# paper_pdf

## Question 3: Are there any group of factors that predict eligibility in the school free lunch program.

We are interested in predicting the eligibility of students for the Free Lunch Program. As we only have county-level child demographic data from the years 2009 - 2010, we will choose the Free Lunch Program eligibility data from 2009. There are lots of different families of variables that could determine eligibility. Some of these include:

- Assistance : How much access to resources do vulnerable groups in the county have? Are there resources like SNAP (Supplemental Nutritional Assistance Program) benefits or WIC (Women, Infants, Children) benefits available?
- Food source availability: How many stores and restaurants are in the county?
- Socio-economic: Does the county suffer from signficant adult and/or child poverty rates.

As the data is census level data, it does not seem to be completely randomly collected. Thus we are not aiming to make causal inferences from our analysis, rather understand the factors at play.

### Exploratory Data Analysis

Using multiple linear regression, we are interested in predicting the Percentage Eligibility of Students in the Free Lunch program (out of the total number of students attending school). The independent variables we are interested in are the following:

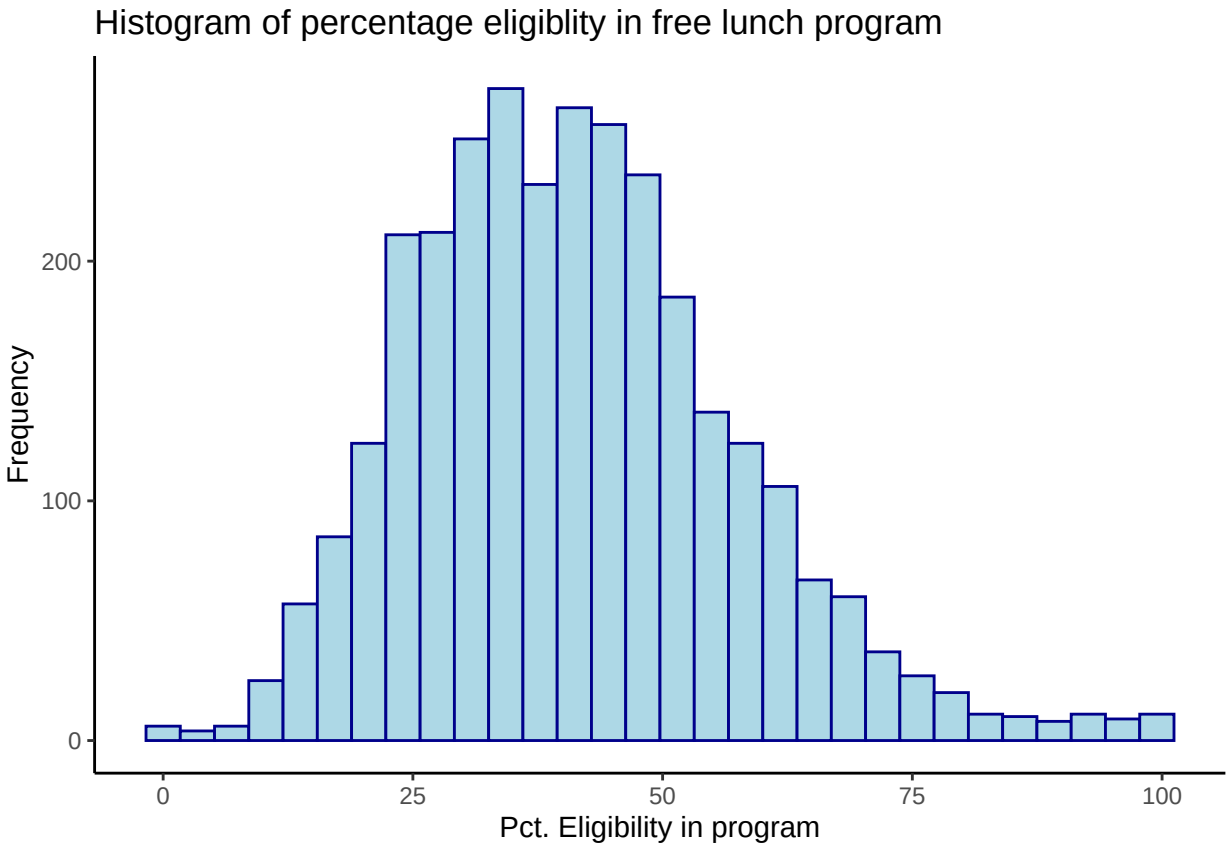
- a
- b
- c
- d
- e
- f
- g
- h
- i
- j
- k
- l

#### Data Preparation

Initially, there are 3140 observations. However, we decide to remove observations that contain null values for Percent eligibility for the Free Lunch Program. This reduces our observations to 3065. Before proceeding to model fitting, we will do some initial data exploration of our variables.

**Exploratory Data Analysis**

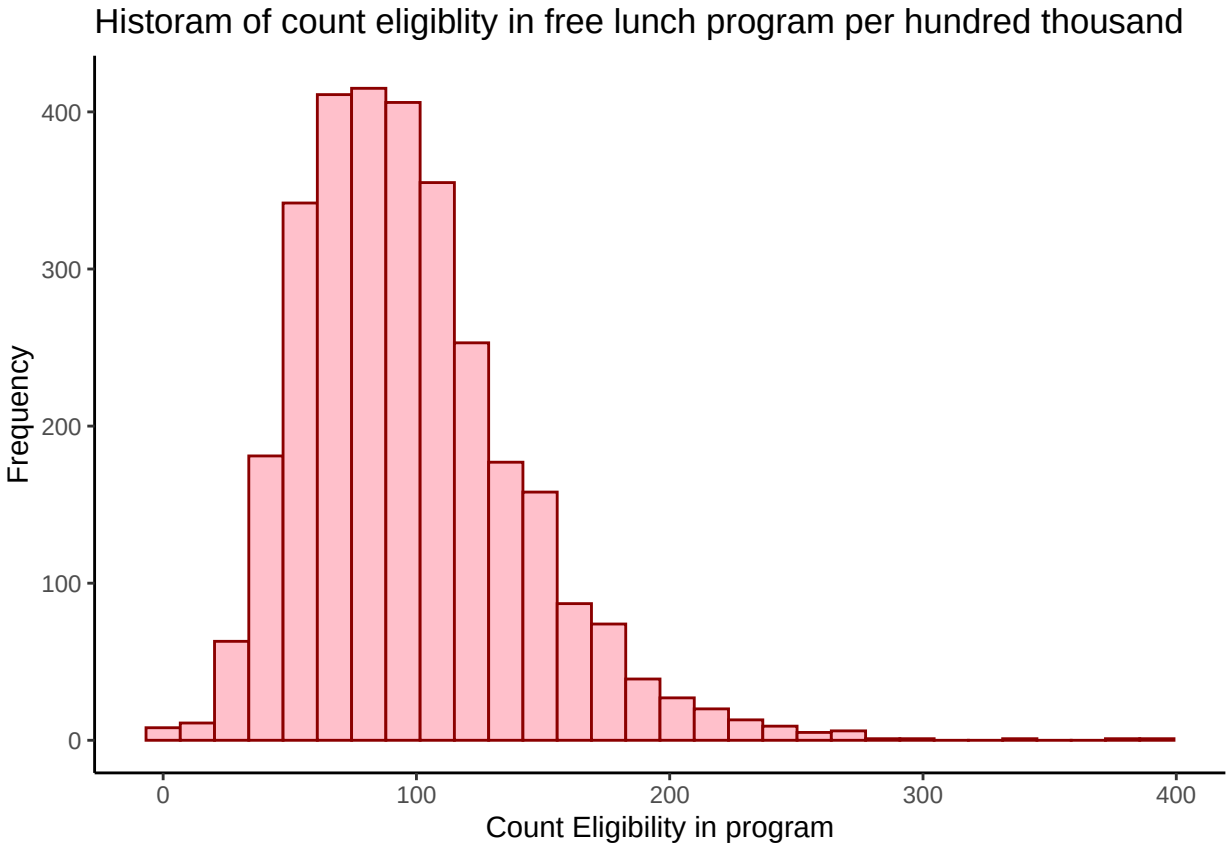
Below we have the histogram of percent eligibiltiy in the free lunch program. While it does not follow a Gaussian distribution exactly, it seems to follow it roughly enough that we can attempt the regression.




We can see that the histogram is not exactly normal but not so far off that we cannot attempt the regression.

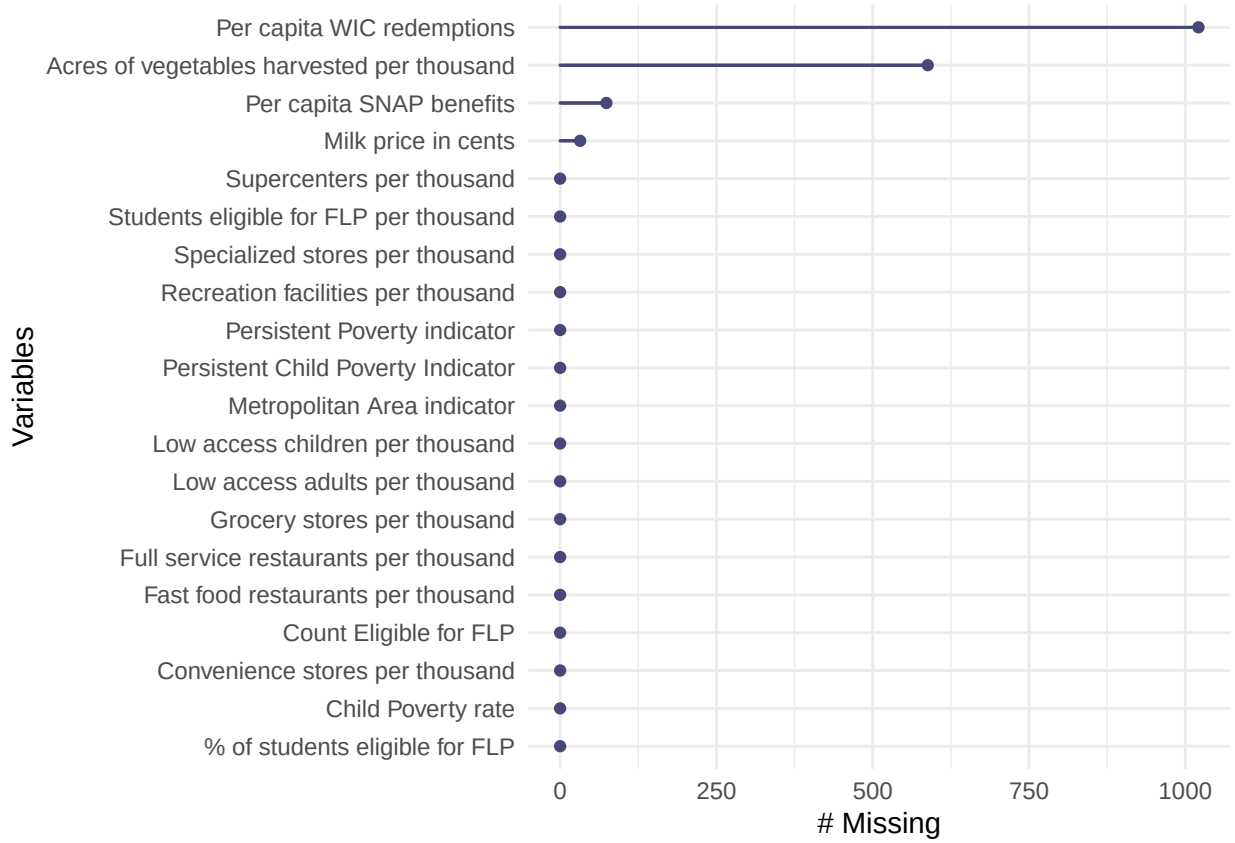
We are also interested in another related dependent variable: Count of children eligible for the free lunch program per hundred thousand. This is calculated by the authors by first finding out the total student population by multiplying the proportion of population aged under 18 years old with the population of the county (estimated using census methods) in 2018. Next, the number of students are multiplied by the Percent Eligibility for the Free Lunch Program to find the Count of children eligible for the Free Lunch Program. The final step is to divide this figure by the county population to get the Count of Children eligible for the Free Lunch Program per thousand. One key assumption and thus future model limitation here is that in the absence of better data, we assume that the population under 18 is equal to the population of students of school going age.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



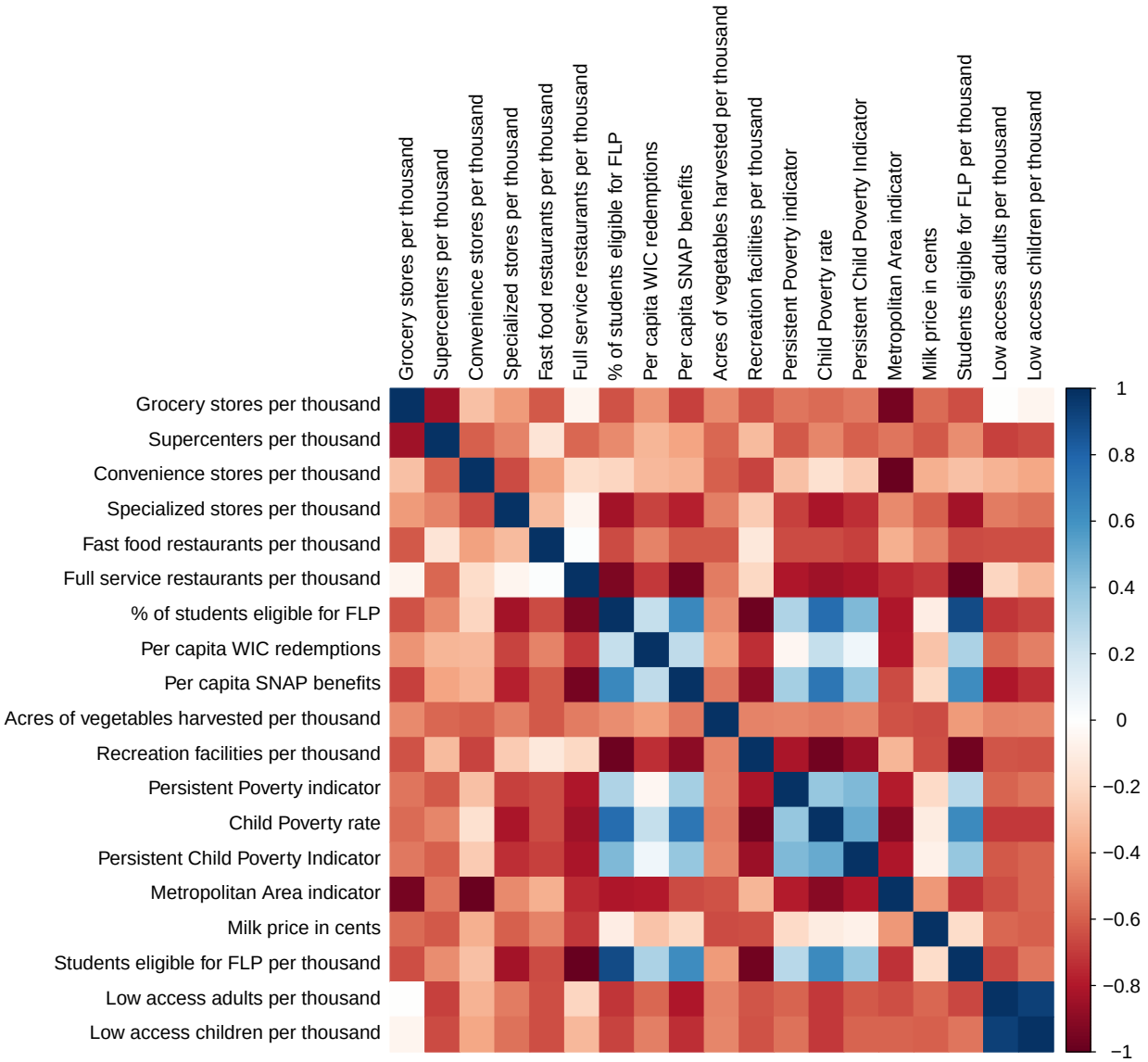

This also seems like highly skewed data, which is expected as there will be some outlier counties with more than the usual amount of resources dedicated to the program and thus having close to all of the school-going population eligible for the free lunch program.

Before we attempt with creating any more regression we should analyze the variables and make sure there is not too much data missing.




From the above we can see the variables of Milk price, Per capita WIC redemptions, Acres of vegetables harvested, and Per capita SNAP benefits all have some missing values. We will replace these missing values with medians of the variables (imputing via mean may not be a good choice because these variables will likely be skewed).

After accounting for the missing values, we also want to know which variables are closely related to our two independent variables of _____ and _____. We can do this with a simple correlation heatmap.




Thus we can see that we expect there to be a fairly strong positive correlation between Percent Eligibility of the Free Lunch Program with per capita WIC redemptions and SNAP benefits, as well all the poverty variables. Meanwhile when we analyze the Students eligible for Free Lunch Program per thousand variable, we can see that it is similarly correlated with the same variables. Thus we expect these variables to be significant in any regression we attempt.

## Linear Regression

We will first attempt a linear regression with Percent of Students eligible for the Free Lunch Program as the dependent variable.

However, we also have to keep in mind that counties receive their resources from the states they are in, thus observations within states will be correlated with each other. Thus, we will report the robust standard error version of this model.

Table 1:

| | *Dependent variable:* | |
|---|---|---|
| | '% of students eligible for FLP' | |
| | (1) | (2) |
| 'Grocery stores per thousand' | $-2.375^{***}$ | $-2.375^{*}$ |
| | (0.758) | (1.280) |
| 'Supercenters per thousand' | $-3.443$ | $-3.443$ |
| | (7.516) | (7.090) |
| 'Convenience stores per thousand' | $1.696^{***}$ | $1.696^{**}$ |
| | (0.533) | (0.710) |
| 'Specialized stores per thousand' | $-2.975$ | $-2.975$ |
| | (2.070) | (2.202) |
| 'Fast food restaurants per thousand' | 0.556 | 0.556 |
| | (0.568) | (0.721) |
| 'Full service restaurants per thousand' | $-1.447^{***}$ | $-1.447^{***}$ |
| | (0.326) | (0.514) |
| 'Per capita WIC redemptions' | $0.120^{***}$ | $0.120^{***}$ |
| | (0.021) | (0.035) |
| 'Per capita SNAP benefits' | $0.271^{***}$ | $0.271^{***}$ |
| | (0.030) | (0.040) |
| 'Acres of vegetables harvested per thousand' | $0.002^{***}$ | $0.002^{***}$ |
| | (0.001) | (0.001) |
| 'Recreation facilities per thousand' | $-9.146^{***}$ | $-9.146^{***}$ |
| | (2.057) | (2.237) |
| 'Persistent Poverty indicator' | $-0.101$ | $-0.101$ |
| | (0.635) | (0.856) |
| 'Child Poverty rate' | $1.025^{***}$ | $1.025^{***}$ |
| | (0.033) | (0.044) |
| 'Persistent Child Poverty Indicator' | $2.898^{***}$ | $2.898^{***}$ |
| | (0.520) | (0.552) |
| 'Metropolitan Area indicator' | $-0.225$ | $-0.225$ |
| | (0.349) | (0.382) |
| 'Milk price in cents' | $0.064^{***}$ | $0.064^{***}$ |
| | (0.013) | (0.012) |
| 'Low access adults per thousand' | $-0.010^{***}$ | $-0.010^{**}$ |
| | (0.003) | (0.004) |
| 'Low access children per thousand' | $0.041^{***}$ | $0.041^{**}$ |
| | (0.012) | (0.019) |
| Constant | $4.498^{***}$ | $4.498^{***}$ |
| | (1.379) | (1.357) |
| F Statistic (df = 3; 360) | $12.879^{***}$ | $7.73^{***}$ |
| Observations | 3,065 | 3,065 |
| $R^2$ | 0.760 | 0.760 |
| Adjusted $R^2$ | 0.758 | 0.758 |
| Residual Std. Error (df = 3047) | 7.976 | 7.976 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Thus we can see that the following variables are significant using both the robust and the non-robust standard error option: - Grocery Stores per thousand (negatively associated) - Convenience stores per thousand (positively associated) - Full service restaurants per thousand (negatively associated) - Per capita WIC redemptions (positively associated) - Per capita SNAP benefits (positively associated) - Acres of vegetables harvested per thousand (positively associated) - Recreation facilities per thousand (negatively associated) - Child poverty rate (positively associated) - Persistent child poverty indicator (positively associated) - Milk price in cents (positively associated) - Number of low access adults per thousand (negatively associated) - Number of low access children per thousand (positively associated)

**Coefficient Interpretation**

As the model is a linear regression, interpretation is straightforward, with one unit increase in the independent variable corresponding with a (in this case) percent change in Percent Eligibility of Enrollment in the Free Lunch Program. The largest positive coefficient is of the Persistent Child Poverty indicator variable. Thus if a county faces persistent child poverty, the Percent Eligibility increases by 2.89 %. The largest negative coefficient is that of recreation facilities. If there is an increase of one recreation facility per thousand people, percent eligibility decreases by 9.14 %. This seems like a larger than usual effect, although one possible hypothesis is that areas with more recreation facilities may be richer and and thus have stricter eligibility requirements. However, this is just conjecture.

Out of the poverty variables, only child poverty rate and persistent child poverty indicator significant, with the latter having a larger coefficient. This indicates that the general poverty of the area is more useful in predicting the percent eligibility enrollment than the exact childhood poverty rate.

Out of the store and restaurant availability variables, it seems that grocery stores and full service restaurants correspond with decreased eligiblity for the Free Lunch Program while the opposite is true for convenience stores. This may be because the former stores' abiity to provide nutrition to children may outweigh that of convenience stores.

Most of our variables are significant with this linear regression, bringing the Multiple R-squared to a 75.96% and adjusted R squared not far behind at 75.82%.

**Goodness of fit: Hypothesis Test**

We want to test the hypothesis that the above model is more useful than a null model. We can let our null hypothesis be

$$H_0 : \text{Model with no independent variables fits data better than linear regression model}$$

reduced model in predicting the Percent Eligibility for the Free Lunch Program. We can let our null hypothesis be that

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 3064 | 806174.77 | | | | |
| 2 | 3047 | 193832.11 | 17 | 612342.67 | 566.23 | 0.0000 |

As the F statistic is large and the p value is smaller than our significance level of 0.05, we reject the null hypothesis. This means we find evidence that there is at least one predictor whose slope is not 0.

Suppose we want to conduct another hypothesis test, this time letting the null model be just the predictors for the stores, restaurants, and vegetable harvests. We want to test whether these variables are better than our full regression.

$H_0$ : Model with only store, restaurant, and vegetable variables is a better predictor than our full model

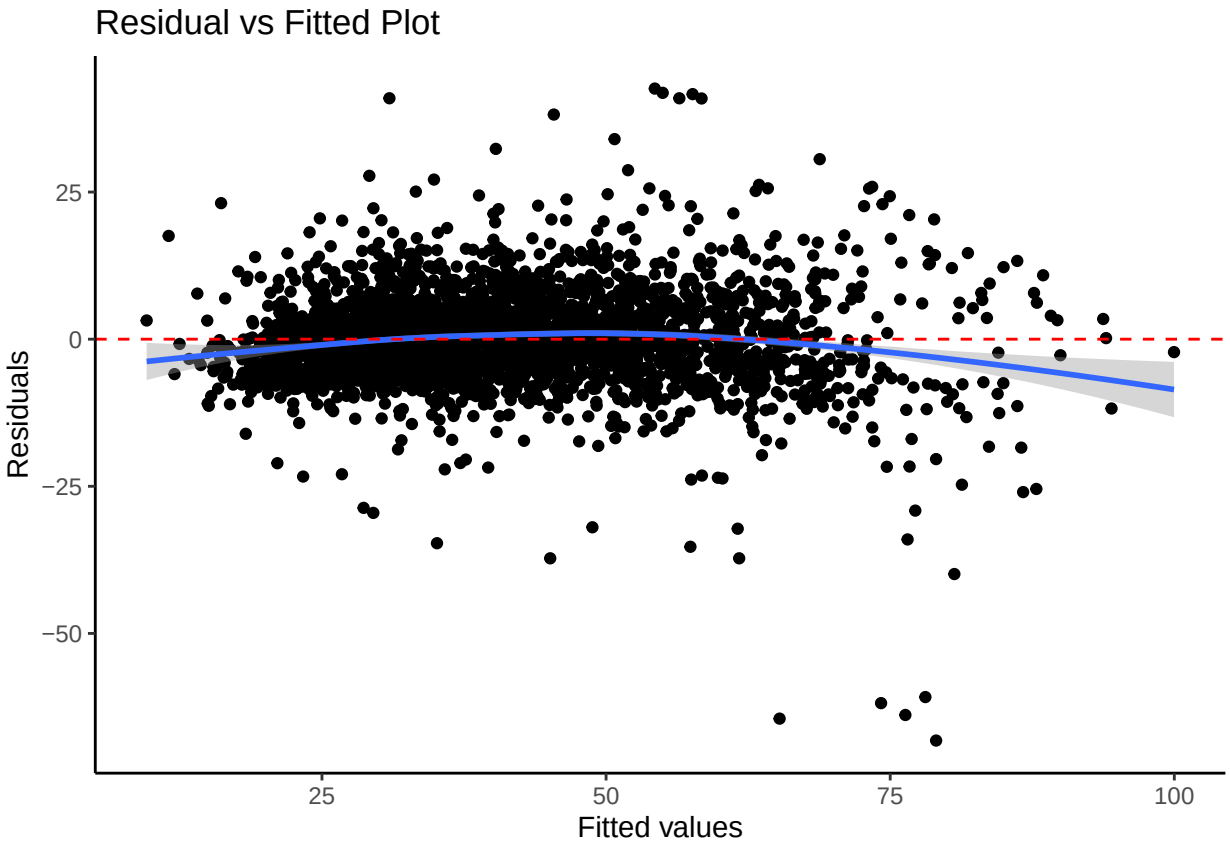
| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 3058 | 681494.35 | | | | |
| 2 | 3047 | 193832.11 | 11 | 487662.24 | 696.90 | 0.0000 |

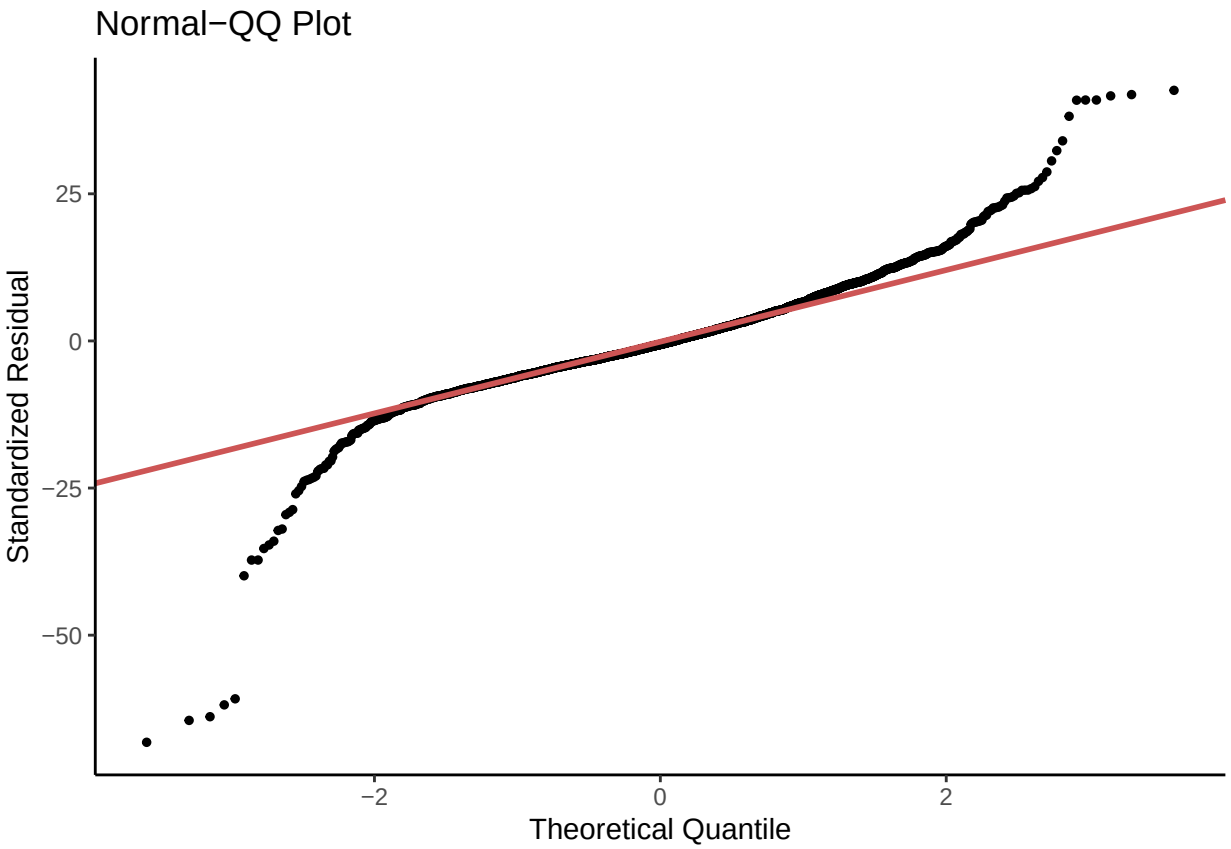
As the F statistic is large and the p value is smaller than our significance level of 0.05, we reject the null hypothesis. This means we find evidence that there is at least one predictor apart from the store and restaurant variables that makes our full regression a better predictor of Percent Eligibility of Enrollment.

However, we must be careful with making too many hypotheses as our probability of reaching a false positive or our type 1 error rate increases for every additional hypothesis we make.

**Goodness of fit plots**

It is important to analzye the assumptions of linear regression and see if our model meets them.






While the residuals look as if they are largely following the assumption of constant variance of residuals, there are still some outliers for larger values where the predictions get more extreme. This shows an error with the model assumptions. In addition, the quantile-quantile plot's deviating tails indicate the presence of non-normality in error residuals as well. While the latter can be accepted if we consider our sample to be large enough, the former violates the validity of the model

## # End of writeup