Quantitative Data Analysis: A Companion for Accounting and Information Systems Research

Teaching Materials

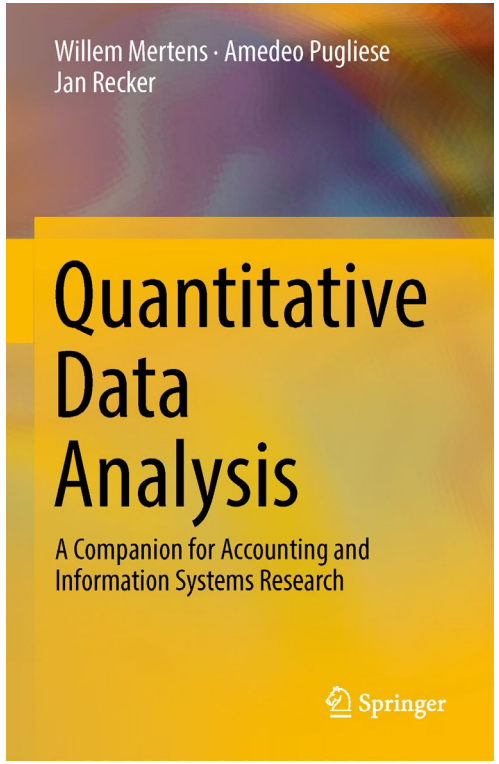
Created by Willem Mertens, Amedeo Pugliese & Jan Recker

Copyright Notice

What these materials are about

Offering a guide through the essential steps required in quantitative data analysis

1. **Introduction**
2. Comparing Differences Across Groups
3. Assessing (Innocuous) Relationships
4. Models with Latent Concepts and Multiple Relationships: Structural Equation Modeling
5. Nested Data and Multilevel Models: Hierarchical Linear Modeling
6. Analyzing Longitudinal and Panel Data
7. **Causality: Endogeneity Biases and Possible Remedies**
8. How to Start Analyzing, Test Assumptions and Deal with that Pesky p-Value
9. Keeping Track and Staying Sane



Part 7:
Endogeneity & Self-Selection: Propensity Score Matching & Selection Models

Agenda

Making Causal claims with Observational Data

- Randomized assignment vs Observational data
- Sources of Endogeneity & Self-Selection Problem
- Specifying the right model

Instrumental Variable & 2 Stage Least Squares

- Concepts and Applications

Propensity Score Matching

- Concepts and Applications

Summary and Takeaways

Why are you here?

1. **Our RQ is a causal-like (e.g.):**

- Does giving incentives to CEOs improve firm performance?
- Does adopting ERP system reduce faulty manufacturing?

We wish to assess whether offering $1 in stock option (adopting an ERP system) improves performance (reduces faults) **– everything else being equal**

2. **We have observational data** (e.g. survey or archival), hence no-random assignment of your units to the treatment / control conditions

Making causal claims with observational data

Units are non-randomly assigned to the (levels of) treatment/control condition:

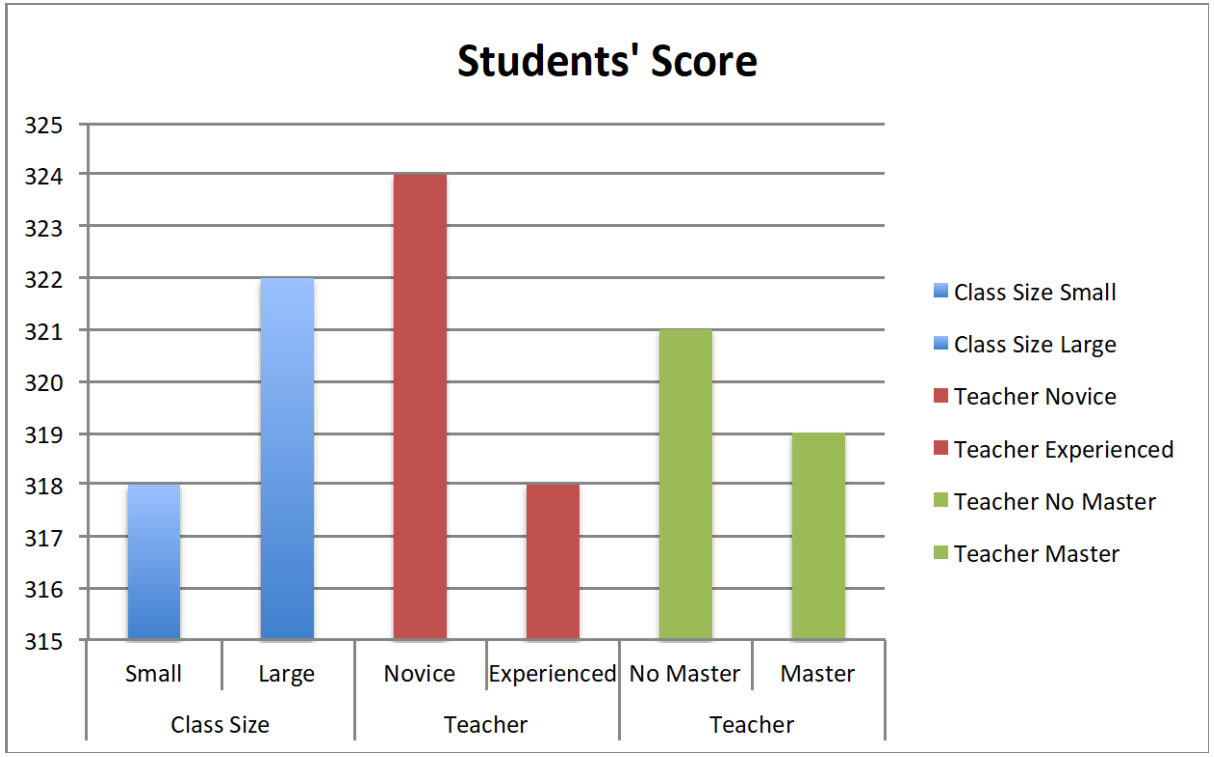
- Self-Select themselves (sick people going to hospital have poorer health afterwards)
- Differ across characteristics correlated with the outcome (wealthy families send kids to good schools)

Making causal claims requires meeting stringent conditions:

1. The cause precedes the effect (time)
2. ###
3. 333

A classic in education research

1. Does class size affect students' performance?
2. Does teacher experience (qualification) affect students' performance?



A classic in education research

Early results suggested that:

- Students in larger classes perform better
- Students in classes with novice and less qualified teachers have better results

Does it make sense? What justifies these results?

Class size and teacher allocation are **endogenous** & depend on factors affecting the outcome:

- Poorly performing students are allocated to smaller classes
- More experienced teachers are assigned to larger and more problematic cohorts

The WAJ problem: refinancing the w/shops

HoS ask for evidence of the effectiveness of w/shops prior to refinancing them

Proxy for Effectiveness: # of individual journal submissions after the w/shops

Six months later, WAJ present their evidence:

- S1: Staff (not) attending w/shops have submitted (1.9) 2.9 articles
- S2: Staff (not) attending w/shops submitted (1.9) 2.0 articles

Would you re-finance w/shops based on results from S1 or S2?

The WAJ problem: attendance is a choice

We are comparing two groups that are potentially different”

1. Staff choose to attend the w/shops based on their need to acquire training
   - Staff familiar with stats are more likely to attend (inflating our findings)
   - Staff less familiar with stats are more likely to attend (works against our findings)

(Selection on dependent variable)

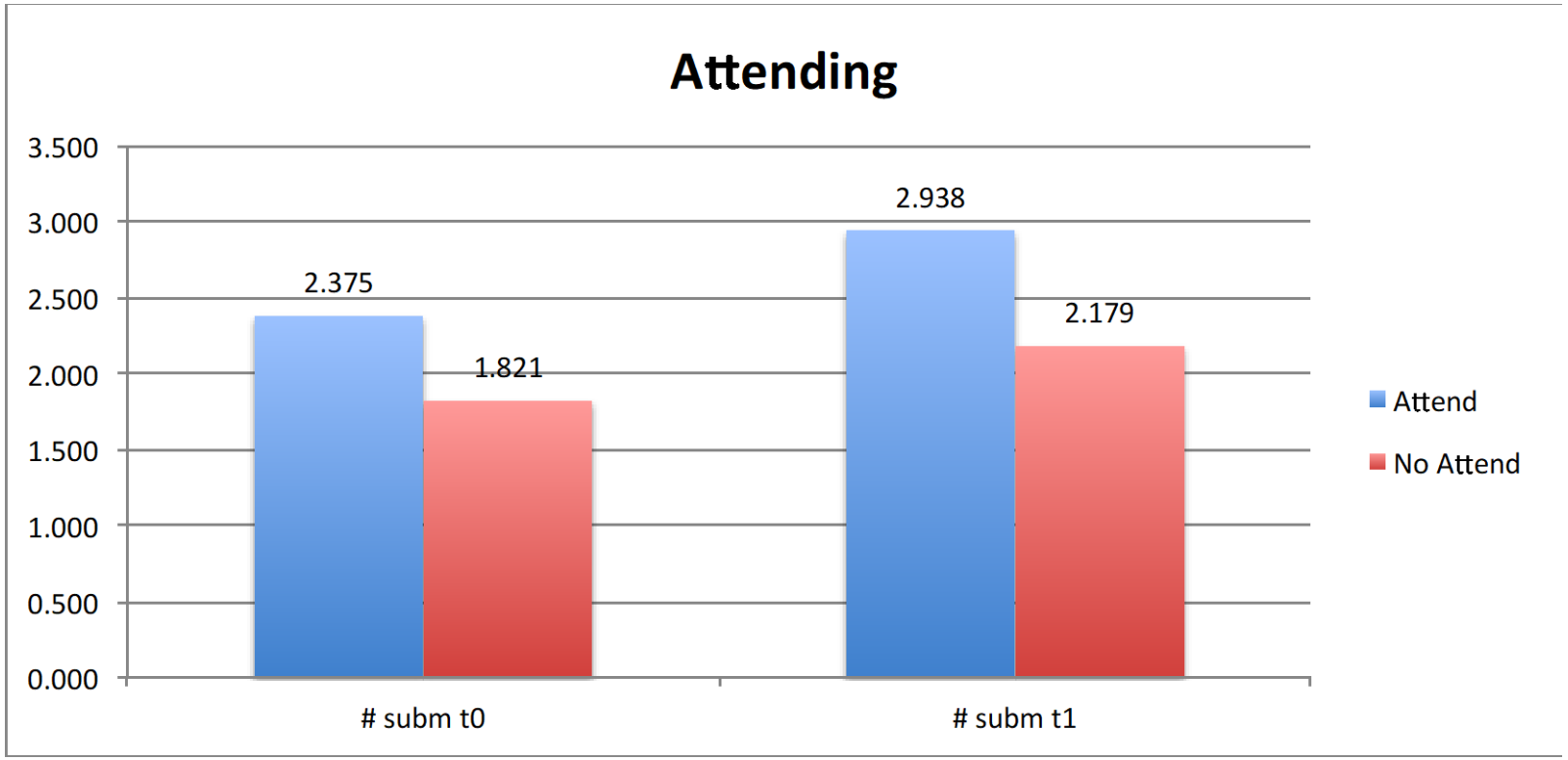
The WAJ problem: attendance is a choice

We are comparing two groups that are potentially different

2. Attending and not-attending staff differ along other dimensions:
    - Background
    - Nationality
    - Gender

(Selection on covariates)

Self-selection: who goes to w/shops?



The two groups are not identical – staff attending w/shops are more productive ex-ante

Overestimate the effectiveness of w/shops: it is confounded with the ex-ante ability of attending staff

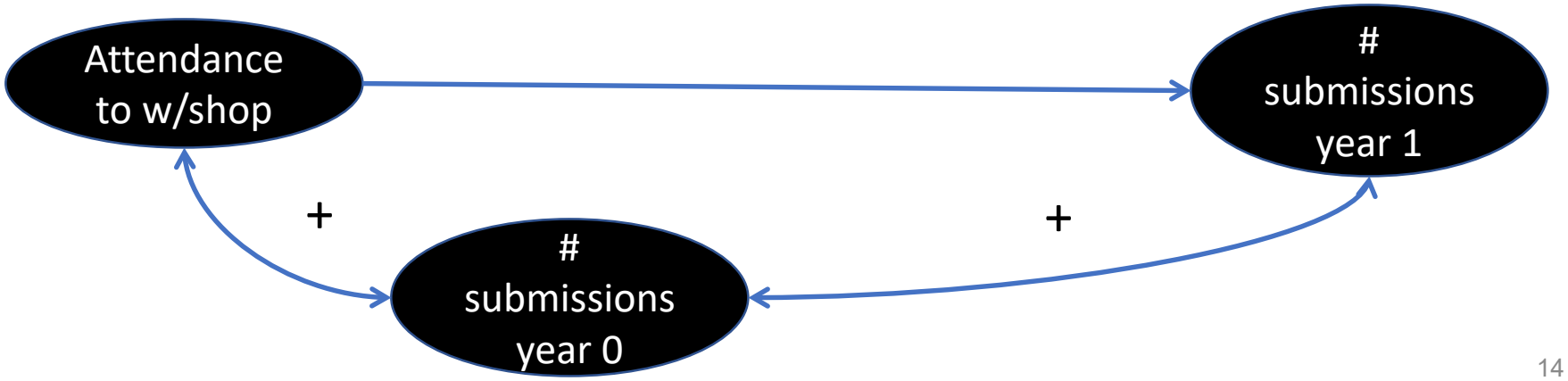
The WAJ problem (3): the ideal state

Test the effectiveness of w/shops in an ideal state:

i. Randomly select staff attending (treatment) and non-attending (control) the w/shops

ii. Compare the # journal submissions between the two groups (t-test)

What do we have instead?



The WAJ problem (4): identification strategy

We cannot compare Attending vs Not Attending Staff but…

We have the following data points for each staff member:

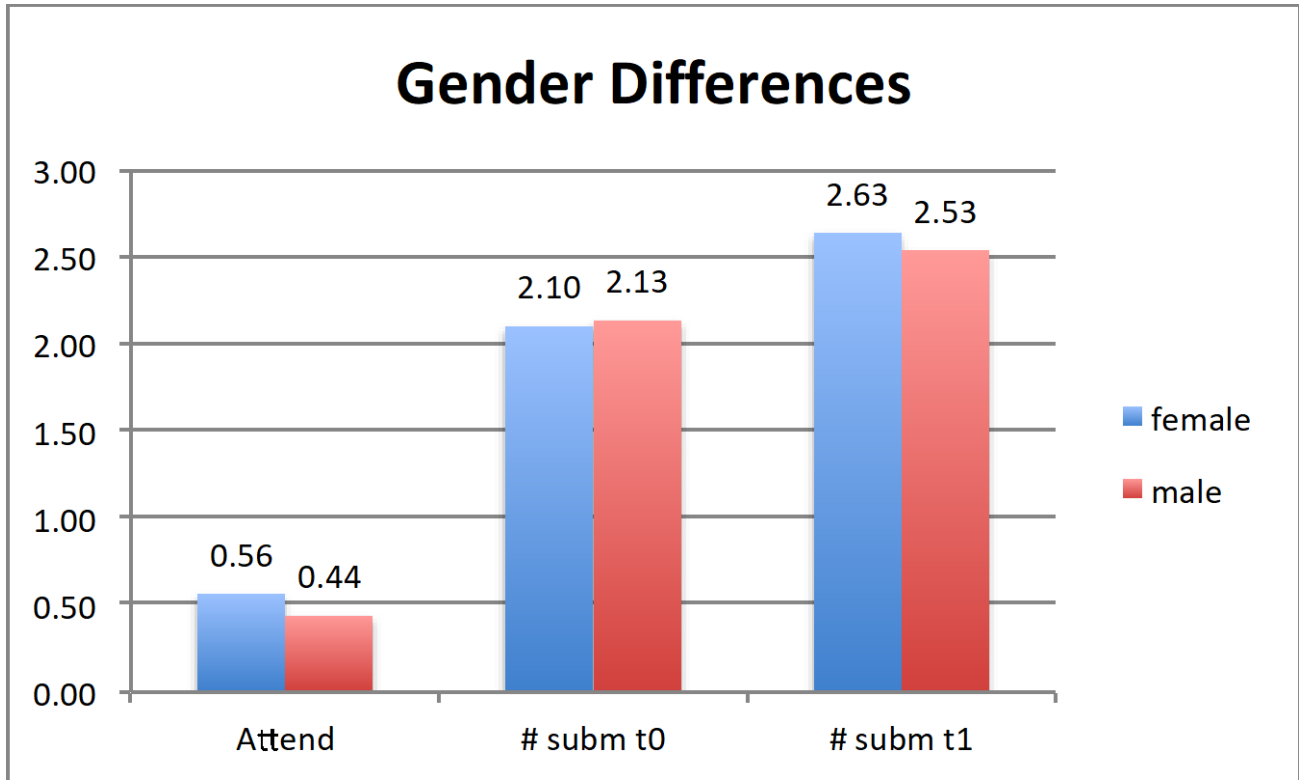
- # jnl submissions at T1
- Attendance to the w.shops (dummy)
- # jnl submissions at T0
- Gender (dummy for female)
- Nationality (dummy for local)
- Background (Science) (dummy if BA in Science)

First stop is a correlation matrix

First Stop: The Correlation Matrix

	submis~1	attended	submis~0	female	local	bkg_sc~e
submission~1	1.0000					
attended	0.3116*	1.0000				
	0.0154					
submission~0	0.6152*	0.2757*	1.0000			
	0.0000	0.0330				
female	0.0411	0.1336	-0.0166	1.0000		
	0.7549	0.3087	0.8996			
local	0.3545*	-0.0357	0.2792*	0.1001	1.0000	
	0.0054	0.7868	0.0308	0.4469		
bkg_science	0.5787*	0.3031*	0.6866*	-0.0334	0.2013	1.0000
	0.0000	0.0186	0.0000	0.8003	0.1229	

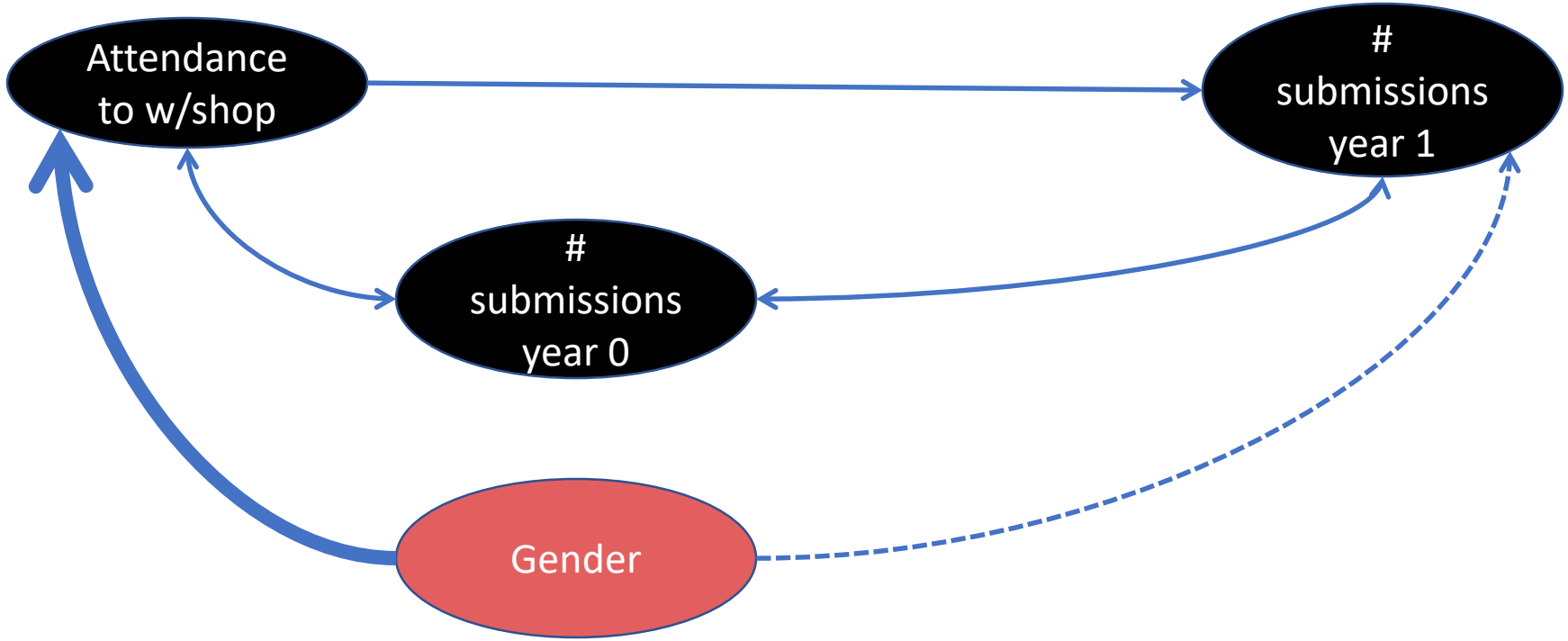
Gender differences



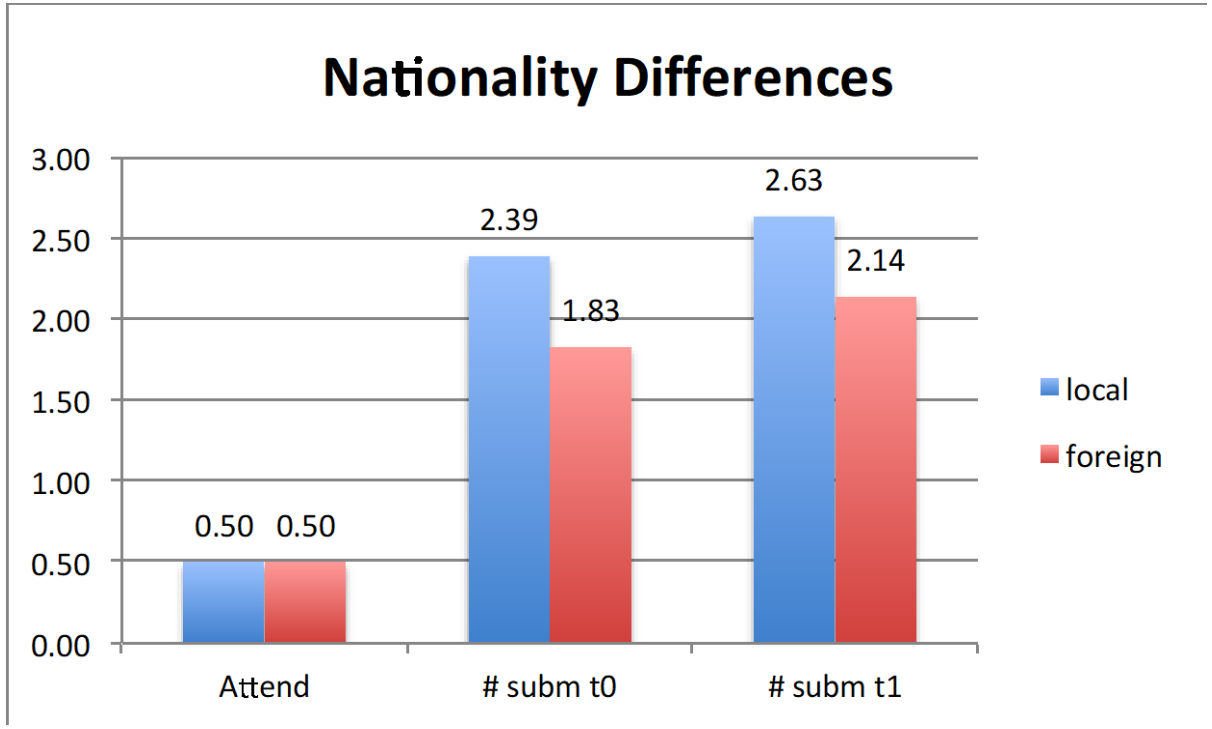
There are gender differences in terms of attendance but not in terms of submissions.

Not a control variable in the model

Model Specification 1



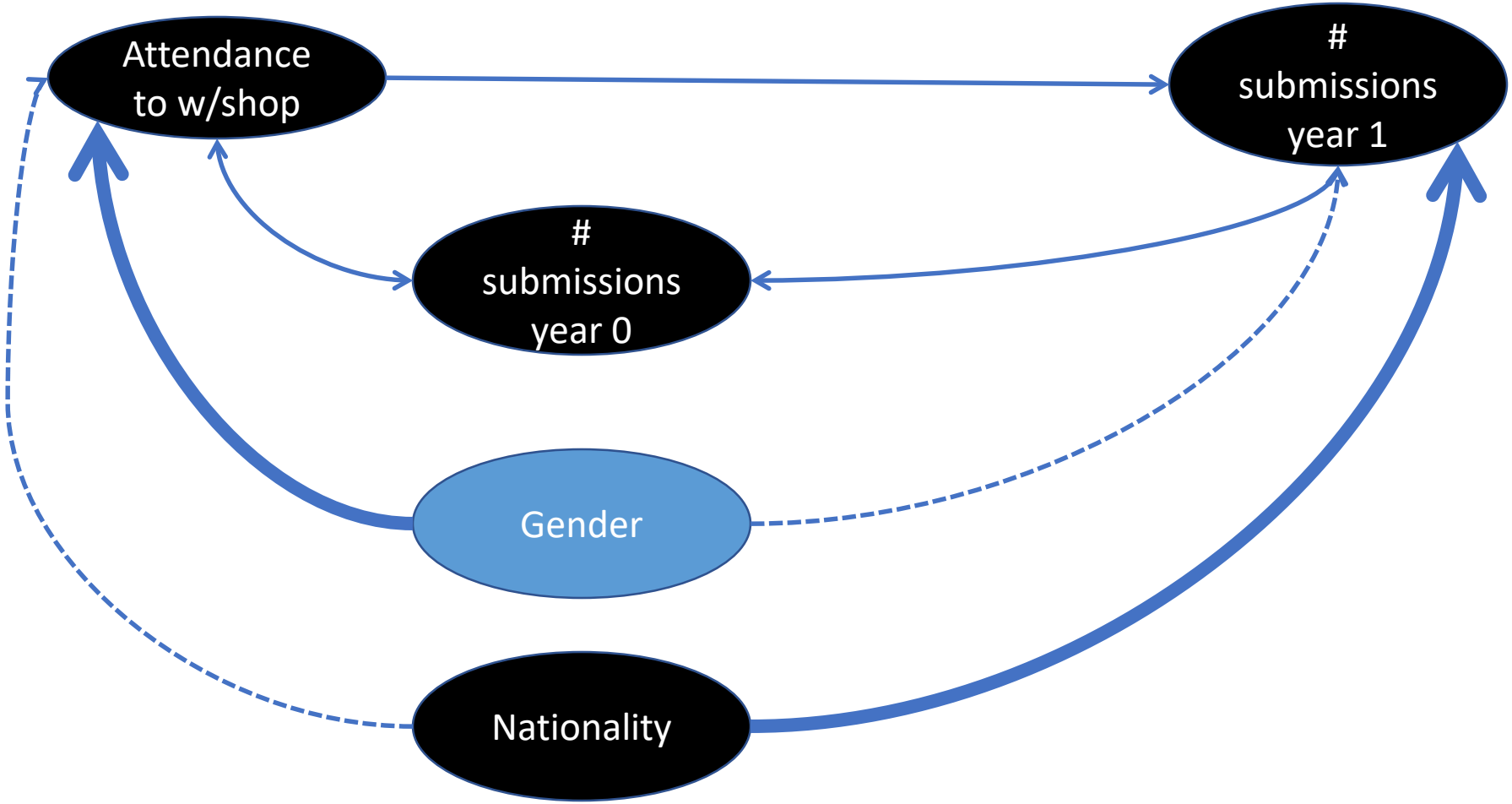
Nationality differences



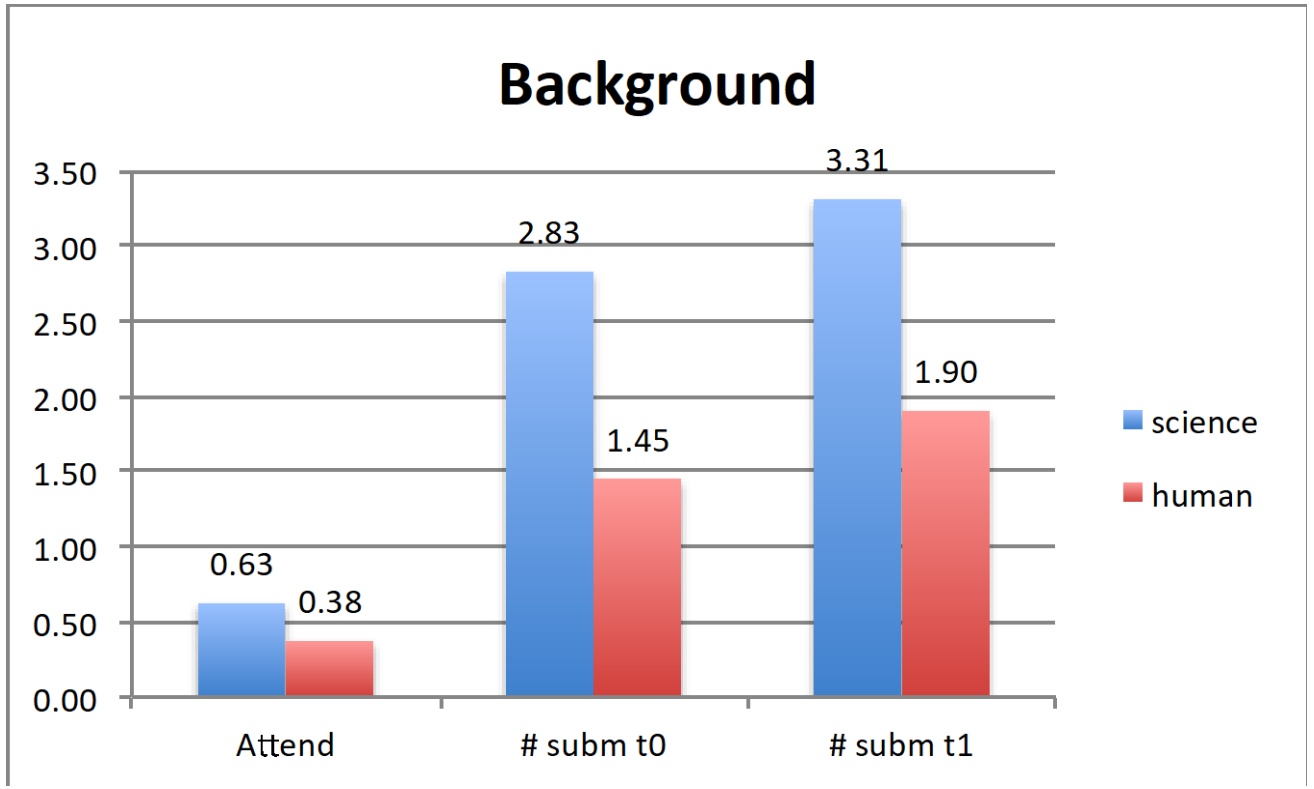
There are differences across nationalities in terms of performance but not attendance.

Advised to include nationality as a control variable to improve estimation precision (R^2) but not to address endogeneity issue. This is a commonly mis-placed criticism in empirical research

Model Specification 2



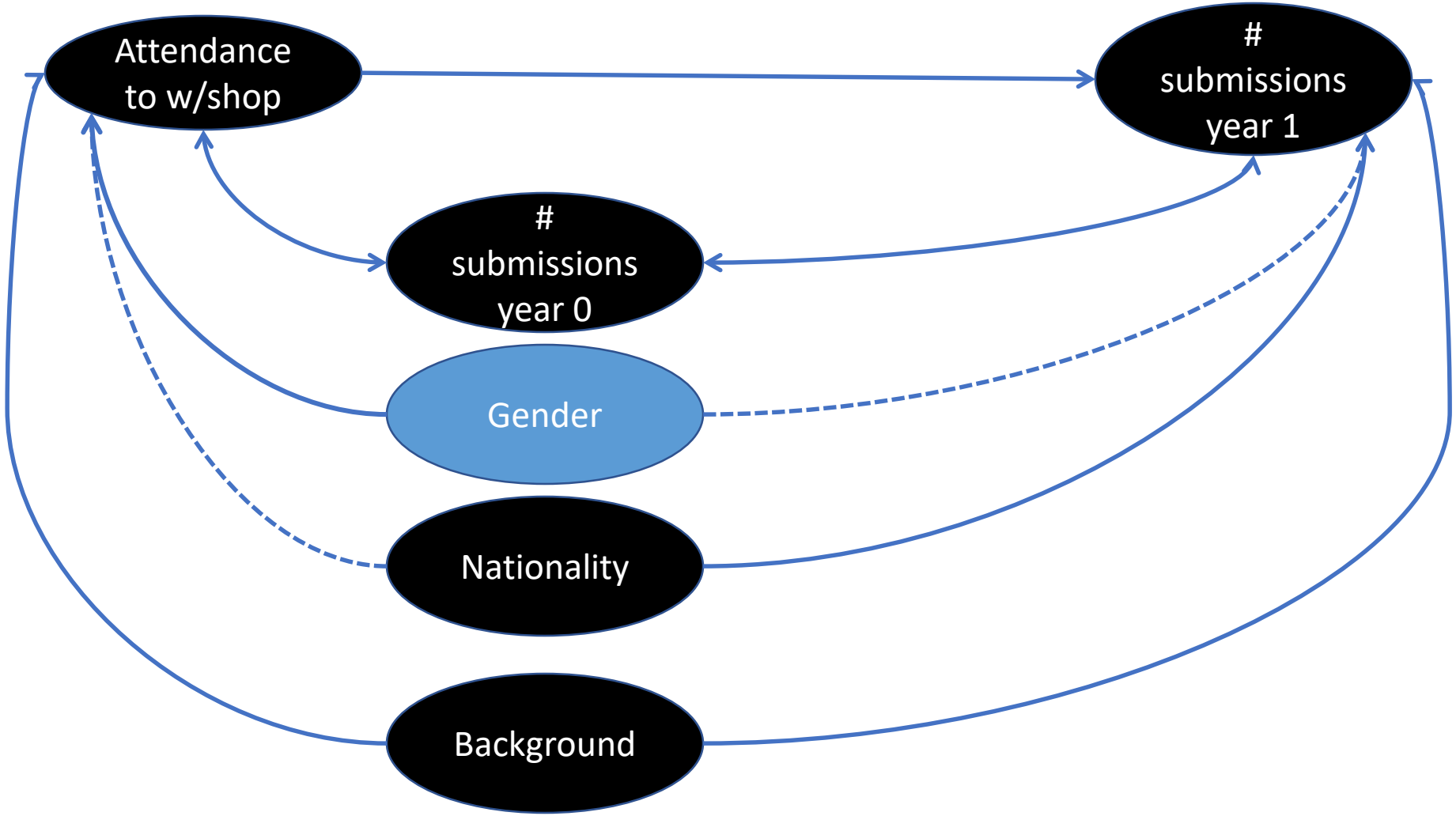
Background Differences



There are differences across background affecting both performance and attendance.

Background is a necessary covariate otherwise an omitted correlated variable problem arises

Model Specification 3



Model specification: selecting covariates

Covariate	Correlated with DV	Correlated with Main Predictor	Included in the model?	
Gender	No	No	No	Not necessary
Nationality	Yes	No	Yes	Improves estimation. Does not affect the bias
Bckg_Science	Yes	Yes	Yes	Biased estimation: omitted correlated variable
Submissions_T0	Yes	Yes	Yes	Biased estimation: reverse causality

Choosing the right covariates involves two steps:

- What is theoretically relevant? – prior literature
- What correlation patterns emerge in the data? – project specific

Towards a less biased model

Model	Dependent Variable	Main Predictor	Covariates	Inclusion/Exclusion
1	# journal submission T1	Attendance W/shops	Gender	Should not be included
2	# journal submission T1	Attendance W/shops	Nationality	Included to improve estimation precision
3	# journal submission T1	Attendance W/shops	Background	Must be included. Otherwise omitted correlated variable
4	# journal submission T1	Attendance W/shops	# journal submission T0	Must be Included. Otherwise reverse causality

$$Sub\_T1 = \alpha + \beta_1 Att + \beta_2 Bckg + \beta_3 Local + \beta_4 Sub\_T0 + \varepsilon$$

Model Comparisons

	Model 1	Model 2	Model 3	Model 4	Model 5
Attended	.76(3.56)	.76(3.52)	.80(4.02)	.43(2.34)	.35(1.99)
Female		-.001(-.01)	-.097(-.48)	.016(.009)	.036(.22)
Local			.90(4.52)	.64(3.62)	.51(2.96)
Background				1.15(6.15)	.63(2.77)
Submission_T0					.41(3.53)
Intercept	2.17(14.0)	2.18(12.01)	1.73(9.94)	1.45(8.31)	.94(4.21)
Adj-R-sqr	.097	.097	.21	.42	.48
	Biased	Biased	Bias – but improved precision	Bias – but control for omitted correlated variable	Unbiased – control for omitted correlated variable and selection on outcome

Potential sources of endogeneity

1. Reverse causality
2. Omitted correlated variable

Ex:

Asses if Sales F.casts issues in addition to Earnings F.casts improves info quality.

1. Reverse causality: analysts issue Sales forecasts when info is more precise (and feel more comfortable doing so)
2. Confounds: better analysts issue Sales F.cast (Sales F.cast do not really add value) and improved estimation

Endogeneity causes a mis-attribution problem

Conditions to Claim Causality

The gold standard requires a time machine…

Give someone a fast swimming suit (treatment) & take time on 100 m stroke then..

… go back in time, give the same subject a standard swimming suit and take the time

Can we do that? Not really … I guess ☺

We choose groups that are ***equal in expectations*** (on the outcome) & assign to treatment or control conditions

The techniques are useful to determine whether we can meet this condition or not.

Self-Selection: theory

Means that the observed units non-randomly assigned:

- Random assignment vs Choice

'Selection occurs when observations are non randomly sorted into discrete groups, resulting in the potential coefficient bias in estimation procedures such as OLS'

(Maddala 1991)

The choice of allocating or not a unit in a group may depend on attributes and characteristics that can potentially affect the outcome.

When do we worry?

Our RQs imply some causality:

- Establish whether X affects Y
- Test whether X has an impact on Y

Some examples:

- Do incentives to middle managers improve productivity?
- Does hiring a Big-N auditor improve audit quality?
- Do support structures affect employee outcomes (MIS Q)

The majority of our research does not imply causality nor we should claim it

A fundamental problem: our data

We are comparing units (e.g. individuals, groups, firms, systems) that are different:

- Along dimensions other than our treatment
- In ways that can affect our outcome

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Endogeneity: a technical note

$$Y = \alpha + \beta_1 X1 + \beta_2 Cov\_1 + \beta_3 Cov\_2 + \varepsilon$$

β_1 is unbiased only if COV (X1, ε) $\neq 0$

The unexplained portion of variance of DV is not correlated with X1 – this is hard to verify ex-post (see WAJ data)

Ex:

$$EARN\_t1 = \alpha + \beta_1 EDU\_t0 + \beta_2 Cov\_1 + \beta_k Cov\_k + \varepsilon$$

EDU is not randomly assigned: it depends on Ability and ex-ante wealth (EARN_t0)

Ok, we have a problem. And now?

Understand the source of endogeneity:

- Observable – data are available
  - OLS Multiple Regression
  - Stratification
  - Propensity Score Matching
  - Instrumental Variable Estimation
  - 2 Stages Least Squares

- Non-observable – no data available
  - 2 Stages Least Squares – Inverse Mills Ratio
  - Modeling change (fixed-effects)

Available instruments

Making sure that we can compare the groups and estimate a B-coeff that is unbiased and reflects the relationship between theoretical constructs

1. OLS – Multiple Regression
2. Propensity Score Matching (PSM)
3. Instrumental Variable Estimation (IVE)
4. 2 Stages Least Squares
5. Simultaneous Equation Modeling
6. Panel and fixed effects (modeling change)
7. Difference in Difference

Recap

1. Our RQ is causal like: X → Y – potentially impactful contribution

2. Observed data do not ensure randomization in the assignment of units to treatment/control groups to the (levels of) main predictor

3. Equality in expectations is not met: units self-select into treatment/control groups

4. Compare groups that are different along dimensions that affect our outcome

5. As a result: mis-attribution of an effect to the treatment

IVE & 2-Stage Least Squares: an overview

The equality in expectation condition is not met: $COV(X1, \varepsilon) \neq 0$

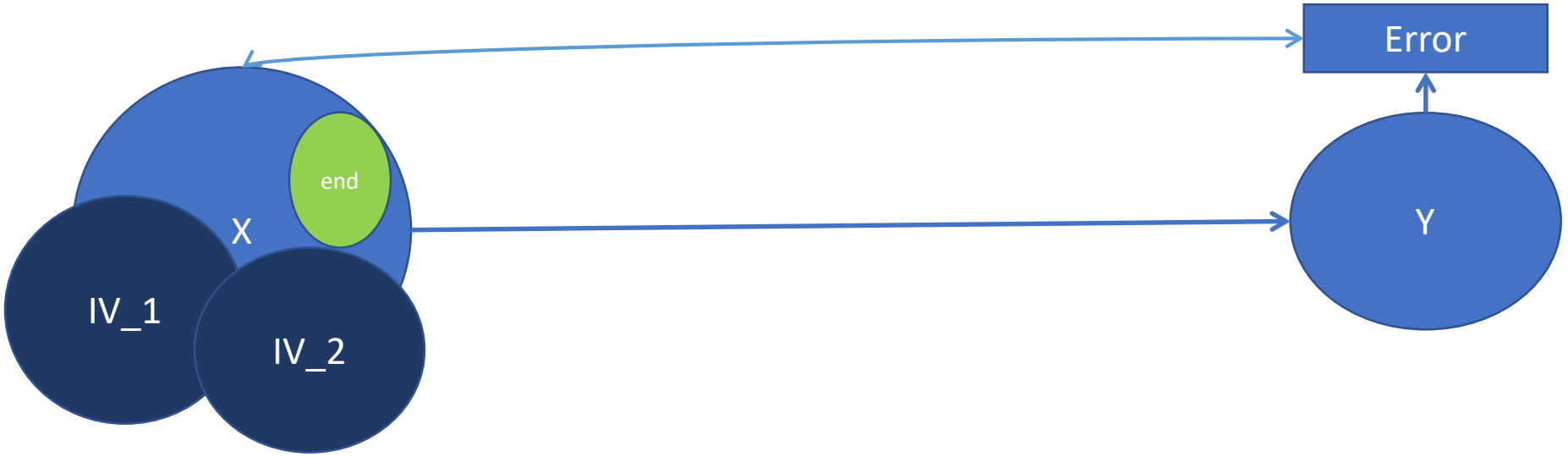
B-coeff on main predictor is biased because ε correlate with main predictor

A solution is finding another (or multiple) variable (Instrument) that:

- Correlates with the predictor
- Does not (conceptually) correlate with the outcome

How feasible is finding such a variable that satisfies both conditions?

What is an *Instrument*?



Instrument is correlated with X but not with Y

Example

Q: Does having a Health Insurance affect medical expense?

Use data from the US Social Security Survey - people >65 yrs old

OLS model: $Med\_Exp = \alpha + \beta_1 Health\_Insurance + \beta_k Controls + \varepsilon$

Subscribing a Health Insurance is a choice and our OLS estimation will suffer from endogeneity problems.

A number of tools help detecting endogeneity and ‘correct’ for it

Model Comparison

	OLS	2SLS (one IV)		2SLS (two IVs)	
		First Stage	Second Stage	First Stage	Second Stage
DV	Medical Expenses	Health Insurance	Medical Expenses	Health Insurance	Medical Expenses
Health Insurance	**.074 (2.89)**				
Health Insurance_HAT			**-.85 (-4.30)**		**-.96(-5.20)**
Illness	**.44 (47.06)**	**0.11(3.12)**	**.44 (43.59)**	**.011(3.25)**	**.44(43.12)**
Age	-.003 (-1.35)	**-.01(-11.97)**	**-.012 (-4.23)**	**-.008(-11.06)**	**-.012(-4.75)**
LN(Income)	0.17 (1.27)	**.05(9.64)**	**.097 (4.35)**	**.05(8.99)**	**.11(4.95)**
SSIncome (IV_1)		**-.19(-14.1)**		**-.19(-13.43)**	
Firm_Location (IV_2)				**.11(5.78)**	
Intercept	**5.7 (37.53)**	**.95(16.86)**	**6.58 (28.09)**	**.91(15.91)**	**6.69(29.97)**
Adj-R^2	.1749	.068	.079	.071	.049
Durbin-Chi2		25.09(p=0.000)		37.21(p=0.000)	
Wu-Husman		25.13(p=0.000)		37.40(p=0.000)	

The Mechanics of IVE and 2SLS

First Stage: $X = \alpha + \beta_1 IV\_1 + \beta_k IV\_k + \beta_n COVARIATES + \varepsilon$

The predicted values ($\bar{X}$) are exogenous and can be used as predictor in the Second Stage regression.

Second Stage: $Y = \alpha + \beta_1 \bar{X} + \beta_k COVARIATES + \delta$

Caveats:

- The higher the R^2 in 1st stage, the more robust the results – otherwise a weak instrument
- The stronger the theoretical and empirical support for the lack of correlation between IV and Y, the more convincing the model (no third path assumption)
- The instruments are not included in the second-stage (exclusion restriction)

IVE and 2SLS: a note of caution

These tools do not solve the endogenity problems – at best they mitigate it or indicate there is little we can do about it.

A few things help in the write up of the method/analysis section:

1. Understand the source of endogeneity (theory): what drives the selection?
2. Find a credible & robust Instrument (s) that predict a large portion of the Main Predictor (F-Stat>10) and high R^2
3. Specify alternative tests with different Instruments (first-stage) and exclusion restrictions (second-stage) and **justify the choice**
4. Report test for Multicollinearity from the second-stage regression

Propensity Score Matching (PSM): overview

- Observations are not randomly assigned to the treatment/control conditions
  - B-coeff will result in a biased under (over) estimation of the relationship IV → DV
- No Instrument available to predict our endogenous variable
- PSM means finding a ‘match’ for treated units (e.g. controls) to enable comparison and estimate the effect
- The more accurate the matching procedure, the better the estimation

The logic of PSM

Matching means ‘reverse engineering’ the randomization process.

Instead of assigning units to treatment/control – we identify controls ex-post.

This is necessary in two cases:

1. Data include treated and control units that have not been randomly assigned (e.g. kids attending Catholic or Public Schools)
2. Data include only treated units (e.g. firms subject to ATO investigation) but no controls

Q: How to find the right peers?

How to find non-treated units that are similar across dimensions that are likely to affect the outcome variable?

Example

In 80's a debate in the US about whether family should be given vouchers to enrol kids into Catholic schools.

Q: do Catholic schools enhance students' achievements?

The National Educational Longitudinal Studies (1988) has the following data:

- Student's attendance to Catholic or Public School
- Student's achievement on Math tests in yr8 and yr12
- Family-related variables (e.g. income)
- Parents' characteristics (e.g. education, employment conditions)

A substantive question of interest

A simple comparison across the two groups suggests that Catholic schools enhance students' ability (in Math tests)

- Catholic (N=592; Math_12= 54.53)
- Public (N=5079; Math_12=50.64)

Large and sig difference suggests that Catholic schools offer better education

Students are not randomly assigned to the type of Schools – self-selection

- Students attending Catholic or Public schools are not comparable.

Data Example

Model Comparison

	OLS	OLS with Family Income	OLS with Family Income Math8	PSM Average Treatment Effect	PSM Average Treatment on the Treated
Catholic	3.89(9.51)	2.68(6.82)	1.67(7.29)	2.66 (4.16)	1.66 (4.34)
Fam_Income		1.29(23.88)	.33(10.13)		
Math_8			.78(104.1)		
Intercept	50.64(382.81)	38.42(72.90)	7.15(16.72)		
Adj-R-sqrd	0.0157	0.11	0.69		

The mechanics of PSM

PSM does two things:

1. First, identify pairs of observations that are comparable. ‘Matching’ is based on our input and knowledge of the selection process.

PSM tells us that obs X_{123} (treated) and X_{456} (non-treated) are comparable across the chosen dimensions: the only difference is the treatment condition.

2. Compares ‘paired’ obs and returns an Average Effect of the Treatment

This gives a ‘bias-correct’ estimation of the effect of the treatment on the outcome

PSM: key issues to remember

1. Identification of peers? What drives the selection process? (theory)
   - Family Income
   - Previous Math Score in Yr 8

2. Specify strong selection model (propensity of going to Catholic school) – R^2

3. How many peers for a comparison? 1 or many?

Takeaways

1. Think about your data: are the groups randomly assigned to the (levels of) treatment? If not you need to have a bias-correction strategy
    - What are the potential confounds?
    - What is the direction of causality?

2. Carefully craft the research design ex-ante
    - Use natural or quasi-experiments (change in legislation, deaths, lottery)
    - Work with the limitations of the data and clearly discuss them in the paper/thesis

3. No statistical remedy helps without a theoretical knowledge of the issue

4. No remedy exists to endogeneity problems: OLS, 2SLS and PSM provide a way to correct the estimation bias. **Sometimes problems are insurmountable**

When to use which tool?

Source of Endogeneity	Examples today	Tool	PROs / CONs
Self-Selection (on the outcome)	Submission at T0 (in the WAJ example)	OLS with lag variable	Parsimonious but not effective. Only if data are non-stationary
	Math Score (in the Catholic School example)	IVE 2Stage Least Square	If a good Instrument is available
		Inverse Mills Ratio (Heckman procedure)	Data is not available
Selection on Observable Covariates	Background (WAJ example) Family Income (Catholic School Example)	OLS IVE & 2SLS PSM	The choice is entirely driven by the quality of the Instrument. OLS can be efficient as well. PSM or 2SLS as robustness tests
Selection on Unobservable Covariates		2SLS IMR Fixed Effect Analysis	All conclusions can be potentially biased

We are not alone…

Colleagues in a number of disciplines have similar problems and share solutions:

- Epidemiology
- Biostatistics
- Education
- Economics

Resources

U of Winsconsin Madison: http://www.ssc.wisc.edu/sscc/

Institute for Digital Research & Education @ UCLA https://idre.ucla.edu

Harvard School Pol Science http://gking.harvard.edu/category/research-interests/methods/causal-inference

End of Part 7