DATA621 Final Project Proposal

Andrew Bowen, Glen Davis, Josh Forster, Shoshana Farber, Charles Ugiagbe

2023-10-14

Introduction

Our dataset comes from OpenData's - School Quality Report for NYC high schools between 2013 and 2014. Our final project will likely include more up-to-date educational information, and potentially geographic data to augment our analysis.

First, let's read in our source CSV file. This is posted in our GitHub repository as well in interest of reproduceability.

```
df <- read.csv("https://raw.githubusercontent.com/andrewbowen19/businessAnalyticsDataMiningDATA621/abowe
```

There's several predictor variables of interest to us. The NYC School Quality Report from the Department of Education included ratings for schools in the following categories. We'd like to see how effective these ratings are at predicting student success:

- *Quality Review Rating*
- *Achievement Rating*
- *Environment Rating*
- *College and Career Readiness Rating*

In addition, it'd be interesting to see if these ratings could be replaced by proxy variables? Doing so could save the Department of Education (DOE) time in not assigning ratings when they could have similar impact by knowing certain values for a school.

Our response variable will be the average student's SAT Score at a given school. SAT Scores are an imperfect metric given their correlation with other socioeconomic factors, but for our purposes can serve as an imperfect benchmark to measure academic performance.

Main research question: Do these DOE-ratings accurately predict whether a school will foster high academic performance in students, and are there other proxy variables that can be used to more accurately predict academic performance (measured in SAT scores)

Data Cleaning

```
# Renaming some dataframe variables
df$math_score_8 <- df$Average.Grade.8.Math.Proficiency
df$english_score_8 <- df$Average.Grade.8.English.Proficiency
df$avg_sat_score <- df$Average.SAT.Score
```

EDA

First, let's plot the counts of different rating combinations by school. This should give us a sense if schools with one rating for a given bucket tend to have similar ratings for other categories (i.e., schools with high *Quality Review Rating* values tend to have high *Achievement Ratings* as well).

```
# Group schools by ratings
ratings <- df %>% group_by(Achievement.Rating, Quality.Review.Rating) %>% summarise(count_schools=n())
```

```
## `summarise()` has grouped output by 'Achievement.Rating'. You can override
## using the `.groups` argument.
```

```
head(ratings, 5)
```

```
## # A tibble: 5 x 3
## # Groups:   Achievement.Rating [2]
##   Achievement.Rating Quality.Review.Rating count_schools
##   <chr>              <chr>                         <int>
## 1 Approaching Target Developing                       46
## 2 Approaching Target N/A                              15
## 3 Approaching Target Proficient                       74
## 4 Approaching Target Well Developed                    3
## 5 Exceeding Target   Developing                        3
```

```
ggplot(ratings, aes(x=Quality.Review.Rating,
                    y=Achievement.Rating,
                    fill=count_schools)) + geom_tile()
```

Modeling

First, let's create a basic linear model between our rating variables, and our dependent variable (Average SAT Score of a school)

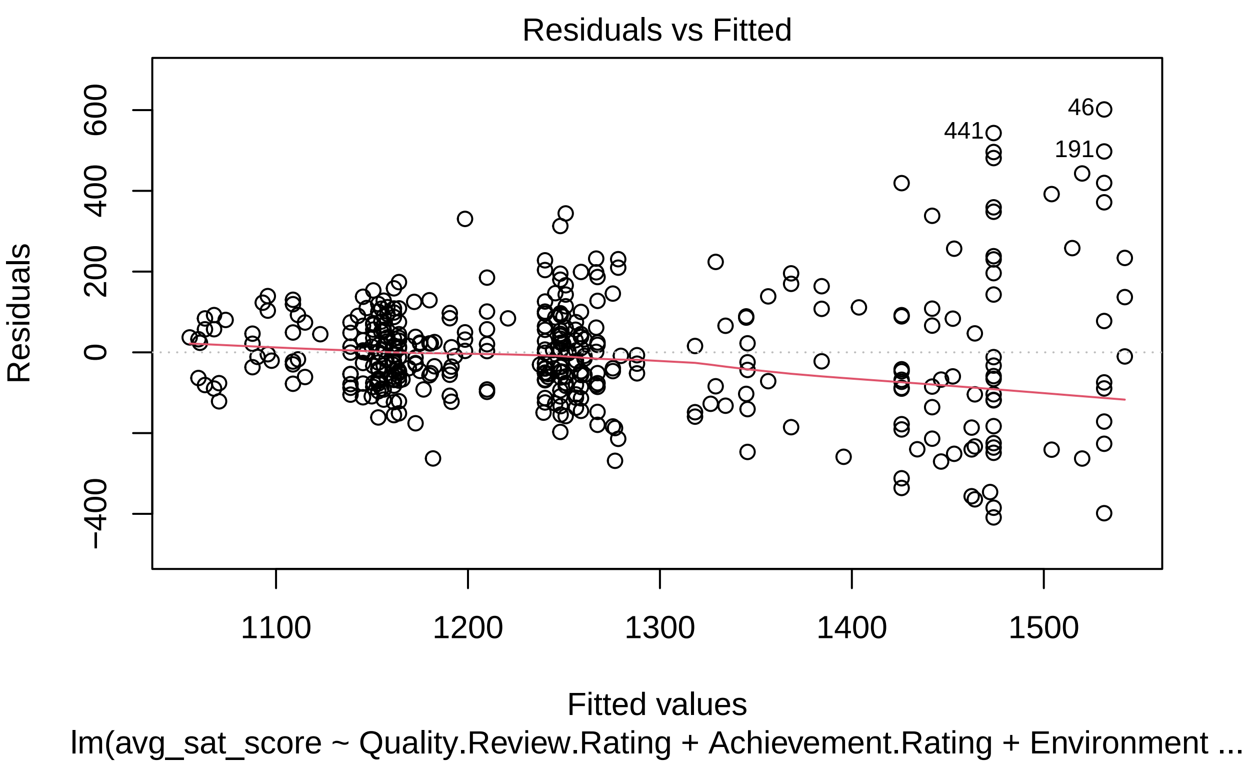
```
lm_ratings <- lm(avg_sat_score ~ Quality.Review.Rating +
  Achievement.Rating +
  Environment.Rating +
  College.and.Career.Readiness.Rating, df)
summary(lm_ratings)
```

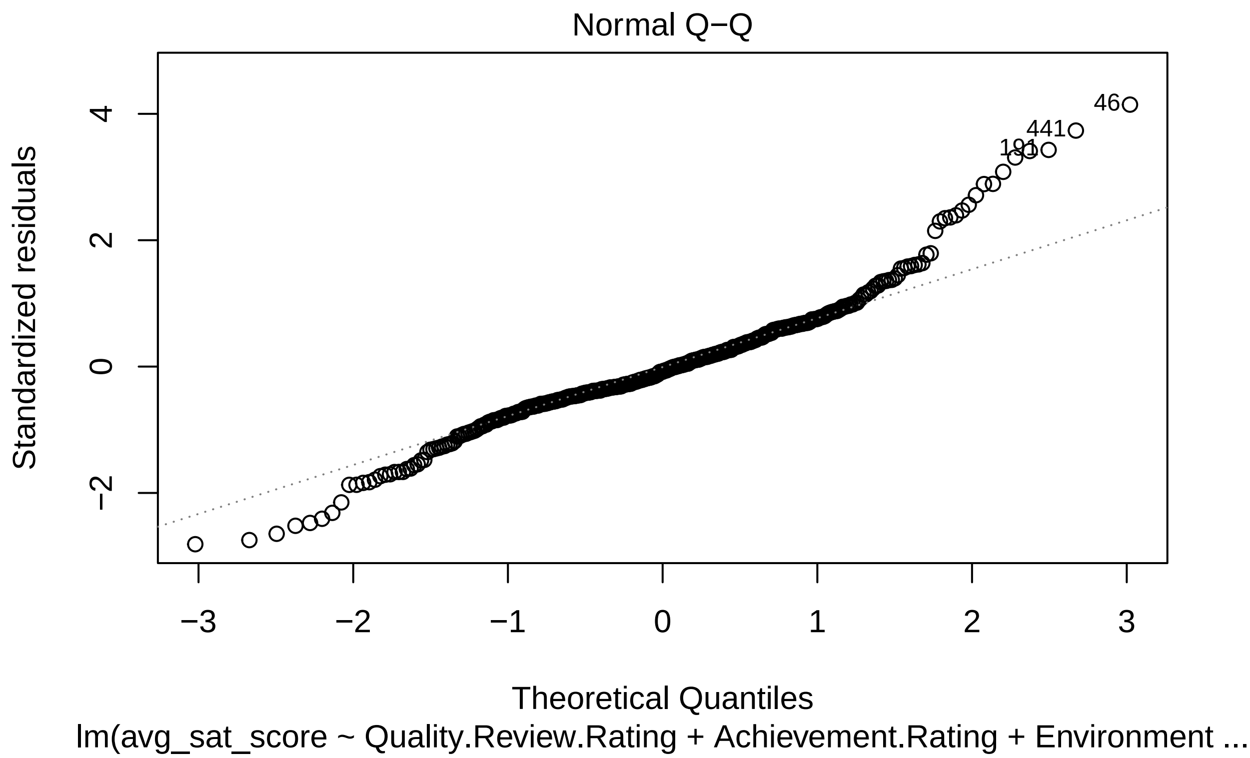
```
##
## Call:
## lm(formula = avg_sat_score ~ Quality.Review.Rating + Achievement.Rating +
##     Environment.Rating + College.and.Career.Readiness.Rating,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -408.85  -76.50  -11.34   74.20  601.57
##
## Coefficients: (2 not defined because of singularities)
##                                                        Estimate Std. Error
## (Intercept)                                            1161.433     21.007
## Quality.Review.RatingN/A                                 29.982     25.603
## Quality.Review.RatingProficient                           2.619     21.343
## Quality.Review.RatingUnderdeveloped                      66.181     93.359
## Quality.Review.RatingWell Developed                      18.524     31.231
## Achievement.RatingExceeding Target                       70.174     27.462
## Achievement.RatingMeeting Target                         -7.970     19.529
## Achievement.RatingN/A                                   -52.614     40.889
## Achievement.RatingNot Meeting Target                     -3.415     44.386
## Environment.RatingExceeding Target                      -68.346     23.762
## Environment.RatingMeeting Target                        -10.762     18.402
## Environment.RatingN/A                                        NA         NA
## Environment.RatingNot Meeting Target                     -6.577     37.735
## College.and.Career.Readiness.RatingExceeding Target     280.607     26.140
## College.and.Career.Readiness.RatingMeeting Target        94.820     18.966
## College.and.Career.Readiness.RatingN/A                       NA         NA
## College.and.Career.Readiness.RatingNot Meeting Target  -90.310     45.680
##                                                        t value Pr(>|t|)
## (Intercept)                                             55.287  < 2e-16 ***
## Quality.Review.RatingN/A                                 1.171  0.24231
## Quality.Review.RatingProficient                          0.123  0.90239
## Quality.Review.RatingUnderdeveloped                      0.709  0.47883
## Quality.Review.RatingWell Developed                      0.593  0.55344
## Achievement.RatingExceeding Target                       2.555  0.01100 *
## Achievement.RatingMeeting Target                        -0.408  0.68340
## Achievement.RatingN/A                                   -1.287  0.19896
## Achievement.RatingNot Meeting Target                    -0.077  0.93872
## Environment.RatingExceeding Target                      -2.876  0.00425 **
## Environment.RatingMeeting Target                        -0.585  0.55900
## Environment.RatingN/A                                       NA       NA
## Environment.RatingNot Meeting Target                    -0.174  0.86173
```

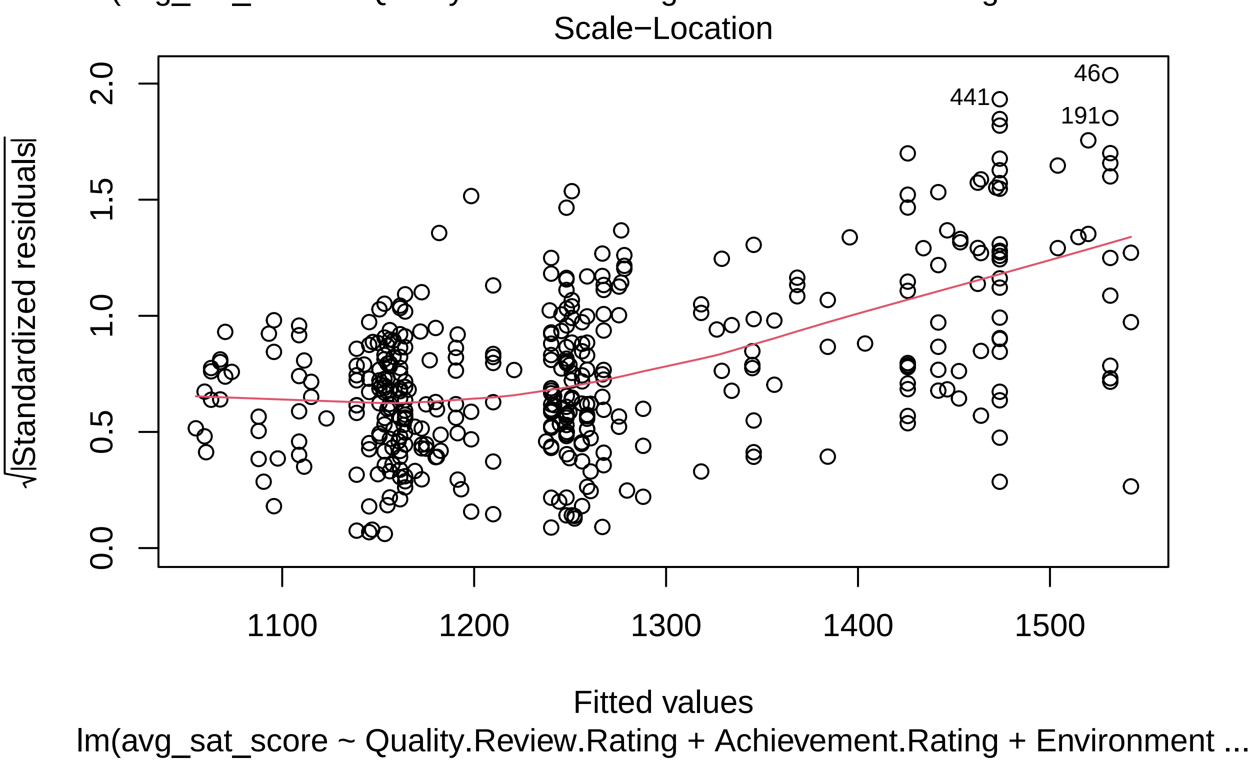
```
## College.and.Career.Readiness.RatingExceeding Target    10.735  < 2e-16 ***
## College.and.Career.Readiness.RatingMeeting Target       4.999 8.78e-07 ***
## College.and.Career.Readiness.RatingN/A                      NA       NA
## College.and.Career.Readiness.RatingNot Meeting Target  -1.977  0.04876 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.1 on 382 degrees of freedom
##   (87 observations deleted due to missingness)
## Multiple R-squared:  0.4206, Adjusted R-squared:  0.3993
## F-statistic:  19.8 on 14 and 382 DF,  p-value: < 2.2e-16
```

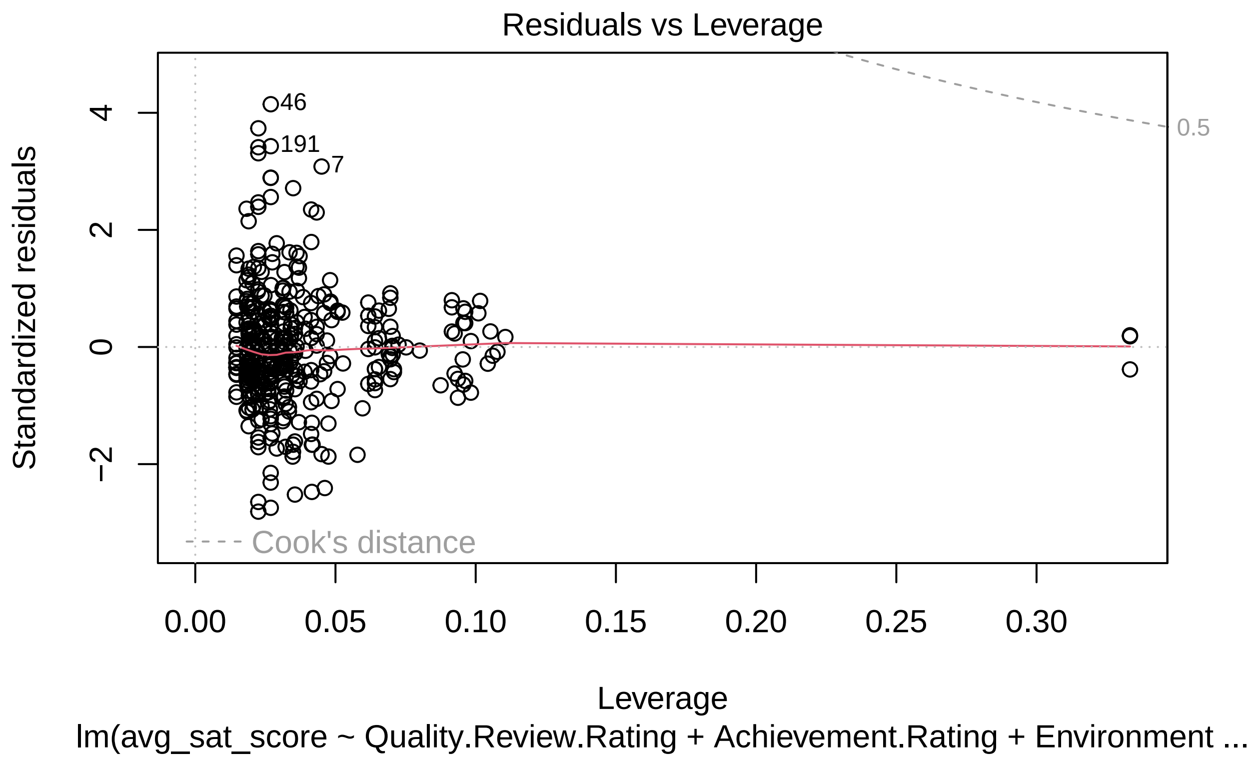
Let's plot our ratings-only model

```
plot(lm_ratings)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(avg_sat_score ~ Quality.Review.Rating + Achievement.Rating + Environment ...
46
441
191
−3
−2
−1
0
1
2
3
4


Scale−Location
√|Standardized residuals|
Fitted values
lm(avg_sat_score ~ Quality.Review.Rating + Achievement.Rating + Environment ...
46
441
191
0.0
0.5
1.0
1.5
2.0
1100
1200
1300
1400
1500



There seems to be some pattern in our residuals vs fitted, as well as some tail behavior in our QQ plot indicating this may not be an ideal fit.

Outside of the DOE ratings, we can create some very basic linear models to help us identify potential predictor variables. Two variables seem like they could potentially be useful, as past academic performance seems like it would correlate tightly with future success:

- `Average.Grade.8.Math.Proficiency`
- `Average.Grade.8.English.Proficiency`

```
lm_english_math <- lm(Average.SAT.Score~ math_score_8 + english_score_8, df)
summary(lm_english_math)
```

```
##
## Call:
## lm(formula = Average.SAT.Score ~ math_score_8 + english_score_8,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -144.146  -35.601   -1.924   33.716  187.616
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       141.15      17.12   8.246 2.52e-15 ***
## math_score_8      237.08      19.27  12.303  < 2e-16 ***
## english_score_8   228.96      19.28  11.873  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 54.44 on 391 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.9182, Adjusted R-squared:  0.9178
## F-statistic:  2195 on 2 and 391 DF,  p-value: < 2.2e-16
```

On the surface, our adjusted R^2 for the model based solely on past performance is significantly better than our model using the school's DOE ratings (`lm_ratings`). While it may seem like comparing apples to oranges

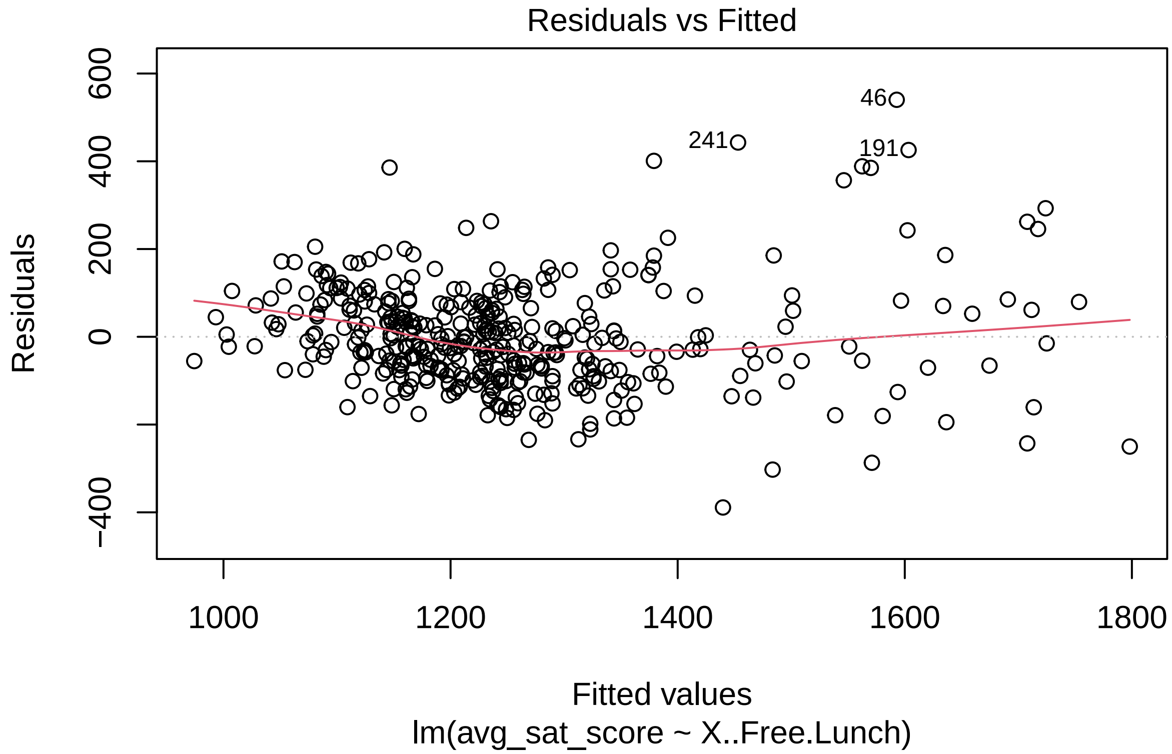
For instance, we see a jump in R^2 as well by using a model solely fit to predict average SAT score from the percentage of students who receive free lunch.

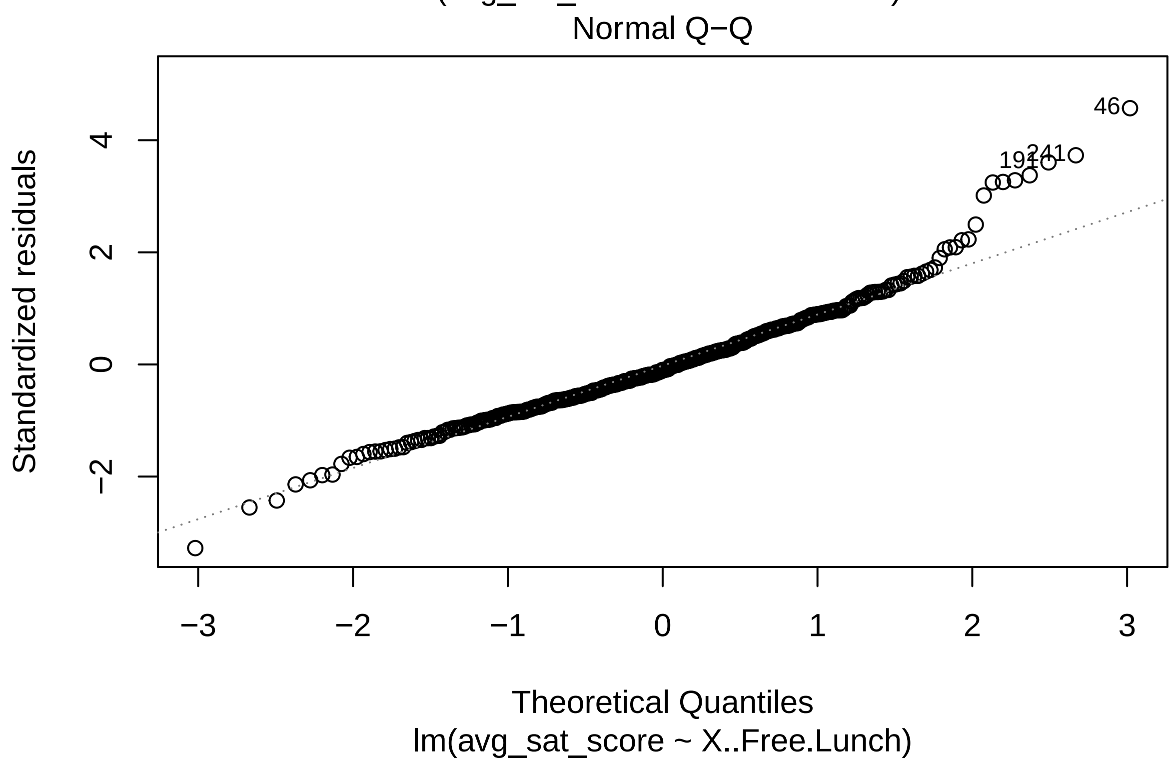
```
lm_free_lunch <- lm(avg_sat_score ~ X..Free.Lunch, df)
summary(lm_free_lunch)
```

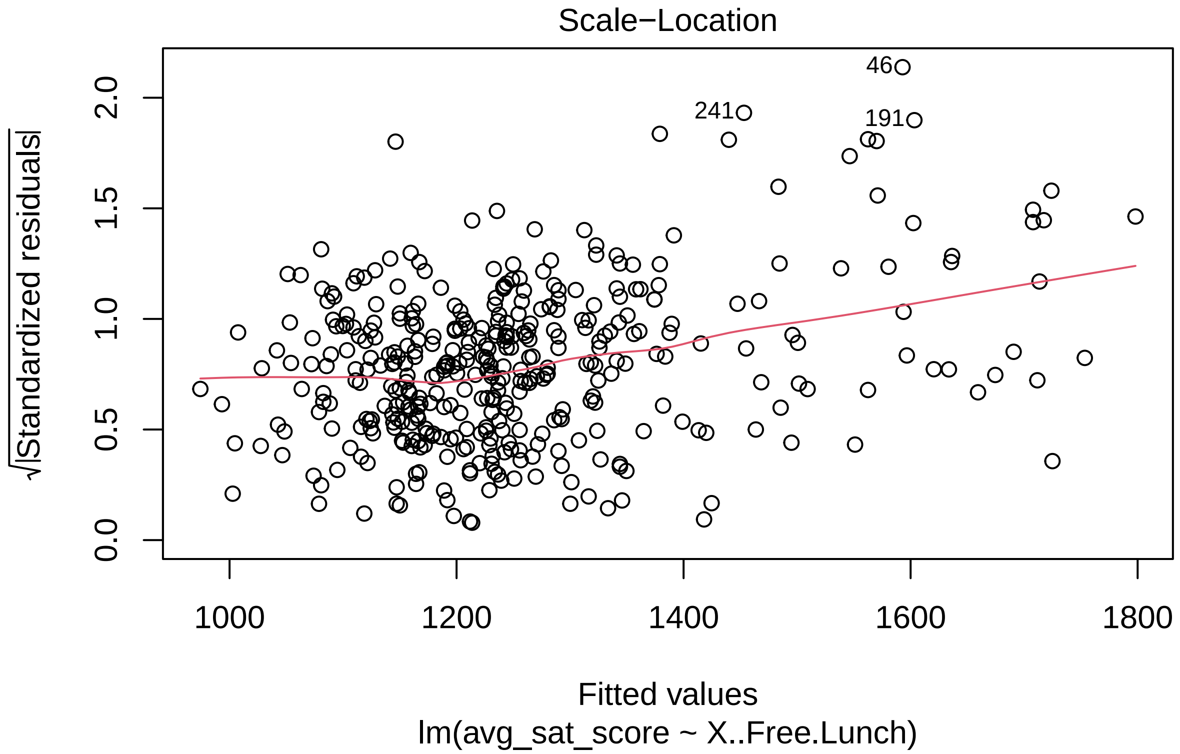
```
##
## Call:
## lm(formula = avg_sat_score ~ X..Free.Lunch, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.89  -75.78  -11.90   70.18  540.12
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1916.92      27.54   69.60   <2e-16 ***
## X..Free.Lunch  -950.25      38.57  -24.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 119.1 on 392 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.6066
## F-statistic:   607 on 1 and 392 DF,  p-value: < 2.2e-16
```

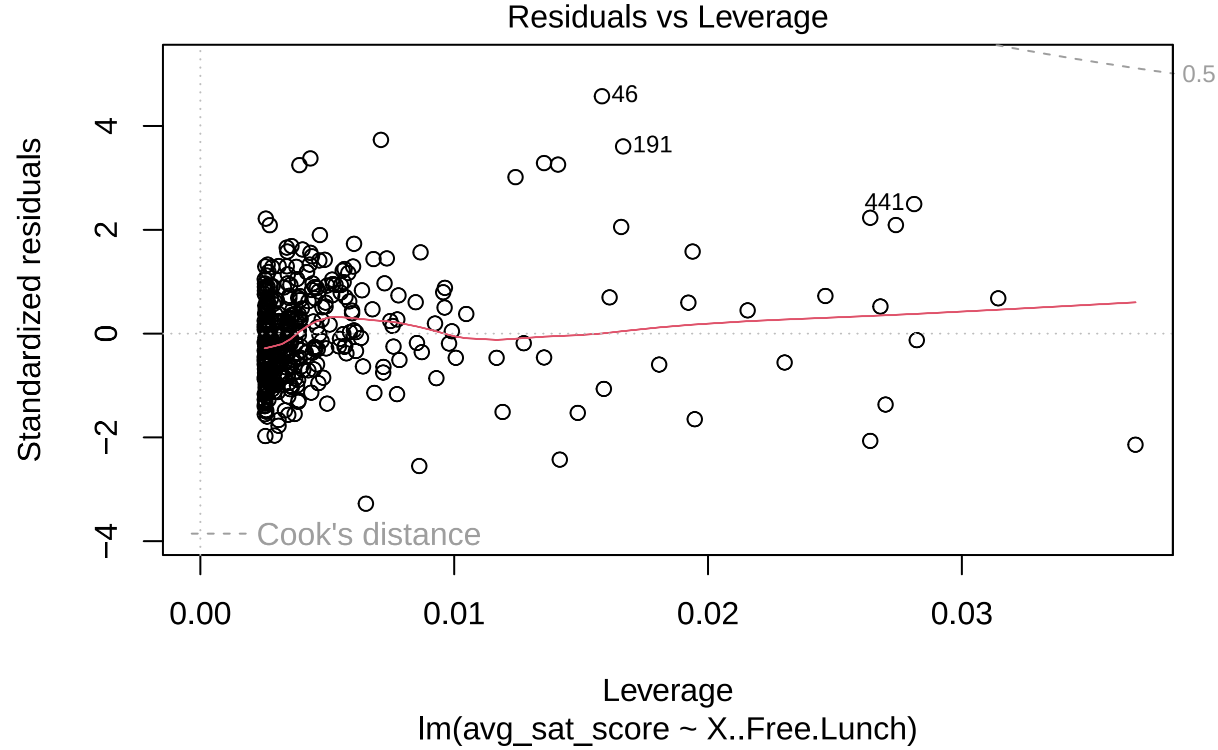
We can plot this free lunch model as well:

```
plot(lm_free_lunch)
```


Residuals vs Fitted
Residuals
Fitted values
lm(avg_sat_score ~ X..Free.Lunch)
46
241
191
−400
0
200
400
600
1000
1200
1400
1600
1800


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(avg_sat_score ~ X..Free.Lunch)
46
191
241
−2
0
2
4
−3
−2
−1
0
1
2
3





This Free Lunch plot isn't ideal, but better behaved than our Rating model plot from above.

Going further in model evaluation, we can use other metrics of interest when evaluating a model: Root mean squared error (RMSE - calculated via the `modelr` package) or AIC, which measures model performance with an added penalty for overly complex models (more input params).

```r
# Calculate RMSE and AIC for both models
rmse_fl <- rmse(lm_free_lunch, df)
rmse_ratings <- rmse(lm_ratings, df)
print(glue("Free Lunch Model RMSE: {rmse_fl}"))
```

```
## Free Lunch Model RMSE: 118.795653099892
```

```r
print(glue("Ratings Model RMSE: {rmse_ratings}"))
```

```
## Ratings Model RMSE: 144.276661079985
```

```r
aic_fl <- AIC(lm_free_lunch)
aic_ratings <- AIC(lm_ratings)
print(glue("Free Lunch Model AIC: {aic_fl}"))
```

```
## Free Lunch Model AIC: 4888.71855932368
```

```r
print(glue("Ratings Model AIC: {aic_ratings}"))
```

```
## Ratings Model AIC: 5106.19297013857
```

From these simple evaluation metrics, we can see the free-lunch only model performs a bit better, however, we can likely find a more optimized fit across our variable space in the source dataset.

Further Work

Below are some additional modeling points to consider for the final project, as well as to augment our dataset for our final project:

- SAT Scores are not a perfect indicator of future academic []. Using other response variables could paint a more complete picture on the educational variables that most impact outcomes
- Include other educational outcome metrics (job placement rates on graduation, income by school) joined to our data
- Fortunately, NYC's DBN (*District-Borough Number*) system allows for easier joining to other education datasets posted on NYC Open Data
- Use a more recent dataset than 2013-2014. NYC Open Data is an excellent tool for this