Omitted Variable Bias

TEXAS A&M UNIVERSITY
Mays Business School

Outline

- Endogeneity, Confounders, and Omitted Variable Bias
- Instrumental Variable (IV) Analysis and OVB
- Impact threshold of a confounding variable (ITCV)
- Concluding Thoughts and Recommendations

Capturing Causality in Organizational Research

Causality represents a challenge for organizational researchers

For example, the strategy field is focused on complex, interrelated factors that influence competitive advantage and firm performance (Bowman, Singh, & Thomas, 2002; Nag, Hambrick, & Chen, 2007)

When examining these relationships, scholars often rely almost exclusively on non-experimental design (Bergh et al., 2004; Bettis et al., 2014; Hamilton & Nickerson, 2003)

Because of this reality, researchers faces several empirical and analytical challenges

Casual Inference and Omitted Variable Bias

"Yes, but have you controlled for…"

(Frank, 2000: 149)

Endogeneity

- Endogeneity occurs when an independent variable is correlated with the error term in a statistical model

- The basic OLS regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$Cov(x_i, \varepsilon_i) \neq 0$$

- Endogeneity may derive from:
  - Omitted variable
  - Measurement error
  - Simultaneous causality
  - Sample Selection

Campbell's Threats to Validity in Econometric Terms

Threats to internal validity	Threats to external validity
• Omitted variables • Trends in outcomes • Misspecified variances • Mismeasurement • Political economy • Simultaneity • Selection • Attrition • Omitted interactions	• Interaction of selection and treatment • Interaction of setting and treatment • Interaction of history and treatment (Meyer, 1995)

Omitted Variable Bias and Confounding

- Omitted variable bias (OVB) arises from not including a relevant variable that belongs in the population model
- OVB = excluding a relevant variable = underspecifying the model = short regression = left out variables error (LOVE)
- Confounding is the bias caused by common causes of a treatment and outcome
  - Produces "spurious correlation" and biased results
- In observational studies, the goal is to avoid confounding
- Pervasive in the organizational research:
  - Director expertise on monitoring (confounding: motivation)
  - CEO influence on firm performance (confounding: managerial ability)
  - Resources and firm performance (confounding: external environment)
- No unmeasured confounding assumes that we've measured all sources of confounding.

Omitting ability when examining the effect of education on wages

- If the true model in the population is:

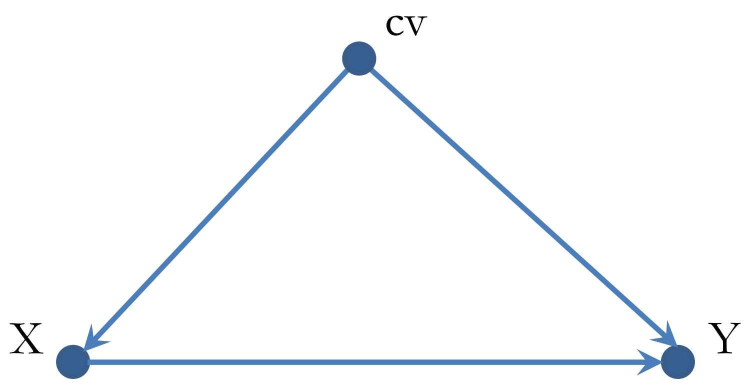
$$\boldsymbol{wages = \beta_0 + \beta_1 education + \beta_2 ability + \mu}$$

- Using a sample, what happens if we only model:

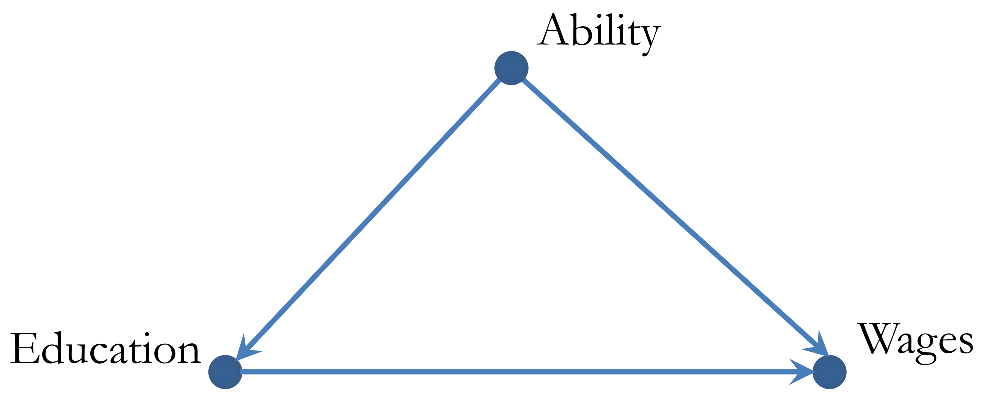
$$\boldsymbol{wages = \beta_0 + \beta_1 education_i + \mu_i}$$

- OVB describes the difference in regression estimates between these two equations (Angrist & Pischke, 2008)

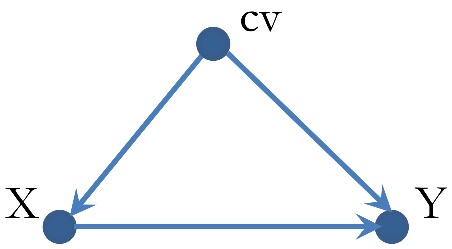
Confounding Variables in DAGs



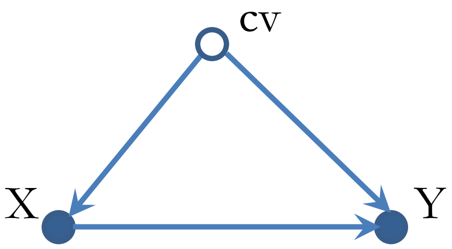
C is a confounder of the proposed causal relationship between X and Y



Two Graphs in which the Causal Effect of X on Y is Confounded by C

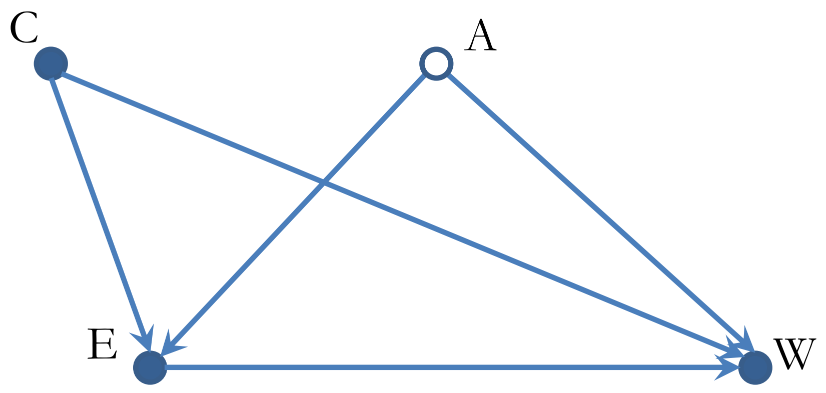


(a)



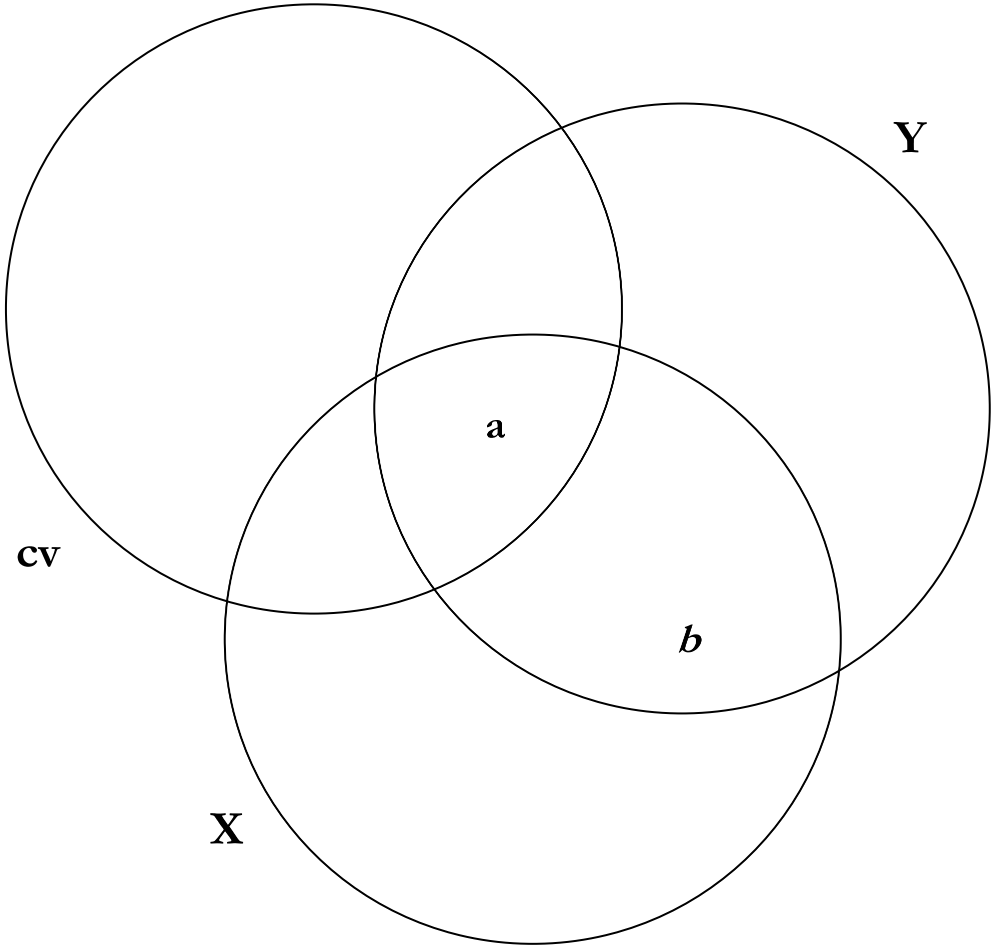
(b)

Expanding the DAG for Wages



$$Wages = \beta_0 + \beta_1 education + \beta_2 C + \beta_3 ability + \mu$$

A Ballantine Illustration of Education on Wages



- Back to the short regression equation:

$$wages = \beta_0 + \ \beta_1 education + \mu$$

- Circles Y and X represent variation $wages$ and $education$, respectively
- The area **a** + **b** represent the overlap in variation between Y and X
- However, the area **a** represent overlap in variation between X $(edcuation)$ and the omitted variable $ability$ and Y($wage$), which creates a correlation between the error term (μ) and X
- If Y were regressed on X, the info in the a + b would be used to estimate $\beta_1 education$.

TEXAS A&M UNIVERSITY
Mays Business School

Example: Omitting Ability when Examining the Effect of Education on Wages

$$\boldsymbol{wages = \beta_0 + \beta_1 education + \beta_2 ability + \mu}$$

Will both be positive

$$\boldsymbol{ability = \delta_0 + \delta_1 education + \vartheta}$$

$$\boldsymbol{wages = (\beta_0 + \beta_2\delta_0) + \boxed{(\beta_1 + \beta_2\delta_1)} + (\beta_2\vartheta + \mu)}$$

The return to wages β_1 will be overestimated because $\beta_2\delta_1 > 0$. It will look as if individuals with more education earn very high wages, but this is partly due to the fact that these individuals with more education also have greater ability on average.

Summary of Bias in $\widetilde{\beta_1}$

If the model in the population is:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

	$Cov(x_1, x_2) > 0$	$Cov(x_1, x_2) < 0$
$\beta_2 > 0$	Positive Bias	Negative Bias
$\beta_2 < 0$	Negative Bias	Positive Bias

When is OVB Not an Issue?

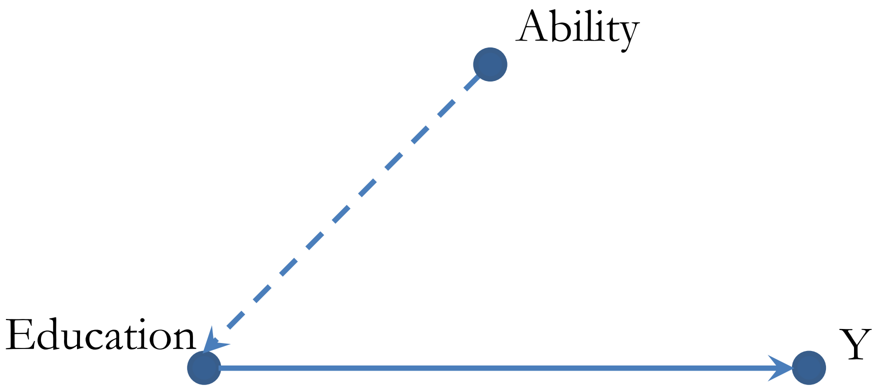
If the model in the population is:

$$\boldsymbol{wages = \beta_0 + \beta_1 education + \beta_2 ability + \varepsilon_i}$$

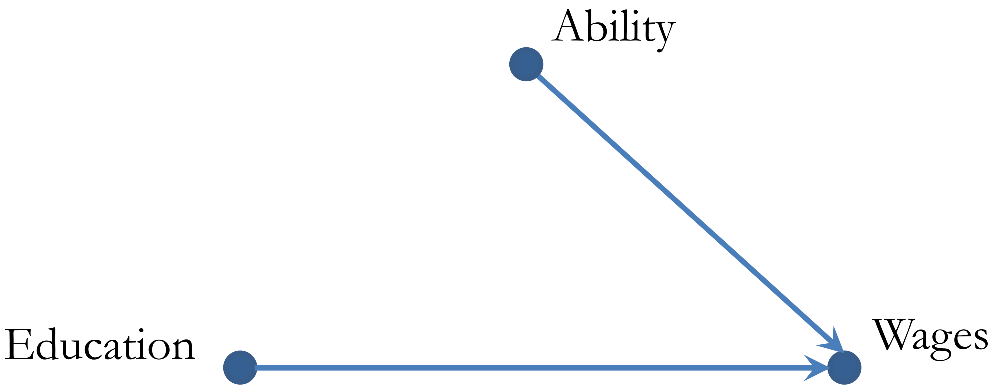
1. If $\beta_2 = 0$ in the population model, $\widetilde{\beta_1}$ is unbiased
2. If education and ability were uncorrelated, $\widetilde{\beta_1}$ is unbiased

When is OVB Not an Issue?

Condition 1: $\beta_2 = 0$ in the population model



Condition 2: Education and Ability are uncorrelated

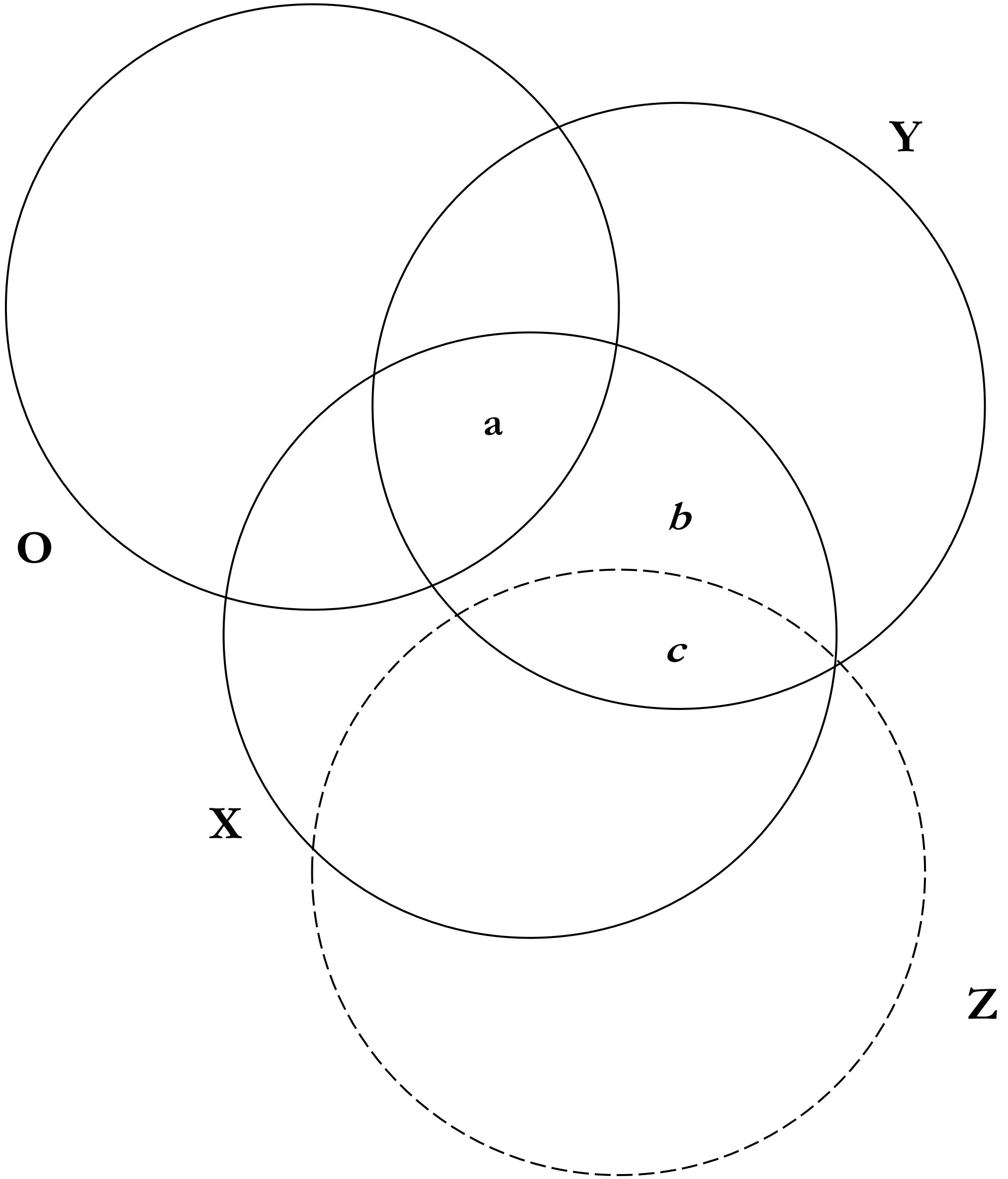


Instrumental Variable (IV) Analysis and OVB

OVB and IV Analysis

- Concern about OVB is often a key motivating reason for adopting instrumental variable techniques
- These techniques typically involve a two-step procedure
- While these techniques can help alleviate the OVB concern, they also have critical assumptions that must be met
- Even when these assumptions are met, instrumental variable techniques are often less efficient than OLS regression

A Ballantine Illustration of IV Analysis Logic



- The Z variable represents an IV
- Suppose X is regressed on Z. The predicted area, $\hat{X}$, is represented by the c area
- Now regress Y on $\hat{X}$ to produce an estimate of $\beta_1 eduction$
- In this case, area c is only used to form the estimate
- Since area c corresponds to variation in Y arising from variation in X, the resulting estimate of $\beta_1 eduction$ is unbiased

An IV Approach to Education on Wages

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

Education variable that is suspected to be endogenous

First stage (= reduced form regression):

The endogenous explanatory variable x_1 is predicted using only exogenous information

$$\hat{x}_1 = \hat{\alpha} + \hat{\alpha}_1 x_2 \cdots + \hat{\alpha}_{k-1} x_{k-1} + \hat{\alpha}_k z_k$$

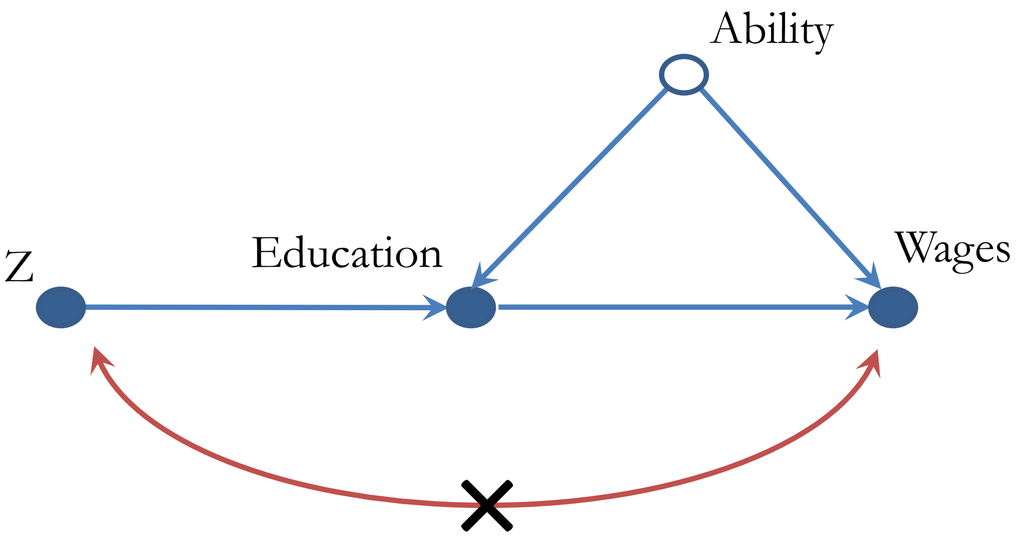
Exogenous variable = distance to a 4-year college

Second stage (= OLS with x_1 replaced by its prediction from the first stage):

$$y = \alpha + \beta_1 \hat{x}_1 + \beta_2 x_2 \cdots + \beta_{k-1} x_{k-1} + \varepsilon_i$$

Intuition: The ***predicted value*** is the value of the endogenous variable as a function of the exogenous instrument. This isolates the variance in the endogenous variable that is exogenous.

Basic IV Setup with DAGs



- Z is the instrument, *education* is the treatment, and *ability* is the unmeasured confounder
- Exclusion restriction
  - No common causes of the instrument and the outcome
  - No direct or indirect effect of the instrument on the outcome not through the *education*.
- First-stage relationship: $Z \longrightarrow$ *education*

Finding Instrumental Variables

Requirements:

- Excluded from the regression model
- Relevance (strong correlation with potential endogenous variable)
  - Multiple strong instruments optimal
- Exogeneity (NOT correlated with error)

Relevance: How Do We Know?

- In the first stage regression (where endogenous IV is the DV), the instrument(s) should explain a significant portion of the DV.

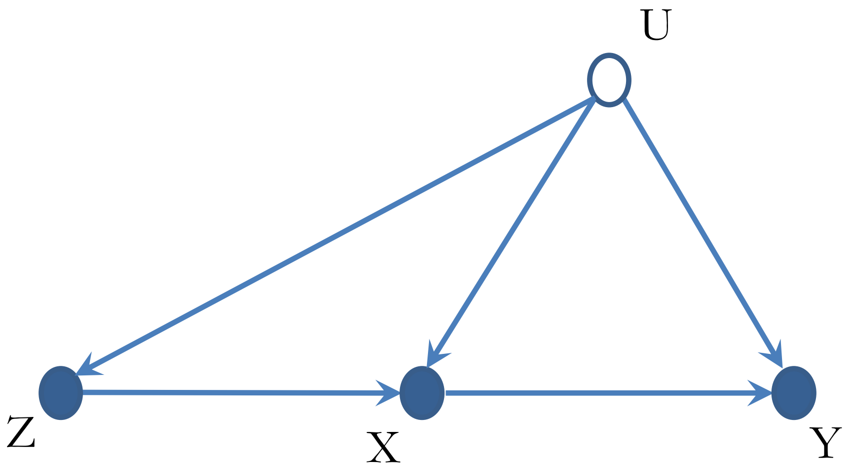
- F-statistics can be used to examine this assumption.

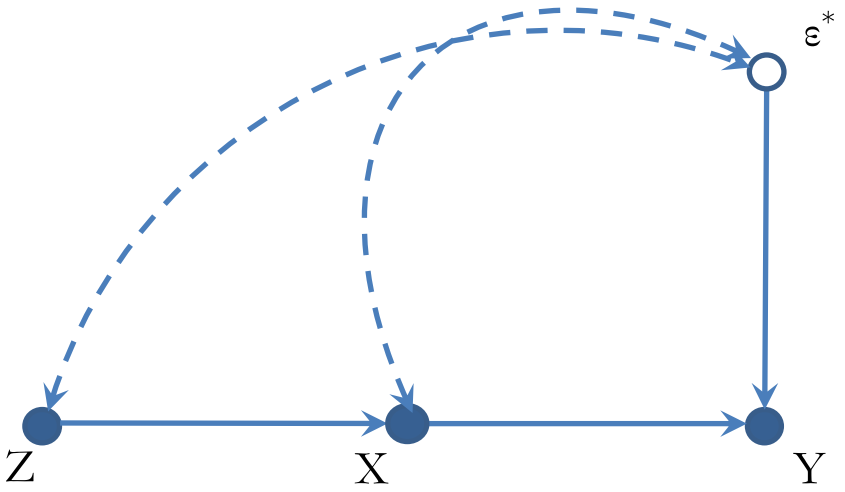
Instrumental Variables: Relevance

Table 1. Main findings OLS vs. single instrument

	OLS	Weak inst.	Moderate inst.	Weak and endo instr.	Moderate and endo instr.
PANEL A: Sample size: 500, True B: 0.1					
No endogeneity					
Beta	0.100	0.088	0.100	1.095	0.399
SE	0.045	0.469	0.138	0.633	0.141
95% interval	95%	100%	96%	84%	43%
% significant	60%	1%	9%	26%	84%
Low endogeneity (0.1)					
Beta	0.200	0.076	0.107	1.073	0.406
SE	0.045	0.478	0.135	0.598	0.139
95% interval	39%	100%	96%	78%	40%
% significant	69%	1%	12%	34%	86%
Medium endogeneity (0.3)					
Beta	0.430	0.097	0.092	1.142	0.404
SE	0.042	0.475	0.136	0.556	0.129
95% interval	0%	100%	95%	66%	36%
% significant	100%	4%	13%	45%	86%
PANEL B: Sample size: 500, True B: 0					
No endogeneity					
Beta	0.001	−0.009	−0.004	0.940	0.335
SE	0.045	0.474	0.137	0.610	0.157
95% interval	95%	100%	95%	84%	44%
% significant	5%	1%	5%	16%	56%
Low endogeneity (0.1)					
Beta	0.100	0.051	−0.009	1.016	0.336
SE	0.044	0.489	0.136	0.608	0.154
95% interval	39%	100%	95%	77%	40%
% significant	61%	0%	5%	23%	60%
Medium endogeneity (0.3)					
Beta	0.330	−0.001	−0.007	0.987	0.338
SE	0.042	0.466	0.135	0.536	0.143
95% interval	0%	98%	95%	63%	35%
% significant	100%	2%	5%	37%	65%

Exogeneity Assumption Violated in DAGs





TEXAS A&M UNIVERSITY
Mays Business School

Tests for Exogeneity

- Testing for exogeneity requires that the number of instruments exceeds the number of endogenous regressors (i.e., the equation is "overidentified").

Bascle (2008) summarizes three tests (and provides Stata code):

– The Sargan or Hansen J-statistic

– The Bassmann statistic

– The difference-in-Sargan statistic

*Note: a *failure* to reject each type of statistic means that the instruments can be considered exogenous.

Instrumental Variables: Exogeneity

Table 1. Main findings OLS vs. single instrument

	OLS	Weak inst.	Moderate inst.	Weak and endo instr.	Moderate and endo instr.
PANEL A: Sample size: 500, True *B*: 0.1					
No endogeneity					
Beta	0.100	0.088	0.100	1.095	0.399
SE	0.045	0.469	0.138	0.633	0.141
95% interval	95%	100%	96%	84%	43%
% significant	60%	1%	9%	26%	84%
Low endogeneity (0.1)					
Beta	0.200	0.076	0.107	1.073	0.406
SE	0.045	0.478	0.135	0.598	0.139
95% interval	39%	100%	96%	78%	40%
% significant	69%	1%	12%	34%	86%
Medium endogeneity (0.3)					
Beta	0.430	0.097	0.092	1.142	0.404
SE	0.042	0.475	0.136	0.556	0.129
95% interval	0%	100%	95%	66%	36%
% significant	100%	4%	13%	45%	86%
PANEL B: Sample size: 500, True *B*: 0					
No endogeneity					
Beta	0.001	−0.009	−0.004	0.940	0.335
SE	0.045	0.474	0.137	0.610	0.157
95% interval	95%	100%	95%	84%	44%
% significant	5%	1%	5%	16%	56%
Low endogeneity (0.1)					
Beta	0.100	0.051	−0.009	1.016	0.336
SE	0.044	0.489	0.136	0.608	0.154
95% interval	39%	100%	95%	77%	40%
% significant	61%	0%	5%	23%	60%
Medium endogeneity (0.3)					
Beta	0.330	−0.001	−0.007	0.987	0.338
SE	0.042	0.466	0.135	0.536	0.143
95% interval	0%	98%	95%	63%	35%
% significant	100%	2%	5%	37%	65%

Mays Business School

Where Do You Find Good IVs?

- Finding IVs that meet the relevance and exogeneity requirements is challenging
- This is the major obstacle with IV techniques
- If you only have one IV, you can never fully know if the criteria are met

"All instruments arrive on the scene with a dark cloud of invalidity hanging overhead. This cloud never goes entirely away, but researchers should chase away as much of the cloud as they can."

Murray (2006: 114)

Impact threshold of a confounding variable (ITCV)

Quantifying the OVB Problem

Impact threshold of a confounding variable (ITCV)

What is the minimum correlation between a confounder variable and the independent variable/dependent variable for OVB to have created a significant effect where none really exists?

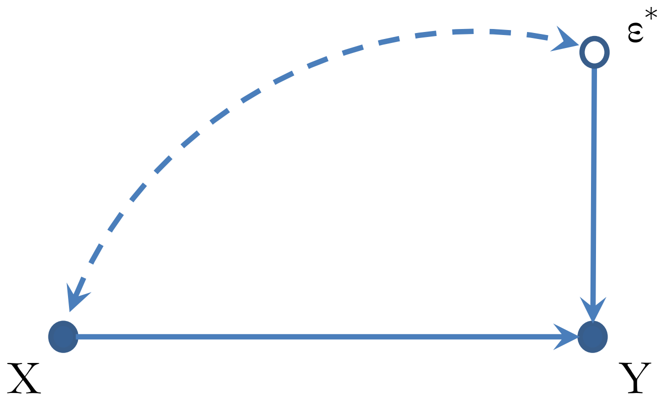
Causal Inferences

- Frank and colleagues suggest that you can make a causal inference from an observational study
- However, you just might be wrong
- The focus shifts from bias to changes in the causal inference
- But what would it take for the inference to be wrong?

Key steps

1. Establish Correlation Between predictor of interest and outcome
2. Define a Threshold for Inference
3. Calculate the Impact Necessary to Invalidate the Inference
4. Multivariate Extension, with other Covariates

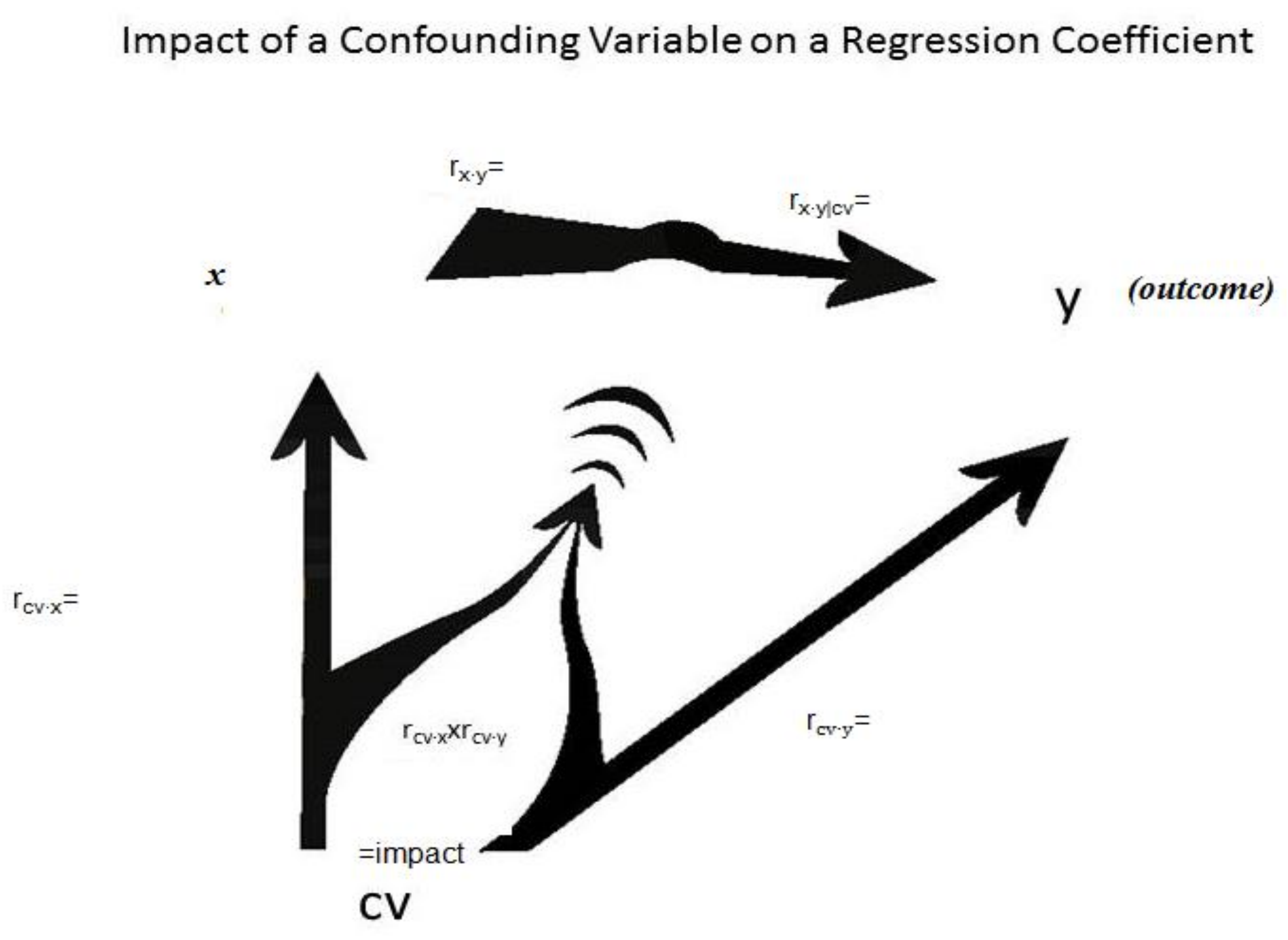
ITCV information



"In settings where **valid instruments are not available**, the question arises how to evaluate OLS estimates…how large does the **endogeneity problem** have to be to make the coefficient statistically insignificant?" (Larcker and Rusticus, 2010: 202)

TEXAS A&M UNIVERSITY
Mays Business School

ITCV Figure



Impact of a Confounding Variable on a Regression Coefficient
$r_{x \cdot y}$=
$r_{x \cdot y|cv}$=
x
y
(outcome)
$r_{cv \cdot x}$=
$r_{cv \cdot x} x r_{cv \cdot y}$
$r_{cv \cdot y}$=
=impact
cv

Influence of Different Inputs on the ITCV

	Increased sample size	Larger Coefficient for IV	Larger SE for IV
Effect on the ITCV	↓	↑	↓
Interpretation of the effect on the ITCV	Increasing the sample size decreases the ITCV since effects are more power with larger samples. This means that it *more likely* a confounding variable exists that can invalidate statistical inference	A larger coefficient for the independent variable increases the ITCV since it move the effect size further away from zero. This means that it is ***less likely*** a confounding variable exists that can invalidate statistical inference	A larger standard error for the independent variable decreases he ITCV since it sampling distribution of the coefficient. This means that it ***more likely*** a confounding variable exists that can invalidate statistical inference

ITCV Resources

- Frank (2000); Frank et al. (2013), Xu et al. (2019)
- Stata, SAS, R code for ITCV
- Kon-Found it! Excel sheet for ITCV
- Website:
  - https://msu.edu/~kenfrank/research.htm

The relationship between IQ and Wages

RQ: What is the relationship between IQ and wages

- Key variables:
- DV – Log Wages
- IV – IQ Score
- Controls – Completed Year of Schooling (s), Experience in Years (expr), Tenure in years (tenure), Residency in Southern US (rns), Reside in Metro Area (smsa), and Year Fixed Effects

Installing Stata Commands

```
ssc install konfound
ssc install moss
ssc install matsort
ssc install indeplist
```

Summary Statistics

```
use http://www.stata-press.com/data/imeus/griliches, clear

sum lw s expr tenure rns smsa iq med kww age mrt, sep(0)
```

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
          lw |        758    5.686739    .4289494      4.605      7.051
           s |        758    13.40501    2.231828          9         18
        expr |        758    1.735429    2.105542          0     11.444
      tenure |        758    1.831135     1.67363          0         10
         rns |        758    .2691293    .4438001          0          1
        smsa |        758    .7044855     .456575          0          1
          iq |        758    103.8562    13.61867         54        145
         med |        758    10.91029     2.74112          0         18
         kww |        758    36.57388    7.302247         12         56
         age |        758    21.83509    2.981756         16         30
         mrt |        758    .5145119    .5001194          0          1
```

Correlation

```
. corr lw s expr tenure rns smsa _I* iq
(obs=758)
```

	lw	s	expr	tenure	rns	smsa	_Iyea~67	_Iyea~68	_Iyea~69	_Iyea~70	_Iyea~71	_Iyea~73
lw	1.0000											
s	0.5027	1.0000										
expr	0.0846	-0.2418	1.0000									
tenure	0.1638	-0.0496	0.2307	1.0000								
rns	-0.1496	-0.0648	0.0058	-0.0366	1.0000							
smsa	0.2156	0.1021	-0.0332	0.0331	-0.1611	1.0000						
_Iyear_67	-0.1920	-0.1104	-0.0819	-0.0668	-0.0749	0.0065	1.0000					
_Iyear_68	-0.1261	-0.0794	-0.1842	-0.1230	-0.0220	-0.0062	-0.1027	1.0000				
_Iyear_69	0.0063	-0.0027	-0.1651	-0.1066	0.0012	-0.0447	-0.1070	-0.1212	1.0000			
_Iyear_70	0.0834	0.0704	-0.1541	0.0023	-0.0452	0.0511	-0.0914	-0.1036	-0.1079	1.0000		
_Iyear_71	0.1386	0.0955	0.1844	0.0496	-0.0069	-0.0426	-0.1119	-0.1268	-0.1321	-0.1129	1.0000	
_Iyear_73	0.3817	0.4398	0.0565	-0.0569	0.0548	0.0761	-0.1545	-0.1750	-0.1824	-0.1558	-0.1907	1.0000
iq	0.3471	0.5131	-0.1663	0.0194	-0.1339	0.0992	-0.0516	-0.0675	0.0087	0.1165	0.0698	0.2021

Regression results

```
reg lw s expr tenure rns smsa _I* iq

      Source |       SS           df       MS      Number of obs   =       758
-------------+----------------------------------   F(12, 745)      =     46.86
       Model |  59.9127611        12  4.99273009   Prob > F        =    0.0000
    Residual |  79.3733888       745  .106541461   R-squared       =    0.4301
-------------+----------------------------------   Adj R-squared   =    0.4210
       Total |   139.28615       757  .183997556   Root MSE        =    .32641

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .0619548   .0072786     8.51   0.000     .0476658    .0762438
        expr |   .0308395   .0065101     4.74   0.000     .0180592    .0436198
      tenure |   .0421631   .0074812     5.64   0.000     .0274763    .0568498
         rns |  -.0962935   .0275467    -3.50   0.001    -.1503719   -.0422151
        smsa |   .1328993   .0265758     5.00   0.000     .0807268    .1850717
   _Iyear_67 |  -.0542095   .0478522    -1.13   0.258    -.1481506    .0397317
   _Iyear_68 |   .0805808   .0448951     1.79   0.073    -.0075551    .1687168
   _Iyear_69 |   .2075915   .0438605     4.73   0.000     .1214867    .2936963
   _Iyear_70 |   .2282237   .0487994     4.68   0.000      .132423    .3240245
   _Iyear_71 |   .2226915   .0430952     5.17   0.000     .1380889     .307294
   _Iyear_73 |   .3228747   .0406574     7.94   0.000     .2430579    .4026915
          iq |   .0027121   .0010314     2.63   0.009     .0006873    .0047369
       _cons |   4.235357   .1133489    37.37   0.000     4.012836    4.457878
------------------------------------------------------------------------------
```

Stata konfound Output

```
. quietly: reg lw s expr tenure rns smsa _I* iq

. konfound iq
------------------
The Threshold for % Bias to Invalidate/Sustain the Inference

For iq:
To invalidate the inference 25.34% of the estimate would have to be due to bias; to invalidate the
inference 25.34% (192) cases would have to be replaced with cases for which there is an effect of 0.
------------------
Impact Threshold for Omitted Variable

For iq:
An omitted variable would have to be correlated at 0.161 with the outcome and at 0.161 with the predictor
of interest (conditioning on observed covariates) to invalidate an inference.
Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be
0.161 x 0.161=0.0260 to invalidate an inference.

These thresholds can be compared with the impacts of observed covariates below.

Observed Impact Table for iq
```

Raw	Cor(v, X)	Cor(v, Y)	Impact
s	.5131	.5027	.258
_Iyear_73	.2021	.3817	.0771
smsa	.0992	.2156	.0214
rns	-.1339	-.1496	.02
_Iyear_67	-.0516	-.192	.0099
_Iyear_70	.1165	.0834	.0097
_Iyear_71	.0698	.1386	.0097
_Iyear_68	-.0675	-.1261	.0085
tenure	.0194	.1638	.0032
_Iyear_69	.0087	.0063	.0001
expr	-.1663	.0846	-.0141

Partial	Cor(v, X)	Cor(v, Y)	Impact
s	.4033	.3564	.1437
_Iyear_70	.0867	.1767	.0153
rns	-.1077	-.1373	.0148
tenure	.0603	.2071	.0125
_Iyear_71	.0544	.1904	.0104
_Iyear_73	.0302	.2809	.0085
smsa	.0339	.1826	.0062
_Iyear_69	.0306	.173	.0053
_Iyear_67	.0137	-.04	-.0005
_Iyear_68	-.0101	.0644	-.0006
expr	-.0608	.1649	-.01

```
X represents iq, Y represents lw, v represents each covariate.
First table is based on unconditional correlations, second table is
based on partial correlations.
```

Spreadsheet

User enters values in yellow

KonFound-it! ©: BASICS

unstandardized estimated effect	standard error	sample size	# of covariates	α (significance level)	null hypothesis
0.0027121	0.0010314	758	12	0.05	0

User enters values in yellow

For a 1-tailed test, double the size of α in cell E4

name of predictor of interest: IQ

Calculated Values

df	default t critical (2-tailed)	t critical
745	1.963	1.963

To override cell C8, type in your own value

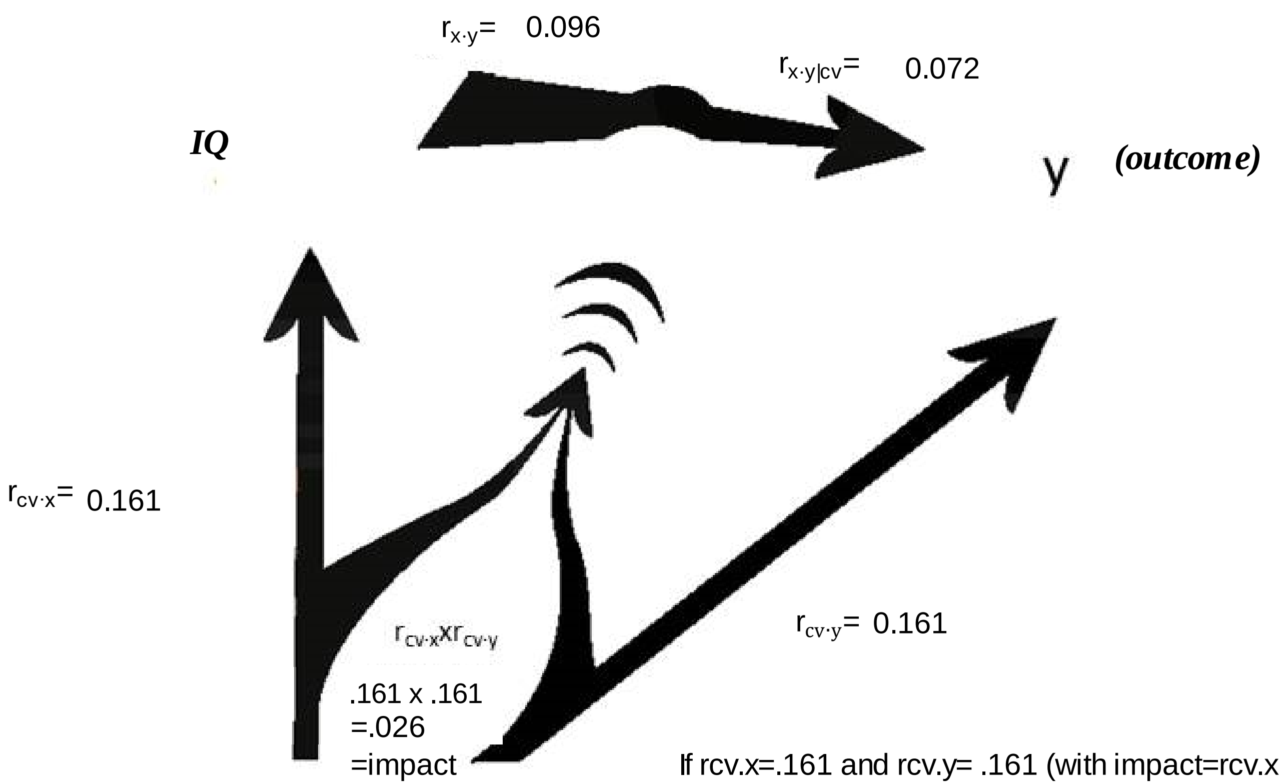
The default sign of t critical is the same as the sign of the estimated effect.

Publishable statements

Replacement of Cases	Correlation Based (linear models only)
To invalidate the inference 25% (192) of the cases would have to be replaced with cases for which there is an effect of zero.	To invalidate the inference an omitted variable would have to be correlated at .161 with IQ and at .161 with the outcome, conditional on covariates.

Impact Figure

Impact of a Confounding Variable on a Regression Coefficient



CV

(omitted confounding variable)

If rcv.x=.161 and rcv.y= .161 (with impact=rcv.x * rcv.y=.026) then the rx.y of .096 would become rx.y|cv of .072 if one were to control for the confound (cv). Therefore because .072 is the threshold for making an inference, To invalidate the inference an omitted variable would have to be correlated at .161 with IQ and at .161 with the outcome, conditional on covariates.

Konfound-it App

KonFound-It!

Quantify the Robustness of Causal Inferences

Click here to view for updates about upcoming workshops and other news through the KonFound-it! mailing list.

KonFound-It! takes four values from many statistical analyses - the estimated effect (such as an unstandardized regression coefficient or the group mean difference), its standard error, the number of observations, and the number of covariates (and, for non-linear models, an additional value). KonFound-It returns output in the forms of publishable statements as well as figures to support the interpretation of the output.

Menu | Linear Models | Non-linear Models (Beta) | Blog | More

Change or set any of the values below and then click run to see output from KonFound-It!

Estimated Effect

.0027121

Standard Error

.0010314

Number of Observations

758

Number of Covariates

12

Note that decimals must be denoted with a period, e.g., 2.1

RUN

Results (Printed) | Threshold Plot | Correlation Plot

Percent Bias Necessary to Invalidate the Inference:

To invalidate an inference, 25.342% of the estimate would have to be due to bias. This is based on a threshold of 0.002 for statistical significance (alpha = 0.05). To invalidate an inference, 192 observations would have to be replaced with cases for which the effect is 0.

See Frank et al. (2013) for a description of the method

Citation: Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. 2013. What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. Education, Evaluation and Policy Analysis, 35 437-460.

Impact Threshold for a Confounding Variable:

The minimum impact to invalidate an inference for a null hypothesis of 0 effect is based on a correlation of 0.161 with the outcome and at 0.161 with the predictor of interest (conditioning on observed covariates) based on a threshold of 0.072 for statistical significance (alpha = 0.05). Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be 0.161 X 0.161 = 0.026 to invalidate an inference for a null hypothesis of 0 effect.

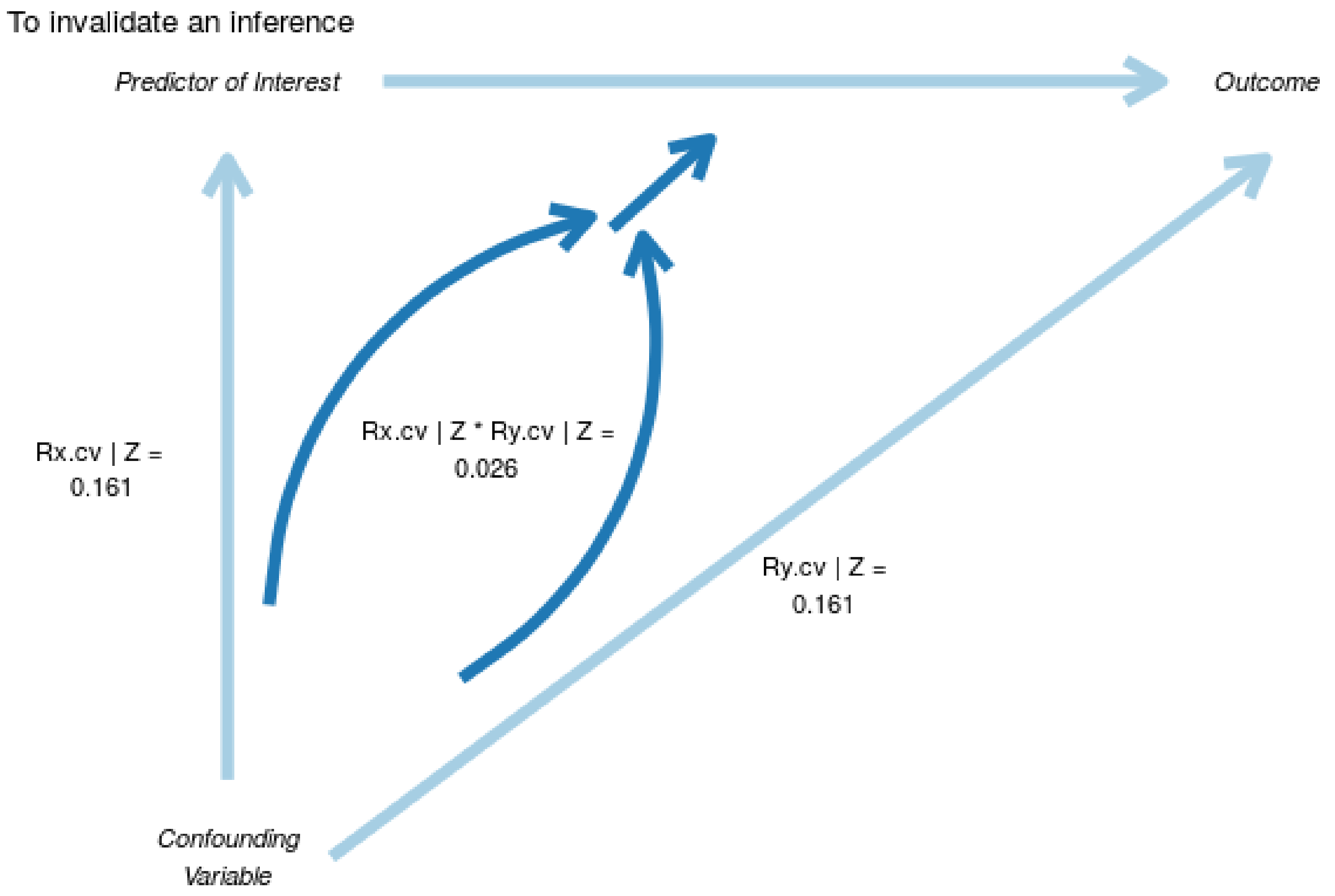
See Frank (2000) for a description of the method

Citation: Frank, K. 2000. Impact of a confounding variable on the inference of a regression coefficient. Sociological Methods and Research, 29 (2), 147-194

- Published empirical examples
- Full publishable write-up (replacement of cases)
- Full publishable write-up (correlation)

http://konfound-it.com/

Impact Figure from Konfound-it App



TEXAS A&M UNIVERSITY
Mays Business School

Study on ITCV's application in management

- Is OVB as big of a problem in management research as extant scholarship suggests?
- Impact Threshold of a Confounding Variable in management journals
- Re-Examining the Semadeni et al. (2014) Simulation

ITCV Values from a Content Analysis of Management Journals

		Panel A: Summary Results				Panel B: Covariates Higher than ITCV										Panel C: ITCV Value					
		Minimum correlation		Path correlation		Minimum correlation					Path correlation										
Journal Outlet	Number of Relationships	% Biased: One Covariate	% Biased: Two Covariates	% Biased: One Covariate	% Biased: Two Covariates	10%	25%	Median	75%	90%	10%	25%	Median	75%	90%	10%	25%	Median	75%	90%	Mean
AMJ	130	26.15%	13.08%	39.23%	26.15%	0	0	0	1	2	0	0	0	2	3	0.040	0.134	0.238	0.415	0.609	0.286
JAP	94	20.21%	11.70%	29.79%	19.15%	0	0	0	0	2	0	0	0	1	2	0.126	0.191	0.345	0.483	0.671	0.367
JOM	47	21.28%	12.77%	38.30%	21.28%	0	0	0	0	2	0	0	0	1	3	0.058	0.193	0.274	0.407	0.603	0.323
SMJ	111	35.14%	23.42%	48.65%	34.23%	0	0	0	1	3	0	0	0	2	4	0.042	0.085	0.148	0.295	0.463	0.211
Overall	**382**	**26.70%**	**15.71%**	**39.53%**	**26.18%**	**0**	**0**	**0**	**1**	**2**	**0**	**0**	**0**	**2**	**3**	**0.046**	**0.132**	**0.245**	**0.400**	**0.597**	**0.288**
Macro	172	33.14%	20.35%	46.51%	31.98%	0	0	0	1	3	0	0	0	2	4	0.045	0.096	0.171	0.308	0.494	0.232
Micro	210	21.43%	11.90%	33.81%	21.43%	0	0	0	0	2	0	0	0	1	3	0.067	0.176	0.310	0.460	0.638	0.335

Notes: "% Biased: One Covariate" and "% Biased: Two Covariates" reflect the proportion of total relationships that are potentially biased if at least one or two control variables, respectively, features the properties of correlations higher than the ITCV value.

Revisiting the Simulation

- Original simulation – Semadeni et al. (2014)
  - Correlated the independent variable with the error term
    - 0.10 = "Low endogeneity"
    - 0.30 = "Moderate endogeneity"
  - Error term represents about 90% of the variance
    - Vector of omitted variables or REALLY strong omitted variable

$$y = \alpha + \beta_{1x1} + \underbrace{\quad\quad e}_{\text{Corr}[x,e] = 0.1 \text{ or } 0.3}$$

Corr[x,e] = 0.1 or 0.3

Revisiting the Simulation

- Original simulation – Semadeni et al. (2014)
  - Correlated the independent variable with the error term
    - 0.10 = "Low endogeneity"
    - 0.30 = "Moderate endogeneity"
  - Error term represents about 90% of the variance
    - Vector of omitted variables or REALLY strong omitted variable

$$y = \alpha + \underbrace{\beta_{1}x_{1} + \beta_{2}x_{2}}_{} + e$$

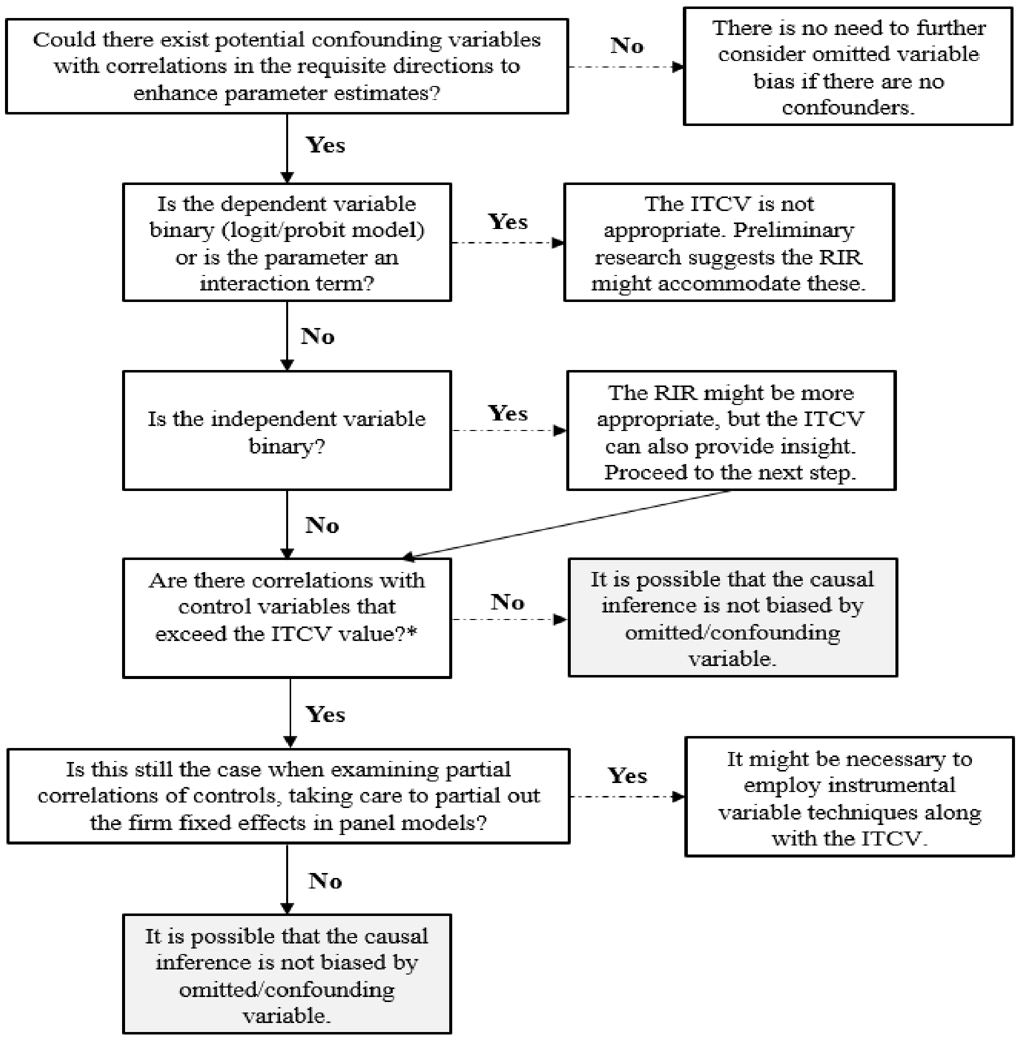
Corr[x1,x2] =

- 0.05 (10th percentile)
- 0.10 (25th percentile)
- 0.15 (50th percentile)
- 0.25 (75th percentile)
- 0.40 (90th percentile)

Simulation Results

Regression Model	Median	10th Pctile	25th Pctile	75th Pctile	90th Pctile	StDev.
Panel A (r=0.00)						
OLS (Fully Specified)	0.127	0.057	0.089	0.162	0.191	0.053
OLS (Not Fully Specified)	0.126	0.058	0.089	0.163	0.189	0.053
2SLS w/ Weak Instruments	0.101	-1.882	-0.710	0.871	1.884	4.104
2SLS w/ Moderate Instruments	0.133	-0.178	-0.023	0.279	0.436	0.239
2SLS w/ Strong Instruments	0.128	0.009	0.071	0.185	0.239	0.091
Panel C (r=0.10)						
OLS (Fully Specified)	0.122	0.057	0.088	0.159	0.193	0.053
OLS (Not Fully Specified)	0.137	0.072	0.103	0.172	0.206	0.053
2SLS w/ Weak Instruments	0.071	-1.581	-0.604	0.906	2.102	2.886
2SLS w/ Moderate Instruments	0.129	-0.170	-0.032	0.295	0.451	0.251
2SLS w/ Strong Instruments	0.120	0.003	0.054	0.180	0.232	0.089
Panel E (r=0.25)						
OLS (Fully Specified)	0.129	0.056	0.092	0.165	0.196	0.054
OLS (Not Fully Specified)	0.163	0.092	0.126	0.195	0.227	0.052
2SLS w/ Weak Instruments	0.113	-1.732	-0.719	0.855	2.075	2.580
2SLS w/ Moderate Instruments	0.108	-0.196	-0.045	0.274	0.422	0.245
2SLS w/ Strong Instruments	0.125	0.020	0.073	0.189	0.245	0.088
Panel F (r=0.40)						
OLS (Fully Specified)	0.127	0.051	0.088	0.165	0.200	0.059
OLS (Not Fully Specified)	0.181	0.111	0.146	0.217	0.245	0.052
2SLS w/ Weak Instruments	0.239	-1.750	-0.586	1.009	2.035	3.318
2SLS w/ Moderate Instruments	0.129	-0.168	-0.022	0.291	0.429	0.235
2SLS w/ Strong Instruments	0.127	0.011	0.065	0.188	0.242	0.091

Decision Tree of When to Employ the ITCV



- It is important to reiterate that the ITCV represents the square root of the product of correlations between a potential omitted variable and both the independent and dependent variables. This is therefore the case when examining control variables as potential proxies for an omitted variable. Specifically, it is essential to compare the square root corr[control, y] times corr[control, x] against the ultimate ITCV value.

TEXAS A&M UNIVERSITY
Mays Business School

Limitations of the ITCV

- The focus is on causal inference, not effect sizes
  - Provides information about statistical significance, not coefficient interpretation
  - Endogeneity can still bias the coefficients
  - Relies on an understanding of the actual DGP (true population model)
- IV techniques may necessary if translated effect sizes matter

Limitations of the ITCV (cont.)

- A confounder may exist that exhibits a correlation with the IV and DV at values greater than control covariates
- The appropriateness of the ITCV also depends on the nature of the empirical estimation procedure and corresponding data
- The ITCV is also currently unable to address interaction terms because the marginal effects of the relationships are contingent on the values of the lower order constituents

Summary

- Confounders and OVB are a major concern in organizational research
- Instrumental Variables Techniques can address the issue, but key assumptions must be met
- The ITCV may help to understand how statistical inferences change because of OVB

Key References

- Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. Strategic organization, 6(3), 285-327.Busenbark, J. R., Frank, K. A., Maroulis, S. J., Xu, R., & Lin, Q. (2021). Quantifying the robustness of empirical inferences in strategic management: The impact threshold of a confounding variable and robustness of inference to replacement. In Research in Times of Crisis. Emerald Publishing Limited.
- Busenbark, J. R., Yoon, H., Gamache, D. L., & Withers, M. C. (2022). Omitted variable bias: examining management research with the Impact Threshold of a Confounding Variable (ITCV). *Journal of Management*, 48(1), 17-48.
- Frank, K. A. (2000). Impact of a Confounding Variable on a Regression Coefficient. *Sociological Methods & Research, 29*(2), 147-194.
- Frank, K. A., Lin, Q., Maroulis, S. J., Mueller, A. S., Xu, R., Rosenberg, J., . . . Zhang, L. (forthcoming). Hypothetical Case Replacement Can Be Used to Quantify the Robustness of Trial Results. *Journal of clinical epidemiology*.
- Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What Would It Take to Change an Inference? Using Rubin's Causal Model to Interpret the Robustness of Causal Inferences. *Educational Evaluation and Policy Analysis, 35*(4), 437-460.
- Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Extended Influence: National Board Certified Teachers as Help Providers. *Education, Evaluation, and Policy Analysis, 30*(1), 3-30.
- Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2021). Endogeneity: A Review and Agenda for the Methodology-Practice Divide Affecting Micro and Macro Research. *Journal of Management*, 0149206320960533.
- Kennedy, P. (2008). *A Guide to Econometrics* (2nd Edition ed.). Oxford, UK: Blackwell.
- Larcker, D. F., & Rusticus, T. O. (2010). On the Use of Instrumental Variables in Accounting Research. *Journal of Accounting and Economics, 49*(3), 186-205.
- Semadeni, M., Withers, M. C., & Certo, S. T. (2014). The Perils of Endogeneity and Instrumental Variables in Strategy Research: Understanding through Simulations. *Strategic Management Journal, 35*(7), 1070-1079.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT press.
- Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*: Nelson Education.
- Xu, R., & Frank, K. A. (forthcoming). Sensitivity Analyses for Network Observations with Applications to Inferences of Social Influence Effects. *Network Science*.
- Xu, R., Frank, K. A., Maroulis, S. J., & Rosenberg, J. M. (2019). Konfound: Command to Quantify Robustness of Causal Inferences. *The Stata Journal, 19*(3), 523-550.