COVID-19
Linear Models & Analysis

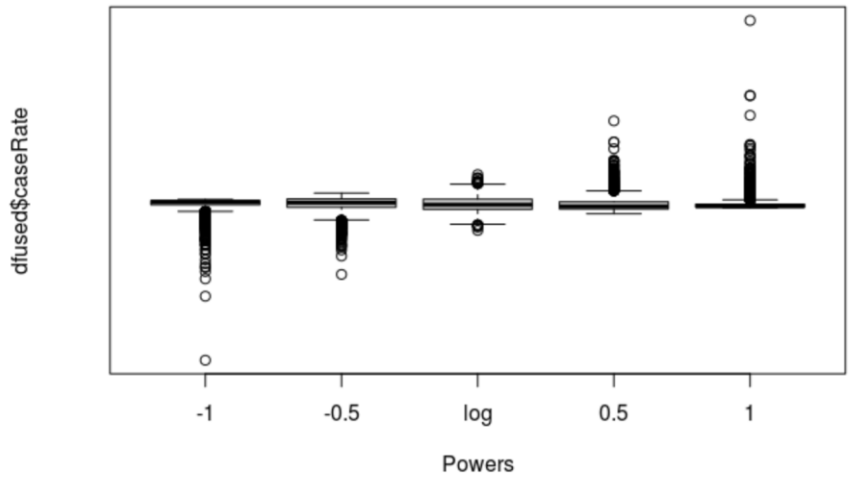
BY
JOHN, SHARON BINU
KHA, WENDY
KYMN, MATTHEW JUNHEONG
MARTINEZ, JONATHAN TONATIUH
SAFIANI, RONEY
XIA, TANGLIN

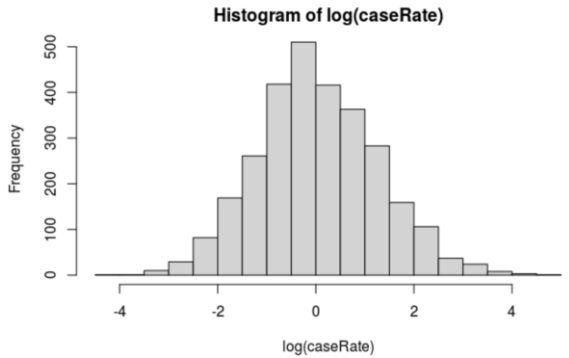
What questions do we have about COVID-19?

- Does race play a role in contracting COVID-19?
- How can the density in one's county/city/household play a role in COVID-19 rates?
- Do areas with more educated individuals have a tendency to have less cases than areas with lower education rates?
- Does weather play a role in contracting COVID-19?

Response Variable

- Case rate per 1000 individuals
- The case rate was very right skewed, with few counties having super high case rate.
- Therefore, we used log transformation to make it normally distributed.





Predictors

- Household density (density) [log transformed for normality, numeric variable]
- Black percentage (black_pct)[numeric variable]
- Percentage of people with bachelor's degree or more (college) [numeric variable]
- Big household size (bighh) [categorical variable, 2 levels]
- The combined effect of big household size and high temperatures (bighh:highTemp) [categorical variables, both are 2 levels]

Predictors continued...

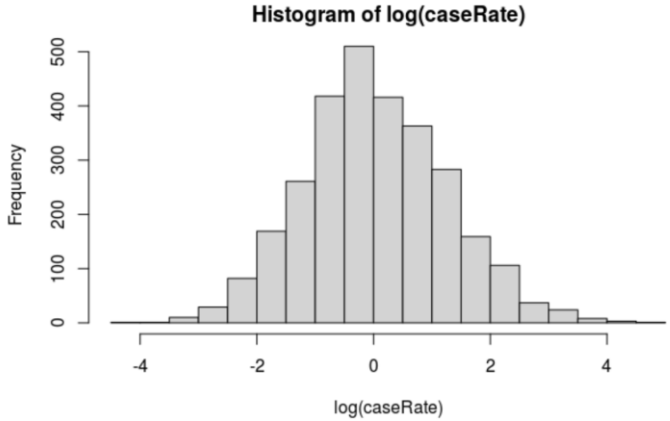
- Household size was a numeric variable that was changed to a factor. This factor was created to have two levels (big household size and small household size). This split was made on the median value of household size. Small household size are all the observations below the median and big household size consists of all the observations that are above the median.
- Temperature was also a numeric variable that was made into a factor of two levels for high temperature and low temperature. The split was once again made on the median to separate high temperature observations from lower temperature observations.

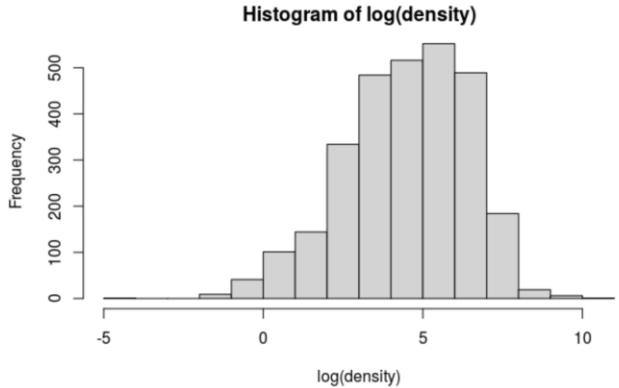
How these predictors can answer our questions...

- Household density and size can point towards a correlation with case rate.
- Black percentage will illustrate whether certain ethnicities are more likely to contract the disease.
- The percentage of people with a bachelor's degree or more will illuminate whether being more educated reduces the amount of cases in that particular area.
- A big household will test whether our expectation of a larger household will contract more cases than a smaller one.
- 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.

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. The histograms below show the distributions of these transformed variables.

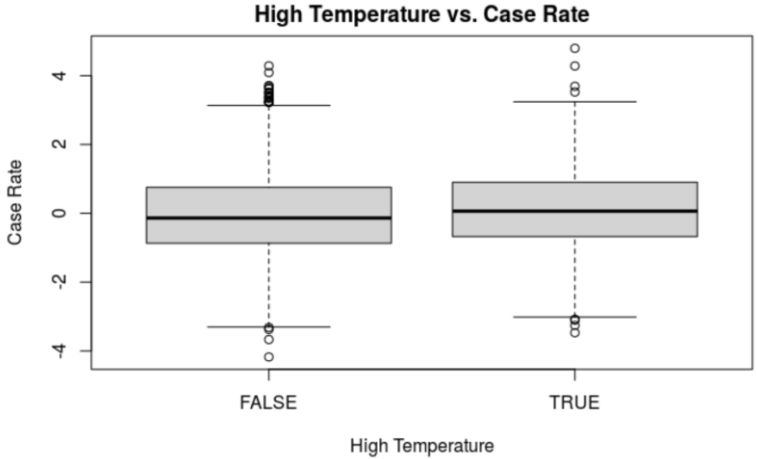




Exploratory Data Analysis

Boxplots of categorical data: big household size & high

Bigger households tends to have higher case rates



Higher temperatures don't raise case rates in a similar way

Exploratory Data Analysis

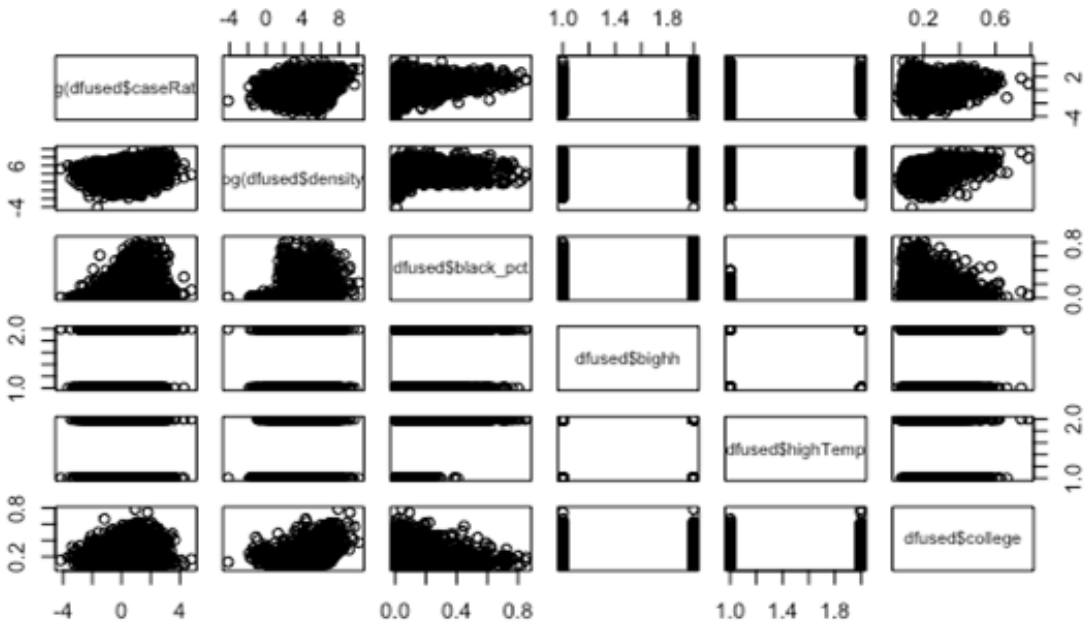
Observing the frequency tables of our categorical variables: bighh & highTemp

- 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.

bighh \ highTemp	False	True
False	923	517
True	507	934

Exploratory Data Analysis

Scatterplot matrix of predictors:



Exploratory Data Analysis

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

- 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.

Model Output

Summary of our model:

- All of our 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.
- 25% of the variance in the rate of cases can be explained by our predictors.

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

Model Output

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		

R-square contribution for each predictor:

log(density):
0.052
black_pct:
0.147
bighh:
0.016
college:
0.014
bighh:highTemp: 0.026

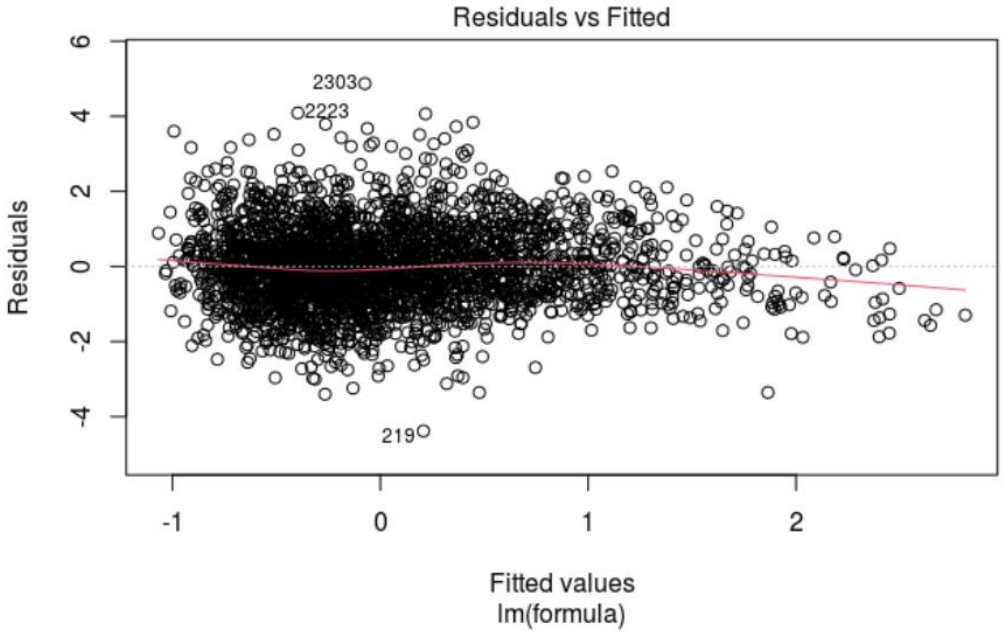
Checking assumption of Regression

Assumption Checklist:

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

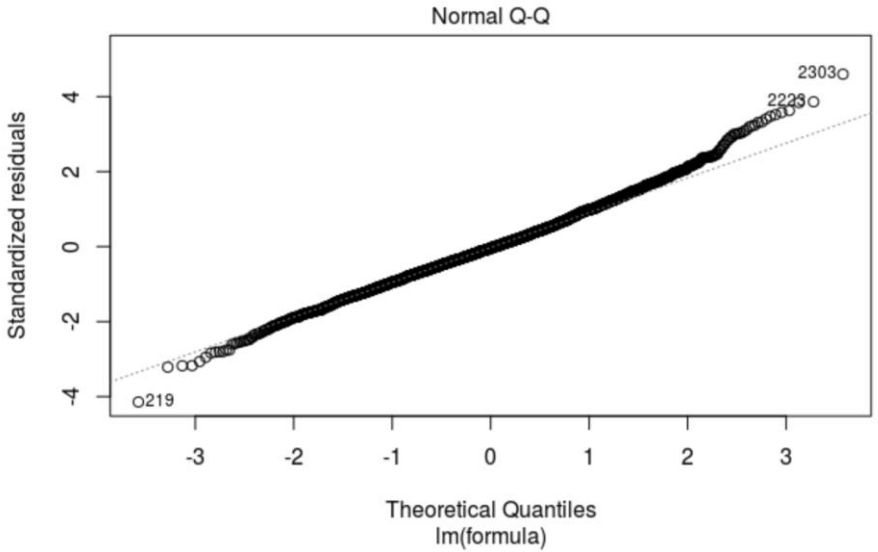
Linearity

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



Normality

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



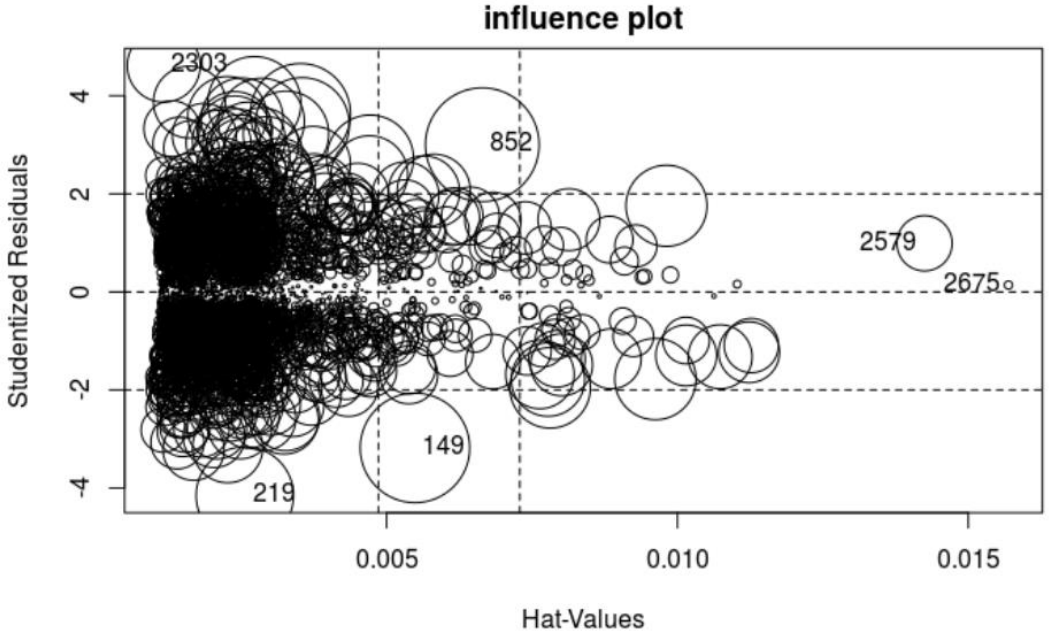
Homoscedasticity

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.

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

Bad Leverages

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 and our leverage boundary is 0.0049 as calculated by hii > 2(k+1)/2881, where k is the number of predictors.

Collinearity

The VIF Test 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.

	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

Model Selection

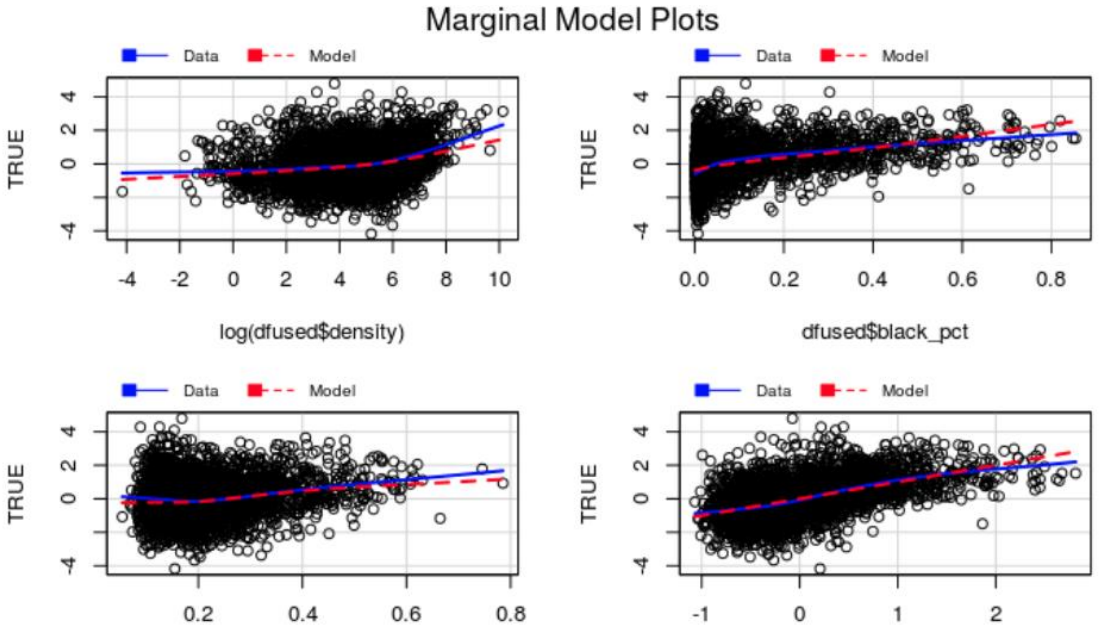
Backward AIC Test:

	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

Our 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.

Model Evaluation

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.



Interpretation of the Results

Coefficient interpretation:
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.

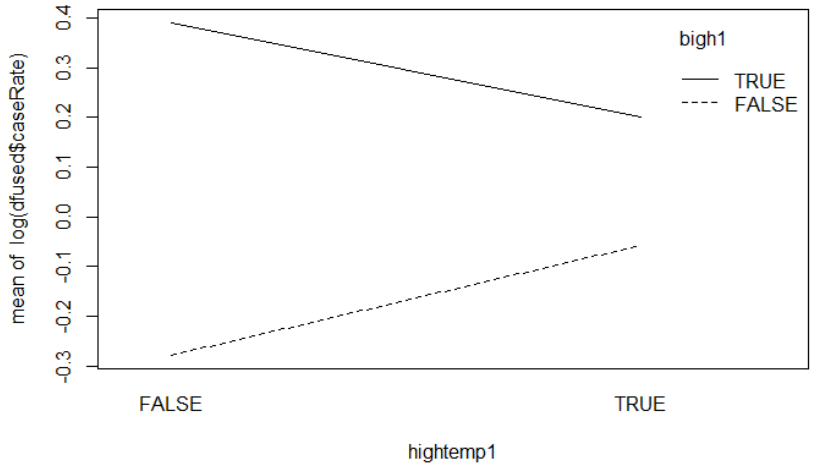
Predictors	Coefficient Est	Std. Error	t-value	p-value
(Intercept)	-1.03478	0.06540	-15.821	< 2e-16
log(density)	0.08486	0.01190	7.129	1.27e-12
black_pct	3.82076	0.15330	24.923	< 2e-16
bighhTRUE	0.57893	0.05874	9.856	< 2e-16
college	1.26106	0.24078	5.237	1.75e-07
bighhFALSE:highTempTRUE	-0.28180	0.06312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

Interpretation of R2:
25.29% of the variance in caseRate is explained by density, black_pct, bighh, and college

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

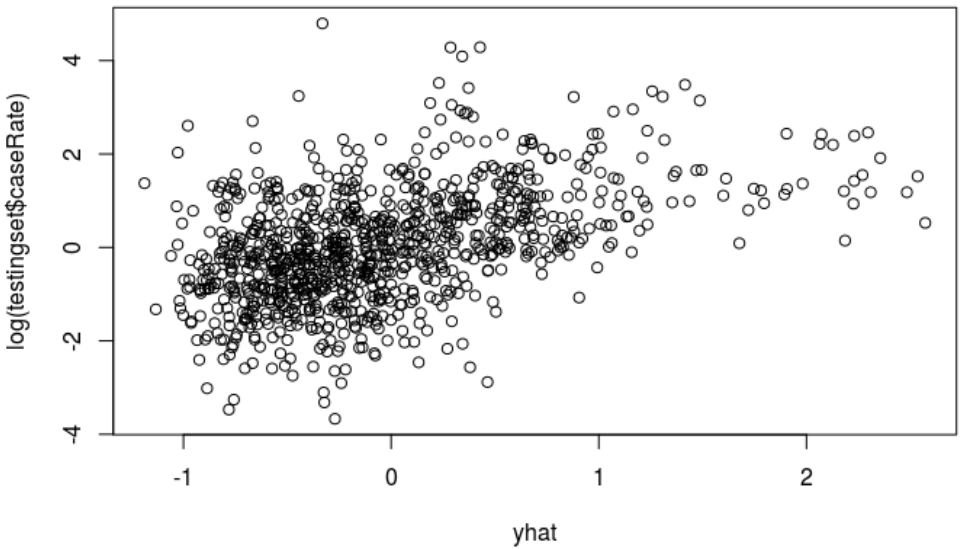
Interpretation of the Interaction

- On average, in a bigger household, higher temperature correlates with a lower infection rate compared to lower temperatures.
- On average, in a smaller household, higher temperature correlates with a higher infection rate compared to lower temperatures.



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.



Logistic Regression Output

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3993.9 on 2880 degrees of freedom
Residual deviance: 3237.9 on 2874 degrees of freedom

AIC: 3251.9

Predictors	Coefficient Estimate	Standard Error	t-value	p-value
(Intercept)	-1.3340118	0.1370485	-9.734	< 2e-16
density	0.0010304	0.0001387	7.427	1.11e-13
black_pct	8.8856263	0.5354708	16.594	< 2e-16
bighhTRUE	0.7356200	0.1199622	6.132	8.67e-10
college	1.5833100	0.5566817	2.844	0.00445
bighhFALSE:highTempTRUE	-0.6577239	0.1448275	-4.541	5.59e-06
bighhTRUE:highTempTRUE	-0.8535041	0.1328468	-6.425	1.32e-10

Interpretations

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 a 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.

Conclusion

- A black racial background has correlation with the contraction of this disease, our analysis confirms reports that African-American communities have been hardest hit in the US.
- Population density within an individual's home/city/county plays a significant role in the rate at which COVID-19 spreads.
- Areas with more educated individuals have a tendency to have an increased number of cases than areas with lower education rates which goes against our initial intuition.
- Higher temperatures are related to lower case rates in regions with large household sizes, and higher temperatures are related to higher case rates in regions with smaller household sizes. Household size and temperatures are inversely correlated with regards to the rate of cases.

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.
  - Data coincided with spike in New York so COVID -19 ends up being correlated with features of New York.
  - Include more variables. Variables such as household income might be a factor, but was not included in the dataset.
- **Predictor Selection and Model Generation**
  - The selection of variables was based on common sense and some random trials. 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.