DATA 605 - Week 11 Discussion Post

Soumya Ghosh

November 10, 2019

Libraries

```
library(ggplot2)
library(grid)
library(gridExtra)
library(dplyr)
library(gvlma)
library(ggResidpanel)
```

Background

Critical Question

How well does the percentage of students in a school who are eligible for free or reduced-price lunch (FRPL, a common measure of poverty) explain the average critical reading SAT score of a school?

Dataset

2006 - 2012 School Demographics and Accountability Snapshot

2012 SAT Results

Findings

Schools with higher poverty (as measured by FRPL) have a lower average critical reading SAT score. However, poverty only explains about 50% of the variation in critical reading score. To make a better model, we need to account for other variables - such as school budget, racial diversity, etc.

Data Preparation

Read in and clean the SAT dataset:

```
url1 <- "C:/CUNY/Semester3 (Fall)/DATA 605/Assignments/Week11/Data/2012_SAT_Results.csv"

sat <- read.csv(url1, fill = TRUE, sep = ",") # Read in dataset
sat <- filter(sat, sat$Num.of.SAT.Test.Takers != "s") # Filter out missing values
# Coerce columns to character and numeric
sat[,1:2] <- lapply(sat[,1:2], as.character)
sat[,3:6] <- lapply(sat[,3:6], as.numeric)
names(sat) <- c("DBN", "school", "num_takers", "reading_avg", "math_avg", "writing_avg")
```

Read in and clean the Demographics dataset, filtering for the 2011-2012 school year (to match the SAT dataset).

```
url2 <- "C:/CUNY/Semester3 (Fall)/DATA 605/Assignments/Week11/Data/2006_-_2012_School_Demographics_and_

demo <- read.csv(url2, fill = TRUE, sep = ",") # Read in dataset
demo <- filter(demo, demo$schoolyear=="20112012") # Filter for the 2011-2012 school year
# Coerce columns to character and numeric
demo[,1:2] <- lapply(demo[,1:2], as.character)
demo[,3:38] <- lapply(demo[,3:38], as.numeric)
demo <- demo[, c(1, 5)] # Choose relevant columns
names(demo)[1] <- "DBN"
```

Join the two datasets together using the DBN (district-borough number) of each school, and view the dataframe.

```
sat_demo <- inner_join(sat, demo, by="DBN")
```

```
## Warning: `chr_along()` is deprecated as of rlang 0.2.0.
## This warning is displayed once per session.
```

Below is a preview if the data set -

```
head(sat_demo)
```

```
##      DBN                                        school num_takers
## 1 01M292 HENRY STREET SCHOOL FOR INTERNATIONAL STUDIES         68
## 2 01M448           UNIVERSITY NEIGHBORHOOD HIGH SCHOOL        166
## 3 01M450                    EAST SIDE COMMUNITY SCHOOL        136
## 4 01M458                     FORSYTH SATELLITE ACADEMY        135
## 5 01M509                       MARTA VALLE HIGH SCHOOL         98
## 6 01M515       LOWER EAST SIDE PREPARATORY HIGH SCHOOL         11
##   reading_avg math_avg writing_avg frl_percent
## 1          34       70          45        88.6
## 2          62       87          48        71.8
## 3          56       68          52        71.8
## 4          93       67          41        72.8
## 5          69       94          66        80.7
## 6          17      153          11        77.0
```

Here is the statistical summary of the data -

```
summary(sat_demo)
```

```
##      DBN                school            num_takers      reading_avg
##  Length:412         Length:412         Min.   :  1.00   Min.   :  1.00
##  Class :character   Class :character   1st Qu.: 68.00   1st Qu.: 47.00
##  Mode  :character   Mode  :character   Median :108.00   Median : 70.00
##                                        Mean   : 99.21   Mean   : 73.22
##                                        3rd Qu.:133.00   3rd Qu.: 94.25
##                                        Max.   :174.00   Max.   :163.00
##     math_avg       writing_avg      frl_percent
##  Min.   :  1.00   Min.   :  1.00   Min.   : 15.80
##  1st Qu.: 38.75   1st Qu.: 42.75   1st Qu.: 58.85
##  Median : 61.00   Median : 64.00   Median : 69.70
##  Mean   : 70.81   Mean   : 69.81   Mean   : 66.11
##  3rd Qu.: 98.00   3rd Qu.: 92.00   3rd Qu.: 76.80
##  Max.   :172.00   Max.   :162.00   Max.   :100.00
```

Data Modeling and Analysis

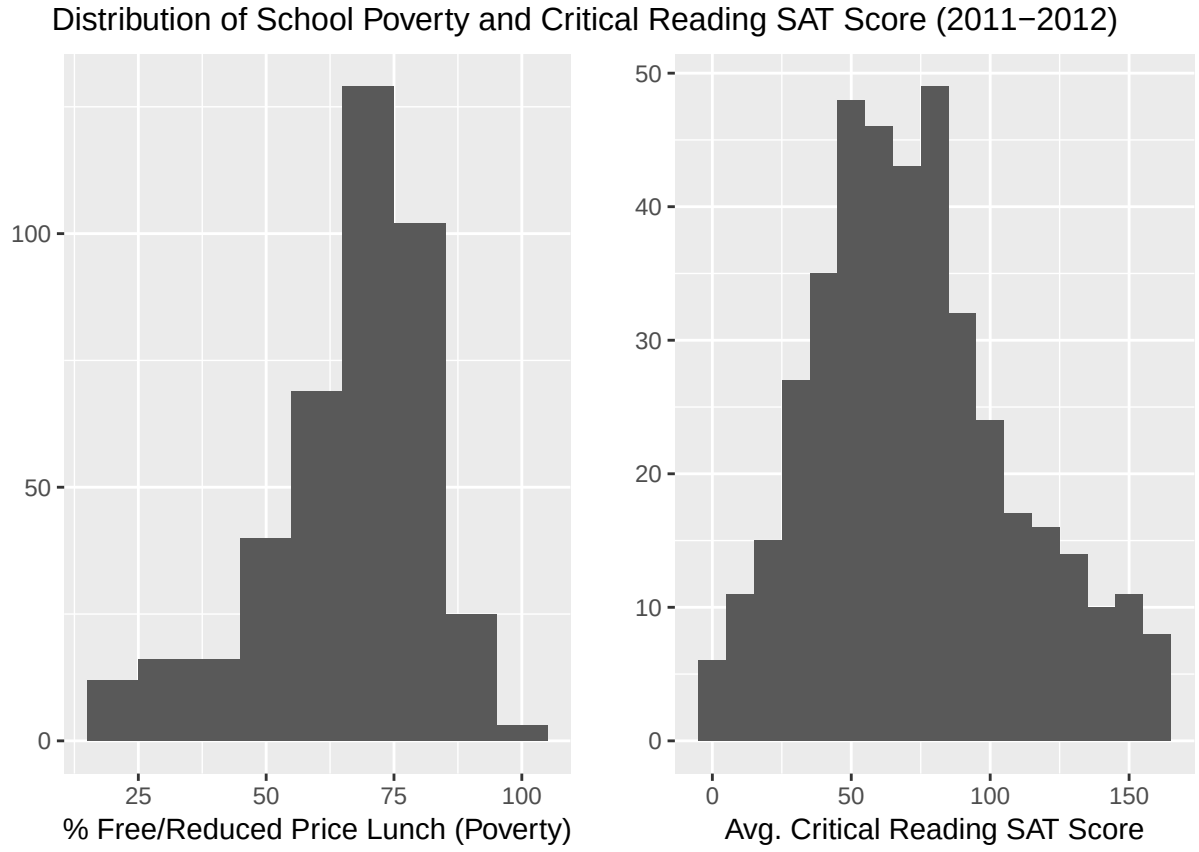
Distributions

When we look at histograms of each variable, we see they are slightly skewed from the normal distribution, especially with FRPL. This suggests bias in the data - that poverty is not evenly distributed across NYC schools. There are more schools with poverty levels between 60-75%. If the distribution was normal, we would expect to see a peak around 50%.

```
y <- sat_demo$reading_avg
x <- sat_demo$frl_percent

plot1 <- qplot(x, geom = "histogram", xlab="% Free/Reduced Price Lunch (Poverty)", binwidth=10)
plot2 <- qplot(y, geom = "histogram", xlab="Avg. Critical Reading SAT Score", binwidth=10)

grid.arrange(plot1, plot2, ncol=2, top="Distribution of School Poverty and Critical Reading SAT Score (
```



Linear Model

I modeled the relationship between FRPL and critical reading SAT score using the lm function and summarized the results below.

Slope: -1.5929. For every 1 percentage-point increase in the proportion of FRPL students in a school, there is a 1.6-point decrease in average critical reading SAT score.

Intercept: 178.5304. A school that has 0% of students eligible for FRPL would be estimated to have an average critical reading SAT score of about 179 points.

Standard error: 5.3885 (intercept) and 0.0792 (slope). These values are much smaller than the corresponding coefficients. Indicates that there is relatively little variability in the estimates of the slope and intercept.

P-value: <2e-16. This tiny value, and three significance stars, means that there is a high likelihood that FRPL is relevant in the model, and the model more accurately predicts it.

R-squared: 0.4966; adjusted R^2: 0.4954. This means that FRPL explains about 50% of the variation in critical reading SAT score. This suggests that we may need more variables than just FRPL to explain SAT score.

Degrees of freedom: 410. There were 412 observations used to generate the model.

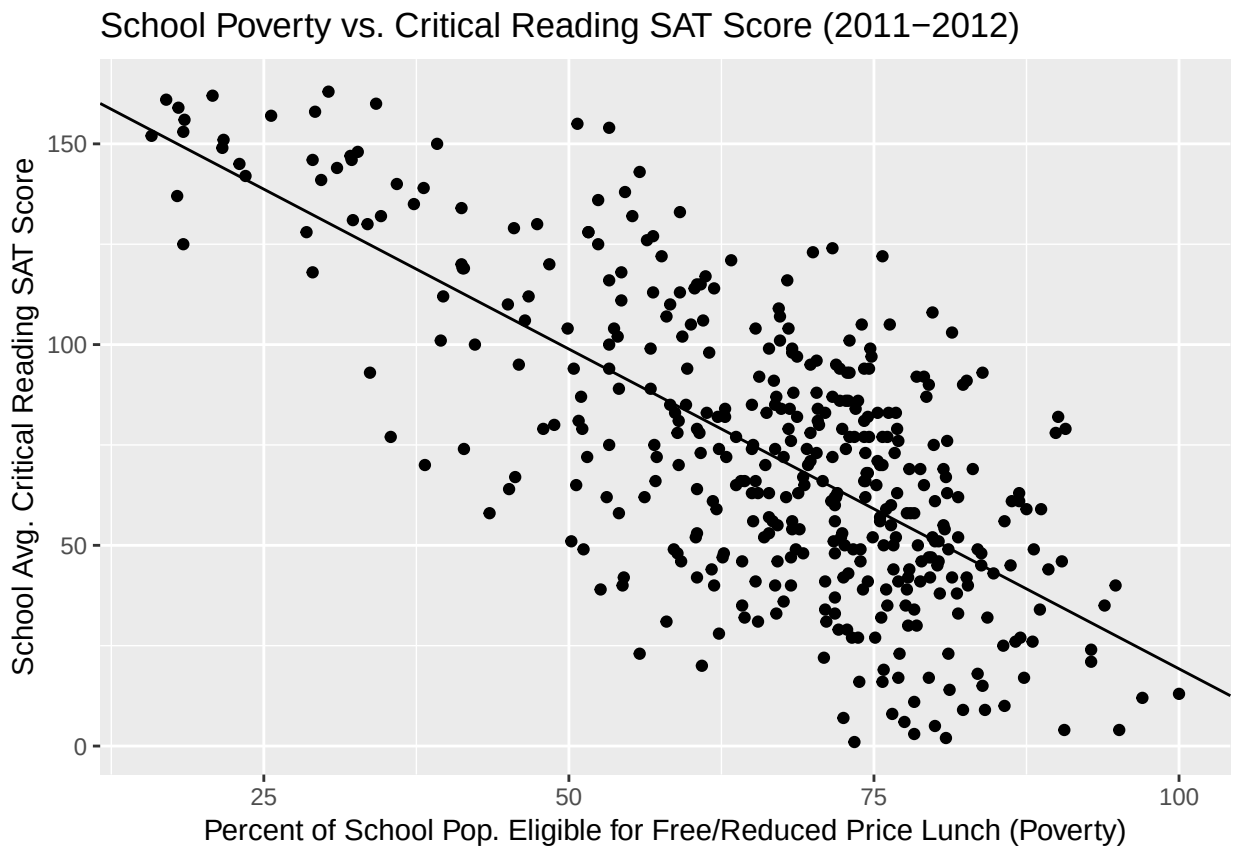
```
a <- lm(y ~ x)
summary(a)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.648 -16.781   0.844  18.789  64.050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 178.5304     5.3885   33.13   <2e-16 ***
## x            -1.5929     0.0792  -20.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.79 on 410 degrees of freedom
## Multiple R-squared:  0.4966, Adjusted R-squared:  0.4954
## F-statistic: 404.5 on 1 and 410 DF,  p-value: < 2.2e-16
```

Visualize the Data

The first step in this one-factor modeling process is to determine whether or not it looks as though a linear relationship exists between the predictor and the output value. Let's inspect through a scatter plot if there is any apparent linear relationship between the Preditor and Response variable -

```
qplot(x, y, ylab="School Avg. Critical Reading SAT Score", xlab="Percent of School Pop. Eligible for Fr
  geom_abline(intercept = a$coefficients[1], slope = a$coefficients[2])
```



Residuals

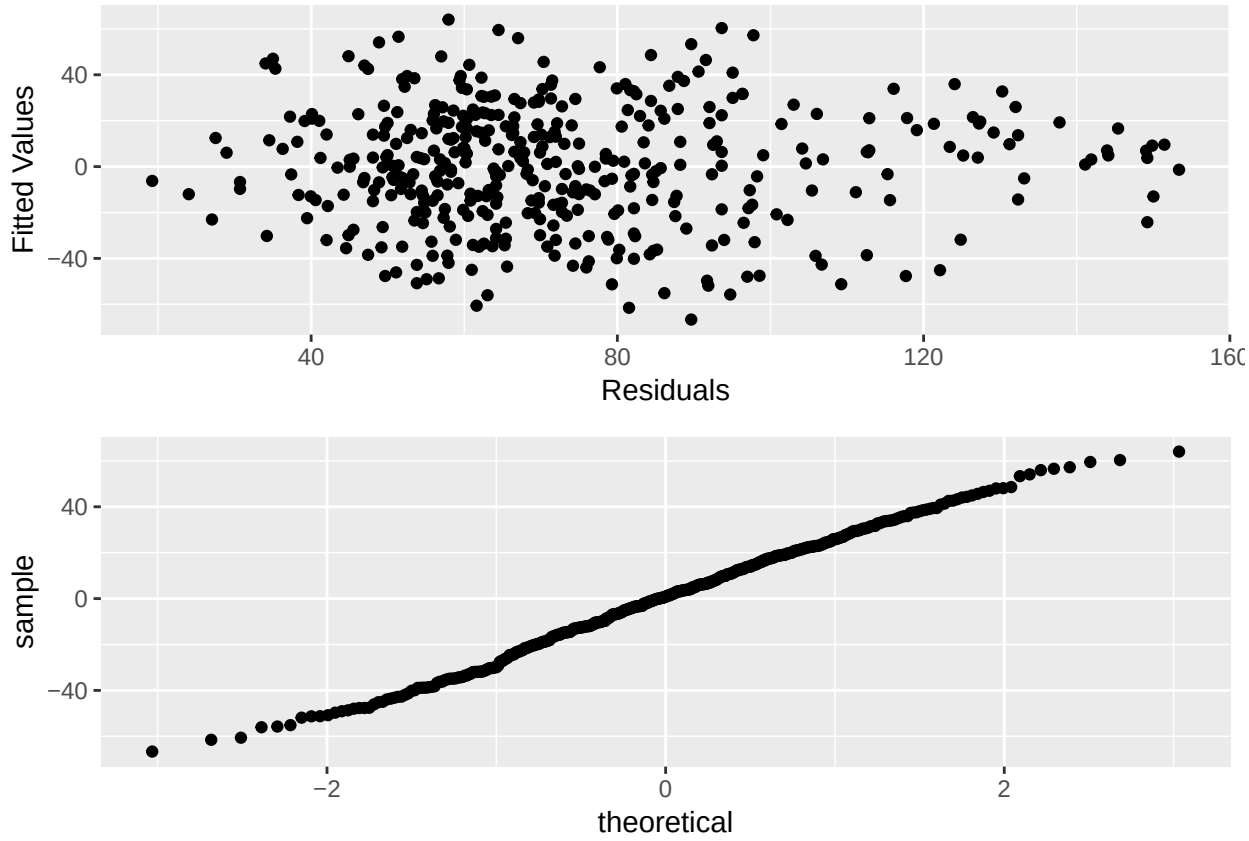
When we plot the residuals of this model, we see a cluster of points between 40 and 80, and slight skew in the ends of the quantile-quantile plot. This reinforces what we saw in the histogram, that there may be bias influencing the data, or that using FRPL alone may not be enough to explain the data.

```
plot4 <- qplot(a$fitted.values, a$residuals, ylab="Fitted Values", xlab="Residuals")

plot5 <- ggplot() + geom_qq(aes(sample = a$residuals))

grid.arrange(plot4, plot5, ncol=1, nrow=2)
```

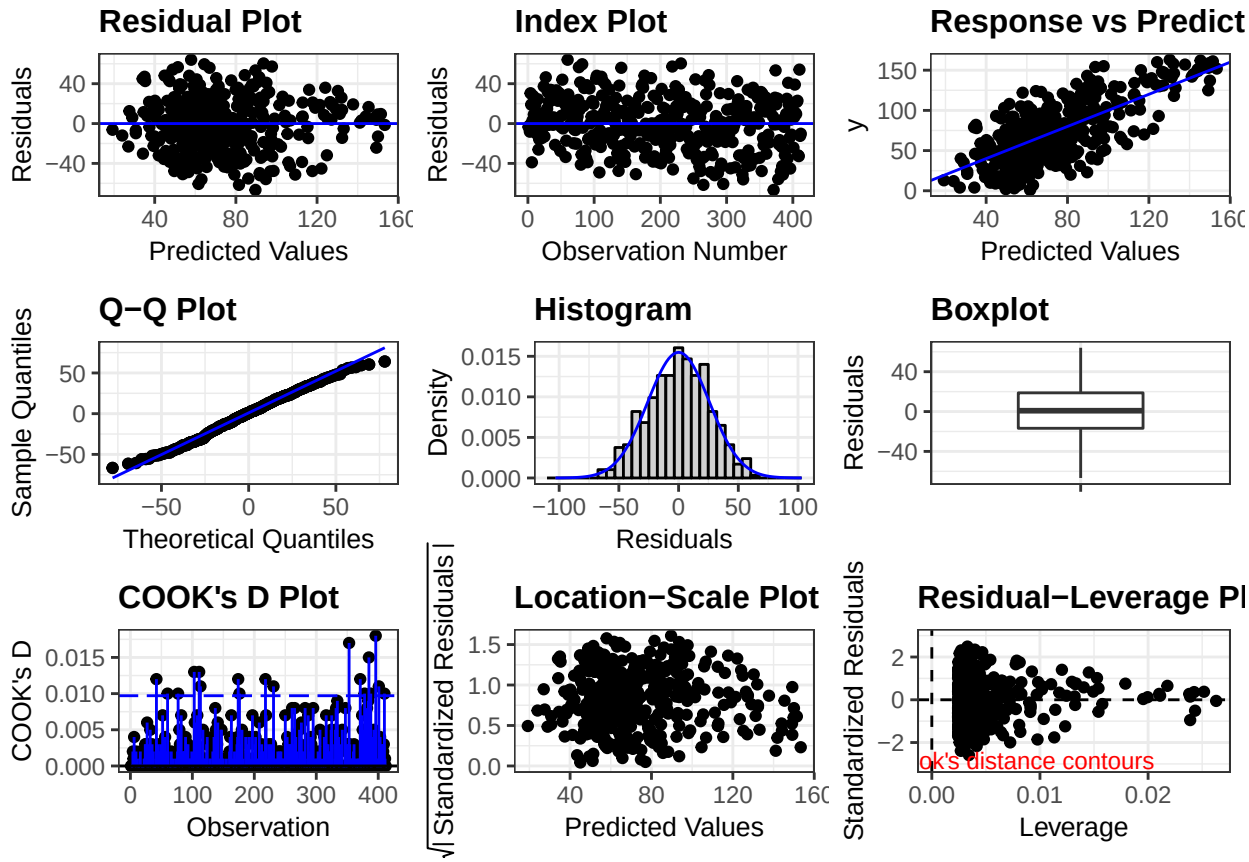
```
## Warning: `list_len()` is deprecated as of rlang 0.2.0.
## Please use `new_list()` instead.
## This warning is displayed once per session.
```



Residial Analysis Summary

Below is a summary of the Residual Analysis using the **resid_panel** function of the **ggResidpanel** package.

```
resid_panel(a, plots='all')
```



Conclusion

Based upon above Model Diagnostics, it can be concluded that the FRPL, a common measure of poverty explains the average critical reading SAT score of a school. Also based on the output of the gvlma function and the corresponding plot, we can find that the conditions for linear regression have been met.

```
gvlma(a)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     178.530       -1.593
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = a)
##
##                     Value p-value                Decision
## Global Stat        8.1109 0.08760 Assumptions acceptable.
```

```
## Skewness           0.5499 0.45837 Assumptions acceptable.
## Kurtosis           3.5254 0.06043 Assumptions acceptable.
## Link Function      1.5418 0.21435 Assumptions acceptable.
## Heteroscedasticity 2.4938 0.11430 Assumptions acceptable.
```

```
plot(gvlma(a))
```