**Principals of instrumental variable analysis practical 2020**

This practical uses a simulated random sample from the population. It simulates a study investigating the effects of two types of anti-inflammatory drug, traditional NSAIDs (e.g. ibuprofen) vs COX-2 selective inhibitors (COX-2s e.g. celecoxib). The dataset contains data from 100,000 patients, it is a patient level file, i.e. each patient has a single row. In the dataset the exposure is indicated by the variable ‘prescribed\_cox\_2’. It equals one if the patient had a COX-2 and zero if they had a traditional NSAID. The outcome of interest is whether the patient subsequently had a gastrointestinal complication (variable ‘has\_gi\_event’), equal to one, or did not have a complication, when the outcome is equal to zero. The dataset is called iv\_practical\_2020.dta. There are 100,000 observations with variables on treatment, physician who prescribed the treatment, age, and sex. We will use the information on the physician who prescribed the treatment to create an instrument and will estimate the effects of prescribing COX-2s versus traditional NSAIDs.

1. Open the dataset and describe the variables

use iv\_practical\_2020, clear

describe

summarize

1. Create the instrument – the physician’s previous prescription. The variable visit order indicates the order in which the patients visited their GP. The command below will create a variable equal to one if the physician previously prescribed a COX-2 and equal to zero if they previously prescribed a traditional NSAID.

bysort physician\_id (visit\_order):gen prior\_prescription= prescribed\_cox\_2[\_n-1]

1. Test the first instrumental variable assumption (the relevance assumption). Are the physicians’ previous prescription associated with their subsequent prescriptions? What happens when you adjust for age and sex? How should estimates from a linear probability model be interpreted?

regress prescribed\_cox\_2 prior\_prescription, robust

regress prescribed\_cox\_2 prior\_prescription age female, robust

1. Investigate the plausibility of the third instrumental variable assumption, independence. Do the instruments associate with the measured confounders?

regress age prior\_prescription, robust

regress female prior\_prescription, robust

1. Estimate the multivariable adjusted linear regression of prescribed COX-2 and the outcome, has GI event. Are prescriptions of COX-2s associated with a higher or lower risk of gastrointestinal events? What happens when you adjust for the observed covariates?

regress has\_gi\_event prescribed\_cox\_2, robust

regress has\_gi\_event prescribed\_cox\_2 age female, robust

1. The instrumental variable ratio estimator (Wald type) is the instrument-outcome association divided by the instrument-exposure association. Estimate this using the instrument you created above (prior\_prescription). Use the following code to save your estimates as scalars. You will need to run this code as a single block for this to work. Compare your instrumental variable estimate with the unadjusted and adjusted estimates above.

regress prescribed\_cox\_2 prior\_prescription, robust

scalar denominator=\_b[prior\_prescription]

regress has\_gi\_event prior\_prescription, robust

scalar numerator = \_b[prior\_prescription]

display scalar(numerator)/scalar(denominator)

1. Now perform two-stage least squares estimation, again using the prior prescription as the instrumental variable. Compare this with the ratio estimate above.

ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription), robust

1. Perform two-stage least squares estimation manually, i.e. by fitting the first stage regression of the exposure (prescription) on the physicians’ prior prescription, saving the predicted values, then regressing the outcome on the predicted values of the exposure. Compare the standard error and confidence interval with those from two-stage least squares.

regress prescribed\_cox\_2 prior\_prescription

predict prescribed\_cox\_2\_hat

regress has\_gi\_event prescribed\_cox\_2\_hat, robust

1. We can potentially increase the power of instrumental variables by including more of the physicians’ prior prescriptions. Create two further variables indicating the physician’s last but two, and last but three prescriptions. Investigate the first stage R2 from the regress of the exposure (prescribed\_cox\_2) on these variables and compare to the R2 from question 3 above.

bysort physician\_id (visit\_order):gen prior\_prescription2= prescribed\_cox\_2[\_n-2]

bysort physician\_id (visit\_order):gen prior\_prescription3= prescribed\_cox\_2[\_n-3]

1. Now perform two-stage least squares estimation using each instrument separately. What do notice about these three estimates compared to the estimate in question 7? What do you notice about the sample size? Why has this happened?

ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription), robust

ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription2), robust

ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription3), robust

1. Now perform two-stage least squares estimation using the three instruments as multiple instruments.

ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription prior\_prescription2 prior\_prescription3), robust

1. Investigate the first-stage diagnostics (partial f-statistic), endogeneity test, and overidentification test using -ivregress- postestimation commands. Interpret the endogeneity test and over-identification test p-values.

ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription\*), robust

estat firststage, all

estat endogenous

estat overid

BONUS QUESTION 1: What 4th instrumental variable assumption could be used here? What treatment parameter can it be used to estimate?

BONUS QUESTION 2: If we included measured confounders in the instrumental variable estimation what interpretation would the causal parameter from two-stage least squares estimate. By reading the helpfile for ivregress can you see how to include confounders?

BONUS QUESTION 3: Using the helpfile for the ivreg2 command\*, can you run all the analyses in question 10 using a single ivreg2 command?

BONUS QUESTION 4: What happens if we run the manual analysis described in Question 8 but include the covariates age and sex in the first stage, but not the second stage?

\* If ivreg2 is not already installed, use the following command:

ssc install ivreg2