KAIST Summer Session 2018

Module 1. Research Design for Data Analytics

Identification Strategy (2) Instrument Variable

KAIST College of Business

Jiyong Park

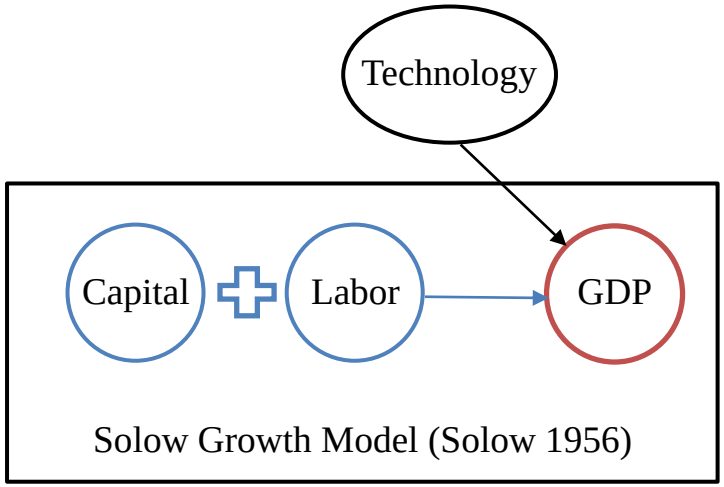
5 July, 2018

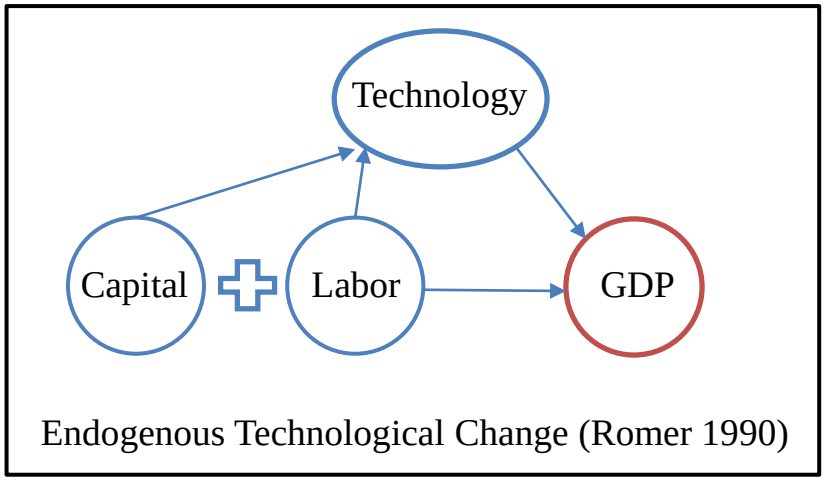
KAIST COLLEGE OF BUSINESS

Sources of Endogeneity

What is Endogeneity?

- Endogeneity (내생성)
  - ➢ Endogeneity issues arise when the focal variables are endogenously determined within a system of interest.
  - ➢ Example



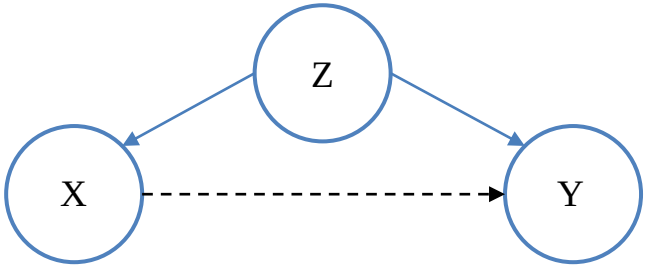


  - ➢ It could lead to a spurious correlation, raising concerns about causality.
  - ➢ Statistically, the variables X_i are endogenous if $Cov(X_i, \varepsilon_i) \neq 0$

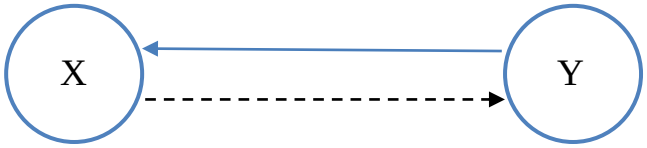
Solow, R. M. 1956. A Contribution to the Theory of Economic Growth. *Quarterly Journal of Economics*, *70*(1), 65-94.
Romer, P. M. 1990. Endogenous Technological Change. *Journal of Political Economy*, 98(5, Part 2), S71-S102.

Sources of Endogeneity

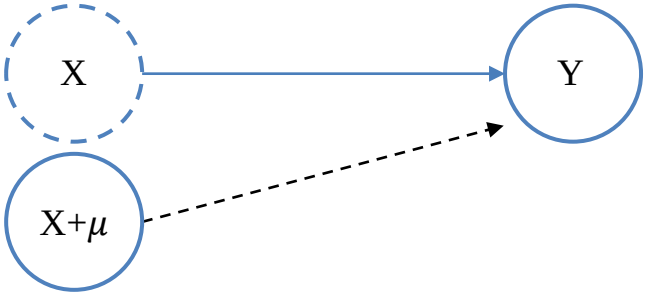
- Omitted variable bias (selection bias)



- Reverse causality bias (simultaneity bias)



- Measurement error bias (attenuation bias)



Sources of Endogeneity

- In practice, there always are sources of endogeneity in your study.

- It is important to infer how endogeneity would influence your estimations
  - Before presenting 2SLS results, it is suggested to report standard OLS results and to discuss why 2SLS is preferred to OLS estimations.
  - Example: Police and crime rate (Levitt 1997)

DV: change in crime rate

Variable	(1) OLS	(2) OLS	(3) 2SLS
ln Sworn officers per capita	0.28 (0.05)	−0.27 (0.06)	−1.39 (0.55)
State unemployment rate	−0.65 (0.40)	−0.25 (0.31)	−0.00 (0.36)
ln Public welfare spending per capita	−0.03 (0.02)	−0.03 (0.02)	−0.03 (0.02)
ln Education spending per capita	0.04 (0.07)	0.06 (0.06)	0.02 (0.07)
Percent ages 15–24 in SMSA	1.43 (1.00)	−2.61 (3.71)	−1.47 (4.12)
Percent black	0.010 (0.003)	−0.017 (0.011)	−0.034 (0.015)
Percent female-headed households	0.003 (0.006)	0.007 (0.023)	0.040 (0.030)
Data differenced?	No	Yes	Yes
Instruments:	None	None	Elections

"Higher crime rates are likely to increase the marginal productivity of police. Cities with high crime rates, therefore, may tend to have large police forces." (p. 270)

"The coefficient on sworn officers now becomes negative, suggesting that unobserved heterogeneity across cities imparts an upward bias on the coefficient." (p. 280)

Levitt, S. D. 1997. Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. *American Economic Review*, 87(3), pp.270-290.

Simple Mathematics of Endogeneity

- It is important to infer how endogeneity would influence your estimations.
  - ➢ Sometimes, mathematics is helpful for this purpose.

A review I received

"How would the reporting bias affect the empirical analysis? I think many of the rapes are not reported, and the reporting rate should be different for people living or visiting different precincts of NYC. I don't know when the authors collected their data, but many rapes are also reported much later than the actual time of rapes. Taken together, the measurement error could be correlated with many things in the econometric model. I am concerned that the inclusion of the precinct fixed effect alone cannot address this issue."

Our response

By nature, we cannot observe sexual assault cases unreported to the police…

We explicitly consider a reporting bias affected by an unobserved factor *K*:

$$\hat{y} = y + \gamma K = \beta_1 + \beta_2 X + \gamma K + \varepsilon$$

where $\hat{y}$ is reported incidents of sexual assaults and γ is assumed to be negative. Then, we can obtain the coefficient of interest from the model with reporting bias not considered, as follows,

$$\hat{y} = \beta_1' + \beta_2' X + \varepsilon$$

$$p\lim_{N\to\infty} \beta_2' = \beta_2 + \gamma \frac{Cov(X,K)}{Var(X)}$$

We believe that $Cov(X,K) < 0$ is more plausible in our research context

Simple Mathematics of Endogeneity

- Omitted variable bias (selection bias)

True model: $y = \beta_1 + \beta_2 X + \gamma Z + \varepsilon$

Estimate model: $y = \beta_1' + \beta_2' X + \varepsilon'$

$$\beta_2' \rightarrow \beta_2 + \frac{Cov(X, \varepsilon')}{Var(X)} = \beta_2 + \gamma \frac{Cov(X, Z)}{Var(X)}$$

➢ Example: Marketing event effectiveness

- One might concern that customer loyalty may lead to selection bias.

Customer loyalty

True model: $y = \beta_1 + \beta_2 X + \gamma Z + \varepsilon$

Estimate model: $y = \beta_1' + \beta_2' X + \varepsilon'$

Sales Event participation

positive

$$\beta_2' \rightarrow \beta_2 + \gamma \frac{Cov(X, Z)}{Var(X)}$$

Estimated coefficient (β_2') is upward-biased than true value (β_2).

Simple Mathematics of Endogeneity

- Reverse causality bias (simultaneity bias)

True model: $y = \beta_1 + \beta_2 X + \varepsilon$

$X = \alpha y + \delta$

Estimate model: $y = \beta_1' + \beta_2' X + \varepsilon$

$$\beta_2' \rightarrow \beta_2 + \frac{Cov(X, \varepsilon)}{Var(X)}$$

$$Cov(X, \varepsilon) = \frac{\alpha}{1 - \alpha\beta_2}$$

➢ Example: Police effectiveness in crime reduction

- High crime rates may lead to more police force in that area.

True model: $y = \beta_1 + \beta_2 X + \varepsilon$

$X = \alpha y + \delta$

Estimate model: $y = \beta_1' + \beta_2' X + \varepsilon$

Crime rate — Police force

$$\beta_2' \rightarrow \beta_2 + \frac{Cov(X, \varepsilon)}{Var(X)}$$

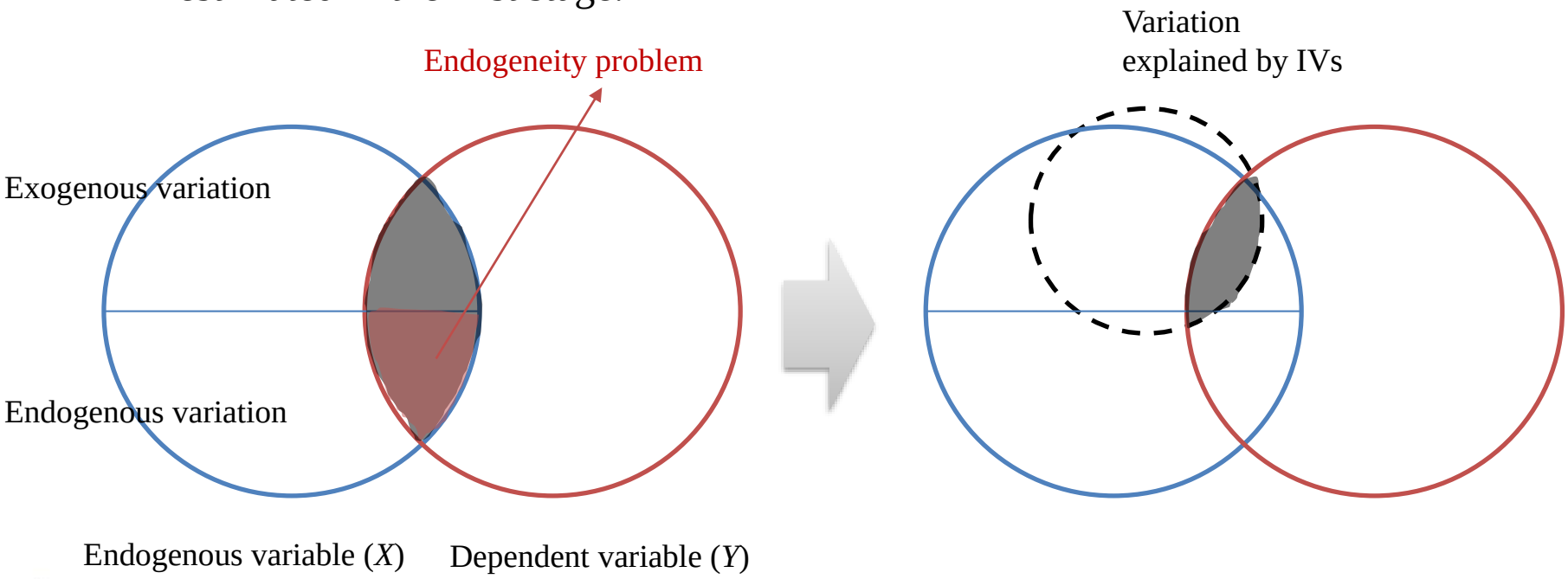
$$Cov(X, \varepsilon) = \frac{\alpha}{1 - \alpha\beta_2}$$ → positive

Although true value (β_2) is negative, estimated coefficient (β_2') can be positive (upward bias).

Instrument Variable Approach

Instrument Variable is a Natural Experiment

- Two-Stage Least Squares (2SLS)
  - The first-stage is to isolate the exogenous variation in an endogenous variable, which is explained by instrument variables (IVs).
  - The second-stage regresses the dependent variable on the predicted value estimated in the first stage.

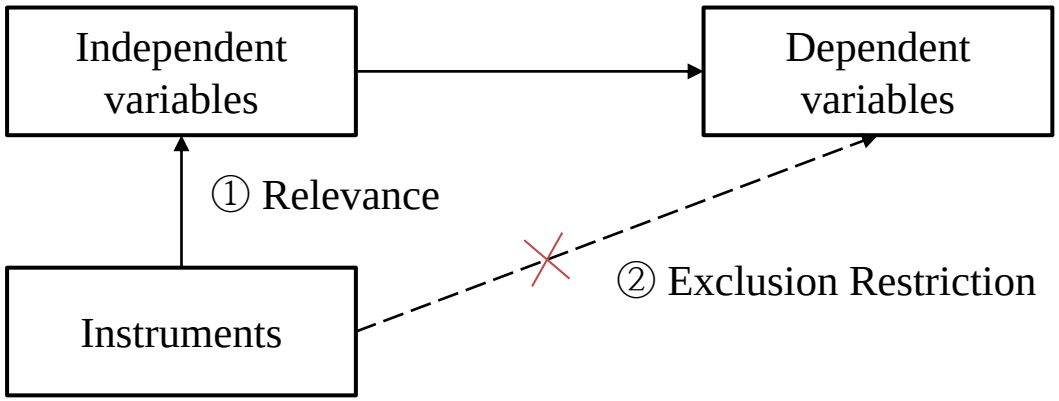


Two-Stage Least Squares (2SLS)

- Two-Stage Least Squares (2SLS)
  - (1) First-stage estimation
    - $X = a + bZ + Controls + \mu$
  - (2) Second-stage estimation
    - $Y = \alpha + \beta\hat{X} + Controls + \varepsilon$

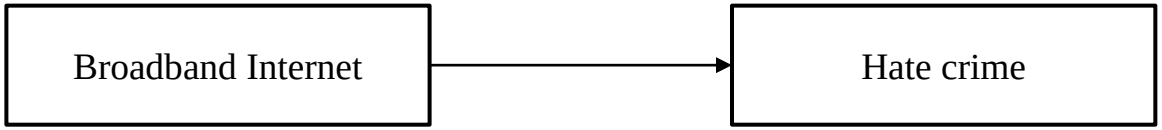
Predicted value from first-stage

- Conditions for successful instruments



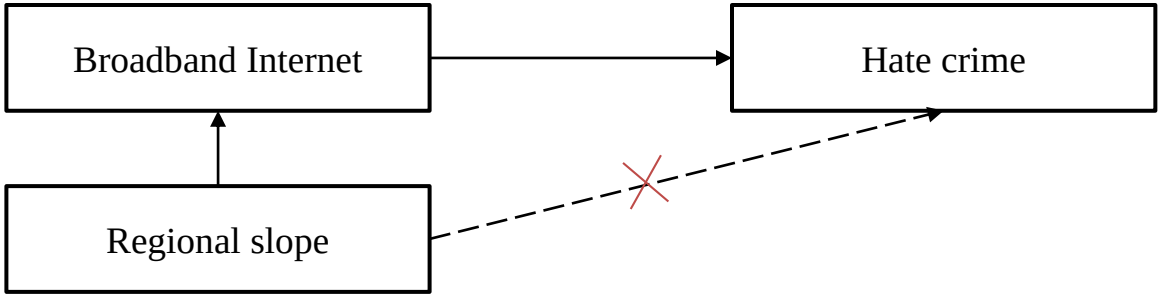
Examples of Instrument Variable Approach

- Example: Internet penetration and hate crime (Chan et al. 2016)
  - ➢ Sources of endogeneity



    - Omitted variables
    - Measurement errors

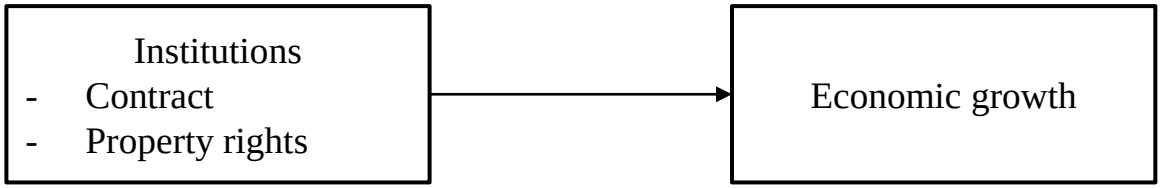
  - ➢ How to address endogeneity
    - Regional slope as instruments for Internet penetration



Chan, J., Ghose, A. and Seamans, R., 2016. The Internet and Racial Hate Crime: Offline Spillovers from Online Access. *MIS Quarterly*, 40(2), pp.381-403.

Examples of Instrument Variable Approach

- Example: Institutions and economic growth (Acemoglu and Johnson 2005)
  - ➢ Sources of endogeneity

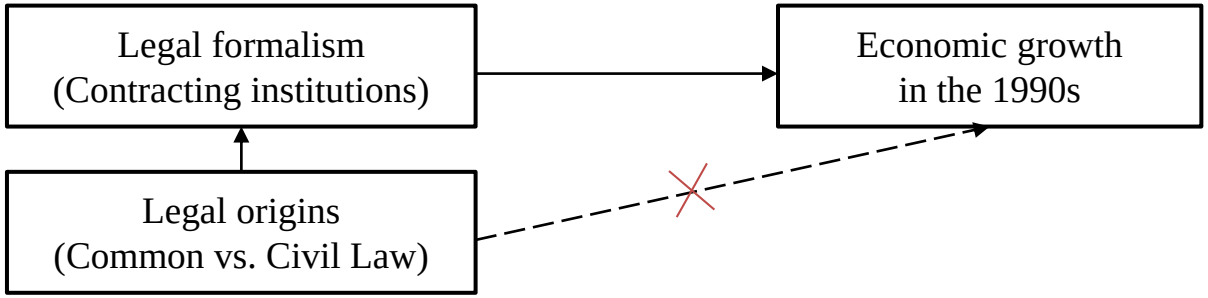


  - Omitted variables
  - Reverse causality
  - Measurement errors

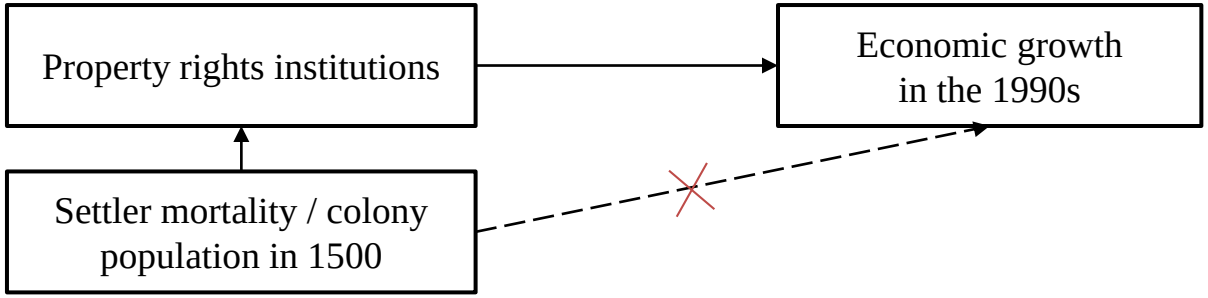
Acemoglu, D. and Johnson, S., 2005. Unbundling Institutions. *Journal of Political Economy*, 113(5), pp.949-995.

Examples of Instrument Variable Approach

- Example: Institutions and economic growth (Acemoglu and Johnson 2005)
  - ➢ How to address endogeneity
    - Legal origins as instruments for legal formalism (contracting institutions)



    - Mortality rates and colony population as instruments for property rights institutions

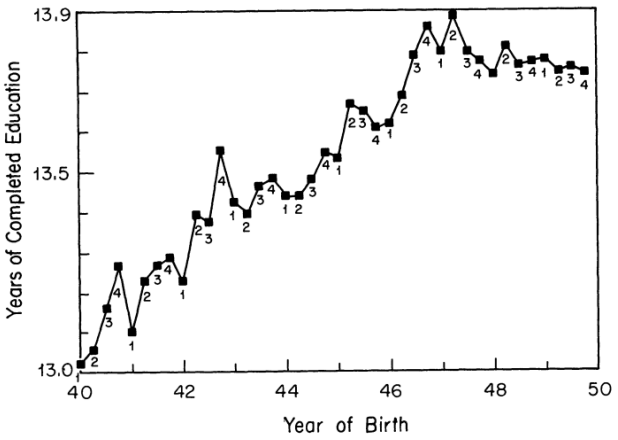


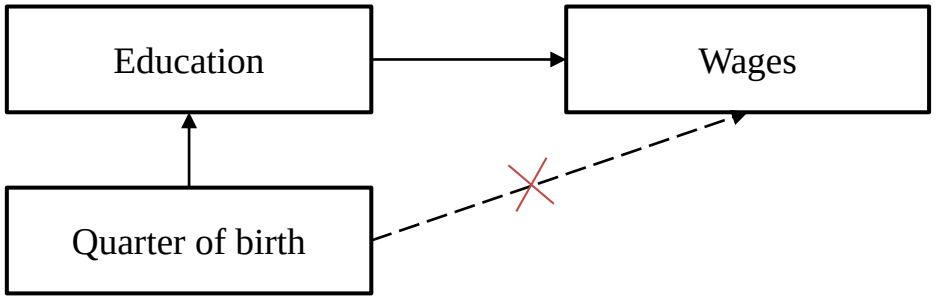
Acemoglu, D. and Johnson, S., 2005. Unbundling Institutions. *Journal of Political Economy*, 113(5), pp.949-995.

Examples of Instrument Variable Approach

- Example: Education and wages (Angrist and Krueger 1991)
  - School entrance in September in the 5th year in the United States
  - Compulsory school attendance until 16 (by law)

  - Quarter of birth (season) as instruments for education

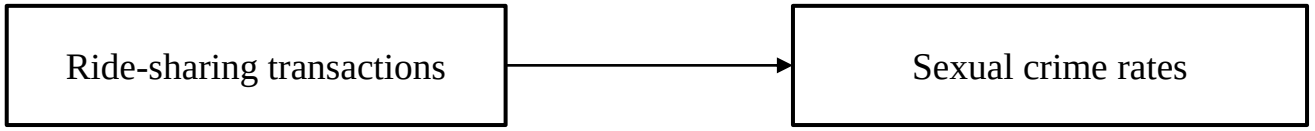




Angrist, J.D. and Keueger, A.B., 1991. Does Compulsory School Attendance Affect Schooling and Earnings?. *Quarterly Journal of Economics*, 106(4), pp.979-1014.

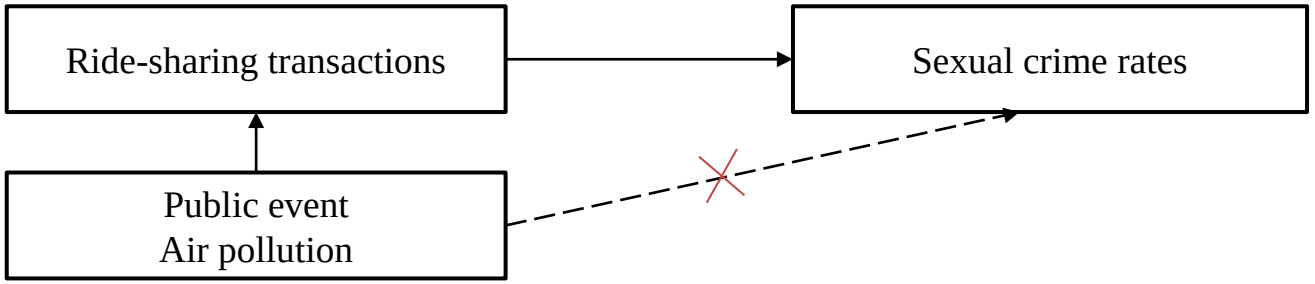
Examples of Instrument Variable Approach

- Example: Ride-sharing and sexual crime (Park et al. 2018)
  - ➢ Sources of endogeneity



    - Omitted variables
    - Reverse causality

  - ➢ How to address endogeneity
    - Public events and air pollution as instruments for ride-sharing (Uber) transactions



Park, J., Kim, J., Pang, M. and Lee, B., 2018. The Deterrent Effect of Ride-Sharing on Sexual Assault and Investigation of Situational Contingencies. *KAIST Working Paper.*

Selection Model

- Selection bias comes up when researchers can only observe the sample after subjects endogenously determine to select into a study (e.g., adopt or not).

- Heckman selection model
  - Additionally control for the probability to be selected
  - (1) First-stage estimation (probit model)
    - $\Pr(X > 0) = \Phi(\gamma Z + Controls + \mu)$
  - (2) Second-stage estimation (truncated model)
    - $Y = \alpha + \beta X + Controls + \varepsilon$

      $\rightarrow \mathrm{E}(Y|X > 0) = \alpha + \beta X + Controls + E(\varepsilon\,|X > 0)$

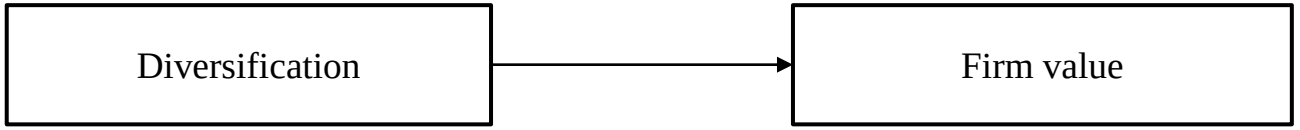
      $\rightarrow \mathrm{E}(Y|X > 0) = \alpha + \beta X + Controls + \rho\sigma_\varepsilon \boldsymbol{\lambda}$

      $\rightarrow Y = \alpha + \beta X + Controls + \rho\sigma_\varepsilon \boldsymbol{\lambda} + \delta$

- $\boldsymbol{\lambda}$ is inverse Mills ratio estimated from the first-stage.
- ρ is proportional to the correlation between unobserved determinants of dependent variable (ε) and unobserved determinants of propensity to X>0 from first stage (μ)

Selection Model

- Example: Diversification and firm value (Campa and Kedia 2002)
  - ➢ Source of endogeneity



  - Omitted variables
  - Reverse causality

  - ➢ Results

DV: firm value	OLS (BO)	OLS	Fixed Effects	IV	Self-selection
Constant	−0.36	−0.75	−0.09	−0.72	−0.68
	(33)	(26)	(1.67)	(30)	(30)
D (diversification)	−0.13	−0.11	−0.06	0.30	0.18
	(10)	(9.13)	(2.88)	(5.03)	(4.03)
Log of total assets	0.04	0.61	0.33	0.55	0.54
	(19)	(36)	(16)	(40)	(39)
EBIT/SALES	1.15	0.69	0.39	0.44	0.44
	(42)	(19)	(13)	(13)	(13)
CAPX/SALES	0.33	0.06	0.19	0.16	0.16
	(15)	(1.96)	(7.25)	(5.65)	(5.68)
Lambda					−0.14
					(6.07)

- Negatively correlated are unobserved determinants of firm value and unobserved determinants of propensity to diversification

Campa, J.M. and Kedia, S., 2002. Explaining the Diversification Discount. Journal of Finance, 57(4), pp.1731-1762.

Check Lists for Instruments

Test Statistics for Instrument Validity

- Endogeneity tests
  - e.g., Hausman test *(not common approach)*
  - **In practice**, describe what could cause the endogeneity concerns, and how the estimations would be biased without controlling for endogeneity.

- Weak identification test
  - e.g., Stock-Yogo test
  - Test for F statistics from the first-stage estimation

- Overidentifcation test
  - e.g., Sargan test / Hansen J test
  - Null hypothesis: Exclusion restriction condition is valid
  - It assumes some instruments are satisfied for exclusion restriction, and tests if the other instruments are correlated with error terms from the second-stage estimations.

Instrument Validity

- Justification for instrument validity is critical in identification
  - Poor instruments will produce poor results.
  - **Theoretical reasoning** is more important, than test statistics.
  - Since untestable assumptions are unavoidable in causal inference, all test statistics are incomplete.

  - In my experiences, reviewers tend not to buy the argument on instrument validity, ultimately the paper itself, unless they are convinced intuitively and theoretically, regardless of statistical tests.

Instrument Validity

- Evidence on theoretical justification for instrument validity

TABLE 2 Characteristics of Published IV Applications Over Time

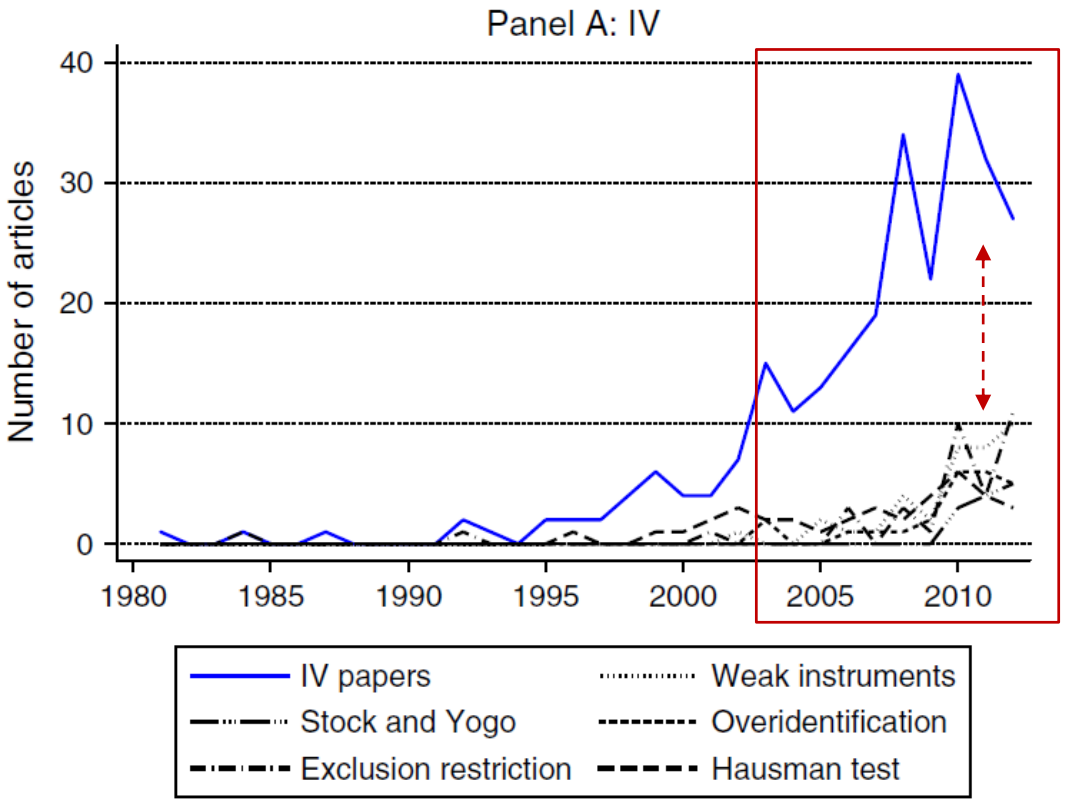
Type of Justification	1985–1990	1991–1996	1997–2002	2003–2008
Experiment	0%	0%	4%	6%
Natural Experiment	0	0	0	3
Theory	9	7	14	31
Lag	9	7	14	11
Reference	5	0	3	5
Empirics	9	0	17	5
None	68	86	48	39
Total Percent	100%	100%	100%	100%
Number of Articles	22	15	29	36
% Just-identified	8	13	24	22
% Report First Stage	23	7	31	33

Table 2 summarizes more than 100 articles published in the *American Political Science Review*, the *American Journal of Political Science*, and *World Politics* over a 24-year span, categorizing them according to the way the IVs are identified. Percentages within each date group add (with rounding error) to 100%.
Explanation of Categories:
Experiment: IVs that were generated through a random assignment process.
Natural Experiment: IVs that were generated through a quasi-random assignment process.
Theory: Articles in which the authors provided a theoretical explanation for the validity of their exclusion restrictions.
Lag: IVs that were generated by lagging the dependent or independent variable.
Empirics: IVs that were selected based on the results of an empirical test (such as regressing Y on X and Z to show no correlation or regressing X on Z to determine the strongest instruments).
Reference: Articles in which the author explains the validity of his or her exclusion restrictions by citing another author's work.
None: No justification provided.

Sovey, A.J. and Green, D.P., 2011. Instrumental Variables Estimation in Political Science: A Readers' Guide. *American Journal of Political Science*, 55(1), pp.188-200.

Instrument Validity

- Evidence on theoretical justification for instrument validity



(Bowen et al. 2016)

Bowen III, D.E., Frésard, L. and Taillard, J.P., 2016. What's Your Identification Strategy? Innovation in Corporate Finance Research. *Management Science*, 63(8), pp.2529-2548.

Simple Mathematics of 2SLS

- Sometimes, mathematics is very helpful in ruling out the possibility of invalid instrument variables (violation of exclusion restrictions).

- 2SLS formula

First stage

$$x = a' + b'z + \tau$$

Second stage

$$y = a + b\hat{x} + \varepsilon$$

x and y are independent and dependent variables, respectively

z is instrument variables

$$\widehat{b_{2SLS}} \rightarrow b + \frac{\sigma_\varepsilon}{\sigma_x}\frac{Corr(z,\varepsilon)}{Corr(z,x)}$$

When $Corr(z,x) > 0$,

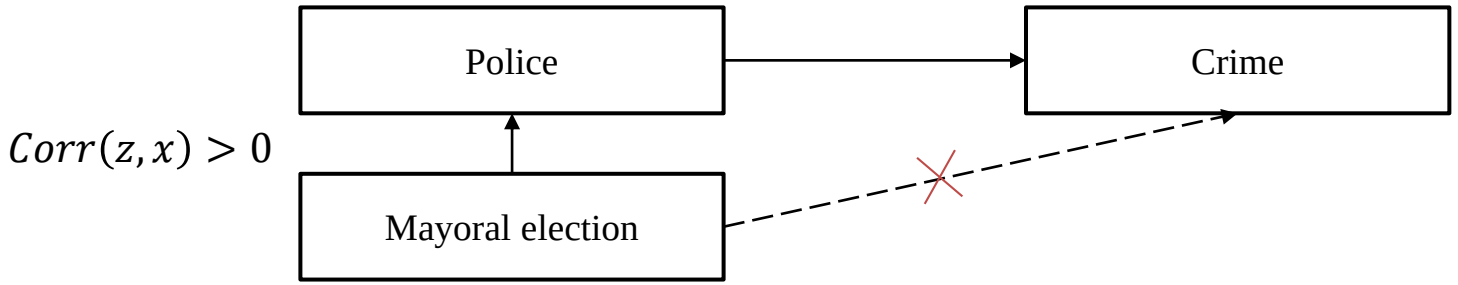
If $Corr(z,\varepsilon) > 0$, $\widehat{b_{2sls}} > b$ (upward bias)

If $Corr(z,\varepsilon) < 0$, $\widehat{b_{2sls}} < b$ (downward bias)

Violation of exclusion restrictions is amplified with weak instruments

Simple Mathematics of 2SLS

- Example: Police and crime (Levitt 1997)
  - How to address endogeneity
    - Mayoral election years as instruments for police hiring



  - Possible violation of exclusion restriction
    - One might concern that mayoral election may increase crime rates, related to political issues and social instability. (i.e., $Corr(z, \varepsilon) > 0$)

$$\widehat{b_{2sls}} \rightarrow b + \frac{\sigma_\varepsilon}{\sigma_x} \frac{Corr(z, \varepsilon)}{Corr(z, x)} \quad \Rightarrow \quad \widehat{b_{2sls}} > b$$

The negative coefficient of $\widehat{b_{2sls}}$ seems even conservative.

Levitt, S. D. 1997. Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. *American Economic Review*, 87(3), pp.270-290.

Checklist of Issues to Address Regarding IV Results

Category	Issues to Address
Model	• What is the estimand? • Are the causal effects assumed to be homogenous or heterogeneous?
Independence	• Explain why it is plausible to believe that the instrumental variable is unrelated to unmeasured causes of the dependent variable.
Exclusion Restriction	• Explain why it is plausible to believe the instrumental variable has no direct effect on the outcome.
Instrument Strength	• How strongly does the instrument predict the endogenous independent variable after controlling for covariates?
Monotonicity	• Explain why it is plausible to believe there are no Defiers, that is, people who take the treatment if and only if they are assigned to the control group.
Stable Unit Treatment Value Assumption (SUTVA)	• Explain why it is plausible to assume that a given observation is unaffected by treatments assigned or received by other units.

Sovey, A.J. and Green, D.P., 2011. Instrumental Variables Estimation in Political Science: A Readers' Guide. *American Journal of Political Science*, 55(1), pp.188-200.

Exploiting Internal Instruments
: Dynamic Panel Model

Internal Instruments

- In most cases, it is quite challenging to find relevant instrument variables.

- Internal variables can be a useful "Plan B" for addressing endogeneity.
  - "This method does not allow us to control for full endogeneity but for a weak type of it." (Levine et al. 2000, p. 50)

- Dynamic panel model
  - Controlling for lagged dependent variable

    : "The key insight is that lagged outcomes are a function of both observable covariates and unobservables." (Keele 2015, p. 322)
  - Using lagged independent variable as internal instruments

Levine, R., Loayza, N. and Beck, T., 2000. Financial Intermediation and Growth: Causality and Causes. *Journal of Monetary Economics*, 46(1), pp.31-77.
Keele, L., 2015. The Statistics of Causal Inference: A View from Political Methodology. *Political Analysis*, 23(3), pp.313-335.

Dynamic Panel Model

- However, including both fixed effects and lagged dependent variables will lead to endogeneity problem.

$$y_{i,t} = \rho y_{i,t-1} + \underset{(1\times K)}{\beta'} \underset{(K\times 1)}{x_{i,t}} + \alpha_i + \varepsilon_{i,t}$$

- Difference GMM
  - To remove fixed effects, take a first-difference in equation

Differenced Equation : $\Delta y_{i,t} = \rho \Delta y_{i,t-1} + \beta' \Delta x_{i,t} + \Delta \varepsilon_{i,t}$

Using lagged level of variables as internal instruments

- System GMM
  - If the variables persist over time, lagged level will be weak instruments for first-differences -> Add another level equation

Differenced Equation : $\Delta y_{i,t} = \rho \Delta y_{i,t-1} + \beta' \Delta x_{i,t} + \Delta \varepsilon_{i,t}$

Level Equation : $y_{i,t} = \rho y_{i,t-1} + \beta' x_{i,t} + \alpha_i + \varepsilon_{i,t}$

Lagged difference of variables as internal instruments

Check Lists for Dynamic Panel Model

- Serial correlation
  - Arellano-Bond test for AR(1) in first differences should be significant
  - Arellano-Bond test for AR(2) in first differences should not be significant

```
Arellano-Bond test for AR(1) in first differences: z =  -3.15  Pr > z =  0.002
Arellano-Bond test for AR(2) in first differences: z =   1.21  Pr > z =  0.225
```

- Over-identification tests
  - Hansen test of over-identification restrictions should not be rejected ($p > 0.05$)

```
Sargan test of overid. restrictions: chi2(78)   = 104.46  Prob > chi2 =  0.024
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(78)   =  45.77  Prob > chi2 =  0.999
  (Robust, but weakened by many instruments.)
```

Check Lists for Dynamic Panel Model

- Problem of too many instruments
  - Too many instruments in difference and system GMM may lead to implausibly high p-value of over-identification J tests. That is, the Hansen test cannot detect the problem of over-identification in these cases (Roodman 2009).
  - More seriously, different numbers of instruments, along with the inability of Hansen test, could arrive at different conclusions, possibly due to over-identification.

> "Researchers should report the number of instruments generated for their regressions. In system GMM, difference-in-Hansen tests for the full set of instruments for the levels equation, as well as the subset based on the dependent variable, should be reported. **Results should be aggressively tested for sensitivity to reductions in the number of instruments.**" (Roodman 2009, p. 156)

Roodman, D., 2009. A Note on the Theme of Too Many Instruments. *Oxford Bulletin of Economics and Statistics*, 71(1), pp.135-158.

Check Lists for Dynamic Panel Model

- Example: Cloud computing and energy efficiency (Park et al. 2018)

Table 5: Sensitivity Analysis to Instruments in System GMM Estima

Dependent variable: Energy efficiency	Standard instruments			Collapsed instruments		
Length of lags as instruments:	One	two	three	One	two	three
	(1)	(2)	(3)	(4)	(5)	(6)
Lagged efficiency	0.755***	0.751***	0.747***	0.944***	0.930***	0.847***
	(0.041)	(0.034)	(0.033)	(0.068)	(0.057)	(0.055)
IT intensity	0.005	0.005	0.000	0.024	0.015	0.013
	(0.008)	(0.004)	(0.004)	(0.032)	(0.027)	(0.029)
Non-IT intensity	0.002	0.001	0.002	0.002	0.005	0.007
	(0.005)	(0.003)	(0.003)	(0.020)	(0.018)	(0.019)
Other intermediate intensity	-0.010	-0.008*	-0.008*	-0.018	-0.020	-0.022
	(0.007)	(0.005)	(0.004)	(0.023)	(0.022)	(0.022)
Data processing and hosting services	0.013***	0.010***	0.008***	0.035**	0.034**	0.035**
	(0.004)	(0.003)	(0.002)	(0.014)	(0.014)	(0.014)
IT systems design services	-0.009**	-0.008**	-0.005**	-0.027*	-0.024	-0.023
	(0.005)	(0.003)	(0.002)	(0.016)	(0.015)	(0.015)
Number of instruments	116	161	203	26	29	32
Serial correlation test	0.458	0.459	0.454	0.449	0.439	0.445
Instrument validity test	1.000	1.000	1.000	0.810	0.444	0.184
Observations	952	952	952	952	952	952

"Roodman (2009b) argues that instrument proliferation, possibly resulting from a long panel, may weaken the Hansen test's ability to detect the problem of over-identification; note that this does not indicate that the condition of over-identification restrictions is violated.

Following the recommendation of Roodman (2009a, 2009b), we check the sensitivity of our findings to the number of instruments." (p. 24)

Park, J., Han, K. and Lee, B., 2018. An Empirical Analysis of Cloud Computing and Energy Efficiency: A Stochastic Frontier Approach. *KAIST Working Paper.*

Concluding Remarks

Think Deeply about the Identification Assumption

- Reasoning about the plausibility of instrument variables (identification strategy) is most critical part of empirical research for causal inference.

"As such, reasoning about the plausibility of an identification strategy in a specific empirical context is a critical part of any statistical analysis that purports to be causal. Since untestable assumptions are unavoidable in causal inference, **it is only through the careful understanding of those assumptions that one can make a case for their plausibility in a given context**. As such, the researcher must think deeply about the assumptions and part of the analysis should be a well-reasoned defense of the identification strategy… Reasoning about assumptions is often not part of a statistical analysis, but **it must be when the goal is to identify causal effects**." (Keele 2015, p. 323-324)

Keele, L., 2015. The Statistics of Causal Inference: A View from Political Methodology. *Political Analysis*, 23(3), pp.313-335.

End of Document