This document illustrates the procedure of estimating potential supply. The potential supply is calculated as a proportion of licensed capacity where the proportion cutoffs, i.e., low, medium, high, are estimated directly from the provider data. In order to do so, I follow several steps.

The first step is to extrapolate average daily attendance for providers who are open but do not report their attendance. [1] Let Y denote average daily attendance, W denote continuous attributes, X denote discrete attributes. Let I denote the group of providers who are open and reporting average daily attendance, and J denote the group of providers who are open but NOT reporting average daily attendance. The expected average daily attendance for providers who are open but not reporting, $\hat{E}[Y_j \mid W_j, X_j]$ where $j \in J$, is estimated nonparametrically by a kernel regression of Y_i on W_i and X_i, evaluated at points W_j and X_j, namely:

$$\hat{E}[Y_j|W_j,X_j] = \frac{\sum\limits_{i \in S_Y} Y_i K\left(\frac{W_j - W_i}{h_W}\right) \mathbb{1}(X_j = X_i)}{\sum\limits_{i \in S_Y} K\left(\frac{W_j - W_i}{h_W}\right) \mathbb{1}(X_j = X_i)},$$

where S_y is the overlapping support such that the kernel density is positive. I use a Gaussian kernel function to preserve the sample size. The optimal bandwidth is computed using the Silverman formula. I do this for every provider in group J. The underlying assumption is that once we control additional covariates, which are included in W and X, the selection is eliminated.

The next step is to determine the percentage thresholds for low, medium, and high supply. I define and calculate the percentage as:

$$r = \frac{\text{Attendance}}{\text{Licensed Capacity}}$$

Then for each county k in Texas, we get a distribution of r^k, from which we can calculate the mean

[1] This part is written in Julia as Stata 14 does not allow post estimation of nonparametric regressions. Alternatively, one can do a linear approximation of the nonparametric regression in Stata 14.

and stand deviation of r^k. Notice that some counties only have one childcare provider that is open, in which case it is not feasible to calculate its standard deviation; therefore this county only has the medium supply scenario. I define the low, medium, and high percentage thresholds as below:

$$\begin{aligned} r^k_{Low} &= E[r^k] - Sd[r^k] \\ r^k_{Medium} &= E[r^k] \\ r^k_{High} &= E[r^k] + Sd[r^k], \end{aligned}$$

where $E[\cdot]$ and $Sd[\cdot]$ stand for expectation and standard deviation, respectively, and they can be directly estimated from the data. Notice that the thresholds r^k_{Low}, r^k_{Medium}, and r^k_{High} are county-specific as opposed to a fixed low cutoff for all counties as before.

The last step is to estimate the potential supply. Let C^k denote the total licensed capacity for all open providers in county k, then the potential supply S^k in county k is defined as:

$$\begin{aligned} S^k_{Low} &= C^k \times r^k_{Low} \\ S^k_{Medium} &= C^k \times r^k_{Medium} \\ S^k_{High} &= C^k \times r^k_{High}. \end{aligned}$$

Recall some counties only have the medium supply scenario, their supply and the number of seats per hundred children are NA and coded as -9999.