Regression Approach Revisted

A. Torgovitsky*

May 10, 2022

- The model implies

$$\mathbb{E}[Y|D,X,Z] = B'\theta,$$

where B are random variables generated by the basis functions and θ are unknown coefficients.

- We assume that $\theta \in \Theta$, some subset of $\mathbb{R}^{d_\theta}$ that encodes shape restrictions.
- We want to characterize the identified set

$$\Theta^\star \equiv \Theta \cap \left\{\theta \in \mathbb{R}^{d_\theta} : \mathbb{E}[Y|D,X,Z] = B'\theta \text{ a.s.}\right\}.$$

- Thus far we have been using the following equality:

$$\Theta^\star = \Theta \cap \underset{\theta \in \mathbb{R}^{d_\theta}}{\arg\min}\, \mathbb{E}[(Y - B'\theta)^2].$$

This argument follows from the best linear approximation motivation of linear regression, which implies that

$$\left\{\theta \in \mathbb{R}^{d_\theta} : \mathbb{E}[Y|D,X,Z] = B'\theta \text{ a.s.}\right\} = \underset{\theta \in \mathbb{R}^{d_\theta}}{\arg\min}\, \mathbb{E}[(Y - B'\theta)^2], \tag{1}$$

as long as the model is correctly specified (which we assume).

- Equation (1) is useful because it turns an abstract almost-sure equality into a scalar criterion using squared loss. However, the squared criterion turns out to make the second step of our two-step estimation a bit delicate computationally.

*Kenneth C. Griffin Department of Economics, University of Chicago.

- As an alternative, I claim that

$$\left\{\theta \in \mathbb{R}^{d_\theta} : \mathbb{E}[Y|D,X,Z] = B'\theta \text{ a.s.}\right\} = \left\{\theta \in \mathbb{R}^{d_\theta} : \mathbb{E}[B(Y - B'\theta)] = 0_{d_\theta}\right\}. \quad (2)$$

  That is, the set of θ that reproduce the conditional mean can be represented as the set of solutions to a finite number of moment inequalities.

- If θ were point identified, then (2) would follow immediately from the first-order conditions to (1). In our case, θ is not point identified. In the regression formulation, this shows up as $\mathbb{E}[BB']$ not being positive definite, so that the squared loss criterion function is not strictly convex. However, since $\mathbb{E}[BB']$ is always positive *semi*-definite, the set equality in (2) still holds. Intuitively, the first-order condition still needs to be satisfied for a minimizer of the weakly convex criterion, since otherwise one could move downhill a bit further and arrive at a contradiction.

- Proof just for sanity. One direction is immediate by first-order conditions

$$\theta_0 \in \arg\min_{\theta \in \mathbb{R}^{d_\theta}} \mathbb{E}[(Y - B'\theta)^2] \quad \Rightarrow \quad \mathbb{E}[B(Y - B'\theta_0)] = 0_{d_\theta}.$$

  For the other direction, first notice that

$$\arg\min_{\theta \in \mathbb{R}^{d_\theta}} \mathbb{E}[(Y - B'\theta)^2] = \arg\min_{\theta \in \mathbb{R}^{d_\theta}} - \mathbb{E}[YB']\theta + \frac{1}{2}\theta' \,\mathbb{E}[BB']\theta \equiv \arg\min_{\theta \in \mathbb{R}^{d_\theta}} f(\theta),$$

  where $c \equiv -\,\mathbb{E}[YB]$ and $A \equiv \mathbb{E}[BB']$. Now suppose that

$$\mathbb{E}[B(Y - B'\theta_0)] = \mathbb{E}[YB] - \mathbb{E}[BB']\theta_0 \equiv -c - A\theta_0 = 0.$$

  Then for any other $\theta \in \mathbb{R}^{d_\theta}$

$$\begin{aligned} f(\theta) &= f(\theta_0 + (\theta - \theta_0)) \\ &= c'\theta_0 + c'(\theta - \theta_0) + \frac{1}{2}\theta_0' A\theta_0 + \theta_0' A(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' A(\theta - \theta_0) \\ &= f(\theta_0) + \underbrace{(c + A\theta_0)'}_{= 0 \text{ by hypothesis}}(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' A(\theta - \theta_0) \\ &= f(\theta_0) + \frac{1}{2}(\theta - \theta_0)' A(\theta - \theta_0) \geq f(\theta_0), \end{aligned}$$

where the last inequality used the positive *semi*-definiteness of A. Since $f(\theta) \geq f(\theta_0)$ for any $\theta \in \mathbb{R}^{d_\theta}$,

$$\theta_0 \in \arg\min_{\theta \in \mathbb{R}^{d_\theta}} f(\theta),$$

which establishes the other direction.

- So we have now established that working with (2) does not sacrifice any identifying information. The computational advantage of working with (2) is that we can use any loss function that we want. So, for example the one-norm:

$$Q(\theta) \equiv \sum_{j=1}^{d_\theta} |\mathbb{E}[B_j(Y - B'\theta)]| \tag{3}$$

  The sample analog of this would be

$$\hat{Q}(\theta) \equiv \sum_{j=1}^{d_\theta} \left| \frac{1}{n} \sum_{i=1}^{n} B_{ji}(Y_i - B_i'\theta) \right|. \tag{4}$$

  This can be linearized in the usual way by adding appropriate slack variables.

- The upshot is that if we base the two-step procedure on (4) then we have
  - All of the information contained in the regression approach.
  - A linear program in both steps of the problem.

- Moreover, the procedure is interpretable as a specific case of the "moment–based" approach that uses IV–like estimands. The IV–like estimands here are

$$\{\mathbb{E}[YB_j]\}_{j=1}^{d_\theta}. \tag{5}$$

  Notice that if θ generated the model, then

$$\mathbb{E}[YB_j] = \mathbb{E}[B_j\, \mathbb{E}[Y|D, X, Z]] = \mathbb{E}[B_j B']\theta.$$

  So choosing θ to match the IV–like estimands given in (5) is the same as trying to set

$$\mathbb{E}[YB_j] - \mathbb{E}[B_j B']\theta = 0 \quad \text{for } j = 1, \dots, d_\theta,$$

which is what the criterion (3) is doing too.