COVID-19: Understanding the Factors

Sharon John
Wendy Kha
Matt Kymn
Jonathan Martinez
Roney Safiani
Tanglin Xia

STATS 101A

2020/07/31

Table of Contents

1. Abstract

The spread of COVID-19 has taken the world by storm, causing nations to shut down in order to protect its citizens from sickness and death. When dealing with a worldwide pandemic, it is essential to analyze what the underlying causes are in contracting this disease and how one might be able to avoid it. In our statistical analysis of this COVID-19 dataset, several different factors were taken into account regarding what might be related to the rate of cases for each area in the United States. Race, household sizes and densities, education, and temperatures within those regions were all taken into account when observing its relation to the rate of cases across the country. For such a complex phenomena, one will usually not be able to find all of the variables that relate to the rise in cases, but given the data, we were able to explain 25% of the variance in rate of cases per 1000 people. Through the following analysis, it was made evident that race plays a big factor in the contraction of COVID-19, educated areas do not necessarily have lower case rates, and the household sizes/densities & temperatures play an interactive role in rate of cases.

2. Question Studied

The purpose of this study was to determine what variables affected the number of COVID-19 cases per thousand people in counties across the US. Specifically, we explored the questions:

1. Does race play a role in the number of cases?
2. Does the density of a person's country/city/household correlate to the rate of cases?
3. Does higher education within a certain area play a role in the rate of cases within that area?
4. Does temperature play a notable role in the number of cases per thousand people?

3. Variables

a. Dependent Variable (outcome variable)

The response variable of this study was the number of cases per thousand residents in each county. This number was acquired by dividing the number of confirmed COVID-19 cases by the county's estimated population.

b. Independent Variables (predictors)

The independent variables are as follows:

1. **household density (density)**
    a. household density and size may show a correlation between these variables and the number of cases. This is a numerical variable .
2. **Black percentage (black_pct)**
    a. Black percentage is a numerical variable showing the percentage of people that are black in a certain county. This variable was included to explore the role of ethnicity composition in the contraction of the disease.
3. **percentage of people with with Bachelor's degree or more (college)**
    a. The percentage of people with a bachelor's degree or more was included to determine whether having more education will reduce the amount of cases in a particular area. This is a numerical variable.
4. **big household size (bighh)**
    a. A big household can show whether our expectations of a larger household contracts more cases than a smaller one. This is a categorical variable with 2 levels.
5. **combined effect of big household size and high temperature (bighh:highTemp)**
    a. The combined effect of household size and high weather temperatures can show whether warmer/cooler weather and the number of people in a household correlate with case rates. This is a categorical variable with 2 levels.

4. Exploratory Data Analysis

For two of the numerical variables, caseRate and density, the logarithmic transformation was taken in order to make these variables normally distributed because the regular variables were very skewed. These transformation decisions were made by using the symbox() function in R. Figure 1(a) and 1(b) show the distribution of the two transformed variables. As can be seen from the histograms, the skewness of the original variables was addressed.

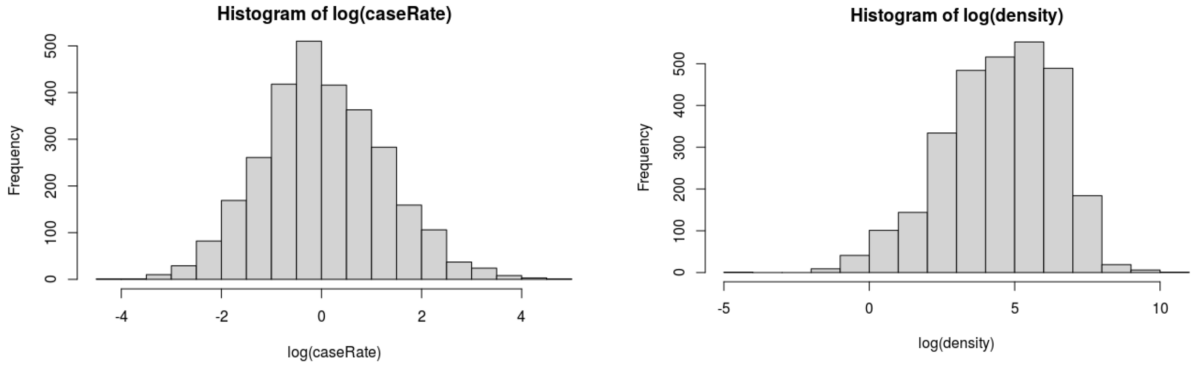

Figure 1(a), 1(b)

Two categorical variables were used in this study, and before we could use them for our study, a contingency table was made to determine whether each category had enough data points.

	High Temperature		
Big Household		False	True
	False	923	517
	True	507	934

Table 1

Observing the frequency tables of our categorical variables: bighh & highTemp

The contingency table between these two categorical variables shows a nice distribution of observations in all four cells. There aren't any cells lacking a significant amount of observations, therefore the frequencies within the cells are ideal.

Scatterplot matrix of predictors:

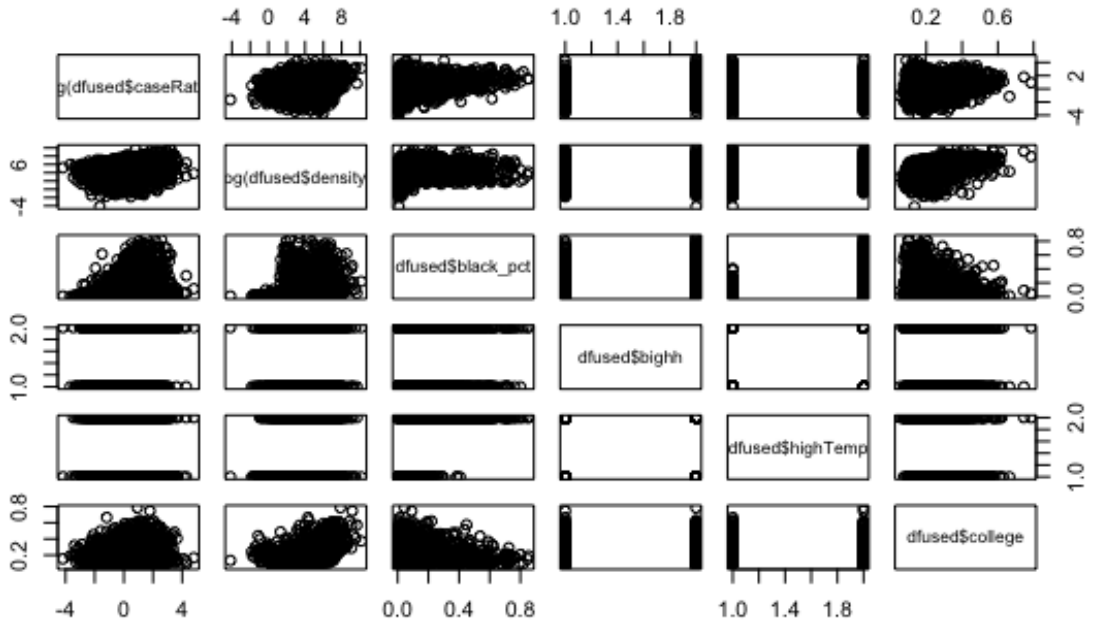

Figure 2. Correlation Matrix Between Predictors

Correlation Matrix for Numerical Predictors:

	caseRate	density	black_pct	college
caseRate	1.00	0.20	0.21	0.04
density	0.20	1.00	0.08	0.42
black_pct	0.21	0.08	1.00	-0.10
college	0.04	0.42	-0.10	1.00

Table 2. Correlation Coefficient Table

There isn't significant correlation among any of the numerical predictors used in our model. Interestingly however, density and college have a correlation value of 0.42 which is much bigger than the correlations among the rest of the variables. This correlation value shouldn't be alarming, however, because areas with greater density surely have more colleges and other education opportunities.

5. Linear Regression

a. Model Validity

Assumptions of MLR:

- Linearity
- Normality
- Homoscedasticity
- Leverage points or influential points
- Collinearity

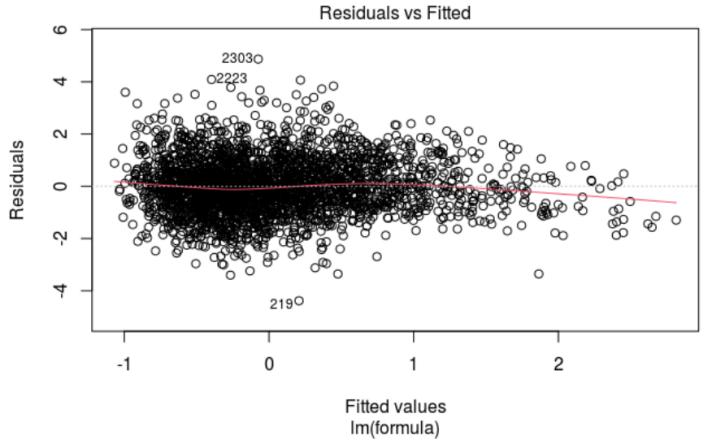

Figure 3

Looking at the residual plot, we can see that there's no clear pattern in the data, and thus, the assumption of linearity is met.

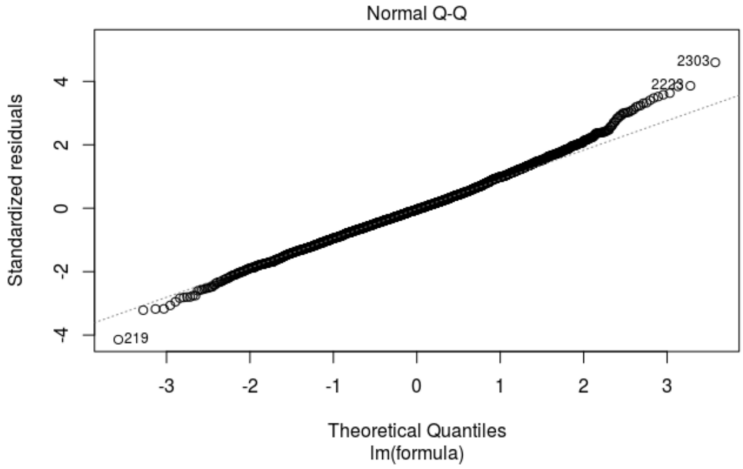

Figure 4

The QQ plot shows that all the data points lie perfectly on a straight line. This indicates that the assumption of normality of the residuals of the data is also met.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.341904, Df = 1, p = 0.067536
```

Figure 5

Homoscedasticity: As displayed above, when conducting an NCV Test, our p-value was 6.8% which is above 5%, thus there's no conclusive evidence to say that the data exhibits heteroscedasticity. The assumption of homoscedasticity is therefore met.

	VIF	Df
log(density)	1.32	1
black_pct	1.30	1
bighh	2.21	1
college	1.37	1
bighh:highTemp	2.90	2

Table 3. VIF output

The VIF Test above illuminates that our predictors do not exhibit multicollinearity since their respective VIF values are less than 5. This means that our predictors are not completely dependent on one another which allows for more accurate analysis.

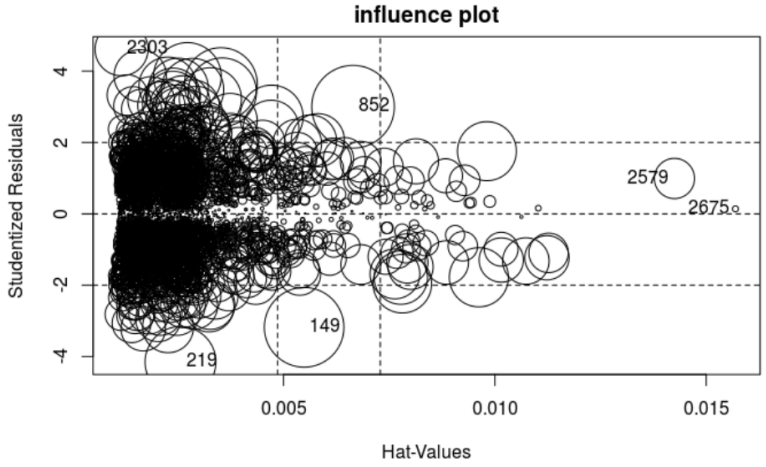

Figure 7. Influence plot

The influence plot of our linear model doesn't show any significant bad leverage points. Only two points stand out as outliers, such as observation 2303 and 219, but these shouldn't be too deterministic for such a big dataset.

For such a large data set we can consider an outlier to be outside of a studentized residual value of +/-4. Our leverage boundary is 0.0049 as calculated by $h_{ii} > 2(k+1)/2881$, where k is the

number of predictors.

b. Model

Summary of our linear model:

Predictors	Coefficient Estimate	Standard Error	t-value	p-value
(Intercept)	-1.03478	0.06540	-15.821	< 2e-16
log(density)	0.08486	0.01190	7.129	1.27e-12
black_pct	3.82076	0.15330	24.923	< 2e-16
bighhTRUE	0.57893	0.05874	9.856	< 2e-16
college	1.26106	0.24078	5.237	1.75e-07
bighhFALSE:bighhTempTRUE	-0.28180	0.06312	-4.464	8.34e-06
bighhTRUE:highTempTRUE	-0.58986	0.06246	-9.444	< 2e-16

F-statistic: 163.5 on 6 and 2874 p-value: <2.2e-16
R-squared: 0.2545 Adjusted R-squared: 0.2529

Table 4. Linear regression output

As shown in the summary of our linear model above, all of the predictors are statistically significant as illustrated by their respective p-values. The F-statistic of the model is also statistically significant, indicating that our model is a good fit. This significance is displayed by the respective p-value of <2.2e-16. 25% of the variance in the rate of cases can be explained by our predictors. 75% of the variance in COVID-19 case rates cannot be explained by our predictors.

Analysis of Variance:

	DF	Sum of Squares	Mean of Squares	F-value	p-value (Pr(>F))
log(density)	1	227	226.7	201.87	<2e-16
black_pct	1	637	636.6	566.93	<2e-16
bighh	1	68	68.0	60.60	9.70e-15
college	1	59	58.6	52.16	6.51e-13
bighh:highTemp	2	112	55.9	49.77	<2e-16
Residuals	2874	3227	1.1		

Table 5. AOV output

The summary of the analysis of error variance illustrates how much each predictor contributed to our R^2 value. The biggest contributing predictor in our model was the percentage of black people and the least contributing predictor was the proportion of college degrees. The proportion of college degrees was still kept as a predictor since our AIC model selection indicated that it was a good fit for our model. The percentage of balck people explained over 14% of the variance in case rates per 1000 people. Each predictors R^2 value is displayed below.

R-square contribution for each predictor:

log(density)	0.052
Black percentage (black_pct)	0.147
Big household (bighh)	0.016
Proportion of college degrees (college)	0.014
Big household & high temperatures (bighh:highTemp)	0.026
Total R-square	0.255

Table 6. R-squared contribution of predictors

c. Interpretation of the Results

Predictors	Coefficient Est	Std. Error	t-value	p-value
(Intercept)	-1.03478	0.06540	-15.821	< 2e-16
log(density)	0.08486	0.01190	7.129	1.27e-12
black_pct	3.82076	0.15330	24.923	< 2e-16
bighhTRUE	0.57893	0.05874	9.856	< 2e-16
college	1.26106	0.24078	5.237	1.75e-07
bighhFALSE:highTempTRUE	-0.28180	0.06312	-4.464	8.34e-06
bighhTRUE:highTempTRUE	-0.58986	0.06246	-9.444	< 2e-16

Table 7. Linear regression output

Coefficient interpretations:
Keeping all else constant...

- For a 1% increase in density, on average, caseRate increases by 0.085%.
- For every 1% increase in the black proportion of the population, on average, there's a e(3.82 * 1%) - 100% = 3.8% increase in the case rate.
- Compared with locations with smaller households, areas with bigger average household size's average case rate is e(0.5789) - 100% = 78.4% higher.
- For every 1% increase in the proportion of people with a college degree, on average, there's a e(1.261 * 1%) - 100% = 1.26% increase in the case rate.
- There's a significant interaction effect between the big household variable and the high temperature variable.

Interpretation of R^2: 25.29% of the variance in rate of cases per 1000 people is explained by density of a household, percentage of blacks, household size, and proportion of college graduates. ~75% of the variance is not explained by these predictors.

d. Interpretation of the Interaction Effect between Household Size & Temperature

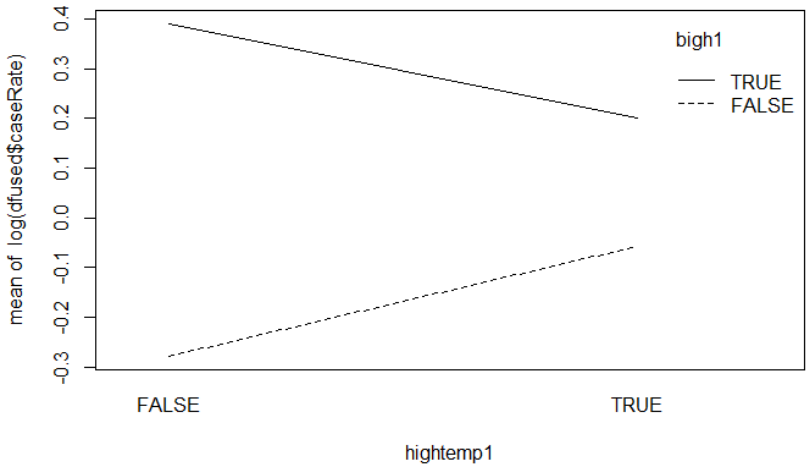

Figure 8. Interaction plot

The interaction effect plot illustrates that on average, in a bigger household, higher temperature correlates with a lower infection rate compared to lower temperatures. Moreover, on average, in a smaller household, higher temperature correlates with a higher infection rate compared to lower temperatures.

e. Model Evaluation

Observing the Marginal Model Plots:

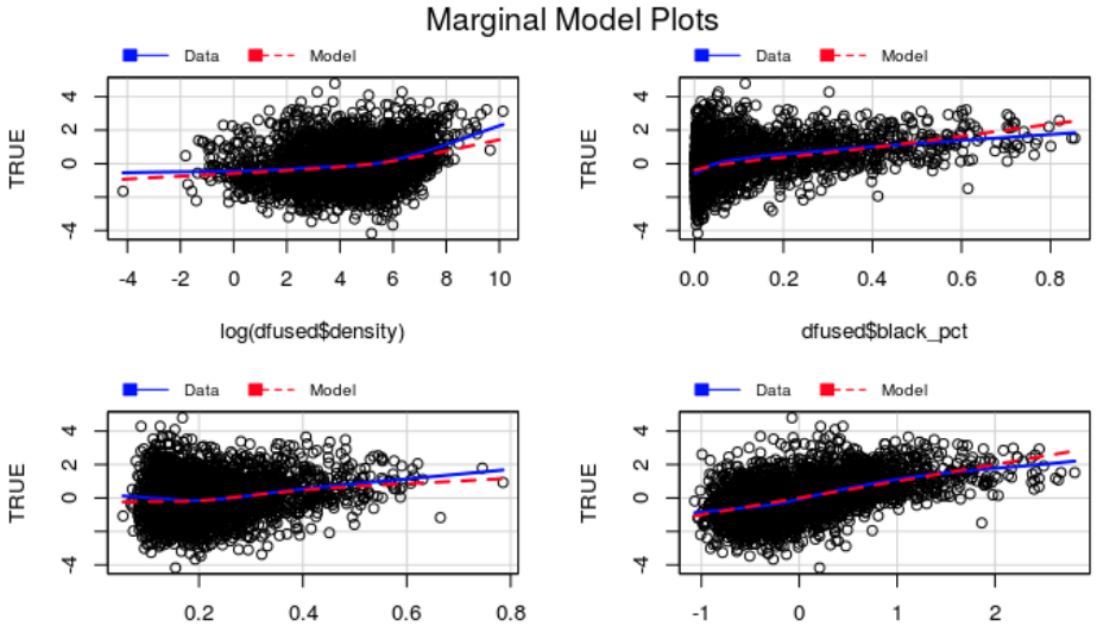

Figure 9. MMP plots

The marginal model plots indicate that our regression model is a good model because the LOESS fit line follows the regression lines decently well for all of our predictors. Only towards the end do the LOESS and regression lines deviate from overlapping, but the deviation is not significant.

f. Model Selection (Backward AIC Test)

	Df	Sum of Squares	RSS	AIC
<none>			3226.9	340.68
college	1	30.80	3257.7	366.05
log(density)	1	57.06	3284.0	389.18
bighh:highTemp	2	111.76	3338.7	434.77
black_pct	1	697.41	3924.3	902.40

Table 8. Backward AIC output

The backward model selection illuminates that our current model is the best model that doesn't deal with any extraneous predictors. As the model is, the AIC value is the lowest, indicating that it's a good fit. If any predictor were to be removed, the AIC value and the residual sum of

squares would increase significantly, resulting in a worse model. Therefore, we conclude that our model is ideal.

g. Cross Validation

In order to assess the results of our statistical analysis, we conducted cross validation on our dataset. We split the data into training and testing samples and found a correlation value of 0.47 between the observed values and predicted values.

Since the correlation value isn't too high, it would not be ideal to conduct predictions based off of this data. This is expected, however, due to the nature of this study. Since this is a social study, there is significant complexity in how this disease spreads and how different variables interact with each other.

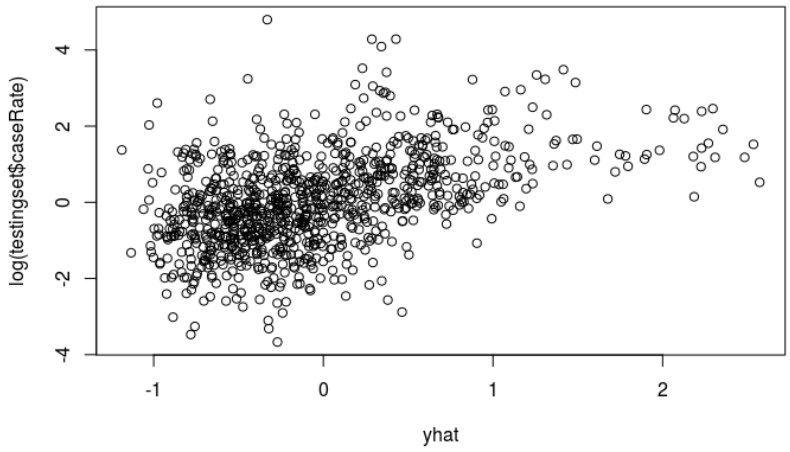

Figure 10. Cross-validation

6. Logistic regression

a. Output and Confidence Interval

The same variables were also used to perform a logistic regression. The AIC score we got for the logistic regression model was 3251.9. This AIC score was much higher for logistic regression than for MLR because we converted a continuous variable such as case rate into a categorical variable, leading to significant information loss.

	Odds Ratio	**95 % Confidence Interval**	
Predictors	**Estimate**	**2.5%**	**97.5%**
(Intercept)	0.2634	0.2010	0.3440
density	1.0010	1.0008	1.0013
black_pct	7227.3396	2599.3286	21225.9431
bighhTRUE	2.0868	1.6505	2.6418
college	4.8711	1.6399	14.5628
bighhFALSE:highTempTRUE	0.5180	0.3889	0.6864
bighhTRUE:highTempTRUE	0.4259	0.3280	0.5522

Table 10. Logistic regression output

Figure 11 below shows an odds ratio plot based on our logistic model. The confidence interval for black_pct was omitted, as if it was included, the huge value would skew the plot by a lot, rendering the plot unclear. As we can see, the most influential predictor is black_pct, which agrees with the MLR model.

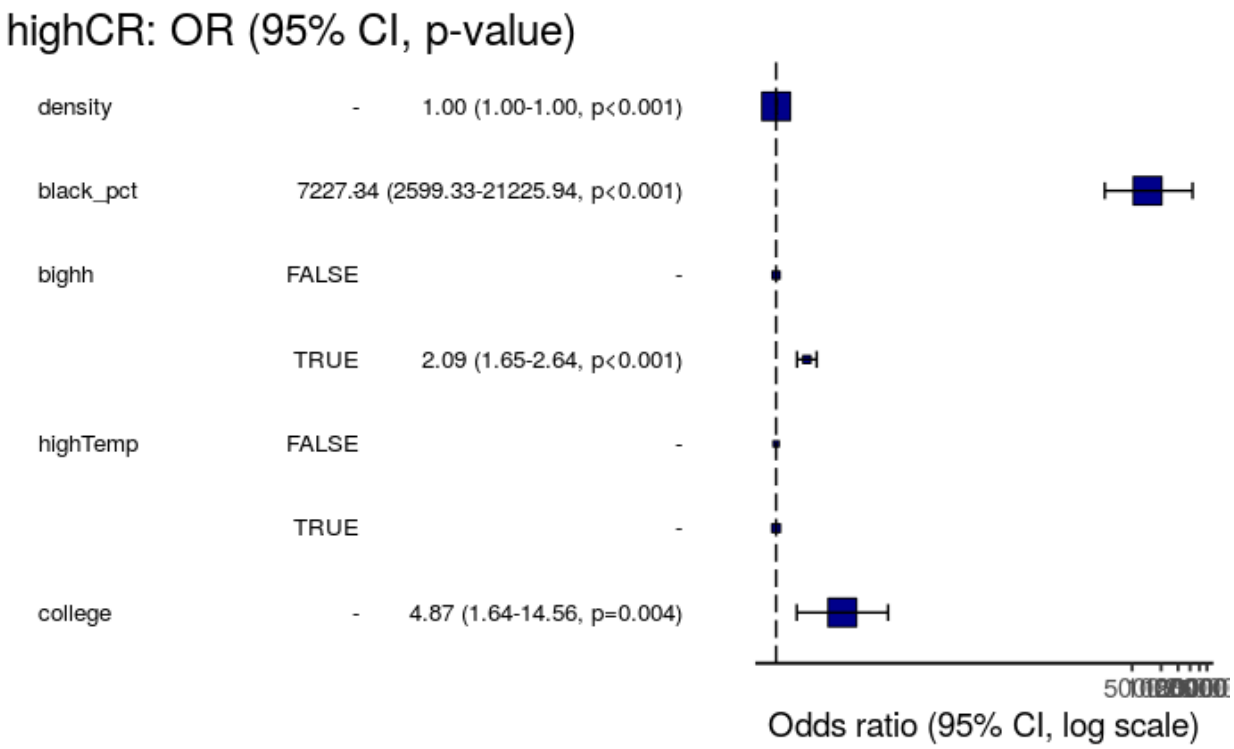

Figure 11. Odds ratio plot

b. Interpretation

Keeping all else constant, on average…

- The odds of having a case rate per 1000 people greater than the median case rate per 1000 people increases by a factor of 1.001 for each unit increase in density.
- The odds of having a case rate per 1000 people greater than the median case rate per 1000 people increases by 9% for each percent increase in black population percentage.
- The odds of having a case rate per 1000 people greater than the median case rate per 1000 people is 1.09 times more for those with big households.
- The odds of having a case rate per 1000 people greater than median case rate per 1000 people increases by 2% for each percent increase in the percentage of people who have received a bachelor's degree or higher.
- The effect of living in different sized households on the case rate per 1000 people is not similar for different temperatures. In particular, those living in a smaller household have a smaller case rate during cooler temperatures instead of warmer temperatures. However, those living in a larger household have a smaller case rate during warmer temperatures instead of cooler temperatures. As the household size increases, the case rate increases

more slowly during warmer temperatures.

7. Conclusion

Based on our regression models, we arrived at the following key conclusions:

- Black racial background in a given neighborhood seems to correlate with increased likelihood of contracting the COVID-19 disease.
- Increased population density correlates with an increase in the spread of the disease. This is in line with what we expected to see.
- Areas with more educated individuals also correlates with higher number of COVID19 cases. This could be due to a number of factors such as educated people traveling for work more, living in cities and more densely populated urban areas or simply a correlation with New York City's mostly professional working population.
- Larger household size is correlated with an increase in the spread of the disease. Large household size with higher temperature however does not correlate with an increase in case rate. In fact, it correlated with a slower spread of disease.
- It is well understood that as we stack interaction effects as we did with household size and temperature, interpretability and attribution becomes more opaque.

With a complex socioeconomic phenomenon such as a pandemic, there are certainly associations and factors that are very complex and these predictors and conclusions serve as a window into further analysis of interaction effects and statistical analysis.

8. Potential Improvements

Data Collection:
Data collected later into the pandemic could be used to improve the model. Many counties did not have COVID-19 cases when this dataset was collected, which skews the data. Also, a lot of the data coincided with a spike in New York so COVID -19 ends up being correlated with features of New York, which isn't necessarily accurate. There must be more variables included because there must be a lot of variables that correlate with the rapid increase in cases in certain parts of the country and world. For example, variables such as household income might be a predictor, but it was not included in the dataset.

Predictor Selection and Model Generation:
The selection of variables was based on common sense and some random trials, therefore, it might be biased. A better model could be generated if we used a neural network. Using gradient

descent, we could potentially find a model with a much higher R^2. However, that would create a lot of interaction effects between predictors, making the result very difficult to interpret.

9. Appendix (R code)

Numeric Data

```{r}
df = read.csv("kolko_covid_shareable_dataset.csv")
dfused = df[which(df$cases != 0),]
dfused = dfused[which(dfused$popestimate2019 != 0),]
dfused = dfused[which(dfused$college != 0),]
dfused = dfused[which(dfused$density != 0),]
```

Adding a categorical variable

```{r}
dfused$highTemp = dfused$temp_mar20 >=
median(dfused$temp_mar20, na.rm = TRUE)
dfused = dfused[!is.na(dfused$highTemp),]
dfused$highTemp <- as.factor(dfused$highTemp)


dfused$bighh <- dfused$hhsize >= median(dfused$hhsize, na.rm =
TRUE)
dfused = dfused[!is.na(dfused$bighh),]
dfused$bighh <- as.factor(dfused$bighh)

```

Create the response variable (case rate per 1000 people)

```{r}
dfused$caseRate = (dfused$cases / dfused$popestimate2019) *
1000
```

Splitting the Data

```{r}
set.seed(1)
```

```
numOfRows = nrow(dfused)
trainingIndices = sample(1:numOfRows, numOfRows*2/3)
trainingset = dfused[trainingIndices,]
testingset = dfused[-trainingIndices,]
```

Looking to see whether transformations are needed.
```{r}
symbox(dfused$density)
symbox(dfused$caseRate)
```

Model
```{r}
formula = log(dfused$caseRate) ~ log(dfused$density) +
dfused$black_pct + dfused$bighh + dfused$highTemp:dfused$bighh
+ dfused$college

#formula = log(dfused$cases) ~ log(dfused$popestimate2019) +
dfused$college + dfused$bighh +
dfused$bighh*log(dfused$popestimate2019)

plot(formula)

model = lm(formula)
summary(model)
vif(model)

backAIC <- step(model, direction = "backward")

summary(aov(model))
```

Check the validity
```{r}
hist(dfused$hhsize)
hist(dfused$temp_mar20)
hist(log(dfused$caseRate), main = "Histogram of log(caseRate)",
xlab = "log(caseRate)")
hist(log(dfused$density), main = "Histogram of log(density)",
xlab = "log(density)")
table(dfused$bighh, dfused$highTemp)
```

```
table(dfused$bighh)
# both histograms look normally distributed

summary(model)

#corplot(model)

plot(model)

library(alr3)
library(car)

mmps(model)

summary(aov(model))

influencePlot(model, main = "influence plot")

vif(model)
ncvTest(model)

#frequency tables
table(dfused$highTemp, dfused$bighh)

round(cor(dfused[,c("caseRate", "density", "black_pct",
"college")]), )
```

Assumption of normality is met as can be seen from the QQ plot since the observations remain on the line.

Assumption of linearity and equality of variance is met as shown by the residual vs fitted plot since the pattern is random and doesn't follow any pattern. Furthermore, the ncv

The MMP plots confirm that our model is a pretty good fit since the loess lines do a decent job of following the regression lines.

According to the influence plot there are no more than a few bad leverage points, but for a dataset with more than 2800 observations, this shouldn't be too deterministic.

The vif function shows that the predictors are not correlated with one another since the vif value is less than 5 for all predictors.

Calculating how much each predictor contributes to R^2
```{r}
#Total Sum of Squares =  4328
227 + 112 + 637 + 59 + 68 + 3227

#Total R^2 = 0.25

#R^2 of log(density) = 0.052
227/4330

#R^2 of bighh = 0.025
68/4330

#R^2 of black_pct = 0.135
637/4330

#R^2 of college = 0.013
59/4330

#R^2 of the combination of bighh and highTemp = 0.025
112/4330

# How much variance in case rate we cannot explain by our predictors (1-R^2) = 0.75
3227/4330
```

Logistic Regression

```{r}
#formula = log(dfused$caseRate) ~ log(dfused$density) + dfused$black_pct + dfused$bighh + #dfused$highTemp:dfused$bighh + dfused$college
dfused$highCR = dfused$caseRate >= median(dfused$caseRate, na.rm = TRUE)
dfused = dfused[!is.na(dfused$highCR),]
dfused$highCR <- as.factor(dfused$highCR)
```

```{r}
lg_model <- glm(dfused$highCR ~ dfused$density +
dfused$black_pct + dfused$bighh + dfused$highTemp:dfused$bighh
+ dfused$college, family = "binomial")
summary(lg_model)
```

#code for exponentiated values for logistic regression
```{r}

round(exp(cbind(Estimate = coef(lg_model),
confint(lg_model))),4)
```

#cross validation

set.seed(1)
numOfRows = nrow(dfused)
trainingIndices = sample(1:numOfRows, numOfRows*2/3)
trainingset = dfused[trainingIndices,]
testingset = dfused[-trainingIndices,]

m1 <- lm(log(trainingset$caseRate) ~ log(trainingset$density) +
trainingset$bighh + trainingset$black_pct + trainingset$college
+ trainingset$bighh:trainingset$highTemp)
summary(m1)
yhat = -1.08917 + 0.09043*log(testingset$density) +
0.56125*(as.numeric(testingset$bighh)-1) +
3.98541*testingset$black_pct + 1.34216*testingset$college -
0.23382*(as.numeric(testingset$highTemp)-1) -
0.59393*(as.numeric(testingset$bighh)-1)*(as.numeric(testingset
$highTemp)-1)
cor(yhat, log(testingset$caseRate))
plot(yhat, log(testingset$caseRate))

#Odds ratio plotting
library(finalfit)
dependent = c("highCR")
independent = c("density", "black_pct", "bighh",
"bighh:highTemp", "college")
or_plot(dfused, dependent, independent, table_text_size=3.5)