OLS for the demographic historian

Auke Rijpma, a.rijpma@uu.nl

October 8, 2025

Introduction

Introduction

- Who am I
- Who are you?

Introduction: who this is for

- This will be a very general and gentle introdution to OLS models.
- Assuming that you either:
  - Had some (not a lot) statistics in your BA/MA, but are not very experienced using OLS in your research
  - Are a more seasoned demographic historian and use lots of survival models, but are interested in the "grass in the neighbour's yard".
- So assumed to be familiar with basic statistical concepts, and will keep everything relatively surface-level.

Introduction: what will we cover

- Introduction to the OLS model, concepts and interpretation.
- Closer look at estimation of OLS models in R, in particular using the fixest library.
- Introduction to "causal inference" and an overview of some popular strategies to deal with it in the context of OLS.
- A closer look at one of them, difference-in-differences.

Introduction to the OLS model

Introduction to the OLS model

- Regression analysis, OLS in particular, is the workhorse of the social sciences.
- Quantify strength of relationship between variables, with the possibility of "keeping other variables constant".
- Can be interpreted as so-called "marginal effects", effect of a change in one variable (Δx) on the outcome variable (Δy).
- Can perform statistical inference on the results to account for uncertainty.

Introduction to the OLS model: advantages

- Extremely flexible method, can be written to fit many types of models.
- Many important extensions
  - clustered research designs (individuals in municipalities; municipality over time, etc)
  - causal inference strategies:
    - time invariant-confounders: FE models
    - instrumental variables
    - difference-in-differences
    - regression discontinuity designs
- Many of these can be estimated in non-OLS context, but typically OLS is first, most complete, and will have good (up-to-date, complete, well-documented) packages.

Introduction to the OLS model: advantages

- Survival models are very often the appropriate model for modern longitudinal datasets in demographic history (like HSN).
- Still, OLS might be worth considering:
  - Sometimes data is not suited for survival models (e.g. CDRs for municipalities).
  - Ability to tap into causal inference tools.

The basics of the OLS model

Regression analysis: the regression equation.

- Population model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- Y and X are random variables. α is the "intercept" en β the slope.
  - Y: dependent, outcome, or response variable.
  - X independent or predictor variable.
- ϵ is the error term, capturing parts of the data that cannot be explained by the deterministic part of the model.

Estimating a model: base-R

```
m = lm(Fertility ~ Agriculture, data = swiss)
```

- Uses `lm()` command
- Uses built-in `swiss` data from the Princeton Fertility project.
- Assigns output with =.

Estimating a model: fixest

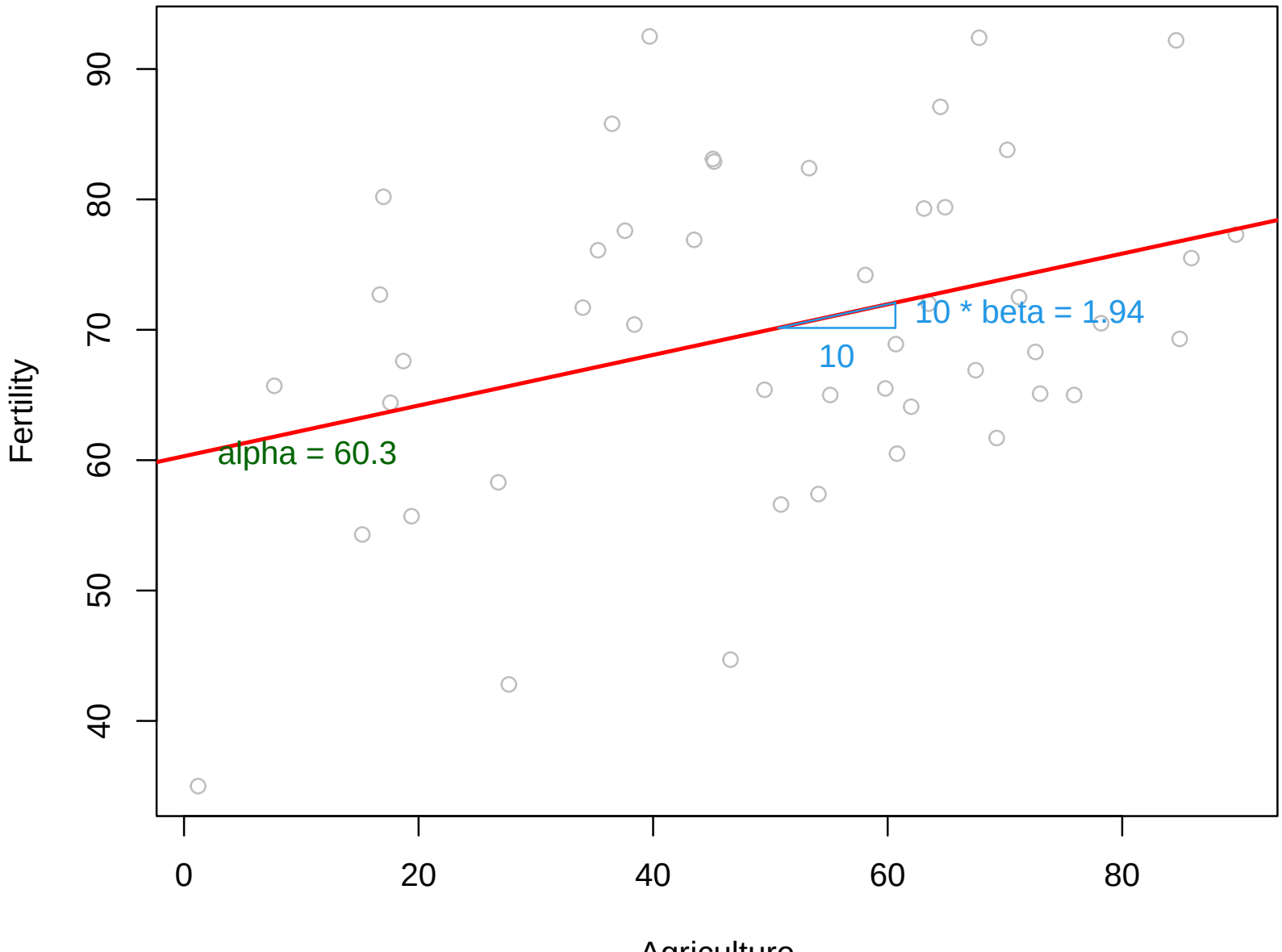
- `lm()` is a perfectly fine command, and hundres of extensions have been written for it (sandwich, lmtest, CAR, texreg being important ones).
- However, later on I will extensively use the `fixest` library.
- Very efficient estimation of an important class of models, "Fixed effects" models.
- Everything you would need in one place if you want to use the methods in the average JEH paper.

Estimating a model: fixest

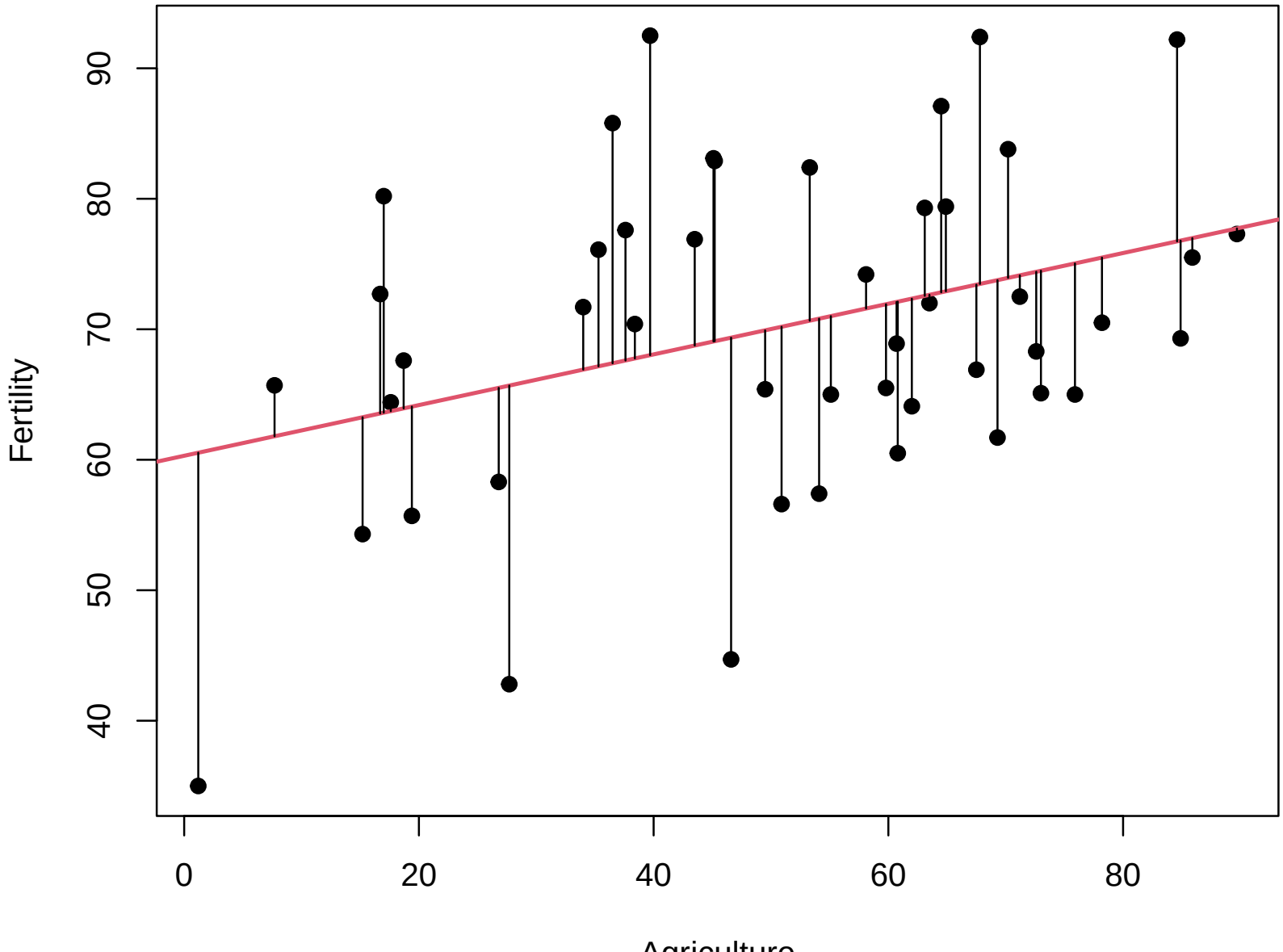
```
library("fixest")
m = fixest::feols(Fertility ~ Agriculture, data = swiss)
```

- Core syntax is the same though, just using `fixest::feols()` rather than `lm()`.

The regression fit



The regression fit: residuals



Regression model: fitting the line

How would you determine what the intercept and the slope of the line should be?

Regression model: fitting the line

- Done by the "least squares" criterion.
- Hence the name "Ordinary Least Squares" regression.
- Under a number of assumptions it can be shown that this OLS fit is an optimal one:
  - Unbiased: expected value of β equals the true population parameter.
  - Most efficient: lowest variance of β.

OLS assumptions

- OLS assumptions for β to be unbiased and efficient.
- Unbiased, $E\hat{\beta} = \beta$ if
  1. linear in its parameters
  2. mean zero error term
  3. no perfect collinearity
  4. no correlation independent variable and error term.
- Efficient, that is lowest $Var(\beta)$ if
  1. no serial correlation
  2. constant variance error
- Of these, linear parameters and no cor independent variables and error term (meaning no omitted variables) are the most important in practice.
- Others are easy to fix (constant variance error term) or fixed for you (mean zero error term, no perfect collinearity)

Regression model: statistical inference

- $\hat{\beta}$ is a sample estimate of a population and has a sampling distribution.
- We can use this to calculate a confidence interval or do a significance test and answer questions:
  - What is the chance that we get this result or something more extreme if the actual population value is zero? (p-value)
- Calculate a standard error of $\hat{\beta}$.

$$se(\hat{\beta}_k) = \frac{\hat{\sigma}}{\sqrt{1 - R_k^2} TSS_{X_k}}$$

- $\hat{\sigma}$: "based on residuals"
- R_k^2: R2 from an auxiliary regression of Xk on all other independent variables
- TSS_{X_k}: total sum of squares for independent variable X_k
- Consequences for "statistical power": more observations, high variance X are nice for higher precision.

Regression model: statistical inference

- Use this standard error to calculate a t-statistic on $\hat{\beta}$

$$t = \frac{\hat{\beta} - \beta}{se(\hat{\beta})}$$

- Where β is the value of the null hypothesis, usually zero.
- H_0 in a regression analysis is that there is no relationship between the variables (in the population).
- H_A is that there is a relationship:
  - Positive: $H_A : \beta > 0$ ($H_0 : \beta <= 0$)
  - Negative: $H_A : \beta < 0$
  - Any: $H_A : \beta \neq 0$
- The t-statistic again implies a p-value which we can use to reject H_0 or not.
- Or use it to construct a CI.

Regression analysis: R example

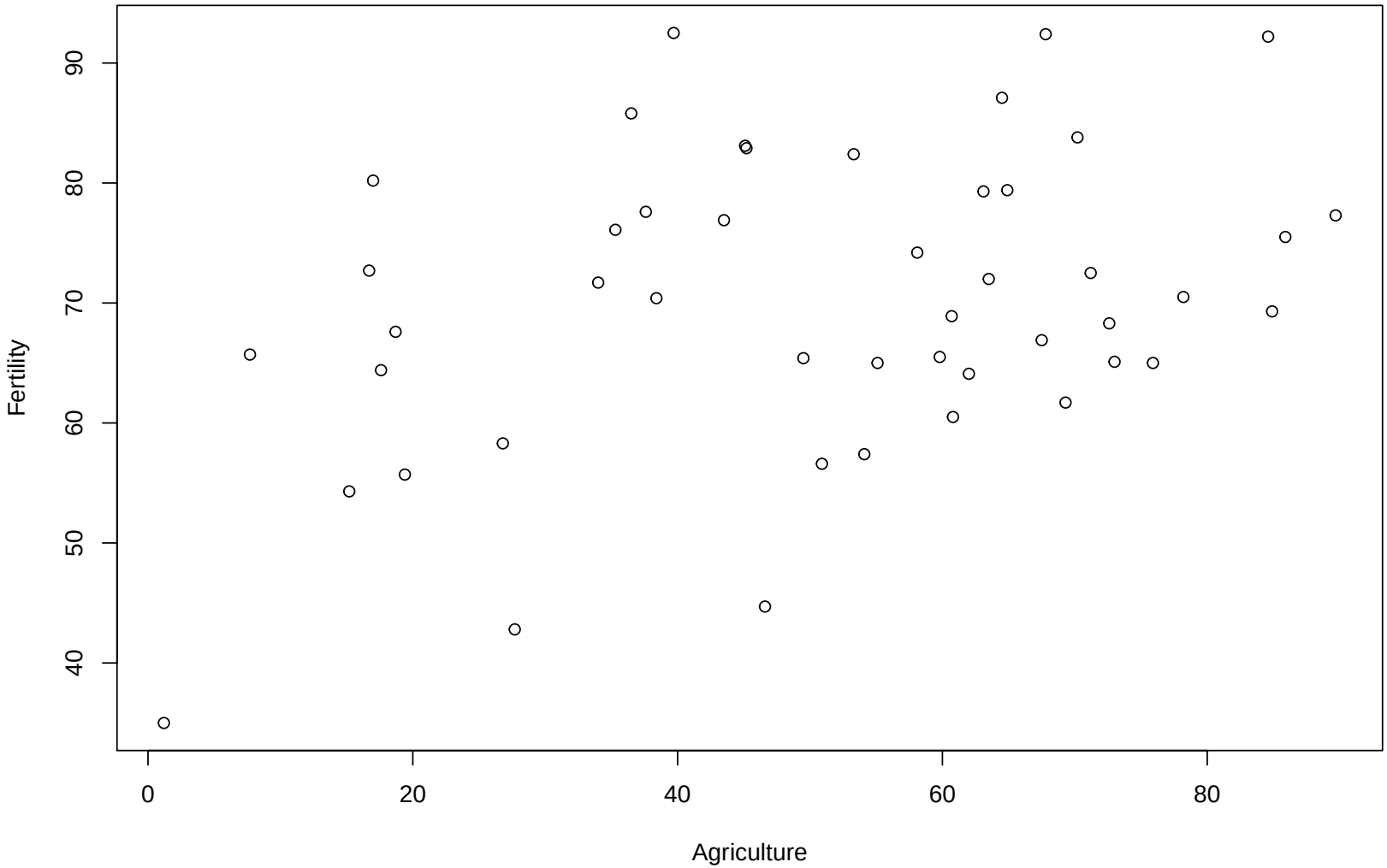
```
head(swiss)
```

```
##                Fertility Agriculture Examination Educatio
## Courtelary          80.2        17.0          15        12
## Delemont            83.1        45.1           6         9
## Franches-Mnt        92.5        39.7           5         5
## Moutier             85.8        36.5          12         7
## Neuveville          76.9        43.5          17        15
## Porrentruy          76.1        35.3           9         7
##                Infant.Mortality
## Courtelary                 22.2
## Delemont                   22.2
## Franches-Mnt               20.2
## Moutier                    20.3
## Neuveville                 20.6
## Porrentruy                 26.6
```

Regression analysis: R example I

```
plot(Fertility ~ Agriculture, data = swiss)
```

Regression analysis: R example II



Regression analysis: R example

- OLS done with `lm()` function – "linear model"

```
lm(Fertility ~ Agriculture, data = swiss)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture, data = swiss)
##
## Coefficients:
## (Intercept)  Agriculture
##     60.3044       0.1942
```

Regression analysis: R example

- In R we assign and re-use things

```
m = lm(Fertility ~ Agriculture, data = swiss)
```

Regression analysis: R example

```
summary(m)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5374  -7.8685  -0.6362   9.0464  24.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.30438    4.25126  14.185   <2e-16 ***
## Agriculture  0.19420    0.07671   2.532   0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 11.82 on 45 degrees of freedom
```

Regression analysis: R example

- Key output:
  - Estimate
  - Std. Error
  - t value
  - Pr(>|t|)
  - Multiple R-squared:

Regression analysis: R example

```
library("fixest")
m2 = feols(Fertility ~ Agriculture, data = swiss)
etable(m2)
```

```
##                                    m2
## Dependent Var.:             Fertility
##
## Constant             60.30*** (4.251)
## Agriculture          0.1942* (0.0767)
## ____________ _________________
## S.E. type                         IID
## Observations                       47
## R2                            0.12466
## Adj. R2                       0.10521
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Multiple regression

- So far we had one independent/predictor variable, but regression analysis allows for multiple independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- A few things change when we do this:
  - We can no longer think of the regression in terms of a scatter plot and a regression line.
  - The interpretation changes: we now have to say something about the other variables as well.
  - We interpretation of a slope variable is now with "keeping the other variables (all else) constant".
  - This means we can correct the relationship between Y and X1 for the effect of X2 on both (it's being kept constant, after all)

Understanding multiple regression: the formula

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Saying what the increase in Y (ΔY) for a one unit-increase in X_1 (ΔX_1) no longer makes sense as long as the other variable can still move.

Understanding multiple regression: the (hyper)plane

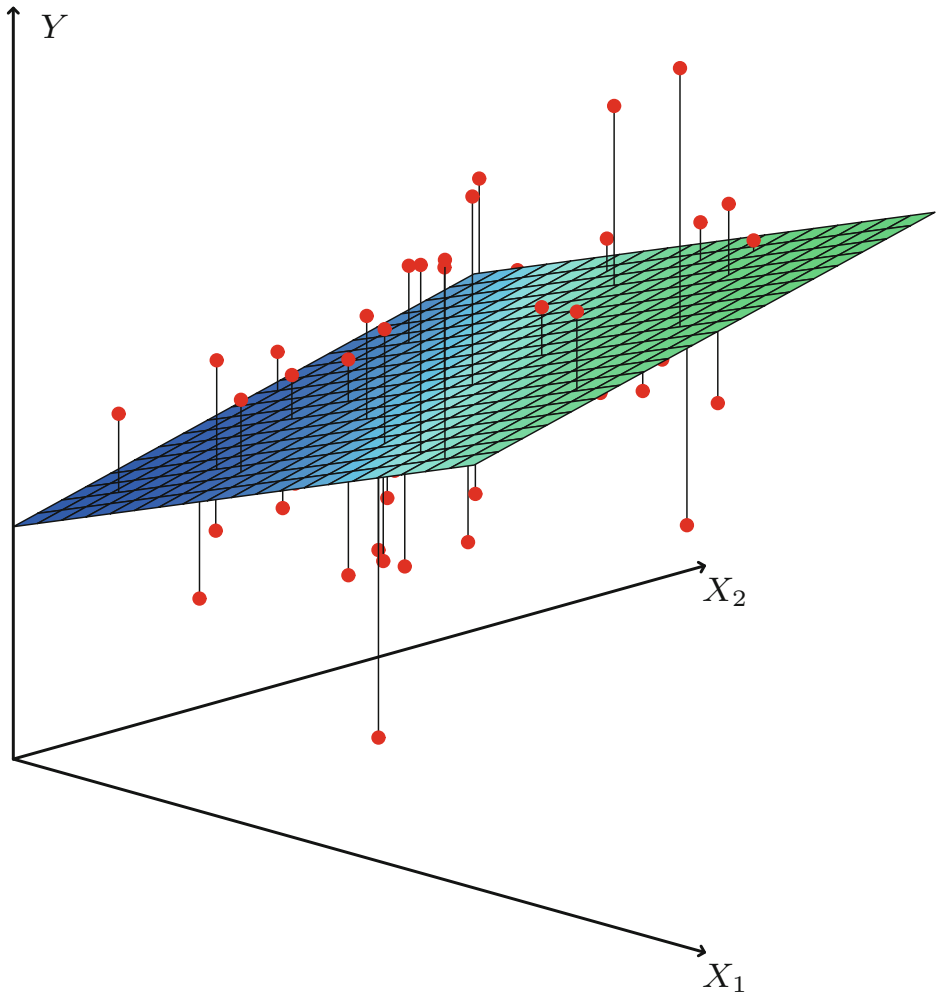


FIGURE 3.4. *In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation*

Understanding multiple regression: the (hyper)plane

- If you look at such a plane, you can see that the slope for the relation between Y and X_1 is the same for all values of X_2.
- Imagine if the slope for X_1 and Y changed for value of X_2 – would it be a plane?

Interpretation of multiple regression I

```
m1 = feols(Fertility ~ Infant.Mortality, data = swiss)
m2 = feols(Fertility ~ Infant.Mortality + Education, data =
etable(m1, m2)
```

- Model 1: a one unit increase in infant mortality predicts a 1.7 increase in standardised fertility.
- Model 2: a one unit increase in infant mortality predicts a 1.5 increase in standardised fertility, keeping education constant.

Correlation and causation

Correlation and causation

- Cliche that correlation does not imply causation.
- Let's talk through an example.
- Say we have the following regression result.

$$\hat{IceCreamSale} = 2 + 80.1\hat{PercentShorts}$$

- Does the percent of men wearing short cause ice cream sales?
- Why not?

Correlation and causation

- Let's do this on a slightly more serious model as well:

```
m1 = feols(Fertility ~ Infant.Mortality, data = swiss)
m2 = feols(Fertility ~ Infant.Mortality + Education, data =
etable(m1, m2)
```

```
##                              m1                  m2
## Dependent Var.:       Fertility           Fertility
##
## Constant          34.52** (11.71)    48.82*** (8.890)
## Infant.Mortality 1.786** (0.5812)   1.519*** (0.4287)
## Education                          -0.8167*** (0.1298)
## ________________ _______________ ___________________
## S.E. type                    IID                 IID
## Observations                  47                  47
## R2                       0.17352             0.56478
## Adj. R2                  0.15515             0.54500
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Correlation and causation

- If our original model is "infant mortality causes higher fertility (e.g. because families strive for a certain number of surviving children)", this could be confounded by education:
  - More education will lead to lower fertility
  - More education can lead to lower infant mortality
- Adding education as a control reduced the strength of the relation between infant mortality and fertility
- If that was the only confounder we worry about, we would now have a causal estimate!

Correlation and causation

- These issues – temperature causing both ice cream sales and short wearing, Education causing both fertility and infant mortality – is called "omitted variable bias".
- It is a crucial assumption underlying (a causal interpretation of) the regression model.
- Adding omitted variables to a regression will resolve this bias, and can give you a causal interpretation of your coefficients!

Multiple regression: model building

- Understanding why these estimates change when adding controls can be difficult.
- Knowing which variables to include is as well.
- Note that there is such a thing as a "bad control", which will actually bias your estimates.
- We also want to be parsimonious.
- Kitchen sink regression considered bad practice.

Multiple regression: model building

- Model building is still difficult.
- One tip to understanding how the variables work is by building models stepwise.
- Start with the independent variable you are interested in, and add relevant control variables one step at a time.
  - Don't keep this up until you have used all your variables, rather use it as a tool to understand what is going on.
  - Also shows reader your results hold across a wide range of specifications.
    - But of course it can be that the estimate adjusting for X_p is the correct one!
- Below model building in Java forced labour and mortality paper.

Multiple regression: model building

TABLE 2
THE IMPACT OF THE CULTIVATION SYSTEM ON CRUDE DEATH RATES

	(1)	(2)	(3)	(4)	(5)	(6)
log forced labor	0.121**	0.190***	0.191***	0.163***	0.147**	0.186**
	(0.045)	(0.055)	(0.057)	(0.060)	(0.067)	(0.079)
log buffalo			–0.193	–0.224*	–0.208	–0.126
			(0.124)	(0.119)	(0.131)	(0.102)
log production				0.062**	0.066**	0.110**
				(0.028)	(0.029)	(0.044)
log rice price					0.040	0.096
					(0.060)	(0.062)
log crop payment						–0.043
						(0.061)
officials						–3.023*
						(1.614)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year-fixed effects	No	Yes	Yes	Yes	Yes	Yes
Observations	547	547	546	546	468	393
R^2	0.27	0.45	0.45	0.46	0.47	0.51

Notes: The results were obtained estimating Equation (1) using log crude death rates as dependent variable. All variables, except for *officials*, are log-transformed and the panel is unbalanced. Driscoll-Kraay standard errors in parentheses.
*p<0.10, **p<0.05, ***p<0.01
Sources: Boomgaard and Gooszen (1991); Van Baardewijk (1994); *Kultuur verslag* (1834–1854)

Extensions

Solutions to the causal interpretation issue

- Naive correlation or regression unlikely to support causal interpretation of parameter of interest.
- Solutions include:
  - Have a completely model of the data generating process, that is all the control variables in the right functional form (difficult!).
  - Panel data methods.
  - Difference in differences.
    - Synthetic control methods
  - Instrumental variables
  - Regression discontinuity designs.
- Exciting, because there are ways to get causal statements from observational data!
- Also dangerous, because they can easily be misapplied!
- Lots of caveats, technical problems to check, which will be completely ignored here.

Panel/FE models

- In the OLS context, nested data is often dealt with through "fixed effects" and "clustered standard errors".
- Nested data: individuals over time, individuals within municaplities, municipalities within provinces, etc.
- Clustered standard errors deal with within-group correlation.
- Fixed effects are essentially dummy variables for the nesting groups, so individual dummies if we're following individuals over time.
- Causal perspective: this allows us to control for all time-invariant variables.
- Individuals over time example above: parental characteristics are now controlled for!
- Implementation above, also coming up in diff-in-diff.

Difference-in-differences

Difference-in-Differences via graphs

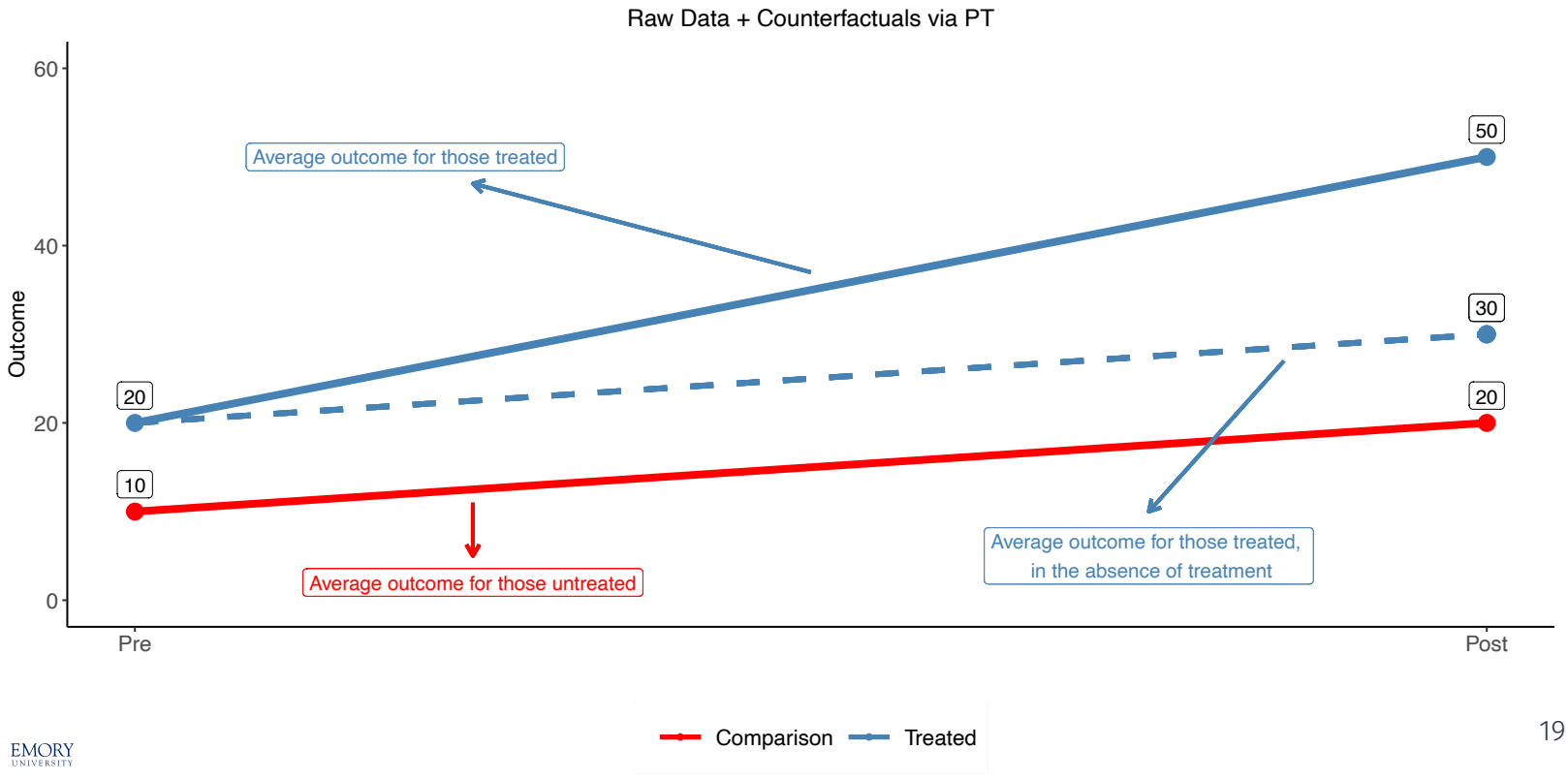


Figure 3: https://psantanna.com/did-resources/

Difference in differences

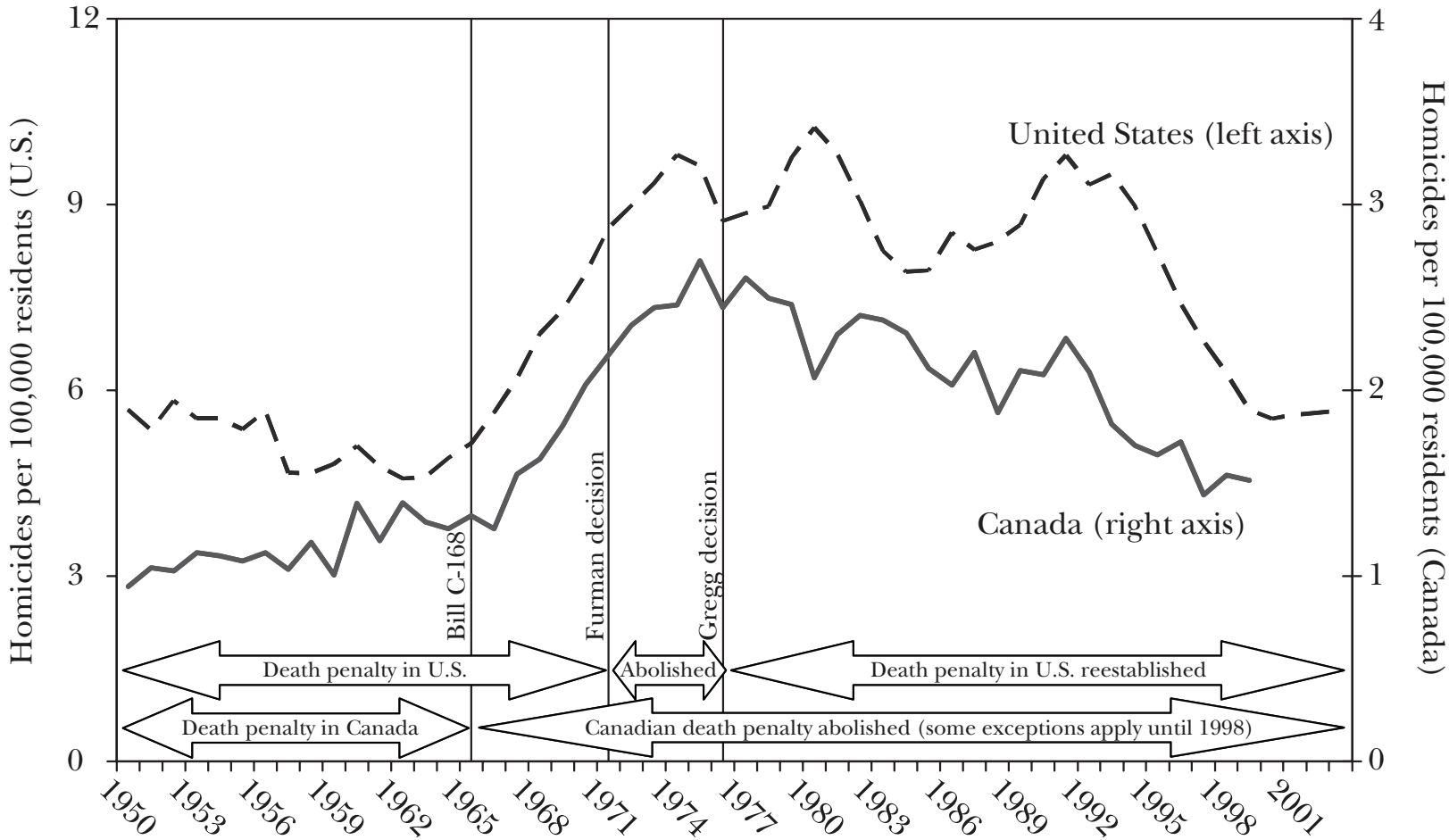
- Compare two groups that shared a trend in an outcome of interest until some treatment/shock (e.g. new law enacted) occurs.
- Then look for diverging trends afterwards.
- "Parallel trend" before treatment interpreted as meaning the trend in the non-treated group can serve as a counterfactual for the treated group.
- Focusing on changes means that time-invariant changes do not matter.

Difference-in-differences

Figure 1

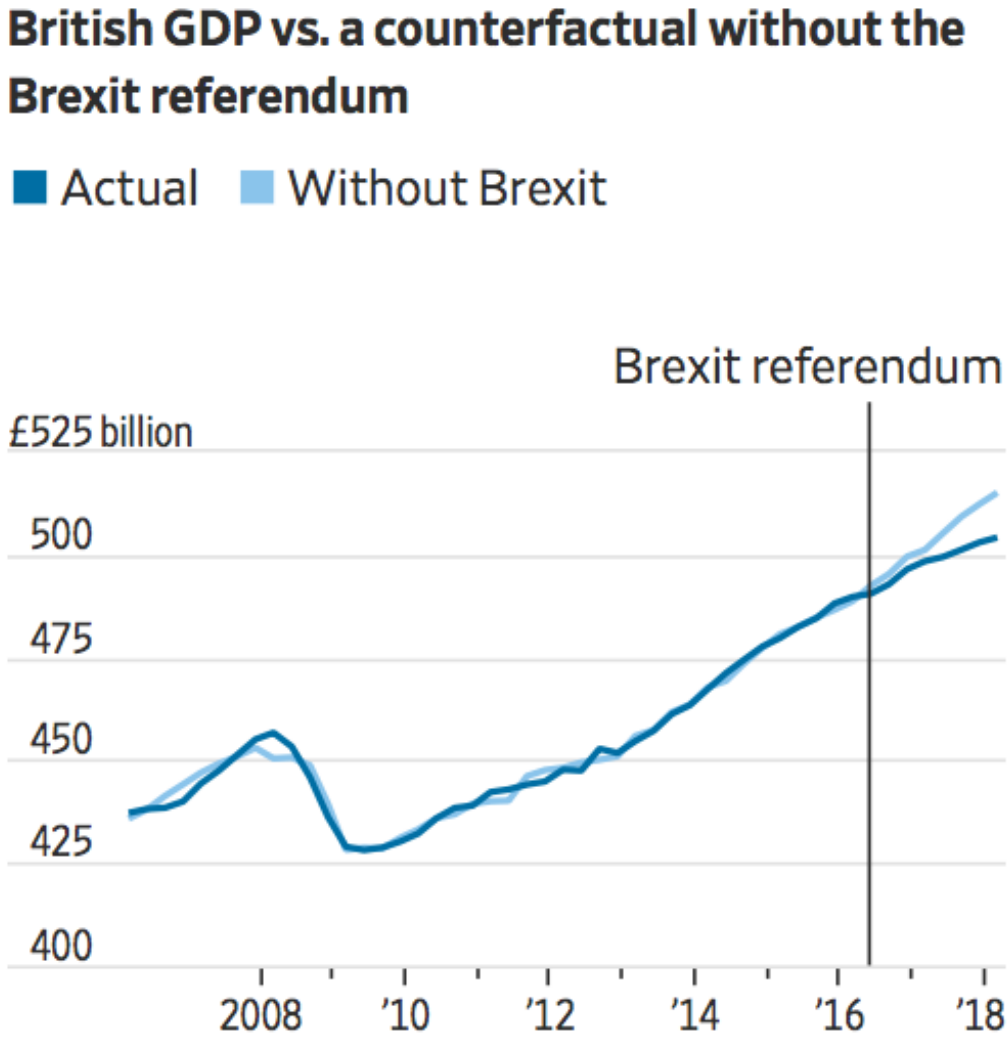
Homicide Rates and the Death Penalty in the United States and Canada

(U.S. and Canada rates on the left and right y-axes, respectively)



Source: Donohue and Wolfers (2005).

DID example



*Counterfactual based on basket of economies whose growth tracked the UK's up to the Brexit referendum
Note: Figures are constant 2016 pounds
Source: UBS

DID example

Figure 6: Mariel Boatlift

DID example

- Mariel Boatlift: 100k+ Cubans are allowed into US, especially Miami (Florida).
- Sudden 7 percent increase in labour force Miami.
- Card (1990) compare before/after Miami with before/after other US cities that did not experience this shock.
- (No impact on wages and unemployment of less-skilled workers.)
  - NB: Card did not use formal DID regression analysis, just compared befores and afters in two groups. Interesting read!

DID example

- See how to implement this in R.
- Contrived example: Dutch municipalities are treated with "mining", as coal mines in Limburg start being exploited from c. 1900 onward.
- Expect mortality to be higher compared to no-mines counterfactual due to e.g. miners' lung.
- Note:
  - We'll see this example does not meet a crucial assumption.
  - Also in reality not a clean treatment (0/1, all at once in 1900).
  - All the technicalities like staggered treatment, conditional parallel trends, etc. are ignored here.
- Maybe there's actually a good DID study of Limburg mining & mortality possible, but this is definitely not it!
- Point is to show how to do these things in `R+fixest`, and in particular how little this method requires.
- Switch to R!

DID example: Limburg mining

- Load packages and read in (old version of) HDNG

```
library("data.table")
library("tinyplot")
library("fixest")

hdng = fread("~/data/hdng/HDNG+prov+missing.txt")
hdng[, dec := round(year, -1)] # decadal data, but decades
hdng[, value := as.numeric(value)]

## Warning in eval(jsub, SDenv, parent.frame()): NAs introc
```

DID example: Limburg mining

- Extract pop, miners, deaths from HDNG, sum to decadal avarage if relevant.

```
mines = hdng[variable_name == "Mijnen, veenderij" & !is.na(ACODE),
    list(miners = sum(value)),
    by = ACODE]
pop = hdng[variable_name == "Bevolking 31-12" & !is.na(ACODE) & is.na(s
    list(pop = mean(value)),
    by = list(ACODE, dec)]
mor = hdng[variable_name == "Sterfte" & !is.na(ACODE) & is.na(sex),
    list(deaths = mean(value)),
    by = list(ACODE, dec)]
```

DID example: Limburg mining

- ID places with substantial mining share in employmenbt.
- 5% eyeballed, and probably too low.

```
mines = merge(mines, pop[dec == 1930], by = "ACODE")
mines[, minershare := miners / pop]
mines[, miningplace := minershare > 0.05]
mining_places = mines[miningplace == TRUE, ACODE]
```

DID example: Limburg mining

- Prep DID dataset by merging in mortality (dependent variable, CDR, probably better w/o inf mort), mining munic dummy, post-treatment period (1900) dummy.

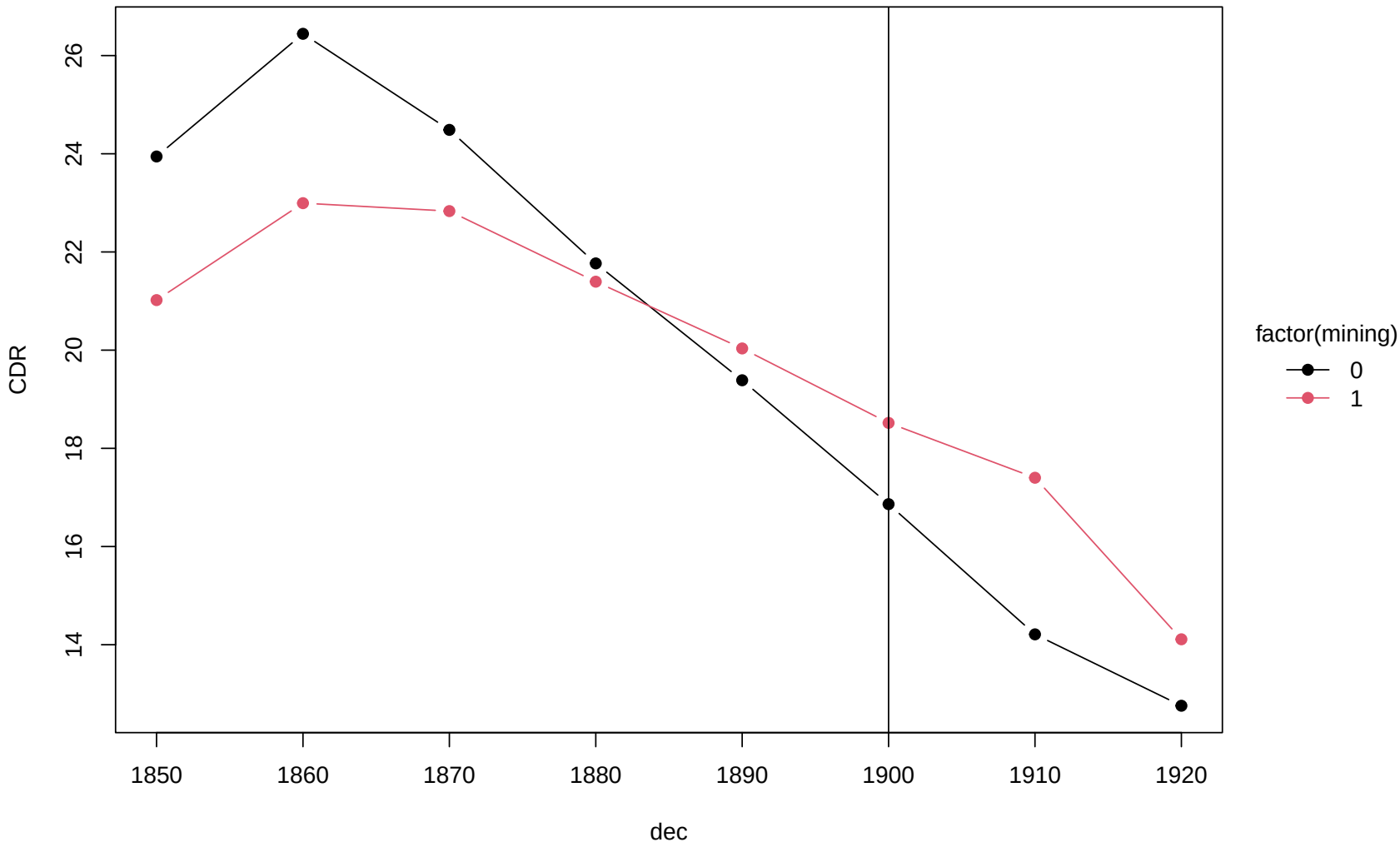
```
dat = merge(pop, mor, by = c("dec", "ACODE"), all = TRUE)
dat = dat[dec <= 1920 & pop > 0]
dat[, mining := as.integer(ACODE %in% mining_places)]
dat[, post := as.integer(dec >= 1900)]
dat[, cdr := (deaths / pop) * 1e3]
```

DID example: Limburg mining I

- Assumption is parallel trends, that is, w/o treatment trends between treated and control would have been the same.
- Weak check you can do: parallel trends before treatment.
- Note: we **fail** this test.
- Possible to have "conditional parallel trends" (so maybe after adjust for urbanisation or health care availability trends are parallel), but is not straightforward and not explored here.

```
toplot = dat[pop > 0, mean(cdr), by = list(dec, mining)]
tinyplot::plt(V1 ~ dec | factor(mining),
    data = toplot,
    ylab = "CDR",
    type = "b", pch = 19)
abline(v = 1900)
```

DID example: Limburg mining II



DID example: Limburg mining I

- Estimate a number of models: naive dummy, simple did, did with ind/place FE, did with TWFE (hairy), did with single FE, clustered SE (preferred)

```
m_nodid = feols(cdr ~ mining, data = dat)
m_did = feols(cdr ~ mining + post + mining*post, data = dat)
m_did_fe = feols(cdr ~ mining + post + mining*post | factor(ACOD
    data = dat)
```

```
## The variable 'mining' has been removed because of collinearit
```

```
m_did_twfe = feols(
    cdr ~ mining + post + mining*post | factor(ACODE) + factor(d
    data = dat)
```

```
## The variables 'mining' and 'post' have been removed because o
```

```
m_did_fe_clus = feols(cdr ~ mining + post + mining*post | factor
    data = dat, vcov = "cluster")
```

```
## The variable 'mining' has been removed because of collinearit
```

DID example: Limburg mining

```
etable(
    list(m_nodid, m_did, m_did_fe),
    digits = 2)
```

```
##                          model 1         model 2         model 3
## Dependent Var.:              cdr             cdr             cdr
##
## Constant          20.2*** (0.07) 23.3*** (0.07)
## mining              -0.40 (0.38) -1.6*** (0.39)
## post                             -8.7*** (0.13) -8.4*** (0.14)
## mining x post                     3.7*** (0.64)  3.5*** (0.27)
## Fixed-Effects:  ------------- ------------- -------------
## factor(ACODE)               No              No             Yes
## _______________ _____________ _____________ _____________
## S.E. type                  IID             IID     by: ACODE)
## Observations             8,942           8,942           8,942
## R2                     0.00012         0.35120         0.62978
## Within R2                   --              --         0.46449
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

DID example: Limburg mining

```
etable(
    list(m_did_twfe, m_did_fe_clus),
    digits = 2)
```

```
##                         model 1         model 2
## Dependent Var.:             cdr             cdr
##
## mining x post   3.5*** (0.26)  3.5*** (0.27)
## post                          -8.4*** (0.14)
## Fixed-Effects:  ------------- --------------
## factor(ACODE)              Yes             Yes
## factor(dec)                Yes              No
## _______________ _____________ ______________
## S.E.: Clustered    by: ACODE)      by: ACODE)
## Observations            8,942           8,942
## R2                    0.71566         0.62978
## Within R2             0.00744         0.46449
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

DID example: Limburg mining

- We would conclude that CDR in mining municipalities is 3–4 deaths per 1000 inhabitants higher than it would have been if mining had never started.
- This is a nice, direct, causal conclusion!
- But note: parallel trends assumption does not seem to hold. Graph and DID logic suggests this would roughly have been the difference regardless of mining.
- Key point is to show how little we need in terms of data and modelling.
- If PT held we would have had a causal effect estimate using nothing more than OLS and the HDNG.

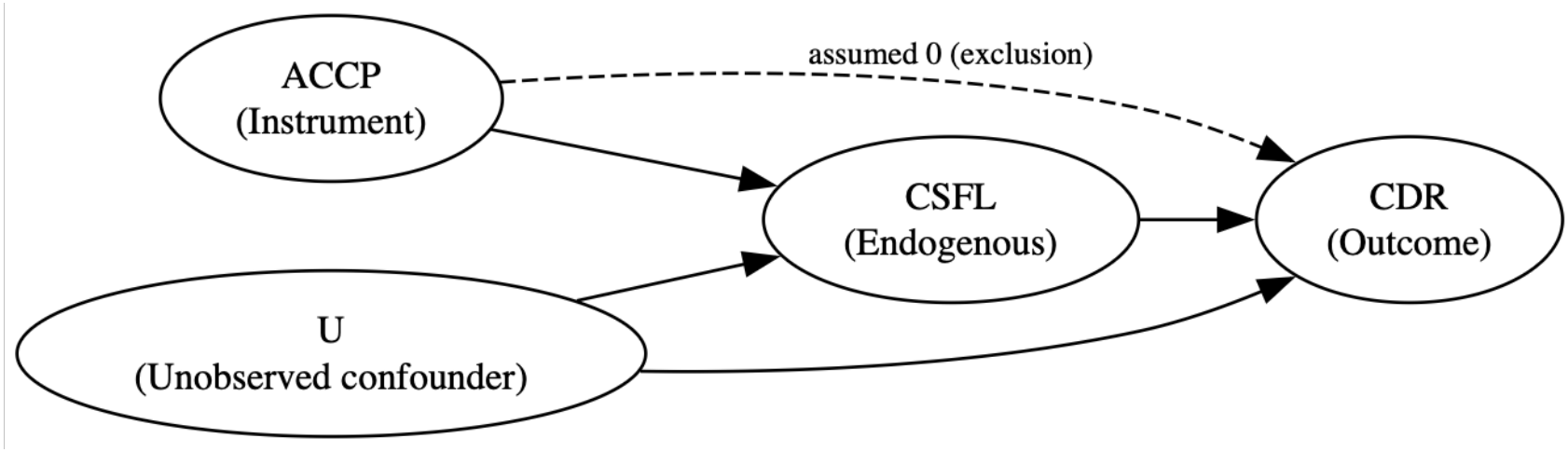
Instrumental variables

- Instrumental variables harder to understand IMO, but very popular.
- Rather than find direct randomisation of "treatment", use "quasi-randomisation" of something that influences treatment, but nothing else (directly).
- Indirect randomisation, if you will.
- Think of it as allowing you to "change only the variable of interest" as if it was an experiment.

Java example

- Back to the Java example.
- What if high-mortality residencies are worked harder to meet desired output?
- What if low-mortality residencies are worked harder because they are more productive?
- Measurement error in forced labour?
- Instrument forced labour with Amsterdam coffee and sugar prices:
  - Higher prices give incentive to increase production (can test).
  - Assumed that Amsterdam prices do not directly impact CDRs.
  - If these hold, we can get a causal effect of forced labour on mortality.

IV for Java


ACCP
(Instrument)
assumed 0 (exclusion)
CSFL
(Endogenous)
CDR
(Outcome)
U
(Unobserved confounder)

IV for Java

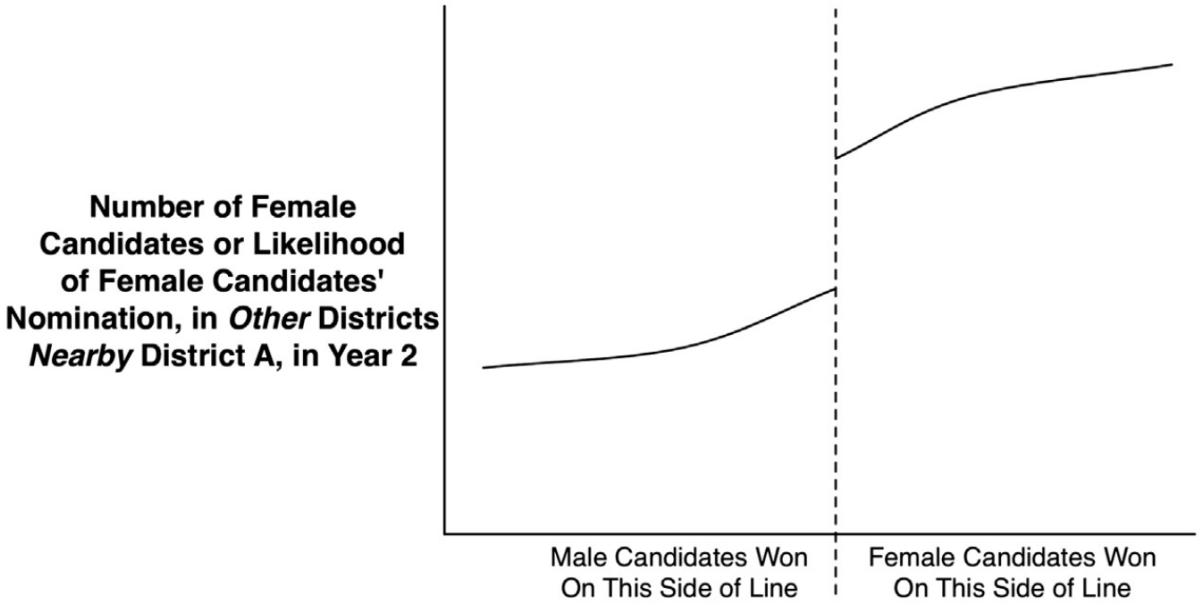
TABLE 5
THE IMPACT OF THE CULTIVATION SYSTEM ON CRUDE DEATH RATES
(IV ESTIMATES)

	(1)	(2)	(3)
Panel A: Fixed Effects			
log forced labor (β)	0.270***	0.238***	0.329***
	(0.075)	(0.073)	(0.098)
Panel B: Second Stage			
log forced labor (β_{IV})	0.590***	0.667**	0.647**
	(0.176)	(0.272)	(0.324)
Panel C: First Stage (dep. var. log forced labor)			
log Amsterdam prices (α)	–0.428***	–0.359***	–0.286***
	(0.080)	(0.076)	(0.108)
F-statistic on excluded instrument	28.86	22.18	7.01
Province- and year-fixed effects	Yes	Yes	Yes
Buffalos	Yes	Yes	Yes
Rice price	No	Yes	Yes
Crop payments	No	No	Yes
Officials	No	No	Yes
Observations	473	407	407

Regression discontinuity design

- Regression discontinuity designs are extremely popular, for economic history in particular since the Dell paper on the Mita.
- RDD exploits fact that sometimes small change in a variable of interest pushes subjects in another category.
- At such small difference in variable of interest, surely groups on both sides of dividing line are similar.

RDD examples

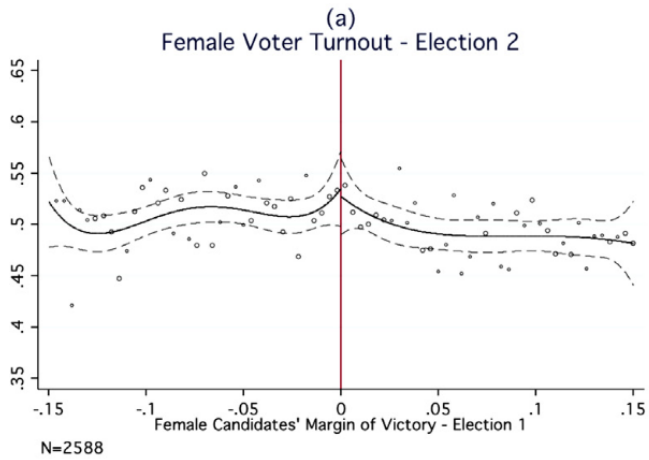


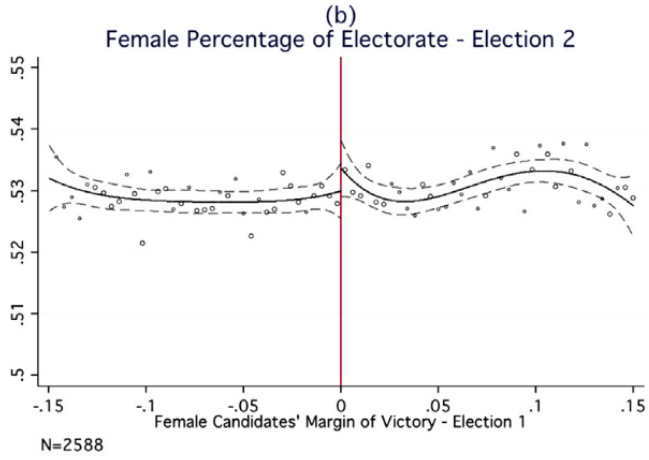
Notes: The Figure visualizes how a regression discontinuity design would discover the effect of a woman's victory on other women's candidacies in subsequent elections. The X-axis describes a woman's share of the two party vote in a district at Time 1; women won races to the right of the discontinuity. Even though the overall relationship is endogenous, the difference between the estimates of the data's true underlying form at the limit captures the causal effect of electing a woman in other districts near district A at Time 2.

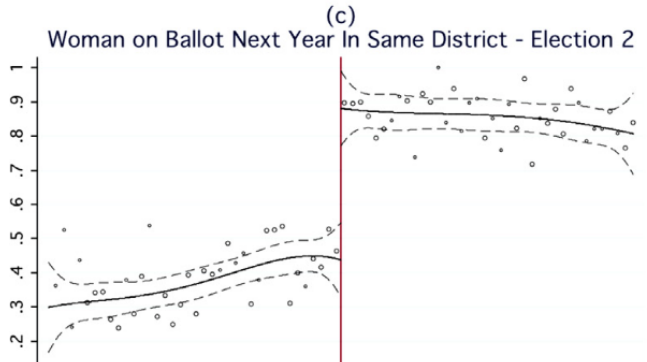
Fig. 1. Hypothesized "breaking of the glass ceiling" pattern.

Figure 7: Brookman, "Do female politicians empower women to vote or run for office? A regression discontinuity approach"

RDD examples







RDD examples

- Example: impact of female electoral victories on women's political participation.
- Issue is that kind of districts that elect women are probably very different from other districts in many ways, and we will probably not have data on all of them.
- RDD solution (informally): compare close elections
- Would not expect districts voting 49.7 % for woman to be systematically different from those voting 50.3 %. This difference is arguably up to chance.
- Very popular variant: two sides of (historical) border.
- Comes with lots of technicalities though.

Conclusion

Conclusion

- Hope I've showed you some basics of the regression model.
- And some important extensions that might be of interest in applied research.
- Huge bit in the middle missing though!
  - OLS diagnostics
  - Variable transformations & interpretation
  - Mechanics of OLS models.
  - Time series
  - Robust and clustered standard errors