Causal inference week 8: Treatment effect heterogeneity

Andy Eggers

Oxford DPIR

HT 2018

Introduction

Overview

Strategies for estimating effects of treatments:

- Randomize treatment and take the DIGM
- Identify and control for confounding variables such that the CIA holds
- Identify an instrumental variable and use two-stage-least-squares to estimate average treatment effect for compliers
- Identify a situation in which the treatment depends on a cutoff
- Use observations at more than one point in time in a diff-in-diff

Congratulations if you can convincingly do **any one** of these for **any sample**

Today: Going further, can we measure how treatment effects **vary across subgroups**?

What do we want to know?

Recall the **conditional average treatment effect** where $X = x$:

$$\text{CATE}_x \equiv E[Y_i(1) - Y_i(0)|X_i = x] \equiv E[\tau_i|X_i = x]$$

In weeks 2 & 3, CATE_x was a means to an end: if treatment is as-if random conditional on X, then we estimate CATE_x for each $X = x$ (e.g. by sub-classification) and average to get the ATE, ATT, ATC.

What do we want to know? (2)

Today we're talking about measuring and comparing different CATE_x as a goal in itself.

e.g. Who do GOTV interventions affect more? In in what conditions does incumbency affect elections more?

Why?

- Sometimes because D is a policy and you want to know who it helps/hurts.
- Usually because your **theory/explanation** for the effect of D on Y implies treatment effect heterogeneity.

Plan

By example, explore four methods for studying treatment effect heterogeneity:

- Analysis in separate subgroups
- Treatment-covariate interactions in regression
- Heterogeneity two-step: estimate treatment effects for subsets, regress on subset characteristics ($\rightarrow$ hierarchical models)
- Machine-learning methods

Consider two pitfalls of studying treatment effect heterogeneity:

- **Multiple comparisons problem**: risk of concluding there are differences across subgroups when it's really random variation
- **Confusing your treatment and your covariates**: risk of concluding that you've measured the effect of X when really you've measured how the effect of D varies with X

Incumbency advantage: background

Party incumbency advantage, generally:

- The effect of holding office on electoral success
- How much better does a party do in constituency i at time t if it won in that constituency at time $t-1$ than if it lost?

Reminder: How would you estimate this by RDD?

Eggers & Spirling (2017): why is heterogeneity interesting?

Research question: Do voters pay more attention to candidate characteristics when partisan stakes are lower?

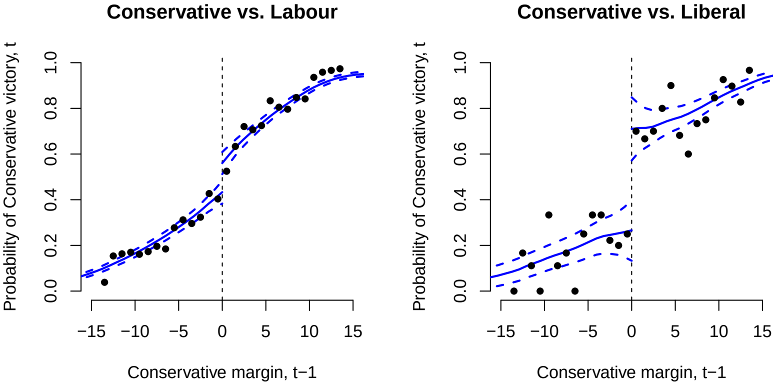
Research design: Compare partisan incumbency advantage in

- Lab-Con battlegrounds (where partisan stakes are higher)
- Lib Dem-Con battlegrounds (where partisan stakes are lower)

Eggers & Spirling (2017): basic analysis

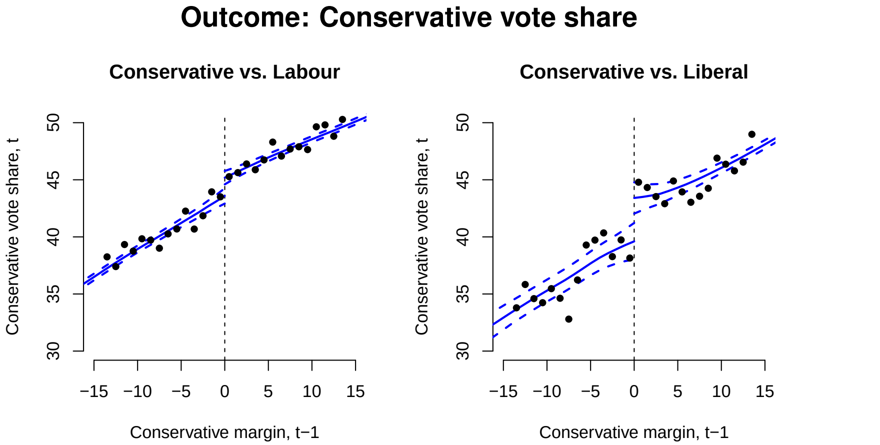
- Split sample based on identity of top two parties (Con & Lab, Con & Lib Dem)
- Estimate incumbency advantage via RDD in each subgroup

Outcome: Pr(Conservative victory)



Eggers & Spirling (2017): basic analysis

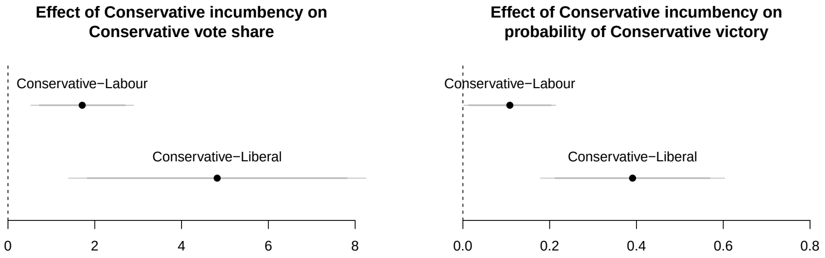
- Split sample based on identity of top two parties (Con & Lab, Con & Lib Dem)
- Estimate incumbency advantage via RDD in each subgroup



Hypothesis testing

Suppose we have estimates from separate subgroups, as below.

How do we know if the difference we find is **statistically significant**?



Treatment-covariate interactions: basics

If D_i is randomly assigned, then the following should yield the same estimate of $\hat{\delta} \equiv E[\tau_i | X_i = 1] - E[\tau_i | X_i = 0]$:

Diff-in-DIGMs:

$$\left\{E[Y_i|D_i = 1, X_i = 1] - E[Y_i|D_i = 0, X_i = 1]\right\} - \left\{E[Y_i|D_i = 1, X_i = 0] - E[Y_i|D_i = 0, X_i = 0]\right\}$$

Interaction in a regression:

$$E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i \times X_i$$

Application to Eggers & Spirling (2017)

Denote by RV_{it} the margin between the Conservatives and the leading non-Conservative candidate in constituency i at time t.

Regression to estimate partisan incumbency effect in the whole sample:

$$E[\text{ConWin}_{it}] = \beta_0 + \beta_1 \text{ConWin}_{i,t-1} + \beta_2 \text{RV}_{i,t-1} + \beta_3 \text{ConWin}_{i,t-1} \times \text{RV}_{i,t-1} + \epsilon_{it}$$

(Restricting to cases where $|\text{RV}_{it}| < h$ and weighting by $h - |\text{RV}_{it}|$.)

- Which coefficient is the incumbency effect?
- What would a plot of the predicted outcome as a function of $\text{ConWin}_{i,t-1}$ look like?

Application to Eggers & Spirling (2017) (3)

How could we extend this regression equation

$$E[\text{ConWin}_{it}] = \beta_0 + \beta_1 \text{ConWin}_{i,t-1} + \beta_2 \text{RV}_{i,t-1} + \beta_3 \text{ConWin}_{i,t-1} \times \text{RV}_{i,t-1} + \epsilon_{it}$$

to measure different treatment effects for $X_{i,t-1} = 1$ (Con-Lab constituencies) vs. $X_{i,t-1} = 0$ (Con-LD constituencies)?

Application to Eggers & Spirling (2017) (4)

Option 1. Assume same relationship between $\text{RV}_{i,t-1}$ and outcome for $X = 1$ and $X = 0$:

$$E[\text{ConWin}_{it}] = \beta_0 + \beta_1 \text{ConWin}_{i,t-1} + \beta_2 \text{RV}_{i,t-1} + \beta_3 \text{ConWin}_{i,t-1} \times \text{RV}_{i,t-1} + \\ + \beta_4 \text{X}_{i,t-1} + \beta_5 \text{ConWin}_{i,t-1} \times \text{X}_{i,t-1} + \epsilon_{it}$$

Application to Eggers & Spirling (2017) (5)

Option 2. Allow relationship between $\text{RV}_{i,t-1}$ and outcome to vary with X:

$$\begin{aligned} E[\text{ConWin}_{it}] &= \beta_0 + \beta_1 \text{ConWin}_{i,t-1} + \beta_2 \text{RV}_{i,t-1} + \beta_3 \text{ConWin}_{i,t-1} \times \text{RV}_{i,t-1} + \\ &\quad + \beta_4 \text{X}_{i,t-1} + \beta_5 \text{ConWin}_{i,t-1} \times \text{X}_{i,t-1} + \\ &\quad + \beta_6 \text{RV}_{i,t-1} \times \text{X}_{i,t-1} + \beta_6 \text{ConWin}_{i,t-1} \times \text{RV}_{i,t-1} \times \text{X}_{i,t-1} + \epsilon_{it} \end{aligned}$$

Application to Eggers & Spirling (2017) (6)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Outcome: Conservative vote share at t (N = 4,030, CCT BW = 16.34)							
Conservative win at $t-1$	1.911***	1.644***	1.642***	.603	1.788***	1.591***	1.899***	1.938***
	(.489)	(.408)	(.464)	(.811)	(.446)	(.410)	(.450)	(.490)
Conservative win at $t-1$ × Liberal opponent		1.782	1.687†	2.013*	1.891	1.988†	2.078†	1.985*
		(1.181)	(1.006)	(.849)	(1.191)	(1.168)	(1.174)	(.993)
	Outcome: Conservative win at t (N = 2,423, CCT BW = 9.99)							
Conservative win at $t-1$	.170***	.133***	.290***	.014	.131***	.137***	.127***	.293***
	(.044)	(.034)	(.039)	(.074)	(.037)	(.034)	(.038)	(.043)
Conservative win at $t-1$ × Liberal opponent		.308**	.384***	.384***	.308**	.314**	.309**	.388***
		(.100)	(.088)	(.079)	(.101)	(.100)	(.100)	(.088)
Covariates (and their interaction with indicator for Conservative win at $t-1$) included:								
Margin at t (running var.)	✓	✓	✓	✓	✓	✓	✓	✓
Decade dummies			✓					✓
Year dummies				✓				
Borough (v. county)					✓		✓	✓
Country						✓	✓	✓
Country × borough							✓	✓

When is winning office in India financially rewarding?

Indian candidates must file financial disclosure forms.

Fisman, Schulz, and Vig's question: do winners accumulate wealth faster than losers?

Going further: is winning office more beneficial in more corrupt states?

Fisman, Schulz, and Vig regression table

TABLE 5
WINNER PREMIUM AND STATE-LEVEL CORRUPTION

	Log(Final Net Assets)				
	BIMARU	Non-BIMARU			
VARIABLE	(1)	(2)	(3)	(4)	(5)
Winner	.257***	.122**	.121**	.104*	.188***
	(.026)	(.051)	(.051)	(.054)	(.045)
Log(Initial Net Assets)	.681***	.743***	.721***	.720***	.718***
	(.022)	(.040)	(.029)	(.030)	(.031)
Winner × BIMARU			.136**		
Winner × BIMAROU				.156***	
				(.059)	
Winner × TI Corruption					.063**
					(.027)
Constant	5.697***	4.672***	5.033***	5.051***	5.080***
	(.324)	(.612)	(.450)	(.454)	(.471)
Observations	386	754	1,140	1,140	998
R^2	.842	.83	.833	.834	.833

Who is affected by GOTV mobilization?

Since Gerber & Green (2000), dozens of GOTV experiments (many with available replication data).

Enos, Fowler, and Vavreck (EFV)'s question: do these interventions remedy existing inequalities in participation ("participation gap"), or exacerbate them?

Who is affected by GOTV mobilization? (2)

Could ask about specific covariates available in several datasets, e.g. gender, age, partisanship.

Enos, Fowler, and Vavreck (EFV)'s clever approach: Denote $Y_i \in \{0, 1\}$ whether unit *i* voted or not. Then:

- Using only units in the control group and *whatever covariates are available in the data*, fit a model like

$$E[Y_i] = \alpha_0 + \alpha_1 \text{Age}_i + \alpha_2 \text{Gender}_i + \dots$$

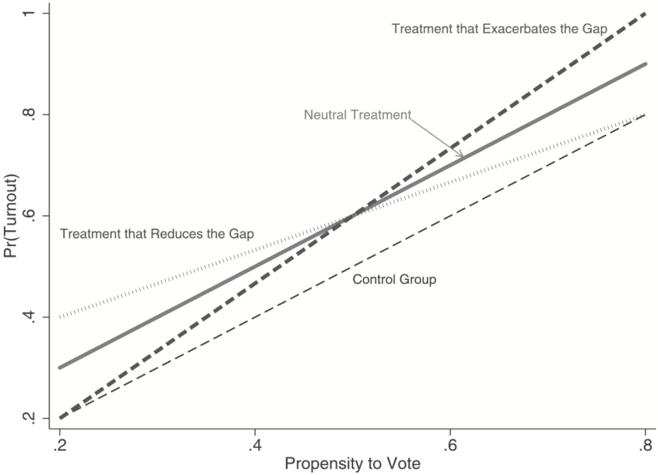
- Use the above model to estimate a propensity score $\hat{p}(X_i)$ for **all** units in the dataset
- Regress turnout decision on treatment interacted with $\hat{p}(X_i)$:

$$E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 \hat{p}(X_i) + \beta_3 D_i \times \hat{p}(X_i) + \epsilon_i$$

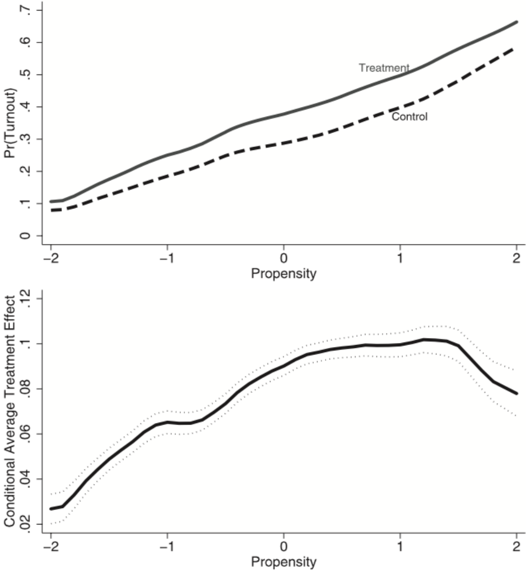
How do we interpret β_3 in the above regression?

Who is affected by GOTV mobilization? (3)

$$E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 \hat{p}(X_i) + \beta_3 D_i \times \hat{p}(X_i) + \epsilon_i$$



EFV: results from one experiment



EFV: results from 24 experiments

Intervention	Treatment	Treatment*Propensity	N-Treated	N-Control	Study
Mail	.016 (.008)*	.002 (.007)	7,679	11,665	GG00
Door	.040 (.011)**	-.006 (.009)	2,877	11,665	–
Mail+Door	.037 (.012)**	.004 (.010)	1,853	11,665	–
Phone+Mail+Door	.031 (.015)*	.026 (.013)*	1,207	11,665	–
Bridgeport	.049 (.020)*	.052 (.025)*	895	911	GGN03
Detroit	.027 (.009)**	-.020 (.006)**	2,472	2,482	–
Minneapolis	.027 (.013)*	.027 (.010)**	1,409	1,418	–
St. Paul	.035 (.016)*	-.015 (.011)	1,104	1,104	–
Stonybrook	.071 (.031)*	-.014 (.031)	680	279	N06
Volunteer	.008 (.004)*	-.004 (.004)	26,565	27,221	N07
Professional	.016 (.004)**	.001 (.004)	27,496	27,221	–
Prof.+Vol.	.015 (.004)**	-.003 (.004)	27,452	27,221	–
Civic Duty	.018 (.003)**	.002 (.003)	38,218	191,243	GGL08
Hawthorne	.025 (.003)**	.008 (.003)**	38,204	191,243	–
Self	.048 (.003)**	.008 (.003)**	38,218	191,243	–
Neighbors	.080 (.003)**	.016 (.003)**	38,201	191,243	–
MoveOn	.016 (.004)**	-.010 (.005)*	23,384	22,893	MG08
Minneapolis	.038 (.016)*	.051 (.017)**	876	1,748	N08
Text Message	.030 (.010)**	-.017 (.010)	4,007	4,046	DS09
Civic Duty	.017 (.008)*	.017 (.009)	3,238	353,341	GGL10
Shame	.064 (.006)**	.029 (.007)**	6,325	353,341	–
Pride	.041 (.006)**	.005 (.006)	6,307	353,341	–
Party Reg.	.034 (.008)**	.011 (.010)	1,173	1,175	GHW10
Planning	.010 (.004)**	.000 (.003)	19,411	228,995	NR10
Pooled	**.033 (.001)****	**.005 (.001)****	**319,251**	**848,521**	–

Heterogeneity two-step: basic idea

So far, looking at how treatment effect depends on features of the **unit/individual**.
Now: do treatment effects depend on features of the **setting**?

Step one: measure the treatment effect in M different settings.

Step two: regress the M treatment effects on **setting characteristics**.

Heterogeneity two-step: more formally

Step one: using some method (RCT, regression, IV), measure ATE in each setting m: $\hat{\tau}_m$.

Step two: let $X_{1m}, X_{2m}, \ldots, X_{Km}$ denote covariates describing setting m. Estimate

$$E[\hat{\tau}_m] = \beta_0 + \beta_1 X_{1m} + \beta_2 X_{2m} + \ldots + \beta_K X_{Km} + \epsilon_m$$

to **describe** how $\hat{\tau}_m$ varies with covariates.

King (1991): testing the “constituency service” hypothesis

Question: Does constituency service explain the incumbency advantage?
Approach: Use the heterogeneity two-step:

- Estimate incumbency advantage in state legislature for each U.S. state and year
- Regress state-year incumbency advantage on state-year legislative operating budget and controls

Step 1: State-year estimates of incumbency advantage

TABLE 1 *Incumbency Advantage in State Legislative Elections (with Standard Errors in Parentheses)*

	1970	1974	1976	1978	1980	1984	1986
California	0.100		0.049	0.092	0.087		0.080
	(0.033)		(0.022)	(0.027)	(0.031		(0.021)
Colorado	0.048		0.043	0.017	0.089	0.007	−0.015
	(0.029		(0.021)	(0.034)	(0.026)	(0.038)	(0.034)
Connecticut	0.003	0.034	−0.015	0.048	0.051	0.030	0.079
	(0.009)	(0.013)	(0.014)	(0.015)	(0.015)	(0.014)	(0.018)
Delaware	0.083	0.019	0.073	0.009	0.101	0.155	0.116
	(0.025)	(0.037)	(0.041)	(0.053)	(0.032)	(0.043)	(0.083)
Iowa	0.029	0.039	0.038	0.074	0.051	0.044	0.103
	(0.026)	(0.016)	(0.015)	(0.017)	(0.026)	(0.019)	(0.022)
Michigan	0.032	0.052	0.021	0.038	0.038	0.025	0.093
	(0.013)	(0.018)	(0.024)	(0.020)	(0.025)	(0.027)	(0.023)
Missouri	−0.006	0.006	0.060	−0.102	0.108	0.084	−0.000
	(0.016)	(0.016)	(0.020)	(0.073)	(0.026)	(0.033)	(0.032)
New York	0.042	0.073	0.079	0.120	0.126	0.044	0.119
	(0.018)	(0.021)	(0.020)	(0.023)	(0.024)	(0.023)	(0.026)
Ohio	0.027	0.059	0.050	0.064	0.057	0.052	0.014
	(0.020)	(0.022)	(0.017)	(0.024)	(0.025)	(0.021)	(0.017)
Pennsylvania	−0.022	0.025	0.005	−0.009	0.071	0.146	0.091
	(0.013)	(0.013)	(0.013)	(0.016)	(0.019)	(0.024)	(0.019)
Rhode Island	0.006	0.033	0.011	0.039	0.050	0.094	0.054
	(0.018)	(0.041)	(0.024)	(0.036)	(0.029)	(0.030)	(0.037)
Utah	0.014	0.024	0.024	0.012	0.005	0.038	−0.020
	(0.016)	(0.036)	(0.022)	(0.031)	(0.021)	(0.024)	(0.024)
Wisconsin	0.013	0.051	0.013	0.077	0.044	0.002	0.045
	(0.020)	(0.038)	(0.038)	(0.030)	(0.033)	(0.050)	(0.026)

Step 2: incumbency advantage as dependent variable

TABLE 2 *Weighted Least Squares Regressions of Incumbency Advantage*

Variables	*b*	s.e.	*b*	s.e.
Constant	−0.1157	0.0508	−0.2513	0.0989
Budget	0.0154	0.0040	0.0230	0.0068
Salary	0.0045	0.0039	0.0098	0.0067
Colorado	0.1288	0.0464	0.2427	0.0830
Conecticut	0.1217	0.0462	0.2430	0.0848
Delaware	0.1777	0.0499	0.2962	0.0917
Iowa	0.1427	0.0447	0.2674	0.0842
Michigan	0.0698	0.0341	0.1281	0.0511
Missouri	0.1099	0.0436	0.2397	0.0780
New York	0.0381	0.0225	0.0967	0.0311
Ohio	0.1098	0.0403	0.2032	0.0715
Pennsylvania	0.0424	0.0296	0.0944	0.0452
Rhode Island	0.1409	0.0518	0.2772	0.0998
Utah	0.1232	0.0510	0.2472	0.0976
Wisconsin	0.1012	0.0419	0.2021	0.0709
Lag(IncAd)			−0.0694	0.1389
n	88		52	

Kasara and Suryanaranan (2014): explaining turnout inequality

Observation: Although political scientists assume the rich vote more than the poor, in many places (e.g. India) this is reversed.

Question: When do the rich vote less than the poor and why?

Hypothesis: Political involvement of the rich depends on their *potential tax exposure*.

Approach: Use the heterogeneity two-step:

- Estimate relationship between income and turnout using surveys (e.g. Latinobarometro, Afrobarometer) in 76 countries
- Regress that relationship on (1) political salience of redistribution in party politics and (2) state's bureaucratic capacity.

Step 1: country-survey estimate of turnout inequality

Let

- Y_i indicate whether respondent i voted,
- IncomeQuintile_i denote respondent i's income quintile (1-5),
- $\mathbf{Z}_i$ denote control variables for respondent i (age, education, gender, residence in urban area)

Turnout inequality in a given country-survey is measured by β in regression*

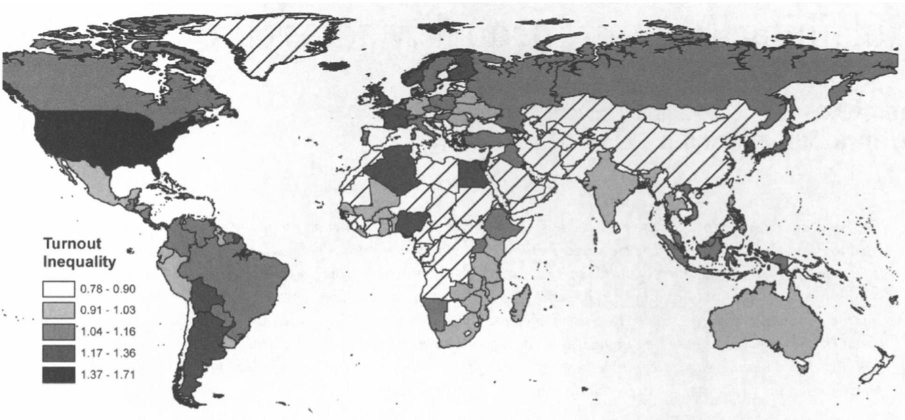
$$E[Y_i] = \alpha + \beta \text{IncomeQuintile}_i + \mathbf{Z}_i\theta + \epsilon_i$$

Denote by $\hat{\beta}_j$ the estimated turnout inequality in country-survey j.

*Actually they run a logistical regression i.e. logit.

Step 1: estimates of turnout inequality

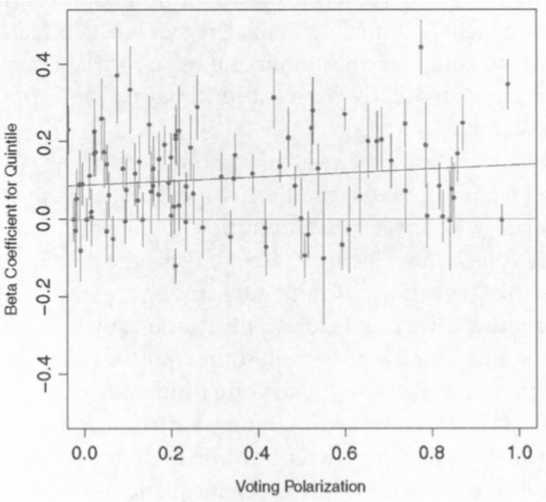
FIGURE 1 Turnout Inequality across the World



Notes: Ratio of turnout among the top quintile to turnout among the bottom quintile on a wealth index. Data are missing for countries with a cross-hatch. The construction of the wealth index is described in the main text.

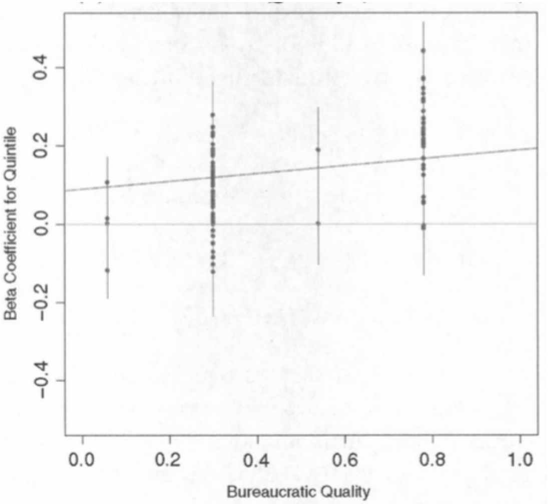
Step 2: turnout inequality and voting polarization

Voting polarization: a measure of how well income quintile predicts vote choice.



Step 2: turnout inequality and bureaucratic quality

Bureaucratic quality: measure by Political Risk Services Group assessing independence and effectiveness of the bureaucracy.



DV: $\hat{\beta}_j$	[1]	[2]	[3]	[4]	[5]	[6]
Voting Polarization	0.045**					
	(0.022)					
Electoral Distance Q1 and Q5		0.046**				
		(0.019)				
Bureaucratic Quality			0.065**			
			(0.029)			
Government Effectiveness				0.084***		
				(0.027)		
Direct Taxes/Revenue					0.053*	
					(0.031)	
Log. GDP per capita						0.13***
						(0.030)
PR	0.0033	0.0062	0.023	0.0056	0.043	0.0066
	(0.035)	(0.035)	(0.040)	(0.037)	(0.032)	(0.034)
Concurrent Elections	0.069***	0.073***	0.044	0.061***	0.090***	0.058**
	(0.025)	(0.025)	(0.027)	(0.027)	(0.032)	(0.027)
Compulsory Voting	−0.050	−0.052	−0.032	−0.051	0.0059	−0.058
	(0.040)	(0.038)	(0.031)	(0.037)	(0.055)	(0.037)
Polity	0.039	0.045	0.023	0.00049	0.065**	0.014
	(0.030)	(0.030)	(0.029)	(0.032)	(0.029)	(0.031)
Infant Mortality	−0.024	−0.025	0.036	0.015	−0.035	0.068***
	(0.024)	(0.023)	(0.032)	(0.026)	(0.025)	(0.024)
Gini (Gross)	−0.022	−0.032	−0.029	−0.027	−0.093***	−0.043
	(0.030)	(0.031)	(0.031)	(0.031)	(0.033)	(0.033)
Homicide Rate	−0.019	−0.016	−0.0053	0.0068	0.034	0.0056
	(0.029)	(0.028)	(0.030)	(0.029)	(0.031)	(0.030)
Ethnic Fractionalization	−0.011	−0.0066	−0.034	−0.029	−0.0099	−0.022
	(0.028)	(0.029)	(0.033)	(0.029)	(0.035)	(0.028)
Intercept	0.049	0.043	0.045	0.047	0.0041	0.050
	(0.037)	(0.038)	(0.039)	(0.037)	(0.030)	(0.035)
N	169	167	183	178	102	187
Countries	60	60	62	63	45	64

Towards hierarchical models

The **two-step** allows us to model treatment effects measured in each setting as a function of features of the settings.

To make more efficient and flexible: a hierarchical/multilevel approach, such as (roughly)

$$E[Y_i] = \alpha_{g[i]} + \beta X_i + \tau_i D_i + \epsilon_i$$
$$\tau_i = \gamma V_{g[i]} + \omega Z_i + \psi V_{g[i]} \times Z_i + \eta_i$$

Key features:

- Treatment effect assumed to vary across individuals
- Treatment effect assumed to depend on features of setting, individual, and interaction between them.

The feature selection problem

You have data from a randomized experiment:

- outcome
- treatment
- covariates

Your questions: How much do treatment effects vary with covariates? What treatment combination is most effective? Who is most affected by the treatment?

The problem: Many treatment combinations and subgroups that could be compared.

A solution: Use *feature selection* techniques developed in machine learning for e.g. genetics, speech/image recognition.

Some recent work on machine learning & causal inference

- Green and Kern (2012) use Bayesian Additive Regression Trees (BART) as a predictive model, then estimate CATEs by simulation
- Imai and Ratkovic (2013) use LASSO-like techniques to identify
  - most effective GOTV intervention
  - most affected subgroups (in job training program)
- Grimmer, Messing, Westwood (2017) use “ensemble methods”

Thought experiment

You do a randomized experiment.

Suppose that

- the treatment D_i has no effect for any subject
- you have 20 subgroups indicated by subgroup dummies $X_i \dots X_{20}$

You run the regression

$$
\begin{aligned}
E[Y_i] \quad = \quad & \tau D_i + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{20} X_{20} + \\
& \gamma_1 X_1 \times D_i + \gamma_2 X_2 \times D_i + \ldots + \beta_{20} X_{20} \times D_i
\end{aligned}
$$

Questions:

- What should the coefficients be if you ran this regression in the population, i.e. with infinite data?
- In a sample, what is the probability of finding at least one interaction significant at the .05 (95%) level?

Thought experiment (2)

The significance level (here, .05) is the probability of a false positive **for a single coefficient**.

The probability of not getting any false positives, given k independent attempts, is $(1 - .05)^k$.

Thus the probability of getting at least one false positive from 20 interaction coefficients is

$$1 - (1 - .05)^2 0 = .642$$

This is the **multiple comparisons** problem.

Detecting the multiple comparisons problem

- Someone tests 12 different interactions and focuses on one barely significant result
- Someone runs a regression with 35 explanatory variables and focuses on one barely significant result
- Someone shows insignificant average effects but then focuses on a subgroup without strong theoretical justification

Correcting the multiple comparisons problem

- **Bonferroni correction**: Given k tests, reject null at α level if p-value is below α/k.
- Many other less-conservative (but less straightforward) methods: see Juliet Schaffer, “Multiple hypothesis testing” (1995)

Remember what is causally "identified"

If D_i is randomly assigned in an experiment and you run this regression

$$E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i \times X_i$$

you cannot interpret β_2 as the effect of X_i, or β_3 as the difference in the effect of X_i for treated and control units.

Example: Eggers & Spirling (2017)

Eggers & Spirling (2017) show that incumbency effect is larger in Con-LD seats than Con-Lab seats.

This could be because of differences in partisanship, but what else may vary with this X_i?

How could we address alternative explanations?

Eggers & Spirling (2017) regression table

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Outcome: Conservative vote share at t (N = 4,030, CCT BW = 16.34)							
Conservative win at $t-1$	1.911***	1.644***	1.642***	.603	1.788***	1.591***	1.899***	1.938***
	(.489)	(.408)	(.464)	(.811)	(.446)	(.410)	(.450)	(.490)
Conservative win at $t-1$ × Liberal opponent		1.782	1.687†	2.013*	1.891	1.988†	2.078†	1.985*
		(1.181)	(1.006)	(.849)	(1.191)	(1.168)	(1.174)	(.993)
	Outcome: Conservative win at t (N = 2,423, CCT BW = 9.99)							
Conservative win at $t-1$	.170***	.133***	.290***	.014	.131***	.137***	.127***	.293***
	(.044)	(.034)	(.039)	(.074)	(.037)	(.034)	(.038)	(.043)
Conservative win at $t-1$ × Liberal opponent		.308**	.384***	.384***	.308**	.314**	.309**	.388***
		(.100)	(.088)	(.079)	(.101)	(.100)	(.100)	(.088)
Covariates (and their interaction with indicator for Conservative win at $t-1$) included:								
Margin at t (running var.)	✓	✓	✓	✓	✓	✓	✓	✓
Decade dummies			✓					✓
Year dummies				✓				
Borough (v. county)					✓		✓	✓
Country						✓	✓	✓
Country × borough							✓	✓

More generally

Suppose given random D_i you run

$$E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i \times X_i$$

but someone says, "X_i and Z_i are related, so maybe Z_i is the real reason why β_3 is not zero."
Then you can run

$$E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i \times X_i + \beta_4 Z_i + \beta_5 D_i \times Z_i.$$

Note that we need no covariates to estimate effect of D_i but potentially many covariates to convince anyone that X_i causes heterogeneity in effect of D_i.

Treatment effect heterogeneity: concluding thoughts

- All datasets can be **described** (by e.g. DIGM, regression), but most don't allow any credible estimates of (causal) effects
- Variation in treatment effects can be **described** (by e.g. analysis in separate subgroups, two-step), but *explaining* this variation is tricky.
- CATEs are measures like GDP, corruption, etc; explaining CATEs is like explaining GDP, corruption, etc.

This course: concluding thoughts

What else is there?

- More causal inference: better regressions for control; sensitivity analysis; synthetic control methods and their generalizations
- Descriptive techniques: measurement models for text, votes, etc; machine learning for prediction, classification, etc; hierarchical models
- Statistical inference: randomization inference, bootstrap; Bayesian methods; more on cluster- and heteroskedasticity-robust standard errors