Model Selection and Regularization

Post-Selection Inference: The Partialling Out Approach (2 Extra Points)

Let x be treatment, y be outcome, and $z \in \mathbb{R}^k$ be a set of control variables. Let our model of their relationship be given by

$$y = \beta_0 + \beta_1 x + \gamma' z + e \tag{1}$$

, where γ is a $k \times 1$ vector of parameters.

We are interested in the causal effect of x on y given z. Therefore, we are interested mainly in the estimation and inference of β_1, while γ can be treated as a nuisance parameter.

In the high-dimensional controls setting, where there is a large number of *potential* control variables so that k is large relative to sample size, we can estimate and obtain valid inference on β_1 using the following method:

Algorithm. *Partialling Out*

Stage 1 *Estimate the following two models by lasso or post-lasso OLS:*

$$y = \alpha' z + \epsilon$$
$$x = \lambda' z + \xi$$

, from which we obtain $\widehat{\epsilon}$ and $\widehat{\xi}$.

Stage 2 *Run the following residual-on-residual regression by OLS:*

$$\widehat{\epsilon} = \beta_1 \widehat{\xi} + e$$

, from which we obtain $\widehat{\beta}_1$ and its asymptotic variance.

The approach has been called **partialling out** and utilizes the principal of **Neyman-orthogonality**. More generally, we can use this approach in situations in which we have a linear model with high-dimensional regressors but are interested in obtaining valid inference only on a low-dimensional subset of the model parameters (e.g., the parameters associated with the treatment variables).

Challenge

1. Write an introduction on the partialling out approach to post-selection inference.
2. Can we estimate (1) directly and obtain inference on β_1 using post-lasso OLS? Will this be a valid approach? Can you use simulation to compare the performance of these two approaches?

Reference

- The main reference for the partialling out approach of post-selection inference is Belloni et al. (2014). For an introduction and summary of the methodology, see Chernozhukov et al. (2017).
- For simulation, you can use the `R` package hdm, which stands for "high-dimensional metrics". Chernozhukov et al. (2016) provides an overview as well as examples of the various methods implemented in the package.