Building a Robot Judge: Data Science for Decision-Making

4. Regression Discontinuity and Diff-in-Diff

Machine Learning vs Causal Inference

Machine learning:

- $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.

Machine Learning vs Causal Inference

Machine learning:

- $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
- learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function

$$y_i = h(\boldsymbol{x}; \theta)$$

that predicts y given covariates $\boldsymbol{x}$.

Machine Learning vs Causal Inference

Machine learning:

- $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
- learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function
$$y_i = h(\boldsymbol{x}; \theta)$$
that predicts y given covariates $\boldsymbol{x}$.
- if we collected more data on $\boldsymbol{x}$, we could predict the associated $\hat{y}$.

Machine Learning vs Causal Inference

Machine learning:

- $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
- learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function
$$y_i = h(\boldsymbol{x}; \theta)$$
that predicts y given covariates $\boldsymbol{x}$.
- if we collected more data on $\boldsymbol{x}$, we could predict the associated $\hat{y}$.
- but $h(\cdot)$ does not provide a *counterfactual prediction* – that is, how the outcome would change if $\boldsymbol{x}$'s were exogenously shifted.

Econometrics (applied statistical causal inference):

- y is one-dimensional, x is low-dimensional.

Machine Learning vs Causal Inference

Machine learning:

- $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
- learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function
$$y_i = h(\boldsymbol{x}; \theta)$$
that predicts y given covariates $\boldsymbol{x}$.
- if we collected more data on $\boldsymbol{x}$, we could predict the associated $\hat{y}$.
- but $h(\cdot)$ does not provide a *counterfactual prediction* – that is, how the outcome would change if $\boldsymbol{x}$'s were exogenously shifted.

Econometrics (applied statistical causal inference):

- y is one-dimensional, x is low-dimensional.
- estimate a low-dimensional **causal parameter** ρ using
$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$
where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.

Machine Learning vs Causal Inference

Machine learning:

- $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
- learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function
$$y_i = h(\boldsymbol{x};\theta)$$
that predicts y given covariates $\boldsymbol{x}$.
- if we collected more data on $\boldsymbol{x}$, we could predict the associated $\hat{y}$.
- but $h(\cdot)$ does not provide a *counterfactual prediction* – that is, how the outcome would change if $\boldsymbol{x}$'s were exogenously shifted.

Econometrics (applied statistical causal inference):

- y is one-dimensional, x is low-dimensional.
- estimate a low-dimensional **causal parameter** ρ using
$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$
where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.
- ρ gives a prediction how outcome y would change if treatment variable x were **exogenously shifted**.
- useful for policy evaluation.

- **Glossary for machine learning vs causal inference terms:** `https://bit.ly/ML-Econ-Glossary`.

Outline

Regression Discontinuity Design (RDD)

- The speciality of applied economists is **natural experiments**.
  - contexts using observational data that approximate a randomized trial.

Regression Discontinuity Design (RDD)

- The speciality of applied economists is **natural experiments**.
  - contexts using observational data that approximate a randomized trial.
- RDD is one of those approaches:
  - threshold rules that are based on some ex-ante score: “running variable”

Regression Discontinuity Design (RDD)

- The speciality of applied economists is **natural experiments**.
  - contexts using observational data that approximate a randomized trial.
- RDD is one of those approaches:
  - threshold rules that are based on some ex-ante score: "running variable"
- Example "running variables" (also called forcing or assignment variable):
  - Score in entry exams
  - Income for subsidy eligibility
  - Age limit for alcohol consumption
  - Votes in an election

Regression Discontinuity Design (RDD)

- The speciality of applied economists is **natural experiments**.
  - contexts using observational data that approximate a randomized trial.
- RDD is one of those approaches:
  - threshold rules that are based on some ex-ante score: "running variable"
- Example "running variables" (also called forcing or assignment variable):
  - Score in entry exams
  - Income for subsidy eligibility
  - Age limit for alcohol consumption
  - Votes in an election
- If there is some randomness in the running variable, being just above or just below the threshold is randomly assigned.

Example: Effect of Minimum Legal Drinking Age on Death Rates

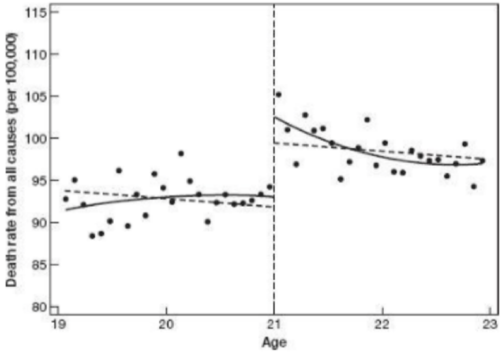
Carpenter and Dobkin (2009)

- outcome variable Y_i : death rate
- running variable x_i : age
- cutoff: $c = 21$, age where minors can suddenly drink legally
- treatment $D = \mathbb{I}[x_i > c]$: legal drinking status

Example: Effect of Minimum Legal Drinking Age on Death Rates

Carpenter and Dobkin (2009)

- outcome variable Y_i : death rate
- running variable x_i : age
- cutoff: $c = 21$, age where minors can suddenly drink legally
- treatment $D = \mathbb{I}[x_i > c]$: legal drinking status



RDD Estimation

- OLS regression:

$$Y_i = \alpha + \rho \mathbb{I}[x_i > c] + f(x_i)'\beta + \epsilon_i$$

- $f(x_i)$ includes polynomials in the forcing variable
  - generally linear or quadratic
  - can also interact with being above or below the cutoff

```
rdd = smf.ols(formula="death_rate ~ above_21 + age + age_squared", data=df).fit()
```

Localizing around cutoff

- Standard practice is to limit sample to a small bandwidth around the cutoff point
  - treatment more likely to be exogenous.

```
df_rdd = df[(df.age >= 19) & (df.age <= 22)]
rdd = smf.ols(formula="death_rate ~ above_21", data=df_rdd).fit()
```

Localizing around cutoff

- Standard practice is to limit sample to a small bandwidth around the cutoff point
  - treatment more likely to be exogenous.

```
df_rdd = df[(df.age >= 19) & (df.age <= 22)]
rdd = smf.ols(formula="death_rate ~ above_21", data=df_rdd).fit()
```

- How to choose the bandwidth?
  - Trade-off: the closer you get the better it is for identification, but the less data you have.
  - there are formulas for "optimal bandwidth" (e.g.: Imbens-Kalyanaraman 2011, Calonico, Cattaneo and Titiunik 2014).
  - can use the rdrobust package.
  - should also explore robustness to different bandwidths

Testing the validity of RDD

- RD Design can be invalid if individuals can precisely manipulate the assignment variable x_i in order to get (or to avoid) treatment.
- Testing for validity:
  1. Density of the running variable should be continuous (McCrary test)
  2. Predetermined characteristics should have the same distribution just above and just below the cut off

Testing the validity of RDD

- RD Design can be invalid if individuals can precisely manipulate the assignment variable x_i in order to get (or to avoid) treatment.
- Testing for validity:
  1. Density of the running variable should be continuous (McCrary test)
  2. Predetermined characteristics should have the same distribution just above and just below the cut off
- Another problem: other important variables are changing at the cutoff besides the treatment you had in mind.
  - have to think carefully / check if observable / run placebos.

Testing the validity of RDD

- RD Design can be invalid if individuals can precisely manipulate the assignment variable x_i in order to get (or to avoid) treatment.
- Testing for validity:
  1. Density of the running variable should be continuous (McCrary test)
  2. Predetermined characteristics should have the same distribution just above and just below the cut off
- Another problem: other important variables are changing at the cutoff besides the treatment you had in mind.
  - have to think carefully / check if observable / run placebos.

- **What are some validity issues for drinking-age cutoffs and death rates?**

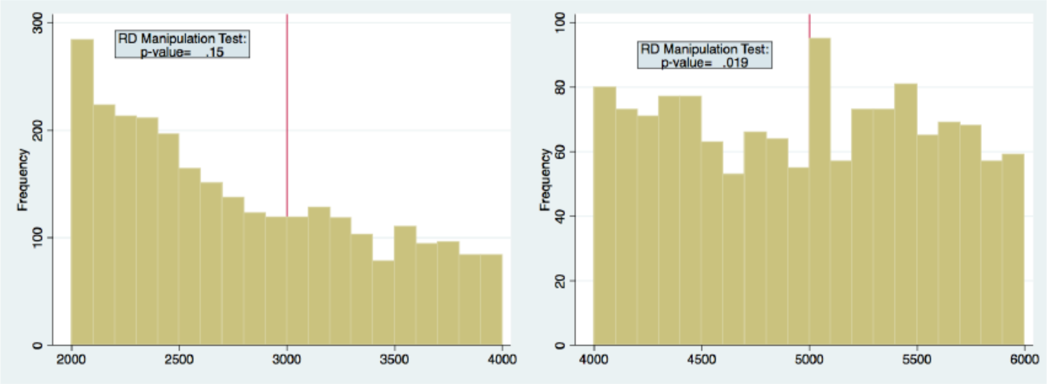
Example: Bagues and Campa (2017)

"Can Gender Quotas in Candidate Lists Empower Women? Evidence from a Regression Discontinuity Design"

- *We provide a comprehensive analysis of the short- and medium-term effects of* ***gender quotas*** *in candidate lists using evidence from Spain, where* ***quotas were introduced in 2007 in municipalities with more than 5,000 inhabitants****, and were* ***extended in 2011 to municipalities with more than 3,000 inhabitants****. Using a* ***Regression Discontinuity Design****, we find that* ***quotas raise the share of women among council members but they do not affect the quality of politicians****, as measured by their education attainment and by the number of votes obtained.*

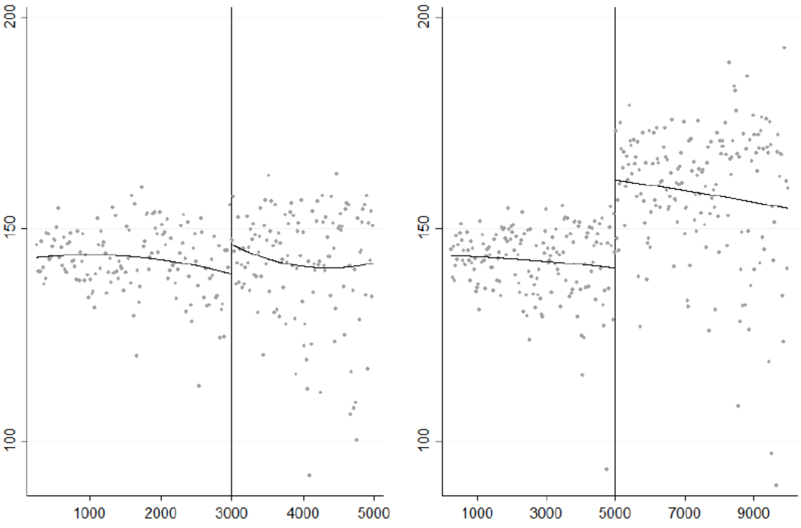
Manipulation Test: Density Around Cutoff

Bagues and Campa (2017): Histograms of Population Around Population Thresholds



Manipulation Test: Effect on Past Covariates

Bagues and Campa (2017): Federal Transfers Per Capita



RDD: Recap

- Useful method to analyze the impact of treatment when the assignment varies discontinuously due to some rules!
  - (test score, electoral results, income threshold, etc.)
- Graphical analysis is key, and can be very convincing
- Need a large sample around the threshold
- Have to check for manipulation at the threshold

Activity: Think of an RD Design

- 4 minutes:
  - Think of an idea for a regression discontinuity design
    - something from your field/hobby/etc
  - write down the associated variables:
    - outcome, running variable, threshold
- 6 minutes, with a partner:
  - take turns describing your RDD idea
  - then, for your partner's design, try to propose potential problems:
    - how would manipulation around the cutoff happen in your partner's example?
    - could other relevant variables be changing at the cutoff besides the treatment you had in mind?
  - discuss together: how would you test/fix these problems?

Outline

Week 2 Recap: Adjusting for Confounders

- Want to estimate effect of an explanatory variable D on an outcome Y.
  - but there is an observed confounder A that would bias an estimate of a causal relationship.

Week 2 Recap: Adjusting for Confounders

- Want to estimate effect of an explanatory variable D on an outcome Y.
  - but there is an observed confounder A that would bias an estimate of a causal relationship.
  - e.g. effect of drinking coffee on study productivity; confounders could be the time of day.

Week 2 Recap: Adjusting for Confounders

- Want to estimate effect of an explanatory variable D on an outcome Y.
  - but there is an observed confounder A that would bias an estimate of a causal relationship.
  - e.g. effect of drinking coffee on study productivity; confounders could be the time of day.
- If confounders are observed, can identify effect of D on Y by "adjusting for" or "controlling for" A.
- two ways to do that:
  1. residualize D and Y on A and estimate relationship between $\tilde{D}$ and $\tilde{Y}$.
  2. include A in a linear regression with outcome Y and predictor D.

Fixed Effects: Intuition

- Most of the time, there are many potential confounders that cannot be observed.
- in the coffee-productivity example, for each person i:
  - whether i's parents drink coffee
  - how close i live to a coffee shop
  - etc.

Fixed Effects: Intuition

- Most of the time, there are many potential confounders that cannot be observed.
- in the coffee-productivity example, for each person i:
  - whether i's parents drink coffee
  - how close i live to a coffee shop
  - etc.
- What if we can observe i's productivity multiple times?
  - sometimes i had coffee, and sometimes not.
  - then could "control" *for the person themselves*, rather than *their individual characteristics*.
  - this adjusts for everything unique to the individual i, whether it is observed or not.

Fixed Effects: Residualization Approach

In Week 2 we had outcome Y, treatment D, confounder A. We adjusted for A by:

1. learn the function $\hat{D}(A)$, compute residual $\tilde{D} = D - \hat{D}$
2. learn the function $\hat{Y}(A)$, compute residual $\tilde{Y} = Y - \hat{Y}$
3. if A is the only confounder, the relationship between $\tilde{D}$ and $\tilde{Y}$ is causal.

Fixed Effects: Residualization Approach

In Week 2 we had outcome Y, treatment D, confounder A. We adjusted for A by:

1. learn the function $\hat{D}(A)$, compute residual $\tilde{D} = D - \hat{D}$
2. learn the function $\hat{Y}(A)$, compute residual $\tilde{Y} = Y - \hat{Y}$
3. if A is the only confounder, the relationship between $\tilde{D}$ and $\tilde{Y}$ is causal.

With fixed effects, we have N individuals, indexed by i, and T periods, indexed by t:

1. de-mean (center) D_{it} for each i – i.e., form $\bar{D}_i = \frac{1}{T}\sum_t D_{it}$, then compute residual $\tilde{D}_{it} = D_{it} - \bar{D}_i, \forall i$.
2. de-mean Y_{it} the same way $\rightarrow \tilde{Y}_{it}$

Fixed Effects: Residualization Approach

In Week 2 we had outcome Y, treatment D, confounder A. We adjusted for A by:

1. learn the function $\hat{D}(A)$, compute residual $\tilde{D} = D - \hat{D}$
2. learn the function $\hat{Y}(A)$, compute residual $\tilde{Y} = Y - \hat{Y}$
3. if A is the only confounder, the relationship between $\tilde{D}$ and $\tilde{Y}$ is causal.

With fixed effects, we have N individuals, indexed by i, and T periods, indexed by t:

1. de-mean (center) D_{it} for each i – i.e., form $\bar{D}_i = \frac{1}{T}\sum_t D_{it}$, then compute residual $\tilde{D}_{it} = D_{it} - \bar{D}_i, \forall i$.
2. de-mean Y_{it} the same way $\rightarrow \tilde{Y}_{it}$
3. if all confounders are at the level of i (there are no confounders that vary over time within i), the relationship between $\tilde{D}_{it}$ and $\tilde{Y}_{it}$ is causal.

Fixed Effects: Regression Approach

- In Week 2 we had the linear model

$$Y_i = \alpha + \beta D_i + \gamma a_i + \eta_i$$

  - could adjust for observed confounder a_i by including it as a linear predictor in the regression.

```
ols = smf.ols(formula="product ~ coffee + educ", data=df).fit()
```

Fixed Effects: Regression Approach

- In Week 2 we had the linear model

$$Y_i = \alpha + \beta D_i + \gamma a_i + \eta_i$$

  - could adjust for observed confounder a_i by including it as a linear predictor in the regression.

```
ols = smf.ols(formula="product ~ coffee + educ", data=df).fit()
```

- Now we have

$$Y_{it} = \alpha_i + \beta D_{it} + \epsilon_{it}$$

  where t indexes time, and α_i is a "fixed effect" for person/group i.
  - α_i includes a set of binary variables that equal one for observations in i.
  - in machine learning this is called a one-hot-encoded categorical variable.

```
fe = smf.ols(formula="product ~ coffee + C(person_id)", data=df).fit()
```

Notes on fixed effects

- Can be used in many contexts:
  - the "entity" i could be people or firms or cities or countries, etc
- Usually, there are many confounders in a regression, many of which we can't measure.
  - Fixed effects adjust for **all** of them at the level of i.
  - we are comparing i to itself at a different time – i is its own control group!

Notes on fixed effects

- Can be used in many contexts:
  - the "entity" i could be people or firms or cities or countries, etc
- Usually, there are many confounders in a regression, many of which we can't measure.
  - Fixed effects adjust for **all** of them at the level of i.
  - we are comparing i to itself at a different time – i is its own control group!
- With the regression approach, we can add multiple sets of fixed effects, e.g.:

$$Y_{it} = \alpha_i + \alpha_t + \beta D_{it} + \epsilon_{ict}$$

  where now we have α_t, a "time fixed effect" which for example could represent time of day or day of the week – a set of dummies for observations at period t.

```
fe2 = smf.ols(formula="product ~ coffee + C(person_id) + C(time)", data=df).fit()
```

- this is a "two-way fixed-effects" model, which we will come back to shortly.

Randomization Blocks

- Consider an outcome Y_{ijc} in case i for judge j on court c, e.g. guilty/innocent.
- We want to estimate the effect of judge characteristic D_j, e.g. political party.
- If judges get different types of cases, estimating $\hat{\beta}$ from

$$Y_{ijc} = \alpha + \beta D_j + \epsilon_{ijc}$$

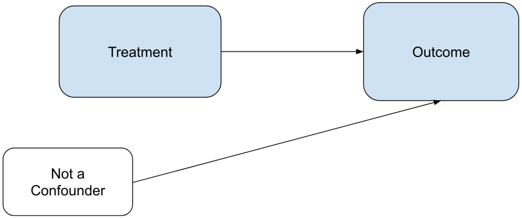
would be biased ($\text{cov}(D_j, \epsilon_{ijc}) \neq 0$) because the unobserved case characteristics (in ϵ_{ijc}) are affecting both D_j and Y_{ijc}.

Randomization Blocks

- Consider an outcome Y_{ijc} in case i for judge j on court c, e.g. guilty/innocent.
- We want to estimate the effect of judge characteristic D_j, e.g. political party.
- If judges get different types of cases, estimating $\hat{\beta}$ from

$$Y_{ijc} = \alpha + \beta D_j + \epsilon_{ijc}$$

  would be biased ($\text{cov}(D_j, \epsilon_{ijc}) \neq 0$) because the unobserved case characteristics (in ϵ_{ijc}) are affecting both D_j and Y_{ijc}.
- But say judges are randomly assigned within court.
  - Then, after conditioning on a court fixed effect α_c, there is no influence of the case-type confounders on the assigned judge characteristic (the treatment):



  - Hence, we get causal estimates of $\hat{\beta}$ from

$$Y_{ijc} = \alpha_c + \beta D_j + \eta_{ijc}$$

Ash et al (2022), “In-group bias in the Indian judiciary Evidence from 5 million criminal cases”

- Collect records from 5 million criminal cases in India trial courts, 2010-2020.

Ash et al (2022), “In-group bias in the Indian judiciary Evidence from 5 million criminal cases”

- Collect records from 5 million criminal cases in India trial courts, 2010-2020.
- Use an LSTM neural network to classify judge and defendant names by religion/gender.

Ash et al (2022), “In-group bias in the Indian judiciary Evidence from 5 million criminal cases”

- Collect records from 5 million criminal cases in India trial courts, 2010-2020.
- Use an LSTM neural network to classify judge and defendant names by religion/gender.
- Use random assignment of judges to defendants to look at in-group bias effects when judge and defendant identity match.

Assignment of cases to judges in the lower judiciary

Cases are assigned to judges following a clear set of rules:

- Police station location determines courthouse.
- Within court, courtroom is assigned based on station and charge.
  - More serious charges $\rightarrow$ more senior judge required.

Assignment of cases to judges in the lower judiciary

Cases are assigned to judges following a clear set of rules:

- Police station location determines courthouse.
- Within court, courtroom is assigned based on station and charge.
  - More serious charges $\rightarrow$ more senior judge required.
- Judges rotate through rooms every few months, so same station/charge leads to different judge.
- Forum-shopping is prohibited and by all reports, not practiced.

Assignment of cases to judges in the lower judiciary

Cases are assigned to judges following a clear set of rules:

- Police station location determines courthouse.
- Within court, courtroom is assigned based on station and charge.
  - More serious charges $\rightarrow$ more senior judge required.
- Judges rotate through rooms every few months, so same station/charge leads to different judge.
- Forum-shopping is prohibited and by all reports, not practiced.

Conditional on court-time and charge fixed effects, cases are as good as randomly assigned.

Estimating equation

We model outcome Y_{ict} for case i in court c at time t as

$$Y_{i,c,t} = \alpha + \beta_1 \text{judge\_male}_{i,c,t} + \beta_2 \text{def\_male}_{i,c,t} + \beta_3 \text{judge\_male}_{i,c,t} * \text{def\_male}_{i,c,t} + \phi_{c,t} + \delta\chi_{i,c,t} + \epsilon \quad (1)$$

- **everything analogous for Muslim/non-Muslim**
  - β_1 = effect of male judge on female defendant
  - $\beta_1 + \beta_3$ = effect of male judge on male defendant
  - β_3 = gender in-group bias
- $\phi_{c,t}$: court-time fixed effect (month or year)
- $\delta\chi_{i,c,t}$: other covariates, including act-section fixed effects and other defendant characteristics
- standard err. clustered by judge (this does not matter much)

Testing exogenous judge assignment

Table 2: Balance test for assignment of judge identity

	(1) Female judge	(2) Female judge	(3) Muslim judge	(4) Muslim judge
Panel A: Sample including observations with no decision				
Female defendant	0.000	0.000	0.001	0.001
	(0.001)	(0.001)	(0.000)	(0.000)
Muslim defendant	0.001	0.001	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	5412789	5425998	5498486	5511735
Fixed Effect	Court-month	Court-year	Court-month	Court-year

Standard errors in parentheses

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

Notes: This table reports results from a formal test of random assignment of judges to cases in the study sample. For specification details, see Equations 2 and 3. Columns 1–2 report the likelihood of being assigned to a female judge relative to a male judge using court-month, and court-year fixed effects. Columns 3–4 report the likelihood of being assigned to a Muslim judge relative to a non-Muslim judge using court-month, and court-year fixed effects. Charge section fixed effects have been used across all columns reported. Heteroskedasticity robust standard errors are reported below point estimates.

No gender in-group bias

	Outcome variable: Acquittal rate					
	(1)	(2)	(3)	(4)	(5)	(6)
Male judge on female defendant	-0.0077***	-0.0071**	—	-0.0072***	-0.0068**	—
	(0.0026)	(0.0032)		(0.0025)	(0.0031)	
Male judge on male defendant	-0.0062***	-0.0056**	—	-0.0056***	-0.0053**	—
	(0.0022)	(0.0027)		(0.0020)	(0.0026)	
Difference = Own gender bias	0.0015	0.0014	-0.0002	0.0016	0.0015	-0.0003
	(0.0017)	(0.0017)	(0.0015)	(0.0017)	(0.0018)	(0.0016)
Reference group mean	0.1756	0.1767	0.1766	0.1761	0.1771	0.1771
Observations	5221156	5127538	5126026	5234591	5141055	5139253
Demographic controls	No	Yes	Yes	No	Yes	Yes
Judge fixed effect	No	No	Yes	No	No	Yes
Fixed Effect	Court-month	Court-month	Court-month	Court-year	Court-year	Court-year

Notes: Standard errors in parentheses. $*\ p < 0.10$, $**\ p < 0.05$, $***\ p < 0.01$.
Reference group: Female judges, female defendants.
Specification: $Y_{i,c,t} = \alpha + \beta_1 judge\_male_{i,c,t} + \beta_2 def\_male_{i,c,t} + \beta_3 judge\_male_{i,c,t} * def\_male_{i,c,t} + \phi_{c,t} + \delta\chi_{i,c,t} + \epsilon$

No religion in-group bias

	Outcome variable: Acquittal rate					
	(1)	(2)	(3)	(4)	(5)	(6)
Non-Muslim judge on Muslim defendant	0.0079**	0.0075	—	0.0073*	0.0061	—
	(0.0040)	(0.0050)		(0.0038)	(0.0049)	
Non-Muslim judge on non-Muslim defendant	0.0071**	0.0072	—	0.0065**	0.0062	—
	(0.0034)	(0.0044)		(0.0030)	(0.0041)	
Difference = Own religion bias	-0.0008	-0.0004	0.0016	-0.0008	0.0001	0.0020
	(0.0027)	(0.0028)	(0.0021)	(0.0028)	(0.0029)	(0.0021)
Reference group mean	0.1804	0.1836	0.1837	0.1807	0.1839	0.184
Observations	5644234	5212279	5210766	5657299	5225791	5223976
Demographic controls	No	Yes	Yes	No	Yes	Yes
Judge fixed effect	No	No	Yes	No	No	Yes
Fixed Effect	Court-month	Court-month	Court-month	Court-year	Court-year	Court-year

Notes: Standard errors in parentheses. $*\ p < 0.10$, $**\ p < 0.05$, $***\ p < 0.01$.
Reference group: Muslim judges, Muslim defendants.
Specification: $Y_{i,c,t} = \alpha + \beta_1 judge\_nonmus_{i,c,t} + \beta_2 def\_nonmus_{i,c,t} + \beta_3 judge\_nonmus_{i,c,t} * def\_nonmus_{i,c,t} + \phi_{c,t} + \delta\chi_{i,c,t} + \epsilon$

Outline

Regression Discontinuity Design

Fixed Effects

Panel Data / Differences-in-Differences

Panel Data (Longitudinal Data) is Data Over Time

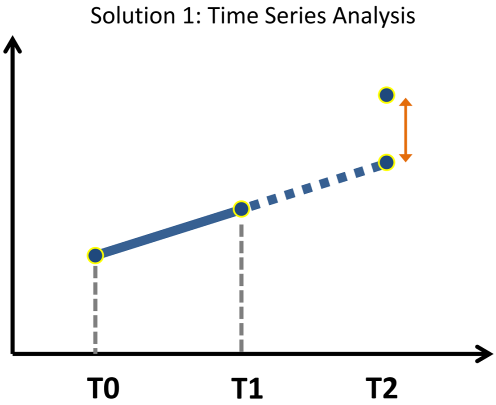
- we have outcomes y_{it} for "unit" (individual/group) i at time t
- N units and T time periods
  - a "balanced" dataset will have NT observations.
  - "unbalanced" panel data means that some unit-period pairs are missing – e.g. due to entering or leaving the sample. this is not a problem in practice.
- The goal of panel data methods is to construct counterfactuals using the longitudinal structure of the data.

What if there is only one unit? Time Series Analysis

- In macroeconomics (analysis of the whole economy), you only observe one unit (the economy).
  - How to estimate causal effect of a macroeconomic policy like changing interest rates?

What if there is only one unit? Time Series Analysis

- In macroeconomics (analysis of the whole economy), you only observe one unit (the economy).
  - How to estimate causal effect of a macroeconomic policy like changing interest rates?



- "time series analysis" assumes economy continues on current trend in absence of intervention.

Source: Yixing Zu slides.

Example where previous methods fail

- Example: taxes raised in canton A, but **not** in canton B
  - we observe employment Y_{jt} in time periods t before and after the reform in both cantons j
  - what is the effect of the tax rise D_{jt} on employment?

Example where previous methods fail

- Example: taxes raised in canton A, but **not** in canton B
  - we observe employment Y_{jt} in time periods t before and after the reform in both cantons j
  - what is the effect of the tax rise D_{jt} on employment?
- naive OLS approach:

$$Y_{jt} = \alpha + \gamma D_{jt} + \varepsilon_{jt}$$

  - there are canton-level confounders biasing the estimate.

Example where previous methods fail

- Example: taxes raised in canton A, but **not** in canton B
  - we observe employment Y_{jt} in time periods t before and after the reform in both cantons j
  - what is the effect of the tax rise D_{jt} on employment?
- naive OLS approach:

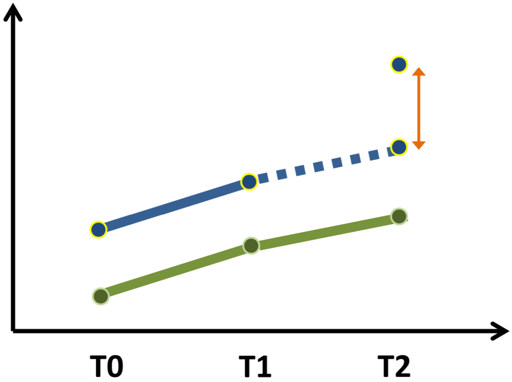
$$Y_{jt} = \alpha + \gamma D_{jt} + \varepsilon_{jt}$$

  - there are canton-level confounders biasing the estimate.
- fixed effects approach:

$$Y_{jt} = \alpha_j + \gamma D_{jt} + \varepsilon_{jt}$$

  - $\hat{\gamma}$ estimates the pre/post change in employment for canton A
  - **but**:
    - what if employment was already going up over time in all of switzerland?
    - the post-treatment estimate $\hat{\gamma}$ is biased upward by the time confounder.

Differences-in-Differences

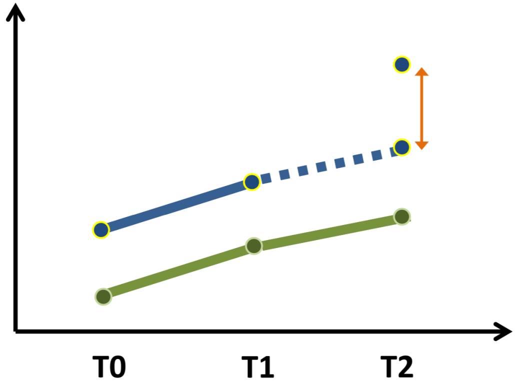


- use canton B as a counterfactual to adjust for the time trend.
- In this example, the DD estimator is

$$[Y_{A1} - Y_{A0}] - [Y_{B1} - Y_{B0}]$$

$=$ **employment change in <u>treated</u> canton, relative to employment change in <u>comparison</u> canton.**

Differences-in-Differences



- use canton B as a counterfactual to adjust for the time trend.
- In this example, the DD estimator is

$$[Y_{A1} - Y_{A0}] - [Y_{B1} - Y_{B0}]$$

= **employment change in treated canton, relative to employment change in comparison canton.**

- in regression form, we estimate

$$Y_{jt} = \alpha_j + \alpha_t + \gamma D_{jt} + \varepsilon_{jt}$$

where α_t is a **time fixed effect** – an indicator variable for each time period t.

```
dd = smf.ols(formula="emp ~ tax + C(canton) + C(time)", data=df).fit()
```

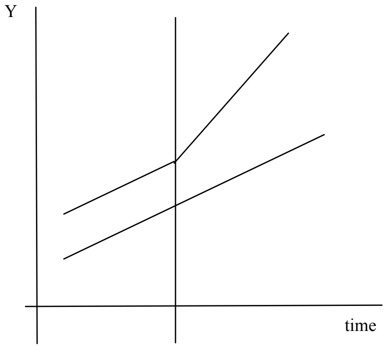
Diff-in-diff: Checking for Parallel trends

- The identification assumption for diff-in-diff is **"parallel trends"** :
  - e.g., absent tax change, trend in employment would have been the same in cantons A and B.

Diff-in-diff: Checking for Parallel trends

- The identification assumption for diff-in-diff is **"parallel trends"** :
  - e.g., absent tax change, trend in employment would have been the same in cantons A and B.

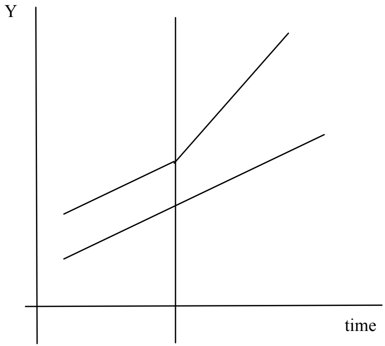
Good



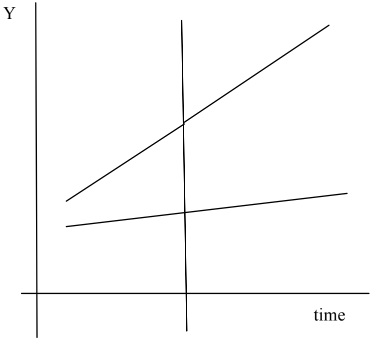
Diff-in-diff: Checking for Parallel trends

- The identification assumption for diff-in-diff is **"parallel trends"** :
  - e.g., absent tax change, trend in employment would have been the same in cantons A and B.

Good



Not Good



Two-Way Fixed-Effects Regression

- The regression form
$$Y_{jt} = \alpha_j + \alpha_t + \gamma \mathsf{D}_{jt} + \varepsilon_{jt}$$
generalizes to > 2 groups and > 2 periods

Two-Way Fixed-Effects Regression

- The regression form
$$Y_{jt} = \alpha_j + \alpha_t + \gamma \mathrm{D}_{jt} + \varepsilon_{jt}$$
generalizes to > 2 groups and > 2 periods
- TWFE is an empirical workhorse.
  - e.g., in our eample, taxes and employment across cantons could be correlated for many confounding reasons.
  - TWFE / Diffs-in-diffs holds constant many of the most important confounders:
    - time-invariant canton-level factors
    - nationwide time-varying factors

Two-Way Fixed-Effects Regression

- The regression form
$$Y_{jt} = \alpha_j + \alpha_t + \gamma D_{jt} + \varepsilon_{jt}$$
generalizes to > 2 groups and > 2 periods
- TWFE is an empirical workhorse.
  - e.g., in our eample, taxes and employment across cantons could be correlated for many confounding reasons.
  - TWFE / Diffs-in-diffs holds constant many of the most important confounders:
    - time-invariant canton-level factors
    - nationwide time-varying factors
- Potential confounders must
  - vary over time by canton
  - be correlated with outcome variable
  - be correlated with the timing of treatment/reforms

Threats to validity for TWFE regression

- Can check that treatment cantons evolved similarly to comparison cantons before reform.
  - can also add canton-specific trends.

Threats to validity for TWFE regression

- Can check that treatment cantons evolved similarly to comparison cantons before reform.
  - can also add canton-specific trends.
- Skeptical questions to ask:
  - Why did the treatment group adopt the policy, and not the control group?
  - Were other policies adopted at the same time that might also affect the outcome?
  - Could the treatment spill over into the comparison cantons?

A note on standard errors

- Consider the regression for cantonal tax cuts and employment. We have 26 cantons.
  - the default standard errors formula for OLS assume that all observations are independent realizations.
- Compare the following analyses:
  - including the 10 years before and after the reform ($N = 260$)
  - including the 20 years before and after ($N = 520$)

A note on standard errors

- Consider the regression for cantonal tax cuts and employment. We have 26 cantons.
  - the default standard errors formula for OLS assume that all observations are independent realizations.
- Compare the following analyses:
  - including the 10 years before and after the reform ($N = 260$)
  - including the 20 years before and after ($N = 520$)
- Using the default SE's, the second analysis would give much more precise estimate, even though the data contain nearly equivalent information.

Clustering Standard Errors

Cluster standard errors:

- statistically acknowledges how many independent sources of information there are in the data.
- the standard approach is to cluster at the unit where treatment is assigned.
  - in this example, by canton.

```
dd = smf.ols(formula="emp ~ tax + C(canton) + C(time)", data=df)
result = dd.fit(cov_type="cluster",cov_kwds={"groups":df["canton"]})
```

- for city-level reforms cluster by city, etc.

Event Study: Dynamic Treatment Effects

- So far we have estimated regressions like

$$Y_{jt} = \alpha_j + \alpha_t + \beta D_{jt} + \varepsilon_{jt}$$

  - $\hat{\beta}$ will give us the average effect in the post-treatment period, relative to pre-treatment and to the control group.
- What if we care about the dynamics of the effect? How it changes over time?

Event Study: Dynamic Treatment Effects

- So far we have estimated regressions like

$$Y_{jt} = \alpha_j + \alpha_t + \beta \mathrm{D}_{jt} + \varepsilon_{jt}$$

  - $\hat{\beta}$ will give us the average effect in the post-treatment period, relative to pre-treatment and to the control group.
- What if we care about the dynamics of the effect? How it changes over time?
- The proper way to do this is with a "panel event study", where we estimate

$$Y_{jt} = \alpha_j + \alpha_t + \sum_{\tau=-W,\tau\neq-1}^{W} \beta_\tau \mathrm{D}_{jt}^{\tau} + \varepsilon_{jt}$$

  - here, each item D_{jt}^{τ} represents a "lead" or a "lag" of treatment time. so, e.g., $\tau = 0$ for the period of treatment, $\tau = 1$ is the year after, $\tau = -2$ is two years before, etc.
  - $\tau = -1$, the year before treatment, is dropped $\rightarrow$ it is the reference year, and $\hat{\beta}_\tau$ measures the difference relative to $\tau = -1$.
- see "The Effect", Section 18.2 and 18.3 for more detail.

Video Presentation: Stevenson and Doleac (2022)