ECON 453
In-Class Exercise 9
November 14, 2023

ANSWER KEY

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.

```
Q1: OLS, using observations 1-9227
Dependent variable: register

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       0.574061     0.00714090    80.39    0.0000     ***
  college     0.176930     0.00965436    18.33    1.03e-073  ***

Mean dependent var   0.670857   S.D. dependent var   0.469927
Sum squared resid    1965.823   S.E. of regression   0.461625
R-squared            0.035128   Adjusted R-squared   0.035024
```

This is just a simple OLS regression. Since the y-variable is binary, this is a linear probability model. This means we should interpret the results as telling us that having attended college increases the probability that someone is registered to vote by about 17.7 percentage points. From our model we can tell that the estimated probability of being registered to vote is 57.4% for those that did not attend college and about 75.1% for those that have.

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?

In order for distance to be a valid instrument we would need to argue that (1) living closer to a college as a kid makes you more likely to attend college, and (2) living closer to a college as a kid does not influence the likelihood that you register to vote as an adult (other than through the college attendance path). The first one seems likely to me (and we can check and prove this with the data). The second part is more complex. If you live closer to college because, for example, your family members work at a college, this might affect other outcomes (like whether you feel the need to vote).

   b. Run the first-stage regression. This means regressing **college** on **distance**.
      i. Interpret the coefficient on the distance variable.

```
Q2b: OLS, using observations 1-9227
Dependent variable: college

             coefficient   std. error    t-ratio    p-value
  ----------------------------------------------------------
  const       0.609118     0.00772873     78.81    0.0000     ***
  distance   -0.00637097   0.000591878   -10.76    7.35e-027  ***

Mean dependent var   0.547090   S.D. dependent var   0.497805
Sum squared resid    2257.930   S.E. of regression   0.494734
R-squared            0.012404   Adjusted R-squared   0.012297
F(1, 9225)           115.8637   P-value(F)           7.35e-27
Log-likelihood      -6598.189   Akaike criterion     13200.38
Schwarz criterion    13214.64   Hannan-Quinn         13205.23
```

This coefficient says that as the distance between your home and the nearest college increases by 1 mile, the likelihood you will attend college decreases by about 0.64 percentage points. This seems pretty small, but if you scale it, it might make more sense. For example, if you move 10 miles further away from a college, the likelihood you attend college drops by about 6.4 percentage points.

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?

Our first stage F-statistic is 115.8637. This is well above the threshold of 10 and indicates that we have a strong instrument. This is another way of saying that the distance to the nearest college does, indeed, influence the likelihood you attend college.

iii. Save the fitted values from your first-stage regression.

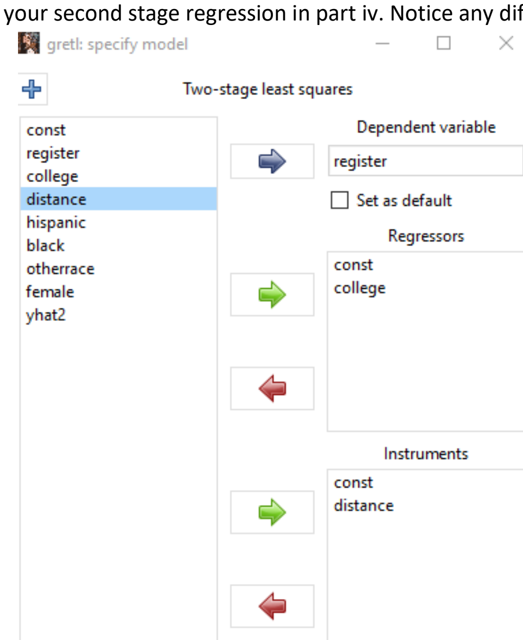
To do this, go to the window with your regression results and choose <u>Save -> Fitted values</u>.

c. Run the <u>second-state</u> 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.

```
Q2c: OLS, using observations 1-9227
Dependent variable: register

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       0.515653     0.0485001     10.63    3.00e-026 ***
  yhat2       0.283691     0.0881993     3.216    0.0013    ***

Mean dependent var   0.670857   S.D. dependent var   0.469927
Sum squared resid    2035.111   S.E. of regression   0.469690
R-squared            0.001120   Adjusted R-squared   0.001012
F(1, 9225)           10.34574   P-value(F)           0.001302
Log-likelihood      -6118.855   Akaike criterion     12241.71
Schwarz criterion    12255.97   Hannan-Quinn         12246.56
```

The coefficient on our fitted values (what I called yhat2), provides our instrumental variable estimate of the impact of college attendance on registering to vote. In this case, we estimate the probability you register to vote increases by about 28.4 percentage points if you attended college. This is a much larger estimate than the OLS estimate in question 1 (17.7 percentage points).

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?



```
Q2d: TSLS, using observations 1-9227
Dependent variable: register
Instrumented: college
Instruments: const distance

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       0.515653     0.0479822    10.75     8.84e-027 ***
  college     0.283691     0.0872575     3.251    0.0012    ***

Mean dependent var   0.670857   S.D. dependent var   0.469927
Sum squared resid    1991.882   S.E. of regression   0.464674
R-squared            0.035128   Adjusted R-squared   0.035024
F(1, 9225)           10.57027   P-value(F)           0.001153
Log-likelihood      -79091.91   Akaike criterion     158187.8
Schwarz criterion    158202.1   Hannan-Quinn         158192.7
```

The coefficient is the same as when we did the instrumental variable estimation in 2 stages (the coefficient on yhat2 in question 2c). However, the standard errors (and thus the t-stats and p-values) are different. It is better to let the statistical programs run the 2-stage least squares for us, they calculate the standard errors correctly based on what we are trying to do.

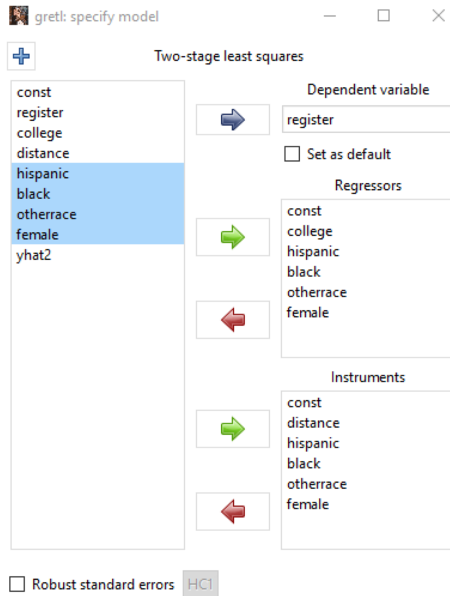
3. Let's practice adding more variables to the model.
    a. Estimate an OLS model using **register** as the dependent variable and **college**, **female**¸ and the race/ethnicity dummies as the explanatory variables. Briefly summarize the findings.

```
Q3a: OLS, using observations 1-9227
Dependent variable: register

             coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const       0.562484     0.00950056   59.21     0.0000    ***
  college     0.179941     0.00967507   18.60     7.98e-076 ***
  hispanic    0.0214140    0.0123818     1.729    0.0838    *
  black       0.0590949    0.0148402     3.982    6.88e-05  ***
  otherrace  -0.102980     0.0224401    -4.589    4.51e-06  ***
  female      0.00645460   0.00960642    0.6719   0.5017

Mean dependent var   0.670857   S.D. dependent var   0.469927
Sum squared resid    1956.401   S.E. of regression   0.460617
R-squared            0.039753   Adjusted R-squared   0.039232
F(5, 9221)           76.34784   P-value(F)           1.12e-78
Log-likelihood      -5936.880   Akaike criterion     11885.76
Schwarz criterion    11928.54   Hannan-Quinn         11900.30
```

When we add the control variables for gender and race, we first notice that the "college" coefficient is very similar to the regression we did in question 1. Our estimate increased slightly to 18 percentage points, instead of 17.7. We learn that Hispanic and black individuals are more likely to be registered to vote (than white) and other race are individuals are less likely to be registered (than white). There is not a significant difference between males and females.

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.

See the image below for how to enter this under <u>Model -> Instrumental Variables-> Two-Stage least squares</u>:



Here are my results:

```
Q3b: TSLS, using observations 1-9227
Dependent variable: register
Instrumented: college
Instruments: const distance hispanic black otherrace female

              coefficient    std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        0.524756      0.0452997     11.58     8.10e-031 ***
  college      0.248386      0.0809269      3.069    0.0022    ***
  hispanic     0.0283147     0.0148242      1.910    0.0562    *
  black        0.0616661     0.0151834      4.061    4.92e-05  ***
  otherrace   -0.106407      0.0228578     -4.655    3.28e-06  ***
  female       0.00405072    0.0100373      0.4036   0.6865

Mean dependent var    0.670857   S.D. dependent var    0.469927
Sum squared resid     1967.019   S.E. of regression    0.461865
R-squared             0.039484   Adjusted R-squared    0.038964
F(5, 9221)            9.012950   P-value(F)            1.48e-08
Log-likelihood       -79021.01   Akaike criterion      158054.0
Schwarz criterion     158096.8   Hannan-Quinn          158068.6
```

The estimated coefficients for Hispanic, black, female, and other race are all very similar in this model to what we saw in the OLS model in part a. One thing we notice is that the IV estimate of the effect of college attendance on registration is reduced (as compared to question 2, part c and d). Now the estimated effect is around 24.8 percentage points. The IV estimate is still larger than the OLS estimate, but the gap between the two is not as large as in the case where we have added the other demographic information.