Lecture 2.A. Understanding the Inference Strategy via Partialling Out: Theory

VC

May 13, 2016

Regression in Population

Let Y be a scalar random variable and X be a p-vector of covariates called regressors. We also assume that $\mathrm{E}Y^2$ and $\mathrm{E}XX'$ are finite.
We then define least squares or projection parameter β in the *population* as the solution of the following prediction problem:

$$\beta := \arg\min_{b \in \mathbb{R}^p} \mathrm{E}(Y - X'b)^2 = (\mathrm{E}XX')^{-1}\mathrm{E}XY.$$

The explicit solution follows from the first order condition:

$$\mathrm{E}(Y - X'\beta)X = 0,$$

provided that $\mathrm{E}XX'$ is of full rank, which amounts to absence of the multicollinearity. Defining $\varepsilon = Y - X'\beta$, we obtain the decomposition identity

$$Y \equiv X'\beta + \varepsilon, \quad \mathrm{E}\varepsilon X = 0.$$

Double Partialling Out. Frisch-Waugh-Lovell Theorem

This is an important tool that provides conceptual understanding of least squares as well as a very practical tool for estimation and visualization of results. We partition vector of regressors X into two groups:

$$X = (D', W')',$$

where D represents "target" regressors of interest, and W represents other regressors, sometimes called the controls. For example, in program evaluation context, D is a treatment or policy variable and W are controls. Write

$$Y = D'\beta_1 + W'\beta_2 + \varepsilon. \tag{1}$$

Double Partialing Out in Population

In *population*, define the partialling out operator that takes a random variable V such that $\mathrm{E}V^2 < \infty$ and creates $\tilde{V}$ according to the rule:

$$\tilde{V} = V - W'\gamma_{VW}, \quad \gamma_{VW} = \arg\min_{b \in \mathbb{R}^p} \mathrm{E}(V - W'b)^2.$$

When V is a vector, we interpret the application of the operator as componentwise.

It is not difficult to see that the partialling-out operator is linear on the space of random variables with finite second moments, i.e. if for V and U such that $\mathrm{E}U^2 + \mathrm{E}V^2 < \infty$,

$$Y = V + U \implies \tilde{Y} = \tilde{V} + \tilde{U}.$$

Thus we apply this operator to both sides of the identity (1) to get:

$$\tilde{Y} = \tilde{D}'\beta_1 + \tilde{W}'\beta_2 + \tilde{\varepsilon},$$

which implies that

$$\tilde{Y} = \tilde{D}'\beta_1 + \varepsilon, \quad \mathrm{E}\varepsilon\tilde{D} = 0. \tag{2}$$

The last line follows from $\tilde{W} = 0$, which holds by definition, and $\tilde{\varepsilon} = \varepsilon$, which holds because of the orthogonality $\mathrm{E}\varepsilon X = 0$; moreover, since $\tilde{D}$ is a linear combination of components of X, we have that $\mathrm{E}\varepsilon\tilde{D} = 0$.

Equation (2) states that $\mathrm{E}\varepsilon\tilde{D} = 0$ is the first-order condition for the population regression of $\tilde{Y}$ on $\tilde{D}$. That is, the projection coefficient β_1 can be recovered from the regression of $\tilde{Y}$ on $\tilde{D}$:

$$\beta_1 = (\mathrm{E}\tilde{D}\tilde{D}')^{-1}\mathrm{E}\tilde{D}\tilde{Y}.$$

This is a remarkable fact, known as Frisch-Waugh-Lovell (FWL) theorem. It asserts that β_1 is a regression coefficient of Y on D after partialling-out the linear effect of W from Y and D.

Theorem (Frisch-Waugh-Lovell)

The population projection coefficient β_1 can be recovered from the population regression of $\tilde{Y}$ on $\tilde{D}$:

$$\beta_1 = (\mathrm{E}\tilde{D}\tilde{D}')^{-1}\mathrm{E}\tilde{D}\tilde{Y}.$$

How to do Estimation?

- In the sample, we will need to mimic partialling out in population.
- Can do by OLS, which would work well when $p \ll n$, but would obviously fail when $p \gg n$.
- When p is large, $p \propto n$, or $p \gg n$, high-quality partialling out can be done via regularization to prevent overfitting. Selection by Lasso is one way to regularize.
- Other Machine Learning methods provide other ways of regularizing.

How to do Estimation?

- We can form:
$$\hat{\beta}_1 = (\mathbb{E}_n \check{D}_i \check{D}_i)^{-1} \mathbb{E}_n \check{D}_i \check{Y}_i.$$
where $\check{M}_i$ is the residual left after predicting M_i with controls W_i using a regularized estimator (e.g., lasso or post-lasso) when p is large; also can use OLS when $p \ll n$.
- This involves **"Double Prediction"** or **"Double Machine Learning"**.

Theorem (Inference Based on High-Quality Partialling Out)

If partialling out is done by a high-quality regularization procedure, then asymptotically the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\check{\beta}_1$, and

$$\sqrt{n}(\check{\beta}_1 - \beta_1) \overset{a}{\sim} N(0, V_{11})$$

where V_{11} is the "standard" robust variance expression:

$$V_{11} = (\mathrm{E}\tilde{D}\tilde{D}')^{-1} \mathsf{Var}(\sqrt{n}\mathbb{E}_n \tilde{D}_i \epsilon_i)(\mathrm{E}\tilde{D}\tilde{D}')^{-1},$$

as if we worked with true residuals.

- See Belloni, Chernozhukov, Wang (Annals of Stat, 2014) for sufficient conditions for the use of lasso and post-lasso as providing high-quality partialing out under approximate sparsity conditions.
- Similar strategy applies for instrumental variable regression models – see Chernozhukov, Hansen, Spindler (ARE, 2015).

The partialling-out extends to IV models. Consider the following model for simplicity:

$$\begin{aligned} Y &= \alpha_1 D + \alpha_2' W + U, \quad U \perp (W', Z')', \\ D &= \beta_1 Z + \beta_2' W + V, \quad V \perp (W', Z')', \end{aligned} \tag{IVM}$$

where W includes a constant.
Application of the partialling out operator to both sides of each of the equations in (IVM) gives us a much simpler system of equations:

$$\begin{aligned} \tilde{Y} &= \alpha_1 \tilde{D} + U, \quad U \perp \tilde{Z}, \\ \tilde{D} &= \beta_1 \tilde{Z} + V, \quad V \perp \tilde{Z}. \end{aligned} \tag{3}$$

Theorem (IV with Partialing Out)

$$\alpha_1 = (\mathrm{E}\tilde{D}\tilde{Z})^{-1}(\mathrm{E}\tilde{Z}\tilde{Y})$$

How to do Estimation?

- We can form:

$$\hat{\alpha}_1 = (\mathbb{E}_n \check{D}_i \check{Z}_i)^{-1} \mathbb{E}_n \check{Z}_i \check{Y}_i.$$

  where $\check{V}_i$ denotes the residual left after predicting V_i with controls W_i using a regularized estimator (e.g., lasso or post-lasso) when p is large (can use OLS when $p \ll n$).
- This involves **"Double Prediction"** or **"Double Machine Learning"**.

Theorem (Inference Based on High-Quality Partialling Out)

If partialling out is done by a high-quality regularization procedure, then asymptotically the estimation error in $\check{D}_i$, $\check{Z}_i$, and $\check{Y}_i$ has no first order effect on $\check{\alpha}_1$, and

$$\sqrt{n}(\check{\alpha}_1 - \alpha_1) \overset{a}{\sim} N(0, V_{11})$$

where V_{11} is the "standard" robust variance expression:

$$V_{11} = (\mathrm{E}\tilde{D}\tilde{Z})^{-1}\,\mathrm{Var}(\sqrt{n}\mathbb{E}_n \tilde{Z}_i \epsilon_i)(\mathrm{E}\tilde{D}\tilde{Z})^{-1},$$

as if we worked with true residuals.

- Reference: see Chernozhukov, Hansen, Spindler (ARE, 2015), which also considers the case of high-dimensional Z.