Does education increase civic engagement?

11/28/22

Social scientists have long been interested in the causal effects of education. We've seen a ton of examples of economists looking at the causal effect of education on wages or income. Political scientists, not surprisingly, are less interested in income and more interested in the effect of education on civil behavior. For instance, does education make people vote more?

For this example, we'll use a subset of data from Dee (2004) (preprocessed and cleaned by Miller (2021)) to explore whether college education causes people to register to vote. (See another version of this at Steve's website)

- Treatment = `college`: A binary variable indicating if the person attended a junior, community, or 4-year college by 1984
- Outcome = `register`: A binary variable indicating if the person is currently registered to vote
- Instrument = `distance`: Miles from respondent's high school to the nearest 2-year college

```
library(tidyverse)
library(haven)
library(broom)
library(ggdag)
library(estimatr)
library(fixest)
library(modelsummary)

voting <- read_stata("data/Dee04.dta")
```

Exploratory data analysis

What proportion of college attendees are registered to vote? Group by `register` and `college`, summarize to get the number or rows in each group, then add a column that calculates the proportion.

```
# TODO
```

Visualize these proportions with `ggplot()` and `geom_col()`

```
# TODO
```

Naive model

Run a super wrong and naive model that estimates the effect of college attendance on voter registration (`register ~ college`).

```
# TODO
```

For fun, make a scatterplot of this relationship:

```
# TODO
```

TODO: Interpret this finding. Why is this estimate wrong?

Distance to college as an instrument

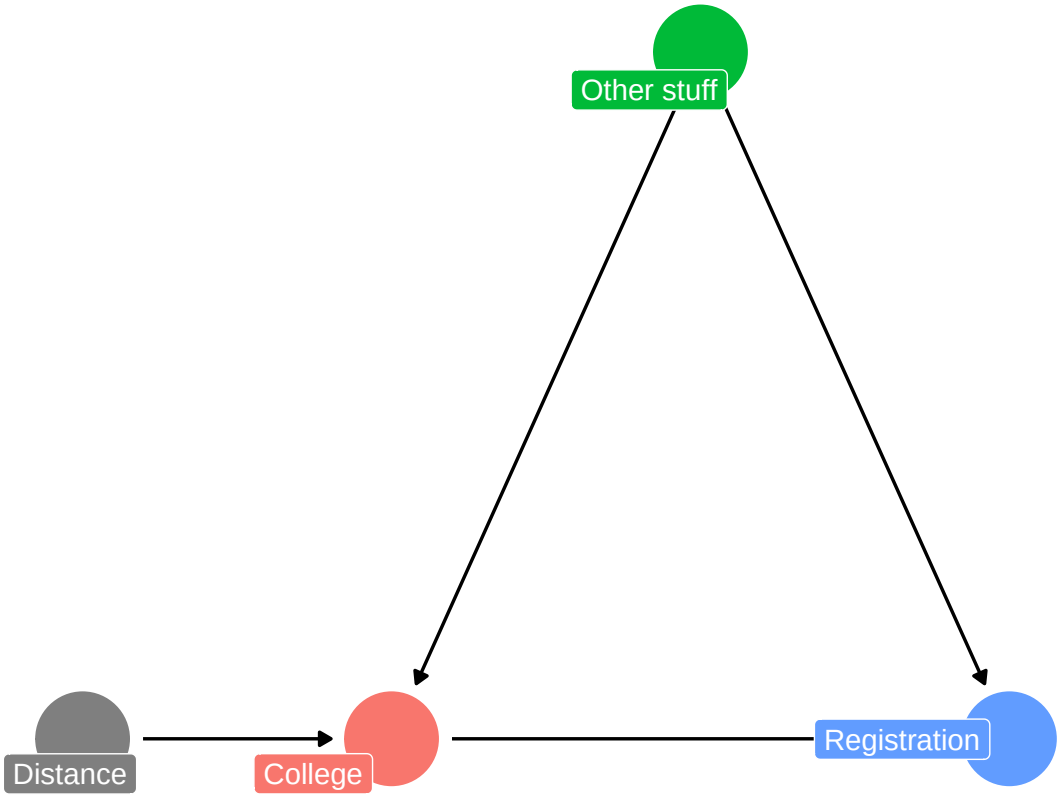
We're stuck with endogeneity. There are things that cause both education and voter registration that confound the relationship, and we can't control for all of them.

In his paper, Dee (2004) uses distance to the nearest college as an instrument to help remove this exogeneity. He essentially creates this DAG (though without actually making a DAG):

At first glance, this feels like it could be a good instrument:

1. **Relevance** (Z $\rightarrow$ X *and* cor(Z, X) 0): Distance to college should be associated with college attendance. The closer a college is, the cheapter it is to attend, and the more opportunity there is to attend.
2. **Excludability** (Z $\rightarrow$ X $\rightarrow$ Y *and* Z not $\rightarrow$ Y *and* cor(Z, Y | X) = 0): Distance affects college attendance which affects voting registration. But distance should influence voting registration *only because* people go to college (and no other reason).
3. **Exogeneity** (U not $\rightarrow$ Z *and* Cor(Z, U) = 0): Colleges should exist before students exist; students and their voting patterns don't influence whether pre-existing colleges exist (i.e. there should be no arrows going into the instrument node).

Let's check these conditions


Other stuff
Distance
College
Registration

Relevance

See if there's correlation between the instrument (`distance`) and the treatment (`college`). Use the `cor()` function (and `cor.test()` if you want a p-value):

```
# TODO
```

Plot the relationship between the instrument and treatment:

```
# TODO
```

TODO: Is this relevant?

Exclusion

See if there's a relationship between the instrument (`distance`) and the outcome (`register`).

```
# TODO
```

To help check the "only through" condition, see if there's a relationship between `distance` and other possibly confounding variables (like `black`, `female`, `hispanic`, and so on). If it's related, that's a good sign that there's an arrow between those nodes, thus breaking the exclusion requirement.

```
# TODO
```

TODO: Does this meet the exclusion requirement?

Exogeneity

There's no statistical test here. Instead we have to tell a theoretical story that distance is uncorrelated with anything else in the DAG.

TODO: Does this meet the exogeneity requirement?

IV estimation

Let's pretend that this is a good instrument, regardless of whatever we concluded above.

By hand, to make your life miserable

Make a first stage model that predicts college attendance based on distance to college. Control for `black`, `hispanic`, and `female`, since they're potential confounders (and because Dee originally did that too).

```
# TODO
```

Plug the original dataset into the first stage model with `augment_columns()` to generate predicted values of `college` (or the exogenous part of `college`):

```
# TODO
```

Make a second stage model that estimates the effect of *predicted* college on voter registration, also controlling for `black`, `hispanic`, and `female`.

What is the causal effect of attending college?

```
# TODO
```

All at once, to make your life wonderful

Use `iv_robust()` to run a 2SLS model all at the same time:

```
# TODO
```

Causal effect

TODO: What is the final causal effect of college attendance on registering to vote?

```
# TODO: Use modelsummary() to show the results from your manual and automatic 2SLS regress
```

References

Dee, Thomas S. 2004. "Are There Civics Returns to Education?" *Journal of Public Economics* 88 (9–10): 1697–1720. https://doi.org/10.1016/j.jpubeco.2003.11.002.

Miller, Steve. 2021. *Stevedata: Steve's Toy Data for Teaching about a Variety of Methodological, Social, and Political Topics*. https://CRAN.R-project.org/package=stevedata.