Data documentation

Contents

Note: all data are simulated. They do not intend to replicate results of the studies that were used to describe the policy context and theoretical rationale. The dataset were created for educational purposes only.

1 Instrumental variables

1.1 FatherEduc (lecture data)

Variable name	Description
wage	Wage
educ	Education
ability	Ability (omitted variable)
experience	Experience
fathereduc	Father's education

Father's education is used as instrumental variable to predict wage. Ability is considered the omitted variable (individuals with greater ability are more likely to stay in the education styem longer and receive higher wage). Based on previous studies such as Blackburn & Newmark, 1993 and Hoogerheide et al. 2012.

1.2 PublicHousing (lab data)

Variable names	Description
HealthStatus	Health status on a scale from 1 = poor to 5 = excellent
HealthBehavior	Omitted variable
PublicHousing	Number of years spent in a public house
Supply	Number of available public houses in the city every 100 eligible households
ParentsHealthStatus	Health status of parents
WaitingTime	Average waiting time before obtaining public housing assistance in the city (in months)
Stamp	Dollar amount of food stamps (TANF) consumed each month
Age	Age
Race	Race, 1 = White, 2 = Black, 3 = Hispanic, 4 = Other
Education	Education, 1 = High School, 2 = Diploma, 3 = Bachelor, 4 = Master
MaritalStatus	1 = Single, 2 = Married, 3 = Widow, 4 = Divorced

This labs is loosely based on the paper written by Angela R. Fertig and David A. Reingold (2007) entitled ["Public housing, health, and health behaviors: Is there connection?".](https://onlinelibrary.wiley.com/doi/epdf/10.1002/pam.20288

The scope is to predict Health Status based on the years spent receiving public housing assistance. However there is an omitted variable bias problem as individuals who cares more about their health might be more likely to self-select into public houses and report a better health status.

Simulated data contain 4 possible instrumental variables.

- Supply of public housing measured as the number of housing available every 100 households
- Health status of parents
- Waiting time for a public housing
- Dollar amount of food stamps (TANF) consumed each month

A correlation matrix including these four variables, the omitted variable, the outcome variable, and the policy variable can reveal which instrumental variables are valid.

2 Regression discontinuity

2.1 Attendance (lecture data)

Variable name	**Description**
Performance	Final exam grade, from 0 to 100
Treatment	Dummy variable if Treatment group (=1) or Control group (=0)
Attendance	Percentage of classes attended from 0% (attended no class) to 100% (attended all classes) - Rating variable
Attendance_c	Attendance variable centered around the cutoff. The cutoff was fixed at the median of Attendance

Studies show that class attendance has a positive and significant effect on student performance. These simulated data aim to estimate the effect of a new policy mandating attendance for students with a Attendance lower than the median.

2.2 SummerSchool (lab data)

Variable name	Description
STD_math_7	Standardized math score in 7th-grade, from 0 to 100
GPA_8	Math GPA at the end of 8th-grade (before summer school), from 0 to 4
STD_math_8	Standardized math scores in 8th-grade, from 0 to 100. Variables used to assign to the treatment and non-treatment group.
GPA_9	Math GPA at the end of 9th-grade (after summer school), from 0 to 4

Summer school programs are designed to help students improve their reading and math ability. They are generallly dedicated to students who have not yet achieved the skills required by the next level. There are, however, mixed evidence on whether summer school works.

Simulated data represent observations among 500 students in a US public middle school. Students participated in the summer school program at the end of the 8th-grade (end of middle school) if they had a standardized math score lower than 60. We want to understand whether attending the summer school has an affect on the math GPA at the of the 9th-grade.

Simulated data also include GPA at the end of 8th-grade and standardized test scores in 7th-grade to show the need of a regression discontinuity approach.

3 Time series

3.1 Wellbeing (lecture data)

Column	Variable name	Description
Y	Wellbeing	Wellbeing index (from 0 to 300)
T	Time	Time (from 1 to 365)
D	Treatment	Observation post (=1) and pre (=0) intervention
P	Time Since Treatment	Time passed since the intervention

Simulated dataset describing the wellbeing of a class of students. The dataset contains 365 daily observations of the wellbeing of a class of students. Wellbeing is measured by a index from 0 to 300. At t = 201 students will start attend a mental health education class. We want to understand how (and if) their wellbeing improves.

The model is based on the equation:

$$Y = b_0 + b_1 * Time + b_2 * Treatment + b_3 * TimeSinceTreatment + e \quad (1)$$

3.2 Ridership (lab data)

Variable name	Description
Passengers	Daily passengers on the buses (in thousands)

Simulated data on ridership. The variable "Passengers" represents the number of daily passengers on all buses in the city (in thousands). Dataset contains 365 observations (hypothetically from January 1st to December 31st). The intervention was implemented on May 1st (day 121 is the first day of the new schedule).

Students need to create other variables to use the dataset: a time variable counting days from 1 to 365; a time since treatment variable starting at obs 121, and a treatment variable to indicate time before (=1) and after treatment (=0).

4 Difference in difference

4.1 Housing data (lecture data)

Variable name	Description
House Price	House pricing
Group	Houses close to a subsized housing site (=1) and houses far from a subsized housing site (=0)
Post_Treatment	Subsized housing site is not constructed (=0) or subsized housing site is open (=1)

Simulated data were loosely based on this work conducted by Ellen and colleagues in 2007.

Data include housing prices data for 2 groups of houses: houses that are located close to a subsized housing site (treatment group) and houses that are far away from a subsized housing site (control group). The data examine whether prices have changed before and after the creation of the subsized housing site.

4.2 Greening (lab data)

Variable name	Description
Vandalism	Number of calls about vandalism acts near to the vacant lots
Group	Treatment (=1) and control group (=0)
Post_Treatment	Observation three years before the treatment (=0) and three years after the treatment (=1)

This lab is loosely based on Branas et al. (2011) article, entitled "A Difference-in-Differences Analysis of Health, Safety, and Greening Vacant Urban Space".

Simulated data on vandalism acts include the total number of calls about vandalism acts that the police has received during the year before and after the greening from households near the vacant lots.

The 'broken windows' theory suggests that "vacant lots offer refuge to criminal and other illegal activity and visibly symbolize that a neighborhood has deteriorated, that no one is in control, and that unsafe or criminal behavior is welcome to proceed with little if any supervision" (p. 1297). To prevent these problems, city A has decided to implement a public program that will green some of the city's vacant lots in 4 different neighborhoods. 'Greening' includes removing trash and debris and planting grass and trees to create a park.

Three years after its implementation, the city wants to know whether the program has been successful and propose to compare lots that have been 'greened' and lots that are still vacants within the same neighborhoods but could have been chosen for greening.

5 Logit

5.1 LSAT (lecture data)

Variable name	Description
LSAT	LSAT score (from 120 to 180)
Essay	Essays score (from 1 to 365)
GPA	Average GPA (from 0 to 5)
Admission	The student has been admitted (=1) or not (=0)

Simulated data on the probability that a student will be admitted to law school based on LSAT score, essay score, average GPA.

5.2 Vaccination (lab data)

The lab is loosely based on the research conducted by Doan & Kirkpatrick (2013) entitled "Giving Girls a Shot: An Examination of Mandatory Vaccination Legislation" and published on Policy Studies Journal, 41(2), 295-318

Variable name	Description
Adoption	Whether a state legislature has considered a mandatory vaccine law (=1) or not (=0)
Democrats	Percentage of House Democrats in the state legislature
Evangelic	Percentage of Evangelic population
Catholics	Percentage of Catholic population
Media	Number of articles covering the HPV vaccine issue in major news sources in a state in the past year
Merck	Average total dollar amount of contributions given to candidates for state offices

Hypotheses:

- Media salience is negatively correlated with mandatory legislation.
- Percentage of democrats is positively correlated with mandatory legislation.
- Religiosity (percentage of catholic and evangelic population) is negatively correlated with mandatory legislation.
- Merck PAC contribution is positively correlated with mandatory legislation.

6 Fixed effect

6.1 Companies (lecture data)

Variable name	Description
RDPub	Level of public investment in R&D
RD	Company investment in R&D
Company	Set of dummy variables indicating the company
Year	Set of dummy variable indicating the year

Simulated panel data about 10 companies over a five-year period. In this example, we want to assess the impact of public expenditure on research and development (R&D) on companies' investments.

6.2 Beer taxes (lab data)

Variable name	Description
state	Each state is indicated with a number, from 1 to 7
taxes	Beer taxes in percentage %, ranges from 0 to 1
year	Year in which observations were collected
accidents	Number of car accidents

Simulated data for 7 southern US states. For each state we have observations across 7 years. The data are structured as a panel dataset.

The lab aims to estimate the effect of beer taxes on mortality rates due to car accidents during night time. Sample study here. Beer taxes are a way for states to control consumption of alcohol; higher beer taxes increase prices, which in turn decrease consumption. Lower consumption of alcohol is expected to decrease drunk driving and therefore accidents, especially at nigth time.