Homework 3
Intermediate Econometrics

November 10, 2023

Exercise 1

Ordinary Least Squares Approach

Question 1

We consider the following regression model:

$$log(wage)_i = \beta_0 + \beta_1 education_i + \beta_2 unemp_i + \beta_3 ethnicity_i + \beta_4 gender_i + \beta_5 urban_i + \varepsilon_i$$

where we treat ($unemp$, $ethnicity$, $gender$, $urban$) as **exogenous** regressors.

1. We estimate model 1 by OLS (standard errors are in parentheses):

	Dependent variable:
	logwage
education	0.0014 (0.0011)
unemp	0.013*** (0.001)
ethnicity	−0.048*** (0.006)
genderfemale	−0.009** (0.004)
urbanyes	0.003 (0.005)
Constant	2.135*** (0.017)
Observations	4,739
R^2	0.083
Adjusted R^2	0.082
Residual Std. Error	0.138 (df = 4733)
F Statistic	85.365*** (df = 5; 4733)
Note:	*p<0.1; **p<0.05; ***p<0.01

Thus, $\hat{\beta}_1 = 0.0014$ and $s.e.(\hat{\beta}_1) = 0.0011$.

2. We conduct a significance test for β_1:

- The hypotheses are:

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

- We use the following test statistic:

$$\hat{t} = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}.$$

  Under H_0, $\hat{t} \xrightarrow{d} \mathcal{N}(0,1)$.

- The rejection rule for a 5% level test is to reject H_0 if:

$$\left|\hat{t}\right| > z_{1-\frac{0.05}{2}\ =\ 1-0.025\ =\ 0.975} = 1.96 \text{ (also given by the R ouput)}$$

- The value of the test statistic is:

$$\hat{t} = \frac{0.0014}{0.0011} = 1.248508 \text{ (also given by the R ouput)}$$

- Since 1.248508 is smaller than 1.96, we cannot reject the null (insignificance) in favor of the alternative at 5%. Thus, we cannot conclude that *education* has an effect on *wage*.

3. In modern econometrics, a variable x is said to be endogenous when $E(\varepsilon \mid x) \neq 0$, i.e. x is correlated with the error term. In model 1, *education* may be endogenous because there may be potential omitted variables (*example*) explaining (log)*wage*, which are not included in the model but are rather "represented" by the error term ($\varepsilon_i = example_i + error_i$), and which influence *education*. Thus, $E(\varepsilon \mid education) \neq 0$ because *education* and *example* are correlated, and this violates the mean independence assumption.

Instrumental Variable Approach I

Question 2

Considering the endogeneity issue, we use *distance* as an instrumental variable.

1. When we have one endogenous regressor and one instrument, it is clear that the instrument is valid if it explains some of the variation in the endogenous regressor (relevance) and is unrelated to the error term (exogeneity). **RELEVANCE:** we expect

distance to be (negatively) correlated to *education* because we expect that the farther a student lives from a 4-year college, the less chance they will have to complete one more year of education. **EXOGENEITY:** we can assume that *distance* is unrelated to the error term.

2. In the first stage, we should regress the endogenous variable on the set of all exogenous variables and on the instrument. The first stage equation for *education* is:

$$education_i = \alpha_0 + \alpha_1 unemp_i + \alpha_2 ethnicity_i + \alpha_3 gender_i + \alpha_4 urban_i + \alpha_5 distance_i + error_i.$$

3. We estimate the first stage by OLS (standard errors in parentheses):

	Dependent variable:
	education
unemp	0.008 (0.010)
ethnicity	-0.457^{***} (0.071)
genderfemale	-0.022 (0.052)
urbanyes	-0.128^{**} (0.065)
distance	-0.089^{***} (0.012)
Constant	14.029^{***} (0.083)
Observations	4,739
R^2	0.020
Adjusted R^2	0.019
Residual Std. Error	1.772 (df = 4733)
F Statistic	18.924^{***} (df = 5; 4733)
Note:	*p<0.1; **p<0.05; ***p<0.01

4. We conduct a significance test for α_5 (the coefficient for *distance*):

- The hypotheses are:

$$H_0 : \alpha_5 = 0 \text{ against } H_1 : \alpha_5 \neq 0.$$

- We use the following test statistic:

$$\hat{t} = \frac{\hat{\alpha_5}}{s.e.(\hat{\alpha_5})}.$$

  Under H_0, $\hat{t} \xrightarrow{d} \mathcal{N}(0,1)$.

- The rejection rule for a 5% level test is to reject H_0 if:

$$\left|\hat{t}\right| > z_{1-\frac{0.05}{2}} = 1-0.025 = 0.975 = 1.96 \text{ (also given by the R ouput)}$$

- The value of the test statistic is:

$$\hat{t} = \frac{-0.089}{0.012} = -7.305953 \text{ (also given by the R ouput)}$$

- Since $|-7.305953| = 7.305953$ is greater than 1.96, we reject the null in favor of the alternative at 5%. Thus, α_5 is significant at 5%. So we can conclude that *distance* is (negatively) correlated to *education*. This means that *distance* is a valid instrument because it is "relevant".

5. We use the ivreg() function to apply 2SLS in one step and get correct standard errors. We get the following estimation results:

	Dependent variable:
	logwage
education	0.0709^{***} (0.0142)
unemp	0.014^{***} (0.001)
ethnicity	-0.018^{*} (0.010)
genderfemale	-0.007 (0.005)
urbanyes	0.003 (0.006)
Constant	1.163^{***} (0.200)
Observations	4,739
R^2	-0.661
Adjusted R^2	-0.663
Residual Std. Error	0.185 (df = 4733)
Note:	*p<0.1; **p<0.05; ***p<0.01

Thus, $\hat{\beta}_1 = 0.0709$ and $s.e(\hat{\beta}_1) = 0.0142$.

We conduct a significance test for β_1:

- The hypotheses are:

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

- We use the following test statistic:

$$\hat{t} = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}.$$

  Under H_0, $\hat{t} \xrightarrow{d} \mathcal{N}(0,1)$.

- The rejection rule for a 5% level test is to reject H_0 if:

$$\left|\hat{t}\right| > z_{1-\frac{0.05}{2}\ =\ 1-0.025\ =\ 0.975} = 1.96 \text{ (also given by the R ouput)}$$

- The value of the test statistic is:

$$\hat{t} = \frac{0.0709}{0.0142} = 4.96 \text{ (also given by the R ouput)}$$

- Since 4.96 is greater than 1.96, we reject the null (insignificance of β_1) in favor of the alternative at 5%. Thus, we can conclude that *education* has a (positive) effect on *wage*. In fact, on average and *ceteris paribus*, one more year of education increases state hourly wage -in US dollars) by 7.09%.

6. We conduct a Hausman test to check the endogeneity of *education*:

- The hypotheses are:

$$H_0: \text{ the variable } \textit{education} \text{ is exogenous}$$

against

$$H_1: \text{ the variable } \textit{education} \text{ is endogenous.}$$

- We use the following test statistic:

$$\hat{t} = \frac{\hat{\beta}_{1,2SLS} - \hat{\beta}_{1,OLS}}{\sqrt{\hat{Var}(\hat{\beta}_{1,2SLS}) - \hat{Var}(\hat{\beta}_{1,OLS})}}.$$

Under H_0, $\hat{t} \xrightarrow{d} \mathcal{N}(0,1)$.

- The rejection rule for a 5% level test is to reject H_0 if:

$$\left|\hat{t}\right| > z_{1-\frac{0.05}{2}\,=\,1-0.025\,=\,0.975} = 1.96 \text{ (also given by the R ouput)}$$

- The value of the test statistic is:

$$\hat{t} = 4.87666 \text{ (given by the R ouput)}$$

- Since 4.87666 is greater than 1.96, we reject the null (exogeneity of β_1) in favor of the alternative at 5%. Thus, we can conclude that *education* is an endogenous variable. Thus we can think that we could add another (potentially viable) instrumental variable on top of *distance*, which gives us the next approach.

Instrumental Variable Approach II

Question 3

We know use both *distance* and *mcollege* as instrumental variables.

	Dependent variable:
	logwage
educ_fit_2	0.0218*** (0.0046)
unemp	0.013*** (0.001)
ethnicity	−0.039*** (0.006)
genderfemale	−0.008** (0.004)
urbanyes	0.003 (0.005)
Constant	1.850*** (0.065)
Observations	4,739
R^2	0.087
Adjusted R^2	0.086
Residual Std. Error	0.137 (df = 4733)
F Statistic	89.833*** (df = 5; 4733)
Note:	*p<0.1; **p<0.05; ***p<0.01

1. We estimate β_1 using the manual 2SLS method. We get the following results at the end of the second stage (TABLE ABOVE):

$\hat{\beta}_1 = 0.0218$ and $s.e(\hat{\beta}_1) = 0.0046$.

2. We conduct a Sargan test to check the exogeneity of the two instruments *distance* and *mcollege*:

- The hypotheses are:

  $$H_0: \text{ the instruments are exogenous}$$

  against

  $$H_1: \text{ at least one instrument is endogenous.}$$

  The Sargan test is based on the observation that the residuals should be uncorrelated with the set of exogenous variables if the instruments are truly exogenous. First, we regress the residuals from the 2SLS regression on all exogenous variables. Then we test the joint nullity of the coefficients of all the instrumental variables using a Wald test. We thus have:

  $$H_0 : \beta_{distance} = \beta_{mcollege} = 0$$

  against

  $$H_1 : \beta_{distance} \neq 0 \text{ or } \beta_{mcollege} \neq 0.$$

- Let $l = 2$ be the number of instruments. Then the Wald test statistic is $J = l \cdot F$ where F is the F-statistic for testing the equality of the l parameters to 0. Under H_0, $J \xrightarrow{d} \mathcal{X}^2_{\text{nb of instruments - nb of endogenous var}} = \mathcal{X}^2_{2-1} = \mathcal{X}^2_1$.

- The rejection rule for a 5% level test is to reject H_0 if:

$$J > \mathcal{X}^2_{1,0.95}.$$

  But here we will reject if p-value < 0.05.

- The value of the test statistic is:

$$J = 24.814 \text{ (given by the R ouput)}$$

- The computation of the p-value in R gives 6.313723e-07 < 0.05. So we reject the null (exogeneity of the instruments) in favor of the alternative at 5%. Thus, we can conclude that AT LEAST ONE instrument is an endogenous variable and thus not valid, but we don't know which one at this point, and it does not necessarily mean that the 2 instruments are invalid.

3. We suppose *distance* is a valid instrument. From the previous question, we know that AT LEAST ONE instrument is invalid among the two. So it can only be *mcollege* that is invalid, it does not satisfy the exogeneity condition.

Exercise 2

Question 1

We consider an instrument to be weak when it is weakly correlated with the endogenous variable(s). As a consequence, weak instruments are less precise in their ability to predict the endogenous variable. Thus, the estimates become imprecise and the standard errors of the estimated coefficients are inflated. Finally, the 2SLS estimator tends to converge towards OLS estimates, which can itself be biased due to endogeneity, so the 2SLS estimator does not provide a great improvement over OLS in this case.

Question 2

We suppose we have a just-identified model with a single endogenous variable and no exogenous regressors. In this very simple case, we can directly look at the R^2 (when the endogenous variable is regressed on the instrument): if it is small, then we can consider the instrument to be weak. Moreover, we can also look at the first-stage F-statistic: if it is small (smaller than 10 according to Stock, J. & Yogo, M. (2002). *Testing for Weak Instruments in Linear IV Regression*), then we can consider the instrument to be weak.