Analysis Write-up

Gleb Furman[1]

[1] Who Kneads a PH.D. Bakery

Analysis Write-up

Cluster Analysis

Population Frame

The population frame is composed of data from three sources: (1) the Common Core of Data (CCD), (2) publicly available accountability data, and (3) the U.S. Census. The CCD is a comprehensive database housing annually collected national statistics of all public schools and districts. Accountability data was used to calculate the proportion of students within each school performing at or above proficiency in Math and ELA. Finally, local median income was obtained from the U.S. Census and was matched to each school by zip code. Weighted means of school level variables were calculated using school size (number of students) to generate district level covariates. In this sense, district covariates describe the population of students rather than the population of schools. These are reported in Table 1

Covariates. Selection of covariates was driven by prior research on district and school participation behavior in RCTs (Stuart et al., 2017; Tipton et al., 2016a; Fellers, 2017). Districts and schools with higher proportions of students who are English language learners (ELL), economically disadvantaged (ED), non-White, and living in urban settings are more likely to participate, as are larger districts and schools. It is important to note, however, that some of these characteristics might also make it more likely that researchers would recruit these districts and schools in the first place. Anecdotal evidence from several research teams also suggests that schools are less willing to participate in experimental interventions for subjects in which their students are already excelling, therefore math and ELA achievement covariates were also included.

Omitted Variable. Neighborhood median income was selected to serve as the omitted variable for several reasons. First, it is related to several of the other variables and therefore likely plays a role in the likelihood of a school participating. Second, because it comes from a non-educational data source (the census) and requires additional work to

Table 1

Descriptives of variables

	School		District Weighted	
Variables	Mean	SD	Mean	SD
Number of Schools	NA	NA	9,875.00	0.00
School Size	579.07	203.19	534.77	225.34
Median Income	60,084.98	25,007.61	56,710.63	20,804.82
Average Proportions				
ELA Proficiency	0.59	0.23	0.60	0.20
Math Proficiency	0.53	0.29	0.54	0.26
Economically Disadvantaged	0.59	0.29	0.54	0.24
English Language Learners	0.23	0.21	0.14	0.17
Minority Status	0.65	0.30	0.46	0.32
Total/Free/Reduced Lunch	0.59	0.29	0.53	0.24
Indicators				
Urban	0.40	0.49	0.15	0.33
Suburban	0.41	0.49	0.33	0.44
Town or Rural	0.19	0.39	0.51	0.48

Note. District variables are derived as aggregate means of school variables

include in the population frame it is likely to be omitted in practice.

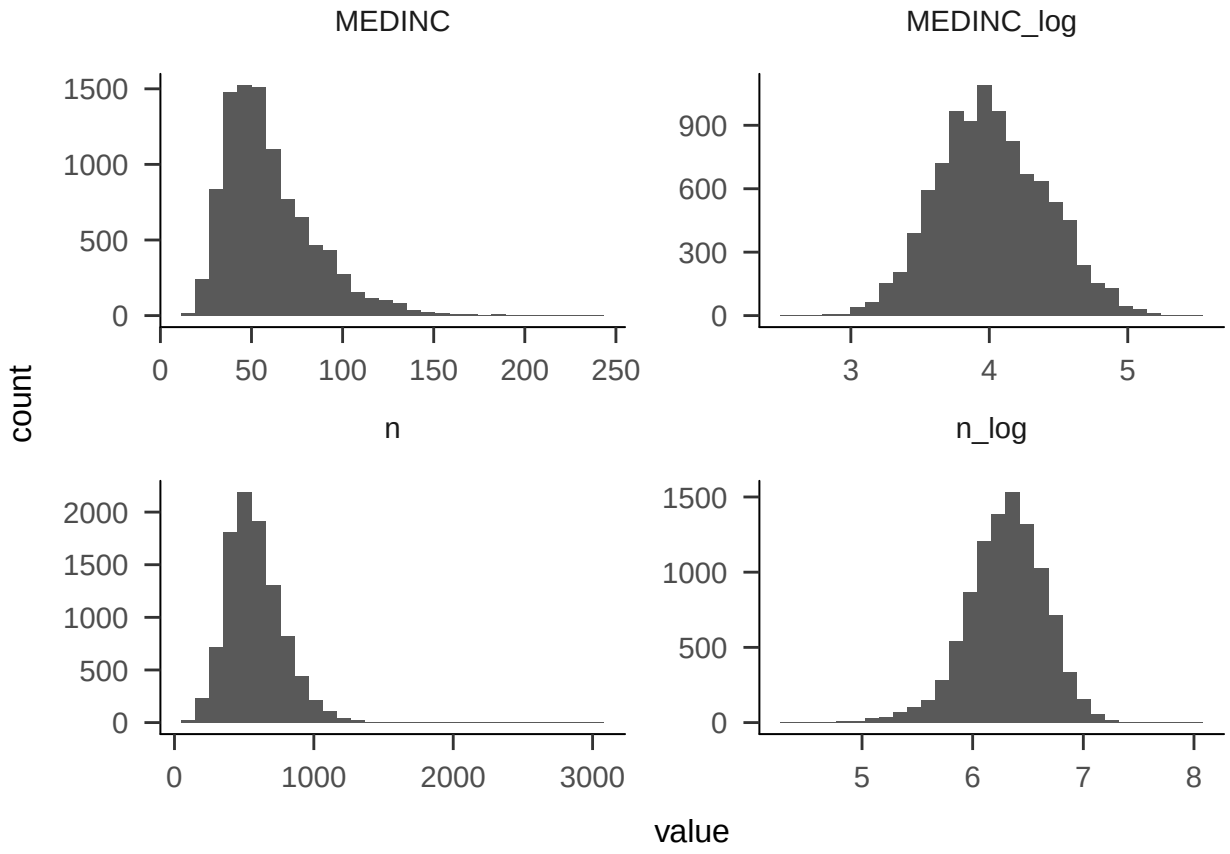

Figure 1. Comparison of covariate distributions and their log transformations.

Variable Transformations. Log-transformation was used on school size (number of students) and median income. This is done to allow proportional comparisons at the extremes of the distributions (Hennig and Liao - 2013). For instance, the difference between two schools with 4000 and 3000 students should be weighed as much as the difference between two schools with 400 and 300 students when generating clusters. Figure 1 displays a comparison of the distribution of these variables and their logs. Figure 2 displays the distribution of the remaining variables.

SUBS

Stratification using balanced sampling (SUBS) was performed prior to simulation because the group of schools in each strata would be static across conditions except where

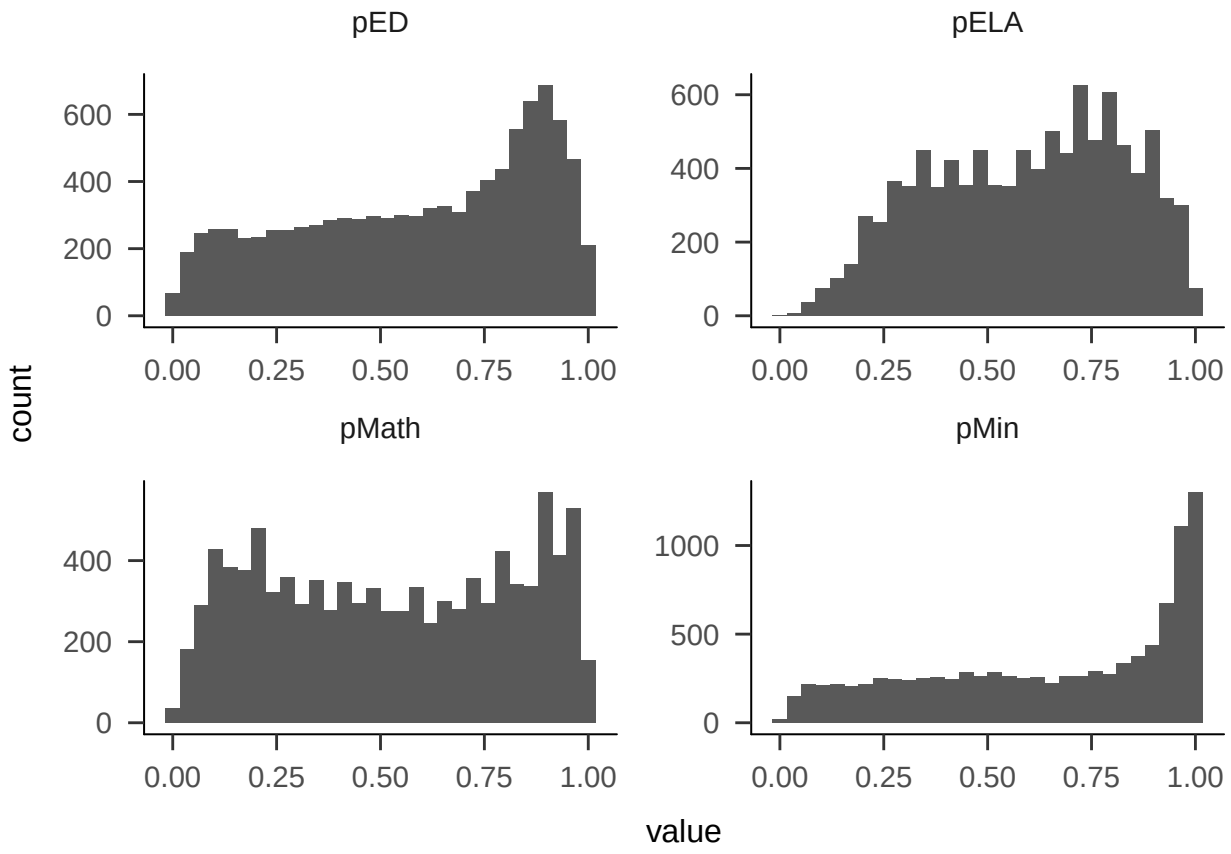

Figure 2. Distributions of the remaining continuous covariates.

the balancing model is manipulated by omitting median income. Per Tipton's (2013) original recomendation, we use k-means clustering to partition the population into strata. This requires selecting a distance metric and choosing the number of strata.

Distance Metric. The set of covariates in both the full model (SUBS-F) and the omitted variable model (SUBS-OV) include continuous covariates as well as binary indicators for urbanicity (urban, suburban, and town/rural). Within this context it is generally recommended to use Gower's (1971) gemeral disimilarity distance (Everitt, 2011) which Tipton (2013) echos. This method relies on different calculations of distance depending on the type of covariates. Let $d_{ii'h}$ be the distance between observed values of covariate X_h for unit i and unit i' where $i \neq i'$. For categorical or dummy coded variables, $d_{ii'h} = 1$ if $X_{ih} = X_{i'h}$ and $d_{ii'h} = 0$ otherwise. For continuous covariates, we use the following formula:

$$d_{ii'h} = 1 - \frac{|X_{ih} - X_{i'h}|}{R_h} \tag{1}$$

where $|.|$ indicates absolute value, X_{ih} and $X_{i'h}$ are values of the h^{th} covariate for units i and i', and R_h is the range of observations for covariate X_h. This method restricts the range of $d_{ii'h}$ to [0,1]. Finally, we calculate the general similarity between each unit pair by taking the weighted average of the distances between all covariates. Let $d^g_{ii'}$ be the general similarity between unit i and unit i' where $i \neq i'$.

$$d^g_{ii'} = \frac{\sum_{h=1}^{p} w_{ii'h} d_{ii'h}}{\sum_{h=1}^{p} w_{ii'h}} \tag{2}$$

where $w_{ii'h} = 0$ if X_h is missing for either unit and $w_{ii'h} = 1$ otherwise.

Number of Clusters. Selecting the number of clusters, k, is one of the most difficult problems in cluster analysis (Steinley, 2006). To date, the most extensive investigation of methods for determining k was conducted by Milligan and Cooper (1985) who analyzed 30 methods. However, aside from the limited generalizability of this study, many methods are also inappropriate in the context of non-hierarchical clustering and thus do not support k-means clustering. Tipton (2013) states that both statistical and practical criteria should be used in selecting the number of clusters. Specifically, a large number of clusters would result in more homogeneous strata and, in turn, a more robust sample. However as strata become smaller they also become more difficult to addequately sample from. Hennig and Liao (2013) also argue that the method of selecting k should depend on the context of the clustering and frame the issue as one of obtaining an appropriate subject-matter-dependent definition of rather than a statistical estimation. Ultimately three considerations were used to select the number of clusters: the ratio of variability between clusters to the sum of within and between cluster variabiltiy as recommended by Tipton (2013), a generalized form of the Calinski-Harabasz index (Calinski and Harabasz, 1974) proposed by Hennig and Liao (2013), and the practicality of sampling from fewer clusters.

- Everitt (2011), p126
- clusterSim
- Continuous data?
  - Calinski and Harabasz (1974)
  - Duda and Hart (1973)
- Steinley, D. (2006) K-means clustering: a half-century synthesis. British Journal of Mathematical & Statistical Psychology, 59, 1–34.
- Milligan and Cooper (1984)
- list 30
- Sugar and James (2003) via Hennig & Liao 2013 p 314
- Modern look

Cluster Analysis. <<<<<<< HEAD

Cluster analysis was performed using the cluster package (Maechler et. al. 2017) in R. First, the *daisy* function is used to compute an n by n pairwise distance matrix across all observations. This function requires two parameters: (1) the data set, and (2) the distance metric. For the full model, the data set included the full list of school level covariates presented in table **??**. For the omitted variable model, median income was omitted from the data set. For both models, the metric was set to "gower". Next the *kmeans* function is used to generate clusters. This method uses an optimization algorithm to classify units into k clusters by minimizing the total within cluster sum of squares. This function also requires two parameters: (1) the distance matrix, and (2) the number of clusters to generate (k). For each k, it is recomended to run *kmeans* at least 10 times, and select the clustering that results in the smallest total within-cluster sum of squares.

Taking the ratio of between to within and plotting it against number of clusters creates a chart allowing us to determine if the added homogeneity of clusters is worth the difficulty of working with a larger of clusters. This chart can be read as an upsidedown elbow graph, where as the slope decreases the tradeoff becomes less worthwile. Additionally, Tipton (2013) recommends selecting the number of clusters such that at least 80% of the variability is between clusters. Figure 4 displays this for both the full and omitted models. In both models at least 10 clusters is necessary for achieving at least 80% between cluster variability. However each additional cluster after 6 seems to add very little benefit individually.

Figures 3, 4, and **??**

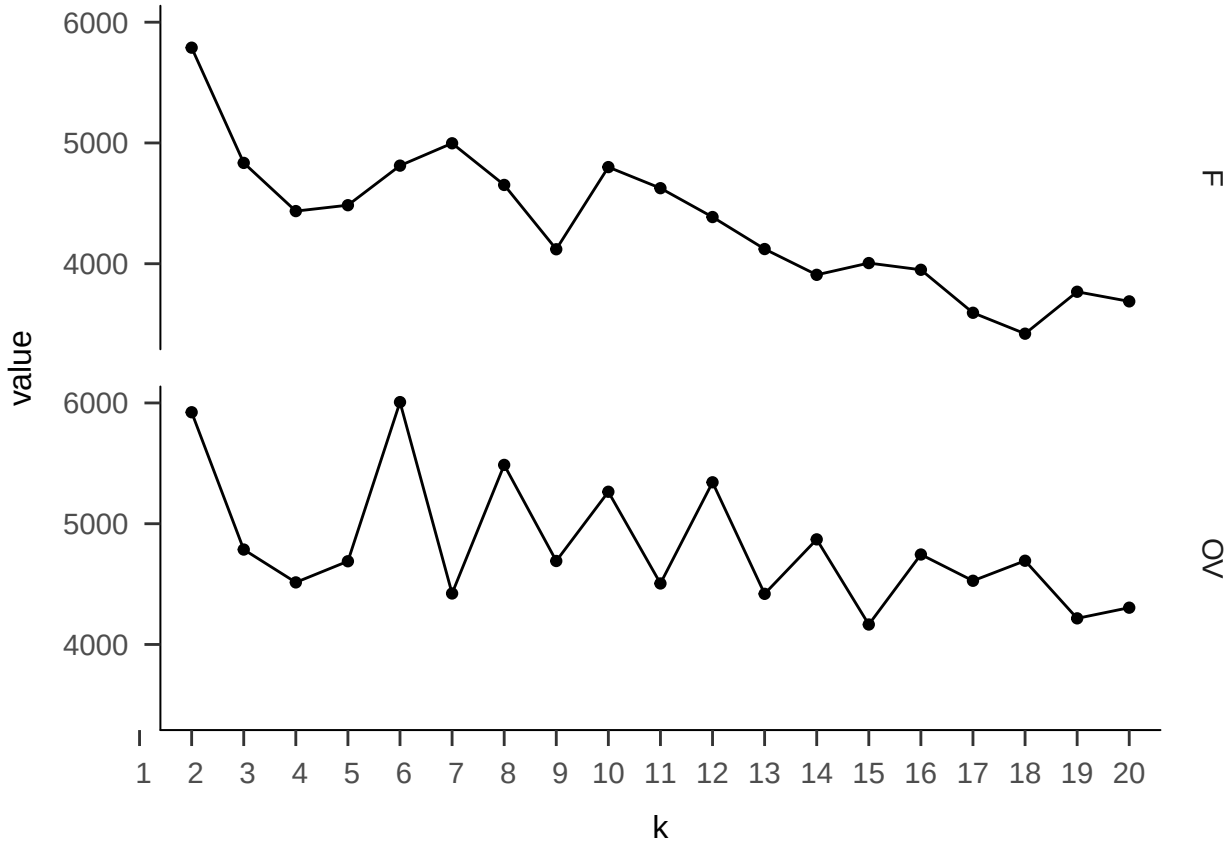

Figure 3. Generalizd Calinski-Harabasz index

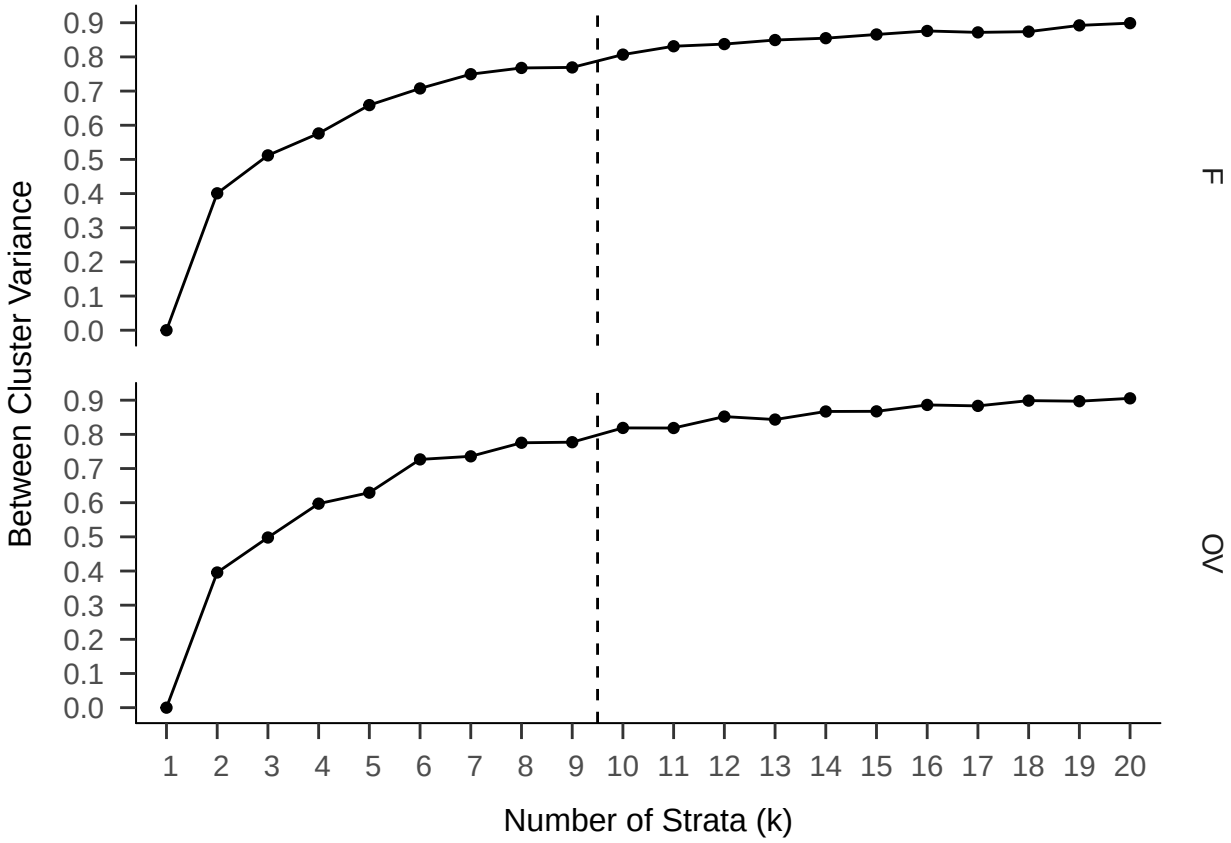

Figure 4. Ratio of between cluster sum of squares to total cluster sum of squares

References

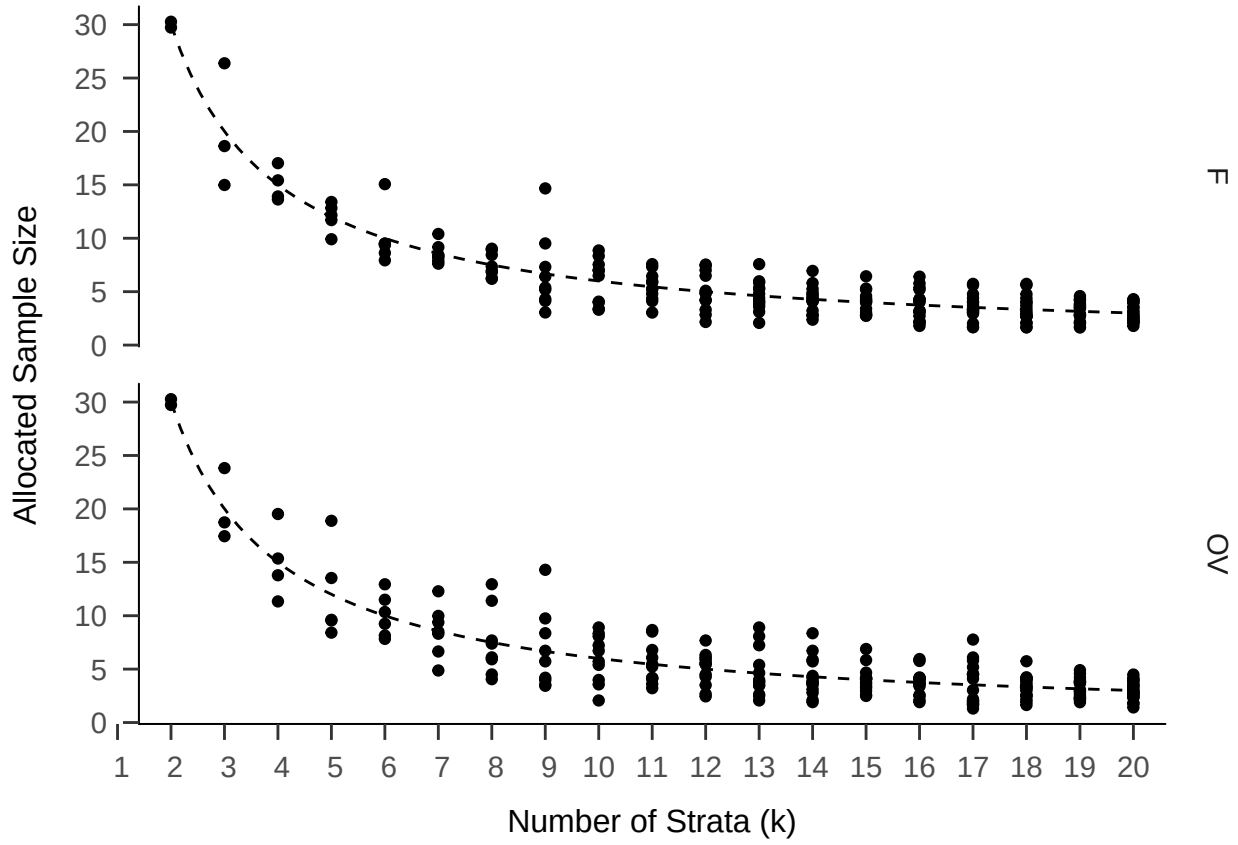
Allocated Sample Size
Number of Strata (k)
F
OV
0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Figure 5