Spring 2018 DATA606 Final Project - Looking at past NYC School Progress Reports

Jonathan Hernandez

April 30, 2018

Introduction

This Final project will focus on looking at historical data for NYC school

progress reports for public schools throughout 2006 - 2007.

The research question I would like to address is:

Is the overall school rating predictive of various categorical scores as well as

predictive of Borough of the school, grades and school level?

I will attempt to use multiple linear regression to see what variables affect

the overall score of a school progress report

I care for this kind of information as I feel that many parents should be aware

and understand the progress and overall performance of the schools they send their

children. Others not only as parents can benefit from this and recommend schools

who are not making enough progress and can reach out and try to make a difference.

Also this dataset and doing analysis may be possibly used to predict the scores

of NYC public schools for other years ahead.

Data Acquisition

```
# Load some packages beforehand
if (!require(plyr)) install.packages('plyr')
if (!require(dplyr)) install.packages('dplyr')
if (!require(ggplot2)) install.packages('ggplot2')
```

Load the dataset

```
school_progress_report_scores <- read.csv("2006-2007_School_Progress_Report.csv")
```

Data

- Data Collection: Data were collected from the DOE (Department of Education) and

  is freely available to the public.

  https://data.cityofnewyork.us/Education/2006-2007-School-Progress-Report/weg5-33pj
  (https://data.cityofnewyork.us/Education/2006-2007-School-Progress-Report/weg5-33pj)

- Cases: There are 1262 cases in this dataset

```
nrow(school_progress_report_scores)
```

```
## [1] 1262
```

- Variables: The variables that will be looked at are as follows:
  - DBN (District Borough Number) - categorical
  - SCHOOL LEVEL - categorical
  - GRADE - categorical ordinal
  - ENVIRONMENT CATEGORY SCORE - numerical continous
  - PERFORMANCE CATEGORY SCORE - numerical continous
  - PROGRESS CATEGORY SCORE - numerical continous
  - QUALITY REVIEW SCORE - categorical ordinal
  - OVERALL SCORE - numerical continous

- Type of Study: This was a observational study as the data were collected from the DOE in various NYC schools and observed the scores and grades of performance/progress.
- Scope of Inference - generalizability: The population of interest is all the public schools of NYC during 2006-2007. The findings of this analysis can be generalized to this population as we will be looking at the entire dataset and can see if we can predict a overall score of a school that was built during that time.
- Scope of Inference - causality: The data and the model can be used to show some type of relationship between the independent variables and the response variable I will show that the scores and location and type of education show a strong relationship towards the overall score NYC schools get.

lets look at the structure and summary as well as a preview of the dataset in question

```
str(school_progress_report_scores)
```

```
## 'data.frame':    1262 obs. of  14 variables:
##  $ DBN                               : Factor w/ 1226 levels "01M015","01M019",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ DISTRICT                          : int  1 1 1 1 1 1 1 1 1 ...
##  $ SCHOOL                            : Factor w/ 1224 levels "47 THE AMERICAN SIGN LANGUAGE AND ENGLISH DUAL
LAN",..: 502 513 518 558 640 643 749 808 812 821 ...
##  $ SCHOOL.SUPPORT.ORGANIZATION.NETWORK: Factor w/ 74 levels "AED1","CEIPEA1",..: 45 16 45 45 45 16 16 45 55 26
 ...
##  $ PROGRESS.REPORT.TYPE              : Factor w/ 2 levels "ESMS","HS": 1 1 1 1 1 1 1 1 1 ...
##  $ SCHOOL.LEVEL.                     : Factor w/ 5 levels "Elementary School",..: 1 1 1 3 1 1 1 1 3 ...
##  $ PEER.INDEX.                       : Factor w/ 812 levels " 10.20 "," 10.44 ",..: 659 425 536 572 443 572 3
93 492 553 605 ...
##  $ GRADE                             : Factor w/ 6 levels "A","B","C","D",..: 2 2 2 3 3 3 2 2 2 2 ...
##  $ OVERALL.SCORE                     : Factor w/ 1113 levels "-0.447","100",..: 513 483 517 367 302 403 480 5
71 728 631 ...
##  $ ENVIRONMENT.CATEGORY.SCORE        : Factor w/ 599 levels "-0.008","-0.019",..: 93 301 354 207 44 349 546 1
42 198 372 ...
##  $ PERFORMANCE.CATEGORY.SCORE        : Factor w/ 602 levels "0.006","-0.025",..: 46 219 418 127 261 330 384 2
97 123 195 ...
##  $ PROGRESS.CATEGORY.SCORE           : Factor w/ 589 levels "-0.021","0.03",..: 477 304 200 323 272 205 124 3
57 518 384 ...
##  $ ADDITIONAL.CREDIT                 : Factor w/ 21 levels "0","0.75","1",..: 9 6 6 2 1 1 8 6 2 6 ...
##  $ QUALITY.REVIEW.SCORE              : Factor w/ 4 levels "","Proficient",..: 3 2 4 2 2 4 2 4 3 2 ...
```

Data Cleaning

- The custom-built function below will clean up the dataset removing rows containing a ‘Under Review’ as well as convert the DBN values into borough names and convert the scores to numerical values for computation and plotting.

(A DBN consist of the following format:)

```
source("Clean_progress_report.R")
progress_report_cleaned <-
  clean_school_progress_report(school_progress_report_scores)

summary(progress_report_cleaned)
```

```
##           Borough               School_Level          Grade
##  Bronx        :290   Elementary School:585   A           :292
##  Brooklyn     :390   High School      :237   B           :487
##  Manhattan    :257   K-8 School       :122   C           :327
##  Queens       :263   Middle School    :294   D           :102
##  Staten Island: 61   Transfer School  : 23   F           : 53
##                                               Under Review:  0
##  Overall_Score     Environment_Score Performance_Score Progress_Score
##  Min.   : -0.447   Min.   :-0.1410   Min.   :-0.1090   Min.   :-0.263
##  1st Qu.: 44.100   1st Qu.: 0.3620   1st Qu.: 0.4260   1st Qu.: 0.385
##  Median : 54.040   Median : 0.4970   Median : 0.5520   Median : 0.506
##  Mean   : 54.079   Mean   : 0.5011   Mean   : 0.5526   Mean   : 0.502
##  3rd Qu.: 63.300   3rd Qu.: 0.6350   3rd Qu.: 0.6750   3rd Qu.: 0.618
##  Max.   :117.420   Max.   : 1.1330   Max.   : 1.1350   Max.   : 1.281
##      Quality_Review_Score
##                :  1
##  Proficient    :703
##  Undeveloped   : 95
##  Well-Developed:448
##  NA's          : 14
##
```

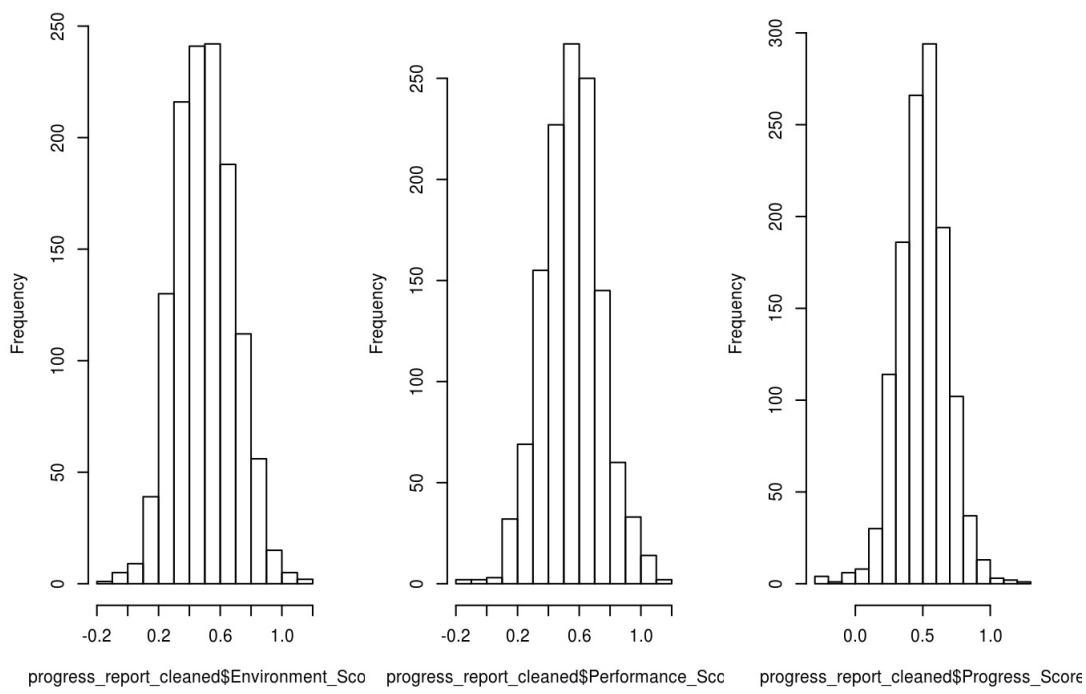
Exploratory Data Analysis

Let’s look at some of the data visually to understand what kind of distribution and behavior

they follow and create some plots.

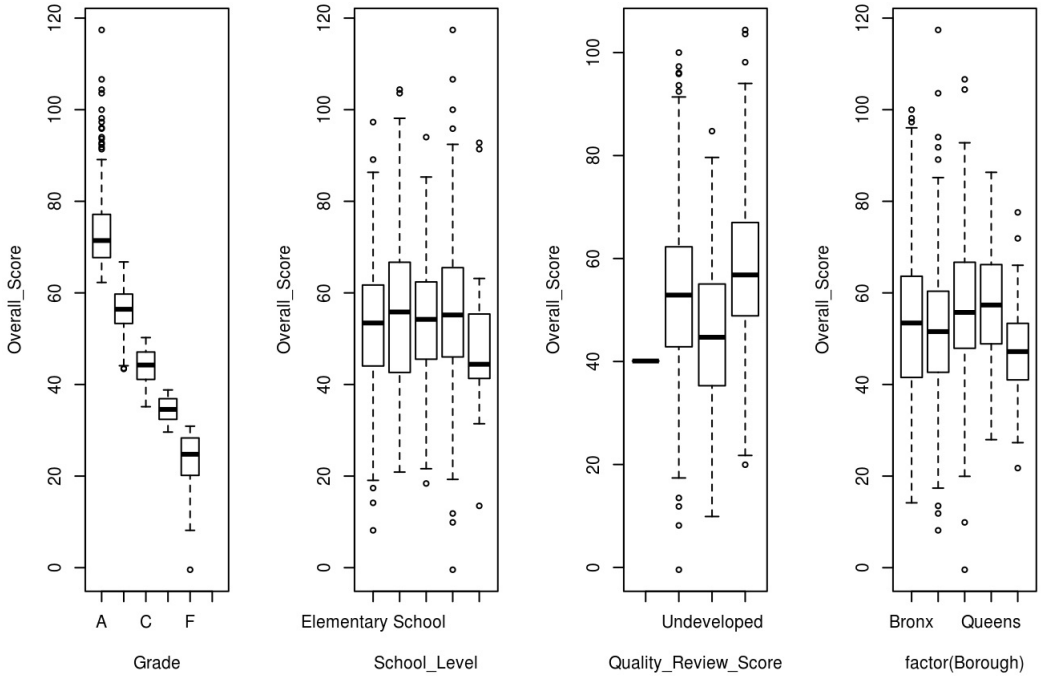
- Some Histograms

```
par(mfrow=c(1,3))
hist(progress_report_cleaned$Environment_Score)
hist(progress_report_cleaned$Performance_Score)
hist(progress_report_cleaned$Progress_Score)
```



- Some scatterplots of overall score vs the category scores

```
par(mfrow=c(1,4))
with(progress_report_cleaned, plot(Overall_Score ~ Grade))
with(progress_report_cleaned, plot(Overall_Score ~ School_Level))
with(progress_report_cleaned, plot(Overall_Score ~ Quality_Review_Score))
with(progress_report_cleaned, plot(Overall_Score ~ factor(Borough)))
```




```
# find correlations on the response variable and the explanatory variabls for the
# overall score and the environment, performance and progress score
summary(progress_report_cleaned$Overall_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -0.447  44.100  54.040  54.079  63.300 117.420
```

```
summary(progress_report_cleaned$Progress_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -0.263   0.385   0.506   0.502   0.618   1.281
```

```
summary(progress_report_cleaned$Environment_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1410  0.3620  0.4970  0.5011  0.6350  1.1330
```

```
summary(progress_report_cleaned$Performance_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1090  0.4260  0.5520  0.5526  0.6750  1.1350
```

```
with(progress_report_cleaned, cor(Overall_Score, Environment_Score))
```

```
## [1] 0.4961702
```

```
with(progress_report_cleaned, cor(Overall_Score, Progress_Score))
```

```
## [1] 0.84446
```

```
with(progress_report_cleaned, cor(Overall_Score, Performance_Score))
```

```
## [1] 0.6029361
```

It seems like the scores are normally distributed and the distribution of

overall scores and borough are each nearly normal as well as most of the grades.

This suggests that the scores and the other variables have a strong effect

Statistical Data Analysis

To see which variables are strong predictors of overall grade score, I will use the concept of multiple linear regression to create a linear equation best fitting the data.

we will add all the variables to the equation as follows:

$$\hat{overallscore} = \beta_0 + \beta_1 * \hat{borough} + \beta_2 * \hat{schoollevel} + \beta_3 * \hat{grade} + \beta_4 * \hat{enviornment} + \beta_5 * \hat{performance} + \beta_6 * \hat{progress} + \beta_7 * \hat{qualityreview}$$

and by looking at such characteristics like correlation and p-value and the beta value parameter esitmates, we can find an equation that best fits the model we are trying to accomplish while minimizing any residuals.

To get our linear model we'll use the lm() function to get coefficent esitmates as well as R-squared and p-values.

```
# fitting our linear model
lm_overall_score <- lm(Overall_Score ~ Borough + School_Level +
                         Grade + Environment_Score +
                         Performance_Score +
                         Progress_Score + Quality_Review_Score,
                       data = progress_report_cleaned)

# look at the summary of our model
summary(lm_overall_score)
```

```
##
## Call:
## lm(formula = Overall_Score ~ Borough + School_Level + Grade +
##     Environment_Score + Performance_Score + Progress_Score +
##     Quality_Review_Score, data = progress_report_cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.146 -1.290 -0.238  1.080  9.659
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       11.61226    2.19311   5.295 1.41e-07
## BoroughBrooklyn                   -0.54072    0.16166  -3.345 0.000848
## BoroughManhattan                  -0.80194    0.18027  -4.449 9.42e-06
## BoroughQueens                     -0.63809    0.18019  -3.541 0.000413
## BoroughStaten Island              -0.29747    0.29786  -0.999 0.318129
## School_LevelHigh School            0.81194    0.16310   4.978 7.33e-07
## School_LevelK-8 School             0.73323    0.20973   3.496 0.000489
## School_LevelMiddle School          0.85982    0.15163   5.671 1.77e-08
## School_LevelTransfer School       -0.69934    0.44538  -1.570 0.116625
## GradeB                            -3.35530    0.22487 -14.921  < 2e-16
## GradeC                            -5.60791    0.33799 -16.592  < 2e-16
## GradeD                            -7.15577    0.46061 -15.535  < 2e-16
## GradeF                            -8.37566    0.60825 -13.770  < 2e-16
## Environment_Score                 12.76324    0.41115  31.043  < 2e-16
## Performance_Score                 25.68389    0.48089  53.409  < 2e-16
## Progress_Score                    51.42630    0.67326  76.384  < 2e-16
## Quality_Review_ScoreProficient    -0.20759    2.06448  -0.101 0.919922
## Quality_Review_ScoreUndeveloped   -0.45210    2.07329  -0.218 0.827418
## Quality_Review_ScoreWell-Developed -0.09802    2.06799  -0.047 0.962205
##
## (Intercept)                       ***
## BoroughBrooklyn                   ***
## BoroughManhattan                  ***
## BoroughQueens                     ***
## BoroughStaten Island
## School_LevelHigh School           ***
## School_LevelK-8 School            ***
## School_LevelMiddle School         ***
## School_LevelTransfer School
## GradeB                            ***
## GradeC                            ***
## GradeD                            ***
## GradeF                            ***
## Environment_Score                 ***
## Performance_Score                 ***
## Progress_Score                    ***
## Quality_Review_ScoreProficient
## Quality_Review_ScoreUndeveloped
## Quality_Review_ScoreWell-Developed
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.048 on 1228 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.9802, Adjusted R-squared:  0.9799
## F-statistic:  3374 on 18 and 1228 DF,  p-value: < 2.2e-16
```

While the R-Squared value is very close to 1, we can remove predictors that have a large p-value as a large p-value would indicate that the null hypothesis should not be rejected and the predictor can be removed from the model.

Looking at the summary of the linear model, the Quality Review Score can be removed from the model as they have large p-values and may not have no relationship in the overall score.

Using the updated model below we get new R-squared, p-values, coefficents etc:

```
lm_overall_score_updated <- lm(Overall_Score ~ Borough + School_Level +
                                 Grade + Environment_Score + Performance_Score +
                                 Progress_Score, data = progress_report_cleaned)
summary(lm_overall_score_updated)
```

```
##
## Call:
## lm(formula = Overall_Score ~ Borough + School_Level + Grade +
##     Environment_Score + Performance_Score + Progress_Score, data = progress_report_cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.024 -1.299 -0.252  1.066  9.706
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  10.9854     0.7566  14.520  < 2e-16 ***
## BoroughBrooklyn              -0.5095     0.1597  -3.189 0.001461 **
## BoroughManhattan             -0.7871     0.1787  -4.405 1.15e-05 ***
## BoroughQueens                -0.5967     0.1781  -3.349 0.000834 ***
## BoroughStaten Island         -0.2064     0.2912  -0.709 0.478641
## School_LevelHigh School       0.8071     0.1618   4.989 6.94e-07 ***
## School_LevelK-8 School        0.7295     0.2044   3.568 0.000373 ***
## School_LevelMiddle School     0.8431     0.1482   5.689 1.59e-08 ***
## School_LevelTransfer School  -0.7014     0.4416  -1.588 0.112497
## GradeB                       -3.2987     0.2200 -14.994  < 2e-16 ***
## GradeC                       -5.5018     0.3265 -16.851  < 2e-16 ***
## GradeD                       -7.0165     0.4438 -15.810  < 2e-16 ***
## GradeF                       -8.1709     0.5824 -14.029  < 2e-16 ***
## Environment_Score            12.9888     0.3855  33.695  < 2e-16 ***
## Performance_Score            25.8871     0.4660  55.557  < 2e-16 ***
## Progress_Score               51.6717     0.6371  81.102  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.043 on 1245 degrees of freedom
## Multiple R-squared:  0.9808, Adjusted R-squared:  0.9806
## F-statistic:  4241 on 15 and 1245 DF,  p-value: < 2.2e-16
```

Thus the new updated fitted model is:

$$\begin{aligned}\hat{overallscore} = 10.985 - 0.510 * \hat{borough}_{Brooklyn} - 0.787 * \hat{borough}_{Manhattan} - 0.597 * \hat{borough}_{Queens} \\ + 0.807 * \hat{schoollevel}_{HS} + 0.730 * \hat{schoollevel}_{Kto8} + 0.843 * \hat{schoollevel}_{MS} \\ - 3.299 * \hat{grade}_{B} - 5.502 * \hat{grade}_{C} - 7.017 * \hat{grade}_{D} - 8.170 * \hat{gradeF} \\ + 12.990 * \hat{environment} + 25.890 * \hat{performance} + 51.670 * \hat{progress}\end{aligned}$$

Note that for the Borough, school level and grade, the values are either 1 or 0

depending on the value of the categorical variables.

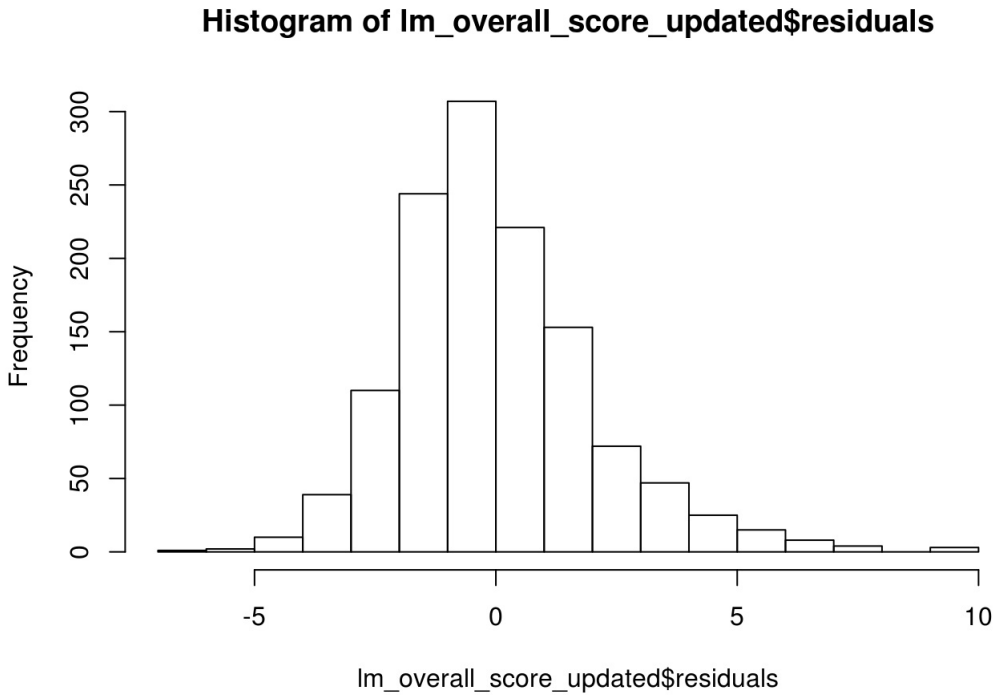
For exmaple, if a Middle School is added at the time to reside in Brooklyn and got a grade of

say B,

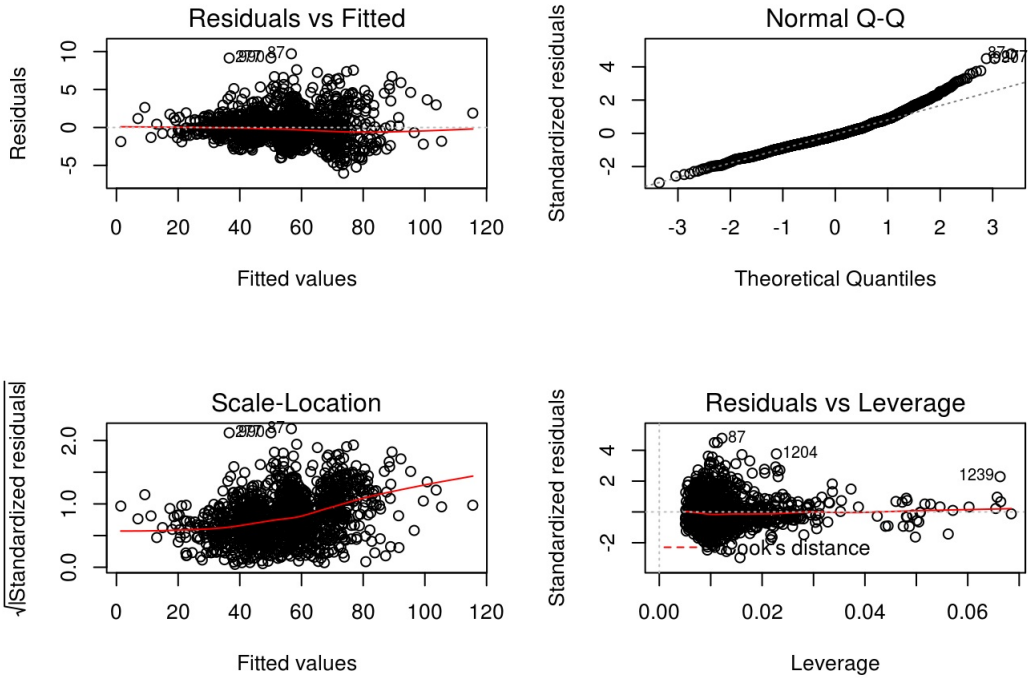
the equation would be as follows:

$$\begin{aligned}\hat{overallscore} &= 10.985 - 0.510 * (1) - 0.787 * (0) - 0.597 * (0) \\ &\quad + 0.807 * (0) + 0.730 * (0) + 0.843 * (1) \\ &\quad - 3.299 * (1) - 5.502 * (0) - 7.017 * (0) - 8.170 * (0) \\ &\quad + 12.990 * \hat{environment} + 25.890 * \hat{performance} + 51.670 * \hat{progress} \\ &= (10.985 - 0.510 + 0.843 - 3.299) + 12.990 * \hat{environment} + 25.890 * \hat{performance} + 51.670 * \hat{progress}\end{aligned}$$

```
# plot the distirbution of residuals
# if a valid model the distribution should be nearly normal (bell-shaped)
hist(lm_overall_score_updated$residuals)
```




```
#plotting a linear model gives us normal probability plots, residual plots and
# residual vs leverage plots
par(mfrow=c(2,2))
plot(lm_overall_score_updated)
```



We see the distribution of residuals is nearly normal, few outliers and the normal probability plot and leverage are fairly reasonable for this model.

Computing 95% confidence intervals for each predictor Intervals show that we are 95% confident the true parameter estimates/slopes lie within the ranges below:

Let H_0 = parameter estimates = 0 Let H_A = parameter estimates != 0

```
confint(lm_overall_score_updated)
```

```
##                                 2.5 %     97.5 %
## (Intercept)                  9.5010989 12.4696910
## BoroughBrooklyn             -0.8228776 -0.1960865
## BoroughManhattan            -1.1376935 -0.4365194
## BoroughQueens               -0.9461373 -0.2471655
## BoroughStaten Island        -0.7776265  0.3649014
## School_LevelHigh School      0.4896703  1.1244697
## School_LevelK-8 School       0.3283743  1.1305729
## School_LevelMiddle School    0.5523675  1.1339043
## School_LevelTransfer School -1.5678490  0.1650328
## GradeB                      -3.7302995 -2.8670862
## GradeC                      -6.1424019 -4.8612665
## GradeD                      -7.8871967 -6.1457892
## GradeF                      -9.3136197 -7.0282709
## Environment_Score           12.2325730 13.7450925
## Performance_Score           24.9729678 26.8012628
## Progress_Score              50.4217622 52.9216480
```

We looked at each p-value in the original model and removed predictors that had a high p-value meaning that we would fail to reject the null hypothesis and that predictor has no influence on the response variable.

Conclusion

- We saw in our model that there is a strong positive linear relationship between overall scores and explanatory variables grades, borough, quality review, type of school and various scores. What has been shown is that various scores had the lowest p-values meaning they contribute the most to the overall scores along with grades. If we simulated sample test schools whose scores and other categorical explanatory variables follow the similar distribution and mean/standard deviation as the original dataset, we would expect to get a overall score of that school that follows the distribution of that school and with good prediction.
- One idea to further test this model is to also look at a similar dataset containing NYC school progress reports for later years like 2007-2008, 2008-2009 and so on. I could then build a linear (or non-linear) model to not only accomidate for the scores/grades/location but factoring in time as well and build a model that will predict the overall scores of NYC schools for years down the road. It would be helpful for parents and staff to know which schools will have a good chance of having good progress and ones below the standard.