Advanced Quantitative Methods

Thinking about causal inference

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Today

- A brief history of social science research (1854–present)
- A brief introduction to LaTeX

In 1854, a cholera epidemic hit London

- In 1848–49: 54,000 die in England and Wales
- In 1854: Another 20,000 dead

The prevailing theories

- From bad air rising from organic matter (miasma theory)
- Cholera correlated with lower altitude
  - Cleaner air at higher altitudes?
- Cholera is correlated with one's occupation (e.g. egg merchants, fish mongers, cowkeepers)
- No germ theory of disease

A government report concludes that cholera was caused by:

1. Stagnant air due to a lack of wind
2. High barometric pressure
3. High river water temperature at night

But an important English physician thinks otherwise

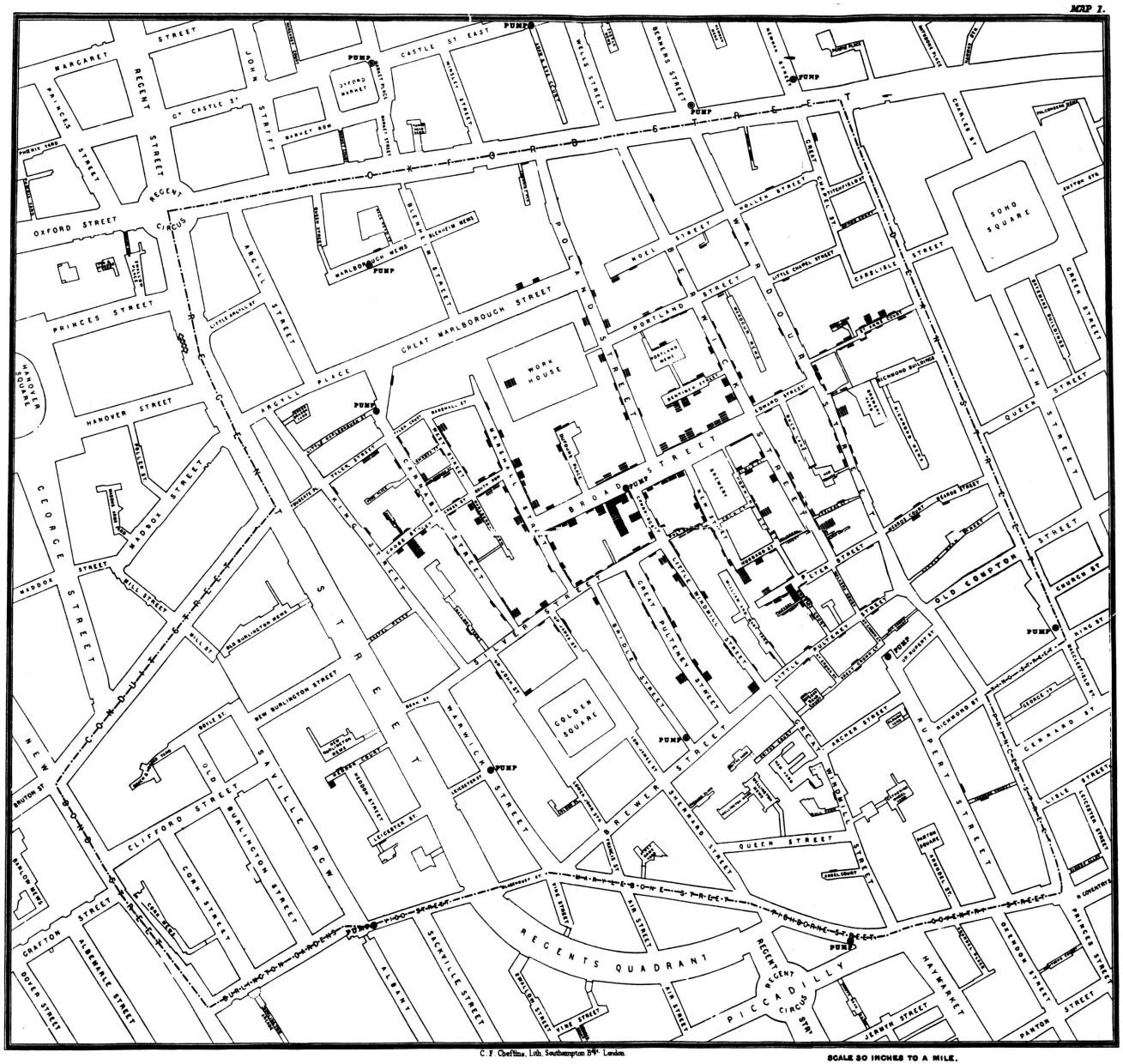
- John Snow, M.D., believes that cholera was transmitted through water

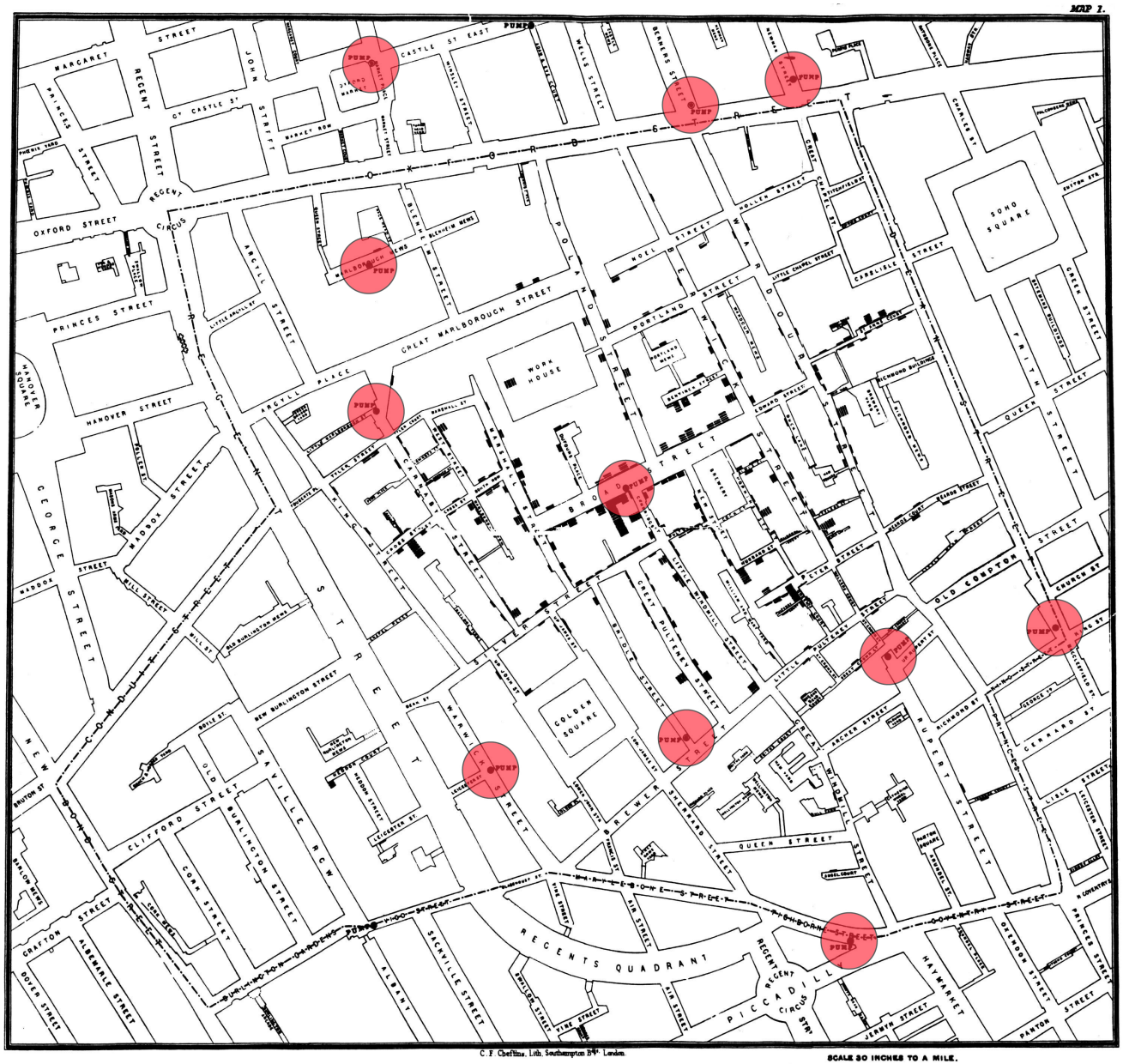
In 1850s London, there were many different water companies

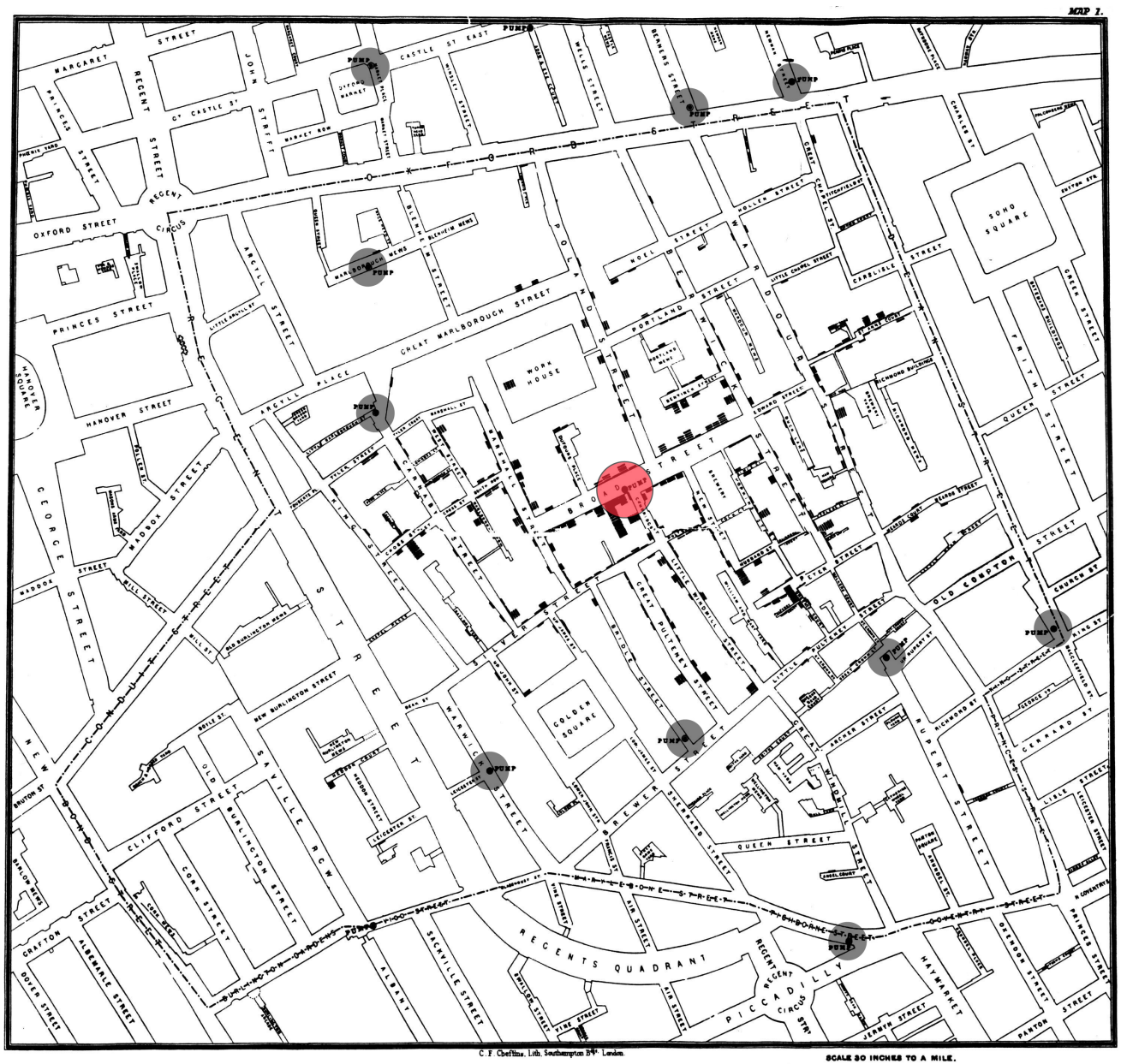
- Many take water from the Thames: a river heavily polluted by sewage
- In 1849, Lambeth water company moved upstream
- Southwark and Vauxhall water company stayed downstream

John Snow's insight

- The location of these companies was a "natural" or "quasi-" experiment
- But this would be only to the extent that people served by Lambeth Co., and by Southwark and Vauxhall Co. was *as-if* randomly assigned to people in the area


MAP 1.
OXFORD STREET
REGENT STREET
PRINCES STREET
HANOVER STREET
GEORGE STREET
MADDOX STREET
NEW BURLINGTON STREET
CLIFFORD STREET
SAVILLE ROW
BOND STREET
GREAT MARLBOROUGH STREET
ARGYLL PLACE
BROAD STREET
GOLDEN SQUARE
WARWICK STREET
REGENTS QUADRANT
PICCADILLY
REGENT CIRCUS
HAYMARKET
SOHO SQUARE
OLD COMPTON STREET
RUPERT STREET
QUEEN STREET
WORK HOUSE
PUMP
C. F. Cheffins, Lith. Southampton Bldgs. London.
SCALE 30 INCHES TO A MILE.


MAP 1.
PUMP
BROAD STREET
GOLDEN SQUARE
REGENTS QUADRANT
PICCADILLY
OXFORD STREET
REGENT STREET
SOHO SQUARE
WORK HOUSE
C. F. Cheffins, Lith. Southampton Bldgs. London.
SCALE 30 INCHES TO A MILE.


MAP 1.
OXFORD STREET
REGENT STREET
PRINCES STREET
HANOVER STREET
GEORGE STREET
BROAD STREET
GOLDEN SQUARE
SOHO SQUARE
REGENTS QUADRANT
PICCADILLY
PUMP
C. F. Cheffins, Lith. Southampton Bldgs. London.
SCALE 30 INCHES TO A MILE.

Snow investigates differences between the houses supplied by each company:

- 300,000 people served by the companies in total
- Rich and poor
- From houses small and large
- People of similar profiles across sex, age, & occupation
- A single house often has a water supply different on either side

"As there is no difference whatever either in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of Cholera than this."

– John Snow (1855)

Snow's results:

Company	Number of houses	Cholera deaths	Deaths per 10,000 houses
Lambeth ($T_i = 0$)	26,107	98	37 / 10,000
Southwark & Vauxhall ($T_i = 1$)	40,046	1,263	315 / 10,000

What is stunning is that Snow had these insights within an inauspicious context:

- Before knowledge of the benefits of randomization
- Before much knowledge of statistical inference
- Before widespread use of graphing

"Snow did some brilliant detective work on non-experimental data. What is impressive is not the statistical technique, but the handling of the scientific issues. He made steady progress from shrewd observation through case studies to analyze ecological data. In the end, he found and analyzed a natural experiment."

– David Freedman (1991)

Fast-forward two centuries..

Meat

Uproar after research claims red meat poses no health risk

One expert says findings by international experts represent 'egregious abuse of evidence'

Sarah Boseley *Health editor*

Mon 30 Sep 2019 22.00 BST

1,496

But also impressive large-scale social science field experiments

RESEARCH ARTICLE

POLITICAL SCIENCE

Reverse-engineering censorship in China: Randomized experimentation and participant observation

Gary King,[1]* Jennifer Pan,[1] Margaret E. Roberts[2]

Existing research on the extensive Chinese censorship organization uses observational methods with well-known limitations. We conducted the first large-scale experimental study of censorship by creating accounts on numerous social media sites, randomly submitting different texts, and observing from a worldwide network of computers which texts were censored and which were not. We also supplemented interviews with confidential sources by creating our own social media site, contracting with Chinese firms to install the same censoring technologies as existing sites, and—with their software, documentation, and even customer support—reverse-engineering how it all works. Our results offer rigorous support for the recent hypothesis that criticisms of the state, its leaders, and their policies are published, whereas posts about real-world events with collective action potential are censored.

The growth in empirical research

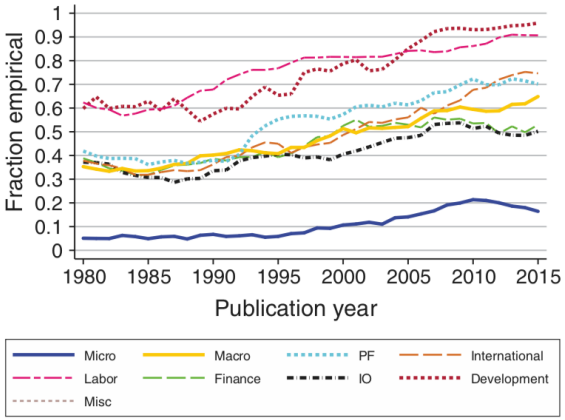


FIGURE 4. WEIGHTED FRACTION EMPIRICAL BY FIELD

Note: Five-year moving averages of the weighted fraction of publications in each field that are empirical.

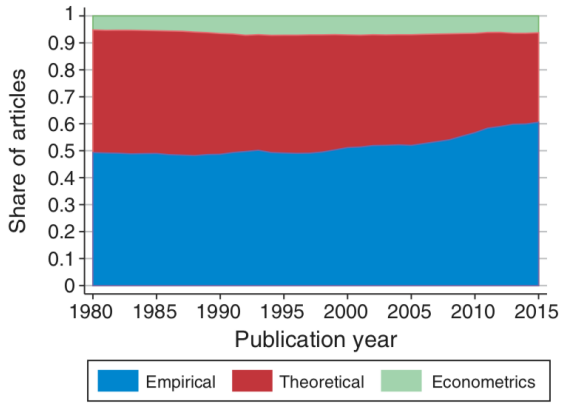


FIGURE 5. PUBLICATIONS BY STYLE

Note: Five-year moving averages of unweighted publication shares in each style.

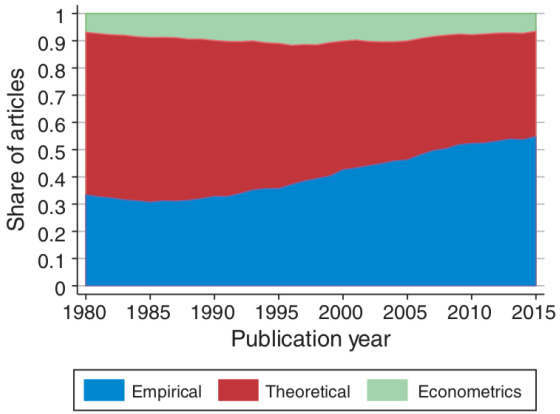


FIGURE 6. WEIGHTED PUBLICATIONS BY STYLE

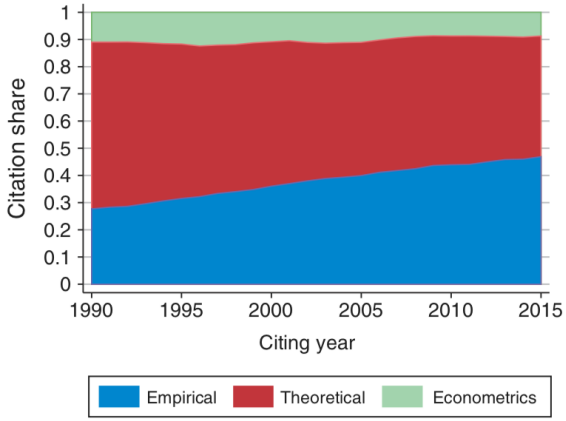


FIGURE 7. WEIGHTED CITATIONS BY STYLE

1970s and 1980s

"The 1970s and early 1980s saw rapid growth in mainframe computer size and power. Stata had yet to appear, but magnetic tape jockeys managed to crunch more and more numbers in increasingly elaborate ways. For the most part, however, increased computing power did not produce more credible estimates."

– Angrist & Pischke (2010)

Garbage-can regressions:

"Many social scientists believe that dumping long lists of explanatory variables into linear regression, probit, logit, and other statistical equations will successfully 'control' for the effects of auxiliary factors. Encouraged by convenient software and ever more powerful computing, researchers also believe that this conventional approach gives the true explanatory variables the best chance to emerge."

– Achen (2005)

e.g. Robert Putnam (2007) on neighborhood trust

Table 3. Predicting Trust in Neighbours from Individual and Contextual Variables

	B	S. E.	Beta	t	Sig.
(Constant)	0.79	0.11		7.0	0.0000
R's age	0.01	0.00	0.15	21.4	0.0000
R owns home (v. rent)	0.25	0.01	0.13	19.7	0.0000
R's education (years)	0.04	0.00	0.13	19.1	0.0000
R's ethnicity: black	−0.31	0.02	−0.12	−18.6	0.0000
Census tract poverty rate	−0.66	0.09	−0.08	−7.1	0.0000
R's satisfaction with current finances	0.10	0.01	0.08	12.4	0.0000
R's ethnicity: Latino	−0.24	0.02	−0.07	−9.8	0.0000
R's household income ($100,000)	0.14	0.02	0.05	7.5	0.0000
County: Non-violent Crimes per Capita	−2.57	0.41	−0.05	−6.2	0.0000
Census tract Herfindahl Index of Ethnic Homogeneity	**0.18**	**0.04**	**0.04**	**5.1**	**0.0000**
Census Tract Population Density (100,000 per sq. mi)	−0.39	0.08	−0.04	−4.8	0.0000
Census Tract Percent Living Same Town as Five Years Earlier	−0.24	0.04	−0.04	−5.4	0.0000
R's decades in this community	.020	.004	0.04	5.3	0.0000
Census Tract Percent Renters	−0.14	0.04	−0.04	−3.5	0.0006
Census Tract Percent Bachelor's Degree	0.29	0.07	0.03	4.3	0.0000
R is Spanish-speaker	−0.13	0.03	−0.03	−4.1	0.0001
R is female	0.05	0.01	0.03	4.7	0.0000
Census Tract Gini Coefficient for Household Income	0.39	0.15	0.02	2.7	0.0069
Census Tract Average Commute Time (hours)	−0.21	−0.06	−0.02	−3.4	0.0006
R's ethnicity: Asian	−0.09	0.03	−0.02	−3.3	0.0011
Census Tract Percent United States Citizens	0.21	0.09	0.02	2.2	0.0264
County: Violent Crimes per Capita	6.59	3.35	0.02	2.0	0.0489
Census Tract Percent Over 65	0.21	0.10	0.01	2.1	0.0364
R is a citizen	0.06	0.03	0.01	2.1	0.0356
R's average monthly work hours	.002	.001	0.01	1.8	0.0732
R is resident of South	−0.02	0.02	−0.01	−1.2	0.2182
R is resident of Midwest	−0.02	0.02	−0.01	−1.0	0.3296
R is resident of West	0.01	0.02	0.01	0.8	0.4238
R's commuting time (hours)	−0.00	0.01	0.00	−0.2	0.8069

Notes: Question was 'How much can you trust people in your neighbourhood?' N = 23,260. Adj. R^2 = 0.26.

Leamer (1983) proposes sensitivity analysis

- Examine multiple regression specifications
- Examine how parameter of interest varies

This became the norm. e.g. The main table from Fearon & Laitin (2003) on ethnicity and civil war:

TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	Model				
	(1) Civil War	(2) "Ethnic" War	(3) Civil War	(4) Civil War (Plus Empires)	(5) Civil War (COW)
Prior war	−0.954**	−0.849*	−0.916**	−0.688**	−0.551
	(0.314)	(0.388)	(0.312)	(0.264)	(0.374)
Per capita income[a,b]	−0.344***	−0.379***	−0.318***	−0.305***	−0.309***
	(0.072)	(0.100)	(0.071)	(0.063)	(0.079)
log(population)[a,b]	0.263***	0.389***	0.272***	0.267***	0.223**
	(0.073)	(0.110)	(0.074)	(0.069)	(0.079)
log(% mountainous)	0.219**	0.120	0.199*	0.192*	0.418***
	(0.085)	(0.106)	(0.085)	(0.082)	(0.103)
Noncontiguous state	0.443	0.481	0.426	0.798**	−0.171
	(0.274)	(0.398)	(0.272)	(0.241)	(0.328)
Oil exporter	0.858**	0.809*	0.751**	0.548*	1.269***
	(0.279)	(0.352)	(0.278)	(0.262)	(0.297)
New state	1.709***	1.777***	1.658***	1.523***	1.147**
	(0.339)	(0.415)	(0.342)	(0.332)	(0.413)
Instability[a]	0.618**	0.385	0.513*	0.548*	0.584*
	(0.235)	(0.316)	(0.242)	(0.225)	(0.268)
Democracy[a,c]	0.021	0.013			
	(0.017)	(0.022)			
Ethnic fractionalization	0.166	0.146	0.164	0.490	−0.119
	(0.373)	(0.584)	(0.368)	(0.345)	(0.396)
Religious fractionalization	0.285	1.533*	0.326		1.176*
	(0.509)	(0.724)	(0.506)		(0.563)
Anocracy[a]			0.521*		0.597*
			(0.237)		(0.261)
Democracy[a,d]			0.127		0.219
			(0.304)		(0.354)
Constant	−6.731***	−8.450***	−7.019***	−6.801***	−7.503***
	(0.736)	(1.092)	(0.751)	(0.681)	(0.854)

Some take Leamer (1983) to the extreme

I Just Ran Two Million Regressions

By XAVIER X. SALA-I-MARTIN*

Following the seminal work of Robert Barro (1991), the recent empirical literature on economic growth has identified a substantial number of variables that are partially correlated with the rate of economic growth. The basic methodology consists of running cross-sectional regressions of the form

$$\gamma = \alpha + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_n \mathbf{x}_n + \varepsilon \quad (1)$$

where γ is the vector of rates of economic growth, and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are vectors of explanatory variables, which vary across researchers

An initial answer to this question was given by Ross Levine and David Renelt (1992).[1] They applied Edward Leamer's (1985) *extreme-bounds test* to identify "robust" empirical relations in the economic growth literature. In short, the extreme-bounds test works as follows. Imagine that there is a pool of N variables that previously have been identified to be related to growth and one is interested in knowing whether variable z is "robust." One would estimate regressions of the form

$$\gamma = \alpha_j + \beta_{yj}\mathbf{y} + \beta_{zj} z + \beta_{xj}\mathbf{x}_j + \varepsilon \quad (2)$$

Achen (2005) on model specification

Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong

CHRISTOPHER H. ACHEN

Department of Politics
Princeton University
Princeton, New Jersey, USA

Many social scientists believe that dumping long lists of explanatory variables into linear regression, probit, logit, and other statistical equations will successfully "control" for the effects of auxiliary factors. Encouraged by convenient software and ever more powerful computing, researchers also believe that this conventional approach gives the true explanatory variables the best chance to emerge. The present paper argues that these beliefs are false, and that without intensive data analysis, linear regression models are likely to be inaccurate. Instead, a quite different and less mechanical research methodology is needed, one that integrates contemporary powerful statistical methods with deep substantive knowledge and classic data–analytic techniques of creative engagement with the data.

Recall that the functional form of controls need to be correctly specified also

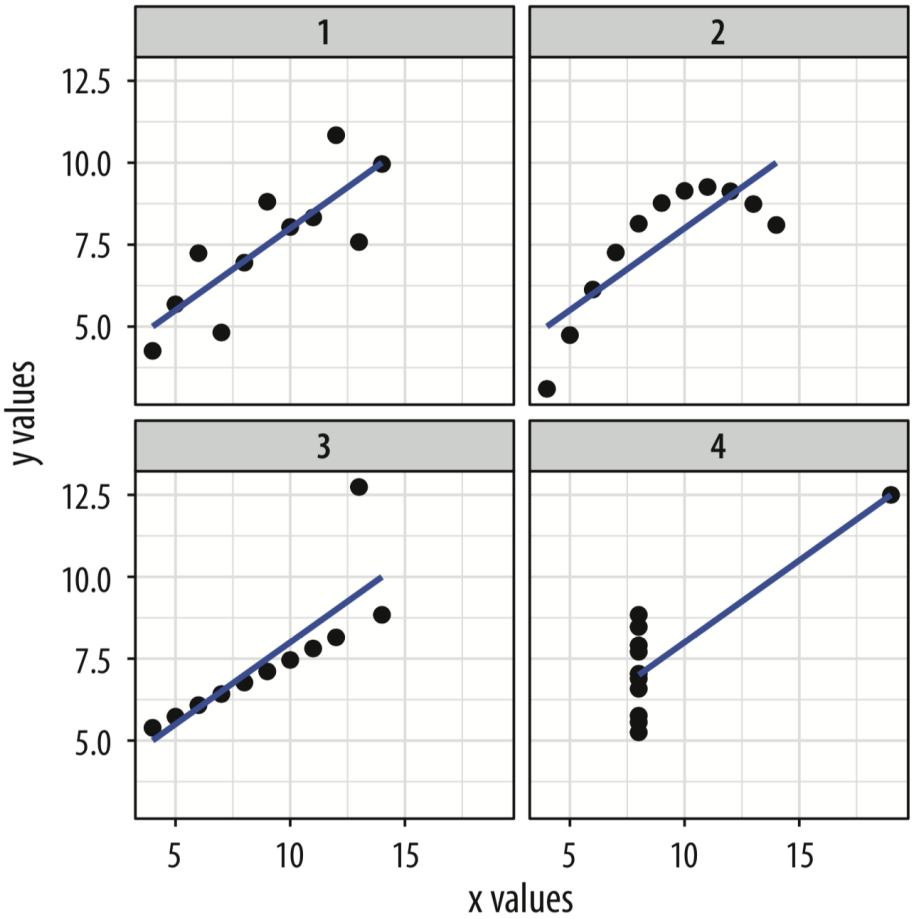


Figure 1.1: Plots of Anscombe's quartet.

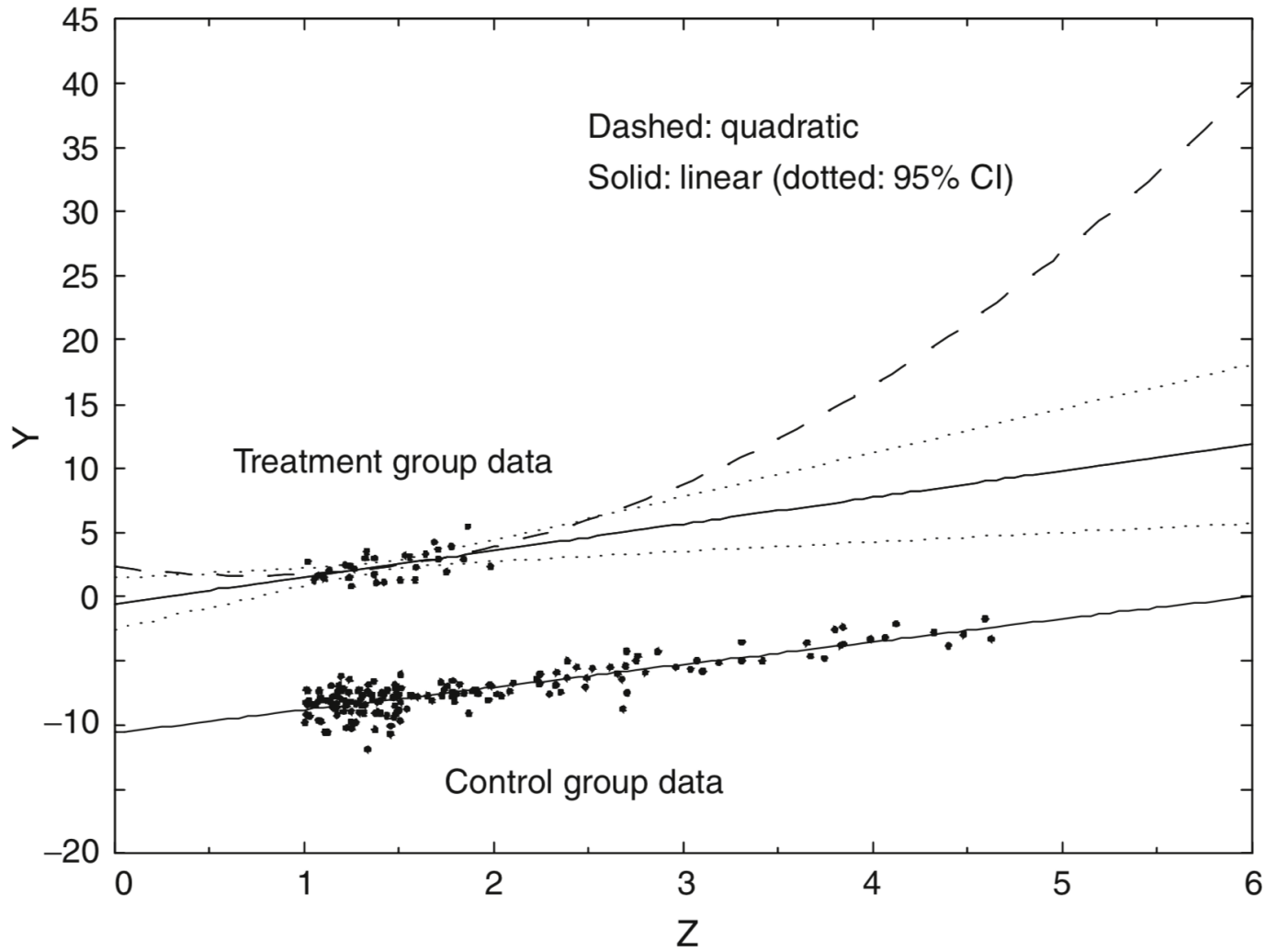


FIG. 6. Illustrating Interpolation and Extrapolation Bias

But then, from the 2010s onward:

- Shift from **model-based** to **design-based** causal inference
- Called the "credibility revolution"
- Focus on clever design, not fancy models
  - Fancy methods now a bit back in fashion to augment these clever designs
- Most top-level political science relies on:
  - Relevant framing, i.e. why your research question and results matter
  - Clarity of writing
  - Good data
  - Clever research design
  - Methodological rigor
- There's a reason we only need OLS in this course

The Credibility Revolution (1990s-present)

"Empirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact. Sensitivity analysis played a role in this, but as we see it, the primary engine driving improvement has been a focus on the quality of empirical research designs. This emphasis on research design is in the spirit of Leamer's critique, but it did not feature in his remedy."

– Angrist & Pischke (2010)

Pseudo-general pseudo-facts

"At the turn of the millennium, the modal quantitative research design was one in which researchers assembled data on theoretically interesting dependent and independent variables ... Researchers then assessed the presumably causal relationships in these data using regressions with informally motivated sets of control variables to reduce the potential for confounding."

– Samii (2016)

The shift toward causal empiricism

"This convention in quantitative causal research appears to be breaking down, and more quantitative causal research is moving toward causal empiricism. This ... represents a major change in what researchers believe are credible ways of doing causal inference."

– Samii (2016)

Why did the Credibility Revolution happen?

- Big papers on IV, diff-in-diff, and RDD in the 1990s
- Frustration with inadequacy of sensitivity analyses
- More transparent discussions of identification strategies
- Better and more data

Increasingly strong preference toward:

- Field
- Lab experiments
- Quasi-experimental designs (thinking like John Snow)
- Clever measurement strategies

The counter-revolution? Some pushed back, and still push back against the credibility revolution

- Estimated treatment effects are often local
  - High internal validity, but low external validity
- Methods often drive the questions we ask
  - Too little theory?
  - No one answering the big questions?
- Angus Deaton was the most vocal critique of the shift toward Randomized Controlled Trials (RCTs) and quasi-experiments

Counter to the counter-revolution:

- Is not a realism and humility about what data can say
  - "Without an experiment, a natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive." (Sekhon, 2009)
  - Regression approaches are still useful, and can be thought of as descriptive
- We still need in-depth contextual knowledge
  - Need to know why treatments might be assigned as they are (e.g. settler mortality)
  - Qualitative research still matters
- Better LATE ("Local Average Treatment Effect") than never

Representative sample ≠ generalizable result

FIGURE 1 Example of nominal and effective samples from Jensen (2003)

Note: On the left, the shading shows countries in the nominal sample for Jensen (2003) estimate of the effects of regime type on FDI. On the right, darker shading indicates that a country contributes more to the effective sample, based on the panel specification used in estimation.

The more that controls explain treatment status, the less that an observation contributes to the "causal" estimate

Post-treatment bias

- Not just omitted variable bias and functional form
- We cannot naively control for all variables that are correlated with both the treatment and the outcomes (else we end up with "bad" controls)

Post-treatment bias example

- The effect of ethnic fractionalization on civil war (Fearon and Laitin, 2003)
- Authors find no "effect" (relationship) (see slide 24)
- But they also control for a strong post-treatment variable: GDP

Fearon & Laitin (2003) replication from Samii (2016):

Table 1. Replication and Auxiliary Analyses for Laitin and Fearon (2003)

	Outcome						
	Civil War Onset						Per Capita Income
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimator	Logit	Logit	Logit	Logit	Logit	Logit	OLS
Prior war	−.95 **				−.24	−.38	
	(.31)				(.23)	(.25)	
Per capita income	−.34***			−.29***		−.29***	
	(.07)			(.07)		(.07)	
Ethnic fractionalization	.17	1.12***	1.12**	.35	1.16**	.40	−4.14***
	(.37)	(.33)	(.42)	(.39)	(.43)	(.40)	(.90)
Observations	6,327	6,610	6,610	6,373	6,610	6,373	6,373
Country-clustered SEs			Y	Y	Y	Y	Y

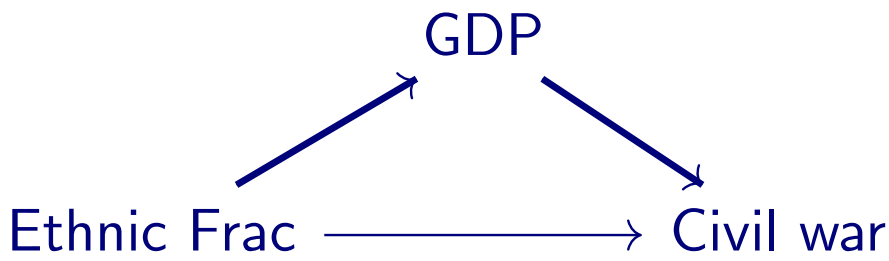
Note. Regression coefficients with standard errors in parentheses. To save space the table omits from column 1 coefficients for the following control variables: log(population), log(% mountainous), noncontinguous state, oil exporter, new state, instability, democracy, religious fractionalization, and the constant term.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

GDP is a "bad" control if the relationship is as follows:



Another example

Former Employees Are Suing Google Over Alleged Gender Discrimination

Melanie Ehrenkranz
9/14/17 4:00pm • Filed to: GOOGLE

18.2K 26 2

Photo: Getty

Google's response

After the *New York Times* detailed the employee spreadsheets on Friday, Google spokesperson Gina Scigliano told Gizmodo that its own data shows, when you take "location, tenure, job role, level and performance" into account, that "women are paid 99.7% of what men are paid at Google." Scigliano described the *Times* story as "extremely flawed."

In sum

- Past regression + controls approaches do not provide inherently generalizable results
  - No paper can easily provide generalizable causal facts without strong assumptions
  - There are no techniques to provide these facts for a population of interest
- Theory is probably better built from facts developed across multiple studies
- Theory still can be developed from descriptive analysis; ethnographies; qualitative work

Observational versus experimental designs in the wild

A yoga class at a New York wellness center. Randomized controlled trials can reverse the conclusions of observational studies. Chad Rhym for The New York Times

Empirical expectations of wellness programs

- Increase well-being
- Increase productivity
- Decrease absenteeism
- Decrease medical spending

Findings from observational designs

- Increase well-being!
- Increase productivity!
- Decrease absenteeism!
- Decrease medical spending!

The problem? Selection. Those who partake in a wellness program are:

- Likely to be healthier to begin with
- Less likely to require medical spending

Controls

Fortunately, observational studies include controls, right?

- Yes.

And properly chosen controls allow us to estimate a causal parameter, right?

- Yes.

The empirics of an observational versus experimental design comparison

- Large-scale randomized controlled trial (RCT).
- Accounts for the fact that not everyone follows through on the treatment
  - i.e. not everyone chooses to use the wellness program
  - We'll learn how to do this soon

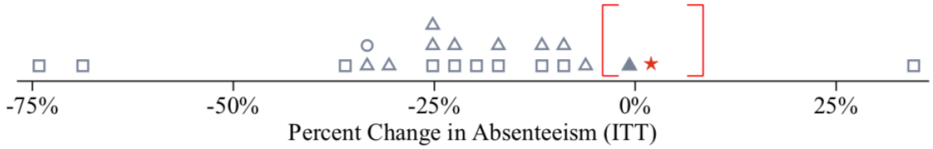
Results

Outcome	Observational result	Experimental result
Well-being	Increase!	
Productivity	Increase!	
Medical spending	Decrease!	
Absenteeism	Decrease!	

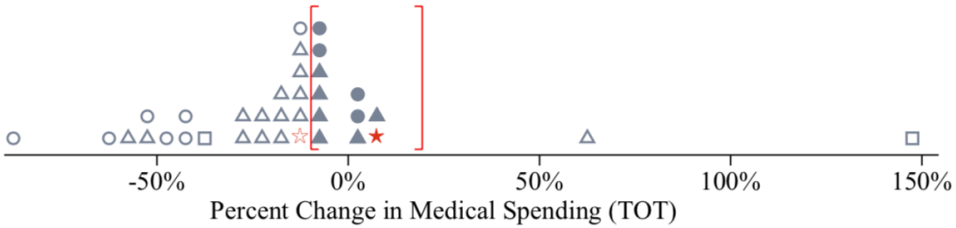
Results

Outcome	Observational result	Experimental result
Well-being	Increase!	**Null**
Productivity	Increase!	**Null**
Medical spending	Decrease!	**Null**
Absenteeism	Decrease!	**Null**

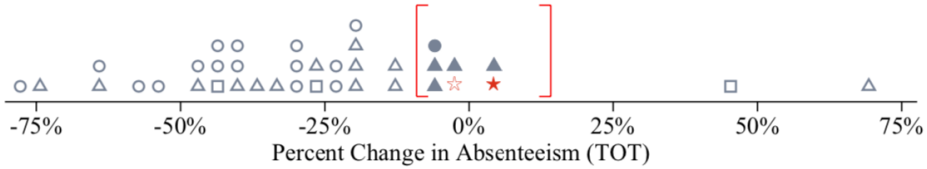
$N = 22$



$N = 22$



$N = 33$



$N = 35$

(a) Medical Spending Estimates

(b) Absenteeism Estimates

IL Estimates: ★ RCT [ ] 95% Confidence Interval ☆ OLS
Prior Studies:
Not Ruled Out: ■ RCT ▲ Pre/Post ● Other
Ruled Out: □ RCT △ Pre/Post ○ Other

Figure 6: Comparison of Experimental Estimates to Prior Studies

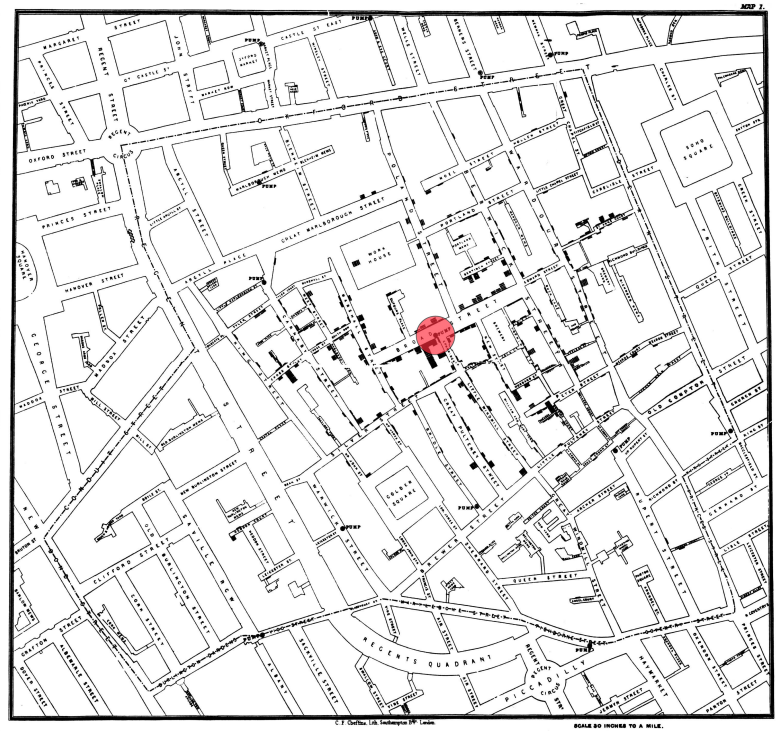
But the authors also analyzed the experiment as if it were an observational study

- Find all the same non-null results as all the other observational studies

"If we had published only these observational analyses, the headline result could have been that even after controlling for a battery of confounding variables, participation in a wellness program was associated with a significant reduction in health care spending, an improvement in exercise, and a lower chance of ceasing employment."

Back to the Future

- **Experiments**
  - Survey
  - Lab
  - Lab-in-the-field
  - Field
- **Panel data**
- **Quasi-experiments**
  - Instrumental variables
  - Difference-in-differences
  - Regression discontinuity



Overleaf exercise

- Create a new LaTeX document, e.g. 1_Main.tex
- Add a title page with your name
- Add an abstract (use fake text, e.g. from a "lorem ipsum" generator website)
- Add two sections ("Introduction" & "Theory")
- Add in a subsection under "Introduction" called "Literature review"
- Add in a figure, and reference the figure in the text, e.g. "As Figure 1 shows..."
- Add in a regression table (use the R code to generate it with modelsummary), and reference the table in text e.g. "As Table 1 shows..."
- Create a bibliography file, e.g. 3_Bibiography.tex; add a source; cite it in the text. Ensure it shows up in the "References" section of the PDF.
- Add a quotation from an author from your bibliography. e.g. As Hansen et al. (2023) state, "populism is now in decline."
    - Note that you need to use two backticks (``) to open a quotation and two single apostrophes ('') to close a quotation. Using the quotation marks from your keyboard results in only closing quotation marks ("Like this", instead of something "like this").