Empirical Bayes Coding Lab 2: Non-parametric EB and Multiple Testing

This coding lab applies non-parametric empirical Bayes and multiple testing methods to data from the employment discrimination experiment of Kline, Rose and Walters (2022, forthcoming).

1. Repeat steps 1-4 from coding lab 1 to obtain a data set of 108 firm-specific contact gap estimates $\hat{\theta}_f$ along with standard errors s_f.
2. Compute a non-parametric deconvolution estimate of the distribution of contact gaps across firms. This can be done with the following steps:
   (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: $g_\theta(\theta) = \int g_\mu(\theta \exp(-t))\, h(t) \exp(-t) dt$, where g_μ is the density function for μ_f and h 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 $\hat{\theta}_f$.
   (e) Use your log-spline estimates to compute non-parametric posterior mean estimates for each θ_f. Plot these against the linear shrinkage estimates from lab 1. What do you make of any differences between these estimates?
3. Conduct a multiple testing analysis to determine which firms can be reliably classified as discriminating against distinctively-Black names while controlling the False Discovery Rate (FDR).
   (a) Use the z_f statistic from part 2(a) to compute the p-value p_f from a one-tailed test of $H_0 : \theta_f = 0$ vs. $H_A : \theta_f > 0$ for each firm.
   (b) Plot a histogram of the firm-specific p-values. What do you notice about this distribution?
   (c) Let $\pi_0 = \Pr[\theta_f = 0] = \int 1[\theta = 0] dG(\theta)$ denote the share of firms in the population that are not discriminating. Use your p-values to compute an upper bound $\hat{\pi}_0$ on π_0. [Hint: compute the average height of the p-value density above some threshold λ. Try $\lambda = 0.5$.]
   (d) Use the bound from part 3(c) to compute q-values as $q_f = [p_f \hat{\pi}_0] / \hat{F}_p(p_f)$, where $\hat{F}_p$ is the estimated CDF of p-values.
   (e) Make a table listing each firm's name, contact gap estimate, standard error, linear shrinkage and non-parametric posterior means, p-value, and q-value. Limiting FDR to 5%, how many firms can you classify as discriminating against distinctively-Black names?