Matching & Regression: Accounting for Rival Explanations

Department of Government
London School of Economics and Political Science

What do you think will be on the exam?

Sample Paper

`https://moodle.lse.ac.uk/mod/resource/view.php?id=534210`

- ST Exam is 50% of overall mark
  - 25%: 5 "shorter answer" questions (of 15); worth 5% of total mark each
  - 25%: 1 essay (of 4)
- Research Design Proposal is 50% of overall mark

Uses of Regression

1. Description
2. Prediction
3. Causal Inference

Descriptive Inference

1. We want to understand a *population* of cases
2. We cannot observe them all, so:
   1. Draw a *representative* sample
   2. Perform mathematical procedures on sample data
   3. Use assumptions to make inferences about population
   4. Express uncertainty about those inferences based on assumptions

Parameter Estimation

- We want to observe population *parameter* β
- If we obtain a representative sample of population units:
  - Our sample statistic $\hat{\beta}$ is an unbiased estimate of β
  - Our sampling procedure dictates how uncertain we are about the value of β

Three Equations

1. Population:
   $Y = \beta_0 + \beta_1 X \;\; (+\epsilon)$
2. Sample estimate:
   $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
3. Unit:
   $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
   $y_i = \bar{y}_{0i} + (y_{1i} - y_{0i})x_i + (y_{0i} - \bar{y}_{0i})$

Mathematically, regression. . .

- . . . describes multivariate relationships in a sample of data points

Mathematically, regression. . .

- . . . describes multivariate relationships in a sample of data points
- . . . depending on sampling procedure, estimates those relationships in the population

Mathematically, regression. . .

- . . . describes multivariate relationships in a sample of data points
- . . . depending on sampling procedure, estimates those relationships in the population
- . . . depending on model fit, provides a way to predict outcome values for new cases

Mathematically, regression...

- ...describes multivariate relationships in a sample of data points
- ...depending on sampling procedure, estimates those relationships in the population
- ...depending on model fit, provides a way to predict outcome values for new cases
- ...depending on model completeness, provides inferences about the effect of X on Y

Causal inference is about comparing an observed outcome to a counterfactual, "potential outcome" for the same cases

Regression provides a "statistical solution" to the fundamental problem of causal inference (Holland)

An Example

- For example, if we think smoking might cause lung cancer, how would we know?
- How would we know if smoking caused lung cancer for an individual who smoked?
    - What's the relevant counterfactual?
- How would we know if smoking causes lung cancer on average across many individuals?
    - What's the relevant counterfactual?

Confounding

- A source of "endogeneity"
- Synonyms: selection bias, omitted variable bias
- In lay terms: the (non)correlation between X and Y does not reflect a causal relationship between X and Y are related for other reasons
    - Most commonly: Some Z causes both X and Y

Addressing Confounding

Addressing Confounding

1. Correlate a “putative” cause (X) and an outcome (Y)

Addressing Confounding

1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)

Addressing Confounding

1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)
3. "Condition" on all confounds
   - Calculate correlation between X and Y at each combination of levels of $\mathbf{Z}$

Mill's Method of Difference

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or an necessary part of the cause, of the phenomenon.

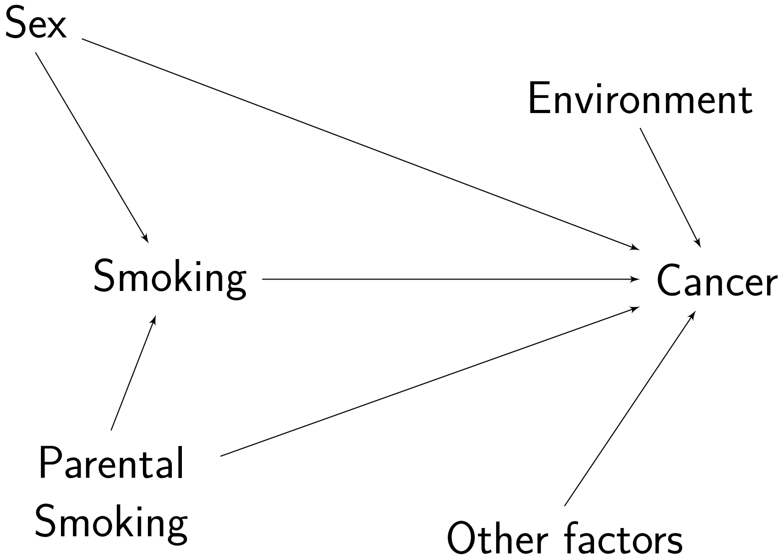
Smoking Example

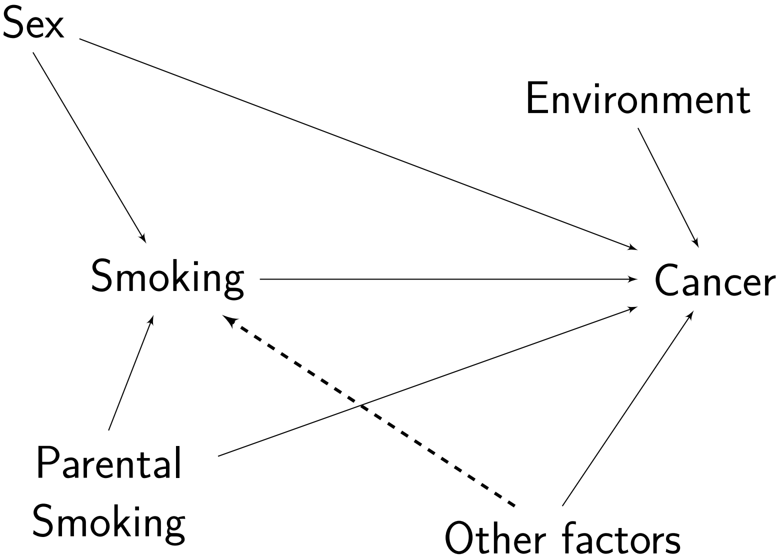
Smoking Example

1. Partition sample into "smokers" ($X = 1$) and "non-smokers" ($X = 0$)

Smoking Example

1. Partition sample into "smokers" ($X = 1$) and "non-smokers" ($X = 0$)
2. Identify possible confounds
   - Sex
   - Parental smoking
   - etc.


Sex
Environment
Smoking
Cancer
Parental Smoking
Other factors


Sex
Environment
Smoking
Cancer
Parental Smoking
Other factors

Smoking Example

1. Partition sample into "smokers" ($X = 1$) and "non-smokers" ($X = 0$)
2. Identify possible confounds
   - Sex
   - Parental smoking
   - etc.

Smoking Example

1. Partition sample into "smokers" ($X = 1$) and "non-smokers" ($X = 0$)
2. Identify possible confounds
   - Sex
   - Parental smoking
   - etc.
3. Estimate difference in cancer rates between smokers and non-smokers within each group of *covariates*

Example I

X	Y **(Cancer)**
Smokers	0.15
Non-smokers	0.05

$$\begin{aligned} ATE &= \bar{Y}_{X=1} - \bar{Y}_{X=0} \\ &= 0.15 - 0.05 \\ &= 0.10 \end{aligned}$$

Example II

Z_1 (Sex)	X	Y (Cancer)
0	Smokers	...
0	Non-smokers	...
1	Smokers	...
1	Non-smokers	...

$$ATE = p_{\text{Male}} * (\bar{Y}_{X=1,Z_1=1} - \bar{Y}_{X=0,Z_1=1}) + p_{\text{Female}} * (\bar{Y}_{X=1,Z_1=0} - \bar{Y}_{X=0,Z_1=0})$$

Example III

Z_2 (Parent)	Z_1 (Sex)	X	Y (Cancer)
0	0	Smokers	...
0	0	Non-smokers	...
0	1	Smokers	...
0	1	Non-smokers	...
1	0	Smokers	...
1	0	Non-smokers	...
1	1	Smokers	...
1	1	Non-smokers	...

$$\begin{aligned} ATE = & p_{\text{Male, Parent non-smoker}} * (\bar{Y}_{X=1,Z_1=1,Z_2=0} - \bar{Y}_{X=0,Z_1=1,Z_2=0}) + \\ & p_{\text{Female, Parent non-smoker}} * (\bar{Y}_{X=1,Z_1=0,Z_2=0} - \bar{Y}_{X=0,Z_1=0,Z_2=0}) + \\ & p_{\text{Male, Parent smoker}} * (\bar{Y}_{X=1,Z_1=1,Z_2=1} - \bar{Y}_{X=0,Z_1=1,Z_2=1}) + \\ & p_{\text{Female, Parent smoker}} * (\bar{Y}_{X=1,Z_1=0,Z_2=1} - \bar{Y}_{X=0,Z_1=0,Z_2=1}) + \end{aligned}$$

Exact Matching

- Repeat this partitioning of the space into "strata" (or "subclasses")
- Requires at least one "treated" and one "untreated" case at every combination of every covariate
- More convenient notation:

$$\text{Naive Effect} = \bar{Y}_{X=1} - \bar{Y}_{X=0}$$
$$\text{ATE} = \bar{Y}_{X=1,\mathbf{z}} - \bar{Y}_{X=0,\mathbf{z}}$$

Note that matching is just a version of Mill's method of difference used for a large number of cases.

Omitted Variables

In the language of potential outcomes:

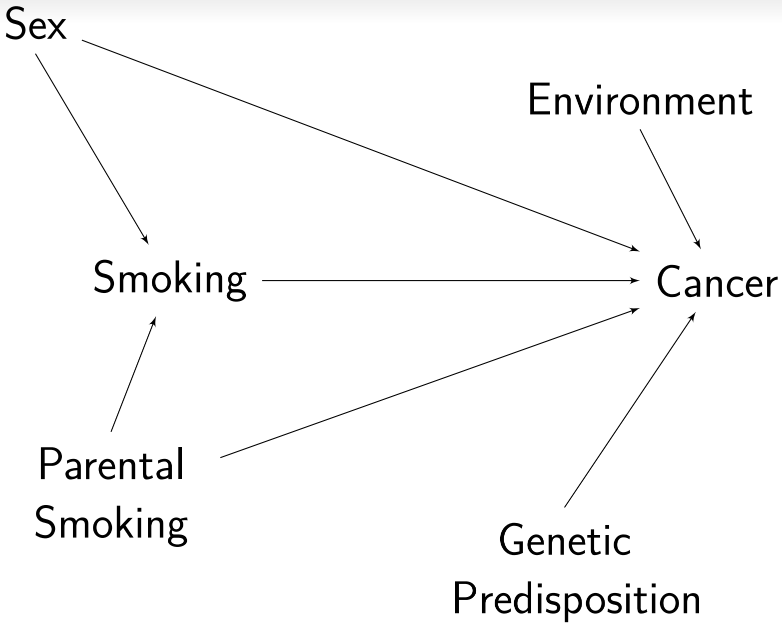
$$\underbrace{E[Y_i|X_i = 1] - E[Y_i|X_i = 0] =}_{\text{Naive Effect}}$$

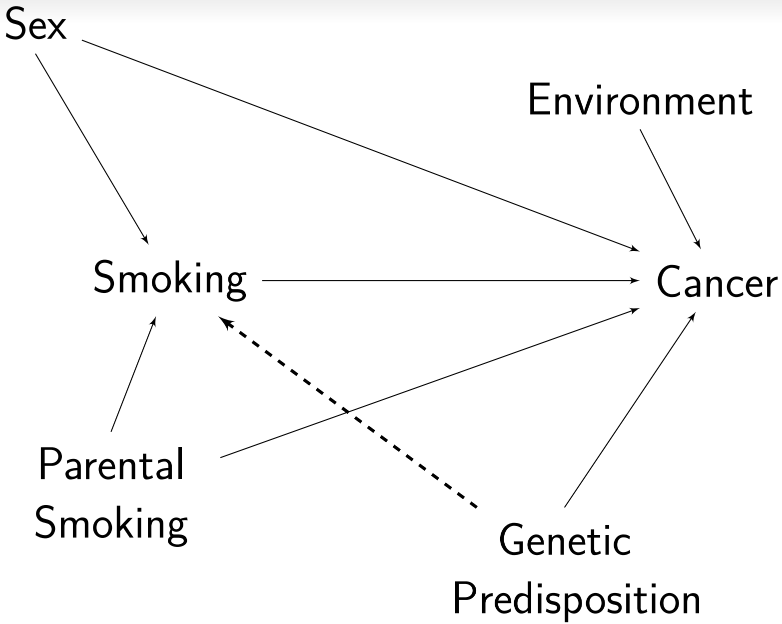
$$\underbrace{E[Y_{1i}|X_i = 1] - E[Y_{0i}|X_i = 1]}_{\text{Treatment Effect on Treated (ATT)}} + \underbrace{E[Y_{0i}|X_i = 1] - E[Y_{0i}|X_i = 0]}_{\text{Selection Bias}}$$

By conditioning, we assert that the potential (control) outcomes are equivalent between treated and non-treated cases, so the difference we observe between treatment and control outcomes is only the average causal effect of the "treatment".

Caveat!

- We can only condition on *observed* confounding variables
- If we think other confounds might exist, but are unobservable, no form of conditioning can help us
    - Example: Tobacco companies argued that an unknown genetic factor was a common cause of both smoking addiction and lung cancer


Sex
Environment
Smoking
Cancer
Parental Smoking
Genetic Predisposition


Sex
Environment
Smoking
Cancer
Parental Smoking
Genetic Predisposition

Common Conditioning Strategies

Common Conditioning Strategies

1. Condition on nothing ("naive effect")

Common Conditioning Strategies

1. Condition on nothing ("naive effect")
2. Condition on some variables

Common Conditioning Strategies

1. Condition on nothing (“naive effect”)
2. Condition on some variables
3. Condition on all observables

Common Conditioning Strategies

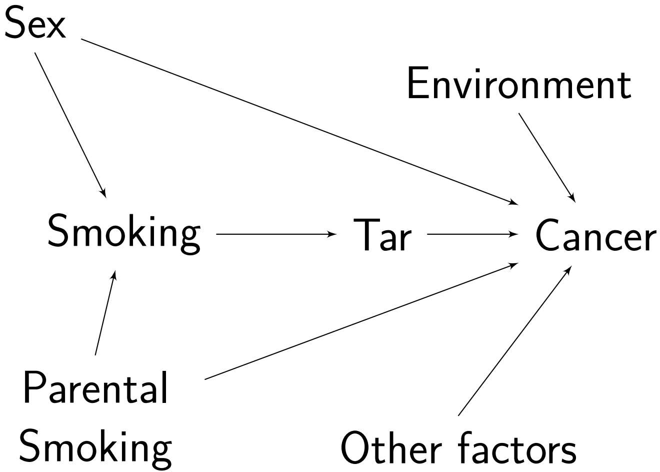
1. Condition on nothing (“naive effect”)
2. Condition on some variables
3. Condition on all observables

Which of these are good strategies?

Post-treatment Bias

- We usually want to know the **total effect** of a cause
- If we include a mediator, D, of the $X \rightarrow Y$ relationship, the coefficient on X:
  - Only reflects the **direct** effect
  - Excludes the **indirect** effect of X through D
- So don't control for mediators!

Post-Treatment Bias



Post-Treatment Bias

D (Tar)	X	Y (Cancer)
0	Smokers	...
0	Non-smokers	...
1	Smokers	...
1	Non-smokers	...

Post-Treatment Bias

D (Tar)	X	Y (Cancer)
0	Smokers	...
0	Non-smokers	...
1	Smokers	...
1	Non-smokers	...

Imagine:

$$ATE_{\text{Tar}} = (\bar{D}_{X=1} - \bar{D}_{X=0}) = 1$$
$$ATE_{\text{Cancer of Tar}} = (\bar{Y}_{D=1} - \bar{Y}_{D=0}) = 1$$

Post-Treatment Bias

D (Tar)	X	Y (Cancer)
0	Smokers	...
0	Non-smokers	...
1	Smokers	...
1	Non-smokers	...

Imagine:

$$ATE_{\text{Tar}} = (\bar{D}_{X=1} - \bar{D}_{X=0}) = 1$$
$$ATE_{\text{Cancer of Tar}} = (\bar{Y}_{D=1} - \bar{Y}_{D=0}) = 1$$

Post-Treatment Bias

D (Tar)	X	Y (Cancer)
0	Smokers	...
0	Non-smokers	...
1	Smokers	...
1	Non-smokers	...

Imagine:

$$
\begin{aligned}
ATE_{\text{Tar}} =& (\bar{D}_{X=1} - \bar{D}_{X=0}) = 1 \\
ATE_{\text{Cancer of Tar}} =& (\bar{Y}_{D=1} - \bar{Y}_{D=0}) = 1 \\
ATE_{\text{Cancer of Smoking}} =& p_{D=1}(\bar{Y}_{X=1,D=1} - \bar{Y}_{X=0,D=1}) + \\
& p_{D=0}(\bar{Y}_{X=1,D=0} - \bar{Y}_{X=0,D=0})
\end{aligned}
$$

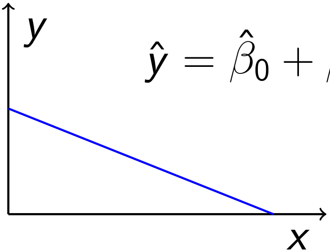
Multiple Regression

- Regression achieves the same objectives as matching
  - Estimate average causal of a variable conditional on other variables
- Requires a *linear* relationship between all RHS (X variables) and Y
  - Can be a set of binary indicator variables
- We interpret coefficient estimates as *marginal* average treatment effects

From Line to Surface I

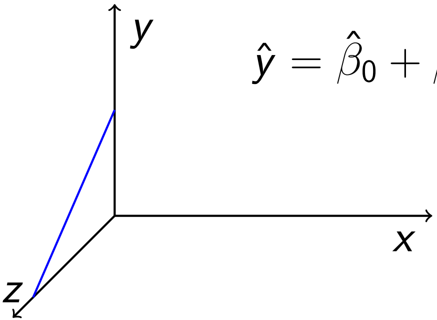
- In simple regression, we estimate a **line**
- In multiple regression, we estimate a **surface**
- Each coefficient is the *marginal effect*, all else constant (at mean)
- This can be hard to picture in your mind

From Line to Surface II



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

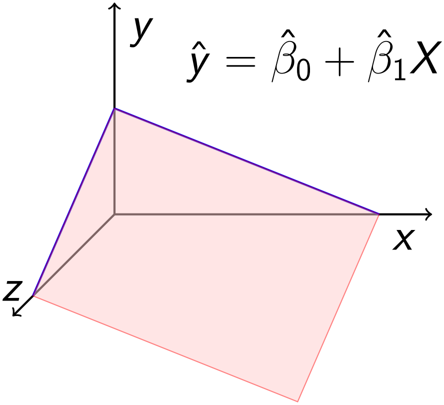
From Line to Surface II



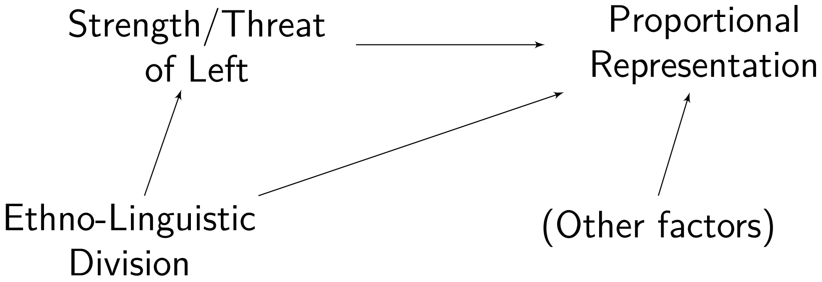
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 Z$$

From Line to Surface II

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z$$



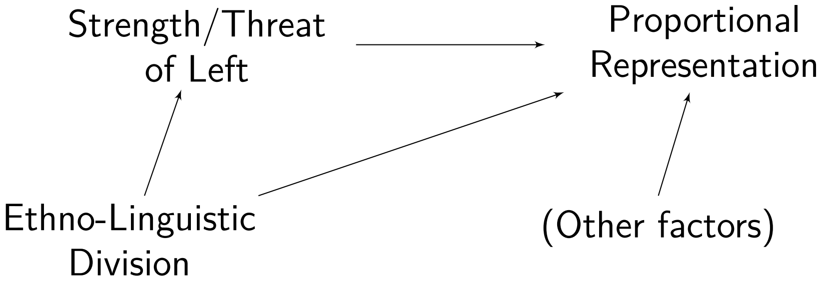
Cusack, Iversen, and Soskice



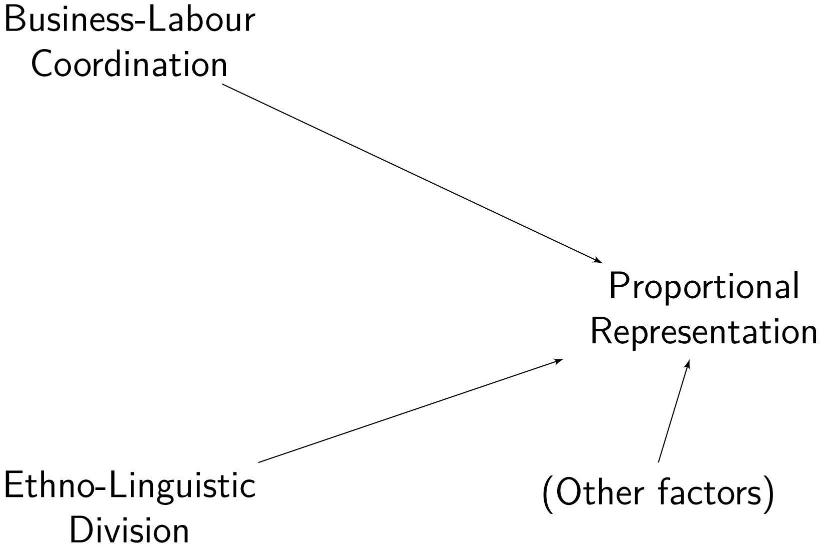
Testing Rival Hypotheses

- Rival hypotheses can be derived from two (or more) different theories
- We can conduct independent tests of each
  - Is there evidence consistent with Hyp 1?
  - Is there evidence consistent with Hyp 2?
- Regression allows us to test both simultaneously on the same data
  - Is the data more consistent with Hyp 1 or Hyp 2?
- Draw inference about causality and about validity of theories based on data

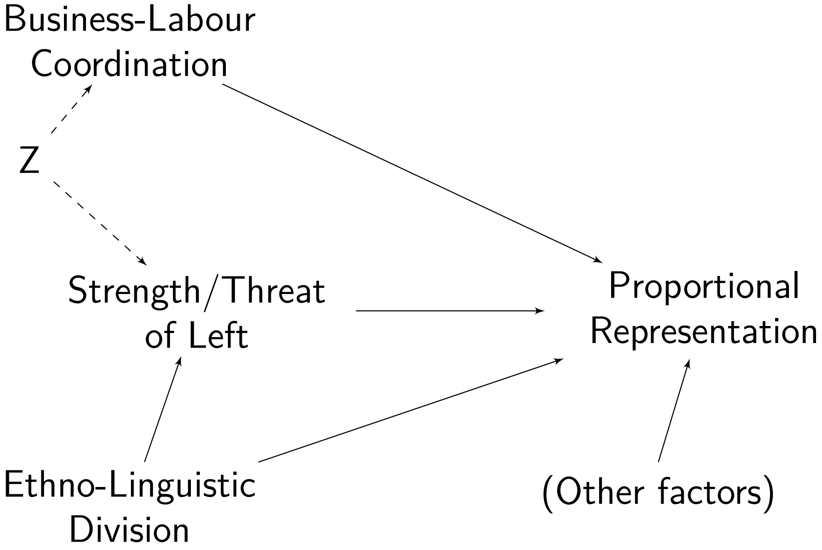
Cusack, Iversen, and Soskice



Cusack, Iversen, and Soskice



Cusack, Iversen, and Soskice



Rival Theories

- Rokkan–Boix:

$$PR = \beta_0 + \beta_1 Threat + \epsilon \tag{1}$$

TABLE 3. Replication and Re-test of Boix's Model on the Choice of Electoral Rules in the Interwar Period

Dependent Variable: Average Effective Threshold in 1919–1939	(1) Replication Using Data Reported in Boix (1999)	(2) Replication as in (1) but with 19 Cases	(3) Replication Using our Timing and	(4) Replication as in (3) but with Dominance-based Threat Score
Constant	31.30*	32.79*	29.64*	24.54*
	(4.68)	(4.93)	(5.48)	(5.82)
Threat	−.134*	−.143*	−.101	−.029
	(.049)	(.052)	(.059)	(.062)
Ethnic–linguistic division	−33.16*	−35.28*	−35.18*	−33.92
X area dummy	(14.75)	(14.74)	(16.48)	(17.84)
Adj. R-squared	.33	.37	.22	.09
SEE	10.57	10.50	11.71	12.67
Number of Obs.	22	19	19	19

* sig. at .05 level.
Note: Cols 2, 3, and 4 exclude Finland, Greece, and Luxembourg from the analysis.

Aside: Interpretation

- All our interpretation rules from earlier still apply in a multivariate regression
- Now we interpret a coefficient as an effect "all else constant"
- Generally, not good to give all coefficients a causal interpretation
  - Think "forward causal inference"
  - We're interested in the $X \rightarrow Y$ effect
  - All other coefficients are there as "controls"

Rival Theories

- Rokkan–Boix:

$$PR = \beta_0 + \beta_1 Threat + \epsilon \qquad (1)$$

Rival Theories

- Rokkan–Boix:

$$PR = \beta_0 + \beta_1 \mathit{Threat} + \epsilon \tag{1}$$

- Cusack, Iversen, and Soskice:

$$PR = \beta_0 + \beta_2 \mathit{Coordination} + \epsilon \tag{2}$$

TABLE 5. Preindustrial Coordination, Disproportionality of Representation, and Electoral System (Standard Errors in Parentheses)

	Dependent Variable: Effective Threshold					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	26.35	31.85*	31.99*	26.71*	−1.90	13.79
	(7.73)	(3.36)	(2.23)	(6.97)	(8.90)	(8.74)
Threat (dominance-based measure)	−0.06	0.02	—	−.22	−0.16	—
	(0.10)	(0.04)		(0.13)	(0.09)	
Coordination	—	−5.30*	−5.46*	—	—	—
		(0.66)	(0.63)			
Pre-1900 Disproportionality	—	—	—	—	0.34*	0.37*
					(0.09)	(0.11)
Ethnic–linguistic division X area dummy	−36.90	−7.10	—	−32.29	−28.39	—
	(20.85)	(9.63)		(22.75)	(14.65)	
Adj. R-squared	0.07	0.83	0.81	0.15	0.65	0.51
SEE	13.47	5.74	5.99	13.60	8.73	10.30
No. of observations	17	17	18	12	12	12

* Significant at .05 level.

Rival Theories

- Rokkan–Boix:

$$PR = \beta_0 + \beta_1 Threat + \epsilon \tag{1}$$

- Cusack, Iversen, and Soskice:

$$PR = \beta_0 + \beta_2 Coordination + \epsilon \tag{2}$$

Rival Theories

- Rokkan–Boix:

$$PR = \beta_0 + \beta_1 Threat + \epsilon \tag{1}$$

- Cusack, Iversen, and Soskice:

$$PR = \beta_0 + \beta_2 Coordination + \epsilon \tag{2}$$

- Combined test:

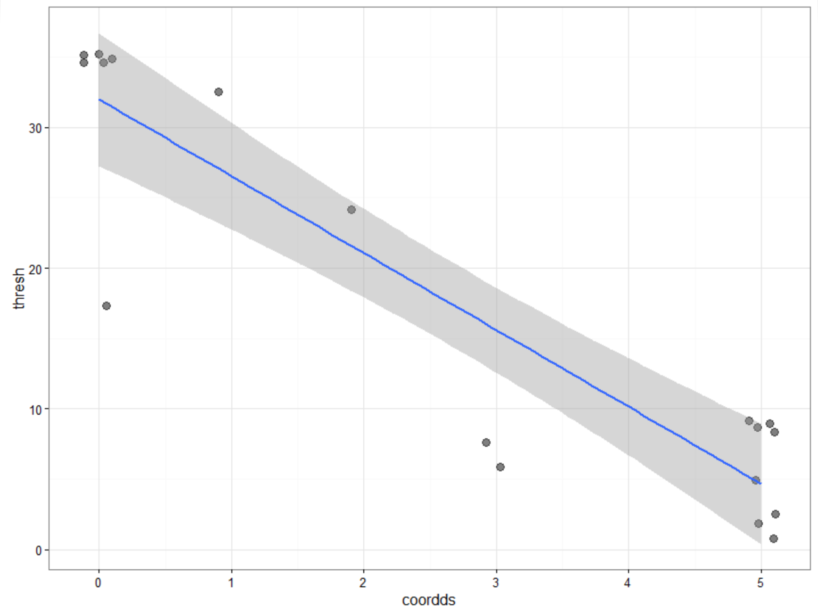
$$PR = \beta_0 + \beta_1 Threat + \beta_2 Coordination + \epsilon \tag{3}$$

TABLE 5. Preindustrial Coordination, Disproportionality of Representation, and Electoral System (Standard Errors in Parentheses)

	Dependent Variable: Effective Threshold					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	26.35	31.85*	31.99*	26.71*	−1.90	13.79
	(7.73)	(3.36)	(2.23)	(6.97)	(8.90)	(8.74)
Threat (dominance-based measure)	−0.06	0.02	—	−.22	−0.16	—
	(0.10)	(0.04)		(0.13)	(0.09)	
Coordination	—	−5.30*	−5.46*	—	—	—
		(0.66)	(0.63)			
Pre-1900 Disproportionality	—	—	—	—	0.34*	0.37*
					(0.09)	(0.11)
Ethnic–linguistic division X area dummy	−36.90	−7.10	—	−32.29	−28.39	—
	(20.85)	(9.63)		(22.75)	(14.65)	
Adj. R-squared	0.07	0.83	0.81	0.15	0.65	0.51
SEE	13.47	5.74	5.99	13.60	8.73	10.30
No. of observations	17	17	18	12	12	12

* Significant at .05 level.

	(1)	(2)
stthroct2	0.047	0.008
	(0.035)	(0.052)
coordds	-6.019^{***}	-5.284^{***}
	(0.706)	(1.008)
dispro2	0.042	0.083
	(0.052)	(0.066)
fragdum	3.624	0.123
	(8.239)	(8.911)
Constant	28.239^{***}	25.211^{***}
	(5.866)	(6.565)
Observations	13	12
R^2	0.947	0.948
Adjusted R^2	0.920	0.919
Residual Std. Error	4.217 (df = 8)	4.207 (df = 7)
F Statistic	35.673^{***} (df = 4; 8)	32.084^{***} (df = 4; 7)
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$


thresh
0
10
20
30
coordds
0
1
2
3
4
5

So the effect found by Rokkan and Boix was confounded by business–labour coordination. What was happening when they omitted the coordination variable?

Omitted Variable Bias

- We want to estimate:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

- We actually estimate:

$$\begin{aligned}\tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x + \epsilon \\ &= \tilde{\beta}_0 + \tilde{\beta}_1 x + (0 * z) + \epsilon \\ &= \tilde{\beta}_0 + \tilde{\beta}_1 x + \nu\end{aligned}$$

- Bias: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$, where $\tilde{z} = \tilde{\delta}_0 + \tilde{\delta}_1 x$

But have Cusack, Iversen, and Soskice considered all possible confounds?

TABLE 4. Indicators of Economic Structure and Organization ca. 1900

	(1) Guild Tradition and Strong Local Economies	(2) Widespread Rural Cooperatives	(3) High Employer Coordination	(4) Industry/ Centralized vs. Craft/ Fragmented Unions	(5) Large Skill-Based Export Sector	(6) Coordination Index
Australia	No	No	No	No	No	0
Canada	No	No	No	No	No	0
Ireland	No	No	No	No	No	0
New Zealand	No	No	No	No	No	0
United Kingdom	No	No	No	No	No	0
United States	No	No	No	No	No	0
France	Yes	No	No	No	No	1
Japan	Yes	No	Yes	No	No	2
Italy	Yes	Yes	Yes	No	No	3
Finland	Yes	Yes	No	No	Yes	3
Austria	Yes	Yes	Yes	Yes	Yes	5
Belgium	Yes	Yes	Yes	Yes	Yes	5
Denmark	Yes	Yes	Yes	Yes	Yes	5
Germany	Yes	Yes	Yes	Yes	Yes	5
Netherlands	Yes	Yes	Yes	Yes	Yes	5
Switzerland	Yes	Yes	Yes	Yes	Yes	5
Norway	Yes	Yes	Yes	Yes	Yes	5
Sweden	Yes	Yes	Yes	Yes	Yes	5

Sources: By column: (1) Crouch 1993; Herrigel (1996); Hechter and Brustein (1980) (2) Crouch 1993; Katzenstein 1985, ch. 4; Symes 1963; Marshall 1958; Leonardi 2006; Guinane 2001; Lewis 1978; (3)–(5) Crouch 1993; Thelen 2004; Swenson 2002; Mares 2003; Katzenstein 1985, ch. 4.
Note: Additive index in column (6) summarized across all indicators with 'Yes' = 1 and 'No' = 0.

TABLE 4. Indicators of Economic Structure and Organization ca. 1900

	(1) Guild Tradition and Strong Local Economies	(2) Widespread Rural Cooperatives	(3) High Employer Coordination	(4) Industry/ Centralized vs. Craft/ Fragmented Unions	(5) Large Skill-Based Export Sector	(6) Coordination Index
Australia	No	No	No	No	No	0
Canada	No	No	No	No	No	0
Ireland	No	No	No	No	No	0
New Zealand	No	No	No	No	No	0
United Kingdom	No	No	No	No	No	0
United States	No	No	No	No	No	0
France	Yes	No	No	No	No	1
Japan	Yes	No	Yes	No	No	2
Italy	Yes	Yes	Yes	No	No	3
Finland	Yes	Yes	No	No	Yes	3
Austria	Yes	Yes	Yes	Yes	Yes	5
Belgium	Yes	Yes	Yes	Yes	Yes	5
Denmark	Yes	Yes	Yes	Yes	Yes	5
Germany	Yes	Yes	Yes	Yes	Yes	5
Netherlands	Yes	Yes	Yes	Yes	Yes	5
Switzerland	Yes	Yes	Yes	Yes	Yes	5
Norway	Yes	Yes	Yes	Yes	Yes	5
Sweden	Yes	Yes	Yes	Yes	Yes	5

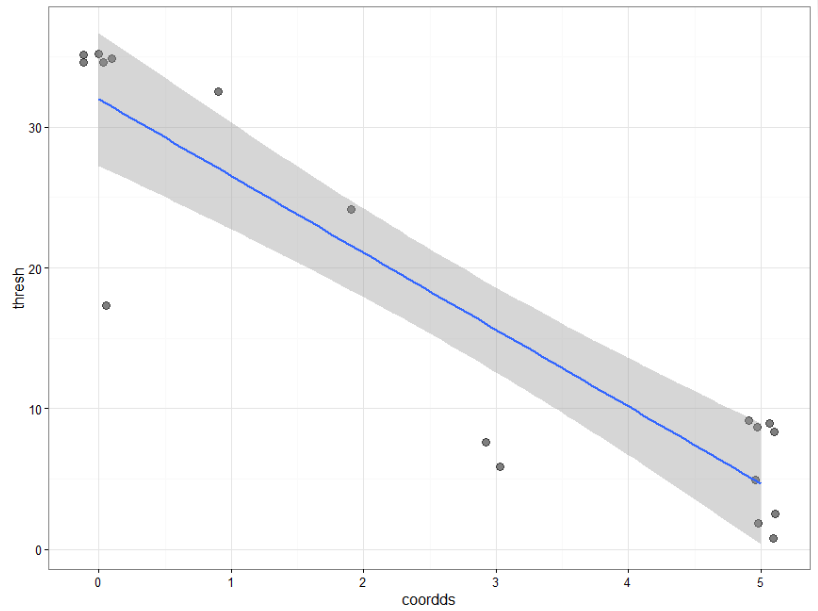
Sources: By column: (1) Crouch 1993; Herrigel (1996); Hechter and Brustein (1980) (2) Crouch 1993; Katzenstein 1985, ch. 4; Symes 1963; Marshall 1958; Leonardi 2006; Guinane 2001; Lewis 1978; (3)–(5) Crouch 1993; Thelen 2004; Swenson 2002; Mares 2003; Katzenstein 1985, ch. 4.
Note: Additive index in column (6) summarized across all indicators with 'Yes' = 1 and 'No' = 0.

	(1)	(2)
stthroct2	0.058	0.006
	(0.048)	(0.043)
coordds	-5.556^{***}	-0.398
	(1.578)	(2.467)
dispro2	0.013	-0.049
	(0.102)	(0.083)
fragdum	4.983	3.366
	(9.642)	(7.465)
brit	4.088	30.412^{*}
	(12.258)	(14.469)
Constant	26.911^{***}	9.390
	(7.388)	(9.253)
Observations	13	12
R^2	0.948	0.970
Adjusted R^2	0.910	0.945
Residual Std. Error	4.472 (df = 7)	3.449 (df = 6)
F Statistic	25.390^{***} (df = 5; 7)	39.083^{***} (df = 5; 6)
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Aside: Interpolation/Extrapolation

In *prediction*, we may want to use our estimated coefficients to predict outcome values for new cases

- *Interpolation* is prediction within the interval covered by our observed data
- *Extrapolation* is prediction outside the interval covered by our observed data



thresh
30
20
10
0
0
1
2
3
4
5
coordds

Lingering Issues

Lingering Issues

1. Inference to a population
   - Inferences from data to population depend on generalizability

Lingering Issues

1. Inference to a population
   - Inferences from data to population depend on generalizability
2. Interactions terms
   - Allow us to test whether than effect varies across values of other variables

$$
\begin{aligned}
PR &= \beta_0 + \beta_1 Threat + \beta_2 Coord + \epsilon \\
&= \beta_0 + \beta_1 Threat + \beta_2 Coord + \beta_3 (Threat * Coord) + \epsilon
\end{aligned}
$$

Lingering Issues

1. Inference to a population
   - Inferences from data to population depend on generalizability
2. Interactions terms
   - Allow us to test whether than effect varies across values of other variables

$$
\begin{aligned}
PR &= \beta_0 + \beta_1 Threat + \beta_2 Coord + \epsilon \\
&= \beta_0 + \beta_1 Threat + \beta_2 Coord + \beta_3 (Threat * Coord) + \epsilon
\end{aligned}
$$

3. Kellstedt and Whitten cover more in-depth mathematical requirements

Preview

- Next week: Experiments and Quasi-Experiments
- Following week: Research ethics
- Last week (Mar. 22):
  - Research Design Proposal Due!
  - Wrap-up course
- I am on leave for 2 weeks from April 21