Sanjay Satish **September 27, 2021**

Duke University – Economics Honors Thesis

I recommend starting with an overview of the goals before getting too far into the details.

Research Plan & Empirical Strategy

At a high level, my main empirical strategy can be described by the following procedure:

1. Given data in the controls set for Project STAR or surrounding private schools, utilize this data to find (or construct) some measure of what the "typical private school student" looks like.

2. Create two partitions of the dataset by selecting a proportion of schools at random. I would aim for the following distribution of observations in the partition: 30% of schools would exist in a "training" dataset and the other 70% in a "testing" dataset.

3. Within the training dataset, utilize a k-nearest neighbors algorithm to classify the students that left the public school system who "look-like" the typical private school student as individuals who may have, in fact, actually gone to a private school. K-nearest neighbors is a classification technique through which for a given binary outcome variable (in this case, an indicator for whether or not a student left for a private school) individuals are assigned an outcome given their degree of "closeness" to its k-nearest neighbors. To explain more rigorously:

   Imagine we have a collection of N students who we know left the public school system $x_1, \ldots, x_n$ denoted by the set X. For each of these students there exists a $d \times 1$ vector $\vec{v_i} \in \mathbb{R}^D$ where each row of the vector represents the value of each covariate for student x_i (i.e. row one is race, row two is gender, etc.). Imagine that we also have a data on N private school students $y_1, \ldots, y_n$ denoted by the set Y which are represented by the set of vectors $W = \{\vec{w_1}, \ldots, \vec{w_n}\}$ where each vector w_i is also $d \times 1$ with each row representing the value of the same data points that are held in each vector v_i. Through taking a measure of "closeness" (e.g. the Euclidean distance between each $\vec{v_i}$ and $\vec{w_i}$) we then look at the k closest distance values and look at whichever group the majority of those k-values belong to (i.e. private school student or not a private school student). We then classify each x_i as belonging to the group which the majority of these values fall under.

4. Given these newly classified students, estimate the lagged-achievement effects of these students leaving by some form of the following regression equation:

$$T_{i,g,c,t,s} = \beta_0 + \beta_1(L_{i,g,c,t-1,s}) + \beta_2(A_{i,g,c,t-1,s}) + \beta_3(PA_{i,g,c,t-1,s}) + \beta_4(X_{i,g,c,t-1,s}) + \alpha_s + \gamma_{t-1} + \epsilon \quad (1)$$

This is great, but

This section i

Also consider Mal

Where:

- $T_{i,g,c,t,s}$ is the test score for student i in grade g in classroom c in school s at time t.
- $L_{i,g,c,t-1,s}$ is the proportion of students in student i's class that left for private schools at time $t-1$.
- $A_{i,g,c,t-1,s}$ is student i's ability, measured by their test score in time $t-1$
- $PA_{i,g,c,t-1,s}$ is the peer ability for student i. That is, the average test scores of the students that left for private school in $t-1$ Should you also control for the lagged test s
- $X_{i,g,c,t-1,s}$ is a vector of controls for each student including observable student, teacher, and school data
- α_s is a school fixed effect
- γ_{t-1} is a year fixed effect
- ϵ is the regression error term

If I am unable to accurately classify students who left Project STAR for private schools, perhaps I may refine my research question to only explore the effect of attrition broadly, or perhaps the effect of attrition across class-types. Evidence that attrition may perhaps have greater effects in smaller classrooms may also result in significant policy consequences.

I recommend starting with this and then trying to identify the "private-looking" student