Student Factors

In this analysis we do not construct student subgroups and then determine the school readiness rate for each of those groups. Had we taken that approach and attempted to create as fine-grained groupings as possible, some of the subgroups would likely have such a small number of students that the statistical power would be limited.

Instead we create a model that best predicts the outcome (school readiness) as a combination of factors. The model we have used is a multivariate linear model of factors because such a model is widely understood and used. This was important to make the analysis as "transparent" as possible and yet still allow us to extract the information you need which is the impact of each Pre-K program. For a definition of best we use a least squares minimization, again because it is so well known.

This multivariate linear model approach "protects" us from the explicit small cell problem because we consider the entire student population. It is true that if there is a specific combination of factors that is rare (equivalent to the small cell problem) it is likely that even if it has significant impact on those few students, it will not have much weight in our model because it will be "diluted" because we are considering the entire set of students.

That said, there is a parallel issue with sample size and number of factors. We are analyzing 21 factors in the model but what we would like to know is how much does random chance impact the ultimate model equation we obtain?

For multivariate linear regression there is a well-known method to detect the impact of such randomness, k-fold cross validation. Essentially, what we did is to randomly allocate every student to 1 of 5 groups of students and perform the complete analysis on each of those 5 groups one at a time. Each of those evaluations result in a measure indicating how well we were able to account for the student outcomes based on our 21 factors (R squared).

First we compared the R squared values for each of the 5 groups to see if there were significant differences. If there are then we know the models derived from each of those are "unstable" meaning they depend on the population of students. The R squared values differed very little (lowest of 26.8% and highest of 28.4%).

We then subjected each of the groups to the equation derived from each of the other groups and compare the R squared for a group across each of the equations to look for differences. The maximum difference for any group across the 5 equations was 0.3%. So, each equation worked essentially as well on any other group as on the group it was derived from.

Lastly, we then generated an equation using the entire population and compared the resulting R squared value with that obtained for each of the groups alone. In this effort we are looking for what is called shrinkage where the R squared value for the whole population "shrinks" as compared to the results for each group. The greatest shrinkage was 0.5%. This is a very small shrinkage.

So we are able to conclude that the model generated is stable and is not vulnerable to random variations. It was only at this point that the model equation was used to evaluate the facilities.

The R squared values for the group analyses are shown below. R squared for the full population model was 27.9%.

		Equation Based On				
		Group 0	Group 1	Group 2	Group 3	Group 4
Data Evaluated	Group 0	26.8%	26.6%	26.6%	26.6%	26.7%
	Group 1	28.1%	28.3%	28.1%	28.2%	28.2%
	Group 2	28.0%	28.0%	28.2%	27.9%	28.0%
	Group 3	28.2%	28.2%	28.1%	28.4%	28.2%
	Group 4	28.0%	28.0%	27.9%	27.9%	28.1%