ECO 375–Homework 3
University of Toronto Mississauga
Due: 5 December, 2023, 5 PM

Notes:

- Late homework assignments—including assignments submitted late on the day of the deadline—will be subject to a late penalty of 10% per calendar day (including weekends) of the total marks for the assignment. Assignments submitted more than 5 calendar days beyond the due date will be assigned a grade of zero. Homework assignments handed in after the work has been returned to the class or sample solutions have been released cannot be marked for credit. Accommodations due to late registration into the course will NOT be approved.
- For full credit, please **show your work**. The correct answer without a clear explanation will receive little credit.
- All problem sets must be legible; if the grader cannot read your submission, you will receive no credit.
- Provide your do file and log file as part of your submission. If you are using Citrix, remember to save your code and log file on your home computer, not the remote computer! You will need to copy-paste your do and log files into Crowdmark.
- When questions involve the use of Stata, as part of the response to each question, you need to submit a write-up that (a) interprets and explains your computer output and (b) includes the code for that answer answer. Without both components, you will receive a mark of 0. For example, if the question asks for the mean of x, your answer might look like this:
  "The mean of x is 12345.
  Code:
  `sum x`"
- You may take as given anything from the textbook or that we learned in class, but not any other information.
- All work MUST be your own. Using generative AI to answer a question constitutes cheating. Posting questions to an online forum or website constitutes cheating and violates copyright law; looking for the answer in an online forum or website constitutes cheating.

Theoretical Problems

1. A company notices that workers who earn a higher wage tend to have higher sales figures. Will raising wages cause their employees to sell more? To figure out, they gather a data set with one observation per worker, with variable y as the amount of sales that person made in a month (say, June), and x as the person's wage in that same month (June). For each of the following potential z variables, briefly explain whether z would be a valid instrument. (Make sure to explain whether or not z satisfies BOTH conditions for being a valid instrument.)

   (a) *(points: 4)* z is a measure of the person's sales ability: for example, the results of a survey of customers asking how much they like the worker.

   (b) *(points: 4)* The company runs an experiment: some randomly-chosen "treated" workers receive \$1 more per hour in this month (June), while other "control" workers keep their old wage. z is a dummy variable that takes a value of 1 if the person is treated, 0 if untreated.

   (c) *(points: 4)* The company runs an experiment: some randomly-chosen "treated" workers receive a class on being a better salesperson just before the month starts (so the end of May), while control workers do not. z is a dummy variable that takes a value of 1 if the person is treated, 0 if untreated. You may assume that the class is effective at making people better at sales.

   (d) *(points: 4)* The company runs an experiment: some randomly-chosen "treated" workers receive an email just before the month starts (so the end of May) reminding them to fill out some short paperwork, while control workers do not. z is a dummy variable that takes a value of 1 if the person is treated, 0 if untreated.

   (e) *(points: 4)* $z = x$. That is, the instrument is exactly the person's wage.

   (f) *(points: 4)* z is the worker's wage in the previous month (May).

2. Suppose you are trying to estimate $y = \beta_0 + \beta_1 x + u$, using z as an instrument. When you estimate $x = \pi_0 + \pi_1 z + v$ in Stata with OLS, you find that the R^2 is .03.

   (a) *(points: 5)* If your sample size is $n = 100$, should you worry about z being a weak instrument?

   (b) *(points: 5)* If your sample size is $n = 1000$, should you worry about z being a weak instrument?

   (c) *(points: 5)* What can you infer from this about the relationship between sample size and instrument weakness?

3. Suppose you are trying to estimate

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{1}$$

where x_1 is exogenous but x_2 is endogenous. Can β_1 be consistently estimated with OLS? In this problem, we will address that question. For this question, suppose $x_2 = \alpha x_1 + z + \delta u + v$, where α and δ are constants and the following random variables are all iid $N(0,1)$ and independent of each other: x_1, u, z, v. For simplicity, also assume that $\beta_0 = 0$.

   (a) *(points: 5)* Suppose we estimate $x_1 = \gamma_0 + \gamma_1 x_2 + r$ with OLS. What is the plim of the OLS estimator of γ_1? Write your answer as a function of the parameters: $\beta_1, \beta_2, \alpha, \delta$.

   (b) *(points: 5)* Now, write a large-sample expression for the $\hat{r}$, the residual in the previous part. That is, write

   $$\hat{r} = x_1 - plim(\gamma_0) - plim(\gamma_1) x_2$$

   as a function of the parameters $\beta_1, \beta_2, \alpha, \delta$ and the random variables x_1, u, z, v.

   (c) *(points: 7)* Use the previous result and the partialling-out results we discussed in class to write the plim of the OLS estimator of β_1 in Equation 1, above. Simplify your answer. Hints:

   - This will involve a lot of algebra; it might be easiest, whenever possible, to combine all terms that multiple each random variable; for example, if an expression includes $Ax_1 + Bx_1$, rewrite as $(A + B)x_1$.
   - You should also take good advantage of the fact that many of the random variables are iid and independent of each other with variance 1.
   - When you are done, you should find that the OLS estimator is consistent for β_1 if $\alpha = 0$ (that is, if x_1 and x_2 are uncorrelated) or if $\delta = 0$ (that is, if x_2 is exogenous).

   (d) *(points: 5)* Since OLS is generally inconsistent, let's try z as an instrument. For the first stage, what is the plim of the predicted $\hat{x}_2$ from estimating with OLS $x_2 = \pi_0 + \pi_1 x_1 + \pi_2 z + e$? Write your answer as a function of the parameters $\beta_1, \beta_2, \alpha, \delta$ and the random variables x_1, u, z, v. (Hint: no algebra is needed for this question; instead, make an argument based on the definition of x_2.)

   (e) *(points: 5)* Suppose we estimate $x_1 = \omega_0 + \omega_1 \hat{x}_2 + s$ with OLS. What is the plim of the OLS estimator of ω_1? Write your answer as a function of the parameters: $\beta_1, \beta_2, \alpha, \delta$.

   (f) *(points: 5)* Now, write a large-sample expression for the $\hat{s}$, the residual in the previous part. That is, write

   $$\hat{s} = x_1 - plim(\omega_0) - plim(\omega_1) x_2$$

   as a function of the parameters $\beta_1, \beta_2, \alpha, \delta$ and the random variables x_1, u, z, v.

   (g) *(points: 5)* Use the previous result and the partialling-out results we discussed in class to write the plim of the 2SLS estimator of β_1 in Equation 1, above, using z as an instrument.

Computer Based Problems

Notes:

- If you are using Citrix, remember to save your code and log file on your home computer, not the remote computer!
- Remember to include the Stata code as part of your writeup. (See "Notes" at the start of this assignment.)

4. **Monte Carlo Simulation**. Many factors influence whether OLS or instrumental variables are the best choice in any particular situation. We saw in class one condition that compares the asymptotic bias of the OLS and IV estimators when neither is consistent. However, we also learned that an OLS estimator always has a lower variance than an IV estimator. Because of that, sometimes we prefer the OLS estimator even if it is inconsistent and the IV estimator is consistent. We will see an example of that here.

   Consider the following model:

$$\begin{aligned} y &= 1 + 2x + \delta p + (2.2 - \delta)u \\ x &= 3 - z + p + v \\ u &\sim N(0,1) \\ v &\sim N(0,1) \\ p &\sim N(0,1) \\ z &\sim N(0,1) \end{aligned}$$

   where δ is a constant that is either 0.2 or 2. Simulate this model with sample size of $n = 10$, 30, and 100, each with $\delta = 0.2$ or $\delta = 2$ (so 6 simulations total). (If you choose, you may instead simulate it twice—once for each δ value—with 100 observations, then take the first 10 or 30 to estimate some properties.) Run the simulation 1000 times. (You should create the code with fewer simulations until you are sure the code works.) Estimate the following model twice: once with OLS, and once with 2SLS, using z as an instrument.

$$y = \beta_0 + \beta_1 x + e$$

   (a) *(points: 4)* Do you expect the OLS estimators to be consistent? Briefly explain.

   (b) *(points: 4)* Do you expect the 2SLS estimators to be consistent? Briefly explain.

   (c) *(points: 10)* Using the results of your simulation, complete the table below four times: showing the average and standard deviation of the estimators of β_1, from using OLS and from using IV.

	$\delta = 0.2$	$\delta = 2$
$n = 10$		
$n = 30$		
$n = 100$		

   (d) *(points: 6)* Using the results above, create one more table with the difference, $MSE(\hat{\beta}_{OLS}) - MSE((\hat{\beta}_{IV}))$. That is, find the MSE of each estimator, and fill the table with the difference. Note that positive values shows situations where IV is preferred, while negative values shows when OLS is preferred.