ISyE 4031 T09 - Georgia Achievement Gaps in K-12 Schools with Regression

Ruwei Ma, Yichen Ma, Yang Yang

1. Introduction

Covid-19 had brought a big impact to the education system across US. National test results for 2022 reveal the pandemic's devastating effects on American schoolchildren, with the performance of 9-year-olds in math and reading dropping to the lowest levels from two decades ago [1]. This lagging effect from the pandemic applies to all races and income levels and sparks a collective decline in academics for the generation that experienced school closures, frequent reliance on virtual and remote learning, and other pandemic effects. The setbacks will occupy the low-performing students for up to 9 months to catch up with the average, prompting an urgent need for the underlying solution to the achievement gap [2]. This setback further adds to, and likely aggravates, the pre-pandemic disparity in student achievement outcomes for vulnerable and at-risk student populations, especially in Georgia. Based on some of my preliminary analysis of the 2021 achievement data across 2,180 schools in Georgia, we found that there are 2 prominent factors that affect achievement rate: the student's economic status and race. The achievement rate in 2021 of economically disadvantaged students is 46.11%, compared to 52.32% across all students. A similar gap can be observed in the difference in achievement rate between white and black students in Georgia, the former as high as 66.99%, compared to the 39.88% of the latter. The gap within the economically-disadvantaged students' group is vast and depends on the county or school they attend. Further analysis at the school level shows strong correlation between achievement rate and the school's other demographics.

2. Problem Goal

We aim to adopt regression modeling to identify gaps in national test achievement rates between different demographic groups in Georgia, and recommend robust strategies to address such disparities. Specifically, the objectives are: (1) visualize the disparities in school resources, such as teacher certifications and FTE (Full-time Equivalent), and quantify its correlation with the student's achievement outcomes, especially among marginalized minority groups (e.g., White, Black, vs. Hispanic students, economically disadvantaged vs. affluent students, and rural vs. Urban schools) (2) quantify the achievement gap at the county level across Georgia's 159 counties at the school level to identify factors that predict student achievement and highlight intervention or resource allocation strategies, and (3) evaluate the impact and predict the trajectory of the policies and strategies produced from step 2 with adjustments.

3. Executive Summary

4. Data Description

```
# write.csv(data$All.Students.Math.Achievement, "all_students_math_achievement.csv")
```

Exporting the data to ExpertFit to fit distribution and test normality.

a. Data Summary

```
library(nortest)
ad.test(data$All.Students.Math.Achievement)
```

```
##
##  Anderson-Darling normality test
##
## data:  data$All.Students.Math.Achievement
## A = 146.39, p-value < 2.2e-16
```

```
##
## Attaching package: 'huxtable'
```

```
## The following object is masked from 'package:dplyr':
##
##     add_rownames
```

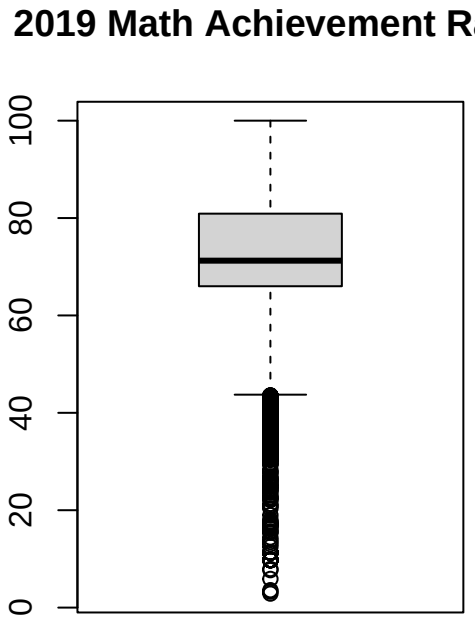
	2019	**2021**
Observations	7208.00	7208.00
Avg. Math achievement	69.8427524972253	56.023041065483
Median Math achievement	71.25	57.9
Lower Bound of Math achievement	2.88	0
Upper Bound of Math achievement	100	100
Standard Deviation	16.9097145750172	20.895088622954

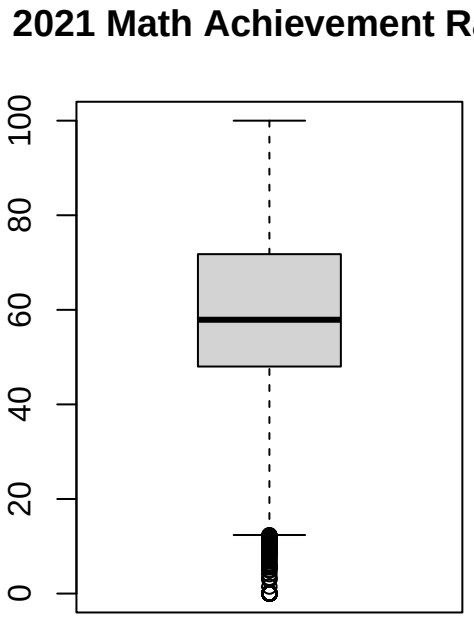
Mean and median Math test achievement rates are higher in 2019 than in 2021.

```
#average change in achievement rate
(52.23121-67.99686)/67.99686
```

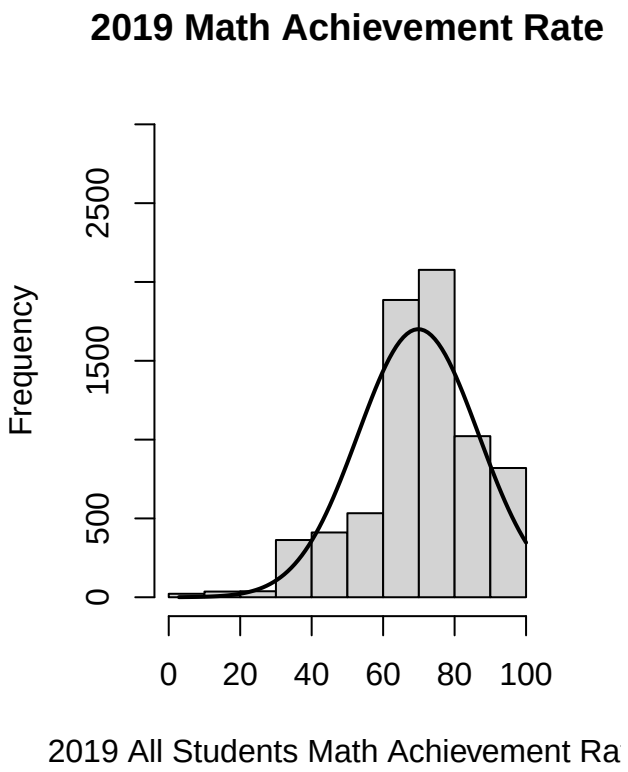
```
## [1] -0.2318585
```

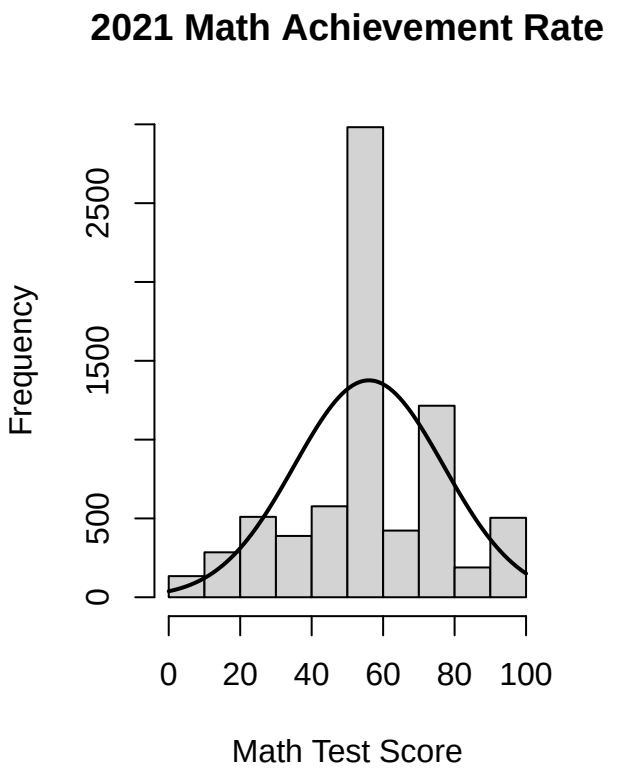
c. Data Visulization





The boxplot of both years' math achievement rate shows that in 2019, the data distribution is more compact, and all quartiles are significantly higher than those in 2021. A tremendous number of outliers are identified in both year's boxplots, suggesting many data points below the lower quartile by more than 1.5 interquartile range (IQR). Achievement rates are highly left skewed.





From both years' histogram, it can be confirmed that there is a very low frequency of math achievement rate between 0-30 for the 2019 data, as compared to the 2021 data. More outliers in the 2019 data could mean a higher . From plain sight, the 2019 data is better approximated by a normal distribution. The 2021 data seems skewed to the center.

d. Table of Variables

Variables	Description	Type
y1	2019 All Students Math Achievement Rate	Quantative
y2	2021 All Students Math Achievement Rate	Quantative
x1	Absent 0-5 Days Percentage	Quantative
x2	Absent 6-15 Days Percentage	Quantative
x3	Absent 15+ Days Percentage	Quantative
x4	Avg. Annual Salaries - Administrators	Quantative
x5	Avg. Annual Salaries - Teachers	Quantative
x6	Avg. Annual Salaries - Support.Personnel	Quantative
x7	Number of Teachers with a phd degree	Quantative
x8	Total Number of Certified Teachers	Quantative
x9	Post Grad Percentage	Quantative
x10	Total Students Enrolled	Quantative
x11	Teacher-Student Ratio	Quantative
x12	White Student Percentage	Quantative
x13	Black Student Percentage	Quantative
x14	Economically Disadvantaged Student Percentage	Quantative
x15	Directly Certified Students Percentage	Quantative
x16	Amount of Money Invested for Students	Quantative
x17	Per-Pupil Expenditure at School Level	Quantative
x18	Rate of Entries and Withdrawls to a School	Quantative
x19	Percentage of Gifted Students	Quantative
x20	Urban/Rural Area of the School	Qualitative

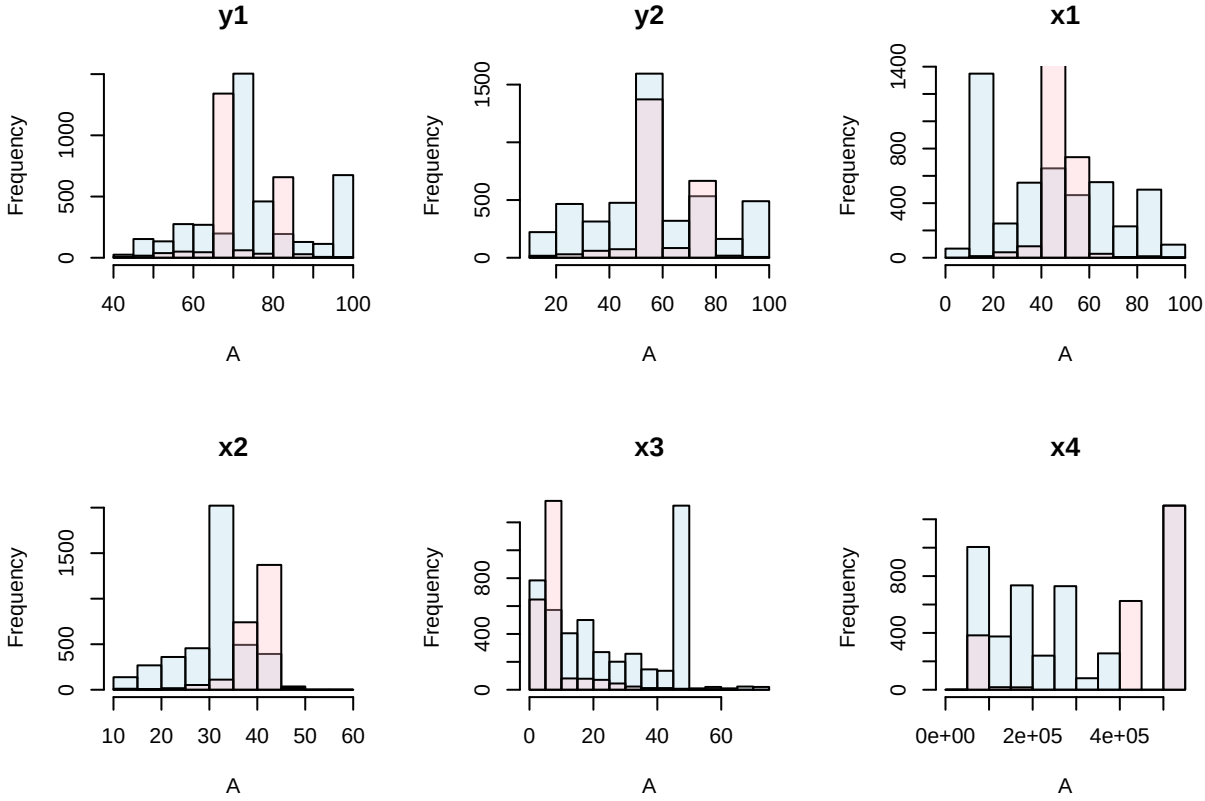
5. Regression Analysis

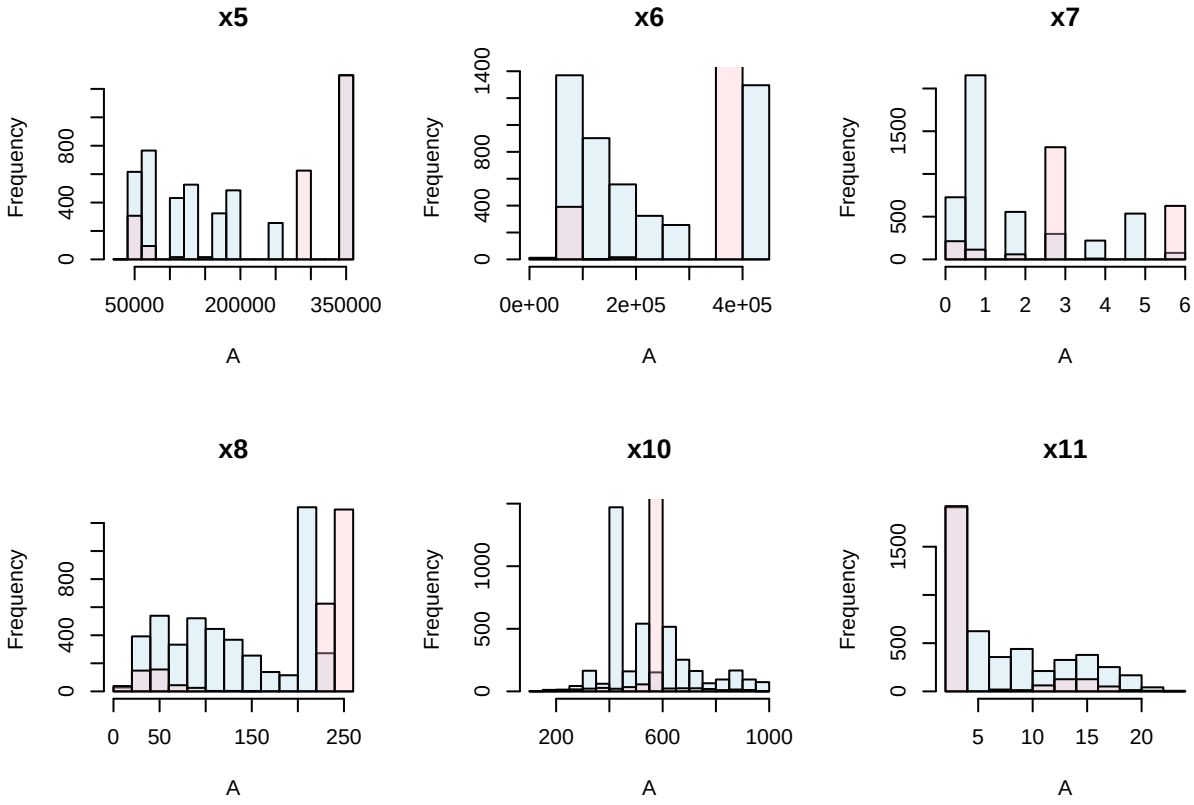
a. Iterations of the analysis process

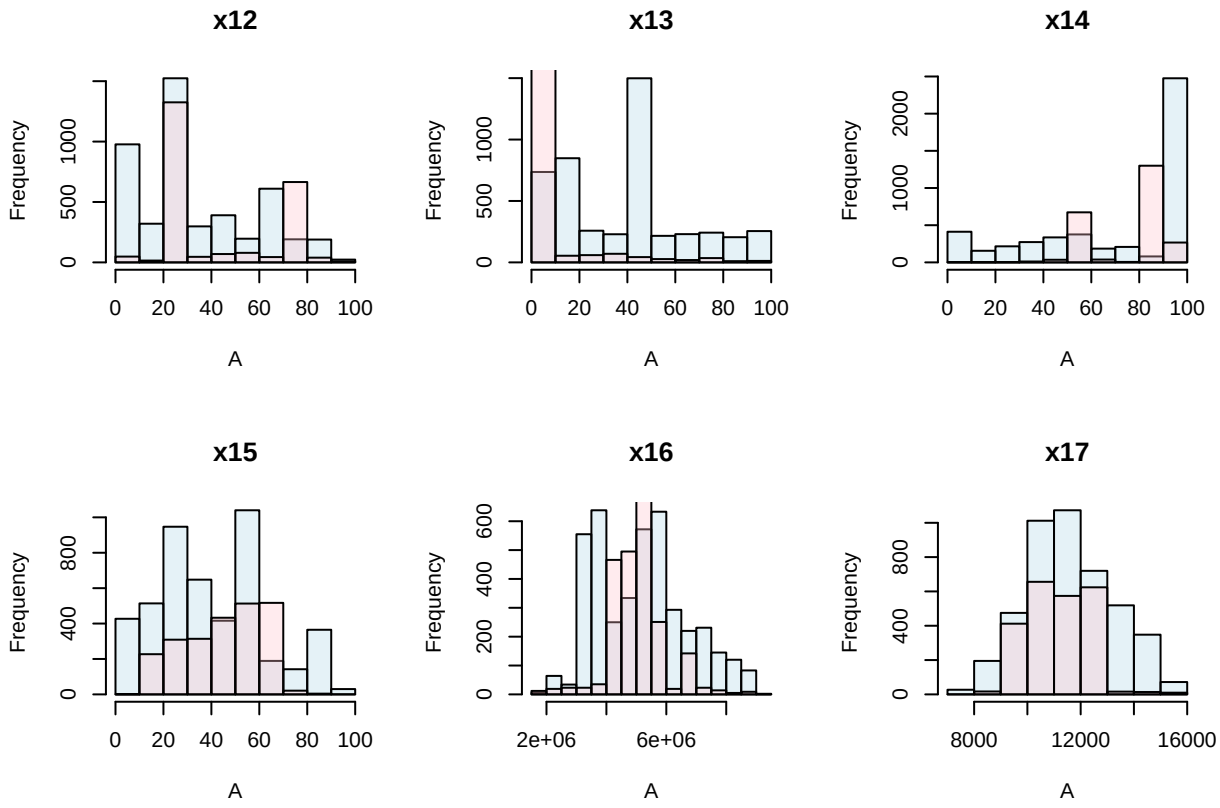
- paragraph description

c. Plots of variables- Scatterplot

For the plots below, a light blue color indicates Urban Area and a light pink color indicates Rural Area.

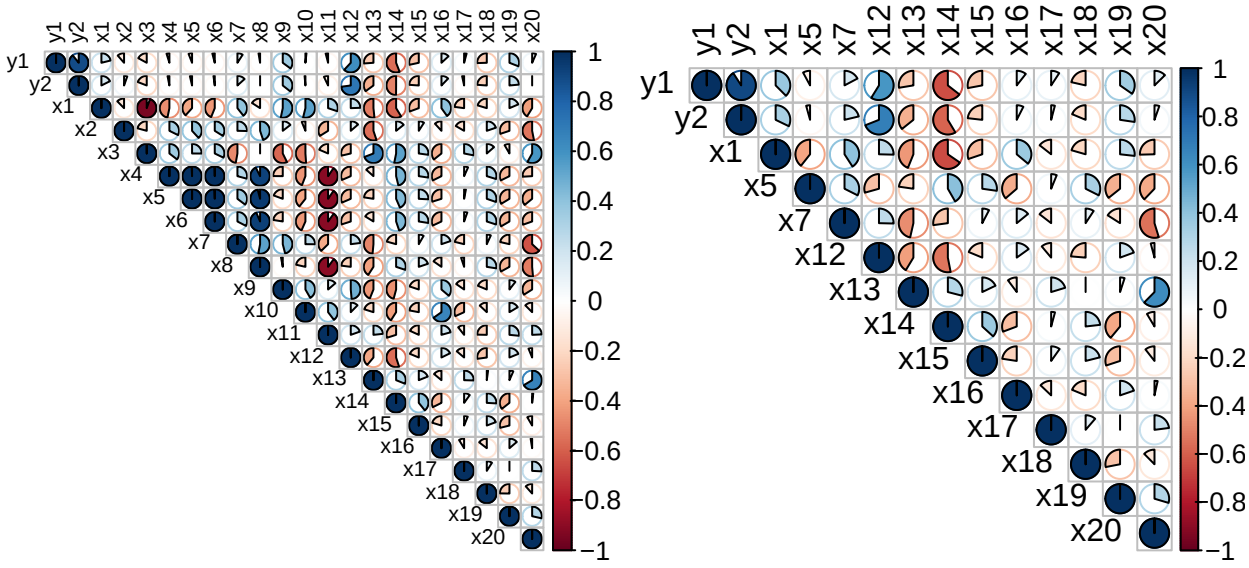



x5
Frequency
0
400
800
50000
200000
350000
A
x6
Frequency
0
400
800
1400
0e+00
2e+05
4e+05
A
x7
Frequency
0
500
1500
0
1
2
3
4
5
6
A
x8
Frequency
0
400
800
0
50
150
250
A
x10
Frequency
0
500
1000
200
600
1000
A
x11
Frequency
0
500
1500
5
10
15
20
A



b. Multicollinearity

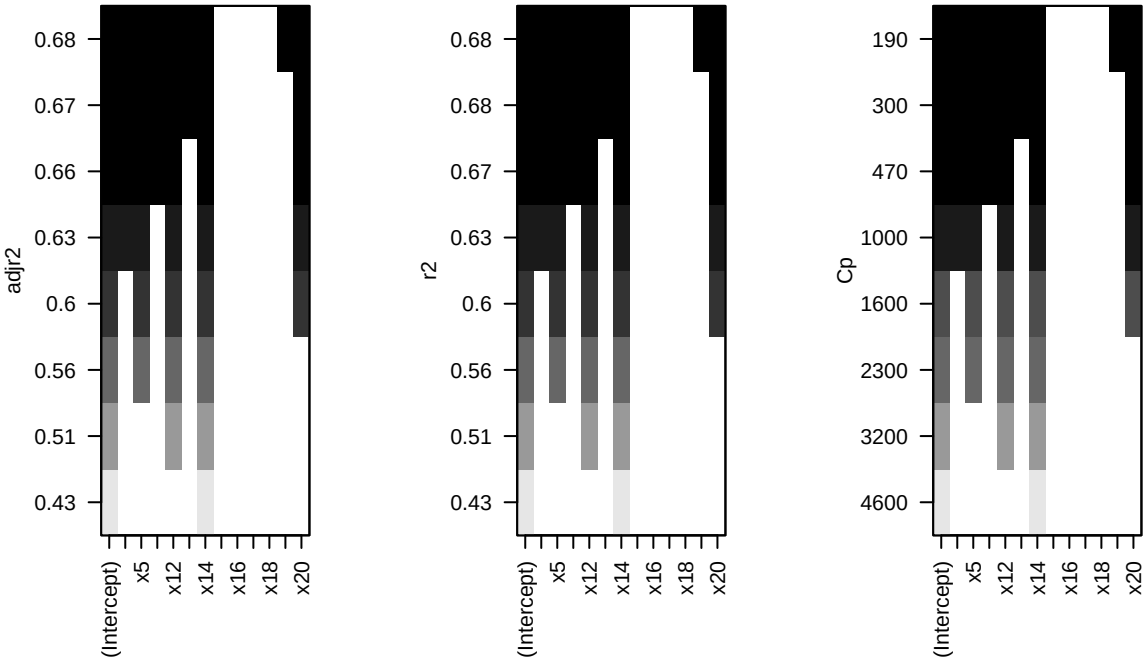
```
## corrplot 0.92 loaded
```



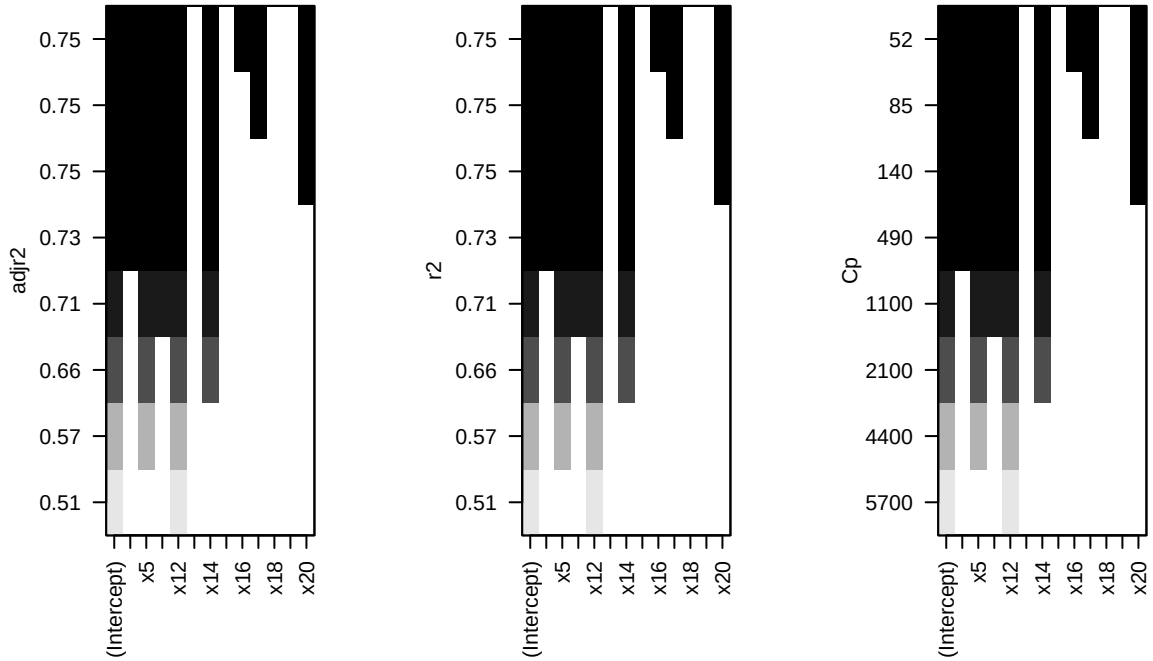
Before doing the model selection process, a Multicollinearity check produces high correlation of (x1:x4), (x4:x6,x7), (x5:x7), (x9:x4,x5,x6), and (x12:x4,x5,x6,x8). And another set of variables that have a high correlation is y1 and y2, since we are modeling them separately as response variables, we do not need to drop any of them. The renewed plot is on the right.

d. Model Selection

2019 Model Selection



2021 Model Selection



d. Best Model

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:huxtable':
##
##     add_footnote

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Based on the model selection, the best model for the 2019 Math Achievement Rate consists of independent variables of 'Absent 0-5 Days Percentage', 'Avg. Annual Salaries for Teachers', 'Number of Teachers with a phd degree', 'White Student Percentage', 'Black Student Percentage', 'Economically Disadvantaged Student Percentage', 'Percentage of Gifted Students', and 'Urban/Rural Area of the School'. The best model for the 2021 Math Achievement Rate consists of independent variables of 'Absent 0-5 Days Percentage', 'Avg. Annual Salaries for Teachers', 'Number of Teachers with a phd degree', 'White Student Percentage', 'Economically Disadvantaged Student Percentage', 'Amount of Money Invested for Students', 'Per-Pupil Expenditure at School Level', and 'Urban/Rural Area of the School'.

	2019 Best Model	2021 Best Model
(Intercept)	49.054	13.842
	(0.923)	(1.680)
x1	0.231	0.283
	(0.007)	(0.010)
x5	0.000 06	0.000 09
	(0.000 001)	(0.000 001)
x7	−1.612	−2.952
	(0.071)	(0.084)
x12	0.261	0.482
	(0.005)	(0.006)
x13	0.052	
	(0.006)	
x14	−0.176	−0.182
	(0.005)	(0.007)
x19	0.231	
	(0.016)	
x20	5.203	5.530
	(0.274)	(0.328)
x16		0.000 000 6
		(0.000 000 1)
x17		0.0007
		(0.000 08)
Num.Obs.	6288	6038
R2	0.714	0.760
R2 Adj.	0.714	0.759
AIC	41 919.9	43 405.1
BIC	41 987.3	43 472.1
Log.Lik.	−20 949.929	−21 692.529
RMSE	6.77	8.79

e. Best Model (Outlier Excluded)

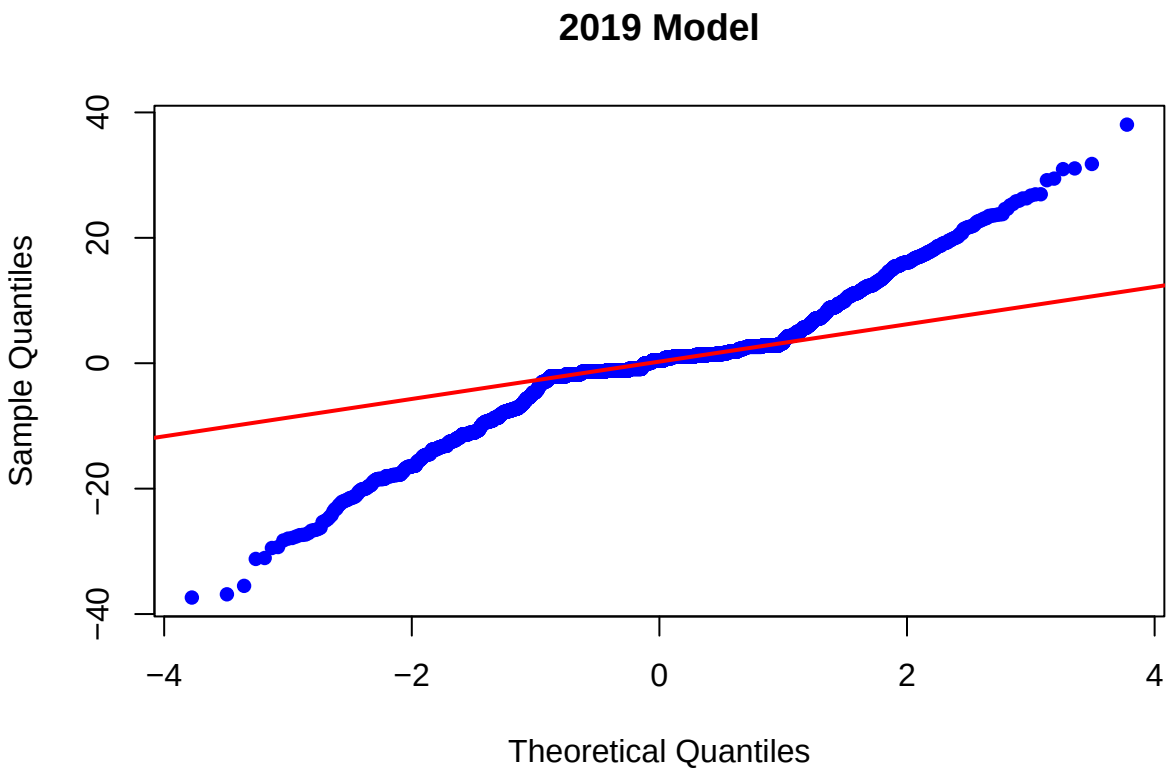
2019

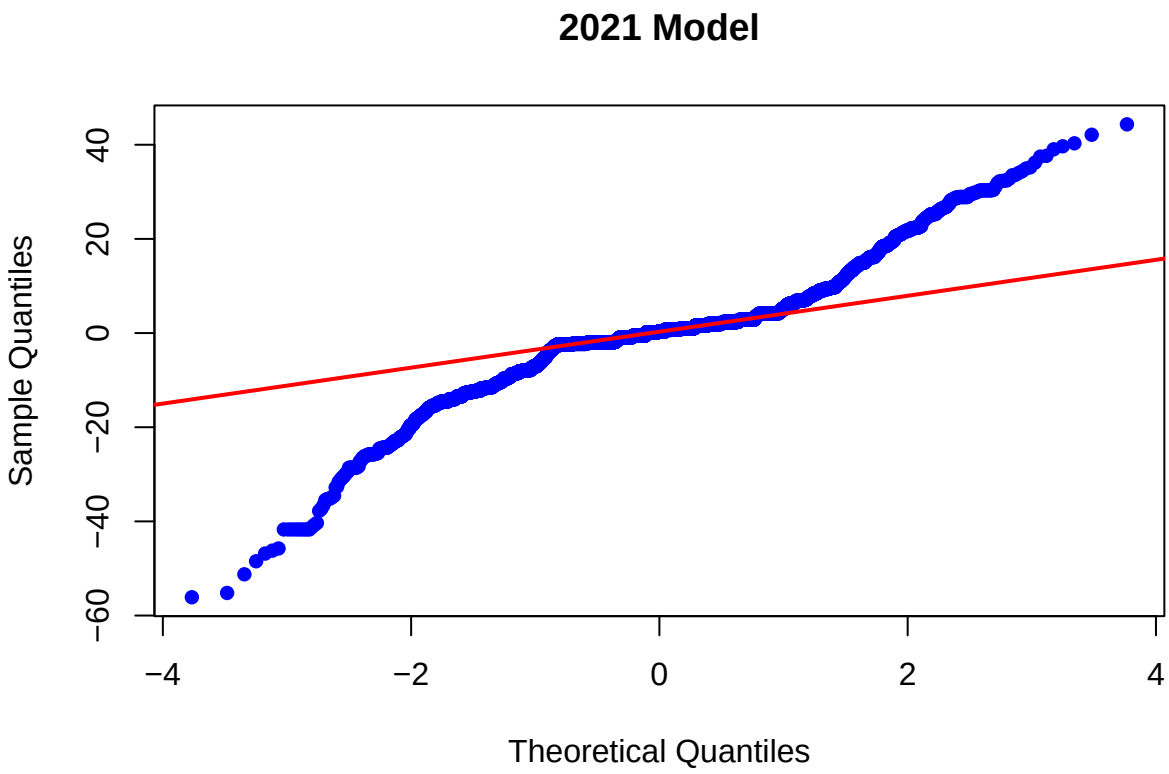
2021

f. Normality Check

```
##
##  Anderson-Darling normality test
##
## data:  resid(best_model_2019)
## A = 244.57, p-value < 2.2e-16
```

```
##
##  Anderson-Darling normality test
##
## data:  resid(best_model_2021)
## A = 211.46, p-value < 2.2e-16
```





g. Transformation

2019

```
##
## Call:
## lm(formula = y1 ~ x1 + x2 + x5 + x7 + x12 + x14 + x17 + x19,
##     data = data_numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.608  -1.823  -0.182   2.115  34.049
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.752e+01  1.210e+00  55.786   <2e-16 ***
## x1           9.901e-02  7.027e-03  14.090   <2e-16 ***
## x2          -3.614e-01  1.670e-02 -21.641   <2e-16 ***
## x5           5.021e-05  1.084e-06  46.321   <2e-16 ***
## x7          -1.771e+00  6.870e-02 -25.785   <2e-16 ***
## x12          2.573e-01  4.777e-03  53.867   <2e-16 ***
## x14         -1.756e-01  5.305e-03 -33.105   <2e-16 ***
## x17          5.987e-04  6.692e-05   8.946   <2e-16 ***
## x19          2.351e-01  1.766e-02  13.312   <2e-16 ***
## ---
```

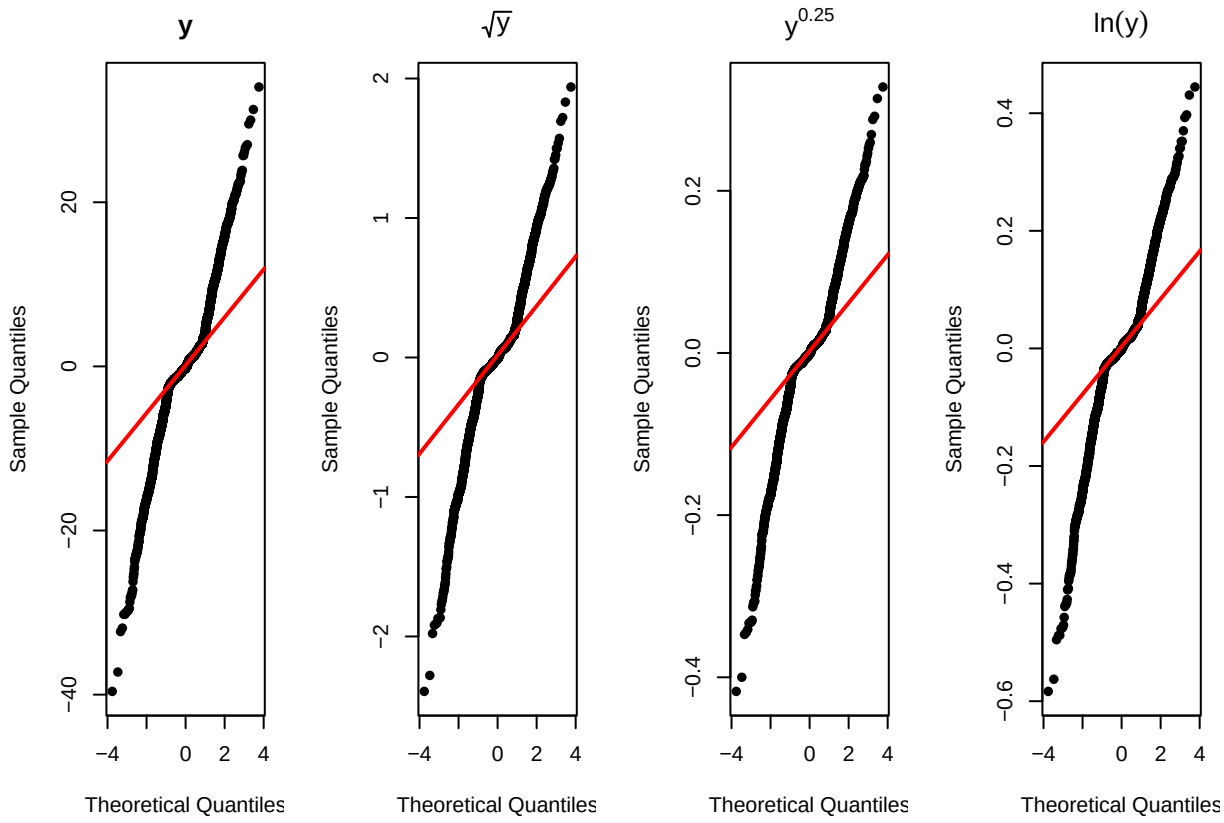
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.799 on 5804 degrees of freedom
##   (1395 observations deleted due to missingness)
## Multiple R-squared:  0.6139, Adjusted R-squared:  0.6133
## F-statistic:  1153 on 8 and 5804 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = trans_y1 ~ x1 + x2 + x5 + x7 + x12 + x14 + x17 +
##     x19, data = data_numeric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39409 -0.10244 -0.00919  0.13583  1.93935
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.157e+00  7.184e-02  113.55   <2e-16 ***
## x1           5.663e-03  4.170e-04   13.58   <2e-16 ***
## x2          -2.067e-02  9.911e-04  -20.85   <2e-16 ***
## x5           3.093e-06  6.433e-08   48.09   <2e-16 ***
## x7          -1.064e-01  4.077e-03  -26.10   <2e-16 ***
## x12          1.539e-02  2.835e-04   54.27   <2e-16 ***
## x14         -1.017e-02  3.148e-04  -32.29   <2e-16 ***
## x17          3.340e-05  3.972e-06    8.41   <2e-16 ***
## x19          1.353e-02  1.048e-03   12.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4035 on 5804 degrees of freedom
##   (1395 observations deleted due to missingness)
## Multiple R-squared:  0.608,  Adjusted R-squared:  0.6074
## F-statistic:  1125 on 8 and 5804 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = trans_y2 ~ x1 + x2 + x5 + x7 + x12 + x14 + x17 +
##     x19, data = data_numeric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41742 -0.01756 -0.00059  0.02243  0.32804
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.851e+00  1.244e-02 229.136  < 2e-16 ***
## x1           9.614e-04  7.223e-05  13.310  < 2e-16 ***
## x2          -3.494e-03  1.717e-04 -20.354  < 2e-16 ***
## x5           5.445e-07  1.114e-08  48.868  < 2e-16 ***
## x7          -1.853e-02  7.061e-04 -26.239  < 2e-16 ***
## x12          2.665e-03  4.911e-05  54.264  < 2e-16 ***
## x14         -1.735e-03  5.453e-05 -31.827  < 2e-16 ***
## x17          5.578e-06  6.879e-07   8.108  6.2e-16 ***
```

```
## x19          2.304e-03  1.815e-04  12.693  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06989 on 5804 degrees of freedom
##   (1395 observations deleted due to missingness)
## Multiple R-squared:  0.6039, Adjusted R-squared:  0.6034
## F-statistic:  1106 on 8 and 5804 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = trans_y3 ~ x1 + x2 + x5 + x7 + x12 + x14 + x17 +
##     x19, data = data_numeric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58336 -0.02410 -0.00081  0.03053  0.44463
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.183e+00  1.730e-02 241.786  < 2e-16 ***
## x1           1.309e-03  1.004e-04  13.036  < 2e-16 ***
## x2          -4.724e-03  2.387e-04 -19.791  < 2e-16 ***
## x5           7.681e-07  1.549e-08  49.576  < 2e-16 ***
## x7          -2.588e-02  9.819e-04 -26.362  < 2e-16 ***
## x12          3.696e-03  6.829e-05  54.120  < 2e-16 ***
## x14         -2.376e-03  7.582e-05 -31.332  < 2e-16 ***
## x17          7.449e-06  9.566e-07   7.787 8.07e-15 ***
## x19          3.147e-03  2.524e-04  12.467  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09719 on 5804 degrees of freedom
##   (1395 observations deleted due to missingness)
## Multiple R-squared:  0.5992, Adjusted R-squared:  0.5986
## F-statistic:  1085 on 8 and 5804 DF,  p-value: < 2.2e-16
```



2021

```
##
## Call:
## lm(formula = y2 ~ x1 + x5 + x7 + x12 + x14 + x16 + x17 + x20,
##     data = data_numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.113  -2.290   0.274   2.861  44.329
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.384e+01  1.680e+00   8.241  < 2e-16 ***
## x1           2.826e-01  9.669e-03  29.228  < 2e-16 ***
## x5           8.643e-05  1.331e-06  64.956  < 2e-16 ***
## x7          -2.952e+00  8.366e-02 -35.285  < 2e-16 ***
## x12          4.823e-01  5.802e-03  83.128  < 2e-16 ***
## x14         -1.817e-01  7.250e-03 -25.060  < 2e-16 ***
## x16          6.068e-07  9.642e-08   6.293 3.33e-10 ***
## x17          6.531e-04  8.420e-05   7.757 1.01e-14 ***
## x20          5.530e+00  3.276e-01  16.881  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

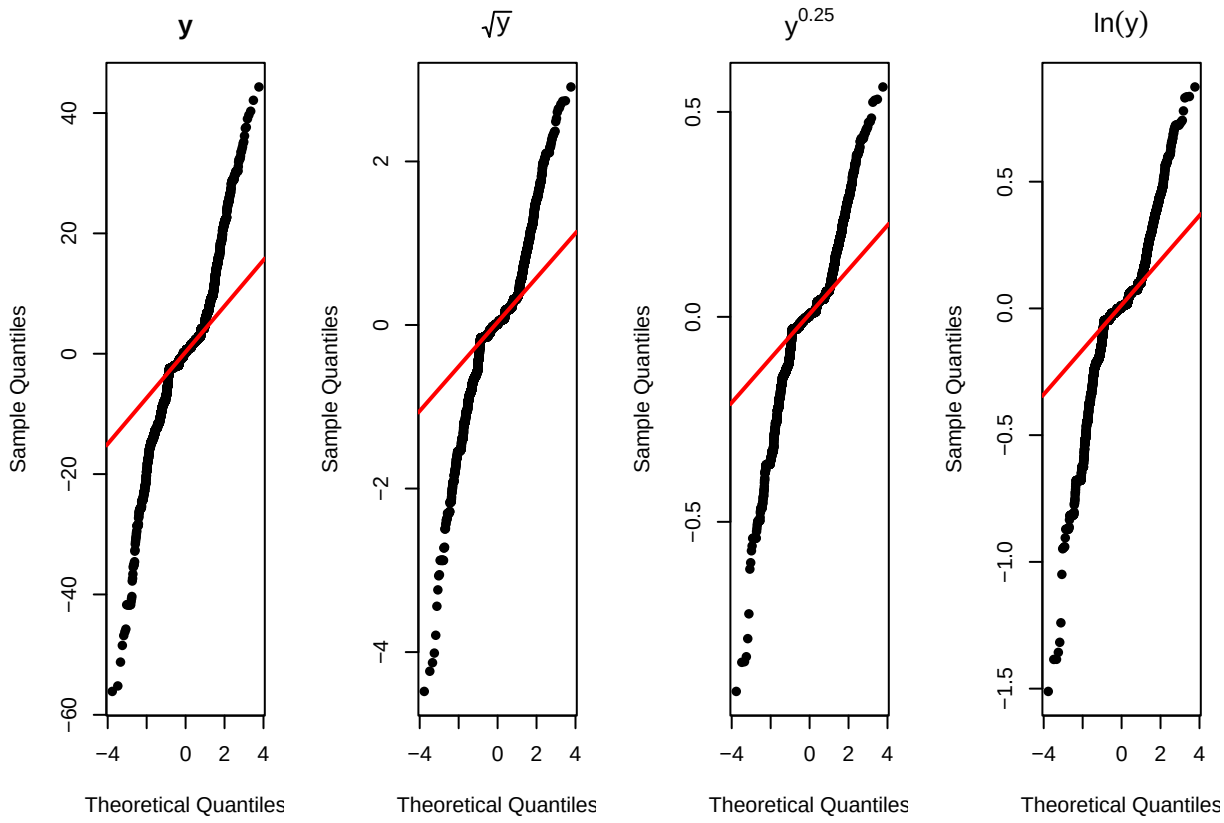
```
## Residual standard error: 8.798 on 6029 degrees of freedom
##   (1170 observations deleted due to missingness)
## Multiple R-squared:  0.7595, Adjusted R-squared:  0.7592
## F-statistic:  2381 on 8 and 6029 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = trans_y1_21 ~ x1 + x5 + x7 + x12 + x14 + x16 + x17 +
##     x20, data = data_numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4808 -0.1482  0.0065  0.2182  2.9106
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.367e+00  1.238e-01  35.289  < 2e-16 ***
## x1           1.937e-02  7.124e-04  27.186  < 2e-16 ***
## x5           6.844e-06  9.804e-08  69.806  < 2e-16 ***
## x7          -2.667e-01  6.164e-03 -43.269  < 2e-16 ***
## x12          3.456e-02  4.275e-04  80.841  < 2e-16 ***
## x14         -1.248e-02  5.342e-04 -23.361  < 2e-16 ***
## x16          6.999e-08  7.104e-09   9.851  < 2e-16 ***
## x17          3.759e-05  6.204e-06   6.060 1.44e-09 ***
## x20          2.674e-01  2.414e-02  11.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6482 on 6029 degrees of freedom
##   (1170 observations deleted due to missingness)
## Multiple R-squared:  0.7464, Adjusted R-squared:  0.746
## F-statistic:  2218 on 8 and 6029 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = trans_y2_21 ~ x1 + x5 + x7 + x12 + x14 + x16 + x17 +
##     x20, data = data_numeric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91372 -0.02947  0.00698  0.04343  0.56079
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.129e+00  2.468e-02  86.272  < 2e-16 ***
## x1           3.651e-03  1.421e-04  25.701  < 2e-16 ***
## x5           1.381e-06  1.955e-08  70.639  < 2e-16 ***
## x7          -5.711e-02  1.229e-03 -46.455  < 2e-16 ***
## x12          6.628e-03  8.525e-05  77.752  < 2e-16 ***
## x14         -2.371e-03  1.065e-04 -22.253  < 2e-16 ***
## x16          1.625e-08  1.417e-09  11.466  < 2e-16 ***
## x17          6.333e-06  1.237e-06   5.119 3.17e-07 ***
## x20          3.793e-02  4.814e-03   7.881 3.83e-15 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1293 on 6029 degrees of freedom
##   (1170 observations deleted due to missingness)
## Multiple R-squared:  0.7329, Adjusted R-squared:  0.7325
## F-statistic:  2068 on 8 and 6029 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = trans_y3_21 ~ x1 + x5 + x7 + x12 + x14 + x16 + x17 +
##     x20, data = data_numeric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51103 -0.04604  0.00955  0.07295  0.87363
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.073e+00  4.044e-02  75.979  < 2e-16 ***
## x1           5.581e-03  2.328e-04  23.972  < 2e-16 ***
## x5           2.252e-06  3.204e-08  70.282  < 2e-16 ***
## x7          -9.840e-02  2.014e-03 -48.849  < 2e-16 ***
## x12          1.027e-02  1.397e-04  73.543  < 2e-16 ***
## x14         -3.670e-03  1.746e-04 -21.023  < 2e-16 ***
## x16          2.988e-08  2.322e-09  12.870  < 2e-16 ***
## x17          8.480e-06  2.027e-06   4.183 2.92e-05 ***
## x20          3.702e-02  7.888e-03   4.693 2.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2118 on 6029 degrees of freedom
##   (1170 observations deleted due to missingness)
## Multiple R-squared:  0.7143, Adjusted R-squared:  0.7139
## F-statistic:  1884 on 8 and 6029 DF,  p-value: < 2.2e-16
```



h. Influential Points

```
## named numeric(0)
```

```
## named numeric(0)
```

VIF

```
## Loading required package: carData
```

```
##
## Attaching package: 'carData'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     Salaries
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
##       x1       x5       x7      x12      x13      x14      x19      x20
## 3.293897 2.439821 2.363011 2.107256 2.713253 3.035327 1.370109 2.371384
```

```
##       x1       x5       x7      x12      x14      x16      x17      x20
## 2.844827 1.993930 1.824783 1.660125 2.828268 1.240032 1.096278 1.966544
```

7. Residual Plot

```
# Residual Plot
# plot(data_numeric$y2, resid(best_model_2021), pch=16, col="blue")
# abline(0, 0, col = "red", lwd = 3)
# plot(fitted(best_model_2021), resid(best_model_2021), pch=16, col="blue", ylab=bquote(paste("e")))
# abline(0, 0, col = "red", lwd = 3)
```

Category

1. Urban & Rural

```
urban = data[data$Urban.Rural == "Urban", ]
rural = data[data$Urban.Rural == "Rural", ]
```

Testing if mean of Urban and Rural Math Achievement Rates are equal

$$H_0 : \mu_{Urban} - \mu_{Rural} = 0$$
$$H_0 : \mu_{Urban} - \mu_{Rural} > 1$$
$$p - value = 0.006737 < \alpha = 0.05 \rightarrow Reject\ H_0$$

```
mean(urban$All.Students.Math.Achievement)
```

```
## [1] 69.73803
```

```
mean(rural$All.Students.Math.Achievement)
```

```
## [1] 70.41292
```

```
t.test(urban$All.Students.Math.Achievement, rural$All.Students.Math.Achievement,
       mu=1, alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  urban$All.Students.Math.Achievement and rural$All.Students.Math.Achievement
## t = -4.7544, df = 7016, p-value = 1
## alternative hypothesis: true difference in means is greater than 1
## 95 percent confidence interval:
##  -1.254423       Inf
## sample estimates:
## mean of x mean of y
##  69.73803  70.41292
```

2. Race

Testing if the difference in mean of White and Black Math Achievement Rates is greater than 13

$$H_0 : \mu_{White} - \mu_{Black} = 0$$
$$H_0 : \mu_{White} - \mu_{Black} > 13$$
$$p - value = 0.004886 < \alpha = 0.05 \rightarrow Reject\ H_0$$

```
mean(data$White.Math.Achievement)
```

```
## [1] 76.35675
```

```
mean(data$Black.Math.Achievement)
```

```
## [1] 50.80237
```

```
t.test(data$White.Math.Achievement, data$Black.Math.Achievement,
       mu=13, alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  data$White.Math.Achievement and data$Black.Math.Achievement
## t = 32.899, df = 12987, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 13
## 95 percent confidence interval:
##  24.92665      Inf
## sample estimates:
## mean of x mean of y
##  76.35675  50.80237
```

```
mean(urban$White.Percentage)
```

```
## [1] 33.77357
```

```
mean(rural$White.Percentage)
```

```
## [1] 41.59337
```

```
mean(urban$Black.Percentage)
```

```
## [1] 39.033
```

```
mean(rural$Black.Percentage)
```

```
## [1] 6.823172
```

3. Economy

```r
# 100% Econ Disadv Percentage
Econ_Dia_100 = data[data$Econ.Disadvantaged.Percentage == '100', ]
Econ_Dia_100_urban = Econ_Dia_100[Econ_Dia_100$Urban.Rural == "Urban",]
Econ_Dia_100_rural = Econ_Dia_100[Econ_Dia_100$Urban.Rural == "Rural",]
# 2019
c(mean(Econ_Dia_100_urban$All.Students.Math.Achievement),
  mean(Econ_Dia_100_rural$All.Students.Math.Achievement))
```

```
## [1] 61.91958 58.68513
```

```r
# 2021
c(mean(Econ_Dia_100_urban$X2021.All.Students.Math.Achievement),
  mean(Econ_Dia_100_rural$X2021.All.Students.Math.Achievement))
```

```
## [1] 47.39643 47.00577
```

$$H_0 : \mu_{Rural\ EconDis} - \mu_{Urban\ EconDis} = 0$$
$$H_0 : \mu_{Rural\ EconDis} - \mu_{Urban\ EconDis} > 15$$
$$p{-}value = 0.04061 < \alpha = 0.05 \rightarrow Reject\ H_0$$

```r
mean(urban$Econ.Disadvantaged.Percentage)
```

```
## [1] 71.63802
```

```r
mean(rural$Econ.Disadvantaged.Percentage)
```

```
## [1] 78.77879
```

```r
t.test(rural$Econ.Disadvantaged.Percentage, urban$Econ.Disadvantaged.Percentage ,
       mu=15, alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  rural$Econ.Disadvantaged.Percentage and urban$Econ.Disadvantaged.Percentage
## t = -13.182, df = 7039.5, p-value = 1
## alternative hypothesis: true difference in means is greater than 15
## 95 percent confidence interval:
##  6.15999     Inf
## sample estimates:
## mean of x mean of y
##  78.77879  71.63802
```

4. Teacher Certificates

$$H_0 : \mu_{Urban\ Certificates} - \mu_{Rural\ Certificates} = 0$$
$$H_0 : \mu_{Urban\ Certificates} - \mu_{Rural\ Certificates} > 10$$
$$p{-}value = 0.001039 < \alpha = 0.05 \rightarrow Reject\ H_0$$

```r
# Number of total certificates at school level
mean(urban$Total)
```

```
## [1] 131.9646
```

```r
mean(rural$Total)
```

```
## [1] 215.9752
```

```r
t.test(urban$Total, rural$Total,
       mu=10, alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  urban$Total and rural$Total
## t = -48.587, df = 3986.5, p-value = 1
## alternative hypothesis: true difference in means is greater than 10
## 95 percent confidence interval:
##  -87.19397       Inf
## sample estimates:
## mean of x mean of y
##  131.9646  215.9752
```

Reference

[1] Mervosh, Sarah. "The Pandemic Erased Two Decades of Progress in Math and Reading." The New York Times, The New York Times, 1 Sept. 2022, https://www.nytimes.com/2022/09/01/us/national-test-scores-math-reading- pandemic.html?smid=nytcore-ios-share&referringSource=articleShare.

[2] Stern, Paul. "The Pandemic Worsened Racial Achievement Gaps. Making up the Difference Won't Be Easy." CT Mirror, 23 May 2022, https://ctmirror.org/2022/05/22/the-pandemic-worsened-racial-achievement-gaps-making-up-the-difference-wont-be-easy/.

[3] Georgia Department of Education. CCRPI Reports. Retrieved from https://www.gadoe.org/CCRPI/Pages/default.aspx

[4] The Governor's Office of Student Achievement. Downloadable Dataset. Retrieved from https://gosa.georgia.gov/dashboards-data-report-card/downloadable-data