

ECON 453
In-Class Exercise 9
November 14, 2023

Please download the file "IC9.gdt", a gretl data file. This dataset comes from the High School and Beyond dataset in the year 2000. During this exercise, we will replicate the methodology used in Dee (2004). This study examines how educational attainment/college attendance affects various outcomes. Here, we will focus on the outcome variable of whether individuals are *registered to vote*.

1. To begin, we will run a simple OLS regression. Use **register** as the dependent variable and **college** as the regressor. Provide an interpretation of the estimated impact of college attendance on registering to vote.
2. We are concerned that there is endogeneity in our estimated relationship. To correct for this, we will run a 2-stage least squares regression using **distance** as an instrumental variable.
 - a. In your opinion, does the **distance** variable fit the description of an appropriate instrumental variable for the **college** variable?
 - b. Run the first-stage regression. This means regressing **college** on **distance**.
 - i. Interpret the coefficient on the distance variable.
 - ii. Examine the F-statistic from your regression. A rule of thumb is that a first-stage F-statistic below 10 is evidence of a "weak instrument". What does the evidence say about our instrument?
 - iii. Save the fitted values from your first-stage regression.
 - c. Run the second-stage regression. This means regressing **register** on the fitted values from your first-stage regression. Interpret the coefficient from this regression (this is the instrumental variable estimate). Compare to the OLS estimate from question 1.
 - d. Let gretl do the work for you. Go to **Model -> Instrumental Variables -> 2SLS**. Compare the results to those of your second stage regression in part iv. Notice any differences?
3. Let's practice adding more variables to the model.
 - a. Estimate an OLS model using **register** as the dependent variable and **college**, **female**, and the race/ethnicity dummies as the explanatory variables. Briefly summarize the findings.
 - b. Estimate the regression with the added control variables using 2SLS and the **distance** instrument. When you do this, you should include all of the control variables (gender/race) in both the regressors and instruments sections. Compare the estimated coefficients in this model to those in part a.