December 11, 2018

Instrumental Variables

Endogeneity

Consider a partitioned regression model:

$$\begin{aligned} Y_i &= \boldsymbol{X}_i'\beta + e_i \\ &= \boldsymbol{X}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{X}_{2i}'\boldsymbol{\beta}_2 + e_i, \end{aligned} \tag{1}$$

where $\boldsymbol{X}_{1i}$ is a k_1-vector and $\boldsymbol{X}_{2i}$ is a k_2-vector of random regressors, $\boldsymbol{\beta}_1$ is $k_1 \times 1$ and $\boldsymbol{\beta}_2$ is $k_2 \times 1$ vectors of unknown parameters, $k_1 + k_2 = k$. We assume that $\boldsymbol{X}_{1i}$ is *endogenous*:

$$\mathbb{E}\left(\boldsymbol{X}_{1i} e_i\right) \neq \mathbf{0},$$

as opposed to *(weakly) exogenous* $\boldsymbol{X}_{2i}$'s:

$$\mathbb{E}\left(\boldsymbol{X}_{2i} e_i\right) = \mathbf{0}.$$

(The assumption $\mathbb{E}\left(e_i | \boldsymbol{X}_{2i}\right) = \mathbf{0}$ is called *strong exogeneity.*) Sources of endogeneity:

- **Omitted variables**. Consider the wage equation:

$$\begin{aligned} \log Wage_i &= \alpha + \beta Education_i + \gamma Gender_i + \delta Ability_i + V_i \\ &= \alpha + \beta Education_i + \gamma Gender_i + U_i. \end{aligned}$$

  Since ability is unobservable, it "goes" to the residuals $U_i = \delta Ability_i + V_i$. We can assume that the gender variable is exogenous, however, education is correlated with the ability, and, therefore, education is endogenous.

- **Errors in variables.** Suppose that the true model is

$$Y_i = \widetilde{\boldsymbol{X}}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{X}_{2i}'\boldsymbol{\beta}_2 + v_i,$$

  however, $\widetilde{\boldsymbol{X}}_{1i}$ is unobservable. Instead, the econometrician observes $\boldsymbol{X}_{1i} = \widetilde{\boldsymbol{X}}_{1i} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ is some noise vector independent of $\widetilde{\boldsymbol{X}}_{1i}$ and $\boldsymbol{X}_{2i}$. Substituting $\widetilde{\boldsymbol{X}}_{1i}$ into the above equation,

$$Y_i = \boldsymbol{X}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{X}_{2i}'\boldsymbol{\beta}_2 - \boldsymbol{\varepsilon}_i'\boldsymbol{\beta} + v_i.$$

  Set $u_i = -\boldsymbol{\varepsilon}_i'\boldsymbol{\beta}_1 + v_i$. While $\boldsymbol{X}_{2i}$ is exogenous, $\boldsymbol{X}_{1i}$ is endogenous, because it is correlated with u_i through $\boldsymbol{\varepsilon}_i$.

- **Simultaneity.** Consider the following equation

$$Hours_i = \beta_1 Children_i + \boldsymbol{X}_{2i}'\boldsymbol{\beta}_2 + U_i,$$

where $Hours_i$ is the hours of work per week, and $Children_i$ is the number of children in the family, and $\boldsymbol{X}_{2i}$ is a vector of exogenous variables. While the number of children affects labor supply, it is reasonable to assume that career decisions affect family size, i.e. there is another equation determining the number of children in the family:

$$Children_i = \gamma_1 Hours_i + \boldsymbol{Z}_{1i}'\boldsymbol{\gamma}_2 + V_i,$$

where $\boldsymbol{Z}_{1i}$ is another vector of exogenous variables. Substituting the expression for the hours into the equation for the number of children, we obtain (assuming that $1 - \beta_1\gamma_1 \neq 0$)

$$Children_i = \boldsymbol{X}_{2i}'\left(\frac{\boldsymbol{\beta}_2\gamma_1}{1-\beta_1\gamma_1}\right) + \boldsymbol{Z}_{1i}'\frac{\boldsymbol{\gamma}_2}{1-\beta_1\gamma_1} + \frac{\gamma_1}{1-\beta_1\gamma_1}U_i + \frac{1}{1-\beta_1\gamma_1}V_i.$$

Assuming that $\boldsymbol{X}_{2i}$, $\boldsymbol{Z}_{1i}$ and V_i are uncorrelated with U_i, we have that

$$\begin{aligned}\mathbb{E}\left(U_i Children_i\right) &= \frac{\gamma_1}{1-\beta_1\gamma_1}\mathbb{E}U_i^2 \\ &\neq 0.\end{aligned}$$

Properties of the OLS under endogeneity

Consider first the OLS estimator of $\boldsymbol{\beta}_1$:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{1n} &= \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{Y} \\ &= \boldsymbol{\beta}_1 + \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{e},\end{aligned}$$

where $\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'$. We have

$$\begin{aligned}n^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1 &= n^{-1}\sum_{i=1}^n \boldsymbol{X}_{1i}\boldsymbol{X}_{1i}' - n^{-1}\sum_{i=1}^n \boldsymbol{X}_{1i}\boldsymbol{X}_{2i}'\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_{2i}\boldsymbol{X}_{2i}'\right)^{-1} n^{-1}\sum_{i=1}^n \boldsymbol{X}_{2i}\boldsymbol{X}_{1i}', \\ n^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{e} &= n^{-1}\sum_{i=1}^n \boldsymbol{X}_{1i}e_i - n^{-1}\sum_{i=1}^n \boldsymbol{X}_{1i}\boldsymbol{X}_{2i}'\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_{2i}\boldsymbol{X}_{2i}'\right)^{-1} n^{-1}\sum_{i=1}^n \boldsymbol{X}_{2i}e_i.\end{aligned}$$

Assume that:

- $\{(Y_i, \boldsymbol{X}_i) : i \geq 1\}$ are iid.
- $\mathbb{E}X_{i,j}^2 < \infty$ for all $j = 1, \ldots k$.
- $\mathbb{E}\boldsymbol{X}_i\boldsymbol{X}_i'$ positive definite.
- $\mathbb{E}e_i^2 < \infty$.

By the WLLN we have

$$\begin{aligned}
n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}' &\to_p \mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}', \\
n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{1i}\boldsymbol{X}_{2i}' &\to_p \mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{2i}', \\
n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{2i}\boldsymbol{X}_{2i}' &\to_p \mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{2i}', \\
n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{2i}e_i &\to_p \boldsymbol{0}, \\
n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{1i}e_i &\to_p \mathbb{E}\boldsymbol{X}_{1i}e_i.
\end{aligned}$$

Thus,

$$\begin{aligned}
n^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1 &\to_p \mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}' - \mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{2i}'\left(\mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{2i}'\right)^{-1}\mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{1i}', \\
n^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{e} &\to_p \mathbb{E}\boldsymbol{X}_{1i}e_i - \mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{2i}'\left(\mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{2i}'\right)^{-1}\mathbb{E}\boldsymbol{X}_{2i}e_i \\
&= \mathbb{E}\boldsymbol{X}_{1i}e_i \\
&\neq \boldsymbol{0},
\end{aligned}$$

and we conclude that $\widehat{\boldsymbol{\beta}}_{1n}$ is inconsistent:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{1n} &\to_p \boldsymbol{\beta}_1 + \left(\mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}' - \mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{2i}'\left(\mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{2i}'\right)^{-1}\mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{1i}'\right)^{-1}\mathbb{E}\boldsymbol{X}_{1i}e_i \\
&\neq \boldsymbol{\beta}_1.
\end{aligned}$$

Inconsistency of the OLS estimator of $\boldsymbol{\beta}_2$ can be shown similarly. We have

$$\widehat{\boldsymbol{\beta}}_{2n} = \boldsymbol{\beta}_2 + \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{e},$$

where $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$. We have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{2n} &\to_p \boldsymbol{\beta}_2 - \left(\mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{2i}' - \mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{1i}'\left(\mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}'\right)^{-1}\mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{2i}'\right)^{-1}\mathbb{E}\boldsymbol{X}_{2i}\boldsymbol{X}_{1i}'\left(\mathbb{E}\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}'\right)^{-1}\mathbb{E}\boldsymbol{X}_{1i}e_i \\
&\neq \boldsymbol{\beta}_2.
\end{aligned}$$

Instrumental Variables estimation

Let $\boldsymbol{Z}_{1i}$ be a k_1-vector of *exogenous* variables:

$$\mathbb{E}\boldsymbol{Z}_{1i}e_i = \boldsymbol{0}.$$

It is important that $\boldsymbol{Z}_{1i}$ is *excluded* from the model (1), i.e. $\boldsymbol{Z}_{1i}$ does not contain any of the elements of $\boldsymbol{X}_{2i}$. Define

$$\begin{aligned}\boldsymbol{X}_i &= \begin{pmatrix} \boldsymbol{X}_{1i} \\ \boldsymbol{X}_{2i} \end{pmatrix}, \\ \boldsymbol{Z}_i &= \begin{pmatrix} \boldsymbol{Z}_{1i} \\ \boldsymbol{X}_{2i} \end{pmatrix}.\end{aligned}$$

Here, $\boldsymbol{X}_i$ is the k-vector of regressors, and $\boldsymbol{Z}_i$ is the k-vector of *Instrumental Variables* (IVs). Note that the exogenous regressors appear again in the vector of IVs, and for each endogenous regressor we bring an exogenous variable (IV) that must be excluded from the model $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + e_i$. When all regressors are endogenous, $k_1 = k$ and we do not have any overlapping elements between X_i and Z_i.

We assume that the IVs are informative about the regressors. This is expressed as the following *rank condition*:

$$\operatorname{rank}\left(\mathbb{E}\boldsymbol{Z}_i\boldsymbol{X}_i'\right) = k. \tag{2}$$

The rank condition in (2) will fail if, for example, $\mathbb{E}\boldsymbol{Z}_{1i}\boldsymbol{X}_i' = \boldsymbol{0}$ ($\boldsymbol{Z}_{1i}$ is exogenous but random noise). The rank condition will also fail if some of the elements of $\boldsymbol{Z}_{1i}$ are linear combinations of the elements of the included exogenous regressors $\boldsymbol{X}_{2i}$.

Example. Consider the Hours/Children example. Angrist and Evans (1998) suggested to use the sex composition of the first two children as an instrument to the number of children in the family (the sample was restricted to women with at least two children). This is motivated by the observation that if the first two children are of the same sex (boy-boy or girl-girl), the family is more likely to have a third child than in the case (boy-girl or girl-boy). Consequently, the dummy variable for the first two children are of the same sex has to be positively correlated with the total number of children. On the other hand, the instrument is uncorrelated with the errors, because sex composition is determined randomly.

We have that

$$\mathbb{E}\boldsymbol{Z}_i e_i = \boldsymbol{0}.$$

The method of moments principle suggests an estimator that solves the following system of k equations:

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\boldsymbol{Z}_i\left(Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_n^{IV}\right) &= \boldsymbol{0}, \text{ or} \\ \widehat{\boldsymbol{\beta}}_n^{IV} &= \left(\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1}\sum_{i=1}^{n}\boldsymbol{Z}_i Y_i \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}. \end{aligned}$$

The estimator $\widehat{\boldsymbol{\beta}}_n^{IV}$ is called the IV estimator of $\boldsymbol{\beta}$.

Next, we show consistency and asymptotic normality of the IV estimator. We assume:

- $\{(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) : i \geq 1\}$ are iid.
- $\mathbb{E}\boldsymbol{Z}_i e_i = \boldsymbol{0}$.

- $\mathbb{E}X_{i,j}^2 < \infty$ for all $j = 1, \ldots k$.
- $\mathbb{E}Z_{i,j}^2 < \infty$ for all $j = 1, \ldots k_1$.
- $\mathbb{E}\boldsymbol{Z}_i\boldsymbol{X}_i'$ is of rank k.
- $\mathbb{E}e_i^2\boldsymbol{Z}_i\boldsymbol{Z}_i'$ is positive definite.

Write

$$\widehat{\boldsymbol{\beta}}_n^{IV} = \boldsymbol{\beta} + \left(n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1} n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i e_i. \tag{3}$$

Note that, under the above assumptions, by the Cauchy-Shwartz inequality

$$\begin{aligned}\mathbb{E}\left|Z_{i,r}X_{i,s}\right| &\leq \sqrt{\mathbb{E}Z_{i,r}^2\mathbb{E}X_{i,s}^2} \\ &< \infty \text{ for all } r, s = 1, \ldots k.\end{aligned}$$

Therefore, by the Continuous Mapping Theorem,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^{IV} &\to_p \boldsymbol{\beta} + \left(\mathbb{E}\boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1}\mathbb{E}\boldsymbol{Z}_i e_i \\ &= \boldsymbol{\beta}.\end{aligned}$$

In order to show the asymptotic normality, we assume in addition that

- $\mathbb{E}Z_{i,j}^4 < \infty$ for all $j = 1, \ldots k$.
- $\mathbb{E}e_i^4 < \infty$.

Write (3) as

$$n^{1/2}\left(\widehat{\boldsymbol{\beta}}_n^{IV} - \boldsymbol{\beta}\right) = \left(n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1} n^{-1/2}\sum_{i=1}^n \boldsymbol{Z}_i e_i.$$

For all $r, s = 1, \ldots, k$,

$$\begin{aligned}\mathbb{E}\left|e_i^2 Z_{i,r}Z_{i,s}\right| &\leq \left(\mathbb{E}e_i^4\right)^{1/2}\left(\mathbb{E}Z_{i,r}^4\mathbb{E}Z_{i,s}^4\right)^{1/4} \\ &< \infty.\end{aligned}$$

Therefore, by the CLT and Slutsky's Theorem,

$$\begin{aligned}n^{1/2}\left(\widehat{\boldsymbol{\beta}}_n^{IV} - \boldsymbol{\beta}\right) &\to_d \left(\mathbb{E}\boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1} N\left(\boldsymbol{0}, \left(\mathbb{E}e_i^2\boldsymbol{Z}_i\boldsymbol{Z}_i'\right)\right) \\ &= N\left(\boldsymbol{0}, \left(\mathbb{E}\boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1}\left(\mathbb{E}e_i^2\boldsymbol{Z}_i\boldsymbol{Z}_i'\right)\left(\mathbb{E}\boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1}\right).\end{aligned}$$

The asymptotic covariance matrix takes the sandwich form and can be estimated consistently by

$$\left(n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1} n^{-1}\sum_{i=1}^n \widehat{e}_i^2\boldsymbol{Z}_i\boldsymbol{Z}_i'\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{Z}_i'\right)^{-1},$$

where $\widehat{e}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_n^{IV}$.

Two-stage Least Squares (2SLS)

Define

$$\begin{aligned}\widetilde{\boldsymbol{X}} &= \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X} \\ &= \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X},\end{aligned}$$

the orthogonal projection of the matrix of regressors $\boldsymbol{X}$ onto the space spanned by the instruments $\boldsymbol{Z}$. Since $\boldsymbol{P}_{\boldsymbol{Z}}$ is idempotent, we can write

$$\widehat{\boldsymbol{\beta}}_n^{2SLS} = \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{Y}.$$

Thus, $\widehat{\boldsymbol{\beta}}_n$ can be obtained using the two-step procedure. First, regress $\boldsymbol{X}$ against instruments, and obtain the fitted values $\widetilde{\boldsymbol{X}}$. The first step removes from $\boldsymbol{X}_i$ the correlation with the error e_i. In the second step, one should run the regression of $\boldsymbol{Y}$ against the fitted values $\widetilde{\boldsymbol{X}}$. Easy to check:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^{2SLS} &= \left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{Z}_i'\left(\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i'\right)^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{Z}_i'\left(\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i'\right)^{-1}\sum_{i=1}^n \boldsymbol{Z}_i Y_i \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}.\end{aligned}$$

We have that

$$n^{1/2}\left(\widehat{\boldsymbol{\beta}}_n^{2SLS} - \boldsymbol{\beta}\right) \to_d N\left(0, \sigma^2\left(\mathbb{E}\boldsymbol{X}_i\boldsymbol{Z}_i'\left(\mathbb{E}Z_iZ_i'\right)^{-1}\mathbb{E}\boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1}\right),$$

under homoskedasticity $\mathbb{E}\left(e_i^2|\boldsymbol{Z}_i\right) = \sigma^2$.