Assignment 3: Instrumental Variable Estimation

Replicating the IV strategy in Stokes (2015)

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

PUBLISHED
February 10, 2026

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): Megan Hessel

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. pretreatment 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`.

```
# Load data
matched_data <- read_csv(here::here("data", "matched_data_subset.csv"))

# `precinct_id` as factor
matched_data$precinct_id <- factor(matched_data$precinct_id)

# set seed
set.seed(241)
```

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.

- Matching can only account for observable confounding variables. IV can help eliminate bias from unobservable confounders that might contribute to turbine location and voting patterns. For instance, there might be political (pre-existing anti-government beliefs in rural areas), economic (economic conditions other than median income), or developmental (city developers might target specific communities when building turbines) unobservable biases.
- IV is a good strategy to eliminate these unobservable biases. Stokes mention that wind speed is unrelated to political/economic preferences, eliminating that selection bias. Using wind speed as the instrument, the only variation in the study is the variation in turbine location driven by wind resources, which reduces bias from other unobservable variables present in the matching method.

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

COME BACK TO AND CHECK WITH OTHERS!!! D

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.

```
# Regress proposed_turbine_3km on log_wind_power with controls
first_stage <- lm(proposed_turbine_3km ~ log_wind_power +
                    #log_home_val_07 + p_uni_degree + log_median_inc + log_pop_denc +
                    mindistlake + mindistlake_sq + long_sq + lat_sq + latitude*longitud
                    as.factor(district_id), # controls ??
                  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_sum <- summary(first_stage)

first_stage_sum$fstatistic[1]
```

```
   value
14.70538
```

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?

- The F-statistic is the variance explained by wind speed / unexplained variance. In the first stage, a F-statistic less than 10 is a weak instrument, which can lead to unreliable inference. In this scenario. a weak instrument would mean wind speed is weakly correlated with turbine placement/ compliance with the Ontario's Green Energy Act policy, leading to biased and

unexplained variance in voting share results. However, the F-statistic is larger than 10, making it a strong and credible instrument. Going forward, "compilers" are those who get turbines if and only if they have good wind resources.

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

```
# get predictions 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 + long_sq + lat_sq + latitude*longitu
                   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") )
```

	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.

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?

- The 2SLS explains turbine placement (that is driven only by wind speed) impacts on voting shares. The estimate means there is a 6.5% decrease in liberal vote shares when wind turbine is proposed within 3km of voters, after isolating for effects driven by wind resources.

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?

- The second stage is isolated with precincts that have optional wind patterns (sub-population of compilers). In other words, the estimate is applied only to locations where turbine placement was determined by wind resources. Therefore, 2SLS estimates a *local* average treatment effect (LATE), not a average treatment effect (ATE) for all precincts.
- This isolation reduces the impact of political, economic, and other factors on the outcome coefficent estimation. However, it limits the generalizabiliy. The generalizability of the IV estimation depends on the similarities of compilers ≈ entire population. In this circumstance, the LATE estimation does not capture any strategic or non-wind driven turbine placement. For instance, if city developers are specifically choosing wind turbine locations based on neighborhood hostility, compliance, grid connectivity, or land availability, this study does not explain those situations.

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,...)

1. Substantial first stage / Relevance
    - The instrument variable must effect/change the treatment.
    - In this study, the wind speed must effect wind turbine placement. This correlation is noted by the F-statistic > 10.
2. Independence / Ignorability:
    - Instrument variable assignment is random and has no selection bias.
    - In this study, some precincts are more likely to have greater wind due to geographical conditions. However, these differences are not due to human selection and/or study design.
3. Exclusion restriction:
    - Instrument change outcomes (Y) must be solely through the treatment variable (D).
    - In this study, the liberal voting shares are dictated by who has wind turbine proposals within 3km of voters (in areas with wind resources).
4. Monotonicity:
    - There are no defiers in the population (individuals/communities that alway do the oppoiste of what the instrument suggest).

- In this study, defiers would be communities that will never put wind turbines up no matter if they have optimal wind patterns. The Green Energy Act took away choice, and communities no longer had a choice to be defiers or not.

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 + long_sq + lat_sq + latitude*longitud
                  data = matched_data
                  )

export_summs(fit_2sls, digits = 3,
             model.names = c("Change in Liberal Vote Share"),
             coefs = c("(Intercept)", "proposed_turbine_3km")
             )
```

	Change in Liberal Vote Share
(Intercept)	16.966
	(15.737)
proposed_turbine_3km	-0.065 *
	(0.027)
nobs	708
r.squared	0.570
adj.r.squared	0.549
sigma	0.082
statistic	27.814
p.value	0.000
df	34.000
df.residual	674.000
nobs.1	708.000

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

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:

1. Informed voting test

   Attempted to understand if voters were informed about the specific governmental responsibility in terms of the wind policy. Found that the voters were mostly informed and their voting behavior in the electoral votes was a reflection of their beliefs about the climate policy. Therefore, treatment (proposed wind turbines) does influence outcomes (liberal voting shares).

2. Distance Gradient Analysis

   Estimated the treatment effect at various distances (0km, 1km, 2km, 3km, 4kn, and 5km) from turbines to examine the spillover effect and the Not in My Backyard (NIMBY) opposition. Found distance from turbines play a huge role in forming public options. Communities in 0-3km of a wind turbines voted strongly against the electoral Liberal candidates, showing that surrounding communities dislike the policy (not just the neighboring communities). Thus, the treatment uses the 3km treatment zones to account for the spatial correlation/ spillover.

END
