Problem Set 4: Instrumental Variables

Gov 51 — Spring 2026

2026-04-17

Reading: Angrist & Pischke, *Mostly Harmless Econometrics*, Chapter 4.1 (IV and the Wald Estimator). Refer to lecture slides from Week 12 as needed.

Data: `card1995.csv` — individual-level data from Card (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp.*

Submission: Push your completed `.qmd` file and rendered PDF to your GitHub repository by 11:59 PM on April 17.

Packages: You will need `tidyverse`, `estimatr`, and `modelsummary`. Install with `install.packages(c("estimatr", "modelsummary"))` if needed.

Context. Does more education raise wages? It seems obvious — but it's harder to answer causally than it looks. People who get more education might also be more motivated, have wealthier parents, or have higher innate ability. An OLS regression of wages on education captures all of these factors together. To isolate the *causal* effect of education, we need variation in education that is unrelated to these confounders.

David Card's insight: **geographic proximity to a college**. If you grew up near a 4-year college, it was cheaper and easier to attend — reducing the effective cost of education. Proximity to a college is not something you chose based on your ability or motivation; it depended on where your family happened to live. Card uses this as an instrumental variable.

In this problem set, you will replicate the core IV analysis from Card (1995), learning the mechanics of instrumental variables from the ground up.

---

1 Setup and Exploration (10 points)

1.1 Load packages and data

Load the `tidyverse`, `estimatr`, and `modelsummary` packages. Then load the dataset from `data/card1995.csv`.

```
library(tidyverse)
library(estimatr)
library(modelsummary)

card <- read_csv("data/card1995.csv")
```

Variable guide:

Variable	Description
`lwage`	Log weekly wage (outcome)
`educ`	Years of completed education (endogenous variable)
`nearc4`	1 = lived near a 4-year college at age 14 (instrument)
`black`	1 = Black
`smsa`	1 = lived in a metropolitan area
`south`	1 = lived in the South
`exper`	Potential labor market experience (years)
`expersq`	Experience squared
`married`	1 = married

1.2 Describe the data

How many observations are in the dataset? What is the unit of observation?

1.3 Distribution of years of education

Create a histogram of `educ`. What is the mean? Are there any unusual spikes? What might explain them?

1.4 The naïve relationship

Create a scatter plot of `lwage` (y-axis) against `educ` (x-axis). Add a linear regression line. Based on this plot, does more education appear to be associated with higher wages?

1.5 Why OLS might be misleading

The OLS regression of `lwage` on `educ` might not give us the causal effect of education. Name two specific reasons why people with more education might earn more *even if education itself had no effect*.

---

2 The Instrument: College Proximity (25 points)

2.1 What the instrument measures

The variable `nearc4` equals 1 if the respondent grew up near a 4-year college. What share of respondents in this sample grew up near a 4-year college?

2.2 Relevance: does proximity affect education?

The **first condition** for a valid instrument is *relevance*: the instrument must be correlated with the endogenous variable (education).

Run the following regression — the **first stage** — and report the coefficient on `nearc4` along with its standard error and t-statistic.

```
first_stage <- lm_robust(educ ~ nearc4 + black + smsa + south +
                           exper + expersq + married,
                         data = card, se_type = "HC2")
```

Interpret the coefficient on `nearc4` in a sentence. Does the sign make sense?

2.3 Checking instrument strength

Report the t-statistic on `nearc4` from the first-stage regression. Compute the **first-stage F-statistic** as $F = t^2$.

Is your instrument strong? (Rule of thumb: F $>$ 10 suggests a strong instrument.) Why does instrument strength matter for the reliability of the 2SLS estimate?

2.4 Exclusion restriction

The **second condition** for a valid instrument is the *exclusion restriction*: college proximity must affect wages **only through** its effect on education — not through any other channel.

(a) State the exclusion restriction formally: what must be true about the relationship between `nearc4` and `lwage` once we control for education?

(b) Can you think of a way the exclusion restriction might be violated — that is, a reason why growing up near a college might affect your wages for reasons other than more education? Be specific.

2.5 Balance check

If the instrument is as-good-as-randomly assigned, respondents who grew up near a college should look similar on *pre-treatment* characteristics to those who did not.

Compute the mean `black`, `smsa`, and `south` for the two groups (`nearc4 == 0` and `nearc4 == 1`). Are the groups balanced? What does any imbalance tell you about the instrument?

---

3 The Wald Estimator (15 points)

The simplest IV estimate is the **Wald estimator**: the reduced-form effect divided by the first-stage effect.

$$\hat{\tau}_{\text{Wald}} = \frac{\text{Reduced Form}}{\text{First Stage}} = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}$$

3.1 Compute the reduced form

Run the **reduced form** regression: regress `lwage` directly on `nearc4` (without `educ`, but including the same controls as the first stage). Report the coefficient on `nearc4` and interpret it: what does it tell us about how proximity to a college relates to wages?

3.2 Calculate the Wald estimator by hand

Using the first-stage coefficient from Section 2.2 and the reduced-form coefficient from Section 3.1, compute the Wald estimator by hand. Show your calculation.

$$\hat{\tau}_{\text{Wald}} = \frac{[\text{reduced form coeff on nearc4}]}{[\text{first stage coeff on nearc4}]}$$

3.3 Interpret the Wald estimator

Interpret your Wald estimate in a sentence. What does it say about the effect of one additional year of education on wages?

---

4 OLS vs. 2SLS (35 points)

4.1 OLS estimate

Run OLS of `lwage` on `educ` and the controls. Report the coefficient on `educ`.

```
ols <- lm_robust(lwage ~ educ + black + smsa + south +
                   exper + expersq + married,
                 data = card, se_type = "HC2")
```

4.2 2SLS estimate

Run the 2SLS regression using `iv_robust()` from `estimatr`. The formula syntax separates the structural equation (left of `|`) from the first-stage instruments (right of `|`):

```
iv <- iv_robust(lwage ~ educ + black + smsa + south +
                  exper + expersq + married |
                  nearc4 + black + smsa + south +
                  exper + expersq + married,
                data = card, se_type = "HC2")
```

Report the coefficient on `educ`. Does it match your Wald estimate from Section 3.2?

4.3 Side-by-side comparison

Display both the OLS and 2SLS estimates in a single regression table using `modelsummary()`. Keep only the rows for `educ` and the model statistics (N, R^2).

```
modelsummary(list("OLS" = ols, "2SLS" = iv),
             coef_map = c("educ" = "Years of Education"),
             gof_map = c("nobs", "r.squared"))
```

4.4 Why do OLS and 2SLS differ?

The OLS and 2SLS estimates of the return to education are likely to differ. Answer the following:

(a) Which is larger — OLS or 2SLS?

(b) Card (1995) interprets the finding that 2SLS > OLS as evidence of **downward bias** in OLS. What could cause OLS to *underestimate* the return to education? *(Hint: think about measurement error in* `educ`*, or about who the compliers are.)*

(c) Which estimate is more credible as a *causal* estimate of the effect of education on wages? Explain briefly.

4.5 Adding IQ as a control

The dataset includes `IQ` scores for a subset of respondents.

(a) Run OLS of `lwage` on `educ`, `IQ`, and all other controls. How does including IQ change the OLS coefficient on `educ`?

(b) Now run 2SLS including `IQ`. Does the 2SLS estimate change substantially when `IQ` is added?

(c) What does the comparison between part (a) and part (b) suggest about the source of the discrepancy between OLS and IV?

---

5 Interpretation and LATE (15 points)

5.1 Who are the compliers?

In a randomized experiment, the ATE averages over *everyone*. But IV estimates a **Local Average Treatment Effect (LATE)** — the average effect for **compliers**: units whose treatment status changes when the instrument changes.

In this context, who are the compliers? Describe them in concrete terms: what type of person's educational attainment was increased because they happened to grow up near a 4-year college?

5.2 Write the LATE interpretation

Write a single careful sentence interpreting the 2SLS coefficient as a LATE. Your sentence should specify:

(i) The causal effect being measured
(ii) For which specific group of people
(iii) What source of variation identifies it

5.3 External validity

Card finds a 2SLS estimate around 12–13% return to each additional year of education. Is this LATE likely to be the same as the ATE (the return for the average person in the sample)?

Think about whether the compliers — men who went to more school because they happened to live near a college — are likely to have *higher* or *lower* returns to education than the typical person. Explain your reasoning.

Submission checklist:

- ☐ All code chunks render without errors
- ☐ Regression tables are legible
- ☐ Interpretations are in complete sentences
- ☐ You answered all sub-parts (a), (b), (c) where applicable
- ☐ Push `.qmd` and rendered `.pdf` to GitHub by 11:59 PM April 17