Empirical Bayes Mixtape Session, Coding Lab 1: School Value-Added

This coding lab will walk you through an example of empirical Bayes estimation of school value-added based on simulated data. Consider a population of students, each attending one of J schools. The variable $D_i \in \{1, ..., J\}$ indicates the school attended by student i. We are interested in the following constant-effects causal value-added model:

$$Y_i = \sum_{j=1}^{J} \theta_j D_{ij} + \beta X_i + \epsilon_i,$$

where Y_i is a test score outcome for student i, θ_j is the causal effect of school j and $D_{ij} = 1\{D_i = j\}$ indicates attendance at j, X_i is an observed control variable (e.g. a lagged test score), and ϵ_i represents unobserved determinants of students' potential outcomes and is assumed to satisfy $E[X_i\epsilon_i] = E[D_{ij}\epsilon_i] = 0 \ \forall j$. The school-level parameters θ_j are assumed to be drawn from a normal mixing distribution:

$$\theta_j \sim N(\mu_\theta, \sigma^2_\theta).$$

Our goal is to learn about the parameters of the mixing distribution and form accurate estimates of the individual school value-added parameters θ_j.

1. Import the data set "vam_example_data.csv" from the course website into a statistical software package of your choice (I recommend Stata or R). This data set includes observations on Y_i, D_i, and X_i simulated from the model above for 2,500 students, each attending one of $J = 50$ schools. Since the data were simulated from a known data generating process, it also includes the true value-added of each student's school, given by $\theta_{d(i)} = \sum_j \theta_j D_{ij}$, which would not be known in a real-world application. Summarize the variables in this data set.

2. Create the school dummy variables D_{ij}. Fit two value-added models by ordinary least squares (OLS):

   (a) An *uncontrolled* OLS regression of Y_i on the set of D_{ij}'s with no constant.

   (b) A *controlled* OLS regression of Y_i on the D_{ij}'s, controlling for X_i (again with no constant).

   For each of these two models, collect the list of estimated value-added coefficients $\hat{\theta}_j$ along with their robust standard errors s_j for each school. Then collapse the data down to a school-level data set with 50 observations on true value-added θ_j, value-added estimates $\hat{\theta}_j$, and standard errors s_j.

3. For each of the two value-added models, use the OLS estimates and standard errors to form estimates $\hat{\mu}_\theta$ and $\hat{\sigma}^2_\theta$ of the mean and variance of value-added. How do the estimated variances differ for the controlled and uncontrolled models? What do you conclude from this comparison?

4. Focus on estimates of the controlled model for the remainder of the question. Using these estimates, form linear shrinkage posteriors $\hat{\theta}^*_j = \left(\frac{\hat{\sigma}^2_\theta}{\hat{\sigma}^2_\theta + s^2_j}\right)\hat{\theta}_j + \left(\frac{s^2_j}{\hat{\sigma}^2_\theta + s^2_j}\right)\hat{\mu}_\theta$. Summarize your estimates by making a plot that includes histograms of the OLS and shrunk estimates overlayed with the estimated mixing distribution (i.e. $N(\hat{\mu}_\theta, \hat{\sigma}^2_\theta)$). Compare standard deviations of the OLS estimates $\hat{\theta}_j$, the shrunk posteriors $\hat{\theta}^*_j$, the estimated mixing distribution, and the true-value-added parameters θ_j. Which standard deviation is biggest, and which is smallest? Provide some intuition.

5. For each school, compute the squared difference between true value-added θ_j and the OLS estimate $\hat{\theta}_j$, as well as the squared difference between true value-added and the linear shrinkage estimate $\hat{\theta}^*_j$. Which estimator has lower mean squared error (MSE) averaged across the 50 schools in the sample? Does shrinkage cause squared error to move in the same direction for all schools? Give some intuition for your results.