Recitation 8: MLE and IV estimation

Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation

November 10th, 2022

Maximum likelihood estimation

Different approach to regression: Maximum likelihood estimators

- Both probit and logit are nonlinear: β_0, β_1 parameters are no longer in linear relationship with the $X_i's$ and subsequently Y_i's
- A **likelihood function** is the conditional density of $Y_1, ..., Y_n$ given $X_1, ..., X_n$ that is treated as the function of the unknown parameters (β_0, β_1 in our case)
- What we are trying to do here is to find the values of β_i's that best matches the values of X_i's and Y_i's
- **Maximum likelihood estimators** is the value of β_i's that best describes the data and maximizes the value of the likelihood function

Maximum likelihood estimators in practice

- Assume Y_i's are IID normal with mean μ and standard error σ (both are unknown)
- The joint probability of Y_i's are (our likelihood function)

$$\begin{aligned}
\Pr(Y_1 = y_1, ..., Y_n = y_n|\mu, \sigma) &= \Pr(Y_1 = y_1|\mu, \sigma) \times .. \times \Pr(Y_n = y_n|\mu, \sigma) \\
&= \prod_{i=1}^{n} f(y_i|\mu, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\mu)^2}{2\sigma^2}} \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(Y_i-\mu)^2}{2\sigma^2}}
\end{aligned}$$

Maximum likelihood estimators in practice

- Calculation is made easier by using log-likelihood functions (take logs to likelihood functions)

$$-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{2\sigma^2}$$

- We differentiate the above with respect to μ and σ to find the MLE of these parameters.
- This gets us

$$\mu_{MLE}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

$$\sigma^2_{MLE}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_{MLE})^2$$

Instrumental variables

Why use IV? Deal with $E[u_i|X_i] \neq 0$

- Recall that OLS estimates can be biased when we have omitted variable bias, measurement error, and simultaneity bias
- Instrumental variable: Allows us to eliminate bias
  - When you have an independent variable X, there are parts of this variable that are correlated with u and the other parts that are independent of u.
  - If we are able to find Z that is correlated with X but not with u, using this Z variable would allow you to sort variable X into what is correlate with u and what is not.
  - Then, using the part that is not correlated with u, variable Z allows us to get unbiased estimates.
- Uses of IV
  - Endogenous variable: X determined as choice, not as given
  - Simultaneity bias: Y can lead to changes in X
  - Omitted variable bias: Unincorporated determinant of Y
  - Measurement error in X

Association vs causation

- Association refers to any type of statistical dependence
  - It mostly refers to correlation, which is

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

- Running an OLS is practically equivalent to finding the correlation
  - From the population regression,

$$\beta_1 = \frac{E[X - E[X]]E[Y - E[Y]]}{(E[X - E[X]])^2} = \text{corr}(X, Y)\sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}$$

    implying that β_1 is obtained by multiplying $corr(X, Y)$ with some constant
  - Same correlation direction (variances are nonnegative)
  - Not good for finding causal effect of X on Y: Switching the two would give same results

What conditions should IV satisfy?

- Must satisfy
  - **Relevance**: Variable Z satisfies relevancy condition if $cov(X, Z) \neq 0$
  - **Exogeneity**: Variable Z satisfies exogeneity condition if $cov(Z, u) = 0$
- In words,
  - Variable Z should be somewhat correlated with the variable X
  - Variable Z should not be correlated with u
  - (For exogeneity): Variable Z should affect Y only through X, or when X is controlled for, Z alone should not affect Y (**exclusion**)
- More on exclusion (on model $Y = \beta_0 + \beta_1 X + u$)

$$
\begin{aligned}
cov(Z, u) &= cov(Z, Y - \beta_0 - \beta_1 X) = 0 \\
&= cov(Z, Y) - cov(Z, \beta_0) - cov(Z, \beta_1 X) = 0 \\
&\implies cov(Z, Y) = cov(Z, \beta_1 X)
\end{aligned}
$$

This condition means that Z is correlated with Y *only through* X

2SLS estimation

- **Regress with X as dependent, Z as independent variable.** Regress

$$X = \delta_0 + \delta_1 Z + v$$

  From this regression, obtain the predicted values of X, denoted as $\hat{X} = \hat{\delta}_0 + \hat{\delta}_1 Z$. This $\hat{X}$ is the part that is related with Z but is uncorrelated with u.

- **Regress with Y as dependent, $\hat{X}$ as independent variable.** Regressing with this $\hat{X}$ will satisfy the $E[u|\hat{X}] = 0$ condition, as the $\hat{X}$ is uncorrelated with u. Thus, your regression equation looks like this:

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

  Then you run a OLS regression on the above equation and get the 2SLS estimator $\hat{\beta}_{\text{TSLS}}$.

Estimating using a covariance method

- Note that
$$cov(Z, Y) = cov(Z, \beta_1 X) \implies cov(Z, Y) = \beta_1 cov(Z, X)$$
- From this, we can get
$$\beta_1 = \frac{cov(Z, Y)}{cov(Z, X)}$$
where division is possible because we require a relevancy condition ($cov(Z, X) \neq 0$)
- Taking the sample counterparts, we get our IV estimator
$$\hat{\beta}_{TSLS} = \frac{s_{ZY}}{s_{ZX}}$$

Reduced form method

- Denote

$$X = \pi_0 + \pi_1 Z + v \text{ (where } cov(Z, v) = 0)$$
$$Y = \gamma_0 + \gamma_1 Z + w \text{ (where } cov(Z, w) = 0)$$

- Rewrite the first equation in terms of Z and get

$$Z = \frac{X}{\pi_1} - \frac{\pi_0}{\pi_1} - \frac{v}{\pi_1}$$

- Then plug this into the second equation. Reorganizing this equation, you should get

$$Y = \left(\gamma_0 - \frac{\pi_0 \gamma_1}{\pi_1}\right) + \left(\frac{\gamma_1}{\pi_1}\right) X + \left(w - \frac{\gamma_1}{\pi_1} u\right)$$

- As a result, β_1 from the equation with X as independent variable is $\beta_1 = \frac{\gamma_1}{\pi_1}$.

IV estimate is consistent!

- Consistency: 2SLS can be written as

$$\hat{\beta}_{1,\text{TSLS}} = \frac{s_{zy}}{s_{zx}} \simeq \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}$$

As $n \to \infty$, we can show that $\hat{\beta}_{1,\text{TSLS}} \to \beta_1$

  - $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y}) \xrightarrow{p} cov(Z, Y)$
  - $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X}) \xrightarrow{p} cov(Z, X)$
  - Note that $cov(Z, Y) = cov(Z, \beta_0 + \beta_1 X + u) = \beta_1 cov(Z, X) + cov(Z, u)$
  - If Z is a valid IV, $cov(Z, u) = 0$
  - So $\hat{\beta}_{1,\text{TSLS}} \to \frac{\beta_1 cov(Z,X)}{cov(Z,X)} = \beta_1$ QED

Multivariate case and key assumptions

- Suppose that we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \delta_1 W_{1i} + ... + \delta_l W_{li} + u_i$$

  where X variables are endogenous and W variables are exogenous.

- Assume that we have found a total of m (not necessarily equal to k) variables that could qualify as IVs, all of them satisfying

  IV1 $E[u_i|W_{1i}, ..., W_{li}] = 0$ (At least for exogenous variables, this is satisfied)

  IV2 $(Y_i, X_{1i}, .., X_{ki}, W_{1i}, .., W_{li}, Z_{1i}, ..., Z_{mi})$ are IID

  IV3 The Y, X, W, Z variables all have nonzero finite 4th moments

  IV4 The instruments are valid. That is $cov(Z_{ji}, u_i) = 0$ for all $j = 1, ..., m$ and relevancy conditions are satisfied for all Z's.

Identification issues on multivariate case

- A parameter is **identified** if different values of the parameter produce different distributions of the data.
- In other words, there is a one-to-one matching of the parameters and the distributions.
- If it is the case that the same distribution can be obtained from different parameter values, we say that the parameters are not identified
- If we have k endogenous regressors and m IV's
  - **Just-identified**: When $m = k$. There are just enough instruments to identify k endogenous variables
  - **Overidentified**: When $m > k$. There are more than enough instruments.
  - **Underidentified**: When $m < k$. There are not enough instruments. The coefficients for X's will not be identified
- Need at least as much instrumental variables as the number of endogenous regressors you have

IV estimators are normally distributed (but only in large samples)

- Break down the 2SLS estimators into

$$\hat{\beta}_{\text{TSLS}} = \frac{s_{zy}}{s_{zx}} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})Y_i}{\sum_{i=1}^{n}(Z_i - \bar{Z})X_i}$$

$$= \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(\beta_0 + \beta_1 X_i + u_i)}{\sum_{i=1}^{n}(Z_i - \bar{Z})X_i} = \beta_1 + \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\sum_{i=1}^{n}(Z_i - \bar{Z})X_i} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})X_i}$$

- After subtracting both sdies by β_1 and multiplying both sides by $\sqrt{n}$, we get

$$\sqrt{n}(\hat{\beta}_{\text{TSLS}} - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})X_i}$$

- Denominator $\xrightarrow{p} cov(Z, X)$ by (weak) law of large numbers
- Numerators $\sim N(0, var[(Z - \mu_Z)u])$ by central limit theorem.
- Thus, $\hat{\beta}_{\text{TSLS}}$ has a normal distribution.

We can directly test relevance condition

- Assume we have

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

where only X_i is endogenous.
- Suppose we have m instruments. Then our first stage equation looks like

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_m Z_{mi} + \pi_{(m+1)i} W_{1i} + \ldots + \pi_{(m+r)i} W_{ri} + v_i$$

- We say our instrument is relevant if at least one of $\pi_1, \ldots, \pi_m$ is statistically nonzero.
- Run the F-test with these null and alternative hypothesis

$$H_0 : \pi_1 = \ldots = \pi_m = 0 \text{ vs. } H_1 : \neg H_0$$

- By rule of thumb, we want F-statistics to be larger than 10. Otherwise, we have a weak instrument problem (and IV estimator may not be normally distributed even in large numbers)

We can get a taste of exogeneity condition with many instruments

- Consider the following case: We want to regress

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \delta_1 W_{1i} + \ldots + \delta_l W_{li} + u_i$$

  and we found m possible candidates for instrumental variables, $Z_1, \ldots, Z_m$
- Overidentifying restriction test we conduct here is a J-test that uses TSLS estimators and not the hypothesized β values. The steps are as follows
  1. Regress the above equation using 2SLS.Obtain the residuals $\hat{u}_i = Y_i - \widehat{Y}_i$
  2. Regress residuals onto instruments and other controls, namely

$$\hat{u}_i = \alpha_0 + \alpha_1 Z_{1i} + \ldots + \alpha_m Z_{mi} + \alpha_{m+1} W_{1i} + \ldots + \alpha_{m+l} W_{li} + v_i$$

  3. Compute the F-statistics from $H_0 : \alpha_1 = \ldots = \alpha_m = 0$ vs. $H_1 : \neg H_0$
  4. Derive the J statistics as follows $J = m \times F$. J follows χ^2_{m-k} distribution
  5. If you have an endogenous IV, you will end up rejecting the null
  6. Then, you need to make a guess on which instrument is violating the exogeneity condition, drop those, and redo the above procedure. (Good luck!)

Overidentification test as a ‘distance’ between different IVs

- Assume a case with one endogenous variable X and two IVs (Z_1, Z_2)
- The estimates for the coefficient for X are

$$\hat{\beta}_{Z_1} = \frac{cov(Z_1, Y)}{cov(Z_1, X)}, \; \hat{\beta}_{Z_2} = \frac{cov(Z_2, Y)}{cov(Z_2, X)}$$

  The numbers look different, although both $\hat{\beta}$’s are suppose to be the same coefficient estimating impact of X_1 on Y.
- The overidentification test checks whether the differences between $\hat{\beta}_{Z_1}, \hat{\beta}_{Z_2}$ are large.
  - If they converge to same item, we do not have to worry. Otherwise, one or more IV might be faulty

What if we have just enough IVs?

- To be very honest, there is not much we can do in terms of rigorous testing approach. This is usually the area of judgement call.
- One way to do this is to use the approach of exclusion restriction - you argue that the instrumental variable Z affects Y only through X.
  - If Z affects Y directly without X, then the exclusion restriction fails.
- You try to persuade others based on logic, previous practices, or intuition (usually a combination of all three).