Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 13: LASSO for Instrumental Variable Models

Instructor: Ma, Jun

Renmin University of China

December 16, 2021

Instrumental variable

- Consider

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

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

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

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

$$\log(Wage_i) = \beta_0 + \beta_1 Education_i + \beta_2 Ability_i + V_i,$$

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

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

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

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,

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

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

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

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

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

- However,

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

- 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 $\log\left(Wage_i\right) = \beta_0 + \beta_1 Education_i + \beta_2 Ability_i + V_i$.
- We need to find an IV variable Z such that $\mathrm{Cov}\left(Ability_i, Z_i\right) = 0$ and $\mathrm{Cov}\left(Education_i, Z_i\right) \neq 0$.
- 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.

Instrumental variable model

- Consider the following model:

$$Y_i = \gamma_0 + \gamma_1 X_{i1} + \ldots + \gamma_k X_{ik} + \beta_1 D_{i1} + \ldots + \beta_m D_{im} + U_i,$$

  where
  - Y_i is the dependent variable.
  - γ_0 is the coefficient on the constant regressor: $\mathrm{E}[U_i] = 0$.
  - $X_{i1}, \ldots, X_{ik}$ are the k exogenous regressors:

$$\mathrm{Cov}[X_{i1}, U_i] = \ldots = \mathrm{Cov}[X_{ik}, U_i] = 0.$$

  - $D_{i1}, \ldots, D_{im}$ are the m endogenous regressors:

$$\mathrm{Cov}[D_{i1}, U_i] \neq 0, \ldots, \mathrm{Cov}[D_{im}, U_i] \neq 0.$$

- Suppose that the econometrician observes l additional exogenous variables (IVs) $Z_{i1}, \ldots, Z_{il}$
- We assume that the IVs $Z_{i1}, \ldots, Z_{il}$ are excluded from the structural equation:

$$Y_i = \gamma_0 + \gamma_1 X_{i1} + \ldots + \gamma_k X_{ik} + \beta_1 D_{i1} + \ldots + \beta_m D_{im} + U_i,$$

  so we still have $k + 1 + m$ structural coefficients to estimate.
- The necessary condition for identification is that the number of IVs is at least as large as the number of unknowns or $l \geq m$.

2SLS

- Consider the first-stage projection models:

$$\begin{aligned} D_{i1} &= \pi_{0,1} + \pi_{1,1} Z_{i1} + \ldots + \pi_{l,1} Z_{il} \\ &\quad + \pi_{l+1,1} X_{i1} + \ldots + \pi_{l+k,1} X_{ik} + V_{i1}, \\ \vdots &\quad \vdots \quad \vdots \\ D_{im} &= \pi_{0,m} + \pi_{1,m} Z_{i1} + \ldots + \pi_{l,m} Z_{il} \\ &\quad + \pi_{l+1,m} X_{i1} + \ldots + \pi_{l+k,m} X_{ik} + V_{im}, \end{aligned}$$

  where $\left(\pi_{0,1}, \pi_{1,1}, ..., \pi_{l+k,m}\right)$ are projection coefficients.
- All right-hand side variables are exogenous.
- The first stage coefficients π's can be estimated consistently by OLS by regressing Y's against Z's and X's.

- After estimating π's, obtain the fitted values for D's:

$$\begin{aligned}
\widehat{D}_{i1} &= \widehat{\pi}_{0,1} + \widehat{\pi}_{1,1} Z_{i1} + \ldots + \widehat{\pi}_{l,1} Z_{il} \\
&\quad + \widehat{\pi}_{l+1,1} X_{i1} + \ldots + \widehat{\pi}_{l+k,1} X_{ik}, \\
\vdots \quad & \vdots \quad \vdots \\
\widehat{D}_{im} &= \widehat{\pi}_{0,m} + \widehat{\pi}_{1,m} Z_{i1} + \ldots + \widehat{\pi}_{l,m} Z_{il} \\
&\quad + \widehat{\pi}_{l+1,m} X_{i1} + \ldots + \widehat{\pi}_{l+k,m} X_{ik}.
\end{aligned}$$

- In the second stage, regress (OLS) the dependent variable Y against a constant, X's, and $\widehat{D}$'s obtained in the first stage:

$$Y_i = \widehat{\gamma}_0^{\text{2sls}} + \widehat{\gamma}_1^{\text{2sls}} X_{i1} + \ldots + \widehat{\gamma}_k^{\text{2sls}} X_{ik} + \widehat{\beta}_1^{\text{2sls}} \widehat{D}_{i1} + \ldots + \widehat{\beta}_m^{\text{2sls}} \widehat{D}_{im} + \widehat{U}_i.$$

- One can show that the resulting 2SLS estimators $\widehat{\gamma}_0^{\text{2sls}}, \widehat{\gamma}_1^{\text{2sls}}, \ldots, \widehat{\gamma}_k^{\text{2sls}}, \widehat{\beta}_1^{\text{2sls}}, \ldots, \widehat{\beta}_m^{\text{2sls}}$ are consistent and asymptotically normal.

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 $Z_i \in \mathbb{R}^l$ ($Z_i = (Z_{i1}, Z_{i2}, ..., Z_{il})^\top$) which satisfies $\mathrm{Cov}[U_i, Z_i] = 0$.
- The first-stage of 2SLS uses the projection model of D_i on Z_i:

$$\begin{aligned} D_i &= Z_i^\top \pi + V_i \\ \mathrm{E}[Z_i V_i] &= 0 \\ \pi &= \underset{a}{\operatorname{argmin}} \mathrm{E}\left[\left(D_i - Z_i^\top a\right)^2\right]. \end{aligned}$$

- Then,

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

  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.
- Under an alternative asymptotic analysis, when the number of IVs l is assumed to be growing $l = l_n \uparrow \infty$ as $n \uparrow \infty$ such that $l_n/n \to c > 0$, the 2SLS estimator is inconsistent.
- LASSO is used for data-driven IV selection.

Optimal instrument

- Suppose that $\mathrm{E}[U_i \mid Z_i] = 0$, then for any function f, $\mathrm{Cov}[f(Z_i), U_i] = 0$ and $\zeta_i = f(Z_i)$ can be used as an IV:

$$\mathrm{E}[\zeta_i Y_i] = \alpha \mathrm{E}[\zeta_i D_i] \Longrightarrow \widehat{\alpha}^{\mathsf{iv}} = \frac{\sum_{i=1}^{n} \zeta_i Y_i}{\sum_{i=1}^{n} \zeta_i D_i}.$$

- Denote $\widehat{\mathbf{D}} = \boldsymbol{P}_{\mathbf{Z}} \mathbf{D}$.

$$\widehat{\alpha}^{\mathsf{2sls}} = \frac{\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}} \mathbf{Y}}{\mathbf{D}^\top \boldsymbol{P}_{\mathbf{Z}} \mathbf{D}} = \frac{\widehat{\mathbf{D}}^\top \mathbf{Y}}{\widehat{\mathbf{D}}^\top \mathbf{D}} = \frac{\sum_{i=1}^{n} \widehat{D}_i Y_i}{\sum_{i=1}^{n} \widehat{D}_i D_i},$$

  where $\widehat{D}_i = Z_i^\top \widehat{\pi}$ and $\widehat{\pi}$ are the first-stage OLS coefficients.
- $\widehat{\alpha}^{\mathsf{2sls}}$ can be viewed as an IV estimator using estimated projection $Z_i^\top \widehat{\pi}$ in lieu of the unknown true projection $Z_i^\top \pi_i$ as the instrument.
- $\widehat{\alpha}^{\mathsf{2sls}}$ summarizes the information in all intruments Z_i and uses a single IV $Z_i^\top \pi$.

- Assume that the model is homoskedastic: $\mathrm{E}\left[U_i^2 \mid D_i\right] = \sigma^2$. We can show that the optimal IV estimator is the one $\widehat{\alpha}^*$ that uses $\zeta_i^* = \mathrm{E}\left[D_i \mid Z_i\right]$:

$$\sqrt{n}\left(\widehat{\alpha}^* - \alpha\right) \to_d \mathrm{N}\left(0, \frac{\sigma^2}{\mathrm{E}\left[\left(\zeta_i^*\right)^2\right]}\right)$$

and

$$\sqrt{n}\left(\widehat{\alpha}^{\mathsf{iv}} - \alpha\right) \to_d \mathrm{N}\left(0, \frac{\sigma^2 \mathrm{E}\left[\zeta_i^2\right]}{\left(\mathrm{E}\left[\zeta_i \zeta_i^*\right]\right)^2}\right).$$

Approximation to the optimal instrument

- The 2SLS uses a linear projection $Z_i^\top \pi$ to approximate $\mathrm{E}[D_i \mid Z_i]$:

$$\begin{aligned}\pi &= \underset{a}{\operatorname{argmin}} \mathrm{E}\left[\left(D_i - Z_i^\top a\right)^2\right] \\ &= \underset{a}{\operatorname{argmin}} \mathrm{E}\left[\left(\mathrm{E}[D_i \mid Z_i] - Z_i^\top a\right)^2\right].\end{aligned}$$

- We generate a dictionary $W_i = \left(W_{i1}, ..., W_{ip}\right)^\top \in \mathbb{R}^p$:

$$W_i = \left(Z_{i1}, Z_{i2}, ..., Z_{il}, Z_{i1}^2, Z_{i1}Z_{i2}, ..., Z_{i1}Z_{il}, Z_{i2}^2, ...\right),$$

  whose dimension p can be larger than n.

- We can also use the linear projection $W_i^\top \delta$ to approximate $\mathrm{E}[D_i \mid Z_i]$, where

$$\begin{aligned}\delta &= \underset{b}{\operatorname{argmin}} \mathrm{E}\left[\left(D_i - W_i^\top b\right)^2\right] \\ &= \underset{a}{\operatorname{argmin}} \mathrm{E}\left[\left(\mathrm{E}[D_i \mid Z_i] - W_i^\top b\right)^2\right].\end{aligned}$$

- It is easy to show that

$$\mathrm{E}\left[\left(\mathrm{E}\left[D_i \mid Z_i\right] - W_i^\top \delta\right)^2\right] < \mathrm{E}\left[\left(\mathrm{E}\left[D_i \mid Z_i\right] - Z_i^\top \pi\right)^2\right].$$

- We assume that if p is very large, then the approximation error is very much close to zero and $W_i^\top \delta$ is the optimal instrument.
- If $p < n$, we can regress D_i on W_i to get the OLS coefficient $\widehat{\delta}$ and uses the estimated optimal instrument $W_i^\top \widehat{\delta}$.
- This procedure is equivalent to 2SLS using all variables in W_i as instruments. We showed that when p is large, the 2SLS estimator may be substantially biased.
- When $p > n$, the 2SLS estimator is not computable. We are forced to select a subset from W_i.

- We assume that the conditional expectation model

$$\begin{aligned} D_i &= W_i^\top \delta + V_i = \sum_{j=1}^{p} \delta_j W_{ij} + V_i \\ \mathrm{E}\left[V_i \mid Z_i\right] &= 0 \end{aligned}$$

  is sparse: $l^* = |\mathcal{A}|$ is a small number, where $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A} = \left\{j : \delta_j \neq 0\right\}$ ($\delta = \left(\delta_1, \delta_2, ..., \delta_p\right)^\top$), although p is very large.
- The IVs in $\mathcal{A}$ are called the effective IVs. Dropping ineffective IVs would not result in loss of efficiency.
- Clearly, there is no difference between the IV estimator using $W_{i,\mathcal{A}}^\top \delta_{\mathcal{A}}$ ($W_{i,\mathcal{A}} = \left\{W_{ij} : j \in \mathcal{A}\right\}$ and $\delta_{\mathcal{A}} = \left\{\delta_j : j \in \mathcal{A}\right\}$) and the IV estimator using $W_i^\top \delta$.
- However, we do not know $\mathcal{A}$ (identities of the effective IVs). We use LASSO selection to find them.

Algorithm

1. LASSO regression of D_i against W_i:

$$\left(\widehat{\delta}_{1,\lambda}, ..., \widehat{\delta}_{p,\lambda}\right) = \underset{b_1,...,b_p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(D_i - \sum_{j=1}^{p} b_j W_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left| b_j \right| \right\}.$$

   Let $\widehat{\mathcal{A}} = \left\{ j : \widehat{\delta}_{j,\lambda} \neq 0 \right\}$ be the selected controls.
   - The dropped IVs are either ineffective or have small coefficients ($\delta_j \propto n^{-1/2}$). In the latter case, it can be shown that such variables do not contribute to the asymptotic variance, so we can drop them without loss of efficiency.
2. Post-LASSO of D_i against $W_{i,\widehat{\mathcal{A}}}$ and get the OLS coefficients $\left\{ \widehat{\delta}^{\mathsf{pl}}_j : j \in \widehat{\mathcal{A}} \right\}$. Generate the fitted value as the estimated optimal IV: $\widehat{\zeta}^*_i = \sum_{j \in \widehat{\mathcal{A}}} \widehat{\delta}^{\mathsf{pl}}_j W_{ij}$.
3. Estimate α using $\widehat{\zeta}^*_i$ as the IV:

$$\widehat{\alpha}^* \left(\widehat{\mathcal{A}} \right) = \frac{\sum_{i=1}^{n} \widehat{\zeta}^*_i Y_i}{\sum_{i=1}^{n} \widehat{\zeta}^*_i D_i}.$$

Model with controls

- The structural model with controls $X_i = (X_{i1}, X_{i2}, ..., X_{ik})^\top$:

$$\begin{aligned} Y_i &= \alpha D_i + X_i^\top \beta + U_i \\ \mathrm{E}\left[U_i \mid X_i, Z_i\right] &= 0. \end{aligned}$$

- The intercept is typically one of the elements in X_i.
- Controls X_i have to be included in the first stage. Consider 2SLS and the following projection models:

$$\begin{aligned} D_i &= Z_i^\top \pi + X_i^\top \gamma + V_i \\ \mathrm{E}\left[V_i \begin{pmatrix} Z_i \\ X_i \end{pmatrix}\right] &= 0 \end{aligned}$$

and

$$\begin{aligned} D_i &= Z_i^\top \widetilde{\pi} + \widetilde{V}_i \\ \mathrm{E}\left[\widetilde{V}_i Z_i\right] &= 0. \end{aligned}$$

- It is easy to show that $\widetilde{\pi} = \Theta\gamma$, where

$$\Theta = \left(\mathrm{E}\left[Z_i Z_i^\top\right]\right)^{-1} \mathrm{E}\left[Z_i X_i^\top\right]$$
$$\widetilde{V}_i = D_i - Z_i^\top \widetilde{\pi} = V_i + \left(X_i^\top - Z_i^\top \Theta\right)\gamma.$$

  $\widetilde{V}_i$ is not correlated with Z_i but it is correlated with X_i.
- If we drop X_i from the first stage,

$$Y_i = \alpha D_i + X_i^\top \beta + U_i$$
$$D_i = Z_i^\top \widetilde{\pi} + \widetilde{V}_i$$
$$\Longrightarrow Y_i = \alpha\left(Z_i^\top \widetilde{\pi} + \widetilde{V}_i\right) + X_i^\top \beta + U_i = \alpha\left(Z_i^\top \widetilde{\pi}\right) + X_i^\top \beta + \alpha \widetilde{V}_i + U_i.$$

  The residual $\alpha \widetilde{V}_i + U_i$ is correlated with X_i.
- Regression of Y_i against $Z_i^\top \widetilde{\pi}$ and X_i does not give consistent estimator for α.

- The 2SLS can be written as an IV estimator:

$$\begin{pmatrix} \widehat{\alpha}^{\text{2sls}} \\ \widehat{\beta}^{\text{2sls}} \end{pmatrix} = \left(\sum_{i=1}^{n} \begin{pmatrix} \widehat{D}_i \\ X_i \end{pmatrix} \begin{pmatrix} D_i \\ X_i \end{pmatrix}^{\top} \right)^{-1} \left(\sum_{i=1}^{n} \begin{pmatrix} \widehat{D}_i \\ X_i \end{pmatrix} Y_i \right),$$

$$= \left(\sum_{i=1}^{n} \begin{pmatrix} \widehat{D}_i \\ X_i \end{pmatrix} \begin{pmatrix} D_i \\ X_i \end{pmatrix}^{\top} \right)^{-1} \left(\sum_{i=1}^{n} \begin{pmatrix} \widehat{D}_i \\ X_i \end{pmatrix} Y_i \right),$$

  where $\widehat{D}_i = Z_i^{\top}\widehat{\pi} + X_i^{\top}\widehat{\gamma}$ denotes the first-stage fitted value.
- The optimal IV: $\zeta_i^* = \mathrm{E}\left[D_i \mid X_i, Z_i\right]$ and the optimal IV estimator:

$$\begin{pmatrix} \widehat{\alpha}^* \\ \widehat{\beta}^* \end{pmatrix} = \left(\sum_{i=1}^{n} \begin{pmatrix} \zeta_i^* \\ X_i \end{pmatrix} \begin{pmatrix} D_i \\ X_i \end{pmatrix}^{\top} \right)^{-1} \left(\sum_{i=1}^{n} \begin{pmatrix} \zeta_i^* \\ X_i \end{pmatrix} Y_i \right).$$

- We need to approximate ζ_i^*.

Many IVs and few controls

- The conditional expectation model for D_i:

$$
\begin{aligned}
D_i &= W_i^\top \delta + X_i^\top \gamma + V_i \\
\mathrm{E}\left[V_i \mid X_i, Z_i\right] &= 0,
\end{aligned}
$$

  where the dictionary W_i contains many polynomials of Z_i and interactions between Z_i and X_i. In this case, we need selection over W_i but need no selection over the controls X_i.
- In the first-stage regression, we force inclusion of X_i by assigning no penalty weights to their coefficients.
- In the second stage, we run IV regression by using the post-LASSO fitted value as the IV.

Algorithm

1. LASSO regression of D_i against W_i and X_i:

$$\left(\widehat{\delta}_{1,\lambda}, ..., \widehat{\delta}_{p,\lambda}, \widehat{\gamma}_{1,\lambda}, ..., \widehat{\gamma}_{k,\lambda}\right)$$

$$= \underset{b_1,...,b_p,d_1,...,d_k}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(D_i - \sum_{j=1}^{p} b_j W_{ij} - \sum_{j=1}^{k} d_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left| b_j \right| \right\}.$$

   Let $\widehat{\mathcal{A}} = \left\{ j : \widehat{\delta}_{j,\lambda} \neq 0 \right\}$ be the selected controls.

2. Run post LASSO of D_i against the instruments in $W_{i,\widehat{\mathcal{A}}} = \left\{ W_{ij} : j \in \widehat{\mathcal{A}} \right\}$ and X_i to get OLS coefficients $\left\{ \widehat{\delta}_j^{\mathsf{pl}} : j \in \widehat{\mathcal{A}} \right\} \cup \left\{ \widehat{\gamma}_1^{\mathsf{pl}}, ..., \widehat{\gamma}_k^{\mathsf{pl}} \right\}$. Construct

$$\widehat{\zeta}_i^* = \sum_{j=1}^{p} \widehat{\delta}_j^{\mathsf{pl}} W_{ij} + \sum_{j=1}^{k} \widehat{\gamma}_j^{\mathsf{pl}} X_{ij}.$$

3. Estimate (α, β) by using $\widehat{\zeta}_i^*$ as the IV:

$$\begin{pmatrix} \widehat{\alpha}^* \\ \widehat{\beta}^* \end{pmatrix} = \left(\sum_{i=1}^{n} \begin{pmatrix} \widehat{\zeta}_i^* \\ X_i \end{pmatrix} \begin{pmatrix} D_i \\ X_i \end{pmatrix}^\top \right)^{-1} \left(\sum_{i=1}^{n} \begin{pmatrix} \widehat{\zeta}_i^* \\ X_i \end{pmatrix} Y_i \right).$$

Partialling out

- Let $\boldsymbol{M}_{\mathbf{X}} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ be the projection matrix on the space that is orthogonal to the column space of $\mathbf{X}$: $\boldsymbol{M}_{\mathbf{X}}\mathbf{X} = \mathbf{0}$.
- Write

$$\begin{aligned} Y_i &= \alpha D_i + X_i^\top \beta + U_i \\ D_i &= W_i^\top \delta + X_i^\top \gamma + V_i \end{aligned}$$

in the matrix form

$$\begin{aligned} \mathbf{Y} &= \alpha \mathbf{D} + \mathbf{X}\beta + \mathbf{U} \\ \mathbf{D} &= \mathbf{W}\delta + \mathbf{X}\gamma + \mathbf{V}. \end{aligned}$$

- Multiply both sides by $\boldsymbol{M}_{\mathbf{X}}$ to get

$$\begin{aligned} \widetilde{\mathbf{Y}} &= \alpha \widetilde{\mathbf{D}} + \widetilde{\mathbf{U}} \\ \widetilde{\mathbf{D}} &= \widetilde{\mathbf{W}}\delta + \widetilde{\mathbf{V}}, \end{aligned}$$

where $\widetilde{\mathbf{Y}} = \boldsymbol{M}_{\mathbf{X}}\mathbf{Y}$, $\widetilde{\mathbf{D}} = \boldsymbol{M}_{\mathbf{X}}\mathbf{D}$, $\widetilde{\mathbf{W}} = \boldsymbol{M}_{\mathbf{X}}\mathbf{W}$, $\widetilde{\mathbf{U}} = \boldsymbol{M}_{\mathbf{X}}\mathbf{U}$ and $\widetilde{\mathbf{V}} = \boldsymbol{M}_{\mathbf{X}}\mathbf{V}$.
- By transforming $(\mathbf{Y}, \mathbf{D}, \mathbf{W})$ into the residuals against $\mathbf{X}$, we have another numerically equivalent way to compute the IV estimator.

The partialling out algorithm

1. Run LASSO regression of $\widetilde{\mathbf{D}} = \left(\widetilde{D}_1, ..., \widetilde{D}_n\right)^\top$ against $\widetilde{\mathbf{W}}$ ($\widetilde{W}_{ij}$ denotes its ij-th element of the $n \times p$ matrix $\widetilde{\mathbf{W}}$) to get $\left(\widehat{\delta}_{1,\lambda}, ..., \widehat{\delta}_{p,\lambda}\right)^\top$. Let $\widehat{\mathcal{A}} = \left\{j : \widehat{\delta}_{j,\lambda} \neq 0\right\}$ be the selected controls.
2. Run post LASSO regression of $\widetilde{\mathbf{D}}$ against the IVs in $\widehat{\mathcal{A}}$ to get OLS coefficients $\left\{\widehat{\delta}_j^{\mathsf{pl}} : j \in \widehat{\mathcal{A}}\right\}$. Construct the estimated optimal IV $\widehat{\zeta}_i^* = \sum_{j \in \widehat{\mathcal{A}}} \widehat{\delta}_j^{\mathsf{pl}} \widetilde{W}_{ij}$.
3. Estimate α by using $\widehat{\zeta}_i^*$ as the IV:

$$\widehat{\alpha}^* = \frac{\sum_{i=1}^n \widehat{\zeta}_i^* \widetilde{Y}_i}{\sum_{i=1}^n \widehat{\zeta}_i^* \widetilde{D}_i},$$

where $\widetilde{\mathbf{Y}} = \left(\widetilde{Y}_1, ..., \widetilde{Y}_n\right)^\top$.

Few IVs and many controls

- In the model

$$\begin{aligned} Y_i &= \alpha D_i + X_i^\top \beta + U_i \\ \mathrm{E}\left[U_i \mid X_i, Z_i\right] &= 0 \\ D_i &= Z_i^\top \pi + X_i^\top \gamma + V_i \\ \mathrm{E}\left[V_i \mid X_i, Z_i\right] &= 0, \end{aligned}$$

  the dimension of X_i is large but the number of IVs is small and we do not use its polynomials and interactions to approximate the optimal IV $\mathrm{E}\left[D_i \mid X_i, Z_i\right]$.
- This is the case, for example, when there is only one binary instrument (a dummy variable, all polynomials are equal) and we do not use its interactions with X_i.
- We assume the outcome equation is sparse: the set of relevant controls $\mathcal{A} = \left\{j : \beta_j \neq 0\right\}$ is small.
- In this case, we need to perform LASSO selection over X_i only.
- We apply the partialling out approach by using LASSO and post LASSO over over X_i.

The partialling out algorithm

1. Perform LASSO and post LASSO of D_i against X_i to generate the residual $\widetilde{D}_i^{\mathsf{pl}}$.
2. Perform LASSO and post LASSO of Y_i against X_i to generate the residual $\widetilde{Y}_i^{\mathsf{pl}}$.
3. Perform LASSO and post LASSO of Z_{ij} against X_i to generate the residual $\widetilde{Z}_{ij}^{\mathsf{pl}}$, for $j = 1, 2, ..., l$.
4. Run OLS of $\widetilde{D}_i^{\mathsf{pl}}$ against $\widetilde{Z}_{i1}^{\mathsf{pl}}, ..., \widetilde{Z}_{il}^{\mathsf{pl}}$ to get the OLS coefficients $(\widehat{\pi}_1, ..., \widehat{\pi}_l)$ and the estimated optimal IV $\widehat{\zeta}_i^* = \sum_{j=1}^{l} \widehat{\pi}_j \widetilde{Z}_{ij}$.
5. Estimate α by

$$\widehat{\alpha}^* = \frac{\sum_{i=1}^{n} \widehat{\zeta}_i^* \widetilde{Y}_i^{\mathsf{pl}}}{\sum_{i=1}^{n} \widehat{\zeta}_i^* \widetilde{D}_i^{\mathsf{pl}}}.$$

Many IVs and many controls

- In the model

$$
\begin{aligned}
Y_i &= \alpha D_i + X_i^\top \beta + U_i \\
\mathrm{E}\left[U_i \mid X_i, Z_i\right] &= 0 \\
D_i &= W_i^\top \delta + X_i^\top \gamma + V_i \\
\mathrm{E}\left[V_i \mid X_i, Z_i\right] &= 0,
\end{aligned}
$$

  where the dictionary W_i contains high-dimensional transformations (polynomials) of the primitive instruments Z_i and interactions of Z_i and X_i.
- Both W_i and X_i are high-dimensional. We need to perform LASSO selections over both.
- The previous partialling out procedure is not practically implementable, since the LASSO and post-LASSO partialling out of many controls from many IVs is computationally hard.
- We do LASSO selection on W_i first to estimate the equation $D_i = W_i^\top \delta + X_i^\top \gamma + V_i$ and find the effective IVs. We then partial out the effects from X_i.

The partialling out algorithm

1. Perform LASSO and post LASSO of D_i against X_i to generate the residual $\widetilde{D}_i^{\mathsf{pl}}$.
2. Perform LASSO and post LASSO of Y_i against X_i to generate the residual $\widetilde{Y}_i^{\mathsf{pl}}$.
3. Perform LASSO and post LASSO of D_i against W_i and X_i to generate the fitted value $\widehat{\zeta}_i^*$ (estimated optimal IV). This step selects a subset from W_i.
4. Perform LASSO and post LASSO of $\widehat{\zeta}_i^*$ against X_i to partial out the effect from X_i and get the residual $\widetilde{\zeta}_i^{\mathsf{pl}}$.
5. Estimate α by

$$\widehat{\alpha}^* = \frac{\sum_{i=1}^n \widetilde{\zeta}_i^{\mathsf{pl}} \widetilde{Y}_i^{\mathsf{pl}}}{\sum_{i=1}^n \widetilde{\zeta}_i^{\mathsf{pl}} \widetilde{D}_i^{\mathsf{pl}}}.$$