Assignment 3: Instrumental Variable Estimation

Replicating the IV strategy in Stokes (2015)

EDS 241 / ESM 244 (DUE: 2/17/2026)

2026-02-16

Assignment instructions:

Working with classmates to troubleshoot code and concepts is encouraged. If you collaborate, list collaborators at the top of your submission.

All written responses must be written independently (in your own words).

Keep your work readable: Use clear headings and label plot elements thoughtfully (where applicable).

Submit both your rendered output and the Quarto (`.qmd`) file.

Assignment submission (YOUR NAME): Leela Dixit

Introduction

In this assignment, you will replicate the instrumental variable analysis from Stokes (2015), which examined how local wind turbine projects influenced electoral outcomes. Building on the matched dataset from Assignment 2, we will use Two-Stage Least Squares (2SLS) to estimate the causal effect of having a wind turbine proposed nearby on the change in Liberal Party vote share between 2007 and 2011. The instrument used in Stokes (2015) is a measure of local wind resource (average wind power, logged), which predicts where wind turbines are proposed. By using this instrument, we aim to isolate the portion of variation in turbine placement that is as-good-as-random, helping to meet the assumptions for causal identification.

Study: Stokes, 2015 – Article

Data source: Dataverse – Stokes, 2015 replication data

> Note: The estimates you obtain may not exactly match the published results in Stokes (2015) due to the alternative matching procedure used for processing the data in the previous assignment. Estimates should approximate the findings reported in *Table 2* of the article.

Load packages

```
library(tidyverse)
library(janitor)
library(here)
library(jtools)   # for export_summs (pretty regression tables)
library(AER)      # for ivreg (2SLS estimation)
```

Load the matched dataset (from Assignment 2)

The matched_data has been preprocessed by matching on key covariates (e.g. pre-treatment home values, education, income, population density) to improve balance between treated and control precincts. We will now use this data for the IV analysis. Make sure to re-code the `precinct_id` variable as a `factor`.

```
matched_data <- read_csv(here::here("data", "matched_data_subset.csv")) %>%
    mutate(precinct_id = as.factor(precinct_id))
```

Part 1: IV Identification Rationale

Intuition for Using an Instrument:

Question 1: After matching on observables, why might we still need to utilize an instrumental variable approach to identify the causal effect of turbine proposals on vote share? In other words, what potential issues remain that an IV method can help address in this context? Use specific examples from the study to illustrate threats to a causal interpretation, then explain how an IV approach is designed to mitigate those threats.

Response: While matching helps to remove bias on observables, it does not control for bias on unobservables. The instrumental variable approach can identify confounds that could influence the treatment effect. In Stokes et. al., many proposed wind turbines surround the Great Lakes due to wind resources, and thus 'distance to lakes', and other geographic controls, are considered instrumental variables. By using the IV approach, the authors can correct for bias in these confounders to get closer to as-if random assignment.

Part 2: Two-Stage Least Squares (2SLS) Step-Wise Implementation

2A. First-Stage Estimation: Regress the treatment (D) on the instrument (Z)

$$D_i = \alpha_0 + \alpha_1 Z_i$$

a. Estimate the first-stage regression of the treatment on the instrument (with controls). Regress `proposed_turbine_3km` on `log_wind_power`.
b. Include the control variables used in Stokes (2015) for both stages: Distance to lakes, geographic coordinates (latitude & longitude) with their squares and interaction, plus district fixed effects.
c. After running the first stage, report the F-statistic for the instrument.

```
first_stage <- lm(proposed_turbine_3km ~ log_wind_power +
                      mindistlake + mindistlake_sq +
                      longitude + latitude +
                      long_sq + lat_sq +
                      long_lat +
                      factor(district_id),
                      data = matched_data)

export_summs(first_stage, digits = 3,
             model.names = c("First stage: Prpoposed Turbine 3km"),
             coefs = c("(Intercept)", "log_wind_power"))
```

	First stage: Prpoposed Turbine 3km
(Intercept)	15.027
	(74.243)
log_wind_power	0.711 ***
	(0.092)
N	708
R2	0.419

*** p < 0.001; ** p < 0.01; * p < 0.05.

Testing Instrument Relevance

Check instrument strength (F-statistic)

```
first_stage_f <- summary(first_stage)$fstatistic[1]
```

The first stage f-statistic for the instrument is ~14.7.

Question 2A: Based on the instrument relevance test reported in the study, would you conclude the instrument is strong enough to be credible? Explain what a weak instrument would mean in this setting: Specifically, what would it suggest about compliance with Ontario's Green Energy Act policy?

Response: We can consider an f-statistic greater than 10 as a strong instrument, and we can conclude the instrument is credible. A weak instrument would yield an f-statistic less than 10, and would ultimately lead to biased estimates and incorrect inferences of regression results. In this example, running a regression with a weak instrument would suggest compliance with Ontario's Green Energy Act policies could be higher or lower than they actually are.

2B. Second Stage Estimation

Regress the outcome (Y) on the fitted values from the 1st stage ($\hat{X}_i$)

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \epsilon_i$$

a. Now estimate the second stage of the 2SLS.
b. First, use the first-stage model to generate the predicted values of proposed_turbine_3km for each precinct (these are $\hat{D}_i$).
c. Add these predicted values as a new column in matched_data (e.g. proposed_turbine_3km_HAT).
d. Then, regress the outcome change_liberal (the change in Liberal vote share from 2007 to 2011) on the predicted treatment (proposed_turbine_3km_HAT), including the same controls and fixed effects as in the first stage.
e. Fill in the code for these steps below to obtain the second-stage regression results.

Save predicted values $\hat{X}_i$ from first stage

```
matched_data$proposed_turbine_3km_HAT <- predict(first_stage)
```

Estimate the second-stage regression

$$LiberalVoteShare_i = \beta_0 + \beta_1 \widehat{ProposedTurbine}_i + ControlVariables... + \epsilon_i$$

```
second_stage <- lm(change_liberal ~ proposed_turbine_3km_HAT +
                     mindistlake + mindistlake_sq +
                     longitude + latitude +
                     long_sq + lat_sq +
                     long_lat +
                     factor(district_id),
                     data = matched_data)

export_summs(second_stage, digits = 3,
             model.names = c("Second stage: Change in Liberal Vote Share"),
             coefs = c("(Intercept)", "proposed_turbine_3km_HAT") )
```

Interpreting the 2SLS Estimate

Question 2B: Imagine you are explaining your 2SLS findings to a policymaker in Ontario. What does the estimated coefficient on `proposed_turbine_3km_HAT` imply about the electoral impact of a local wind turbine proposal (within 3 km) on liberal vote share?

	Second stage: Change in Liberal Vote Share
(Intercept)	16.966
	(15.432)
proposed_turbine_3km_HAT	-0.065 *
	(0.027)
N	708
R2	0.586

*** p < 0.001; ** p < 0.01; * p < 0.05.

Response: Our model predicts that following proposed wind turbine projects in Ontario, the liberal vote share will decrease by an average of ~6.5% compared to districts without a proposed wind turbine project.

Question 2C: Explain what it means that IV identifies a LATE in the context of the wind-turbine voting study. What specific subset of observations does the second-stage 2SLS estimate apply to, and what does it imply about interpretation and generalizability?

Response: The IV approach only estimates the treatment effect for units/subjects that comply with the experimental treatment (LATE). In the context of the wind-turbine voting study, this IV approach identifies the local average treatment effect of for voters who comply with their treatment (proposed turbine in their district or no proposed turbine). The 2SLS estimate applies to the observations of participants who's treatment was determined solely by wind. We can interpret regression coefficients for voting patterns only for areas where wind turbines were proposed because of wind speeds, which should be an unbiased estimate of the treatment effect, but we cannot generalize to observations where this assumption is false.

Part 3: IV Assumptions and Validity

Evaluate Instrument Validity

Question 3: List the four key assumptions required for the IV strategy (2SLS) to identify a causal effect, and briefly explain what each one means in the context of this study. (Hint: think about what conditions a valid instrument must satisfy (relevance, exclusion,…)

Response:

1. Exclusion : Wind speed (instrument) affects liberal vote share only through its effect on proposed wind turbine location.
2. Relevance : Wind speed is correlated with proposed wind turbines. We tested this in the first stage regression, where we saw a f-statistic greater than 10.
3. Independence : Wind speed is randomly assigned and independent of the error term, and uncorrelated with omitted variables.
4. Monotonicity : Wind speed does not negatively influence compliance of proposed wind turbines. Wind speed can encourage proposed wind turbines or have no effect, but never discourage.

Part 4: Estimate 2SLS using `AER::ivreg()`

SEE Documentation for specification details: AER package Viggnette Example

> **Tip**
>
> Syntax for specifying 2SLS using `ivreg()`:
> `ivreg( Y ~ D + CONTROLS  | Z + CONTROLS , data )`
>
> - The first-stage predictor variables go after the `~` symbol
> - The second-stage predictor variables go after the `|` symbol

```
fit_2sls <- ivreg(change_liberal ~
                      proposed_turbine_3km +
                      mindistlake + mindistlake_sq +
                      longitude + latitude +
                      long_sq + lat_sq +
                      long_lat +
                      factor(district_id) |
                      log_wind_power +
                      mindistlake + mindistlake_sq +
                      longitude + latitude +
                      long_sq + lat_sq +
                      long_lat +
                      factor(district_id),
                  data = matched_data)
```

```r
export_summs(fit_2sls, digits = 3,
             model.names = c("Change in Liberal Vote Share"))
```

```r
             #coefs = c("(Intercept)", "proposed_turbine_3km")
```

Robustness checking strategies utilized in Stokes, 2015

Question 4: Choose two *robustness checks* from the paper that the authors use to increase confidence in their causal identification strategy. For each one, summarize the logic and findings from the robustness check in your own words:

Response:

1. For the instrumental variable approach (a robustness check itself), the authors conducted the 2SLS model with many different controls and different specifications, with treatment effect proving robust over these iterations. This validates their empirical results as their causal identification persists when changing control variables, which increases confidence in their results as they are not driven by specific conditions or endogeneity.
2. The authors conducted their 2SLS instrumental variable approach with the pretreatment (liberal vote share in 2003) as the outcome variable. They found their model did not return significant results anymore, suggesting the relationship found is causal and not a confounding factor, and not due to an omitted variable.

END

	Change in Liberal Vote Share
(Intercept)	16.966
	(15.737)
proposed_turbine_3km	-0.065 *
	(0.027)
mindistlake	0.002 **
	(0.001)
mindistlake_sq	-0.000 ***
	(0.000)
longitude	-0.150
	(0.314)
latitude	-1.077
	(0.574)
long_sq	-0.002
	(0.002)
lat_sq	0.007
	(0.007)
long_lat	-0.006
	(0.004)
factor(district_id)10	0.299 ***
	(0.051)
factor(district_id)14	0.046
	(0.075)
factor(district_id)18	0.176 **
	(0.064)
factor(district_id)19	0.214 ***
	(0.064)
factor(district_id)21	-0.050
	(0.072)
factor(district_id)22	0.004
	(0.078)
factor(district_id)28	0.162 *