Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 13: LASSO for Instrumental Variable Models

Instructor: Ma, Jun

Renmin University of China

December 9, 2021

Instrumental variable

- Consider

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + e_i, \\ \mathrm{E}[e_i] &= 0 \\ \mathrm{Cov}[X_i, e_i] &\neq 0. \end{aligned}$$

- An instrument is an variable Z_i which satisfies the following conditions:
  1. The IV is exogenous: $\mathrm{Cov}[Z_i, e_i] = 0$.
  2. The IV determines the endogenous regressor: $\mathrm{Cov}[Z_i, X_i] \neq 0$.
- When an IV variable satisfying those conditions is available, it allows us to estimate the effect of X on Y consistently:

$$\begin{aligned} \mathrm{Cov}[Y_i, Z_i] &= \beta_1 \mathrm{Cov}[X_i, Z_i] + \mathrm{Cov}[e_i, Z_i] \\ &= \beta_1 \mathrm{Cov}[X_i, Z_i] \Longrightarrow \beta_1 = \frac{\mathrm{Cov}[Y_i, Z_i]}{\mathrm{Cov}[X_i, Z_i]}. \end{aligned}$$

Sources of endogeneity

There are several possible sources of endogeneity:

1. Omitted explanatory variables.
2. Simultaneity.
3. Errors in variables.

All result in regressors correlated with the errors.

Omitted explanatory variables

- Suppose that the true model is

$$\ln Wage_i = \beta_0 + \beta_1 Education_i + \beta_2 Ability_i + V_i,$$

  where V_i is uncorrelated with $Education$ and $Ability$.
- Since $Ability$ is unobservable, the econometrician regresses $\ln Wage$ against $Education$, and $\beta_2 Ability$ goes into the error part:

$$\begin{aligned} \ln Wage_i &= \beta_0 + \beta_1 Education_i + U_i, \\ U_i &= \beta_2 Ability_i + V_i. \end{aligned}$$

- $Education$ is correlated with $Ability$: we can expect that $\mathrm{Cov}\left(Education_i, Ability_i\right) > 0, \beta_2 > 0$, and therefore $\mathrm{Cov}\left(Education_i, U_i\right) > 0$.

Simultaneity

- Consider the following demand-supply system:

$$\begin{array}{ll} \text{Demand:} & Q^d = \beta_0^d + \beta_1^d P + U^d, \\ \text{Supply:} & Q^s = \beta_0^s + \beta_1^s P + U^s, \end{array}$$

  where: Q^d =quantity demanded, Q^s =quantity supplied, P=price.
- The quantity and price are determined simultaneously in the equilibrium:

$$Q^d = Q^s = Q.$$

- Note that Q^d and Q^s are not observed separately, we observe only the equilibrium values Q.

$$
\begin{aligned}
Q^d &= \beta_0^d + \beta_1^d P + U^d, \\
Q^s &= \beta_0^s + \beta_1^s P + U^s, \\
Q^d &= Q^s = Q.
\end{aligned}
$$

- Solving for P, we obtain

$$0 = \left(\beta_0^d - \beta_0^s\right) + \left(\beta_1^d - \beta_1^s\right) P + \left(U^d - U^s\right),$$

  or

$$P = -\frac{\beta_0^d - \beta_0^s}{\beta_1^d - \beta_1^s} - \frac{U^d - U^s}{\beta_1^d - \beta_1^s}.$$

- Thus,

$$\mathrm{Cov}\left(P, U^d\right) \neq 0 \text{ and } \mathrm{Cov}\left(P, U^s\right) \neq 0.$$

  The demand-supply equations cannot be estimated by OLS.

- Consider the following labour supply model for married women:

$$Hours_i = \beta_0 + \beta_1 Children_i + \text{Other Factors} + U_i,$$

  where $Hours$=hours of work, $Children$=number of children.
- It is reasonable to assume that women decide simultaneously how much time to devote to career and family.
- Thus, while we may be mainly interested in the effect of family size on labour supply, there is another equation:

$$Children_i = \gamma_0 + \gamma_1 Hours_i + \text{Other Factors} + V_i,$$

  and $Children$ and $Hours$ are determined simultaneously in an equilibrium.
- As a result, $\text{Cov}\left(Children_i, U_i\right) \neq 0$, and the effect of family size cannot be estimated by OLS.

Errors in variables

- Consider the following model:

$$Y_i = \beta_0 + \beta_1 X_i^* + V_i,$$

where X_i^* is the true regressor.

- Suppose that X_i^* is not directly observable. Instead, we observe X_i that measures X_i^* with an error ε_i:

$$X_i = X_i^* + \varepsilon_i.$$

- Since X_i^* is unobservable, the econometrician has to regress Y_i against X_i.

$$
\begin{aligned}
X_i &= X_i^* + \varepsilon_i, \\
Y_i &= \beta_0 + \beta_1 X_i^* + V_i.
\end{aligned}
$$

- The model for Y_i as a function of X_i can be written as

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 (X_i - \varepsilon_i) + V_i \\
&= \beta_0 + \beta_1 X_i + V_i - \beta_1 \varepsilon_i,
\end{aligned}
$$

  or

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + e_i, \\
e_i &= V_i - \beta_1 \varepsilon_i.
\end{aligned}
$$

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + e_i, \\
e_i &= V_i - \beta_1 \varepsilon_i, \\
X_i &= X_i^* + \varepsilon_i.
\end{aligned}
$$

- We can assume that

$$
\mathrm{Cov}\left[X_i^*, V_i\right] = \mathrm{Cov}\left[X_i^*, \varepsilon_i\right] = \mathrm{Cov}\left[\varepsilon_i, V_i\right] = 0.
$$

- However,

$$
\begin{aligned}
\mathrm{Cov}\left[X_i, e_i\right] &= \mathrm{Cov}\left[X_i^* + \varepsilon_i, V_i - \beta_1 \varepsilon_i\right] \\
&= \mathrm{Cov}\left[X_i^*, V_i\right] - \beta_1 \mathrm{Cov}\left[X_i^*, \varepsilon_i\right] \\
&\quad + \mathrm{Cov}\left[\varepsilon_i, V_i\right] - \beta_1 \mathrm{Cov}\left[\varepsilon_i, \varepsilon_i\right]
\end{aligned}
$$

- Thus, X_i is enodgenous and β_1 cannot be estimated by OLS.

Example: Compulsory schooling laws and return to education

- Angrist and Krueger, 1991, *QJE*, suggested using school start age policy to estimate β_1 in $\ln Wage_i = \beta_0 + \beta_1 Education_i + \beta_2 Ability_i + V_i$.
- We need to find an IV variable Z such that $\text{Cov}\left(Ability_i, Z_i\right) = 0$ and $\text{Cov}\left(Education_i, Z_i\right) \neq 0$.
- They argue that due to compulsory schooling laws, the season of birth variable satisfies the IV conditions:
  - A child has to attend the school until he reaches a certain drop-out age.
  - Students born in the first quarter of the year, reach the legal drop-out age before their classmates who were born later in the year.
  - The quarter of birth dummy variable is correlated with education.
  - The quarter of birth is uncorrelated with ability.

Example: Sibling-sex composition and labor supply

- Angrist and Evans, 1998, *AER*, argue that the parents' preferences for a mixed sibling-sex composition can be used to estimate β_1 in $Hours_i = \beta_0 + \beta_1 Children_i + \ldots + U_i$.
- We need to find an IV Z such that $\text{Cov}\left[U_i, Z_i\right] = 0$ and $\text{Cov}\left(Children_i, Z_i\right) \neq 0$.
- Consider a dummy variable that takes on the value one if the sex of the second child matches the sex of the first child.
  - If the parents prefer a mixed sibling-sex composition, they are more likely to have another child if their first two children are of the same sex.
  - The same-sex dummy is correlated with the number of children.
  - Since sex mix is randomly determined, the same sex dummy is exogenous.

2SLS estimation with many IVs

- We consider the simple model (0 intercept):

$$\begin{aligned} Y_i &= \alpha D_i + U_i \\ \mathrm{E}[U_i] &= 0 \\ \mathrm{Cov}[D_i, U_i] &\neq 0. \end{aligned}$$

- Suppose that we have l IVs $Z_i \in \mathbb{R}^l$ which satisfies $\mathrm{E}[U_i \mid Z_i] = 0$.
- Note that this assumption is stronger than what we typically assume ($\mathrm{Cov}[U_i, Z_i] = 0$).
- The first-stage of 2SLS uses the linear projection of D_i on Z_i:

$$\begin{aligned} D_i &= Z_i^\top \pi + V_i \\ \mathrm{E}[Z_i V_i] &= 0. \end{aligned}$$

- Then,

$$\begin{array}{c} Y_i = \alpha D_i + U_i \\ D_i = Z_i^\top \pi + V_i \end{array} \implies Y_i = \alpha Z_i^\top \pi + \alpha V_i + U_i \ .$$

Regression of Y_i on $Z_i^\top \pi$ consistently estimates α.

- $\mathbf{Z}$: the $n \times l$ matrix of IVs; $\mathbf{D} = (D_1, D_2, ..., D_n)^\top$; $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^\top$; $\mathbf{U} = (U_1, U_2, ..., U_n)^\top$; $\mathbf{V} = (V_1, V_2, ..., V_n)^\top$.
- Since π is unknown, we replace it with $\widehat{\pi} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{D}$:

$$\widehat{\alpha}^{\mathsf{2sls}} = \frac{\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}}\mathbf{Y}}{\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}}\mathbf{D}} = \alpha + \frac{n^{-1}\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}}\mathbf{U}}{n^{-1}\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}}\mathbf{D}},$$

  where $\boldsymbol{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$.
- $n^{-1}\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}}\mathbf{D}$ is less variable when n and l are both large. The bias of $\widehat{\alpha}^{\mathsf{2sls}}$ mainly depends on the numerator $n^{-1}\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}}\mathbf{U}$.
- Suppose that $\mathrm{E}[\mathbf{U}\mathbf{V}^\top \mid \mathbf{Z}] = \sigma_{UV}\mathbf{I}_n$ and $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_l$, then

$$\mathrm{E}\left[\frac{1}{n}\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}}\mathbf{U} \mid \mathbf{Z}\right] = \sigma_{UV}\frac{l}{n}.$$

- When the number of IVs is large and comparable to the sample size n, the bias can be substantial.

- In the context of a small and fixed number of IVs, adding one more IV reduces the variance of the 2SLS estimator.
- However, if there are too many IVs, the bias becomes non-negligible and we have to selection a small subset of best IVs out of the long list of potential IVs.
- LASSO is used for data-driven IV selection.