ECON 523: Program Evaluation for International Development
Empirical Exercise 12

In this exercise, we use data form the paper "Enhancing Young Children's Language Acquisition Through Parent-Child Book Sharing: a Randomized Trial in Rural Kenya" by Lia C.H. Fernald, Heather A. Knauer, Pamela Jakiela, and Owen Ozier. The study examines the short-term impact of dialogic reading training combined with mother tongue storybooks. The key outcome variable is `e_zstoryexp`, a measure of the expressive vocabulary children might have picked up from the storybooks. The data set also contains a large set of baseline covariates.

Begin by creating a do file that reads in the data. Familiarize yourself with the dataset and the range of baseline variables included in it. Then extend your do file as you answer the following questions.

1. Regress the outcome `e_zstoryexp` on `treatment`. What is the p-value associated with the test of the hypothesis that treatment has no impact on children's vocabulary?

2. Predict the residuals from the regression above. What is the variance of the residuals?

3. Use `lasso` to identify the baseline covariates that predict `treatment`. Use cross-validation to choose the tuning parameter, λ. Which variables does lasso select as predictors of `treatment`? Store the selected variables in a local named `Tvars`.

4. Use `lasso` to identify the baseline covariates that predict `e_zstoryexp`. Use cross-validation to choose the tuning parameter, λ. Make sure you tell lasso to always include the `treatment` variable. Which variables does lasso select as predictors of the outcome? Store the selected variables in a local named `Yvars`.

5. Regress the outcome `e_zstoryexp` on treatment including all the controls selected by lasso. What is the p-value associated with the test of the hypothesis that treatment has no impact on children's vocabulary?

6. Predict the residuals from the regression that includes the controls selected by lasso. What is the variance of these residuals?