Assignment 3: Instrumental Variable Estimation

Replicating the IV strategy in Stokes (2015)

AUTHOR
Aakriti Poudel

PUBLISHED
February 11, 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 (Aakriti Poudel): ______________________________________

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)
```

```r
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`.

```r
# Read the data
matched_data <- read_csv(here("data", "matched_data_subset.csv")) %>%
    clean_names() %>%
    mutate(precinct_id = factor(precinct_id))
```

```r
names(matched_data)
```

```
 [1] "precinct_id"          "district_id"          "change_liberal"
 [4] "proposed_turbine_3km" "log_wind_power"       "log_home_val_07"
 [7] "p_uni_degree"         "log_median_inc"       "log_pop_denc"
[10] "mindistlake"          "mindistlake_sq"       "longitude"
[13] "long_sq"              "latitude"             "lat_sq"
[16] "long_lat"             "weights"              "subclass"
```

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: Even after matching on variables like home values, education, income, and population density, there could still be unmeasured factors that affect both where turbines are placed and how people vote. Matching only controls for things we can see in the data.

For example, developers chose turbine locations based on things like available land and farmers willingness to lease their property. These factors might also relate to political views in ways we cannot measure. So we cannot be sure if vote changes are due to turbines or due to these hidden differences.

The instrumental variable approach solves this by using wind power as an instrument. Wind power predicts where turbines are built because developers want strong wind to make more profit. But wind power depends on geography and weather, which have nothing to do with how people vote. By using only the variation in turbine placement that comes from wind resources, we remove bias from things we cannot observe.

Matching fixes what we can see. The instrumental variable fixes what we cannot see.

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: Regress treatment on instrument with controls
first_stage <- lm(proposed_turbine_3km ~ log_wind_power +
                      mindistlake + mindistlake_sq +
                      latitude + longitude +
                      lat_sq + long_sq + long_lat +
                      factor(district_id),
                  data = matched_data)

# Display results
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)

```
summary(first_stage)$fstatistic
```

```
    value     numdf     dendf
 14.70538  33.00000 674.00000
```

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: Yes, the instrument is strong enough to be credible. The F-statistic of 14.7 is above the threshold of 10, and Stokes (2015) reported an F-statistic of 68. The coefficient on `log_wind_power` (0.711) is positive and significant, which means areas with more wind are more likely to get turbine proposals.

A weak instrument in this setting would mean wind power does not predict where turbines are built. In this study, that would suggest developers were not building turbines where wind is strongest. Under Ontario's Green Energy Act, developers got a fixed price for electricity, so they should build where wind is best to make the most profit. If the instrument were weak, it would mean developers chose locations for other reasons like political connections or land deals, not wind. This would make it harder to argue that turbine placement is random with respect to politics, and we could not trust our causal estimates.

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

```
# Save predicted values from first stage
matched_data$proposed_turbine_3km_HAT <- fitted(first_stage)
```

Estimate the second-stage regression

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

```
# Second stage: Regress outcome on predicted treatment with controls
second_stage <- lm(change_liberal ~ proposed_turbine_3km_HAT +
                     mindistlake + mindistlake_sq +
```

```r
                    latitude + longitude +
                    lat_sq + long_sq + long_lat +
                    factor(district_id),
                  data = matched_data)

# Display results
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?

Response: The coefficient of -0.065 means that precincts within 3 km of a proposed wind turbine had a 6.5 percentage point drop in Liberal vote share between 2007 and 2011. When turbines were planned near a community, people in that area voted less for the Liberals. Voters blamed the government for putting turbines near their homes. For policymakers, this shows that climate policies can cost votes in communities where the projects are built, even when most people support the policy overall.

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: LATE stands for Local Average Treatment Effect. It means the IV estimate only applies to a specific group, not everyone.

The -0.065 estimate applies only to precincts that got turbine proposals because of strong wind. These are places where developers built turbines because wind resources made it profitable. The estimate does not apply to places that got turbines for other reasons or places that would never get turbines even with good wind.

The effect is for precincts where wind power actually determined whether turbines were proposed. If some areas got turbines because of political deals or other reasons unrelated to wind, our estimate does not capture those cases.

We should be careful about applying these findings everywhere. The -0.065 effect is specific to areas where wind drove turbine placement. It may not apply to areas with weak wind or where turbines were placed for different reasons.

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: The four key assumptions required for the IV strategy (2SLS) to identify a causal effect, in the context of this study, are as follow: **1. Relevance:** - Wind power must predict where turbines are proposed. If wind power does not predict turbine location, we cannot use it as an instrument. The F-statistic of 68 shows this is met. The areas with stronger wind are more likely to get turbines. **2. Exclusion Restriction:** - Wind power must only affect voting because it causes turbines to be built. Wind itself should not directly change how people vote. This makes sense because wind patterns have nothing to do with politics. **3. Ignorability:** - Wind power must not be related to other hidden factors that affect turbine placement. Since wind depends on geography and weather, it should be independent from political or community factors. The study controls for location variables like latitude and longitude to help with this. **4. Monotonicity:** - Wind power must work the same way for everyone. Stronger wind should make turbine proposals more likely for all precincts, not more likely for some and less likely for others. This makes sense because developers always want stronger wind to make more money.

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

```
# Estimate 2SLS using ivreg()
fit_2sls <- ivreg(change_liberal ~ proposed_turbine_3km + mindistlake +
                    mindistlake_sq + latitude + longitude +
                    lat_sq + long_sq + long_lat + factor(district_id) |
                    log_wind_power + mindistlake + mindistlake_sq +
                    latitude + longitude + lat_sq + long_sq + long_lat +
                    factor(district_id),
                  data = matched_data)
```

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

	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)
latitude	-1.077
	(0.574)
longitude	-0.150
	(0.314)
lat_sq	0.007
	(0.007)
long_sq	-0.002
	(0.002)
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 *
	(0.071)
factor(district_id)29	0.239 ***
	(0.061)
factor(district_id)34	0.145 *
	(0.058)

	Change in Liberal Vote Share
factor(district_id)36	0.223 *
	(0.090)
factor(district_id)40	0.077
	(0.068)
factor(district_id)41	0.172 *
	(0.083)
factor(district_id)42	0.178
	(0.094)
factor(district_id)55	0.198 **
	(0.070)
factor(district_id)58	0.151 *
	(0.070)
factor(district_id)67	0.135
	(0.070)
factor(district_id)69	0.082
	(0.049)
factor(district_id)70	0.138 *
	(0.065)
factor(district_id)73	0.139
	(0.072)
factor(district_id)85	0.152 **
	(0.052)
factor(district_id)87	0.132
	(0.109)
factor(district_id)90	-0.033
	(0.140)
factor(district_id)91	-0.117
	(0.128)
factor(district_id)98	0.209 **
	(0.076)
factor(district_id)105	0.113
	(0.066)
nobs	708
r.squared	0.570
adj.r.squared	0.549
sigma	0.082
statistic	27.814
p.value	0.000
df	34.000

	Change in Liberal Vote Share
df.residual	674.000
nobs.1	708.000

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

```
#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: The two robustness checks from the paper that the authors use to increase confidence in their causal identification strategy are:

Robustness Check 1: Federal Election Results (Placebo Test) The authors checked if voters also punished the federal liberal party for wind turbines. But the federal government had nothing to do with Ontario's wind policy. Only the provincial government made that decision. If voters understood this, they should only punish the provincial liberals. The results showed the effect on federal liberals was much smaller, only 1-2% compared to 4-10% for provincial elections. This means voters mostly knew who was responsible. If voters had blamed both parties equally, it would suggest they were just angry at anyone called liberal, not responding to the actual policy.

Robustness Check 2: Distance Effects The authors tested if the effect got smaller as distance from turbines increased. They grouped precincts by distance (1 km, 2 km, 3 km, etc.) and measured the effect for each group. The effect lasted up to 3 km but disappeared at 4-5 km. This makes sense because people farther away cannot see or hear the turbines. This supports the idea that people oppose turbines because of local impacts on their community. If the effect did not shrink with distance, we would question whether turbines were really causing the vote change.

END
