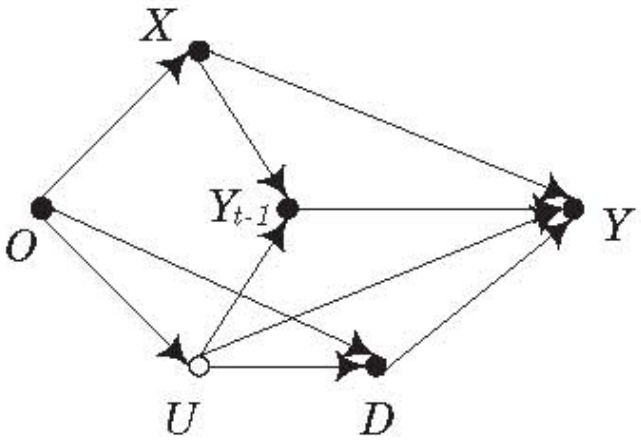


(b) Coleman's solution

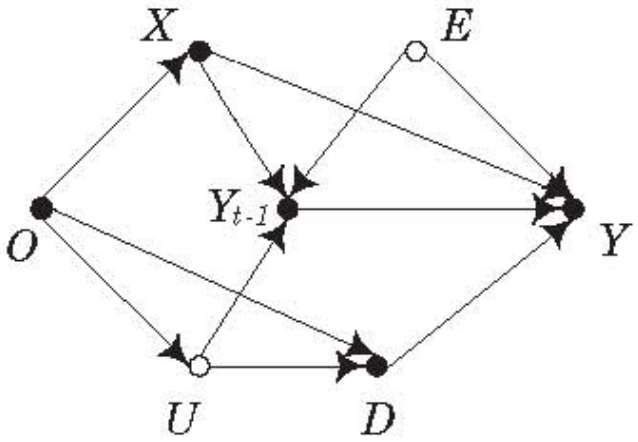
Figure 6.1: Coleman's strategy for the identification of the causal effect of Catholic schooling on learning.

Morgan and Winship, 2007, page 179

… and its inherent problems



(a) Criticism 1: The lagged Y is an imperfect screen

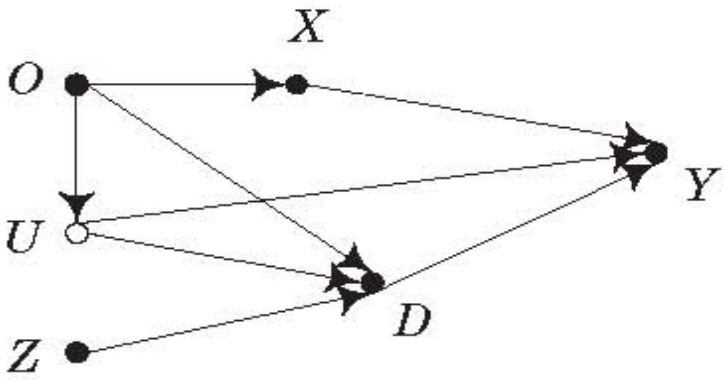


(b) Criticism 2: The lagged Y is a collider, and conditioning on it does not block a back-door path through a variable such as E

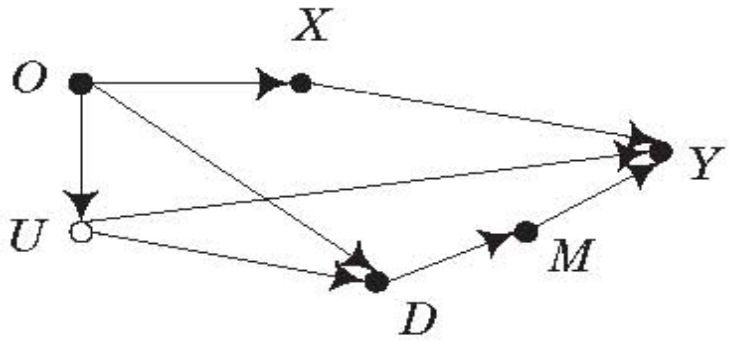
Figure 6.2: Criticism of Coleman's estimates of the effect of Catholic schooling on learning.

Morgan and Winship, 2007, page 181

Alternative Approaches: IVs and mechanisms



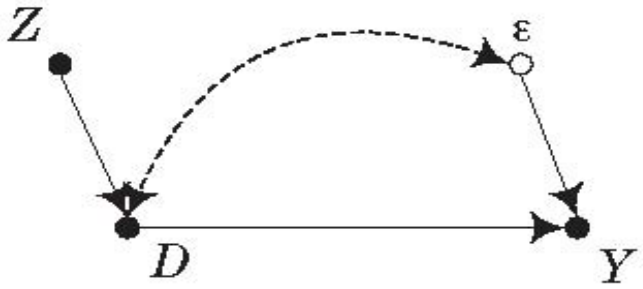
(a) Alternative solution 1: An instrumental variable



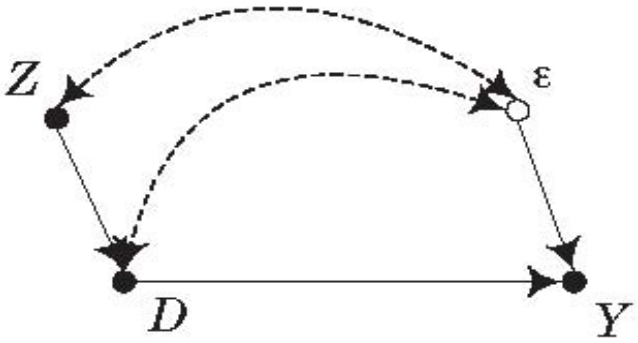
(b) Alternative solution 2: A complete mechanism

Figure 6.3: Alternative identification strategies for the causal effect of Catholic schooling on learning.

Morgan and Winship, 2007, page 182



(a) Z is a valid instrumental variable for D



(b) Z is not a valid instrumental variable for D

Figure 7.1: Causal diagrams for Z as a potential IV.

Morgan and Winship, 2007, page 188

IV in real-life research

Journal of Accounting and Economics 49 (2010) 186–205

Contents lists available at ScienceDirect

Journal of Accounting and Economics

journal homepage: www.elsevier.com/locate/jae

On the use of instrumental variables in accounting research ☆

David F. Larcker [a,*], Tjomme O. Rusticus [b]

[a] Stanford Graduate School of Business, 518 Memorial Way, Stanford, CA 94305-5015, USA
[b] Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Room 6216, Evanston, IL 60208, USA

ARTICLE INFO

Article history:
Received 14 April 2005
Received in revised form
6 May 2009
Accepted 18 November 2009
Available online 26 November 2009

JEL classification:
C30
G30
M41
M43

Keywords:
Endogeneity
Instrumental variables
Disclosure
Cost of capital

ABSTRACT

Instrumental variable (IV) methods are commonly used in accounting research (e.g., earnings management, corporate governance, executive compensation, and disclosure research) when the regressor variables are endogenous. While IV estimation is the standard textbook solution to mitigating endogeneity problems, the appropriateness of IV methods in typical accounting research settings is not obvious. Drawing on recent advances in statistics and econometrics, we identify conditions under which IV methods are preferred to OLS estimates and propose a series of tests for research studies employing IV methods. We illustrate these ideas by examining the relation between corporate disclosure and the cost of capital.

Table 1
Accounting research articles that use instrumental variable methods.

The sample is based on an electronic search for the terms "2SLS," "3SLS," "instrumental variable," and "endogeneity" for papers published in *Journal of Accounting Research* (JAR), *Journal of Accounting and Economics* (JAE), or *The Accounting Review* (TAR) during the time period from 1995 to 2005.

Earnings Management (EM)
Anderson et al. (JAE, 2004)
Barton (TAR, 2001)
Beatty et al. (JAR, 1995)
Darrough and Rangan (JAR, 2005)
DeFond et al. (JAR, 2002)
D'Souza (TAR, 1998)
Haw et al. (JAR, 2004)
Hope (JAR, 2003)
Hunt et al. (JAE, 1996)
Kang and Sivaramakrishnan (JAR, 1995)

Disclosure (DIS)
Barton and Waymire (JAE, 2004)
Bushee et al. (JAE, 2003)
Kasznik (JAR, 1999)
Lang et al. (JAR, 2004)
Leuz and Verrecchia (JAR, 2000)

Auditing (AUD)
Copley et al. (JAR, 1995)
Khurana and Raman (TAR, 2004)
Weber and Willenborg (JAR, 2003)
Whisenant et al. (JAR, 2003)
Willenborg (JAR, 1999)

Other Financial Accounting (OTH)
Aboody (JAE, 1996)
Aboody et al. (TAR, 2004)
Ball and Shivakumar (JAE, 2005)
Barth et al. (JAR, 2001)
Beaver et al. (JAE, 1997)
Bell et al. (TAR, 2002)
Callen et al. (JAE 2005)
Frank (JAR, 2002)
Kothari and Zimmerman (JAE, 1995)
Lev and Sougiannis (JAE, 1996)
Loudder et al. (TAR, 1996)
Phillips (TAR, 2003)
Rajgopal et al. (JAR, 2003)
Shi (JAE, 2003)

Management Accounting/Compensation (MAC)
Abernethy et al. (TAR, 2004)
Hanlon et al. (JAE, 2003)
Holthausen et al. (JAE, 1995)
Keating (JAE, 1997)
Murphy (JAE, 2000)
Nagar (TAR, 2002)
Rajgopal and Shevlin (JAE, 2002)
Roulstone (JAR, 2003)

(Larcker and Rusticus, JAE 2010: 188)

Table 2
Descriptive statistics for the accounting research articles that use instrumental variable methods.
The sample is based on an electronic search for the terms "2SLS," "3SLS," "instrumental variable," and "endogeneity" for papers published in *Journal of Accounting Research*, *Journal of Accounting and Economics*, or *The Accounting Review* during the time period from 1995 to 2005.

Panel A. Type of application	EM	DIS	AUD	OTH	MAC	Total
Recursive models						
Standard two-stage-least-squares	3	1	1	8	2	15
First stage is a probit model (Heckman)	0	2	2	2	1	7
Non-recursive models						
Simultaneous equations models	7	2	2	4	5	20

Panel B. Features of the application	**Two stage** # (%)	**Heckman** # (%)	**Simultaneous** # (%)	**Total** # (%)
Importance of instrumental variables				
Instrumental variables are used in main test	8 (53%)	3 (43%)	12 (60%)	23 (55%)
Instrumental variables are used as robustness test	7 (47%)	4 (57%)	8 (40%)	19 (45%)
Explanation in the paper				
Extensive discussion of model/method	5 (33%)	5 (71%)	16 (80%)	26 (62%)
Discussion of instruments	10 (67%)	1 (14%)	12 (60%)	23 (55%)
Justification of instruments	6 (40%)	0 (0%)	3 (15%)	9 (21%)
Reported empirical work				
Report first-stage coefficients	6 (40%)	5 (71%)	2 (10%)	13 (31%)
Report first-stage explanatory power	5 (33%)	4 (57%)	6 (30%)	15 (36%)
Report explanatory power of instruments	3 (20%)	0 (0%)	1 (5%)	4 (10%)
Standard regression (e.g., OLS) also reported	12 (80%)	6 (86%)	12 (60%)	30 (71%)
Reported tests				
Hausman test (or functional equivalent)	6 (40%)	4 (57%)	15 (75%)	25 (60%)
Over-identifying restrictions test	1 (7%)	0 (0%)	3 (15%)	4 (10%)
Number of papers by category	15	7	20	42

(Larcker and Rusticus, JAE 2010: 188)

The econometrics of IV

(1) $$y = \beta x + \sum_{k=1}^{K} \gamma_k x_k + u$$

(2) $$x = \delta y + \sum_{i=1}^{J} \theta_j x_j + \varepsilon$$

Residualizing (1) and (2) with respect to x_k and x_j

(3) $$y = \beta x + u$$

$$\operatorname{plim} b_{OLS} = \beta + \frac{\operatorname{cov}(x,u)}{\operatorname{var}(x)} = \beta + \frac{\sigma_u}{\sigma_x} \operatorname{corr}(x,u)$$

Using an instrument z (which should be correlated with x but uncorrelated with u):

$$\operatorname{plim} b_{IV} = \beta + \frac{\operatorname{cov}(z,u)}{\operatorname{cov}(x,z)} = \beta + \frac{\sigma_u}{\sigma_x} \frac{\operatorname{corr}(z,u)}{\operatorname{corr}(x,z)}$$

(clipped from Larcker and Rusticus, JAE 2010: 189f.)

How robust is IV? A monte carlo simulation study

- 1,000 observations, 100 rounds for each variable vector, set-up:

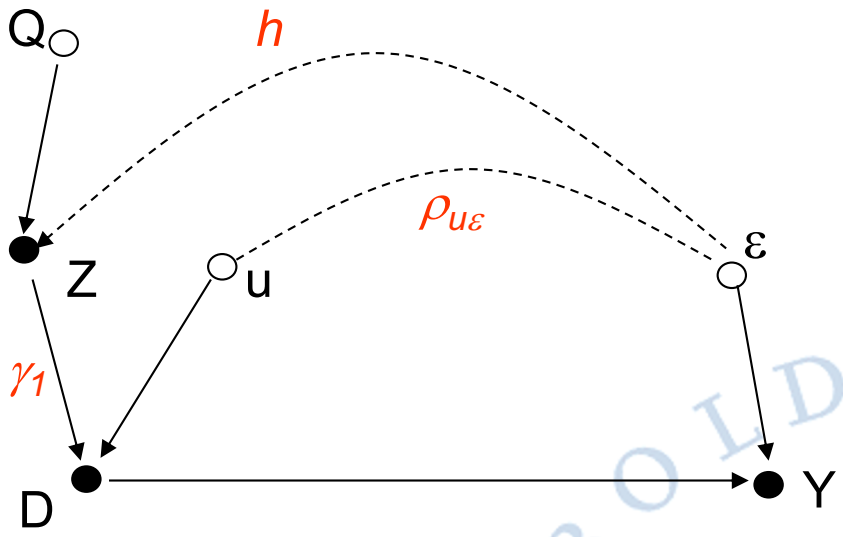
$$Y = \beta_0 + \beta_1 D + \varepsilon$$

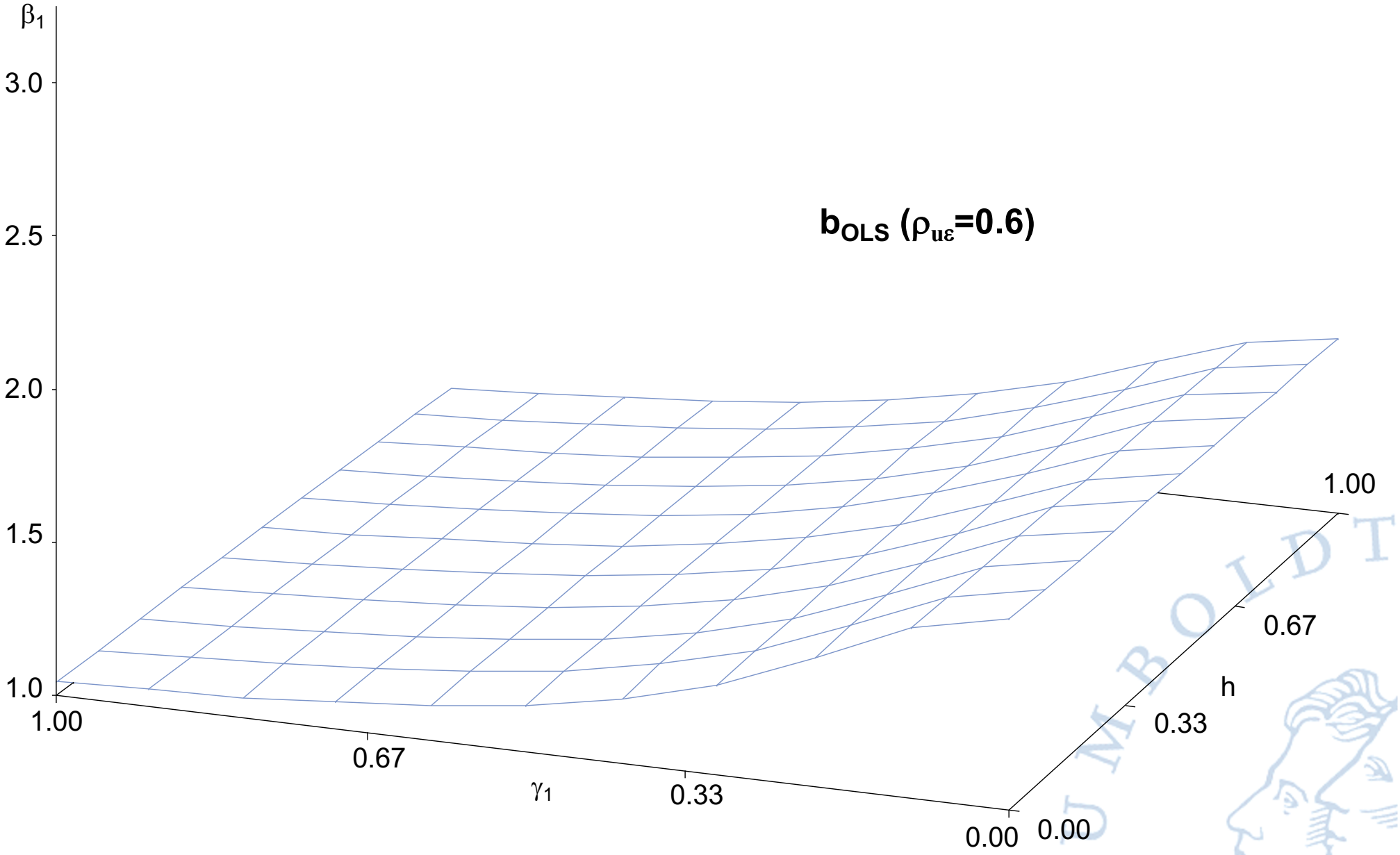
$$D = \gamma_0 + \boxed{\gamma_1} Z + u$$

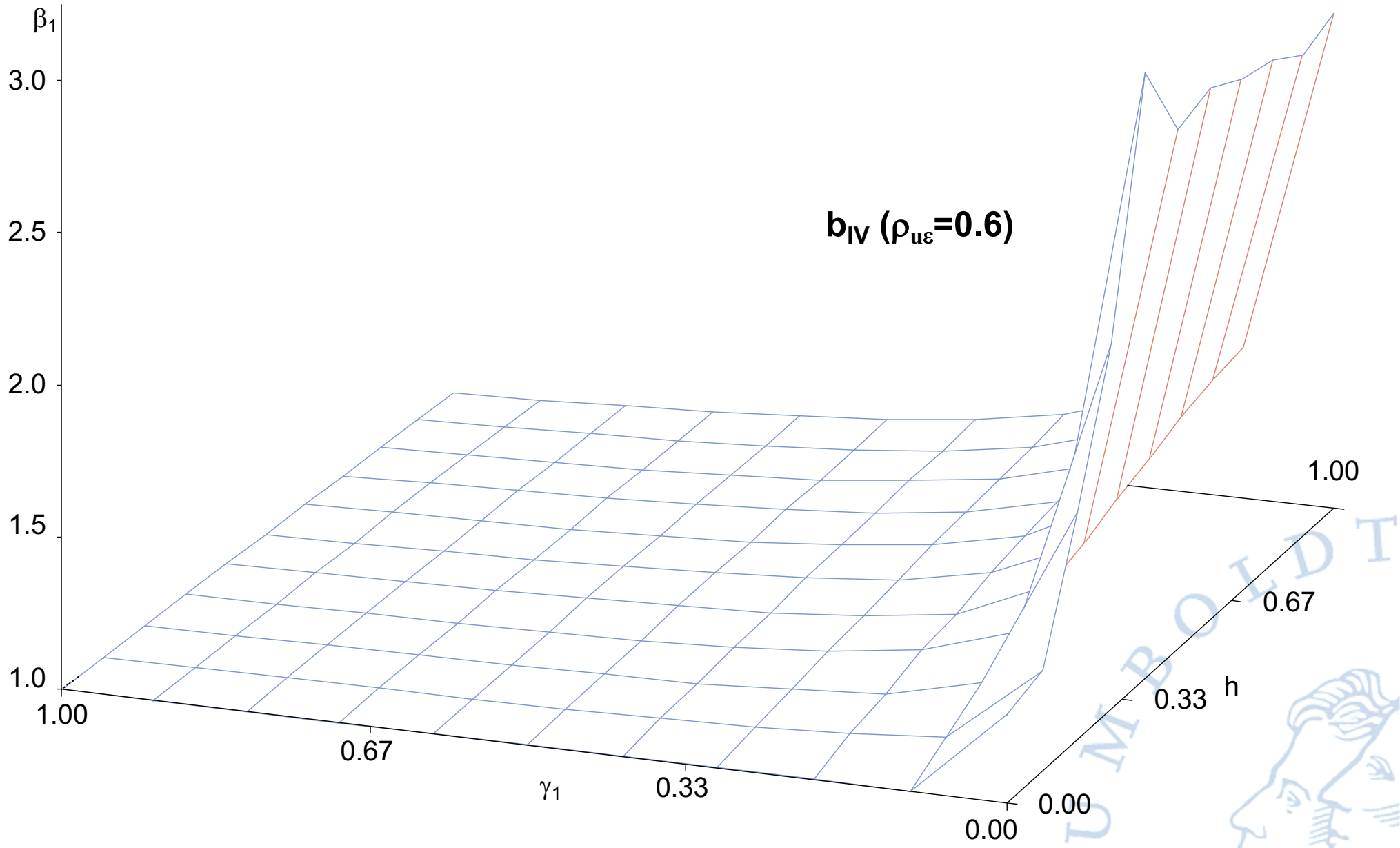
$$Z = Q + \boxed{h}\varepsilon$$

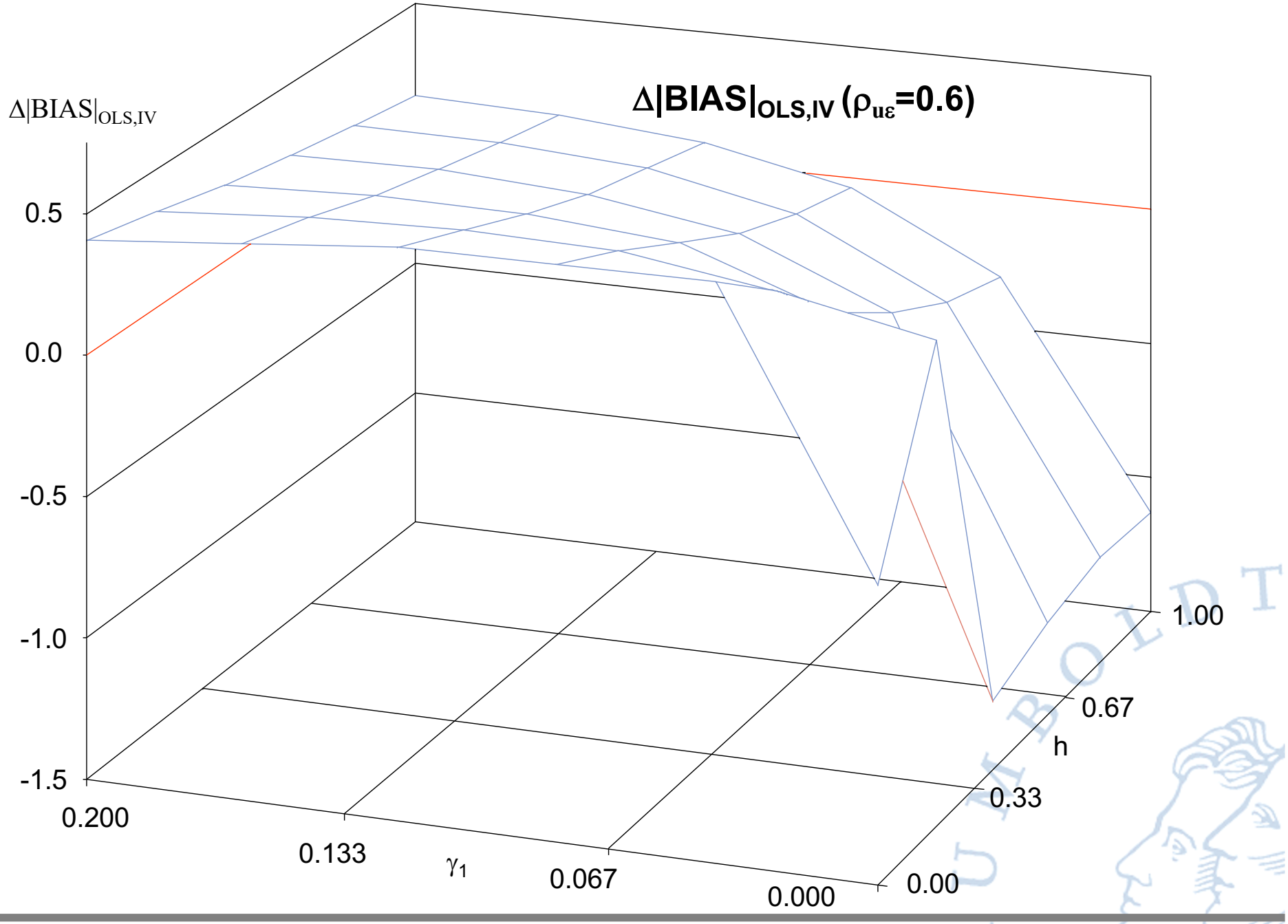
$$Q \sim N[0,1],\ \boxed{\rho_{u\varepsilon}} \geq 0$$

$$\beta_0 = 1,\ \beta_1 = 1$$



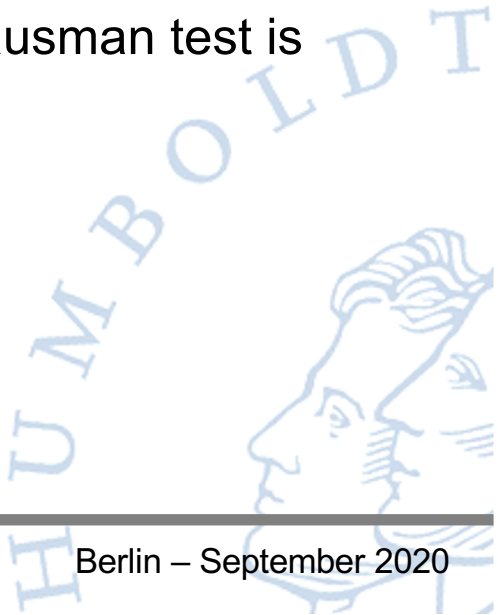

β1
3.0
2.5
2.0
1.5
1.0
bOLS (ρuε=0.6)
1.00
0.67
γ1
0.33
0.00
1.00
0.67
h
0.33
0.00


β1
3.0
2.5
2.0
1.5
1.0
b_IV (ρ_uε=0.6)
1.00
0.67
γ1
0.33
0.00
1.00
0.67
h
0.33
0.00


Δ|BIAS|OLS,IV (ρuε=0.6)
Δ|BIAS|OLS,IV
0.5
0.0
-0.5
-1.0
-1.5
0.200
0.133
γ1
0.067
0.000
0.00
0.33
h
0.67
1.00

But I did a Hausman test…

- Two standard tests are used to assess the validity of the IV approach
- Hausman test
  - tests whether instrumented variable is endogenous to instrument(s)
  - Basically a standard test for the joined significance of the first-stage instruments in the second stage
  - Assumes instruments to be exogenous to the second stage
- Over-identifying restrictions test
  - Tests whether a given instrument is exogenous, using the other instruments
  - Assumes that at least one instrument is exogenous
- If the Over-identifying restrictions test is significant than the Hausman test is pointless.



Suggested steps for dealing with endogeneity problems

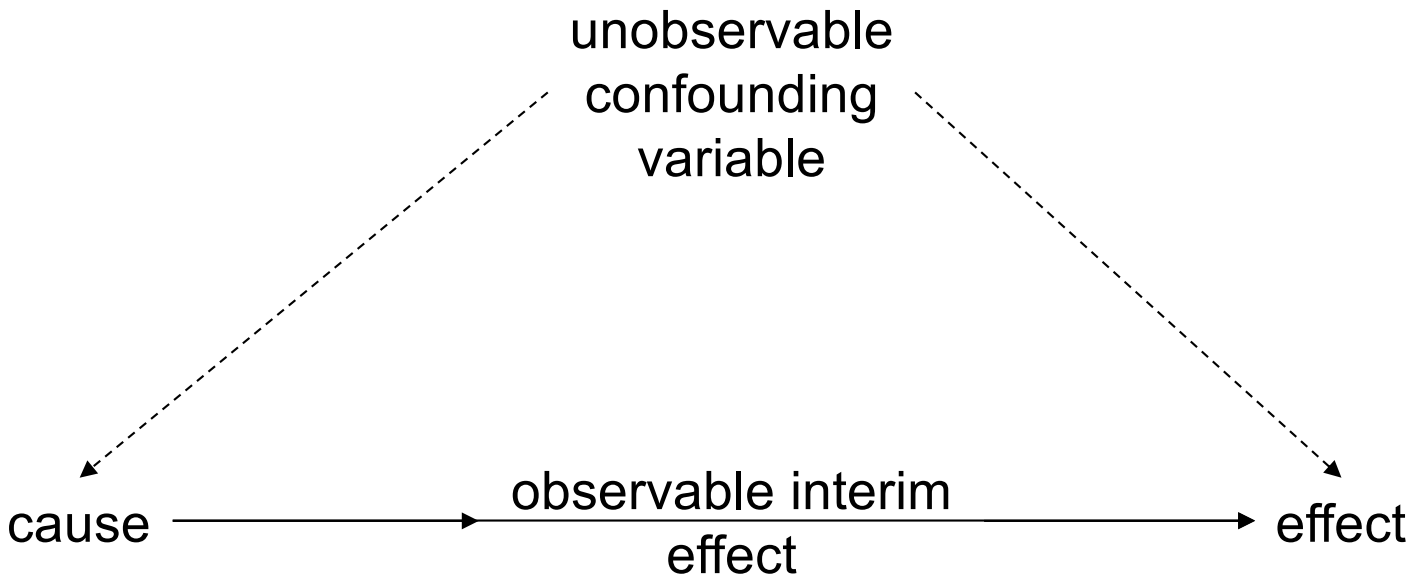
- Addressing endogeneity problems
  - Describe the nature of the endogeneity problem
  - Explore alternative research designs

- Implementation of the instrumental variable estimation
  - Use economic theory to select and justify the choice of instruments
  - Evaluate the first-stage results and diagnostics
  - Evaluate the second-stage results and diagnostics
  - Run a sensitivity analysis on the choice of instruments
  - Compare and contrast the estimates from OLS and 2SLS methods

- Assess the potential impact of unobserved confounding variables

(Quoted from Larcker and Rusticus, JAE 2010: 196, Table 4)

What is a mechanism?



Designing a identification strategy for causal inference

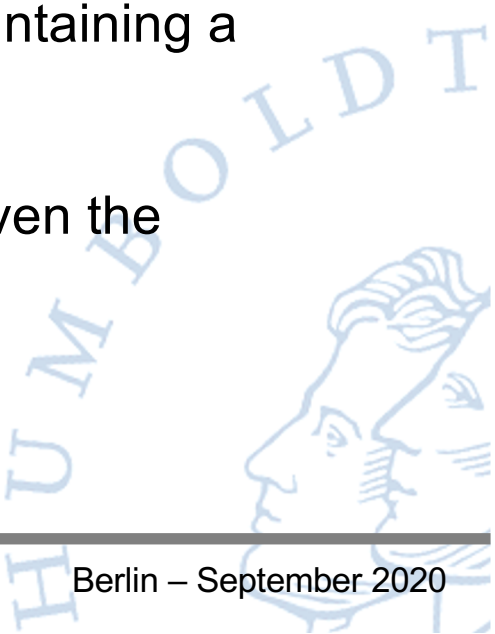
Designing a causal study requires addressing four questions

1. What do you believe is going on? (theory that predicts a causal relationship)

2. What would be a perfect experiment to test your causal prediction (and why is it not feasible)?

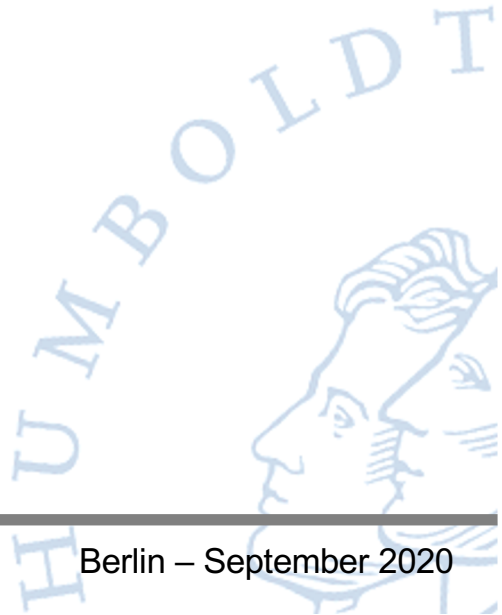
3. What is the second-best quasi experiment that, while feasible, deviates as little as possible from the perfect experiment while maintaining a desired level of external validity?

4. How do you adjust your method of statistical inference given the identification strategy presented under #3?



Wednesday, 09.09.2020

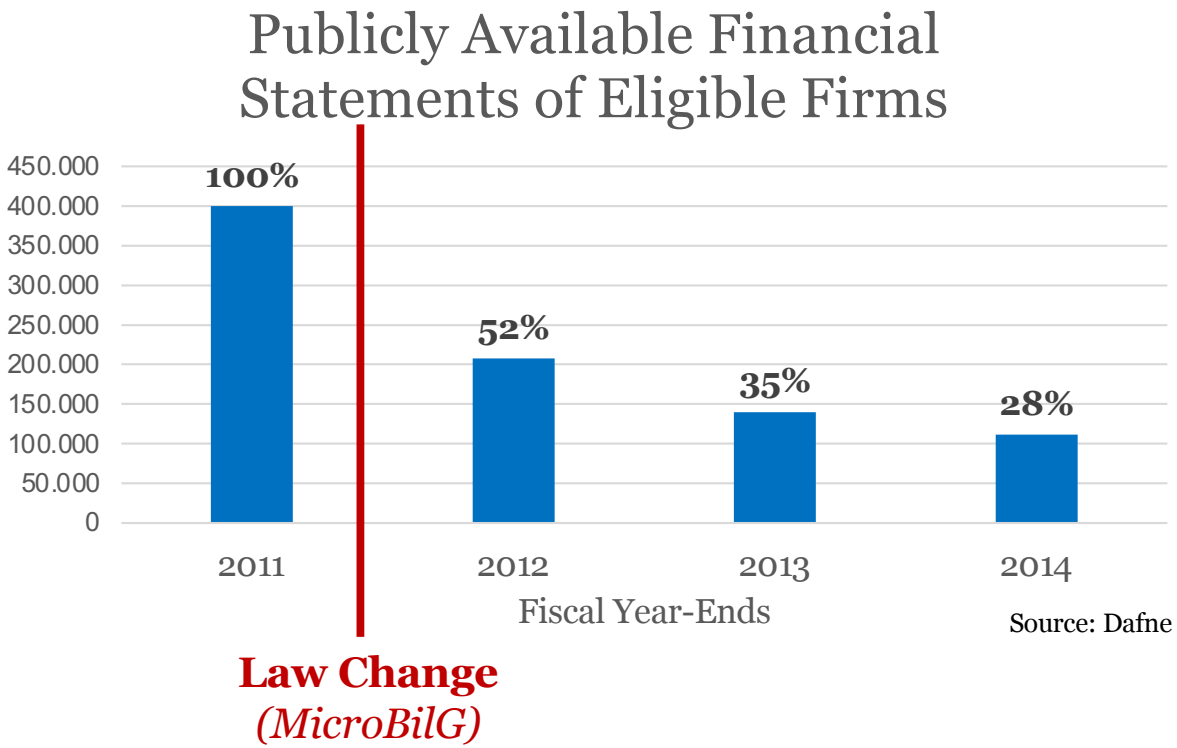
- Paper: Breuer and Windisch (2019)
- Research design
  - Identification strategies
  - From the research question to the research setting
- Paper: Gassen and Muhn (2018)
- Execution
  - Data wrangling
  - Exploration
  - Modelling and testing
- Class project
  - Exploratory data analysis
  - Discussion of potential research questions



Financial Transparency of Private Firms: Evidence from a Randomized Field Experiment

Joachim Gassen
Humboldt University of Berlin
Maximilian Muhn
University of Chicago – Booth School of Business

Motivation



- Majority of eligible firms are choosing the 'restriction'-option (~ 70%)
- However, still more than 100,000 eligible firms are publishing their financial statements without any restrictions

Focus and Approach

Research Questions

1. How important are informational constraints in a (private) firm's decision to disclose financial information?
   - ➢ Connects to Bloom et al. QJE 2013; Atkin et al. QJE 2017; Zwick WP 2018
2. Which stakeholders are influencing a firm's decision to become financially transparent or opaque?
   - ➢ Connects to the large literature on voluntary disclosure

Features of this Project

- Drawing causal inferences via large-scale randomized field experiment
- Information treatment allowing for within-treatment variation
- Using survey results and administrative data for uncovering likely mechanisms

Institutional Setting and Context

Eligibility for *MicroBilG* Exemptions

- Since 2012, firms not exceeding more than one of these three size thresholds:
  1. Total Assets <= €350,000 (~ $446,000)
  2. Sales <= €700,000 (~ $893,000)
  3. Employees <= 10 people

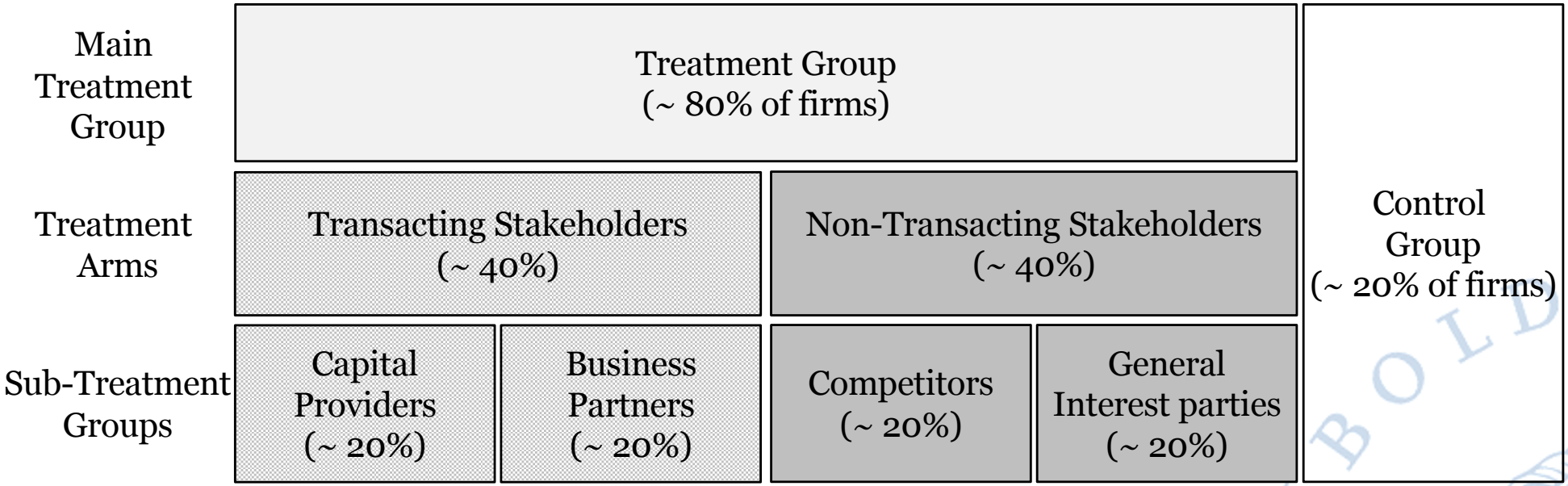
Main Consequence of *MicroBilG*

- *MicroBilG* provides eligible firms the option to restrict access to their financial statements:
  - Financial statements no longer listed in the *Federal Gazette*
  - Access only via the *Company Register* after registration
  - Small fee for each financial statement
  - Data no longer contained in databases such as *Dafne*

Research Design: Treatment and Control Groups

Setup

- Identifying firms which still publicly disclosed their 2014 financial statements



Expectations of Treatment Effect

Control	
Treatment	↑ **Higher Restriction Rate**

- Information treatment provides information about the option
- Information treatment bypasses the tax advisor

Transacting Treatment	↓ **Lower Treatment Effect** (**Capital providers, Customers & suppliers**)
Non-Transacting Treatment	↑ **Higher Treatment Effect** (**Competitors, General interest parties**)

- Within-treatment frames disclosure decision
- Treated firms think more carefully about the impact of the disclosure decision on the "framed"-stakeholder group

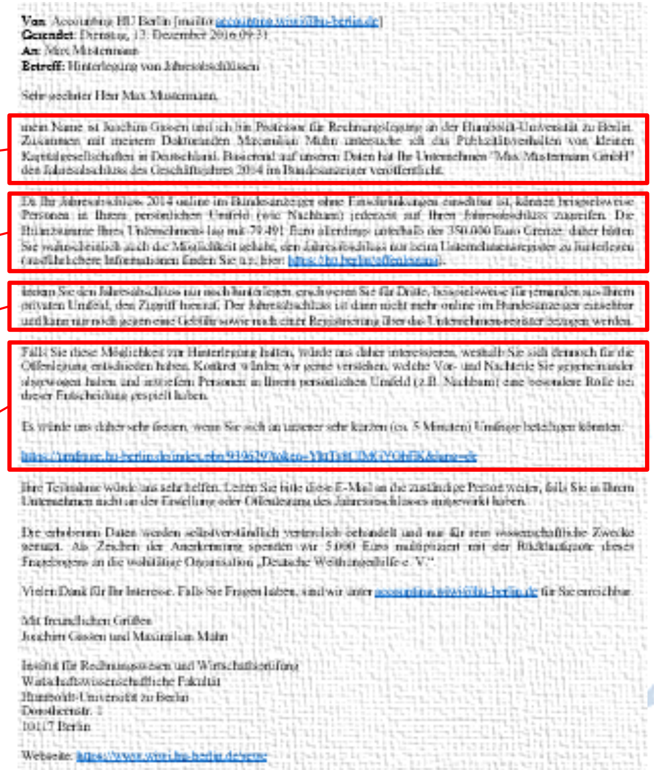
Research Design: Information Treatment

Execution

- Single email providing information via survey participation request

Structure of email

- Four main paragraphs:
  1. Introduction
  2. Eligibility for restriction option
  3. Implications of restriction option
  4. Participation in survey about 2014 decision
- Within-treatment variation:
  - ➢ Almost identical emails
  - ➢ But different stakeholder group is mentioned three times throughout email text


Sehr geehrter Herr Max Mustermann,

Sample Selection and Data

Data Processing

- Crawled *Company Register* until December 12, 2017 (one year after experiment)
  - ➢ Outcome Variable: Restricted (1/0)
  - ➢ Final sample consists of 25,724 firms

Descriptive Statistics

(N = 25,724)	*Mean*	*Std. Dev.*	*P1*	*P25*	*Median*	*P75*	*P99*
Control Variables:							
Total Assets (€)	159,951	97,004	6,399	76,046	149,531	239,677	345,082
Firm Age (Years)	16.31	14.87	3	6	13	22	79
Tangibility Ratio	0.191	0.203	0	0.032	0.124	0.284	0.856
Equity Ratio	0.340	0.294	0	0.034	0.302	0.576	0.966

- Median firm size comparable to US studies using Survey of Small Business Finances data (e.g., Allee and Yohn TAR 2009; Cassar et al. JAE 2015)

Filing Decision by Main Treatment and Treatment Arm Status

	(1) *Restricted*	*(2)* *Restricted*	*(3)* *Restricted*	*(4)* *Restricted*
Constant (= No email)	0.262***	–	0.262***	–
	(42.82)		(42.82)	
Experimental Variables:				
Treatment	0.039***	0.039***	–	–
	(5.69)	(5.57)		
Non-Transacting	–	–	0.046***	0.045***
			(6.00)	(5.84)
Transacting	–	–	0.033***	0.032***
			(4.33)	(4.25)
Controls:	NO	YES	NO	YES
Fixed Effects				
Industry	NO	YES	NO	YES
District	NO	YES	NO	YES
P-value from F-Test:				
Non-Transacting = Transacting	–	–	0.045	0.057
N (= Number of firms)	25,724	25,708	25,724	25,708

- Focus on intention-to-treat effects only (more conservative)
- Robust standard errors in all tests (following Abadie et al. NBER WP 2017)

Summary and Contribution

Contribution

- Informational constraints of the relevant decision maker is an economically relevant determinant for the adoption of certain accounting practices or methods

- Causal evidence that transacting (non-transacting) stakeholders are the main reason why private firms choose to be financially transparent (opaque)

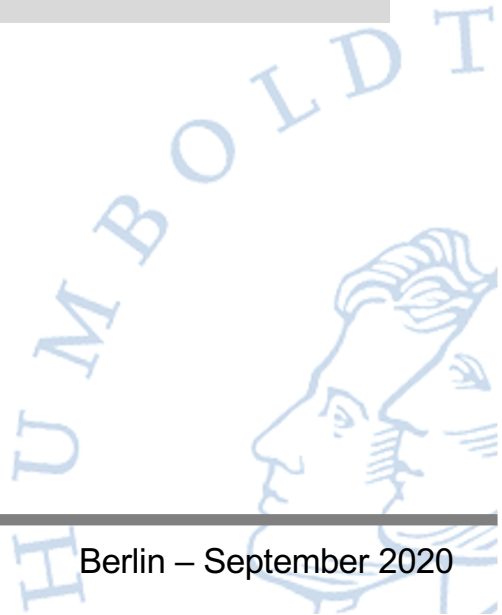
- Exploring the role of non-traditional stakeholders (general interest parties) in a private firm's disclosure decision

Follow-up Project

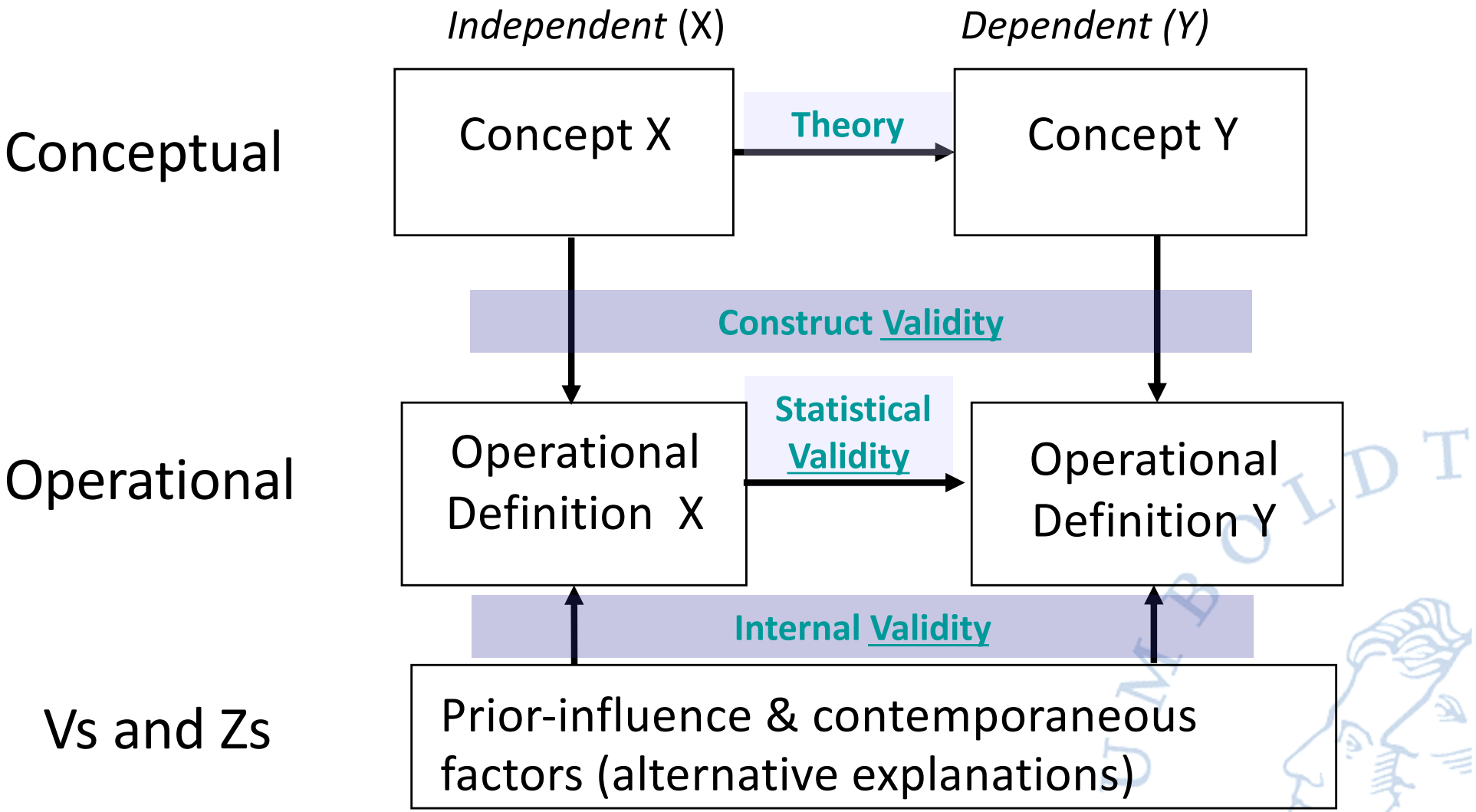
- Focus on causal consequences of this disclosure decision (data available in 2019-2020)

Wednesday, 09.09.2020

- Paper: Breuer and Windisch (2019)
- Research design
  - Identification strategies
  - From the research question to the research setting
- Paper: Gassen and Muhn (2018)
- Execution
  - Data wrangling
  - Exploration
  - Modelling and testing
- Class project
  - Exploratory data analysis
  - Discussion of potential research questions



Predictive validity framework ("Libby boxes")

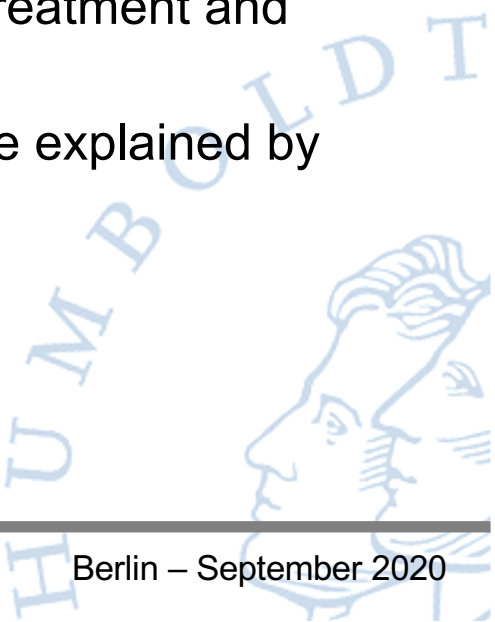


Noisy measures generate noisy estimates...

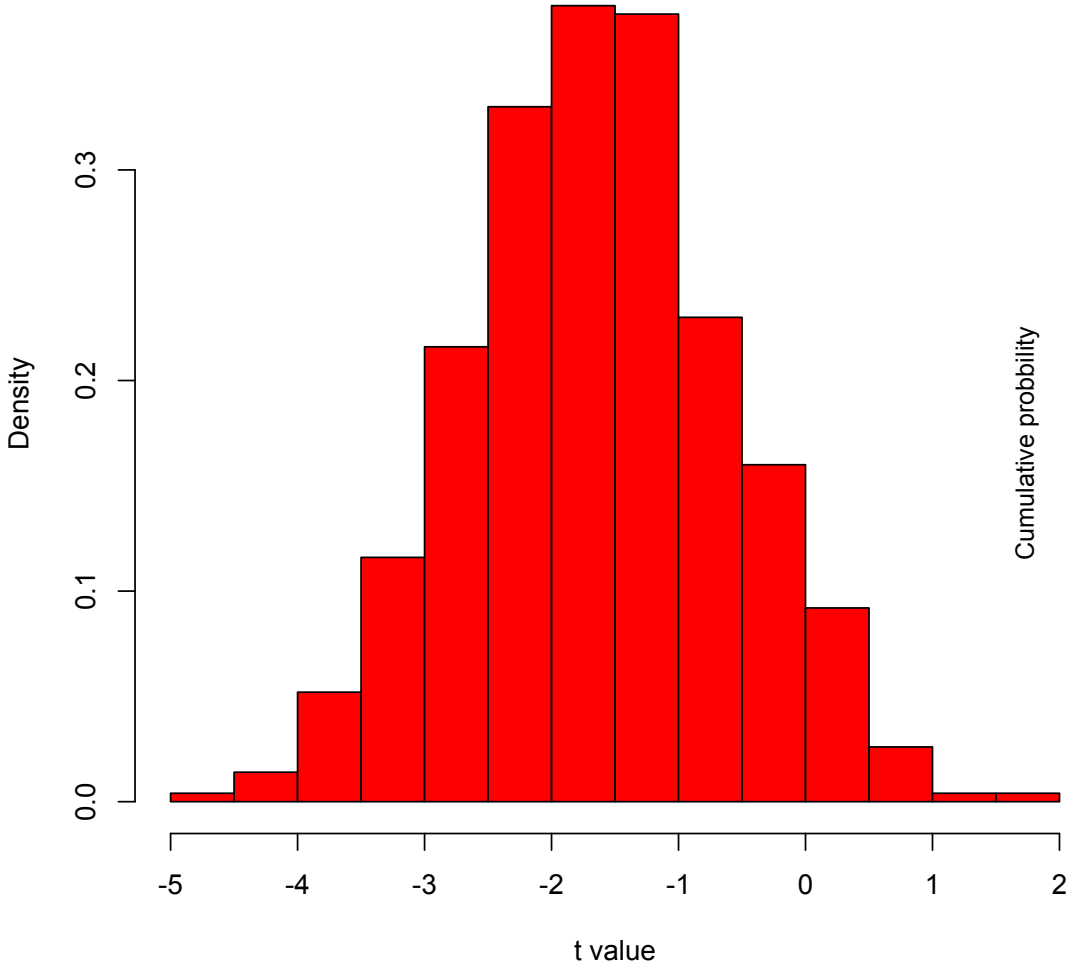
Economic Construct	Empirical Implementation
Earnings Quality	Abnormal accruals, other earnings attributes
Earnings Management	Discretionary accruals, small loss avoidance
Audit Quality	Discretionary accruals, audit fees, gcao, earnings properties, ...
Conservatism	Basu regression coefficients
Transparency of financial accounting narratives	Fog index
Stock prive informativeness	Inverse R2 of CAPM regressions

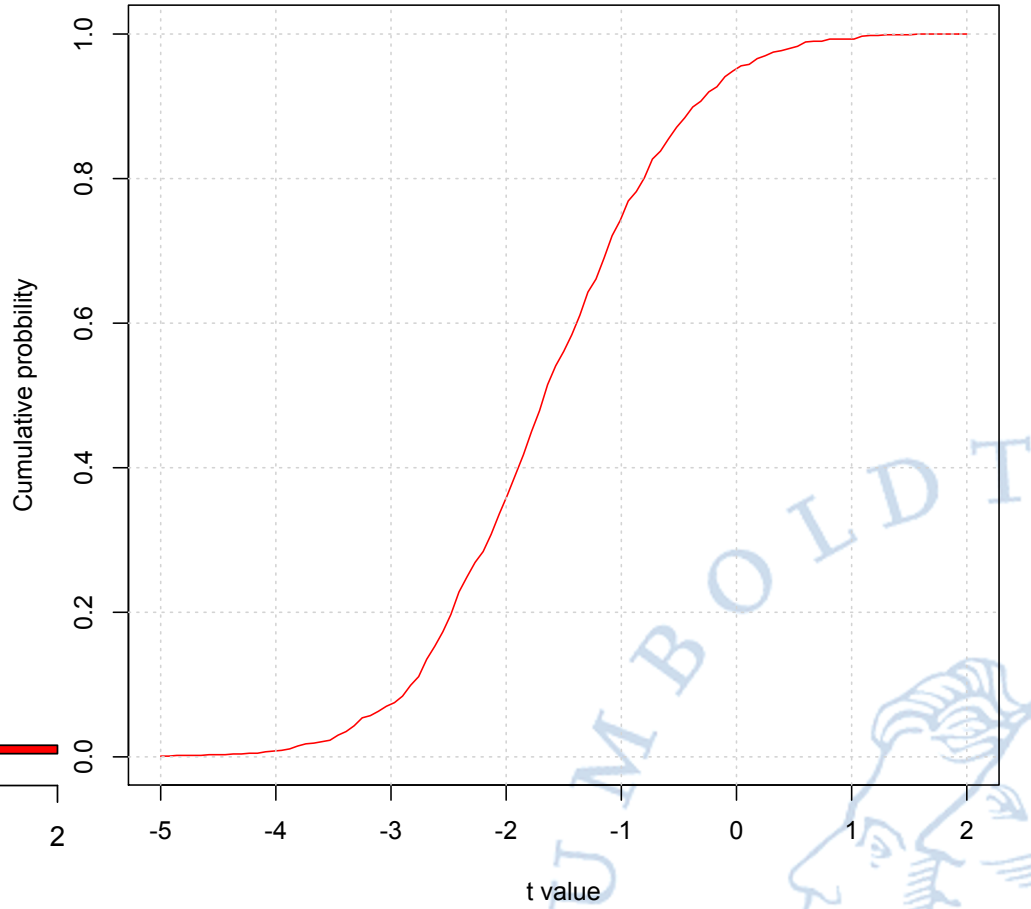
Estimate test power ex ante experimentation: An example

- Assume you want to do an study on cost of equity capital consequences of financial reporting
  - How large is the change in the financial reporting information environment that you want to study?
  - How large is the relative share of financial reporting relative to the total information environment?
  - How large is the information asymmetry component in the cost of equity capital?
  - What is the empirical distribution of cost of equity capital?
  - How large is your sample and how is it distributed across treatment and control observations?
  - How much of the dependent variable variation is likely to be explained by other covariates?



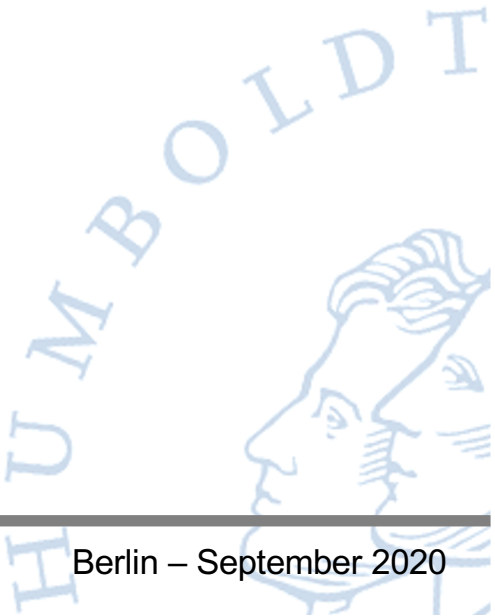
Power simulation
(COEC effect=7bp, n=25,000)



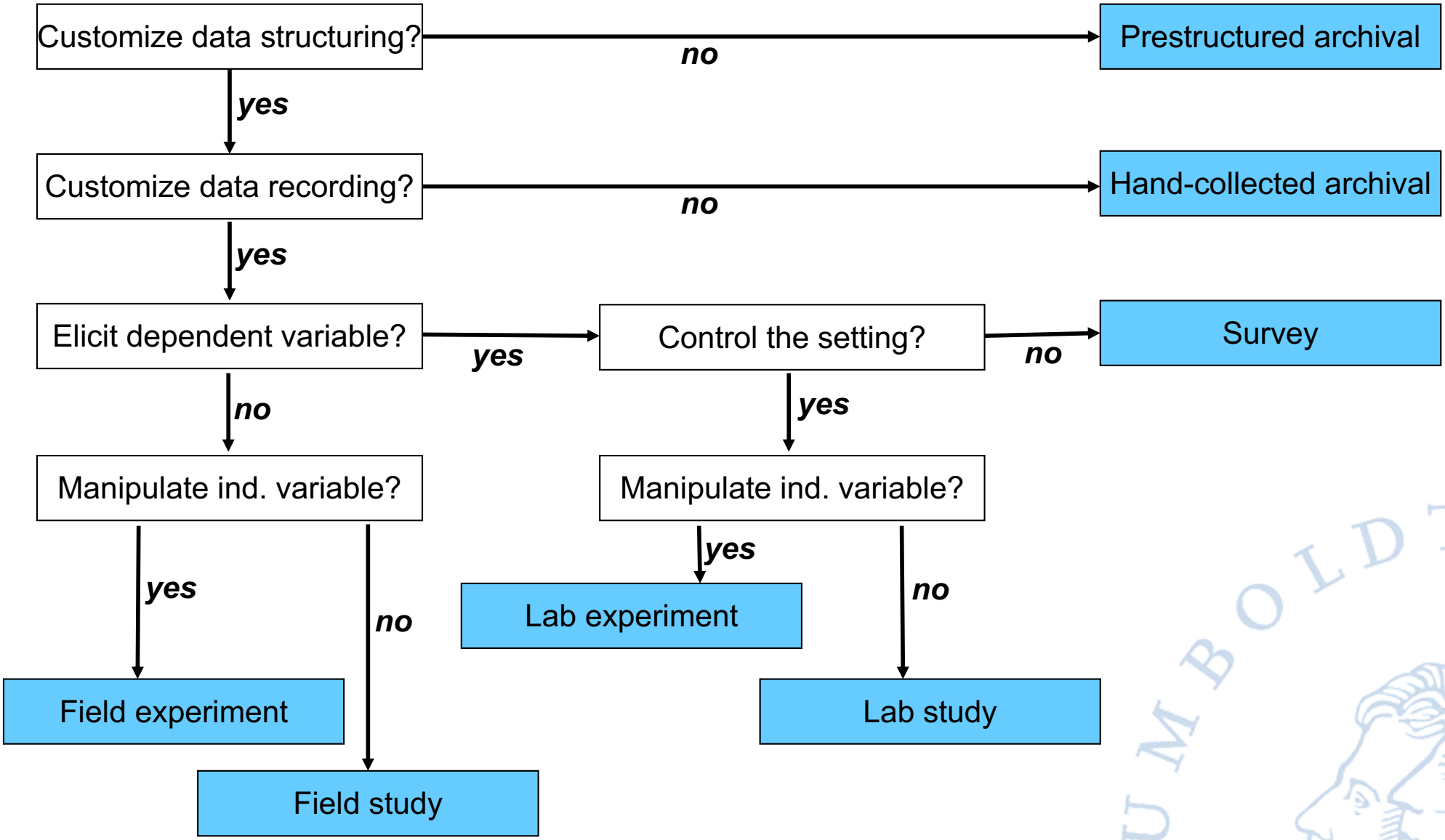


Data collection

„One of the real limitations of empirical research is that we tend to work on problems we are able to study because data are available; we thereby tend to overlook problems that we ought to study, if data for such problems are not easy to obtain. […] Gathering direct and original facts is a tedious and difficult task, and it is not surprising that such work is avoided.“ (Vatter, ***1966****)*



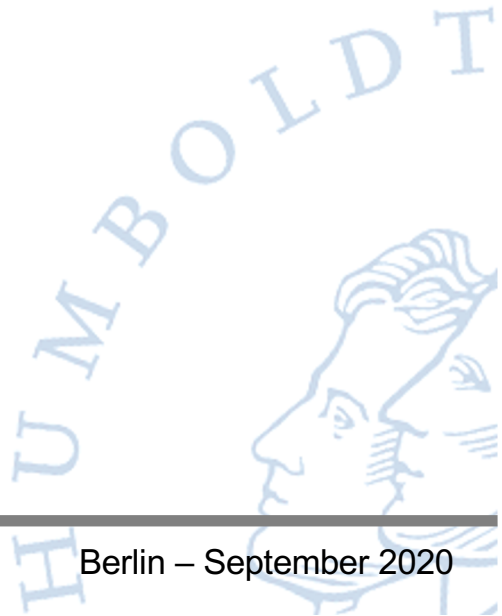
Methods of data collection



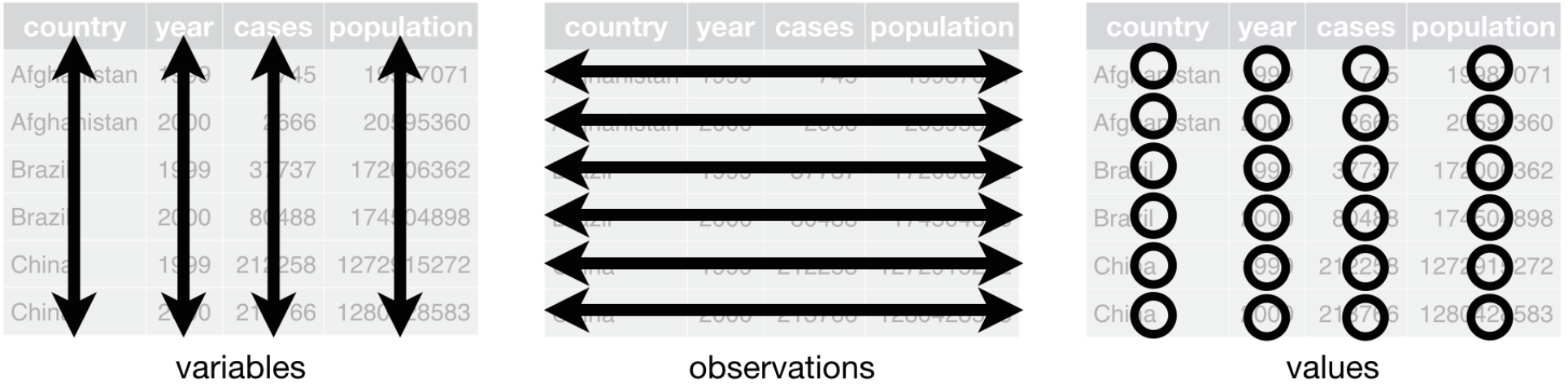
(Bloomfield et al., 2015, Figure 3)

Picking the right data source…

- There are tons of interesting data out there: The best might not be in a database
- Check data from the research data center: http://sfb649.wiwi.hu-berlin.de/fedc/
- Relevant Database providers for firm-level data:
  - Worldscope
  - Compustat Global Vantage
  - Bureau van Dijk Amadeus
- Relevant Database providers for capital market data
  - Datastream (international)
  - CRSP (US)
  - Many small providers for market microstructure data
- Common pitfalls
  - Lacking coverage, dead firms
  - Accounting data is being treated
  - Market microstructure matters for capital market data



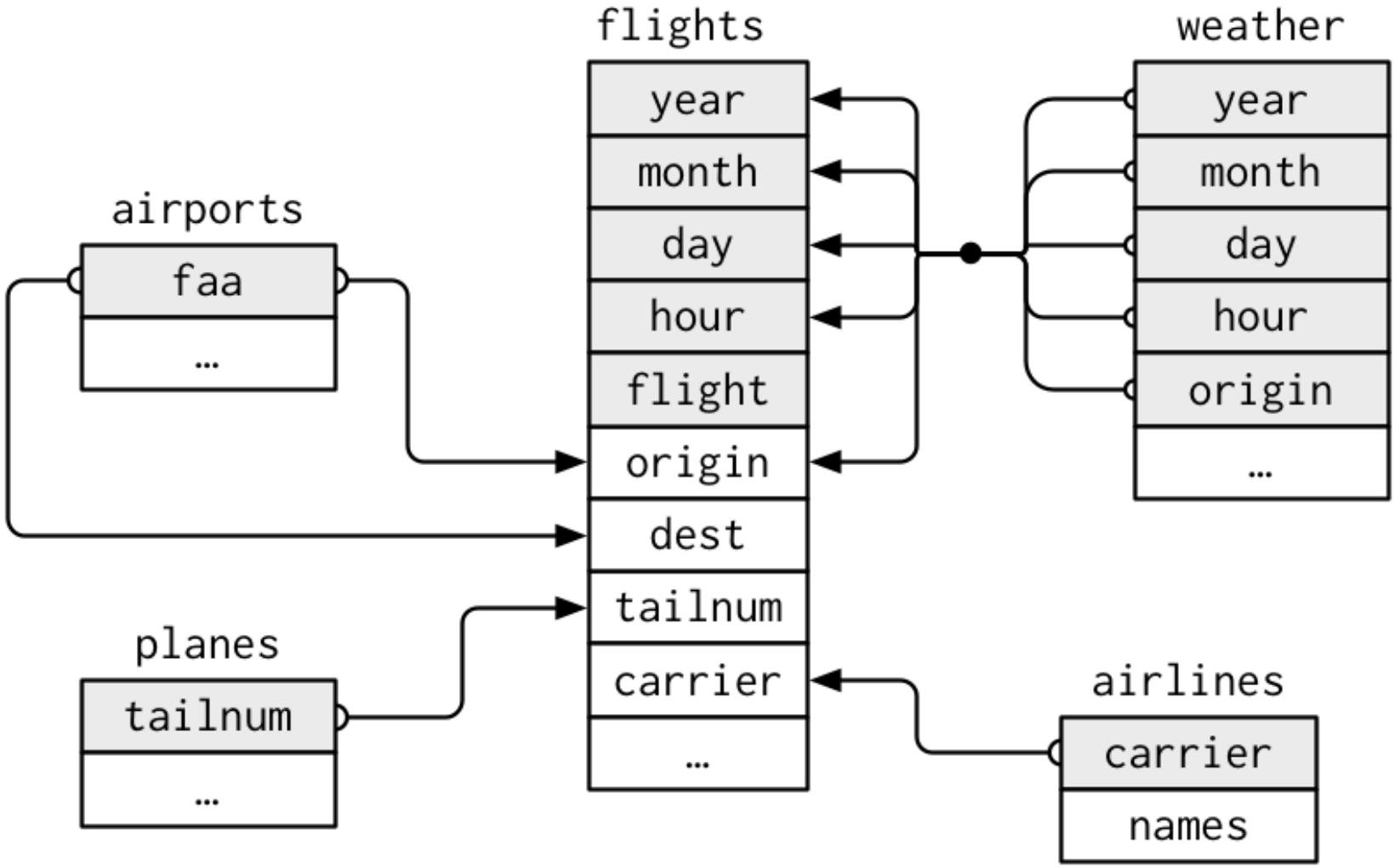
Tidy data (long format)



1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

Grolemund and Wickham (2017), R for Data Science, http://r4ds.had.co.nz/tidy-data.html

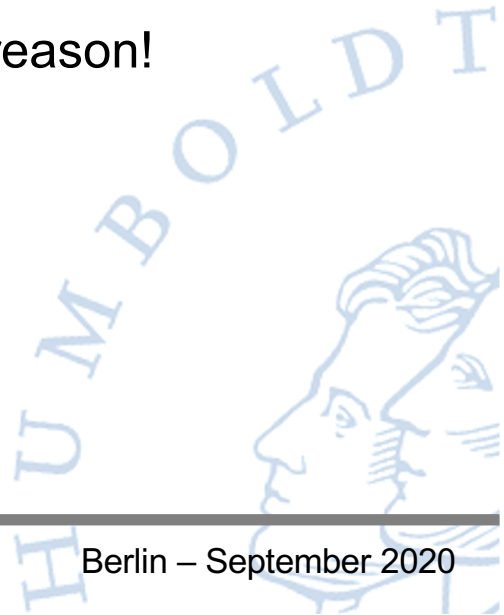
Relational data



Grolemund and Wickham (2017), R for Data Science

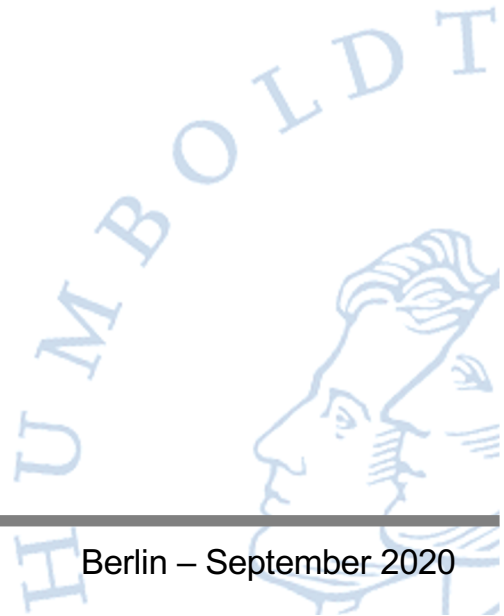
Organizing data

- An Observation…
  - is identified by its primary key
  - is stored in a row
  - can (and likely will) have multiple variables defined
- Some rules
  - Always know what your primary key is!
  - Do not allow duplicates!
  - Store raw data in separate files (make them read-only)!
  - Always keep variable definitions close to your raw data!
  - Do not rename variable names in raw data files without a reason!
  - **Never** change data by editing files!
  - Debatable: Learn SQL!



From raw data to a sample: Some steps

- Developing the research design
- Identifying the raw data to use
- Merging it, thereby defining the structure of the sample (cross-sectional unit, time dimension)
- Defining, calculating and verifying variables
- Decide how to deal with missing values
  - Floating or stable sample?
  - Balanced or unbalanced panel?
- Addressing extreme value cases
- Preparing descriptive statitsics
- Conducting tests
- Return to square one…



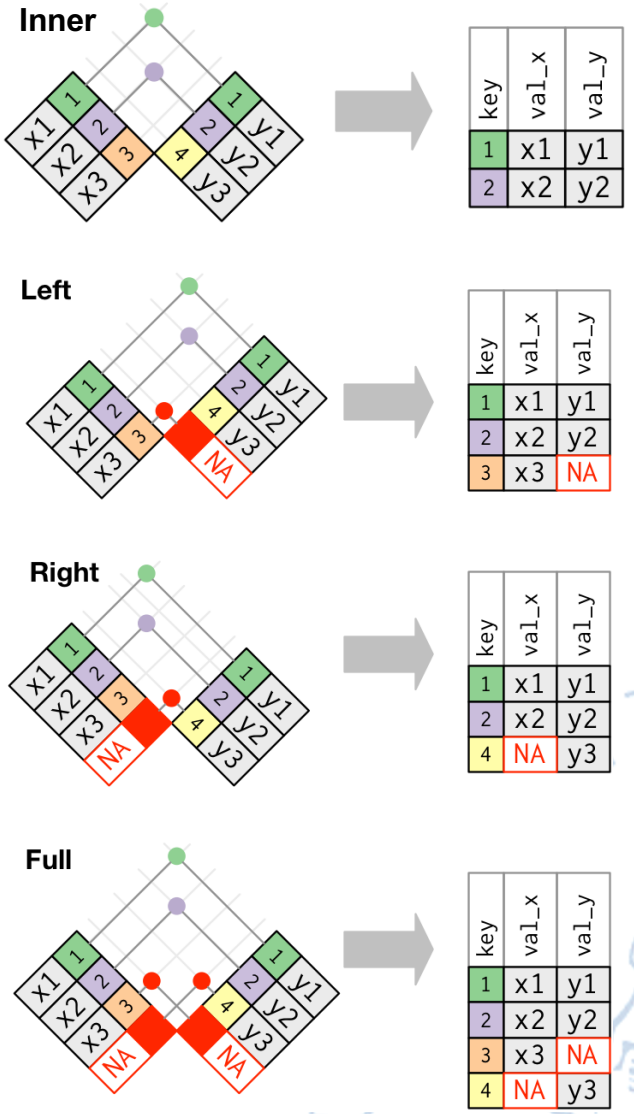
Joining data

x

1	x1
2	x2
3	x3

y

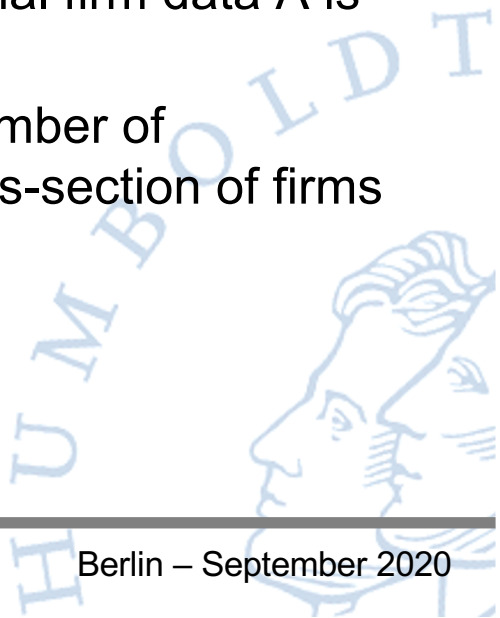
1	y1
2	y2
4	y3



Grolemund and Wickham (2017), R for Data Science

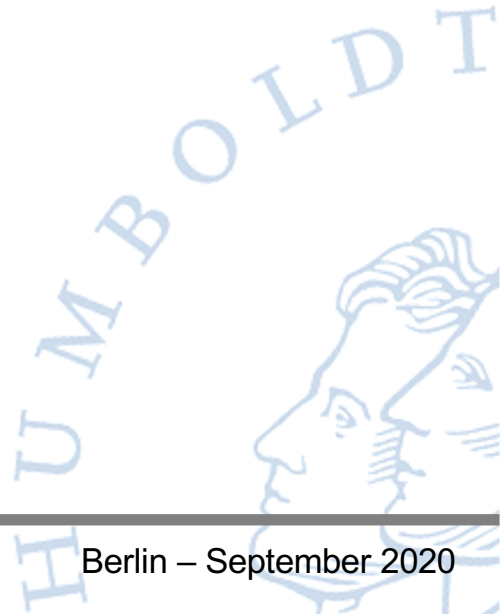
Merging data

- When preparing your sample, you will be merging raw data from different sources
- This is done by identifying matching observations by their keys
- Different matching types
  - **1:1**: Each observation of set A can be linked to a maximum of 1 observation of set B and vice versa (**Example**: Cross-sectional firm data from different data sources)
  - **1:n**: Each observation of set A can be linked to a finite number of observations from set B but every observation of set B has a maximum of 1 matching observation from set A (**Example**: Cross-sectional firm data A is linked to panel data B, which has a time dimension)
  - **n:m**: Each observation of set A can be linked to a finite number of observations from set B **and vice versa** (**Example**: A cross-section of firms A is merged to a dataset of financial analysts B)



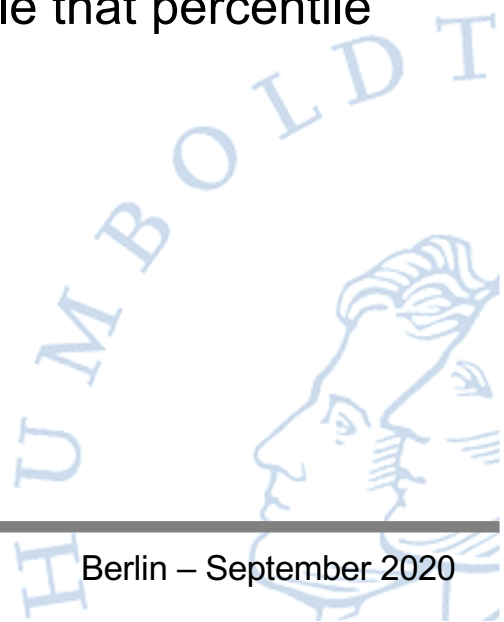
Sampling...

- The research question should determine the setting and the sample (not the other way around!)
- Sampling determines external validity
- Large N is not necessarily good
- Normally there is a tradeoff: N versus construct validity and measurement error
- What do you want your sample to be representative for?
- Do you know the underlying population?
- Do you have a random sample? Most likely not
- If your sample is not random, deal with sample selection bias
- Remember: Matching might be a smart sampling approach
- Weighted test approaches can also be helpful



Addressing extreme values

- No matter how cautiously you define your variables, some observations will show extreme values
- Extreme values can be treated as if they are
  - errors, which you do not want to influence your findings or as
  - extreme values which are valid but might have an overly extreme influence on your test findings
- In the former case you would **truncate** your data, deleting the extreme values (and thereby in most cases the according observation)
- In the latter case, **winsorizing** is the way to move forward. When winsorizing, you set every value which is below or above a certain percentile that percentile
- Winsorizing is more common than truncating
- A common threshold is the 1st and 99th percentile



Defining, calculating and verifying variables

- Follow established literature where possible, do not cook your own
- Deflate! The size effect is huge an renders variables uninformative
- Rethink your definition and the data it is going to produce. Do you expect your data to show a „nice“ distribution? Some examples
  - DEBT/TA vs. DEBT/(TA-DEBT)
  - ROE vs. ROA
- Look at the distribution of your data. Does it make sense? What about the magnitude? What about the extremes?
- Do not use your data in tests until you are satisfied…