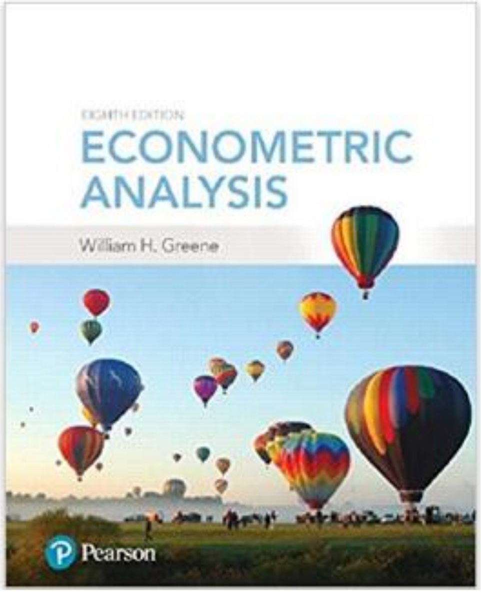
Econometrics I

Professor William Greene

Stern School of Business

Department of Economics



Econometrics I

Part 13 – Endogeneity: Applications

Measurement Error

$y = \beta x^* + \varepsilon$ all of the usual assumptions

$x = x^* + u$ the true x^* is not observed (education vs. years of school)

What happens when y is regressed on x? **Least squares attenutation**:

$$\text{plim } b = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{\text{cov}(x^*+u, \beta x^*+\varepsilon)}{\text{var}(x^*+u)}$$

$$= \frac{\beta \, \text{var}(x^*)}{\text{var}(x^*) + \text{var}(u)} < \beta$$

Why Is Least Squares Attenuated?

$y = \beta x^* + \varepsilon$

$x = x^* + u$

$y = \beta x + (\varepsilon - \beta u)$

$y = \beta x + v$, $\text{cov}(x,v) = -\beta\,\text{var}(u)$

Some of the variation in x is not associated with variation in y. The effect of variation in x on y is dampened by the measurement error.

Measurement Error in Multiple Regression

Multiple regression: $y = \beta_1 x_1^* + \beta_2 x_2^* + \varepsilon$

x_1^* is measured with error; $x_1 = x_1^* + u$

x_2 is measured with out error.

The regression is estimated by least squares

Popular myth #1. b_1 is biased downward, b_2 consistent.

Popular myth #2. All coefficients are biased toward zero.

Result for the simplest case. Let

$\sigma_{ij} = \text{cov}(x_i^*, x_j^*), i, j = 1, 2$ (2x2 covariance matrix)

$\sigma^{ij} =$ ijth element of the inverse of the covariance matrix

$\theta^2 = \text{var}(u)$

For the least squares estimators:

$$\text{plim } b_1 = \beta_1 \left(\frac{1}{1 + \theta^2 \sigma^{11}} \right), \quad \text{plim } b_2 = \beta_2 - \beta_1 \left(\frac{\theta^2 \sigma^{12}}{1 + \theta^2 \sigma^{11}} \right)$$

The effect is called "smearing."

Twins

Application from the literature: Ashenfelter/Krueger: A wage equation for twins that includes “schooling.”

y = earnings

x = education

z = education as reported by sibling

NBER WORKING PAPER SERIES

ESTIMATES OF THE ECONOMIC RETURN TO SCHOOLING FROM A NEW SAMPLE OF TWINS

Orley Ashenfelter
Alan Krueger

Working Paper No. 4143

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1992

Table 3: OLS, GLS, IV, and Fixed Effects Estimates of Log Wage Equations for Identical Twins[a]

Variable	OLS (1)	GLS (2)	GLS (3)	IV[b] (4)	First Difference (5)	First Diff. by IV (6)
Own Education (+100)	8.387 (1.443)	8.744 (1.495)	8.844 (1.515)	11.624 (2.950)	9.157 (2.371)	16.697 (4.311)
Sibling's Education (+100)	--	--	-.665 (1.518)	-3.735 (2.946)	--	--
Age	.088 (.019)	.090 (.023)	.090 (.023)	.088 (.019)	--	--
Age-Squared (+100)	-.087 (.023)	-.089 (.028)	-.090 (.029)	-.087 (.024)	--	--
Male	.204 (.063)	.204 (.077)	.206 (.077)	.206 (.064)	--	--
White	-.410 (.127)	-.417 (.143)	-.424 (.144)	-.428 (.128)	--	--
Sample Size	298	298	298	298	149	149
R^2	.260	.219	.219	--	.092	--

Orthodoxy

- A proxy is not an instrumental variable
- Instrument is a noun, not a verb
- Are you sure that the instrument is really exogenous? The "**natural experiment**."

Some Conventional Approaches

A study of moral hazard
Riphahn, Wambach, Million: "Incentive Effects in the Demand for Healthcare"
Journal of Applied Econometrics, 2003

Did the presence of the ADDON insurance influence the demand for health care – doctor visits and hospital visits?

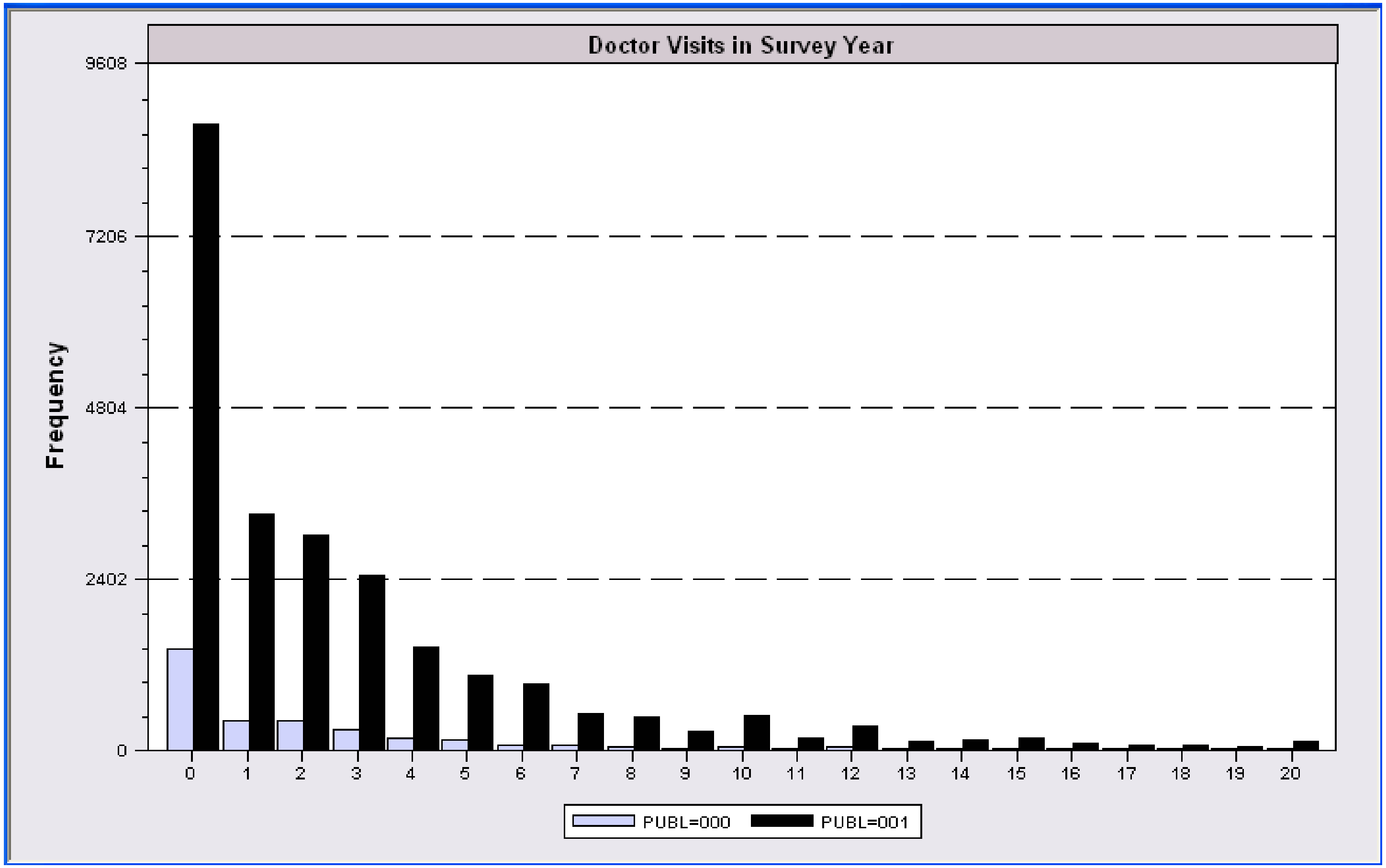
For a simple example, we examine the PUBLIC insurance (89%) instead of ADDON insurance (2%).

Application: Health Care Panel Data

German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods
Variables in the file are
Data downloaded from Journal of Applied Econometrics Archive. This is an unbalanced panel with 7,293 individuals. They can be used for regression, count models, binary choice, ordered choice, and bivariate binary choice. **This is a large data set. There are altogether 27,326 observations. The number of observations ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987).** Note, the variable NUMOBS below tells how many observations there are for each person. This variable is repeated in each row of the data for the person. (Downloaded from the JAE Archive)

DOCTOR	=	**1(Number of doctor visits > 0)**
HOSPITAL	=	**1(Number of hospital visits > 0)**
HSAT	=	**health satisfaction, coded 0 (low) - 10 (high)**
DOCVIS	=	**number of doctor visits in last three months**
HOSPVIS	=	**number of hospital visits in last calendar year**
PUBLIC	=	**insured in public health insurance = 1; otherwise = 0**
ADDON	=	**insured by add-on insurance = 1; otherswise = 0**
HHNINC	=	**household nominal monthly net income in German marks / 10000**. (4 observations with income=0 were dropped)
HHKIDS	=	**children under age 16 in the household = 1; otherwise = 0**
EDUC	=	**years of schooling**
AGE	=	**age in years**
MARRIED	=	**marital status**
EDUC	=	**years of education**

Evidence of Moral Hazard?



Regression Study

```
----------------------------------------------------------------------
Ordinary    least squares regression ............
LHS=DOCVIS  Mean                     =       3.18352
            Standard deviation       =       5.68969
            Number of observs.       =         27326
Model size  Parameters               =             6
            Degrees of freedom       =         27320
Residuals   Sum of squares           =  853326.41135
            Standard error of e      =       5.58878
Fit         R-squared                =        .03533
            Adjusted R-squared       =        .03516
Model test  F[  5, 27320] (prob) =   200.1(.0000)
--------+-------------------------------------------------------------
        |                Standard            Prob.        Mean
  DOCVIS| Coefficient      Error        z    z>|Z|        of X
--------+-------------------------------------------------------------
Constant|     .43660      .29014     1.50  .1324
     AGE|     .06754***   .00304    22.25  .0000       43.5257
  HHNINC|  -1.54898***    .19956    -7.76  .0000        .35208
  FEMALE|     .94128***   .06895    13.65  .0000        .47877
    EDUC|   -.05549***    .01624    -3.42  .0006       11.3206
  PUBLIC|     .59843***   .11370     5.26  .0000        .88571
--------+-------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------
```

Endogenous Dummy Variable

- Doctor Visits = f(Age, Educ, Health,
  Presence of Insurance,
  Other unobservables)
- Insurance = f(Expected Doctor Visits,
  Other unobservables)

Approaches

- (Semiparametric) Instrumental Variable: Create an instrumental variable for the dummy variable (Barnow/Cain/ Goldberger, Angrist, Current generation of researchers)
- (Parametric) Control Function: Build a structural model for the two variables (Heckman)
- (?) Propensity Score Matching (Heckman et al., Becker/Ichino, Many recent researchers)

Instrumental Variable Approach

Construct a prediction for T using only the exogenous information

Use 2SLS using this instrumental variable.

```
--------------------------------------------------------------------
Two stage    least squares regression ............
LHS=DOCVIS   Mean                   =        3.18352
ONE         AGE        HHNINC     FEMALE     EDUC       TFIT
--------+-----------------------------------------------------------
        |                  Standard            Prob.          Mean
  DOCVIS| Coefficient        Error        z    z>|Z|          of X
--------+-----------------------------------------------------------
Constant|   -33.1176***     2.56970   -12.89  .0000
     AGE|     .07535***      .00487    15.47  .0000        43.5257
  HHNINC|    3.17825***      .47734     6.66  .0000         .35208
  FEMALE|     .62839***      .11232     5.59  .0000         .47877
    EDUC|     .92150***      .07802    11.81  .0000        11.3206
  PUBLIC|    23.9012***     1.76483    13.54  .0000         .88571
--------+-----------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
--------------------------------------------------------------------
```

Magnitude = 23.9012 is nonsensical in this context.

Heckman's Control Function Approach

- $Y = x\beta + \delta T + E[\varepsilon|T] + \{\varepsilon - E[\varepsilon|T]\}$
- $\lambda = E[\varepsilon|T]$, computed from a model for whether T = 0 or 1

```
------------------------------------------------------------------
Sample Selection Model.............................
Two step     least squares regression ............
LHS=DOCVIS   Mean                   =        3.18352
Correlation of disturbance in regression
and Selection Criterion (Rho)...........     -.88169
--------+---------------------------------------------------------
        |                  Standard            Prob.        Mean
  DOCVIS| Coefficient       Error        z    z>|Z|        of X
--------+---------------------------------------------------------
Constant|   -14.8749***     1.01175   -14.70  .0000
     AGE|     .07062***      .00348    20.28  .0000     43.5257
  HHNINC|     .58241**       .26463     2.20  .0277      .35208
  FEMALE|    1.00046***      .06885    14.53  .0000      .47877
    EDUC|     .39321***      .03360    11.70  .0000     11.3206
  PUBLIC|    11.1200***      .66997    16.60  .0000      .88571
  LAMBDA|   -5.64728***      .35142   -16.07  .0000    .497D-09
--------+---------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
------------------------------------------------------------------
```

Magnitude = 11.1200 is nonsensical in this context.

Propensity Score Matching

- Create a model for T that produces probabilities for T=1: “Propensity Scores”
- Find people with the same propensity score – some with T=1, some with T=0
- Compare number of doctor visits of those with T=1 to those with T=0.

```
+----------------------------------------------------------------------+
| Estimated Average Treatment Effect (PUBLIC  )  Outcome is DOCVIS     |
| Nearest Neighbor  Using average of  1 closest neighbors              |
| Note, controls may be reused in defining matches.                    |
| Number of bootstrap replications used to obtain variance    =     25 |
+----------------------------------------------------------------------+
  Estimated average treatment effect =       .258108
  Begin bootstrap iterations *****************************************
  End bootstrap iterations   *****************************************
+----------------------------------------------------------------------+
| Number of Treated observations =  24203  Number of controls =   2568 |
| Estimated Average Treatment Effect    =        .258108               |
| Estimated Asymptotic Standard Error   =        .163314               |
| t statistic (ATT/Est.S.E.)            =       1.580447               |
| Confidence Interval for ATT = (     -.061986  to       .578203) 95% |
| Average Bootstrap estimate of ATT     =        .315962               |
| ATT - Average bootstrap estimate      =       -.057853               |
+----------------------------------------------------------------------+
```

Application of a Two Period Model

- “Hemoglobin and Quality of Life in Cancer Patients with Anemia,”
- Finkelstein (MIT), Berndt (MIT), Greene (NYU), Cremieux (Univ. of Quebec)
- 1998
- With Ortho Biotech – seeking to change labeling of already approved drug ‘erythropoetin.’ r-HuEPO

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

The New York Times

Business Day

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search | Global | DealBook | Markets | Economy | Energy | Media | Te

Drug Makers, in Shift, Join Fight Against Doping

LM Otero/Associated Press

Drug makers long paid little attention to how their products could be abused by athletes like Lance Armstrong, the director of the World Anti-Doping Agency said.

By KATIE THOMAS
Published: February 18, 2013

The blood-enhancing drug EPO has improved the lives of millions of anemia patients, but Lance Armstrong and other top cyclists have turned the medicine into a byword for doping.

FACEBOOK
TWITTER
GOOGLE+

QOL Study

- Quality of life study
  - i = 1,… 1200+ clinically anemic cancer patients undergoing chemotherapy, treated with transfusions and/or r-HuEPO
  - t = 0 at baseline, 1 at exit. (interperiod survey by some patients was not used)
- y_{it} = self administered quality of life survey, scale = 0,…,100
- $\mathbf{x}_{it}$ = hemoglobin level, other covariates
  - Treatment effects model (hemoglobin level)
  - Possibly **Endogenous treatment** – r-HuEPO treatment to affect Hg level: Actually not; treatment was not optional and all participated.
- Important statistical issues
  - Unobservable individual effects
  - The placebo effect
  - Attrition – sample selection
  - FDA mistrust of “community based” – not clinical trial based statistical evidence
- Objective – when to administer treatment for maximum marginal benefit

Regression-Treatment Effects Model

$$\begin{aligned}QOL_{it} &= \alpha_t + \text{"other covariates"} \\ &\quad + \beta_7 Hb_{it}^7 + \beta_8 Hb_{it}^8 + \beta_9 Hb_{it}^9 + \ldots \beta_{15} Hb_{it}^{15} \\ &\quad + c_i + \varepsilon_{it}\end{aligned}$$

$Hb_{it} =$ hemoglobin level, grams/deciliter, range 3+ to 15

$Hb_{it}^7 = 1(3 \le Hb_{it} < 7.5)$ (Base case; $\beta_7 = 0$)

$Hb_{it}^8 = 1(7.5 \le Hb_{it} < 8.5)$

$\vdots$

$Hb_{it}^{15} = 1(14.5 \le Hb_{it} \le 15)$

Effects and Covariates

- Individual effects that would impact a self reported QOL: Depression, comorbidity factors (smoking), recent financial setback, recent loss of spouse, etc.
- Covariates
  - Change in tumor status
  - Measured progressivity of disease
  - Change in number of transfusions
  - Presence of pain and nausea
  - Change in number of chemotherapy cycles
  - Change in radiotherapy types
  - Elapsed days since chemotherapy treatment
  - Amount of time between baseline and exit

First Differences Model
Change in r-HuEPO definitely changes Hb
Does change in Hb change QOL?

$$\Delta QOL_i = QOL_{i1} - QOL_{i0}$$
$$= (\alpha_1 - \alpha_0) + \Sigma_{j=8}^{15}\beta_j(Hb_{i1}^j - Hb_{i0}^j) + \Sigma_{k=1}^{K}\delta_k(x_{ik,1} - x_{ik,0}) + \varepsilon_{i1} - \varepsilon_{i0}$$

Regression to the mean (the "tendency to mediocrity")

$\varepsilon_{i0} - \varepsilon_{i1} = u_i - \rho(QOL_{i0} - \overline{QOL_0})$ Expect $0 \leq \rho < 1$

implies

$$\alpha = \alpha_1 - \alpha_0 + \rho\overline{QOL_0}$$

$$\Delta QOL_i = QOL_{i1} - QOL_{i0}$$
$$= \alpha + \Sigma_{j=8}^{15}\beta_j(Hb_{i1}^j - Hb_{i0}^j) + \Sigma_{k=1}^{K}\delta_k(x_{ik,1} - x_{ik,0}) - \rho QOL_{i0} + u_i$$

Dealing with Attrition

- The attrition issue: Appearance for the second interview was low for people with initial low QOL (death or depression) or with initial high QOL (don't need the treatment). Thus, missing data at exit were clearly related to values of the dependent variable.
- Solutions to the attrition problem
  - Heckman selection model (used in the study)
    - Prob[Present at exit|covariates] = $\Phi(\mathbf{z}'\boldsymbol{\theta})$ (Probit model)
    - Additional variable added to difference model $\lambda_i = \Phi(\mathbf{z}_i'\boldsymbol{\theta})/\Phi(\mathbf{z}_i'\boldsymbol{\theta})$
  - The FDA solution: fill with zeros. (!)

Evaluation of an OFT intervention

Independent fee-paying schools

UK Office of Fair Trading, May 2012; Stephen Davies

In this context, the OFT's evaluation team has evaluated the impact of the intervention addressing the anti-competitive practice of 50 independent fee-paying schools in the setting of fees during academic years 2001/02 to 2003/04. This research has been carried out by OFT economists and independently reviewed by Professor Stephen Davies.[1]

The main aim is to understand whether the OFT intervention had an impact, and to estimate this impact in terms of reduced school fees. To do so we have collected data on the evolution of school fees and other variables before and after the OFT's intervention.

http://dera.ioe.ac.uk/14610/1/oft1416.pdf

For the academic years 2001/02 – after the Competition Act came into force – to 2003/04, the OFT held that the exchange of future pricing information between the Sevenoaks Survey schools 'had as its object the distortion of competition within the United Kingdom'.[2] It was not necessary therefore for the OFT to come to a conclusion as to whether the information exchange had an anti-competitive effect.

The schools concerned had exchanged information relating to their intended fee increases and fee levels for boarding and day pupils in relation to the academic years 2001/02, 2002/03 and 2003/04. The information was exchanged through a survey, known as the 'Sevenoaks Survey'. Between February and June of each year, the schools concerned gave details of their intended fee increases and fee levels for the academic year beginning in September. Sevenoaks then collated that information and circulated it, in the form of tables, to the schools concerned. The information in the tables was updated and circulated between four and six times each year as schools developed their fee increase proposals in the course of their annual budgetary processes.

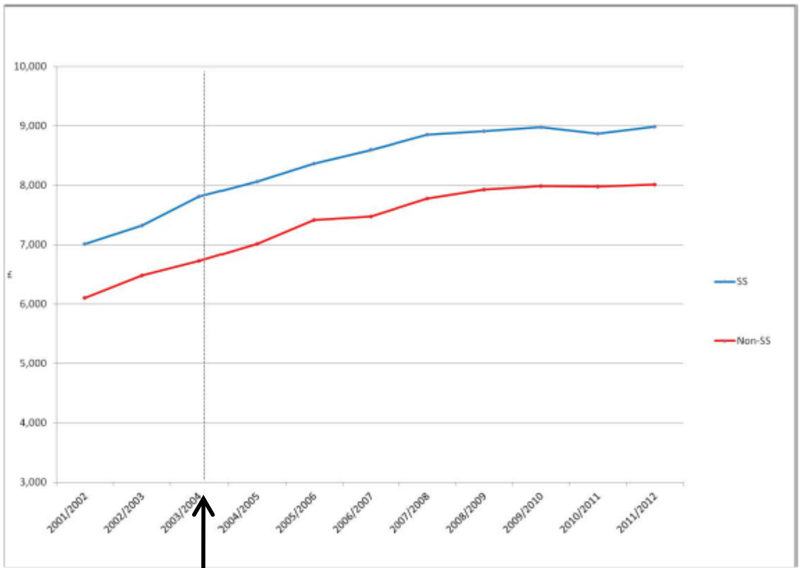
The key features of the infringement that were instrumental in the OFT's assessment of the information exchange as an object offence included:

- The information that was exchanged related to future intentions of price, and was confidential and not publicly available.
- It was done on a regular and highly systematic basis, and for a number of years.
- The timing of the exchange corresponded with the timing in which school fees for the following year were set.

Outcome is the fees charged.

Activity is collusion on fees.

Figure 2: Average fees per term (boarding, deflated)

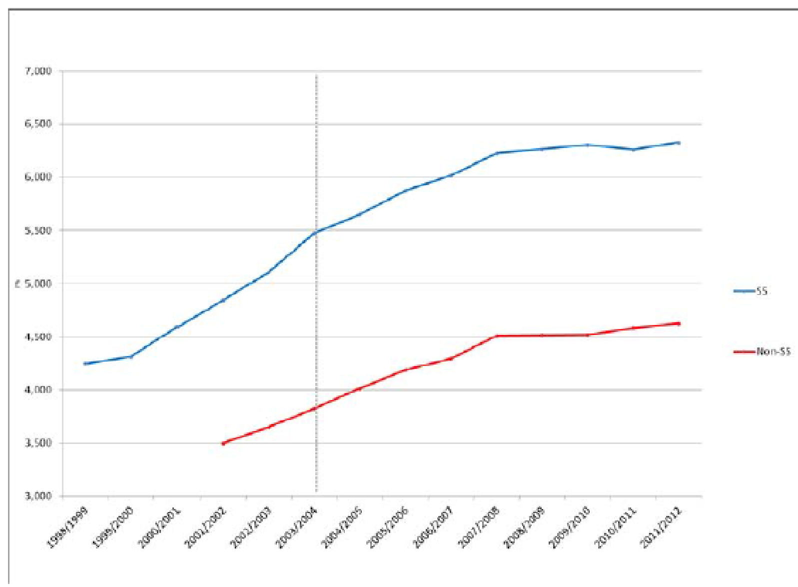


Treatment Schools: Treatment is an intervention by the Office of Fair Trading

Control Schools were not involved in the conspiracy

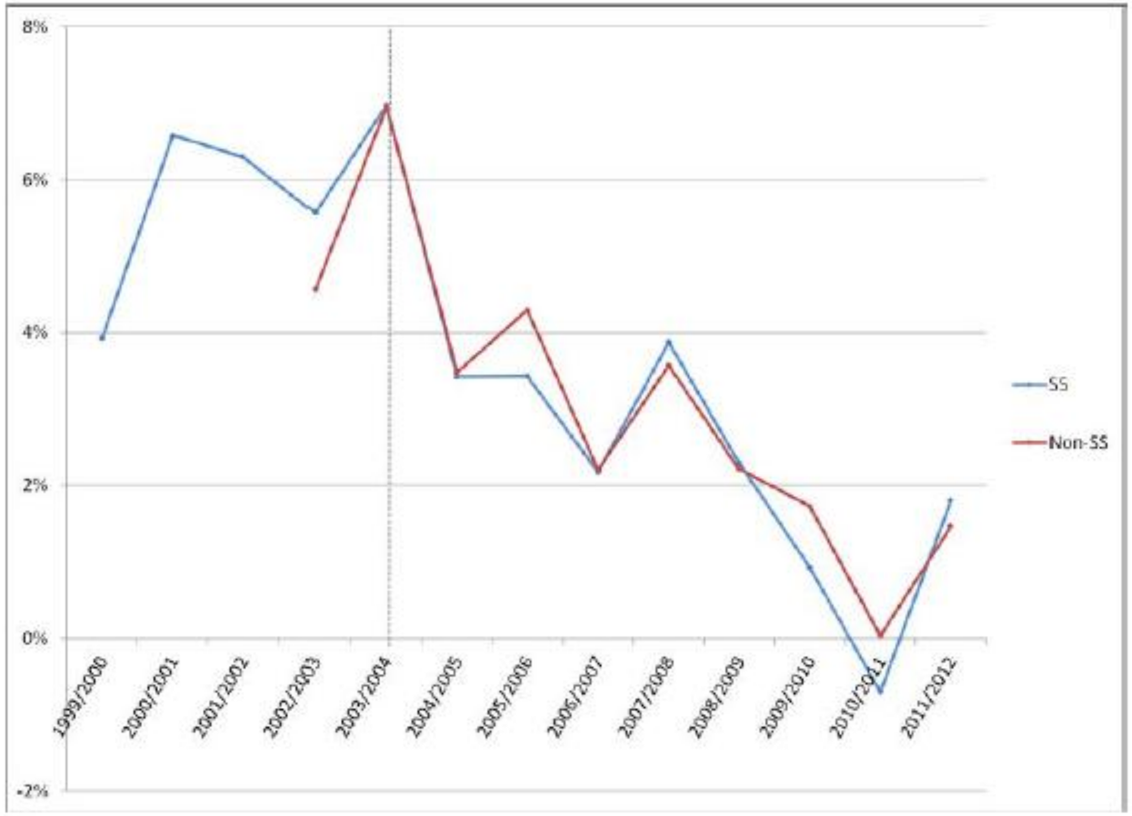
Treatment is not voluntary

Figure 3; Average fees per term (day, deflated)



Apparent Impact of the Intervention

Figure 4: Average annual increase in deflated boarding fees per term (per cent)



Econometric model

5.6 This analysis uses a panel of yearly, school-level data on fees to estimate a fixed effects model. The below econometric model is estimated:

$$\log(Fee_{it}) = \beta_0 + \beta_1.boarder\%_{it} + \beta_2.ranking\%_{it} + \beta_3.\log\,(Pupils_{it}) + \beta_4.year_t + \lambda.postintervention_t + \boxed{\delta.infringe.post_{it}} + S_i + \varepsilon_{it}$$

Treatment (Intervention) Effect = β_1 + β_2 if SS school

- $Boarder\%_{it}$ is the percentage of boarders in school *i* in year t. For example, a school with 75 per cent boarders would have a value of 0.75.
- $Ranking\ \%_{it}$ is the percentile in the Financial Times school rankings for school i in year t. For example, if a school had a ranking in year t which put them at the 80th percentile this variable would equal 0.8.
- $Pupils_{\ it}$ is the number of pupils in school i in year t.
- $Year_t$ is the relevant year and accounts for any linear trend in fees.
- $postintervention_t$ indicates whether or not the observation comes from the post-intervention period and allows for the trend, for all schools, to differ before and after the intervention.
- $infringe.post_{it}$ indicates whether or not the observation is from an SS School in the post intervention period.[30] Under specific assumptions concerning the scope and the duration of the anti-competitive agreement, the estimated result for this variable can provide a basis on which to estimate the impact of the OFT intervention. This is the pivotal variable in the difference in difference approach. A negative and statistically significant coefficient would suggest, consistent with theory that the intervention led to a reduction in fees.

Table 1: Regression results

Dependent Variable	Log(Real Day Fees)		Log(Real Boarding Fees)	
	Fixed Effect (OLS SEs)	Fixed Effect (HAC SEs)	Fixed Effect (OLS SEs)	Fixed Effect (HAC SEs)
Boarder%	0.0773***	0.0773^{+}	0.0367	0.0367
	(0.018)	(0.051)	(0.030)	(0.029)
Ranking%	-0.0147	-0.0147	0.00396	0.00396
	(0.015)	(0.019)	(0.015)	(0.015)
Log(Pupils)	0.0247^{+}	0.0247	0.0291*	0.0291^{+}
	(0.015)	(0.033)	(0.017)	(0.021)
Year	0.0698***	0.0698***	0.0709***	0.0709***
	(0.001)	(0.004)	(0.001)	(0.004)
Post intervention	0.0750***	0.0750***	0.0674***	0.0674***
	(0.005)	(0.027)	(0.006)	(0.022)
Post intervention and	**-0.0149****	**-0.0149****	**-0.0162****	**-0.0162*****
Infringer (DiD)	**(0.007)**	**(0.007)**	**(0.007)**	**(0.005)**
N	1829	1825	1317	1311
R^2	0.949	0.949	0.957	0.957

Standard errors in parentheses

$+$ $p < 0.2$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In order to test robustness two versions of the fixed effects model were run. The first is Ordinary Least Squares, and the second is heteroscedasticity and auto-correlation robust (HAC) standard errors in order to check for heteroscedasticity and autocorrelation.

Key findings

- Following OFT intervention, SS School boarding fees fell by an estimated 1.6 per cent per annum relative to what we would expect had the OFT not intervened.
- This estimate is highly statistically significant (at the 95 per cent level), and robust to a number of different specifications and sensitivity tests, and therefore presents strong evidence that OFT intervention has driven a reduction in consumer harm.
- The impact for SS School day fees is estimated at 1.5 per cent per annum. This finding, although statistically significant at the 90 per cent level, is not as robust as for boarding fees.
- The findings control for the influence of other factors – for instance quality, to the extent that this is captured by the variable 'FT rank' – and are likely to represent a lower bound of impact given the potential for broader impact across the market.

The cumulative impact of the intervention is the area between the two paths from intervention to time T.

Figure 12: Average annual boarding fees of SS Schools: actual and expected in absence of OFT intervention

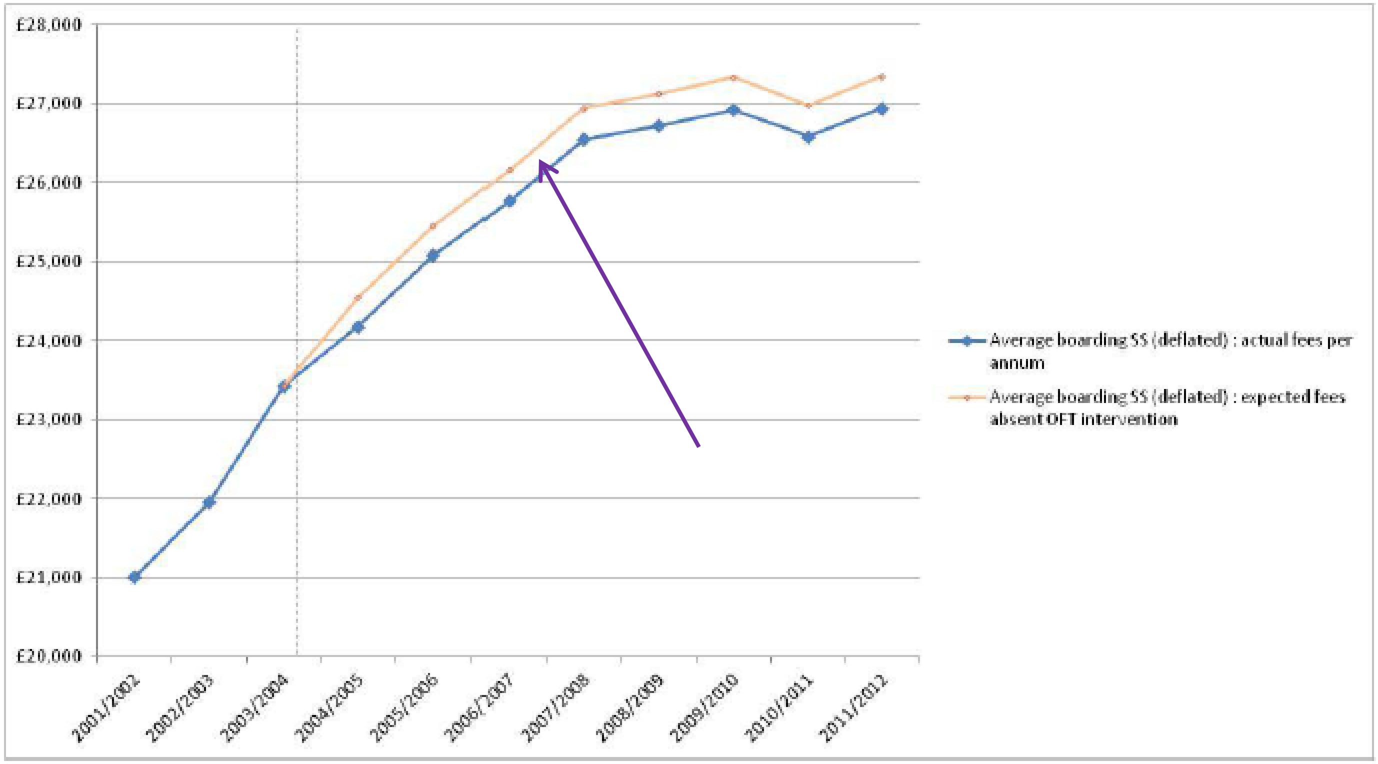
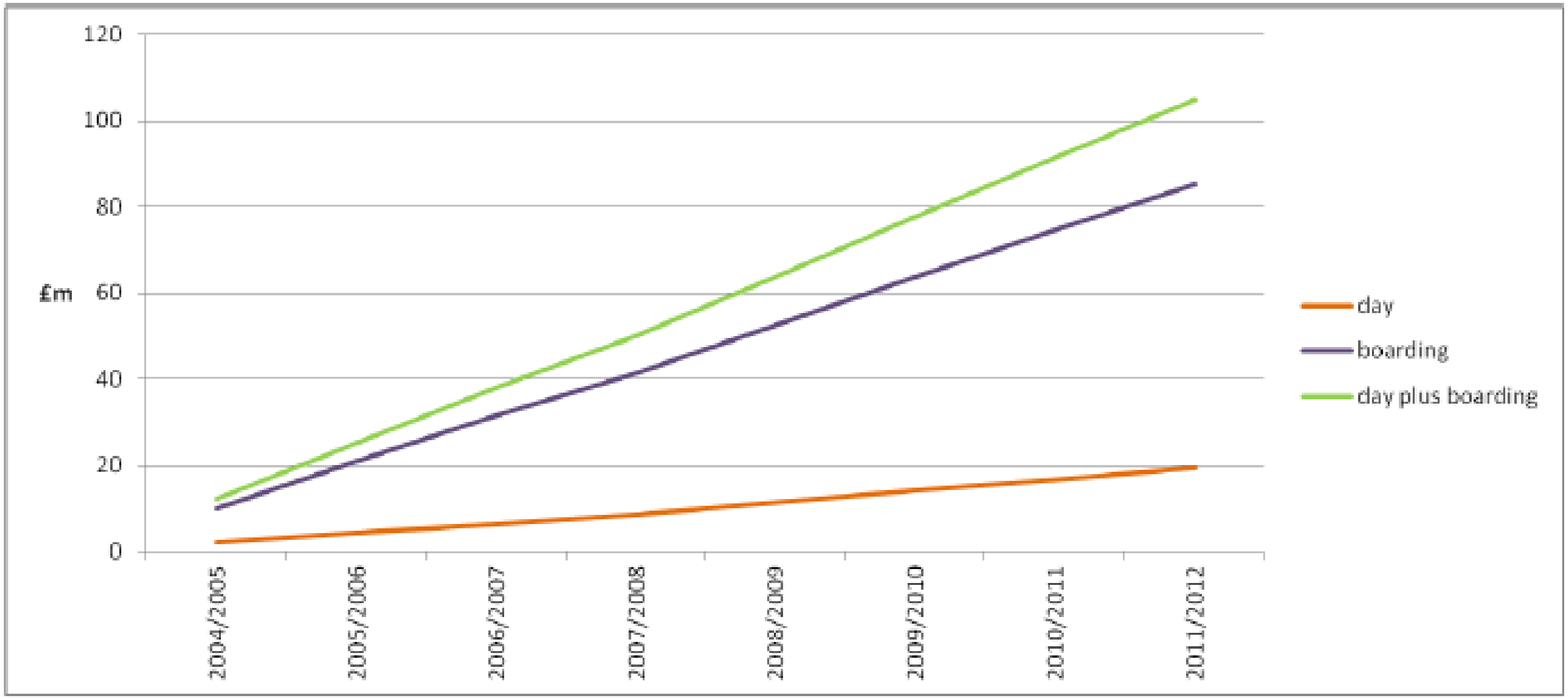


Figure 13: Cumulative savings in fees to the consumer from OFT intervention, 2010 prices, £m, discounted to present



Endogenous Treatment in SAT Tests

Example 6.8 SAT Scores

Each year, about 1.7 million American high school students take the SAT test. Students who are not satisfied with their performance have the opportunity to retake the test. Some students take an SAT prep course, such as Kaplan or Princeton Review, before the second attempt in the hope that it will help them increase their scores. An econometric investigation might consider whether these courses are effective in increasing scores. The investigation might examine a sample of students who take the SAT test twice, with scores y_{i0} and y_{i1}. The time dummy variable T_t takes value $T_0 = 0$ "before" and $T_1 = 1$ "after." The treatment dummy variable is $D_i = 1$ for those students who take the prep course and 0 for those who do not. The applicable model would be (6-3),

$$SAT\ Score_{i,t} = \beta_1 + \beta_2\, 2ndTest_t + \beta_3\, PrepCourse_i + \delta\, 2ndTest_t \times PrepCourse_i + \varepsilon_{i,t}.$$

The estimate of δ would, in principle, be the treatment, or prep course effect.

Using least squares,

$$d_3 = (\overline{Score_2} - \overline{Score_1})_{PrepCourse=1} - (\overline{Score_2} - \overline{Score_1})_{PrepCourse=0}$$

Potential $\mathbf{x}$ = Income, Parents' Education, GPA

Potential **endogeneity**: PrepCourse = $\boldsymbol{\theta}'\mathbf{z}$+w + w, Cov[u,$\varepsilon$] $\neq$ 0

Treatment Effect

- Earnings and Education: Effect of an additional year of schooling

- Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter
  - Philip Oreopoulos
  - AER, 96,1, 2006, 152-175

Treatment Effects and Natural Experiments

THE AMERICAN ECONOMIC REVIEW MARCH 2006

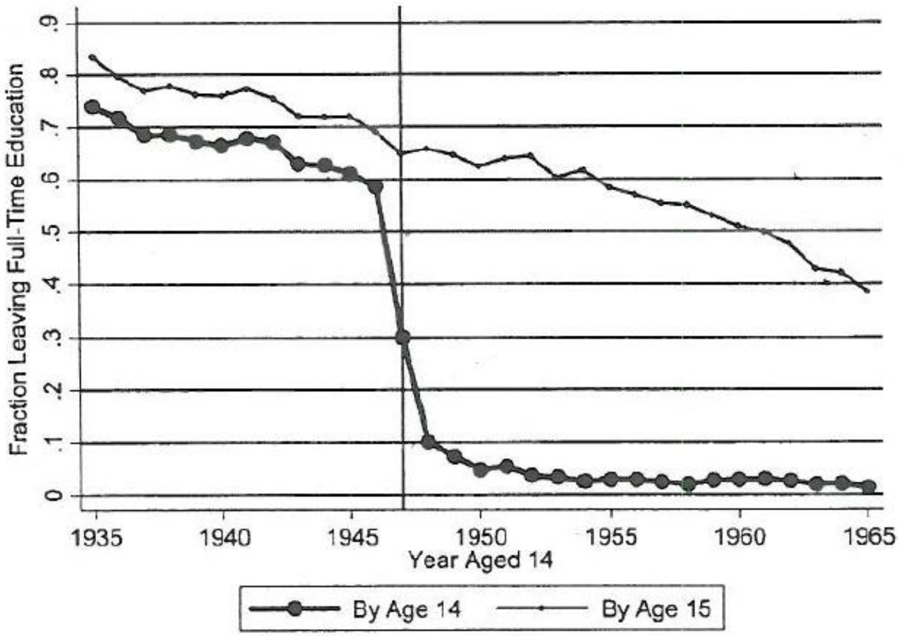


FIGURE 1. FRACTION LEFT FULL-TIME EDUCATION BY YEAR AGED 14 AND 15 (*Great Britain*)

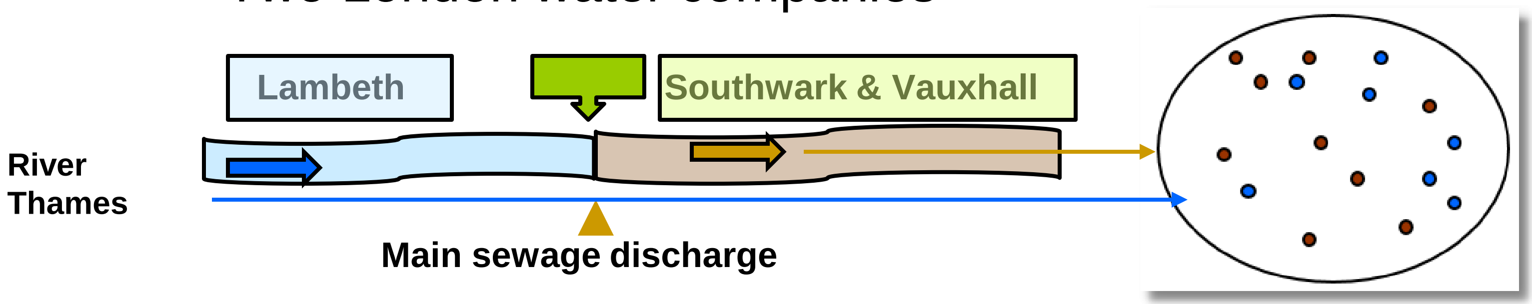
Note: The lower line shows the proportion of British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys who report leaving full-time education at or before age 14 from 1935 to 1965. The upper line shows the same, but for age 15. The minimum school-leaving age in Great Britain changed in 1947 from 14 to 15.

The First IV Study Was a Natural Experiment

(Snow, J., On the Mode of Communication of Cholera, 1855)
http://www.ph.ucla.edu/epi/snow/snowbook3.html

- London Cholera epidemic, ca 1853-4
- Cholera = f(Water Purity,u) + ε.
  - ‘Causal’ effect of water purity on cholera?
  - Purity=f(cholera prone environment (poor, garbage in streets, rodents, etc.). Regression does not work.

    Two London water companies



Paul Grootendorst: A Review of Instrumental Variables Estimation of Treatment Effects…
http://individual.utoronto.ca/grootendorst/pdf/IV_Paper_Sept6_2007.pdf

A review of instrumental variables estimation in the applied health sciences. *Health Services and Outcomes Research Methodology* 2007; 7(3-4):159-179.

Investigation Using an Instrumental Variable

Theory : Cholera $= \beta_0 + \beta_1$BadWater + Other Factors

Model : $C = \beta_0 + \beta_1 B + \varepsilon$ (Stylized)

(C=0/1=no/yes) (B=0/1=good/bad) (ε=other factors)

Interesting measure of causal effect of bad water : β_1

Endogeneity Problem : Cholera prone environment u affects B and ε.

Interpret this to say B(u) and ε(u) are correlated because of u.

$\textbf{Confounding Effect : E[Cholera | Bad Water]} \neq \beta_0 + \beta_1 B$

because there are unmodeled factors that affect cholera and water.

$$
\begin{aligned}
&E[C|B] \quad \neq \ \beta_0 + \beta_1 B \text{ because } E[\varepsilon|B] \neq 0 \\
&E[C|B=1] \ = \ \beta_0 + \beta_1 + E[\varepsilon|B=1] \\
&E[C|B=0] \ = \ \beta_0 \qquad + E[\varepsilon|B=0] \\
&E[C|B=1] - E[C|B=0] = \ \beta_1 + \{E[\varepsilon|B=1] - E[\varepsilon|B=0]\}
\end{aligned}
$$

Conclusion : Comparing cholera rates of those with bad water (measurable) to those with good water, P(C|B=1) - P(C|B=0), does not reveal the water effect.

Instrumental Variable : L = 1 if water supplied by Lambeth

L = 0 if water supplied by Southwark/Vauxhall

Relevant? Is E[B|L=1] $\neq$ E[B|L=0]? That is Snow's theory, that the water supply is partly the culprit, and because of their location, Lambeth provided purer water than Southwark.

Exogenous? Is E[ε|L=1]-E[ε|L=0]=0? Water supply is randomly supplied to houses. Homeowners do not even know which supplier is providing their water. "Assignment is random."

Using the IV in E[C|L] = $\beta_0 + \beta_1 E[B \mid L] + E[\varepsilon \mid L]$:

$$E[C \mid L = 1] = \beta_0 + \beta_1 E[B \mid L = 1] + E[\varepsilon \mid L = 1]$$

$$E[C \mid L = 0] = \beta_0 + \beta_1 E[B \mid L = 0] + E[\varepsilon \mid L = 0]$$

Estimating Equation : $E[C \mid L = 1] - E[C \mid L = 0] = \beta_1 \left\{ E[B \mid L = 1] - E[B \mid L = 0] \right\}$

$+ \left\{ E[\varepsilon \mid L = 1] - E[\varepsilon \mid L = 0] \right\}$ (zero because L is exogenous)

IV Estimator : $E[C\,|\,L=1]-E[C\,|\,L=0]=\beta_1\left\{E[B\,|\,L=1]-E[B\,|\,L=0]\right\}$

$$\beta_1=\frac{E[C\,|\,L=1]-E[C\,|\,L=0]}{E[B\,|\,L=1]-E[B\,|\,L=0]}$$ (Note : nonzero denominator is the relevance condition.)

Operational : P(C|L=1) = Proportion of observations supplied by Lambeth that have Cholera

P(C|L=0) = Proportion of observations supplied by Southwark that have Cholera

$P(B\,|\,L=1)=$ Proportion of observations supplied by Lambeth with Bad Water

$P(B\,|\,L=0)=$ Proportion of observations supplied by Southwark with Bad Water

Estimate : $$b_1=\frac{P(C\,|\,L=1)-P(C\,|\,L=0)}{P(B\,|\,L=1)-P(B\,|\,L=0)}=\text{ (broadly) }\frac{\text{Cov}(C,L)}{\text{Cov}(B,L)}$$ (The Wald estimator)

A Tale of Two Cities

- A sharp change in policy can constitute a **natural experiment**
- The Mariel boatlift from Cuba to Miami (May-September, 1980) increased the Miami labor force by 7%. Did it reduce wages or employment of non-immigrants?
- Compare Miami to Los Angeles, a comparable (assumed) city.
- Card, David, “The Impact of the Mariel Boatlift on the Miami Labor Market,” Industrial and Labor Relations Review, 43, 1990, pp. 245-257.

Difference in Differences

$i =$ individual, T = 0 for no immigration, T=1 for migration

$(Y_i \mid T) = Y_{i,T} = 1$ if unemployed, 0 if employed.

c = city, t = period.

Unemployment rate in city c at time t is $E[Y_{i,0} \mid c,t]$ with no migration

Unemployment rate in city c at time t is $E[Y_{i,1} \mid c,t]$ with migration

$$\text{Assume } E[Y_{i,0} \mid c,t] = \beta_t + \gamma_c$$

$$E[Y_{i,1} \mid c,t] = \beta_t + \gamma_c + \delta$$

$$= E[Y_{i,0} \mid c,t] + \delta$$

$\delta =$ the effect of the immigration on the unemployment rate.

Applying the Model

- c = M for Miami, L for Los Angeles
- Immigration occurs in Miami, not Los Angeles
- T = 1979, 1981 (pre- and post-)
- Sample moment equations: $E[Y_i|c,t,T]$
  - $E[Y_i|M,79] = \beta_{79} + \gamma_M$
  - $E[Y_i|M,81] = \beta_{81} + \gamma_M + \delta$
  - $E[Y_i|L,79] = \beta_{79} + \gamma_L$
  - $E[Y_i|M,79] = \beta_{81} + \gamma_L$
- It is assumed that unemployment growth in the two cities would be the same if there were no immigration.

Implications for Differences

- If neither city exposed to migration
  - $E[Y_{i,0}|M,81] - E[Y_{i,0}|M,79] = \beta_{81} - \beta_{79}$ (Miami)
  - $E[Y_{i,0}|L,81] - E[Y_{i,0}|L,79] = \beta_{81} - \beta_{79}$ (LA)
- If both cities exposed to migration
  - $E[Y_{i,1}|M,81] - E[Y_{i,1}|M,79] = \beta_{81} - \beta_{79} + \delta$ (Miami)
  - $E[Y_{i,1}|L,81] - E[Y_{i,1}|L,79] = \beta_{81} - \beta_{79} + \delta$ (LA)
- One city (Miami) exposed to migration: The **difference in differences** is.
  - $\{E[Y_{i,1}|M,81] - E[Y_{i,1}|M,79]\} - \{E[Y_{i,0}|L,81] - E[Y_{i,0}|L,79]\}$
    $= \delta$ (Miami)

Autism: Natural Experiment

- Autism ←-----→ Television watching
- Which way does the causation go?
- We need an instrument: Rainfall
  - Rainfall effects staying indoors which influences TV watching
  - Rainfall is definitely absolutely truly exogenous, so it is a perfect instrument.
- The correlation survives, so TV "causes" autism.