Thesis - Draft

Nathan Yang

Project Background

The goal of this project is to model academic performance in school districts across the United States through various demographic and socioeconomic factors. The data sources include the American Community Survey (ACS), the Educational Opportunity Project, the Longitudinal School Demographic Dataset (LSDD), the Common Core of Data (CCD), and the Census. The data was joined by school district and county to create a comprehensive dataset for analysis.

I created some preliminary simple linear regression models within an interactive dashboard to explore the relationship between academic performance and various socioeconomic factors. I used the shiny (Chang et al. 2023) package to create the dashboard and the ggplot2 (Wickham 2016) package to create visualizations for the data. I used the rsconnect (Atkins et al. 2024) package to deploy the dashboard onto shinyapps.com where it can be publicly accessible.

Due to a lack of datasets aggregated to the school district level, this project focused on expanding my earlier work by incorporating data sources with county level statistics and an improved method for joining datasets to preserve more records and develop a more comprehensive dataset for analysis. Furthermore, I incorporated hierarchical modeling to account for the nested structure of the data and explored the impact of various predictors on academic performance.

Literature Review

The COVID-19 pandemic has had a significant impact on education, with many students experiencing disruptions in learning due to school closures and shifts to remote learning. Several studies have shown that academic performance declined during the pandemic and students from lower income areas were disproportionately affected. For example, Irwin et al. (n.d.) examined the disruption in postsecondary education plans due to the pandemic and found that lower-income families were more likely to experience disruptions in learning from canceled classes. The Educational Opportunity Project (EOP) by Stanford University created a scale for measuring academic performance across all school districts in the US and found that disadvantaged students suffered larger learning loss E. M. Fahle et al. (2023). A followup

study by the EOP also found that test scores recovered from 2022 to 2023 but that nonpoor students had greater gains than poor students, further widening the achievement gap between the two groups E. Fahle et al. (2024). This project utilizes the dataset aggregated and curated by the EOP to model academic performance across school districts and years.

Methodology

The first step of my project was to identify the datasets that I would be using for this project. I started with only looking at datasets aggregated by school district as that would be the most granular level relevant to my research question. I started with the American Community Survey (ACS), the Educational Opportunity Project, the Common Core of Data (CCD), and the Census. I reviewed the data dictionaries for each dataset to understand the variable encoding and the extent of the data. I also reviewed the data sources to understand the data collection process and the limitations of the data.

Data

Data Resources

The core dataset is from the Educational Opportunity Project Reardon (n.d.) and contains academic performance data across school districts. The academic performance varaibles represent difference in grade level relative to the 2019 national average. This dataset contains academic performance variables from 2016 to 2023 aggregated across different student subgroups and subjects within 7390 school districts.

I then identified a dataset from the Census that contains mappings from school district to county "School Districts and Associated Counties" (2021) . Counties typically contain at least one school district and often several. This dataset contains the mappings for 18998 school districts. This dataset was used to join all of the county-level datasets to the academic performance dataset.

The next core datasets were the Data Profiles (DP) from the American Community Survey (ACS). These DPs contain a selection of features from various ACS datasets that are curated to provide a consistent set of features across counties. These datasets contain demographic and socioeconomic features such as income, poverty, housing, education, and employment. These datasets are aggregated at the county level and represent the 2018 to 2022 5-Year estimated statistics for 3222 counties

I did attempt to use ACS datasets that were aggregated by school district but found that the data was very sparse and did not provide enough information for analysis.

The Common Core of Data (CCD) dataset "Local Education Agency (School District) Finance Survey (f-33) Data, (v.1a)" (2022) contains information on membership, salaries, and revenue from local, state, and federal sources. Additionly, this dataset contains COVID

emergency relief funding from the Elementary and Secondary School Emergency Relief Fund (ESSER), which was allotted funding through the Coronavirus Aid, Relief, and Economic Security (CARES) Act for the purposes of education stabilization during the pandemic. This dataset also contains funding statistics from the Governor's Emergency Education Relief Fund (GEER), which was also allotted from the CARES act and was intended to provide emergency support to schools and higher education institutions. Both of these emergency funds also had extensions (ESSER II and GEER II respectively) that were provided through additional legislation. Both funds were also allocated to schools in alignment with Title 1 funding levels which are intended to provide additional funding to schools with a high percentage of students from low-income families. This dataset is aggregated at the school district level and contains data for 19572 school districts for the 2022 fiscal year.

The Covid School Data Hub (CSDH) dataset "Percentage of School Year Spent in-Person, Hybrid, or Virtual" (2023) contains self-reported data from state education agencies on learning modality and enrollment. This dataset is aggregated at the school district level and contains data for 14967 school districts for the 2020-2021 school year.

The Social Capital Project (SCP) dataset contains information on social capital indicators such as family structure, religious attendance, and social trust. "Social Capital Project" (2023) uses survey data from the ACS to construct these subindices with variables such as births per married woman, religious congregations, voting turnout, and violent crimes per population. This dataset is aggregated at the county level and contains data for 3142 counties generated in 2017.

While these datasets come from a 5-year timeframe, these are variables that are relatively stable over time and are not expected to change significantly within this timeframe. The academic performance data is the most dynamic and will be the focus of the analysis. The other datasets will be used to provide context and additional features for the analysis. Additionally, many of these datasets will have fields that are highly correlated with each other and will need to be pruned down to a more manageable set of features.

Data Curation

I first joined the datasets solely by school district name. This was a simple join that matched the exact names of the school districts. However, this method had issues as many school districts had different names in different datasets due to no common naming convention. This would result in many records not being matched and a loss of data.

Next, I reviewed the data sources I initially picked out and identified additional data sources that could be useful for this project. These came primarily from reading various research papers and reports that studied similar topics. I used an excel spreadsheet to track all of the data sources and variables of interest.

After identifying an exhaustive list of datasets and variables, I began the process of downloading and cleaning the data. I used the tidyverse (Wickham et al. 2019) package to clean

and manipulate the data to prepare the datasets to be joined. For my joining process, I used fuzzy matching techniques to join records that had similar school district names but not exact matches. The metrics I used for fuzzy matching were string distance and Jaccard difference.

String distance is a metric that calculates the number of character changes needed to transform one string into another while Jaccard difference is a metric that compares how many 2-letter pairs are shared between two strings. I used the stringdist (Loo 2014) package to calculate both of these metrics and determined thresholds from examining the distributions and matching strength for each metric. I then joined the datasets purely by matching state and calculated the metrics for every pair of school district names within a state. Once I had this dataset with all the potential matches, I developed an extensive filtering process to ensure I got the most accurate matches possible.

1. Filter for matches that both begin with the same letter: This prevents matches names containing North/South and East/West at the beginning are not accidentally mapped together due to the characters in these cardinal directions being similar
2. Filter for matches that end with the same three letters: This prevents matches such as "Abcdefgh county" and "Abcdefgh city" where the school districts may have the same name but are clearly different entities. This also resolves matching names that have numbers at the end such as "Abcdefgh 231" and "Abcdefgh 562" that clearly represent different school districts
3. For each school district in the academic performance dataset, I find its best match based on string distance with ties broken by Jaccard difference (and ties at this stage decided randomly).

This is an example of a dataset joined between my academic performance data and a dataset from the CCD. Using these string comparison metrics, I was able to preserve many records that would have been unmatched if I performed a direct name join. It is especially noticeable with abbreviated words that these metrics help to identify matches like with "Heights" being reduced to "Hts." or "Community" being abbreviated to "Com" as shown below. Additional common abbreviations found in the school district names are "Saint" written as "St." and cardinal directions only represented by the first letter.

Table 1: **Example of School District Matching**. This table shows the similarity between `seda_district` and `ccd_district` using a distance measure and Jaccard index.

seda_district	ccd_district	dist	jaccard
Beaverton Rural Schools	Beaverton Schools	6	0.2727273
North Daviess Community Schools	North Daviess Com Schools	6	0.2580645
Southern Wells Community Schools	Southern Wells Com Schools	6	0.2580645

seda_district	ccd_district	dist	jaccard
North Lawrence Community Schools	North Lawrence Com Schools	6	0.2500000
South Harrison Community Schools	South Harrison Com Schools	6	0.2500000
Greenfield-Central Community Schools	Greenfield-Central Com Schools	6	0.2285714
Minnetonka Public School District	Minneapolis Public School District	5	0.2857143
Morris Area Public Schools	Moorhead Area Public Schools	5	0.2758621
West St. Paul-Mendota Hts.-Eagan	West St. Paul-Mendota Heights-Eagan	5	0.2432432
Minnesota Public School District	Minneapolis Public School District	5	0.2424242
North Branch Public Schools	North Branch Area Public Schools	5	0.1724138
Ridgefield Park School District	Ridgefield School District	5	0.1666667
Hamilton County CUSD 10	Hamilton Co CUSD 10	4	0.2727273
West Washington County CUD 10	West Washington Co CUD 10	4	0.2142857
Rising Sun-Ohio County Com	Rising Sun-Ohio Co Com	4	0.1818182

By joining datasets by exact district name, I would have only had 4441 records with the CCD data. However, using the fuzzy matching techniques, I was able to match 4576 records. This is a 3% increase in the number of records that were matched.

Through district name joining I was only able to match about 250 school districts with ACS income data. However, using the fuzzy matching techniques, I was able to match 281 school districts. This is a 12% increase in the number of records that were matched.

Early on in my project, I only selected datasets that were aggregated by school district and it unfortunately did not prove fruitful as many of the ACS datasets I investigated had very limited data on school districts. This resulted in poor record retention for future dataset merging in addition to reduced modeling data as demonstrated by the ACS income dataset. I transitioned to identifying the counties for school districts in the academic performance dataset and then joining the datasets by county. This proved to be much more successful as I was able to use many ACS Data Profiles (DP) datasets which are a selection of curated features from

various ACS datasets that have greater consistency in data. This allowed me to retain more records and have a more comprehensive dataset for analysis. Only two datasets had sufficient coverage at the school district level, the CCD and the CSDH datasets.

The new comprehensive dataset is much larger and contains more features than the previous dataset. However, the loss of granularity from school district to county may have an impact on the accuracy of the modeling. I keep this in mind throughout my modeling phase and weigh this in when interpreting the results.

The final step for this comprehensive dataset was to filter for records that did not have any missing data in the key variables. This was to ensure that the data was clean and ready for analysis.

Before filtering for missing data, the dataset contained 10842 records with academic performance data across 2019, 2022, and 2023. After applying all filters for missing data, the dataset contained 2208 records. This was a significant reduction in the number of records as many school districts did not have complete coverage across all datasets.

I analyzed the geographic distribution of the data before and after filtering for missing data.

Table 2: Geographic Distribution of Data by State Before Filtering

stateabb	n
CA	1165
IL	1102
PA	928
OH	916
TX	862
MI	795
NJ	703
WI	607
IN	555
MN	506
MA	435
MO	421
GA	343
WA	328
KY	312
KS	305
AL	273
MS	268
TN	261
CT	234
NC	224

stateabb	n
VA	211
AR	206
AZ	191
NE	184
OK	176
NH	146
SC	145
FL	133
ID	130
LA	127
SD	116
WY	76
UT	67
RI	60
WV	55
MD	48
NV	24
ND	14
DC	1

Table 3: Geographic Distribution of Data by State After Filtering

stateabb	n
MI	644
WI	488
WA	311
NJ	256
NC	193
IN	171
NH	135
IL	4
MA	4
TN	2

The geographic diversity is noticeably affected by this filtering process as many states have a significant reduction in the number of school districts. For example, California had 1165 records before filtering and 0 after filtering. This is due to the high number of missing values in the California data. This will be a limitation in the analysis as the data is not representative of all states. In particular, the ACS and SCP data were only available for counties in 21 states

of which only 14 were in common for both datasets. Further filtering brought this down to the 10 states that were common across all datasets.

Due to this non-random pattern, standard imputation techniques such as mean imputation by state were deemed unsuitable. An alternative approach considered converting revenue data into categorical ranges (e.g., "Not Reported," "0-X," etc.), though this approach risked reducing the informative value of the continuous revenue variable.

Modeling

Preparation

I first transformed the dataset into a long format using the pivot_longer() function. This allowed me to convert the wide-format data on yearly math scores into a format suitable for longitudinal modeling.

Next, I created new variables to facilitate trend analysis. I calculated the number of years since 2019 `yearsince2019` and its square (yearsqrdsince2019) to account for potential nonlinear trends over time as there is a general increase in test scores from 2022 to 2023. To standardize the scale of financial variables and put variables on a more similar scale, I converted all revenue and salary figures from raw values into millions of dollars. I also calculated per-student revenue and salaries by dividing total figures by student membership counts as school financial and membership variables had extremely high correlation with each other. Additionally, key socioeconomic metrics such as median income, mean income, and owner-occupied property values were scaled by dividing by 1,000. Percentages for instructional modes (in-person, hybrid, and virtual) were adjusted to range from 0 to 100 for better interpretability.

Curation of Predictors

When modeling, it is important to check for collinearity between predictors because highly correlated predictors can lead to unstable estimates and inflated standard errors. I used the `cor()` function to calculate the correlation matrix for each set of predictors in the dataset and identified highly correlated variables.

For the school modality variables `share_virtual`, `share_inperson`, and `share_hybrid`, I found that share_inperson and share_hybrid were highly correlated. I decided to remove both of these variables and just keep `share_virtual` in the model as that kept interpretations more straightforward as I could delineate between the effect of learning environments that were fully virtual or had some component of inperson learning.

For the revenue and salary variables, I found that the membership, total revenue, and total salary variables were very highly correlated. I decided to remove all the total revenue and total salary variables from the model except for `total_revenue`. Additionally, none of the per

student variables except for `inst_salaries_per_student` were highly correlated with each other so I kept all but that one in the model.

Final List: - `total_revenue` - `revenue_per_student` - `esser_per_student` - `geer_per_student`

For the SCP variables, I know from the data documention of the SCP dataset that County_Level_index is a linear combination of the other social capital variables. As such, I removed all the other social capital variables from the model except for `County_Level_Index`.

Final List: - `County_Level_Index`

For the ACS social characteristic variables, I found that all the variables regarding marital status were highly correlated with each other. I decided to remove all of these variables except for `married_household`. Additionally, I interestingly found that the variables regarding educational attainment were not highly correlated with each other. As such, I kept `over_25_highschool_degree` and `over_25_bachelors_degree` in the model.

Final List: - `married_household` - `over_25_highschool_degree` - `over_25_bachelors_degree`

For the ACS demographic variables, I found that `native_born`, `only_english`, and `non_english` were all highly correlated with each other. I decided to remove `only_english` and `non_english` from the model. Additionally, I found that `with_internet` and `with_computer` were highly correlated with each other. As such, I elected to keep `with_computer` in the model. The only racial variables that had high correlation were `white_percent` and `black_percent`. I keep all the racial variables in the model however as they are all important for understanding the demographic makeup of the school district.

Final List: - `native_born`: Percentage of people born in the US - `with_computer`: Percentage of households with a computer - `white_percent` - `black_percent` - `hispanic_percent` - `asian_percent`

For the ACS employment variables, I found that `no_workers` and `employment_past_year` were very highly correlated but no other variables were. As such I kept all the employment variables except `employment_past_year` in the model.

Final List: - `no_workers`: Percentage of households with no workers - `one_worker`: Percentage of households with one worker - `unemployment`:

For the ACS income variables, `median_income` and `mean_income` jumped out as being extremely correlated with each other. I decided to stick with `median_income` because it is a more robust measure of central tendency. These income variables were also very highly correlated with `owner_occupied_value`, `SMOC`, `rent`, and `mortgage_percentage` so those variables were removed.

Final List: - `median_income` - `with_health_insurance` - `poverty` - `occupancy`

Once I compiled my list of variables, I then analyzed the correlation matrix for the final set of predictors to ensure that there were no highly correlated variables that could lead to multicollinearity in the model.

Following this final check, `with_health_insurance`, `poverty`, `occupancy`, `over_25_bachelors_degree`, `unemployment`, `one_worker`, and `native_born` were removed due to high correlations with other variables.

Two Level Modeling

I started with fitting two unconditional models: an unconditional means model and an unconditional growth model. These models were helpful in understanding the variation in math scores across school districts and over time. Examining this variance would also determine the necessity for hierarchical modeling.

Unconditional Means Model

The unconditional means model evaluates the variation in math scores across school districts and over time without including any predictors. The model is specified as:

$$Y_{ij} = a_0 + \mu_i + \epsilon_{ij}$$

where Y_{ij} is the math score for school district i in year j, α_0 is the overall mean math score, μ_i is the random effect for school district i, and ϵ_{ij} is the residual error.

To determine the necesity for hierarchical modeling, I calculated the intraclass correlation coefficient (ICC) which is the proportion of between-district variance to total variance:

$$ICC = \frac{\text{Between-district variance}}{\text{Between-district variance} + \text{Within-district variance}} = \frac{1.23}{1.23 + 0.374} = 0.767$$

For this model, 76.7% of the total variance in math scores is due to differences between school districts. As such, this supports the need for hierarchical modeling due to the nested structure of the data.

Unconditional Growth Model

The unconditional growth model (UGM) extends the unconditional means model by including a linear time component `year` to estimate how math scores change over time:

$$Y_{ij} = \alpha_0 + \alpha_1 \times \text{yearsince2019} + \mu_i + \epsilon_{ij}\epsilon_{ij} \sim N(0, \sigma^2)\mu_i \sim N(0, \sigma_\mu^2)$$

where α_1 is the fixed effect of time on math scores.

Following this, I created an additional model that incorporated (County_Level_Index) as a fixed effect to capture regional differences:

$$Y_{ij} = \alpha_0 + \alpha_1 \times \text{yearsince2019} + \beta \times \text{County\_Level\_Index} + \mu_i + \epsilon_{ij} \epsilon_{ij} \sim N(0, \sigma^2) \mu_i \sim N(0, \sigma_\mu^2)$$

Three Level Modeling

I then moved on to fitting a three-level hierarchical model as the data had shown considerable variance at the school district level. Multilevel hierarchical modeling would be able to explain the variance at the school district and county level and provide more accurate estimates of the fixed effects.

Unconditional Means Model

Revising my unconditional means model to account for county-level variation:

$$Y_{ijk} = \alpha_0 + \mu_i + \tau_{ij} + \epsilon_{ijk} \epsilon_{ijk} \sim N(0, \sigma^2) \mu_i \sim N(0, \sigma_\mu^2) \tau_{ij} \sim N(0, \sigma_\tau^2)$$

where Y_{ijk} is the math score for school district i in county j in year k, α_0 is the overall mean math score, μ_i is the random effect for county i, τ_{ij} is the random effect for school district j in county i, and ϵ_{ijk} is the residual error.

Unconditional Growth Model

I then extended the unconditional growth model to account for county-level variation:

TODO: Double check this equation and put it on multiple lines

$$Y_{ijk} = \alpha_0 + \alpha_1 \times \text{yearsince2019} + \mu_i + \tau_j + u_1 \times \text{yearsince2019} + u_2 \times \text{yearsince2019} + \epsilon_{ijk} \epsilon_{ijk} \sim N(0, \sigma^2) \mu_i \sim N(0, \sigma_\mu^2) \tau_j$$

where u_1 and u_2 are the random slopes for year at the school district and county levels, respectively.

Variable Selection

Given my previously curated list of predictors, I created a model that included all of these variables to determine their significance and effect on math scores.

After which I filtered the predictors to only include those with a t-value greater than 2. I then created a new reduced model with these 8 predictors. The model was compared to the previous model using anova to determine if the reduced set of predictors significantly improved the model fit. The drop in deviance test showed that the reduced model had a lower AIC and BIC but a very small p value which indicates that the reduced model did not significantly affect the model fit.

```
Warning: Some predictor variables are on very different scales: consider
rescaling
```

Random Effects

Once I had finalized the fixed effects for my model, I then explored the random effects to understand the variation at the school district and county levels.

For all random effects I tested, I fit a 95% confidence interval to the model coefficients to determine the significance of the random effects. In addition to the random slopes for year, I also explored random intercepts for school district and county level variables. Variables such as `share_virtual` were tested as random slopes because it is reasonable to assume that school districts or counties may have been more or less affected by the shift to virtual learning.

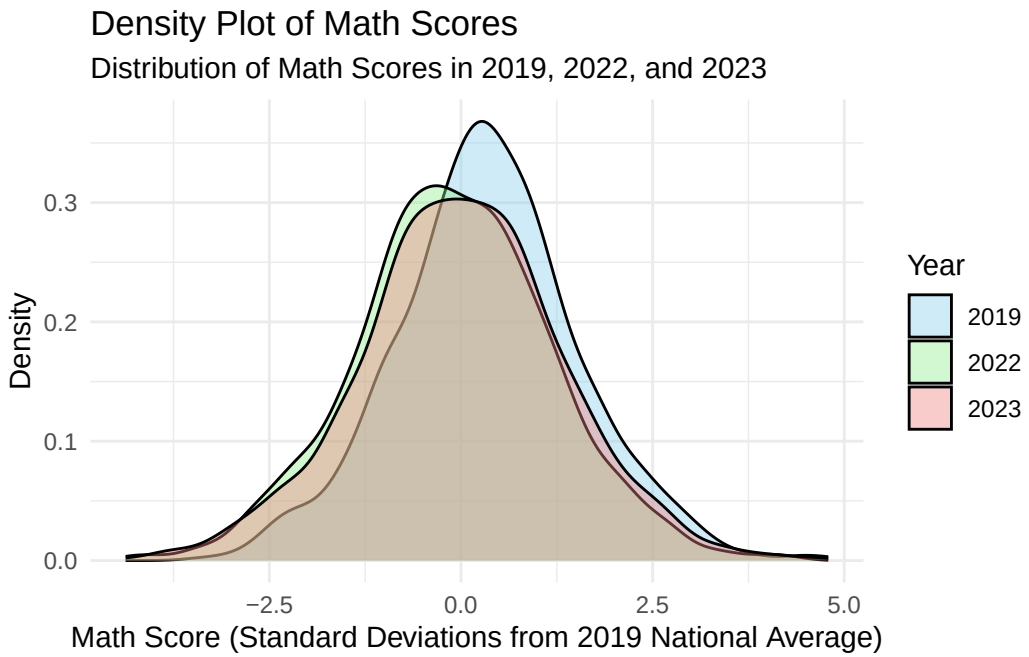
Results

Integrated Dataset Analysis

In addition to the hierachical modeling for predicting academic performance, I needed to understand the distributions and relationships between the features. I used the `ggplot2` (Wickham 2016) package to create histograms, scatter plots, and other visualizations to understand the data better. I also used the `dplyr` (Wickham et al. 2023) package for data manipulation and summarization.

First, I visualized math scores across school districts over time. I created several histograms that showed the distribution of math scores across school districts for each year contained in the data. Additionally, I examined relationships between math scores and my predictors such as school modality, socioeconomic variables, and social capital.

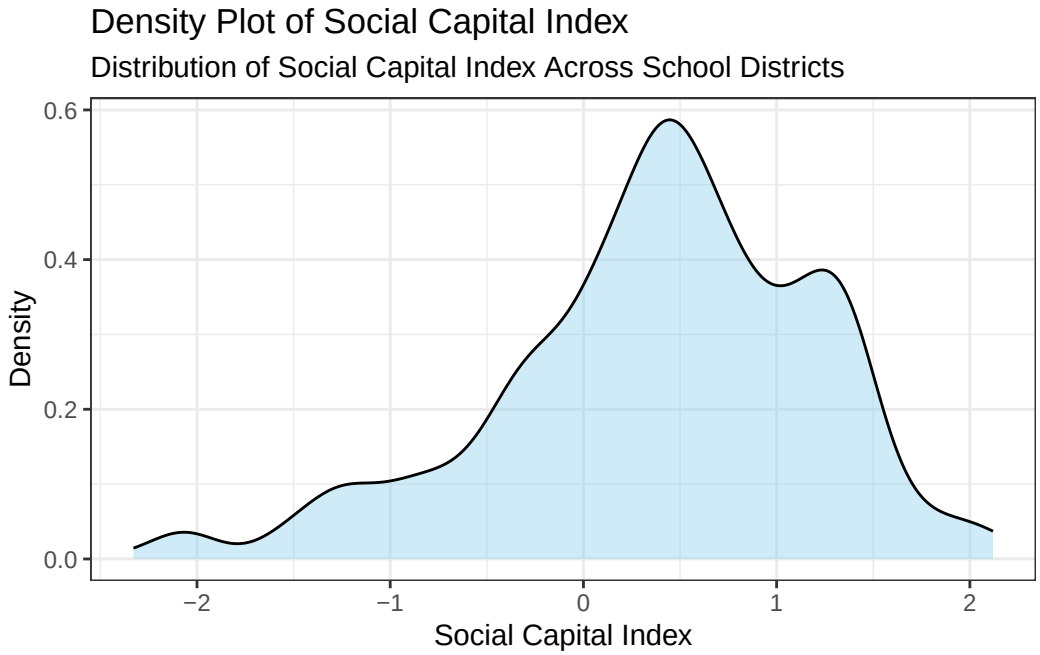
Math Score Distribution

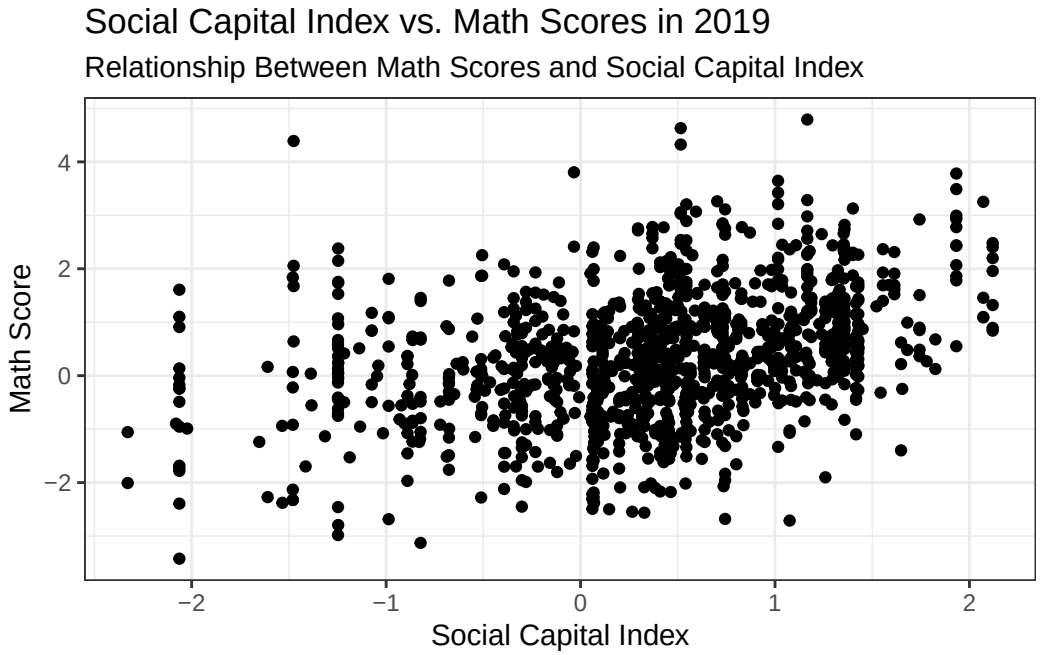


The distribution of math scores is noticeably shifted to the left in 2022 and 2023 compared to 2019. This indicates that the average math scores (in terms of standard deviations from the 2019 national average) across school districts decreased in this time period. This is consistent with the findings of other research that has shown a decline in academic performance during the COVID-19 pandemic.

Social Capital Index

This dataset also introduced new types of data of interest such as the Social Capital Index from the Social Capital Project dataset. This index is a composite measure of social capital that includes indicators such as family structure, religious attendance, social cohesion, and institutional trust. I wanted to explore the relationship between this index and academic performance.

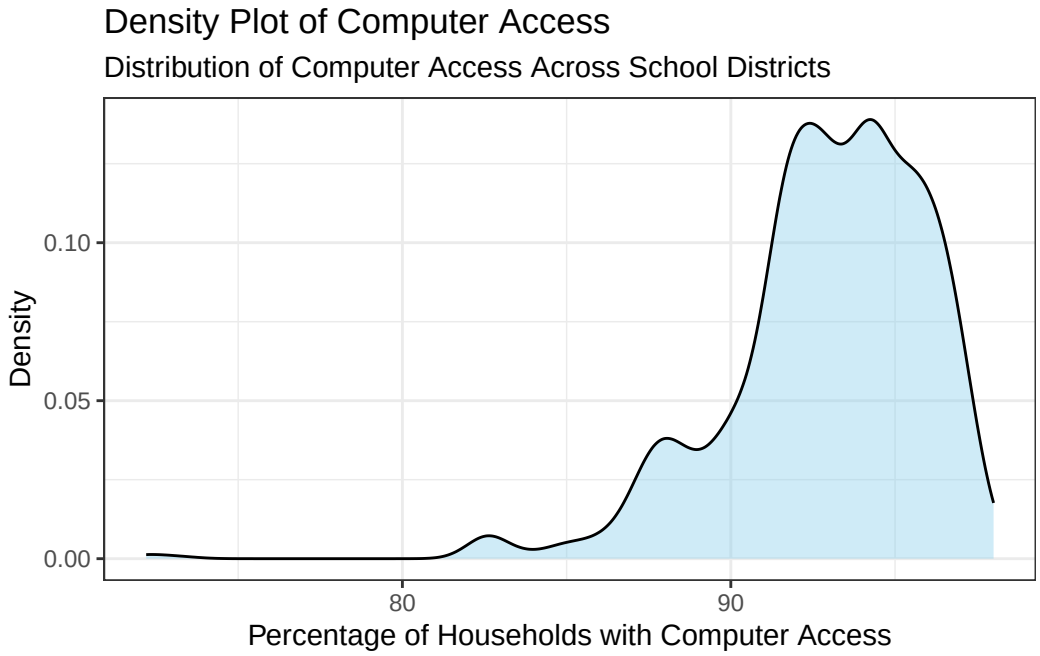


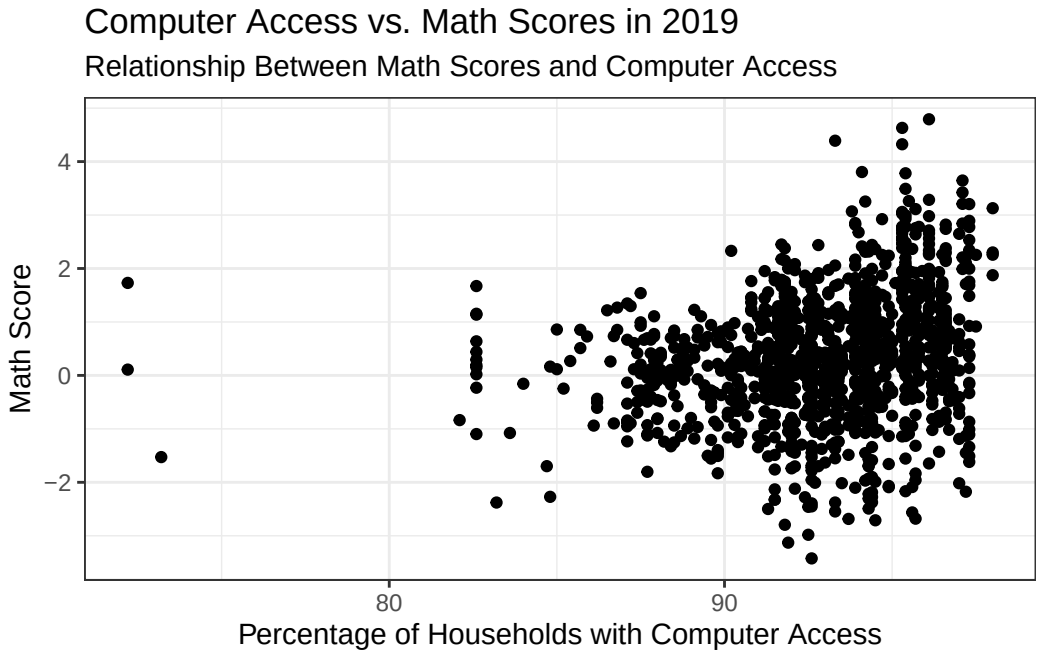


This scatter plot shows a weak positive relationship between the social capital index and math scores in 2019. This suggests that school districts with higher social capital tend to have higher math scores. As such, this remained a key variable of interest for further analysis.

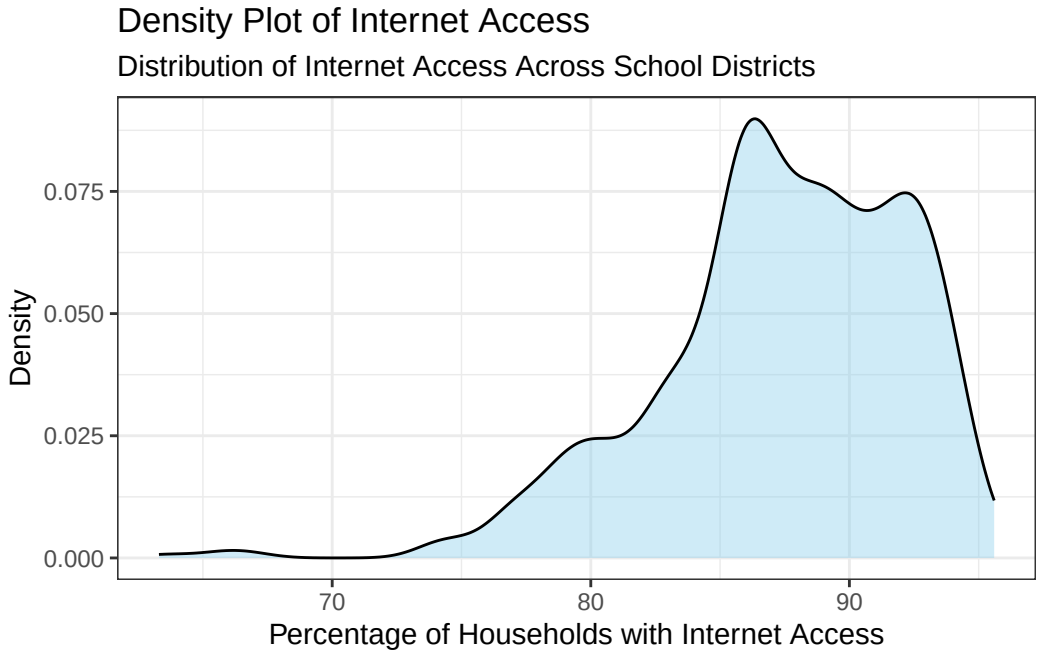
Technology Use

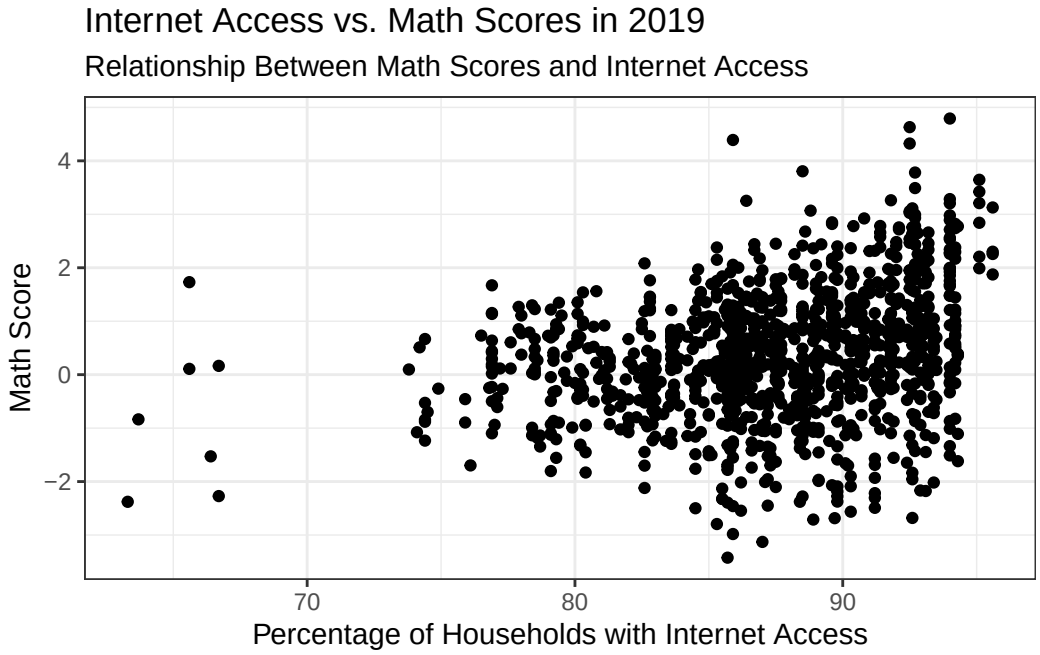
I also wanted to explore the relationship between technology use and academic performance. I wanted to examine how access to technology such as computers and the internet might impact math scores.





There does appear to be a weak positive relationship between computer access and math scores. This suggests that school districts with higher computer access tend to have higher math scores. This relationship is consistent with the idea that access to technology can positively impact academic performance.



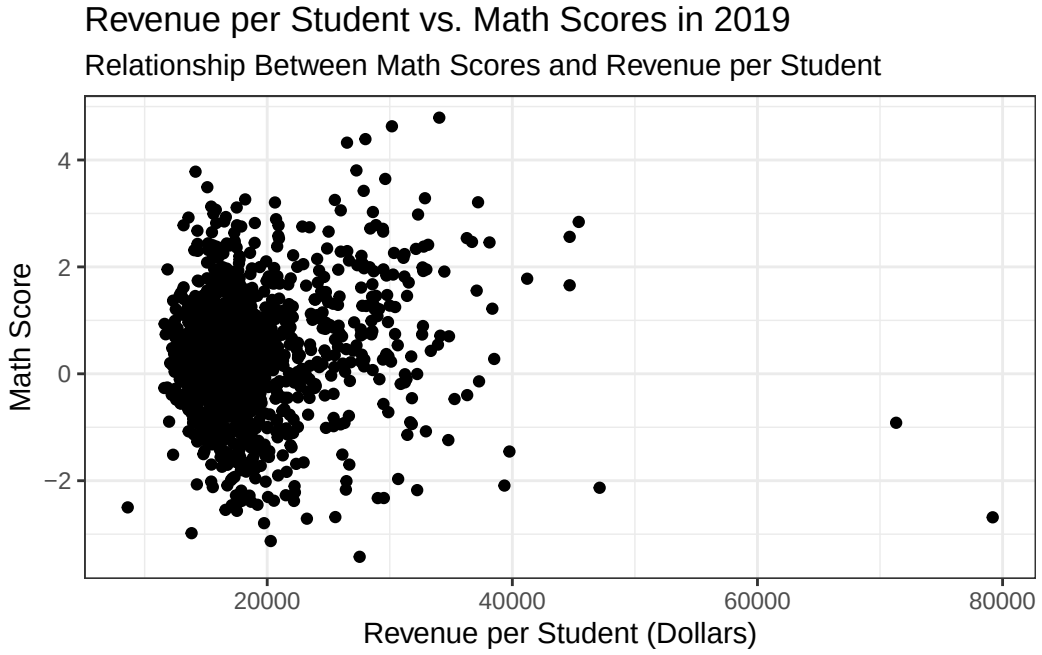


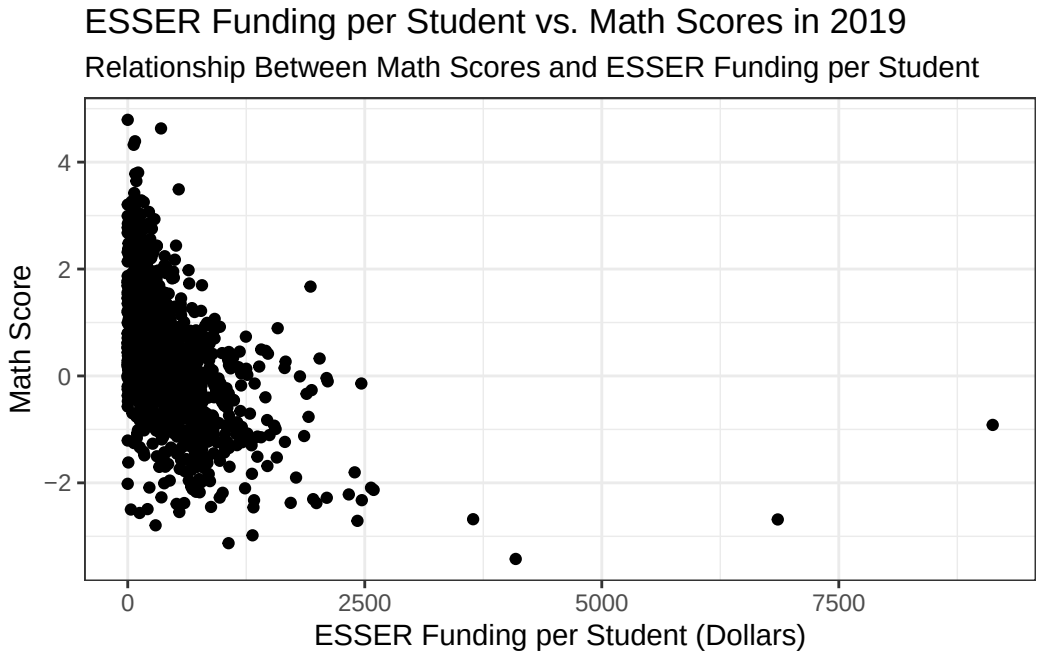
The scatterplot for internet usage also seems to suggest a weak positive relationship with math scores. This indicates that school districts with higher internet access tend to have higher math scores.

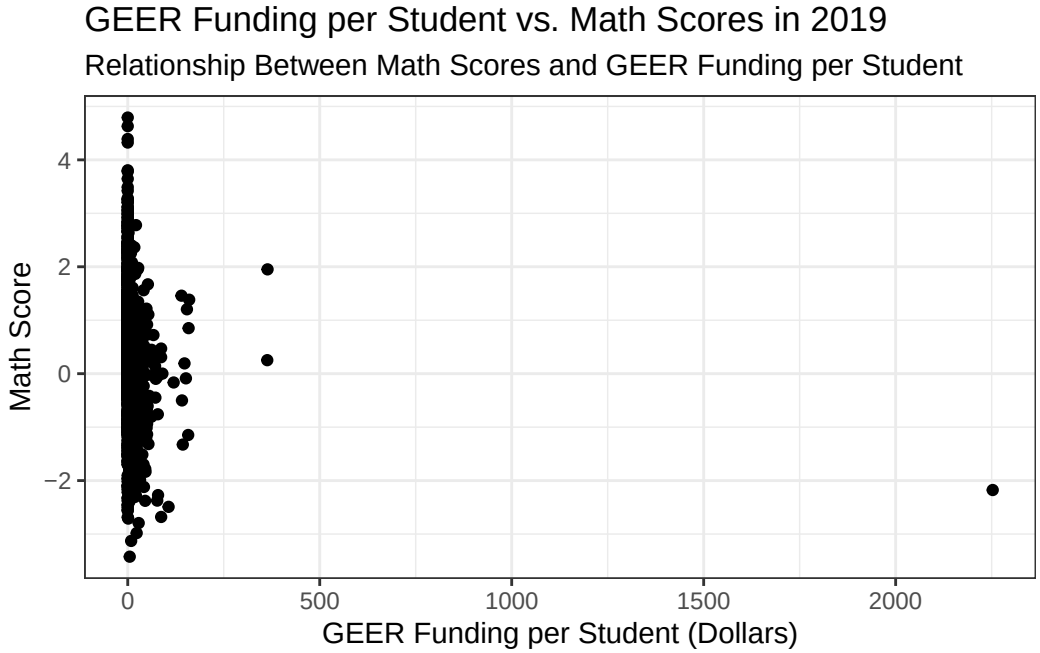
Revenue and Funding

I also wanted to explore the relationship between revenue and funding and academic performance. In particular, I was interested in the COVID-19 emergency relief funding (ESSER and GEER) that was provided to schools during the pandemic. I also wanted to examine the level of funding per student and how that might impact math scores.

revenue_per_student	esser_per_student	geer_per_student
Min. : 8621	Min. : 0.0	Min. : 0.00
1st Qu.:15434	1st Qu.: 170.9	1st Qu.: 0.00
Median :17510	Median : 344.4	Median : 0.00
Mean :19039	Mean : 454.0	Mean : 11.87
3rd Qu.:20650	3rd Qu.: 582.4	3rd Qu.: 12.22
Max. :79193	Max. :9123.4	Max. :2252.76







The summary statistics for each kind of funding variable show that the the vast majority of school districts fall into a similar range except for a small number of outliers. The scatter plots for revenue per student does not indicate any clear relationship with math scores. However, the scatter plots for ESSER and GEER funding per student show a distinct negative relationship

with math scores. This suggests that school districts that received more emergency relief funding were associated with lower math scores.

Two Level Modeling

Unconditional Means Model

```
# A tibble: 3 x 6
  effect   group     term            estimate std.error statistic
  <chr>    <chr>     <chr>              <dbl>     <dbl>     <dbl>
1 fixed    <NA>      (Intercept)       0.0863    0.0368      2.34
2 ran_pars sedaadmin sd__(Intercept)   1.23     NA          NA
3 ran_pars Residual  sd__Observation   0.374    NA          NA
```

The two level unconditional means model produced an intercept of 0.086, representing the mean math score across all districts and years (in terms of standard deviations from the 2019 national average). The variance components were 1.23 and 0.374 for between-district and within-district variance, respectively. As mentioned in the previous section, the ICC was 0.767, indicating that 76.7% of the total variance in math scores was due to differences between school districts and thereby justified the need for hierarchical modeling.

Unconditional Growth Model

```
# A tibble: 4 x 6
  effect   group     term            estimate std.error statistic
  <chr>    <chr>     <chr>              <dbl>     <dbl>     <dbl>
1 fixed    <NA>      (Intercept)       0.319    0.0365      8.74
2 fixed    <NA>      yearsince2019    -0.0947   0.00243   -38.9
3 ran_pars sedaadmin sd__(Intercept)   1.21    NA          NA
4 ran_pars Residual  sd__Observation   0.336   NA          NA
```

The two level unconditional growth model yielded -0.095 as the coefficient for year, showing that math scores have generally been declining over time. The variance components were similar to the UMM model with 1.21 and 0.336 for between-district and within-district variance, respectively. The ICC was 0.783, further supporting the need for hierarchical modeling.

Unconditional Growth Model with County-Level Index

```
# A tibble: 5 x 6
  effect   group     term               estimate std.error statistic
  <chr>    <chr>     <chr>                 <dbl>     <dbl>     <dbl>
1 fixed    <NA>      (Intercept)        191.       4.92         38.9
2 fixed    <NA>      year                -0.0947   0.00243     -38.9
3 fixed    <NA>      County_Level_Index   0.588    0.0405       14.5
4 ran_pars sedaadmin sd__(Intercept)      1.11    NA            NA
5 ran_pars Residual  sd__Observation      0.336   NA            NA
```

The model output showed that the social capital index term was positive, indicating that counties with higher levels of social cohesion and institutional trust tended to have better math scores.

```
refitting model(s) with ML (instead of REML)

Data: integrated_longer
Models:
ugm_model: gys_mn ~ yearsince2019 + (1 | sedaadmin)
ugmc_model: gys_mn ~ year + County_Level_Index + (1 | sedaadmin)
           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
ugm_model     4 9325.6 9352.8 -4658.8   9317.6
ugmc_model    5 9134.2 9168.1 -4562.1   9124.2 193.41  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Furthermore, the anova test showed that the inclusion of the county-level index significantly improved the model fit. The AIC and BIC both dropped considerably along with a very small p-value, indicating a better fit with the additional fixed effect.

Three Level Modeling

Unconditional Means Model

```
# A tibble: 4 x 6
  effect   group        term             estimate std.error statistic
  <chr>    <chr>        <chr>               <dbl>     <dbl>     <dbl>
1 fixed    <NA>         (Intercept)       0.0257    0.0536     0.479
2 ran_pars sedaadmin    sd__(Intercept)   1.03     NA         NA
3 ran_pars County Names sd__(Intercept)   0.627    NA         NA
4 ran_pars Residual     sd__Observation   0.374    NA         NA
```

The three level unconditional means model produced an intercept of 0.0257, representing the mean math score across all districts, counties, and years. The variance components were 0.627, 1.03, and 0.374 for between-county, between-district, and within-district variance, respectively.

Unconditional Growth Model

```
# A tibble: 9 x 6
  effect   group        term                          estimate std.error statistic
  <chr>    <chr>        <chr>                            <dbl>     <dbl>     <dbl>
1 fixed    <NA>         (Intercept)                     0.263    0.0517       5.07
2 fixed    <NA>         yearsince2019                  -0.101    0.00442    -22.9
3 ran_pars sedaadmin    sd__(Intercept)                 0.977   NA           NA
4 ran_pars sedaadmin    cor__(Intercept).yearsince~     0.242   NA           NA
5 ran_pars sedaadmin    sd__yearsince2019               0.0882  NA           NA
6 ran_pars County Names sd__(Intercept)                 0.612   NA           NA
7 ran_pars County Names cor__(Intercept).yearsince~     0.161   NA           NA
8 ran_pars County Names sd__yearsince2019               0.0341  NA           NA
9 ran_pars Residual     sd__Observation                 0.241   NA           NA
```

The three level unconditional growth model yielded a negative coefficient for year since 2019, confirming our previous findings that math scores have been declining over time.

Full Model

```
# A tibble: 27 x 6
   effect group term                   estimate  std.error statistic
   <chr>  <chr> <chr>                     <dbl>      <dbl>     <dbl>
 1 fixed  <NA>  (Intercept)             0.978      2.14         0.458
 2 fixed  <NA>  yearsince2019           -0.343      0.00988    -34.7
 3 fixed  <NA>  yearsqrdsince2019       0.0644     0.00234     27.5
 4 fixed  <NA>  share_virtual           -0.0155     0.00160     -9.69
 5 fixed  <NA>  total_revenue           0.000440   0.000234     1.88
 6 fixed  <NA>  revenue_per_student     13.3        7.41         1.79
 7 fixed  <NA>  esser_per_student      -803.        62.5        -12.9
 8 fixed  <NA>  geer_per_student      -1637.       375.          -4.37
 9 fixed  <NA>  County_Level_Index      0.227      0.0807       2.81
10 fixed  <NA>  married_household       0.00607    0.00982      0.618
# i 17 more rows
```

The full model with all curated variables produced many coefficients that were not signficant when examining the t-value. The reduced model with only the significant predictors from the full model did not significantly improve the model fit as shown by the drop in deviance test.

```
# A tibble: 16 x 6
   effect   group        term                          estimate std.error statistic
   <chr>    <chr>        <chr>                            <dbl>     <dbl>     <dbl>
 1 fixed    <NA>         (Intercept)                   -5.69e-1   0.205       -2.78
 2 fixed    <NA>         yearsince2019                 -3.43e-1   0.00988    -34.7
 3 fixed    <NA>         yearsqrdsince2019              6.44e-2   0.00234     27.5
 4 fixed    <NA>         share_virtual                 -1.59e-2   0.00148    -10.7
 5 fixed    <NA>         esser_per_student             -7.28e+2  52.4        -13.9
 6 fixed    <NA>         geer_per_student              -1.52e+3 376.          -4.04
 7 fixed    <NA>         County_Level_Index             1.46e-1   0.0517       2.83
 8 fixed    <NA>         asian_percent                  3.86e-2   0.0138       2.81
 9 fixed    <NA>         median_income                  1.47e-2   0.00253      5.81
10 ran_pars sedaadmin    sd__(Intercept)                8.44e-1  NA           NA
11 ran_pars sedaadmin    cor__(Intercept).yearsinc~     1.95e-2  NA           NA
12 ran_pars sedaadmin    sd__yearsince2019              1.02e-1  NA           NA
13 ran_pars County Names sd__(Intercept)                3.10e-1  NA           NA
14 ran_pars County Names cor__(Intercept).yearsinc~     1.12e-1  NA           NA
15 ran_pars County Names sd__yearsince2019              3.40e-2  NA           NA
16 ran_pars Residual     sd__Observation                1.87e-1  NA           NA
```

```
refitting model(s) with ML (instead of REML)
```

```
# A tibble: 2 x 9
  term           npar   AIC   BIC logLik deviance statistic    df  p.value
  <chr>         <dbl> <dbl> <dbl>  <dbl>    <dbl>     <dbl> <dbl>    <dbl>
1 reduced_model    16 4659. 4757. -2314.    4627.      NA      NA NA
2 full_model       27 4652. 4818. -2299.    4598.      28.9    11  0.00237
```

AIC increased slightly while BIC decreased slightly with the reduced model. The p-value was very small, indicating that the reduced model did not significantly improve the model fit.

An interesting observation of note was that sex_ratio and black_percent were on the cusp of being significant predictors. Removing one from the model actually made the other significant. This suggests that there may be some correlation between these two variables that was not clearly captured in the EDA.

Discussion

Conclusion

References

Atkins, Aron, Toph Allen, Hadley Wickham, Jonathan McPherson, and JJ Allaire. 2024. "Rsconnect: Deploy Docs, Apps, and APIs to 'Posit Connect', 'Shinyapps.io', and 'RPubs'." https://CRAN.R-project.org/package=rsconnect.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2023. "Shiny: Web Application Framework for r." https://CRAN.R-project.org/package=shiny.

Fahle, Erin M, Thomas J Kane, Tyler Patterson, Sean F Reardon, Douglas O Staiger, and Elizabeth A Stuart. 2023. "School District and Community Factors Associated With Learning Loss During the COVID-19 Pandemic," May.

Fahle, Erin, Thomas J Kane, Sean F Reardon, and Douglas O Staiger. 2024. "The First Year of Pandemic Recovery: A District-Level Analysis," January.

Irwin, Véronique, Ke Wang, Sarah Hein, Jijun Zhang, Riley Burr, Ashley Roberts, Amy Barmer, et al. n.d. "Report on the Condition of Education 2022."

"Local Education Agency (School District) Finance Survey (f-33) Data, (v.1a)." 2022. https://nces.ed.gov/ccd/Data/zip/sdf22_1a_sas7bdat.zip.

Loo, M. P. J. van der. 2014. "The Stringdist Package for Approximate String Matching" 6: 111–22. https://CRAN.R-project.org/package=stringdist.

"Percentage of School Year Spent in-Person, Hybrid, or Virtual." 2023. https://assets.ctfassets.net/9fbw4onh0qc1/XfBEuMLMOBgHrhmjBdVpc/8e555b362876da16ba52c85be5b2effe/District_Overall_Shares_03.08.23.csv.

Reardon, Ho, S. F. n.d. "Stanford Education Data Archive (Version 5.0)." https://purl.stanford.edu/cs829jn7849.

"School Districts and Associated Counties." 2021. https://www2.census.gov/programs-surveys/saipe/guidance-geographies/districts-counties/sdlist-21.xls.

"Social Capital Project." 2023. https://www.jec.senate.gov/public/index.cfm/republicans/socialcapitalproject.

Wickham, Hadley. 2016. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the {Tidyverse}" 4: 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.