Problem Set 3

Goal: Explore Multiple Linear Regression
Application: Agan and Starr (2017)

~~Part 1. Research Question & Data~~

The goal of this problem set is to explore multiple linear regression and the interpretation of coefficients. We will explore the relationship between a saturated linear model and conditional means, and interpret coefficients in the context of this experiment. In doing so, we will also review diff-in-diff.

1. First let's try to understand the goal of this paper. Read the introduction.

   (a) Explain why the creators of Ban the Box believed the policy would reduce racial differences in employment.

   (b) The authors put forward two hypotheses about why the policy might backfire. What are they?

~~2. Load the data that you used for Lab 3. Use `read.table` to import AganStarrQJEData.dta, and following Footnote 22 of the paper, use the `subset` command to drop those observations who have remover=-1. If you need a refresher on the dataset, look back to the summary table you made in Lab 3.~~

~~Part 2. Review of Law of Iterated Expectations~~

1. ~~Recall the Law of Iterated Expectations (LIE): for any random variables X and W,~~

   ~~$E[X] = E\left[E(X \mid W)\right]$~~.

   ~~For example, if W is binary,~~

   ~~$E[X] = \Pr[W = 1]E[X \mid W = 1] + \Pr[W = 0]E[X \mid W = 0]$~~.

   ~~Apply LIE to the callback rate in this dataset. Specifically, use the callback rate for white applicants, the callback rate for black applicants, and the fraction of applicants who are white, to calculate the overall callback rate. Show that your answer agrees with the callback rate calculated directly.~~

Part 3. Equivalence between Linear Regression with Fully Saturated Model and Conditional Means

As reviewed in class, linear projection can be seen as the best linear predictor of the outcome given the regressors, while the conditional expectation function (CEF) is the best (possibly non-linear) predictor of the outcome given the regressors. One can also view linear projection as the best linear approximation to the CEF. If the CEF is linear, then the CEF and linear projection will coincide.

1. ~~Let's show that when the CEF is linear, it coincides with linear projection. Suppose that the CEF is linear in two binary regressors:~~

   ~~$E[Y \mid X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2$~~

   ~~. This is an example of a fully saturated linear model. What conditional means do β_1, β_2, and β_3 correspond to?[1]~~

2. ~~Using only observations in the pre period (`post==0`) that had no "box" in the pre period (`remover==0`), estimate the following model by OLS regression:~~

$$\text{Callback}_i = \alpha_0 + \alpha_1 \text{White}_i + \epsilon_i \tag{1}$$

   (a) Interpret and discuss the estimated α_1. Why can we use $\hat{\alpha}_1$ as a consistent estimator of the effect of race on callback probability without worrying about omitted variable bias? (2 sentences or less)

   ~~(b) Show that $\hat{\alpha}_0$ equals the sample mean of Callback_i among black applicants (in the pre period for employers with no box).~~

   ~~(c) Show that $\hat{\alpha}_0 + \hat{\alpha}_1$ equals the sample mean of Callback_i among white applicants (in the pre period for employers with no box)~~.

3. Using only observations in the pre-period that had a "box" in the pre-period, estimate the following model by OLS regression.

$$\text{Callback}_i = \lambda_0 + \lambda_1 \text{White}_i + \lambda_2 \text{Crime}_i + \lambda_3 (\text{White}_i \times \text{Crime}_i) + \epsilon_i \tag{2}$$

   (Note that this fully saturated model is analogous to a diff-in-diff model, except that there is no difference over time. Instead we have differences in differences over race and over criminal status. Since "white" and "crime" are randomly assigned, we do not need any "common trends" assumption for identification.)

   (a) ~~Give a justification for including Criminal Record status in the regression here for employers with a box, but not in Question 1 for employers without a box.~~

   (b) Interpret and discuss the estimated coefficients, and compare/contrast your result with your result from Question 1. Why can we use $\hat{\lambda}_1$ and $\hat{\lambda}_1 + \hat{\lambda}_3$ as consistent estimator of the effect of race on callback probability on those without and with a criminal record, without worrying about omitted variable bias?

   (c) ~~Show that $\hat{\lambda}_0$ equals the sample mean of Callback_i among black applicants without a criminal record (in the pre period for employers with no box)~~.

   ~~(d) Show that the remaining sample means of Callback_i conditional on race and criminal record status (black applicants with a criminal record, white applicants without a criminal record, white applicants with a criminal record) equal the appro~~priate linear combination of OLS coefficients.

[1] FYI: it can be shown that as a result, the corresponding OLS coefficient estimates will be linear combinations of conditional sample means, and OLS fitted values will be equal to conditional sample means. Specifically, $\hat{\beta}_0$ equals the sample mean of Y_i among observations with $X_i = 0$, and $\hat{\beta}_1$ equals the sample mean of Y_i among observations with $X_i = 1$ minus the sample mean of Y_i among observations with $X_i = 0$.

(e) Consider the estimated average effect of being white on callback probabilities for employers in the pre-period with a box (without conditioning on criminal record status). You could estimate this by running a similar regression as (1), for `remover==1`. But you can also estimate this using only the coefficients you estimated in model (2), plus the law of iterated expectations (LIE). Specifically, use LIE (see Part 2), the fraction of observations with a criminal record, and your $\hat{\lambda}_1$ and $\hat{\lambda}_1 + \hat{\lambda}_3$ estimates to compute the estimated overall average effect of being white in the pre-period for employers with a box.

(f) Summarize your results thus far. Is there evidence of discrimination? Is the discrimination different for employers that have a box vs not? How does the discrimination interact with having a criminal record?

4. Thus far, we have fit two different linear regression models on two different samples, the linear regression model of equation 1 on the pre-period sample with no box, and the linear regression model of equation 2 on the pre-period sample with a box. It is sometimes convenient to combine both regressions into one equivalent regression. For example, if one wishes to test the null that $\alpha_1 = \lambda_1$, it is convenient to estimate both parameters as part of one regression. Now, using all observations from the pre-period, estimate the following model by OLS regression:

$$\begin{aligned}\text{Callback}_i &= \beta_0 + \beta_1 \text{White}_i + \beta_2 \text{Box}_i \\ &\quad + \beta_3(\text{White}_i \times \text{Box}_i) + \beta_4(\text{Crime}_i \times \text{Box}_i) + \beta_5(\text{White}_i \times \text{Crime}_i \times \text{Box}_i) + \epsilon_i \end{aligned} \tag{3}$$

(a) Interpret each β coefficient.

(b) If we define the models of equations 1, 2, 3 by linear projection, what is the relationship between $(\beta_0, \beta_1, ..., \beta_5)$ and (α_0, α_1), $(\lambda_0, ...\lambda_3)$? (*Hint:* write expressions for the conditional means of Callback for no box employers by race, and the conditional means for box employers by race and criminal status. You should get expressions linking β's, α's, and λ's.)

(c) Conjecture a relationship between $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_5)$ and your $(\hat{\alpha}_0, \hat{\alpha}_1)$ and $(\hat{\lambda}_0, ...\hat{\lambda}_3)$ from Questions 1 and 2. Verify your conjecture.

(d) State in words (interpret) the null hypothesis $\alpha_1 = \lambda_1$ from equations 1, 2 in Questions 1 and 2.

(e) In equation 3, testing the null $\alpha_1 = \lambda_1$ corresponds to what null hypothesis about which β parameter(s)?

(f) Why is $\hat{\beta}_2$ not necessarily a consistent estimator of the effect of having a box on callback probabilities for black applicants? If the policy question of interest if the effect of banning the box on callbacks for black applicants, will $\hat{\beta}_2$ be a convincing answer to that question?

5. Now let's do some inference on the coefficients of model 3. Please use standard errors that are robust to heteroscedasticity and clustering by chain.

(a) Consider the effect of being white on the callback rate in the pre-period for employers who don't have a box in the pre-period ($\text{Box}_i = 0$)

i. What is the estimate?

ii. What is the standard error?

iii. Test the null hypothesis of that effect being zero, versus a two-sided alternative, at the 5% significance level.

iv. Construct an asymptotic, two-sided 95% confidence interval.

v. Discuss your results.

(b) Repeat part (a), but now for the effect of being white on the callback rate in the pre-period for applicants without a criminal record applying to employers with a box. To answer this question, run a new regression which directly estimates that effect.

(c) Repeat part (a), but now for the effect of being white on the callback rate in the pre-period for applicants *with* a criminal record applying to employers with a box. To answer this question, run a new regression which directly estimates that effect.

(d) Test the null hypothesis that, for employers with a box, the effect of being white vs black is the same for applicants with vs without a criminal record. Use a two-sided alternative, and an (asymptotic) 5% significance level.

(e) Test the null hypothesis that the effect of being white for employers with no box is the same as the effect of being white for applicants with no criminal record applying to employers with a box. Use a two-sided alternative, and an (asymptotic) 5% significance level.

(f) Test the joint null of no effect of being white on callbacks for employers without a box (no effect for applicants with or without a criminal record) using an (asymptotic) 5% significance level.

6. Again using data from the pre-period, consider the following regression model:

$$\begin{aligned}\text{Callback}_i = {} & \beta_0 + \beta_1 \text{White}_i + \beta_2 \text{Box}_i + \beta_3(\text{White}_i \times \text{Box}_i) + \beta_4(\text{Crime}_i \times \text{Box}_i) \\ & + \beta_5(\text{White}_i \times \text{Crime}_i \times \text{Box}_i) + \beta_6 \text{GED}_i + \beta_7 \text{EmploymentGap}_i + \epsilon_i \end{aligned} \quad (4)$$

(a) We saw that OLS estimation of the regression model of equation 3 is equivalent to OLS estimation of the regression model of equation 1 on employers without a box and separately OLS estimation of the regression model of equation 2 on employers with a box. Is the same statement true for OLS estimation of equation 4? If not, what changes would you have to make to equation 4 for the statement to be true? would there be disadvantages of making those changes?

(b) For the regression models of equations 1 - 3, we saw that the estimated parameters are linear combinations of conditional sample means, and the resulting fitted values are conditional sample means. Are the same statements true for the regression model of equation 4? If not, what changes would you have to make to equation 4 for the statements to be true? would there be disadvantages of making those changes?

(c) Estimate the regression model of equation 4.

(d) Comparing estimates from equations 3 vs 4, does including GED and EmploymentGap as regressors substantially change the estimated coefficients on the other regressors? Use what you know for omitted variable bias to explain the results.