



Stockholm
University

Pär Nyman

Methods in Political Science
30 March 2026

Control variables in applied research

Multiple regression, variable selection and causality

Introduction

The structure of today's lecture

- Different purposes of adding variables.
- How does multiple regression work?
- Selecting control variables.
 - Selection criteria.
 - Trade-offs.
 - Special cases.
- Judging bias/causality.
 - The causal dichotomy.
 - Causal inference with observational data.

Closely matches Section 3 and Section 4 in my text.

Introduction

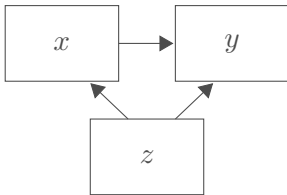
Reasons to include additional variables

1. Control for confounders
2. Mediation analysis
3. Improve precision
4. Better explanation/prediction

Introduction

Potential confounders

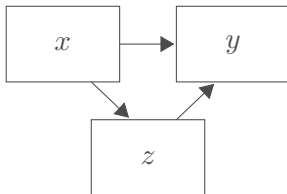
- Included to reduce omitted variable bias.
- Affect both the independent variable of interest (x) and the dependent variable (y).
- The focus of today's lecture. Sometimes these are the only ones referred to as 'control variables'.



Introduction

Variables that mediate the effect

- Included to test a channel through which x affects y .
- Are affected by x and affect y .
- If you include them with the purpose of reducing bias, Angrist and Pischke call them ‘bad controls’.



Introduction

Variables that improve precision

- Explain variation in y that is not explained by x .
- We seek variables that are more strongly correlated with (the unique variation in) y than with (the unique variation in) x and that are not affected by x .
- Common in experiments.

Introduction

Not always sufficient to study a single effect

- Making better predictions.
- Comparing effects between variables.
- Interested in several explanations, or the joint explanatory power of a set of predictors.

How does it work?

Controlling for z as holding it constant

- Controlling for z is often described as holding z constant.
- Easy to use when interpreting regression results.
- Intuitively understood as comparing observations that have the same values of z .
- Two potential limitations with this understanding:
 - Works best with a single categorical control variable.
 - Not fully satisfactory explanation of how it works? How can we hold something constant without an experiment?

How does it work?

My preferred way: unique/residual variation

- When estimating slope coefficients, we only use the unique variation in each independent variable.
- I.e., we take away the variation which is expected given the values on other x , and only use the deviations from what we should expect given the other variables.
- This is not exactly how Stata does it, but it has been my key to understanding multiple regression.

How does it work?

Assume we are interested in the effect of x while controlling for z :

$$y = a + b_1x + b_2z + e$$

We could keep the residuals (e^y and e^x) from these regressions:

$$y = a + bz + e^y \quad \text{and} \quad x = a + bz + e^x$$

And then regress the residual of y on the residual of x :

$$e^y = a + be^x$$

b_1 in the first equation would be the same as b in the last.

How does it work?

- * Open a dataset over car sales

```
sysuse auto, clear
```

- * Regress weight on length and store the residuals

```
reg weight length
```

```
predict weight_res, res
```

- * Regress price on length and store the residuals

```
reg price length
```

```
predict price_res, res
```

- * Note identical slope coefficients for weight

```
reg price weight length
```

```
reg price_res weight_res
```

How does it work?

Things I like with this way of understanding controls

- It works just as well for multiple and continuous control variables.
- It shows how model-dependent the control variable approach is.
- It helps us think about multicollinearity* in terms of how much variation we have left.

* The degree to which the independent variables are correlated with each other. As it is the residual variation which determines precision we should avoid rules of thumb for 'too much' multicollinearity and variable transformations which reduce multicollinearity without increasing the residual variation.

Selecting control variables

Criteria for suitable controls

- Potential confounder.
- Exogenous to x and y ($z \rightarrow x$ and $z \rightarrow y$).
- Not too strongly correlated with x . But ...
- Data quality (observations, variation, validity).

Selecting control variables

Multiple trade-offs

- A variable rarely meets all criteria.
- Controlling for a variable at the cost of
 - Worsened precision.
 - Risk of z being endogenous to x (bias).
- Worsened OVB is always a risk even when criteria are met.
 - The remaining variation may be more susceptible to OVB.
 - We may be removing bias with opposite sign.
- As a result, we don't know the best model.
- We handle this by evaluating a set of specifications.

Selecting control variables

Special cases guided by structure of data

- Time controls.
- Lagged dependent variable.
- Fixed effects.

Sometimes these are referred to as models.

We will talk about these in L7, but FE is good to hear twice.

Fixed effects

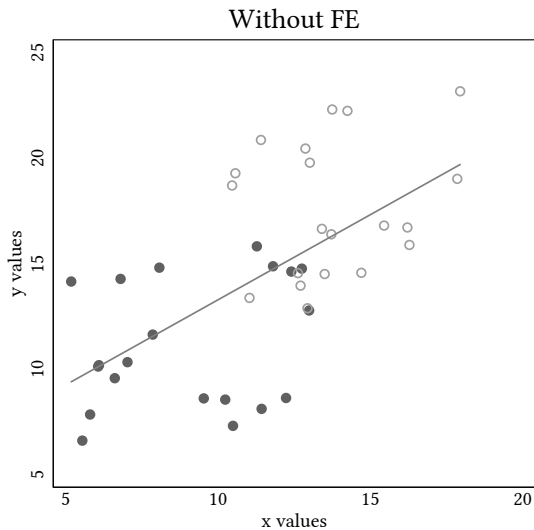
- With clustered data, there is variation *between* as well as *within* clusters.
- If we only want to use the within-variation we can use fixed effects.
 - Maybe less sensitive to omitted variable bias?
 - Less variation may hurt precision.
- Using x -fixed effects is equivalent to controlling for x .
- Common to present results with and without fixed effects.

Fixed effects

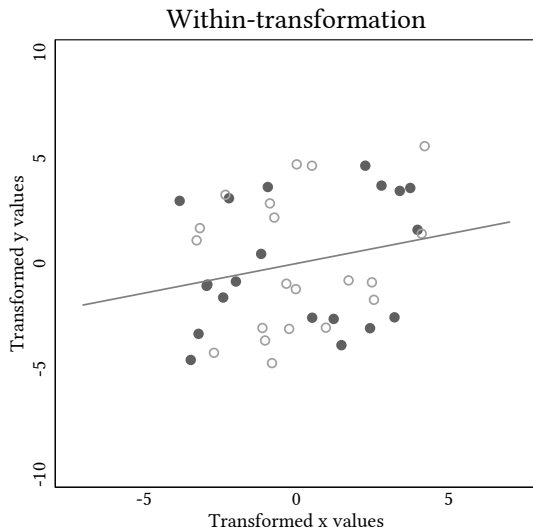
Two ways of conceptualizing fixed effects

- Within-transformation.
 1. Calculate cluster means for all variables.
 2. Subtract the mean from each variable (i.e., remove the between-cluster variation).
 3. Conduct the regression on the transformed variables.
- Dummy variables.
 1. Add a binary indicator per cluster to the regression model (i.e., control for the between-cluster variation).

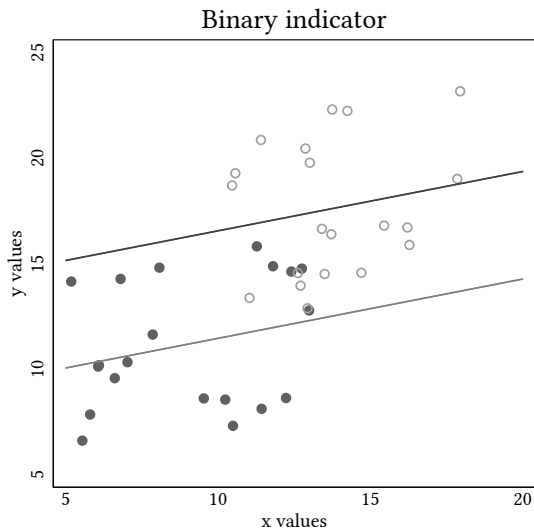
Fixed effects



Fixed effects



Fixed effects



Causal inference

The causal revolution

- Strong trend towards more experiments.
- The causal revolution is reflected in the focus of this course.
 - The problem of violating the exogeneity assumption.
 - Techniques like controlled experiments, RD, IV, etc.
 - Possible criticism.

Causal inference

Today we will discuss these aspects

- How should we view the difference between experimental and other studies?
- When is it reasonable to draw causal inference from observational/traditional/non-experimental studies?
- These are my views.

Causal inference

- The naive researcher
 - Thinks that a set of control variables in general is sufficient to draw causal inference.
 - Puts too much weight into less important aspects (sample size, model choice, etc.)
 - Example: the research about lifestyle and health that you read in the newspaper.
- The binary researcher
 - Sees everything that is not design-based as descriptive.
 - Believes that everyone who does not do design-based inference reasons like the naive researcher.
 - Example: many economists and RCT-fundamentalists.

The dichotomy

- If you take a course or read a textbook in causal inference, it will solely be about controlled and natural experiments.
- Why the binary division into design-based/experimental and traditional/observational studies?
 - They are different with regards to language, tools and paper structure.
 - But the estimate from an experimental study is not necessarily closer to the effect of interest.

Note that natural experiments fall into the first category, even if they are observational in strict sense.

The dichotomy

Unique characteristics of design-based/experimental

- The argument is that the exogeneity assumption holds.
- Specific techniques to validate that assumption.
 - Case-knowledge.
 - Balance tests.
 - Placebo tests.
- Control variables are generally not required (or limited to variables required by design).
- Successful experimental studies allow *unbiased* effect estimates (well, no bias from exogeneity violations).

The dichotomy

But both groups of studies can answer causal questions

- Our estimates are only that – estimates.
- Sampling errors are estimated by the standard errors.
- Bias is arguably more problematic.
 - Difficult to quantify.
 - Cannot be solved by adding data or averaging over studies.
- IMHO, that is not a reason to equal causal with unbiased.
 - Conceptually more reasonable to keep them separate.
 - A slightly biased estimate may be preferable to an imprecise unbiased estimate.
- But (estimated) causal effect has come to mean unbiased.

The dichotomy

How does concurrent elections affect voter turnout in second-order elections?

- A student of mine once compared turnout in Norway and Sweden (in a more reflected way than presented here).
- In Sweden 2018 turnout was 3 p.p. lower in local elections than in national elections.
- In Norway 2017/2019 the difference was 13 p.p.
- His conclusion: Norway could increase local turnout by 10 p.p. by having them the same day as national elections.

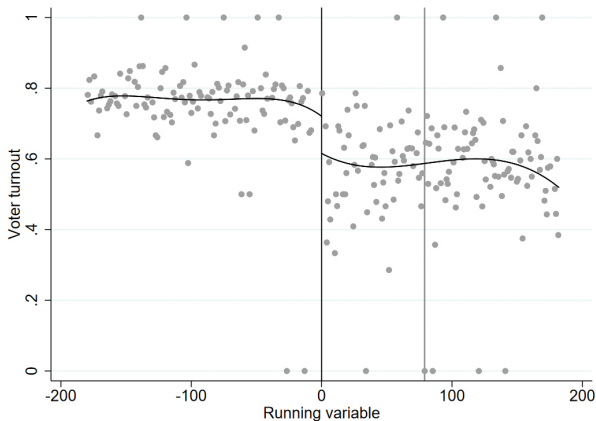
The dichotomy

Mattila (2003) regresses the turnout in EU elections on relevant variables. His estimate: 10–11 percentage points.

Compulsory voting	9.94*** (2.45)	9.58*** (2.51)
Concurrent elections	11.22*** (2.28)	10.10*** (2.75)
Time until next elections	–	–0.06 (0.08)
Weekend voting	13.67*** (2.40)	13.48*** (2.42)
Net contributor	–8.09*** (2.45)	–7.65*** (2.52)
Net beneficiary	3.65 (2.30)	3.55 (2.32)
Public support for the EU	0.15*** (0.04)	0.14*** (0.04)
Constant	–41.71*** (2.58)	–40.19*** (3.29)
Adj. R^2	0.81	0.81

The dichotomy

Aggeborn et al (2020) use an RD design with the days until citizenship as running variable. They estimate that national eligibility increases turnout in local elections with 10–20 p.p.



The dichotomy

Takeaway

- These designs are not equally strong, but even in the weakest design it is difficult to come up with an alternative explanation.
- To me, they all provide reasonable answers – effect estimates – to a causal question.
- The RD study is an important contribution, but it does not transform the field.

The dichotomy

This is not always the case

- New RQ: the effect of education on turnout.
- Very easy to come up with possible confounders.
- Design-based studies have struggled to find effects and we no longer trust the traditional studies.
- Here it is difficult to sell an observational study as anything but descriptive.

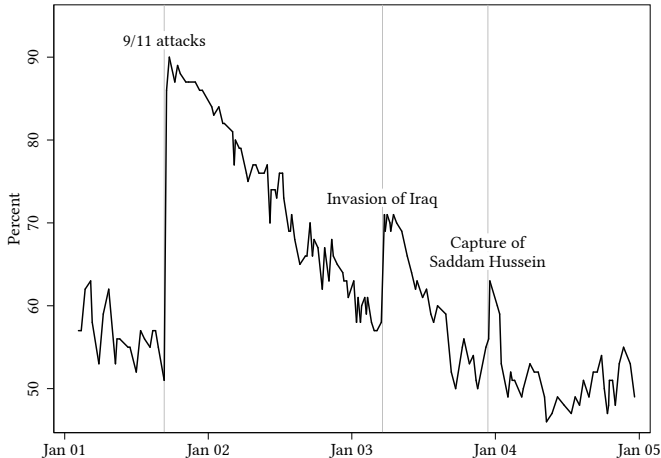
Causal inference without experiments

How do we know?

- Some variables are exogenous.
 - The difference between a descriptive and causal interpretation is often purely semantic.
- Sometimes we can argue that the bias is small in relation to the effect. My criteria:
 - Plausibility of the effect in relation to alternative explanations (yes, a subjective evaluation).
 - Little residual variation.
 - Outcome reacting quickly to sudden changes in explanatory variable.

Causal inference without experiments

Rally around the flag—effects on Bush approval



Causal inference without experiments

How do we know? Continued.

- We need to rely on controls when we suspect non-negligible omitted variable bias.
- When can control variables be convincing?
 - Plausibility of proposed effect.
 - A clear idea of why the identifying variation is reasonably exogenous or unpredictable.
 - Some variables are impossible to isolate.
 - The more coincidental the better.
 - Personally I like to rely on timing.
- Assumptions are less clear than in design-based studies and personal judgement is more important. Hence, it may be difficult to convince a sceptic.

Causal inference without experiments

Expressing the uncertainty

- Causal effect
- Plausibly causal.
- The regression coefficient.
- Descriptive statements.
- Under the assumption ...

Thanks for listening!

Extra material

How can we visualize multiple regression?

- The relation between Y and X can be represented by a line.
- With two (continuous) independent variables we get a plane.
- I will not show how you can create that in Stata.
 - Two user-written packages can produce a still shot of a 3d graph: `graph3d` and `surface`. You need Stata 16, Python and several extra packages to make rotating 3D graphs.
 - Showing 3D graphs in a 2D environment is rarely the best way of making a point.
 - The 3D visualization does not help as we add more variables?
- But *click here* for someone else's example.
- The constant adjusts the height of the plane and fit is about the distance between the observations and the plane.

Extra material

A 2D graph works if one variable is categorical

