Empirical Bayes Coding Lab 1: Quantifying Variation in Discrimination

This coding lab applies empirical Bayes tools to data from the employment discrimination experiment of Kline, Rose and Walters (2022, forthcoming). This experiment submitted fictitious applications to real job vacancies within Fortune 500 firms. Each application was randomly assigned a distinctively-white or distinctively-Black name, stratified by job so that each vacancy received 4 white and 4 Black applications (though a few vacancies closed before all 8 applications could be sent). The key outcome is whether the application received a callback from the employer within 30 days.

1. Download the data set "krw_data.csv". This is an application-level data set including job and firm identifiers, firm names, an indicator for a distinctively-white name, and a callback indicator. Summarize the variables in this data set.

2. Regress an indicator for a callback on an indicator for a white name in the pooled sample of all applications. What is the average effect of a white name on callbacks? Compare robust and job-clustered standard errors for the race coefficient. How do these standard errors differ, and why?

3. Compute the white/Black difference in callback rates (the racial contact gap) separately for each job vacancy in the data set. Let $\hat{\Delta}_{jf}$ denote the contact gap for job j within firm f. Take the average of $\hat{\Delta}_{jf}$ for each firm, resulting in 108 firm-specific estimates $\hat{\theta}_f$. This is an unbiased estimate of the average effect of race at firm f, labeled θ_f.

4. Compute a standard error for each $\hat{\theta}_f$ as $s_f = \sqrt{\frac{1}{n_f(n_f-1)}\sum_{j=1}^{n_f}(\hat{\Delta}_{jf} - \hat{\theta}_f)^2}$, where n_f is the number of jobs for firm f. Collapse the data down to a firm-level data set with 108 observations on $\hat{\theta}_f$ and s_f.

5. Suppose the θ_f's are *iid* draws from a normal mixing distribution: $\theta_f \sim iid\ N(\mu_\theta, \sigma_\theta^2)$. Compute estimates $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$ of the mean and variance of the mixing distribution.

6. Compute linear shrinkage posteriors $\hat{\theta}_f^* = \left(\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2 + s_f^2}\right)\hat{\theta}_f + \left(\frac{s_f^2}{\hat{\sigma}_\theta^2 + s_f^2}\right)\hat{\mu}_\theta$. Summarize your results by making a plot that includes histograms of the OLS and shrunk estimates overlayed with the estimated mixing distribution. Compare standard deviations of the unbiased estimates $\hat{\theta}_f$, the shrunk posteriors $\hat{\theta}_f^*$, and the estimated mixing distribution. Which standard deviation is biggest, and which is smallest?