Empirical Bayes Mixtape Session, Coding Lab 2: Labor Market Discrimination

This coding lab will walk you through an empirical Bayes analysis of employer heterogeneity in labor market discrimination using data from the field experiment of Kline, Rose and Walters (2022). This experiment submitted 83,643 fictitious applications to 11,114 real job vacancies within 108 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. Import the data set "krw_data.csv" from the course website into a statistical software package of your choice (I recommend Stata or R). This is an application-level data set including job and firm identifiers, 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 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 we view the θ_f's as random draws from a mixing distribution G. Using your unbiased estimates and standard errors, compute a bias-corrected estimate of the variance of G, labeled $\hat{\sigma}^2_\theta$. How does the standard deviation $\hat{\sigma}_\theta$ compare to the standard deviation of unbiased estimates $\hat{\theta}_f$? In economic terms, is $\hat{\sigma}_\theta$ big or small?

6. Form linear shrinkage estimates of the effect of race at each firm. Plot histograms of unbiased and linear shrinkage estimates.

7. The last part of this lab asks you to compute a non-parametric deconvolution estimate of G, the distribution of discrimination across firms. This is an advanced exercise, so do not worry if you find it difficult.

   (a) Convert the estimates for each firm to a z-score, $z_f = \hat{\theta}_f/s_f$. Assume $z_f \sim N(\mu_f, 1)$, where $\mu_f = \theta_f/s_f$.

   (b) Compute a log-spline deconvolution estimate of the distribution of μ_f across firms. [Hint: This can be done in R with the **deconvolveR** package.]

   (c) Compute a kernel density estimate of the distribution of log standard errors, $\log s_f$. [Hint: This can be done in R with the **density** command in the **stats** package.]

   (d) If μ_f and $\log s_f$ are independent, the density function for $\theta_f = \mu_f \exp(\log s_f)$ is given by:

$$g_\theta(\theta) = \int g_\mu\left(\theta \exp(-t)\right) f(t) \exp(-t) dt,$$

   where g_μ is the density function for μ_f and f is the density function for $\log s_f$. Use this expression together with your results from parts (b) and (c) to compute an estimate of the distribution of θ_f across firms. Overlay this distribution on the histogram of unbiased estimates from part (6).

   Extra credit: Use your log-spline estimates to compute non-parametric posterior mean estimates for each θ_f. Plot these against the linear shrinkage estimates from part (6). What do you make of any differences between these estimates?