Econometrics: Models with Endogenous Explanatory Variables

Burcu Eke

UC3M

Endogeneity

- Given the following linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

  - If $E[\varepsilon | X_1, X_2, \ldots X_k] = 0 \; \forall X_j$, then we say that we have **explanatory exogenous variables**
  - If, for some reason such as omission of relevant variables, measurement errors, simultaneity, etc., X_j is correlated with ε, we say that X_j is an **endogenous explanatory variable**.

Endogeneity

- OLS estimators of the model parameters are invalid (inconsistent, etc.) under the existence of endogenous explanatory variables.

- In this topic, we will study how to obtain consistent estimators of the model parameters in the presence of endogenous explanatory variables using instrumental variables and applying the two-stage least squares estimation (2SLS).

Endogeneity Example 1: Measurement Error in the Explanatory Variables

- Recall $Y = \beta_0 + \beta_1 X^* + \varepsilon$, where the classical assumptions are satisfied, hence:
  $E[\varepsilon|X^*] = 0 \Rightarrow E[Y|X^*] = L(Y|X^*) = \beta_0 + \beta_1 X^*$. Under this case,
  - $E[\varepsilon] = 0$ and $C(X^*, \varepsilon) = 0$
  - $\beta_0 = E[Y] - \beta_1 E[X^*]$ and $\beta_1 = \dfrac{C(Y, X^*)}{V(X^*)}$

Endogeneity: Example 1

- However, X^* has measurement error. We observe X such that $X = X^* + v_1$, where v_1 is the measurement error
- Substituting $X^* = X - v_1$, we get

$$Y = \beta_0 + \beta_1 X + \underbrace{\varepsilon - \beta_1 v_1}_{u}$$

- Then, $C(X, u) \neq 0 \Rightarrow X$ is endogenous.

Endogeneity Example 2: Omission of Explanatory Variables

- Recall the case of omitting a relevant variable
- Let $Y = \gamma_0 + \gamma_1 X_1 + u$, where $u = \varepsilon + \beta_2 X_2$ and $\beta_2 \neq 0$. Then this model is misspecified by omitting a relevant variable
- In general, $C(X_1, u) \neq 0 \Rightarrow X_1$ is endogenous.

Endogeneity Examples

1. Ability is not observed in a wage equation such as:
   $\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{ability} + \varepsilon$.
   - Since the ability is not observable, we are left with the following simple regression model:
     $\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + u$, where the "ability" is included in the error term u
   - If we estimate this model by OLS, we will obtain a biased and inconsistent estimator of β_1 if $C(\text{educ}, \text{ability}) \neq 0$.
2. Effect of smoking on wages (ignoring the level of education)
3. Effect of smoking on Cancer (ignoring the physical health state)

Endogeneity Example 3: Simultaneity

- It is quite common that the realizations of distict variables are economically related.
- This causes the equation for the dependent variable to be a part of a system of simultaneous equations:
  - Some of the variables on the right side of the equation of interest appear as dependent variables in other equations, and vice versa.

Endogeneity Example 3a: Market Equilibrium Model

- Consider the following system:

$$Y_1 = \alpha_1 Y_2 + \alpha_2 X_1 + u_1 \qquad \text{(Demand)}$$
$$Y_2 = \alpha_3 Y_1 + \alpha_4 X_2 + \alpha_5 X_3 + u_2 \qquad \text{(Supply)}$$

- The endogenous variables are Y_1 =quantity, Y_2 =price are determined by
  - the exogenous variables, X_1 =rent, X_2 =salary, and X_3=interest rate
  - and by the disturbances: u_1 =demand shock, and u_2 =supply shock
- The variables Y_1 and Y_2, both of which appeared on the right hand side of the supply and demand equations, are not orthogonal to their respective shocks:

$$E[Y_1|Y_2, X_1] = \alpha_1 Y_2 + \alpha_2 X_1 + \underbrace{E[u_1|Y_2, X_1]}_{\neq 0}$$

Endogeneity Example 3b: Production Function

- If a company is a profit maximizer, or a cost minimizer
  - The quantities of the inputs are simultaneously determined with the level of output
  - The disturbance, that is, the realization of the technologic shocks, is in general correlated with the quantity of the inputs.

Instrumental Variables

- **Instrumental variables (IV)** approach allows us to get consistent estimators of the population parameters when the OLS estimators are inconsistent (in situations such as omitting a relevant variable, measurement errors, and simultaneity)

Instrumental Variables

- In general, we have to use the model: $Y = \beta_0 + \beta_1 X + \varepsilon$,
  - where $C(X, \varepsilon) \neq 0 \Rightarrow \beta_0$ and β_1 are not the same as the parameters of the linear projection, $L(Y|X)$
  - $\Rightarrow$ The OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) of the linear projection of Y on X are inconsistent estimators of β_0 and β_1: $(x_i = X_i - \bar{X})$, $(y_i = Y_i - \bar{Y})$

Instrumental Variables

$$
\begin{aligned}
p\lim_{n\to\infty}\hat{\beta}_1 &= \frac{p\lim_{n\to\infty}\left(\frac{1}{n}\sum_i x_i y_i\right)}{p\lim_{n\to\infty}\left(\frac{1}{n}\sum_i x_i^2\right)} \\
&= \frac{p\lim_{n\to\infty}\left(\frac{1}{n}\sum_i x_i(\beta_1 x_i+\varepsilon_i)\right)}{p\lim_{n\to\infty}\left(\frac{1}{n}\sum_i x_i^2\right)} \\
&= \beta_1 + \frac{p\lim_{n\to\infty}\left(\frac{1}{n}\sum_i x_i\varepsilon_i\right)}{p\lim_{n\to\infty}\left(\frac{1}{n}\sum_i x_i^2\right)} = \beta_1 + \frac{C(X,\varepsilon)}{V(X)} \neq \beta_1
\end{aligned}
$$

Instrumental Variables: Definition

- In the model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

, where $C(X, \varepsilon) \neq 0$, in order to get consistent estimators of β_0 and β_1, we need **additional information** in the form of additional variables.

- Suppose we have a new variable, Z, which is called the instrumental variable, with the following properties:
  1. It is uncorrelated with the error term: (a) $C(Z, \varepsilon) = 0$;
  2. It is correlated with the endogenous variable X: (b) $C(Z, X) \neq 0$.

Instrumental Variables: IV Estimation in the Simple Model

- Using Z as an instrument, we can obtain consistent estimates β_0 and β_1.
- From (1) we get: $C(Z,Y) = \beta_1 C(Z,X) + C(Z,\varepsilon)$, which, given (a), implies that:

$$\beta_1 = \frac{C(Z,Y)}{C(Z,X)}$$

$$\beta_0 = E[Y] - \beta_1 E[X] = E[Y] - \frac{C(Z,Y)}{C(Z,X)} E[X]$$

Instrumental Variables: IV Estimation in the Simple Model

- Assuming we have a random sample of size n of the population, and by replacing population moments with the sample values in the expression above, (principle of analogy), we get the Instrumental Variable Estimator (IV):

$$\tilde{\beta}_1 = \frac{S_{YZ}}{S_{XZ}} = \frac{\sum_i z_i y_i}{\sum_i z_i x_i}$$

$$\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{X}$$

- $y_i = Y_i - \bar{Y}$, $x_i = X_i - \bar{X}$, and $z_i = Z_i - \bar{Z}$

Instrumental Variables: Properties of IV Estimators in the Simple Model

- Provided that (a) and (b) hold, the IV estimator will be a consistent estimator:

$$
\begin{aligned}
p \lim_{n\to\infty} \tilde{\beta}_1 &= \frac{p \lim_{n\to\infty} \left(\frac{1}{n} \sum_i z_i y_i \right)}{p \lim_{n\to\infty} \left(\frac{1}{n} \sum_i z_i^2 \right)} \\
&= \frac{p \lim_{n\to\infty} \left(\frac{1}{n} \sum_i z_i (\beta_1 z_i + \varepsilon_i) \right)}{p \lim_{n\to\infty} \left(\frac{1}{n} \sum_i z_i^2 \right)} \\
&= = \beta_1 + \frac{C(Z,\varepsilon)}{V(Z)} = \beta_1
\end{aligned}
$$

Instrumental Variables: Properties of IV Estimators in the Simple Model, (a)

- Any instrument or instrumental variable must meet the two properties: (a) and (b). Regarding to this:
  - The condition (a): $C(Z, \varepsilon) = 0$, can not be confirmed. So we must depend our choice of Z on economic behavior or some theory $\Rightarrow$ we must be very careful when choosing Z.

Instrumental Variables: Properties of IV Estimators in the Simple Model, (b)

- The condition (b): $C(Z, X) = 0$ can be tested using the sample data.
  - The simplest way is to consider the linear projection of X on Z: $X = \pi_0 + \pi_1 Z + v$,
  - then, estimate this by OLS and perform the following test: $H_0 : \pi_1 = 0$ vs $H_1 : \pi_1 \neq 0$

- Note: If $Z = X$, we obtain the OLS estimation. That is, when X is exogenous, can be used as its own instrument, and the IV estimator is then identical to the OLS estimator.

Instrumental Variables: The Variance of the IV Estimator

- In general, the IV estimator will have a larger variance than the OLS estimator.

- To see this, let's drive the estimated variance of the IV estimator $\tilde{\beta}_1$, $S^2_{\tilde{\beta}_1}$:

$$S^2_{\tilde{\beta}_1} \equiv \hat{V}\left(\tilde{\beta}_1\right) = \frac{\tilde{\sigma}^2 S^2_Z}{n S^2_{ZX}} = \frac{\tilde{\sigma}^2}{n r^2_{ZX} S^2_X}$$

where $r_{ZX} = \dfrac{S_{ZX}}{S_Z S_X}$, is the sample correlation coefficient between X and Z, which measures the degree of linear relationship between X and Z in the sample.

Instrumental Variables: The Variance of the IV Estimator

- Recall the variance of the OLS estimator of β_1, $\hat{\beta}_1$:

$$S^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{nS^2_X}$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_i \hat{\varepsilon}_i^2$

- If X is really exogenous, then, the OLS estimators are consistent, and $p\lim_{n\to\infty} \hat{\sigma}^2 = p\lim_{n\to\infty} \tilde{\sigma}^2 = \sigma^2 =$
Since $0 < |r_{ZX}| < 1$, this implies that $S^2_{\tilde{\beta}_1} > S^2_{\hat{\beta}_1}$, and the difference will be bigger, the lower the $|r_{ZX}|$ is.

Instrumental Variables: The Variance of the IV Estimator

- Therefore, **if X is exogenous**, using IV estimator instead of the OLS estimator is costly, in terms of efficiency.

- The lower the correlation between Z and X, the greater the difference between the IV variance and the OLS variance, in favor of the OLS variance.
  - In the case where both $\tilde{\beta}_1$ and $\hat{\beta}_1$ are consistent, then asymptotically the IV estimator variance and that of the OLS estimator have the following relationship:

$$p\lim\left(S^2_{\tilde{\beta}_1}/S^2_{\hat{\beta}_1}\right) = 1/\rho^2_{ZX}$$

Instrumental Variables: The Variance of the IV Estimator

- This implies, when $n \to \infty$
- If $\rho_{ZX} = 1\% = 0.01$, then $V(\tilde{\beta}_1) = 10000V(\hat{\beta}_1)$, and therefore, $\sqrt{V(\tilde{\beta}_1)} = 100\sqrt{V(\hat{\beta}_1)}$
- If $\rho_{ZX} = 10\% = 0.1$, then $V(\tilde{\beta}_1) = 100V(\hat{\beta}_1)$, and therefore, $\sqrt{V(\tilde{\beta}_1)} = 10\sqrt{V(\hat{\beta}_1)}$
- Even with a relatively high correlation, $\rho_{ZX} = 50\% = 0.5$, then $V(\tilde{\beta}_1) = 4V(\hat{\beta}_1)$, and therefore, $\sqrt{V(\tilde{\beta}_1)} = 2\sqrt{V(\hat{\beta}_1)}$

Instrumental Variables: The Variance of the IV Estimator

- BUT if **X is endogenous**, the comparison between OLS and the IV variances has no meaning because the OLS estimator is inconsistent.

Instrumental Variables: Inferences with the IV Estimator

- Consider the simple model: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Assume the conditional homoscedasticity assumption holds: $V(\varepsilon|Z) = \sigma^2 = V(\varepsilon)$
- then, it can be shown that:

$$\frac{\tilde{\beta}_1 - \beta_1}{S_{\tilde{\beta}_1}} \overset{\text{asy}}{\sim} N(0,1)$$

Instrumental Variables: Inferences with the IV Estimator

- $S_{\tilde{\beta}_1}$ is the standard error of the IV estimators:

$$\begin{aligned} S^2_{\tilde{\beta}_1} &\equiv \hat{V}\left(\tilde{\beta}_1\right) = \frac{\tilde{\sigma}^2 S^2_Z}{n S^2_{ZX}} \\ &\Rightarrow S_{\tilde{\beta}_1} = \frac{\tilde{\sigma}}{\sqrt{n} S_{ZX}} \end{aligned}$$

- where $\tilde{\sigma}^2 = \frac{1}{n}\sum_i \tilde{\varepsilon}_i^2$, and $\tilde{\varepsilon}_i = Y_i - \left(\tilde{\beta}_0 + \tilde{\beta}_1 X\right) = y_i - \tilde{\beta}_1 x_i$

- This result allows us to construct confidence intervals, and perform hypothesis tests.

Instrumental Variables: Goodness of fit under IV Estimation

- Most econometric programs calculate the R^2 in the IV estimation using the following formula:

$$R^2 = 1 - \frac{\sum_i \tilde{\varepsilon}_i^2}{\sum_i y_i^2}$$

- However, when X and ε are correlated, this formula is not correct. The R^2 of the IV estimation:
  - can be negative if $\sum_i \tilde{\varepsilon}_i^2 > \sum_i y_i^2$.
  - has no natural interpretation, because if $C(X, \varepsilon) \neq 0$, we cannot decompose the variance of Y as $\beta_1^2 V(X) + V(\varepsilon)$.
  - cannot be used to calculate the W^0 test statistic (we have to use SSR).

IV versus OLS

- If our objective is to maximize the R^2, then we should always use the OLS.

- But if our goal is to properly estimate the causal effect of X on Y, that is, β_1, then:
  - If $C(X, \varepsilon) = 0$, we have to use the OLS. (It will be more efficient than any IV estimator Z such that $Z \neq X$).
  - If $C(X, \varepsilon) \neq 0$, OLS will not be consistent, thus using an IV estimator $Z \neq X$ is appropriate. (The goodness of fit, in this context, is not of interest to us).

Instrumental Variables: Inadequate Instruments

- The IV estimator is consistent if (a) $C(Z, \varepsilon) = 0$ and (b) $C(Z, X) \neq 0$.

- If these conditions are not satisfied, the IV estimator has an asymptotic bias, that is bigger than that of the OLS estimator, especially if $|\rho_{XZ}|$ is small.

- We can see this by comparing the $p\lim$ of both estimators: the IV estimator (considering the possibility that Z and ε are correlated) and the OLS estimator (when X is endogenous).

$$
\begin{aligned}
p\lim \tilde{\beta} &= \beta_1 + \frac{C(Z, \varepsilon)}{C(Z, X)} \\
p\lim \hat{\beta} &= \beta_1 + \frac{C(X, \varepsilon)}{C(X)}
\end{aligned}
$$

Instrumental Variables: Inadequate Instruments

- Expressed in terms of correlations and population standard deviations of Z, Y and ε, respectively, are:

$$\begin{aligned} p\lim \tilde{\beta} &= \beta_1 + \frac{\rho_{Z\varepsilon}}{\rho_{ZX}} \frac{\sigma_\varepsilon}{\sigma_X} \\ p\lim \hat{\beta} &= \beta_1 + \rho_{X\varepsilon} \frac{\sigma_\varepsilon}{\sigma_X} \end{aligned}$$

- Therefore, we prefer the IV estimator to the OLS estimator if $\dfrac{\rho_{Z\varepsilon}}{\rho_{ZX}} < \rho_{X\varepsilon}$

Instrumental Variables: Inadequate Instruments

- When Z and X are not correlated at all, then the situation is particularly bad, regardless of whether Z and ε are correlated or not.

- When Z and X have a very small sample correlation r_{ZX}, the problem will be very similar:
  - It may seem like $C(Z, X) = 0$.
  - The estimates will be very inaccurate and may present values that are implausible.

Instrumental Variables: Example of Inadequate Instruments

- Example: Effect of the Mother's cigarette consumption on the birth weight of the baby:
  - Following example illustrates why we should always check whether the endogenous explanatory variable is correlated with the potential instrument.
  - Our estimation results for the effect of several variables on the birth weight, including cigarette consumption of the mother are presented below:

Instrumental Variables: Example of Inadequate Instruments

Model 1: OLS, using observations 1–1388
Dependent variable: bwght

	Coefficient	Std. Error	t-ratio	p-value
PACKS	−0.0837	0.0175	−4.80	0.000
MALE	0.0262	0.0100	2.62	0.009
PARITY	0.0147	0.0054	2.72	0.007
LFAMINC	0.0180	0.0053	3.40	0.001
const	4.6756	0.0205	228.53	0.000

R^2	0.0350

Instrumental Variables: Example of Inadequate Instruments

Where

- LBWGHT = logarithm of the baby's birth weight
- MALE = dummy variable that equals 1 if the baby is male and 0 otherwise,
- PARITY = baby birth order (between siblings)
- LFAMINC = logarithm of family income in thousands of dollars
- PACKS = average number of packs of cigarettes smoked per day during pregnancy.

Instrumental Variables: Example of Inadequate Instruments

- PACKS may be correlated with other health habits and / or with good prenatal care $\Rightarrow$ PACKS and the error term might be correlated.

- A possible instrumental variable for PACKS is the average price of cigarettes per pack: CIGPRICE
  - Assume that CIGPRICE is uncorrelated with the error term (although the availability of the state health care might be correlated with taxes on cigarettes).
  - Economic theory suggests that $C(\text{PACKS, CIGPRICE}) < 0$.

Instrumental Variables: Example of Inadequate Instruments

The Linear projection of PACKS on CIGPRICE and other exogenous variables:

Model 1: OLS, using observations 1–1388
Dependent variable: packs

	Coefficient	Std. Error	t-ratio	p-value
const	0.1374	0.1040	1.32	0.187
CIGPRICE	0.0008	0.0008	1.00	0.317
MALE	−0.0047	0.0159	−0.30	0.766
PARITY	0.0182	0.0089	2.04	0.041
LFAMINC	−0.0526	0.0087	−6.05	0.000

R^2 0.030454

Instrumental Variables: Example of Inadequate Instruments

- The reduced form results indicate that there is no relationship between smoking during pregnancy and the price of cigarettes (that is, the price elasticity of cigarette consumption, which is an addictive good, is not statistically different from zero).

- Since CIGPRICE and PACKS and are not correlated, CIGPRICE does not satisfy the condition $(b) \Rightarrow$ CIGPRICE should not be used as an IV.

- But what happens if we use CIGPRICE as an instrument? The IV estimation results are:

Instrumental Variables: Example of Inadequate Instruments

Model 1: TSLS, using observations 1–1388
Dependent variable: LBWGHT
Instrumented: PACKS
Instruments: CIGPRICE

	Coefficient	Std. Error	t-ratio	p-value
const	4.4679	0.2563	17.43	0.000
PACKS	0.7991	1.1132	0.72	0.474
MALE	0.0298	0.0172	1.73	0.084
PARITY	−0.0012	0.0254	−0.05	0.961
LFAMINC	0.0636	0.0571	1.12	0.265

R^2 . F-statistic 2.50

Instrumental Variables: Example of Inadequate Instruments

- The coefficient of the variable PACKS is very large and it has an opposite sign compared to what is expected. Its standard error is also very large.
- But these estimates are meaningless since CIGPRICE does not satisfy one of the requirements for a valid instrumental variable.

Generalization: The 2SLS Estimator for the Simple Model

- Let the model be: $Y = \beta_0 + \beta_1 X + \varepsilon$ such that $C(X, \varepsilon) \neq 0$,

- We have two possible IVs: Z_1 and Z_2 that meet:

$$C(Z_1, \varepsilon) = 0, \quad C(Z_2, \varepsilon) = 0,$$
$$C(Z_1, X) \neq 0, \quad C(Z_2, X) \neq 0.$$

- We can get two simple and distinct IV estimators: one with Z_1 and another one with Z_2.

- BUT we can also obtain an IV estimator which uses the linear combination of Z_1 and Z_2 as an instrument:
  - We get the estimators via **two-stage least squares estimation** (2SLS)

Generalization: The 2SLS Estimator for the Simple Model

- **Stage 1**: We use OLS to estimate the linear projection of the endogenous explanatory variable X on the instruments Z_1 and Z_2 (known as the reduced form):

$$X = \pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + v \tag{2}$$

  - Let $\hat{\pi}_0$, $\hat{\pi}_1$ and $\hat{\pi}_2$ be the OLS estimators of the reduced form.
  - Then, the estimated values of X are: $\hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z_1 + \hat{\pi}_2 Z_2$

- **Stage 2**: We use OLS to estimate the regression of Y on $\hat{X}$ (hence the name):

$$Y = \beta_0 + \beta_1 \hat{X} + u \tag{3}$$

  - This estimate is equal to estimating β_0 and β_1 by using the IV $Z = \hat{X}$.

Generalization: The 2SLS Estimator for the Simple Model

- Although in both cases the coefficients are the same, the standard errors obtained by 2SLS are incorrect.
  - The reason is that the error term of the second stage, u, includes v, but the standard error should only include the variance of ε.
- The majority of the economic packages have special options for IV estimation, which will conduct the 2SLS, so we don't need to perform the two stages sequentially.

Generalization: The 2SLS Estimator for the Simple Model

- The reduced form (2) decomposes the endogenous explanatory variable into two additive parts:
  - The exogenous part, explained linearly by the instruments, $\pi_0 + \pi_1 Z_1 + \pi_2 Z_2$
  - The endogenous part, the part that could not be explained by the instruments, that is, the error term v

Generalization: The 2SLS Estimator, Interpretation of the Reduced Form

- If the instruments are valid and $V(\varepsilon|Z_1, Z_2)$ is homoscedastic, it can be shown that the 2SLS estimators are **consistent** and **asymptomatically normal**.

- Thus, we can make inferences using an estimator of the population variance

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_i \tilde{u}_i^2$$

  where $\tilde{u}_i^2$ are the residuals from the 2SLS estimation.

- Similar to the simple IV estimator, when instruments are not appropriate (because they're correlated with the error term or weak correlations with the endogenous variable) the 2SLS estimators can be worse than OLS estimator

Generalization: The 2SLS Estimator for the Multiple Model

- For simplicity, consider the following linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where $E[\varepsilon] = 0$, $C(X_1, \varepsilon) = 0$, $C(X_2, \varepsilon) \neq 0$

- That is:
  - X_1 is an exogenous variable
  - But X_2 is endogenous

Generalization: The 2SLS Estimator for the Multiple Model

- Suppose we have an instrumental variable Z such that $C(Z, \varepsilon) = 0$.
- Then, the reduced form is $X_2 = \pi_0 + \pi_1 X_1 + \pi_2 Z + \varepsilon$ In order Z to be a valid instrument, we need $\pi_2 \neq 0$, in other words, $C(Z, X_2) \neq 0$.
- **Very important**: Notice that the reduced form for the endogenous explanatory variable includes the instruments AND all the exogenous explanatory variables of the model.

Generalization: The 2SLS Estimator for the Multiple Model

- **What if we have one more endogenous variable?**
- Suppose $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, where X_1 and X_2 are endogenous, and X_3 is exogenous.
  $E[\varepsilon] = 0$, $C(X_1, \varepsilon) \neq 0$, $C(X_2, \varepsilon) \neq 0$, $C(X_3, \varepsilon) = 0$
- In that case, we will need at least as many additional exogenous variables as endogenous explanatory variables to use as instruments.

Generalization: The 2SLS Estimator for the Multiple Model

- In this case, Let Z_1 and Z_2 be such that $C(Z_1, \varepsilon) = C(Z_2, \varepsilon) = 0$.

- We will have a reduced form equation for each endogenous explanatory variable, where use all the exogenous explanatory variables and the instruments:

$$\begin{aligned} X_1 &= \pi_{10} + \pi_{11} X_3 + \delta_{11} Z_1 + \delta_{12} Z_2 + v_1 \\ X_2 &= \pi_{20} + \pi_{21} X_3 + \delta_{21} Z_1 + \delta_{22} Z_2 + v_2 \end{aligned}$$

  where, at least, $\delta_{11} \neq 0$ and $\delta_{22} \neq 0$ OR $\delta_{12} \neq 0$ and $\delta_{21} \neq 0$

- In general, all instruments are present in the equations for the reduced form of each of the endogenous explanatory variables.

Test of Endogeneity: Hausman Test

- In practice, there are many situations where we do not know whether an explanatory variable is endogenous. One possibility is we can perform a hypothesis test for this.

- For example, for the following model $Y = \beta_0 + \beta_1 X + \varepsilon$, we can consider the following hypothesis:

$$\begin{aligned} H_0 : &\quad C(X, \varepsilon) = 0 \quad \text{(exogenous)} \\ H_1 : &\quad C(X, \varepsilon) \neq 0 \quad \text{(endogenous)} \end{aligned}$$

- How can we perform such a test?

- Suppose we have a valid instrument Z such that $C(Z, \varepsilon) = 0$ and $C(Z, X) \neq 0$

Test of Endogeneity: Hausman Test

- Then, from the reduced form $X = \pi_0 + \pi_1 Z + v$,we can easily get

$$C(X, \varepsilon) = C(\pi_0 + \pi_1 Z + v, \varepsilon) = C(v, \varepsilon) \Rightarrow$$

$$C(X, \varepsilon) = 0 \Leftrightarrow C(v, \varepsilon) = 0$$

- Therefore, if $H_0 : C(X, \varepsilon) = 0$ is true, the coefficient α of $\varepsilon = \alpha v + \xi$ should satisfy $\alpha = 0$, or, equivalently, the coefficient α of

$$Y = \beta_0 + \beta_1 X + \alpha v + \xi \tag{4}$$

Test of Endogeneity: Hausman Test

- Therefore, if you could estimate (4), we could test $H_0 : \alpha = 0$, which is equivalent to $H_0 : C(X, \varepsilon) = 0$.

- In practice, v is not observable. Therefore, it is replaced by the OLS residuals $\hat{v}$ from the reduced model.

- Therefore, the model

$$Y = \beta_0 + \beta_1 X + \alpha \hat{v} + \xi_1 \tag{5}$$

  is estimated by OLS, where $\hat{v} = X - (\hat{\pi}_0 + \hat{\pi}_1 Z)$

- The null hypothesis is that X is exogenous, ie $H_0 : \alpha = 0$.

- Therefore, if we reject that α is 0 in model (5), we conclude that X is endogenous.

Test of Endogeneity: Hausman Test

- **Generalization**:The Hausman test for r potential endogenous variables would be:
  - Estimate r reduced forms corresponding to each of the r variables,
  - Obtain the residuals from each of the reduced forms
  - Include all these r residuals as additional regressors to the main model of Y
  - Test for joint significance of the residuals by W^0 test:

$$W^0 = n\frac{SSR_r - SSR_{un}}{SSR_{un}} \overset{\text{asy}}{\sim} \chi^2_r$$

Test of Endogeneity: Hausman Test

- where
  - SSR_r is the sum of the squares of the residuals of the restricted model
  - SSR_{un} is the sum of the squares of the residues of the full model, which includes the residuals from each of the reduced forms as additional regressors
  - r is the number of potential endogenous variables.

- If the conclusion is in favor of joint significance, then at least one of the explanatory variables is endogenous.

Hausman Test Example

- To illustrate, suppose we have the following model:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$
  where X_1 and X_2 are potentially endogenous, and X_3 is exogenous.

- We will need at least two instruments Z_1 and Z_2 such that $C(Z_1, \varepsilon) = C(Z_2, \varepsilon) = 0$.

- Then, we have the following two reduced form equations:

$$\begin{aligned} X_1 &= \pi_{10} + \pi_{11} X_3 + \delta_{11} Z_1 + \delta_{12} Z_2 + v_1 \\ X_2 &= \pi_{20} + \pi_{21} X_3 + \delta_{21} Z_1 + \delta_{22} Z_2 + v_2 \end{aligned}$$

Hausman Test Example

- The hypothesis of exogeneity is now
  $H_0 : C(X_1, \varepsilon) = C(X_2, \varepsilon) = 0$

- Equivalently, we can use the extended regression with the residuals:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \alpha_1 \hat{v}_1 + \alpha_2 \hat{v}_2 + \xi$, where $\hat{v}_1$ and $\hat{v}_2$ are the residuals from the reduced forms for X_1 and X_2, respectively

- The null hypothesis can be written as $H_0 : \alpha_1 = \alpha_2 = 0$.

- To test this hypothesis with two restrictions, we should estimate both the restricted and unrestricted models, and calculate the sums of squares of residuals, and calculate the test statistic W^0, which is distributed, approximately, χ^2

Testing Overidentification Restrictions: Sargan Test

- If we have only one instrumental variable for each endogenous explanatory variable, we say that the model is "exactly identified".
  - In this case, we can not test the condition (a) ($C(Z, \varepsilon) = 0$).

- BUT if we have more instrumental variables than the potentially endogenous explanatory variables, we say that the model is "overidentified".
  - In this case, we can test whether any of the IVs is correlated with the error term.

- Suppose we have r potentially endogenous explanatory variables and q instruments, where $q > r$
  - $(q - r)$, then, is the number of overidentification restrictions (number of "extra" instruments).

Testing Overidentification Restrictions: Sargan Test

- We do not observe the errors of the equation of interest, u
- But we can implement a test based on residuals of the 2SLS, $\tilde{u}$ (sample realizations of u).
- The test procedure:
  1. Estimate the equation of interest using 2SLS and obtain the 2SLS residuals, $\tilde{u}$.
  2. Regress $\tilde{u}$ on all "exogenous" variables (including IVs). Obtain the R^2 of this regression, say $R^2_{\tilde{u}}$
  3. Under the null hypothesis that none of the IVs is correlated with $\tilde{u}$, we have $nR^2_{\tilde{u}} \overset{\text{asy}}{\sim} \chi^2_{q-r}$

Testing Overidentification Restrictions: Sargan Test

- Intuition of the test: the fitted values of this auxiliary regression, $\hat{\tilde{u}}$, have zero mean and variance $\sigma^2_{\tilde{u}}$. Under conditional homoscedasticity, asymptotically $\sum_i \frac{\hat{\tilde{u}}^2}{\sigma^2_{\tilde{u}}}$ is the sum of squares of $N(0,1)$ random variables, out of which only $q-r$ are independent. Hence, this expression follows an asymptotic χ^2 distribution with $q-r$ degrees of freedom
- In practice, we will replace $\sigma^2_{\tilde{u}}$ by its estimator $s^2_{\tilde{u}} = \frac{1}{n}\sum_i \hat{\tilde{u}}^2$
- Therefore, this statistic also has the same distribution

$$\sum_i \frac{\hat{\tilde{u}}^2}{\frac{1}{n}\sum_i \hat{\tilde{u}}^2} = nR^2_{\tilde{u}}$$

Testing Overidentification Restrictions: Sargan Test

- If $nR^2_{\tilde{u}}$ exceeds the critical value of χ^2_{q-r} at the significance level, we reject H_0 and conclude that there is no evidence for exogeneity.
- Another thing is that this test does not distinguish which variable is the reason for rejecting the null of no correlation.
- This test is also known as Hansen-Sargan test.

Example: The Wage Equation

- Let the model be: $\ln(\text{wage}) = \beta_0 + \beta_1 \text{ educ} + \beta_2 \text{ cap} + \varepsilon$ where $\beta_2 \neq 0$ (ie, the variable capacity, which is unobserved, is a relevant variable).
- If we estimate the model using OLS: $\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + u$, with $u = \beta_2 \text{cap} + \varepsilon$, we will have inconsistent estimates.
- If we have an instrumental variable for educ, we can estimate the model by IV estimation.

Example: The Wage Equation

- What are the conditions for the IV in order our estimators to be consistent?
  1. $C(Z, u) = 0$: IV should not be correlated with the capacity or other unobserved variables that affect wages.
  2. $C(Z, \text{educ}) \neq 0$: IV should be correlated with education.

- Some examples of possible instruments (Z) for education are, mother's education, father's education, number of siblings, distance between school and home, etc.

Example: The Wage Equation, OLS estimation

- We have a sample of 336 married women.
- The OLS estimation results are as follows:
$$\widehat{\ln(\text{wage})} = \underset{(0.120)}{0.286} + \underset{(0.009)}{0.083}\ \text{educ}$$
- The interpretation is that an additional year of schooling, on average, increases the wages by 8.3%.

Example: The Wage Equation, IV estimation (a single instrument)

- Possible instrument: Father's education $\Rightarrow$ educf
- Reduced form:
  $\widehat{\text{educ}} = \underset{(0.198)}{9.799} + \underset{(0.021)}{0.282}\ \text{educf},\ R^2 = 0.196$
- t test statistic for the instrumental variable is
  $t = \dfrac{0.282}{0.021} \approx 13.52$, that is, we reject $H_0 : \pi_1 = 0$
- Therefore, the education of women (educ), is significantly correlated with the education of their father (educf).

Example: The Wage Equation, IV estimation (a single instrument)

- The IV Estimate: $\widehat{\ln(\text{wage})} = \underset{(0.289)}{0.363} + \underset{(0.023)}{0.076}\ \text{educ}$
- By comparing the results of OLS with that of IV, we see that the OLS estimate is higher, which is consistent with a positive bias due to omitting capacity.
- Notice that the standard error of the IV estimators are substantially higher than those of the OLS estimators as the theory suggests (but education still remains significant)

Example: The Wage Equation, IV estimation (a single instrument)

- **Hausman Test**

- From the reduced form, we get the variable $\hat{v}$, the residual of the estimated equation:
  $\hat{v} = educ - (9.799 + 0.282 \text{ educf})$, and perform the regression:
  $\ln(\text{wage}) = \beta_0 + \beta_1 \text{ educ} + \alpha\hat{v} + e$, and obtain:
  $\widehat{\ln(\text{wage})} = \hat{\beta}_0 + \hat{\beta}_1 \text{ educ} + \underset{(0.024)}{0.007\hat{v}}$

- Then we test $H_0 : \alpha = 0$ (educ is exogenous), with
  $t = 0.007/0.024 \approx 0.3 \Rightarrow$ DNR H_0 (DNR exogeneity)

Example: The Wage Equation, IV estimation (multiple instruments)

- Suppose that in addition to the father's education, we also have the mother's education, educm, as an instrument.
- Now, the reduced form is:
$$\widehat{\text{educ}} = \underset{(0.226)}{8.976} + \underset{(0.025)}{0.183}\ \text{educf} + \underset{(0.026)}{0.183}\ \text{educm},\ R^2 = 0.245$$
- Test statistic for the joint significance of educf and educm is $W^0 \approx 243.3$, and has asymptotic χ^2_2 distribution
- The IV Estimate using two IVs is:
$$\widehat{\ln(\text{wage})} = \underset{(0.272)}{0.396} + \underset{(0.023)}{0.074}\ \text{educ}$$

Example: The Wage Equation, IV estimation (multiple instruments)

- In order to implement the Hausman test, we take the residuals $\hat{v}$ from the reduced form:
  $\hat{v} = \text{educ} - (8.976 + 0.183\ \text{educf} + 0.183\ \text{educm})$

- And use OLS to estimate the regression model:$\ln(\text{wage}) = \beta_0 + \beta_1\ \text{educ} + \alpha\hat{v} + e$, and obtain:
  $\widehat{\ln(\text{wage})} = \hat{\beta}_0 + \hat{\beta}_1\ \text{educ} + \underset{0.022}{0.0107}\hat{v}$

- hen we test $H_0 : \alpha = 0$ (educ is exogenous), with
  $t = 0.0107/0.022 \approx 0.5 \ \Rightarrow$ DNR H_0 (DNR exogeneity)

Example: The Wage Equation, IV estimation (multiple instruments)

- **Sargan Test**
- Following the latter case, we have two instruments (educf and educm) for a potentially endogenous variable (educ), with $q - r = 1$ overidentification restrictions.
- We can, therefore, partially assess the validity of instruments (that is, the null hypothesis of exogeneity) by testing that the IVs have no correlation with the error term using a Sargan test.

Example: The Wage Equation, IV estimation (multiple instruments)

- To do this, we, first, calculate the residuals of the 2SLS estimation
  $\tilde{u} = \ln(\text{wage}) - (0.396 + 0.074\ \text{educ})$

- Then, perform the auxiliary regression of the residuals on the exogenous variables and the instruments:
  $\hat{\tilde{u}} = \underset{(0.0703)}{0.0054} + \underset{(0.0075)}{0.0020}\ \text{educf} - \underset{(0.0081)}{0.0025}\ \text{educm},\ R^2 = 0.0003$

Example: The Wage Equation, IV estimation (multiple instruments)

- Therefore, the test statistic, $nR^2_{\tilde{u}} = 0.1008$, has a value less than the critical value from the χ^2_1 distribution. Therefore we do not reject the null of no correlation between the instruments and the error term of the model.

- That is, there is no evidence against the validity of the instruments.

Final Points

- In practice, in many situations, it is difficult to find valid instruments, that is, a variable which is not included in the model, and is highly correlated with potentially endogenous explanatory variables, and is not correlated with the error term of the model.

- The problem is that for the economic variables, most of the available variables are results of agents' decisions, and therefore, their exogeneity is questionable.

Final Points

- Ideally, we would like to use, as instrumental variables, variables that are given to the agent (i.e., exogenous). We have seen an example the price of cigarettes as instrument for the number of packets of cigarettes consumed.

- The problem is that in many contexts (like the example we mentioned above) the quality of the instrument is reduced by the weak correlation between the IV and the endogenous explanatory variable we use the IV for.

Final Points

- EXAMPLE: The availability of information about previous realizations of the variables of interest opens interesting possibilities for possible instruments. Thus, an endogenous explanatory variable could be replaced by the previous realizations of the same variable (since the previous realizations are given before the current values are realized)

  - For example, for the consumption and permanent income equation, we use income since the forme is not available. But, we could also use the disposable income corresponding to the previous period as IV.
  - If we analyze this relationship with aggregated time series, we could use the lagged disposable income as an instrument.
  - If we analyze the relationship with longitudinal family data, lagged disposable income of each family could be used as an instrument.