Assignment 7b

Satvik Saha

2024-10-22

Answer 1

Consider a model $y_i = a + bx_i + cz_i + \epsilon_i$, where x_i, y_i are the pre-test and post-test scores, z_i is the treatment indicator, and ϵ_i are errors. We have fitted $\hat{c} = 0.35$ with a standard error of 0.30.

Now suppose that the sample size is doubled. We do not expect the estimated coefficients to change much. However, the standard error in $\hat{c}$ will go down by a factor of around $\sqrt{2}$, giving 0.21.

Recall that in the linear model $y = X\beta + \epsilon$, we have $\hat{\beta} \sim N(\beta, \sigma^2(X^\top X)^{-1})$. Doubling the sample size almost amounts to stacking two copies of the design matrix together, say $\tilde{X}^\top = (X^\top X^\top)$. With this, $\tilde{X}^\top \tilde{X} = 2X^\top X$, which explains the factor of $\sqrt{2}$. This is not quite right since the values of x_i are not actually duplicated, but should be redrawn from its generating distribution; with a sample size as large as what has been given, we ignore this discrepancy.

Final Project Plan

David and I plan to analyze the results of the 2024 House of Representatives elections district by district and form a simulation-based ranking of the most gerrymandered states. For each state, we will randomly sample the state's voter population into equally-sized partitions of voters and calculate the margin in each of our new randomly-created districts. After enough simulations, we'll be able to tell how likely the true district-by-district results would've been if such a random assignment was actually used to determine congressional districts.

This is difficult because there are numerous confounding variables, such as geography (urban/rural proximity) and racial demographics, that make this an imperfect way of simulating district divisions. In the allotted time for the project, we will have to choose which variables to attempt to take into account and which ones to abstract out.

Through this investigation, we'd like to gain more experience with bootstrapping and simulation, as well as communicating significant results.