Group 9 Final Project: COVID-19 Presentation

By: Changmin Ahn, Cathy Luo, Yiling (Ofelia) Pan, Evan Pickell, Vishal Sundaram

Abstract

Due to the current worldwide pandemic of COVID-19, our group intend to find out the specific county-level demographic features having the strongest relationship with the spread of coronavirus. 25 variables were originally raised as potential leading factors to the numerous cases and deaths of COVID-19. With the help of R and relevant statistical computations, the case rate can be found to be more important than the death rate, which can be affected by a variety of economic and social factors beyond demographic features. Besides, only six of the remaining variables were eventually determined to have significant influences over the cases of COVID-19 after creating a correlation matrix and going through forward and backward selections. the variables utilized in the study include pcases (the per-capita case rate of COVID-19 for each county in the U.S.), temp_mar20 (actual temperature averages for each county in March 2020), dist (distance from the population-weighted county centroid to the population-weighted county centroid of Manhattan), black_pct (percentage of county residents who identify as Black), hisp_pct (percentage of county residents who identify as Hispanic), transit_modeshare (percentage of county residents who commute by public transportation), and hhsize (the average household size of county residents). Among them, hhsize is made into a categorical variable, and divided into hhsize_cLarge, hhsize_cMid and hhsize_cSmall for the MLR, and hhsize_cMid and hhsize_cSmall for the logistic model.

Abstract (Continued)

With six independent variables and one dependent variable, pcases, MLR transformations were utilized to analyze the relationship. By using Symbox function, the log transformation of pcases best fits the data. Assumptions for the linear regression model are also tested. The final conclusion is that $\log(pcases) = 10.717788 - 0.014065 * tempmar20 - 0.427941 * dist + 0.279238 * blackpct + 0.260365 * hisppct + 0.130462 * transitmodeshare - 0.112210$*hhsizecMid-0.444466*hhsizecSmall. We found that all of these variables had a practical effect on increasing COVID-19 case rates. However, the most impactful predictor in the model was the distance of the county's centroid from the centroid of Manhattan, the area with the largest amount of COVID-19 infections at the time of the data.

Data

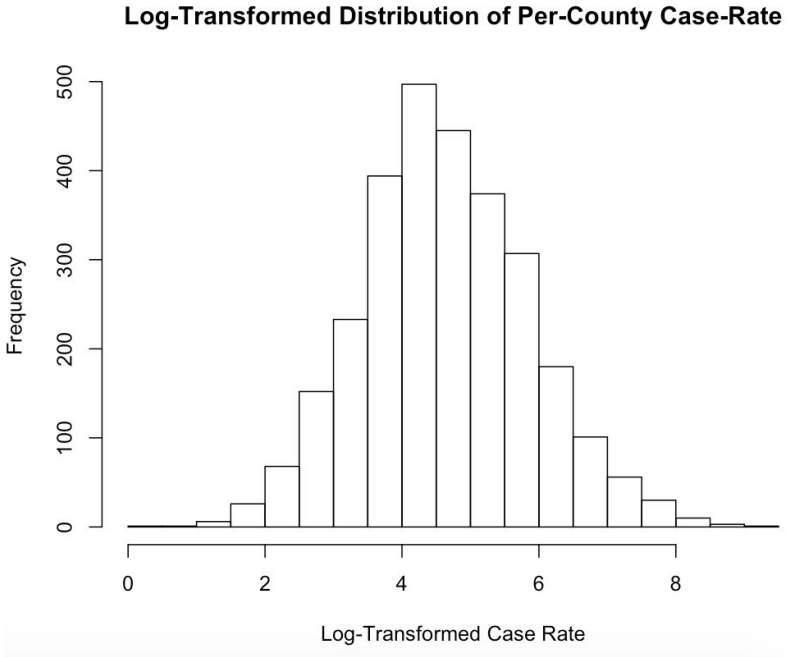
1. Our data comes from The New York Times's published daily counts of COVID-19 cases and deaths by county, as well as the 2018 U.S. Census Demographic data.
2. Each row of our data represents a different county within the United States and their associated COVID-19 data as of May 13th.

Question

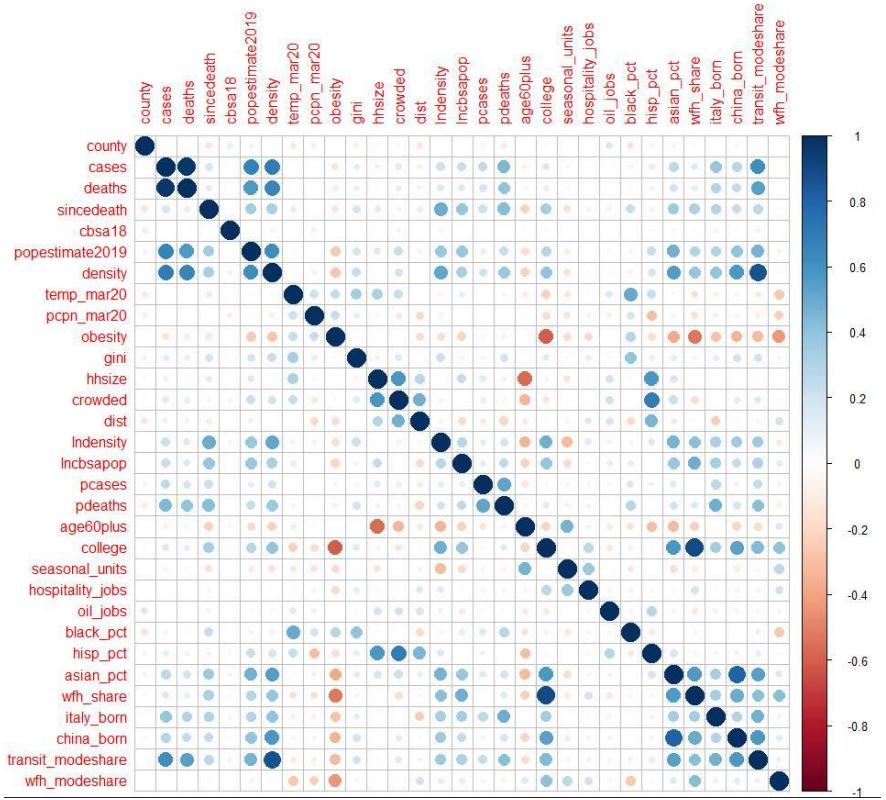
Which county-level demographic features demonstrate the strongest relationships to the spread of COVID-19, as measured by county-level case rates from across the United States?

Outcome Variable: pcases

1. Our dependent variable is the county-level case rate of COVID-19
2. The mean case rate is 212 cases per 100,000 county residents.
3. The case rate ranges from from 0 cases per 100,000 county residents to 12,088 cases per 100,000 county residents .
4. We decided to rescale it into a log(y) version to make it less skewed as seen in the histogram.

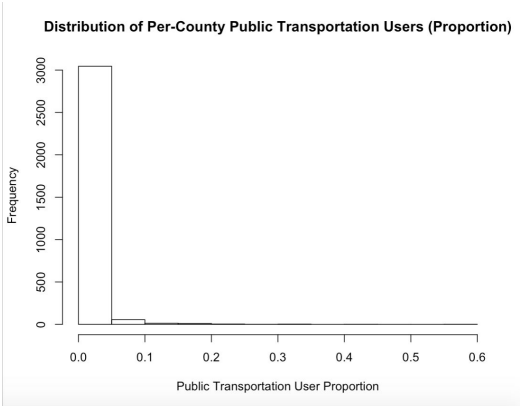


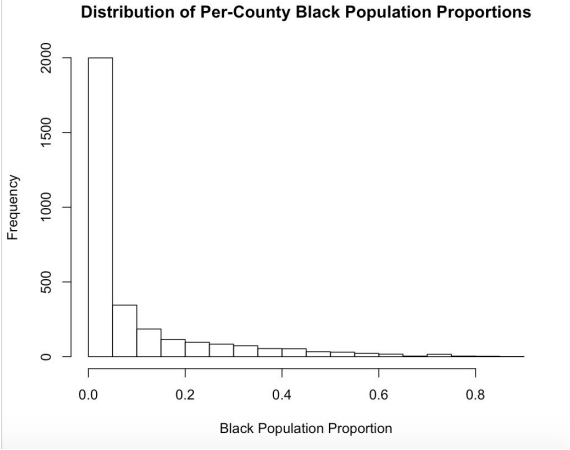
Exploratory Data Analysis: Correlation

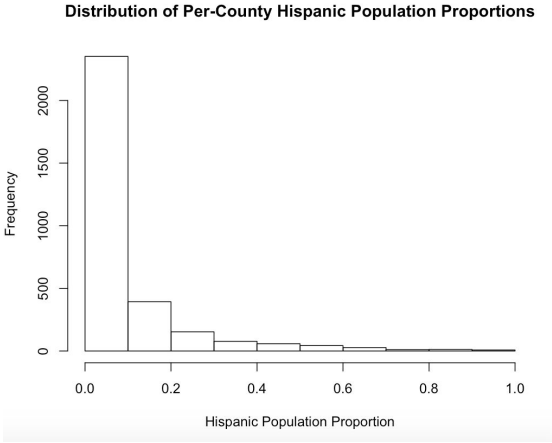


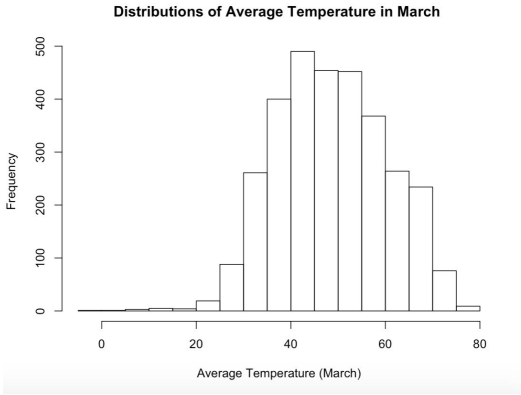
We created a correlation matrix in order to visualize the correlation between the various variables in our data and help start the process of selecting the variables for our final multiple linear regression model.

EDA: Analyzing Potential Predictors





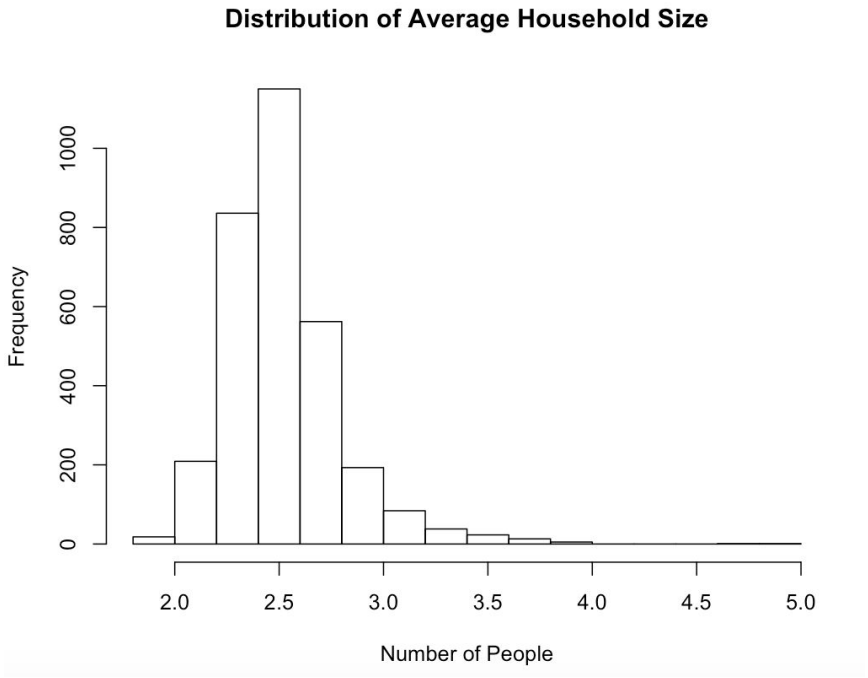




We then created histograms to look at the distributions of significant predictors based off of information from the correlation matrix.

EDA: Creating a Categorical Variable

Our COVID-19 data set did not include any categorical variables, so we had to create one ourselves. We plotted the histogram of the household size variable and ran summary statistics to learn more about its distribution.



EDA - Categorical Variable Continued

Breakdown of hhsize:

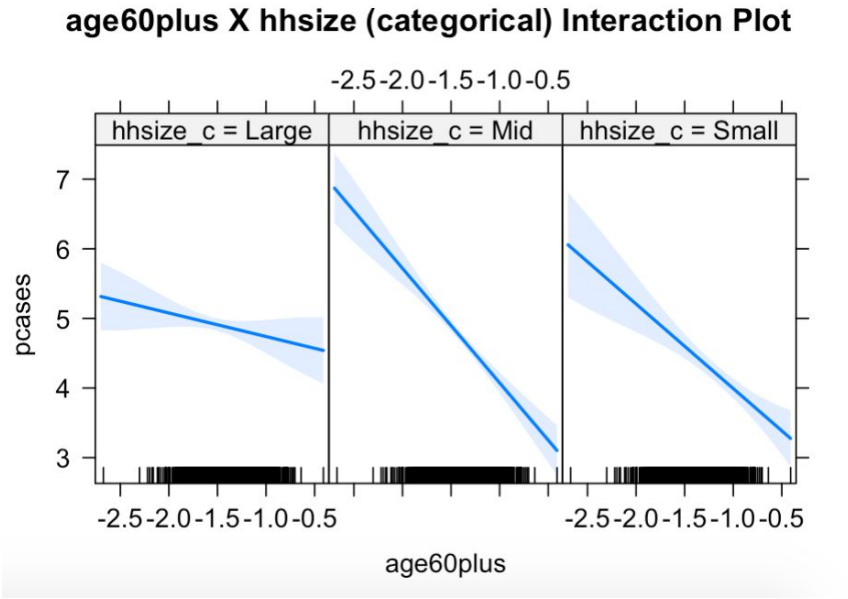
Min	1st Quartile	Median	Mean	3rd Quartile	Max
1.826	2.352	2.484	2.520	2.634	4.971

We then categorized household sizes as small, medium, or large. If the average household size (hhsize) was less than 2.352 people, it was labeled “small.” If the average hhsize was between 2.352 and 2.634, it was labeled “medium.” If the average hhsize was greater than 2.634, it was labeled “large.”

Small	Medium	Large
772	1559	802

Interaction Effect

Combination of Age60Plus and HHsize (categorical)



Predictors	Df	Sum Sq	Mean Sq	F value	p-value
age60plus	1	22647652	22647652	103.323	<2e-16
hhsize_c	3	2774317	924772	4.219	0.0055
Age60plus : hhsize_c	2	250319	125160	0.571	0.5650

The effect of the percentage of the older-than-60 population doesn't vary with the household size. The p-value of the interaction effect of age60plus and hhsize is not statistically significant. The lines in the interaction plot are parallel, indicating that as the percentage of population who are greater than 60 years old increases, the case rate also decreases with a similar rate in counties with different average household size.

Exploratory Data Analysis - Continued

Backward Selection Model:

We used backwards selection to test what variables might be statistically significant for our MLR.

```
Selection Algorithm: backward
          popestimate2019 density temp_mar20 pcpn_mar20 obesity gini crowded dist
1  ( 1 ) " "             " "     " "         " "         " "     " "  " "     " "
2  ( 1 ) " "             " "     " "         " "         " "     " "  " "     " "
3  ( 1 ) " "             " "     " "         " "         " "     " "  " "     " "
4  ( 1 ) " "             " "     " "         " "         " "     " "  " "     " "
5  ( 1 ) " "             " "     " "         " "         " "     " "  " "     "*"
6  ( 1 ) " "             " "     "*"         " "         " "     " "  " "     "*"
7  ( 1 ) " "             " "     "*"         " "         " "     " "  " "     "*"
8  ( 1 ) " "             " "     "*"         " "         " "     " "  " "     "*"
9  ( 1 ) " "             " "     "*"         " "         " "     " "  " "     "*"
          lndensity lncbsapop age60plus college seasonal_units hospitality_jobs oil_jobs
1  ( 1 ) " "       " "       " "       " "     " "            " "             " "
2  ( 1 ) " "       " "       " "       " "     " "            " "             " "
3  ( 1 ) " "       " "       " "       " "     " "            " "             " "
4  ( 1 ) " "       " "       " "       " "     " "            " "             " "
5  ( 1 ) " "       " "       " "       " "     " "            " "             " "
6  ( 1 ) " "       " "       " "       " "     " "            " "             " "
7  ( 1 ) " "       " "       " "       " "     " "            " "             " "
8  ( 1 ) " "       " "       " "       " "     " "            " "             " "
9  ( 1 ) " "       " "       " "       " "     "*"            " "             " "
          black_pct hisp_pct asian_pct wfh_share italy_born china_born transit_modeshare
1  ( 1 ) " "       " "      " "       " "       "*"        " "        " "
2  ( 1 ) "*"       " "      " "       " "       "*"        " "        " "
3  ( 1 ) "*"       " "      " "       " "       "*"        " "        "*"
4  ( 1 ) "*"       "*"      " "       " "       "*"        " "        "*"
5  ( 1 ) "*"       "*"      " "       " "       "*"        " "        "*"
6  ( 1 ) "*"       "*"      " "       " "       "*"        " "        "*"
7  ( 1 ) "*"       "*"      " "       " "       "*"        " "        "*"
8  ( 1 ) "*"       "*"      " "       "*"       "*"        " "        "*"
9  ( 1 ) "*"       "*"      " "       "*"       "*"        " "        "*"
          wfh_modeshare hhsize_cLarge hhsize_cMid hhsize_cSmall age60plus:hhsize_cLarge
1  ( 1 ) " "           " "           " "          " "           " "
2  ( 1 ) " "           " "           " "          " "           " "
3  ( 1 ) " "           " "           " "          " "           " "
4  ( 1 ) " "           " "           " "          " "           " "
5  ( 1 ) " "           " "           " "          " "           " "
6  ( 1 ) " "           " "           " "          " "           " "
7  ( 1 ) " "           "*"           " "          " "           " "
8  ( 1 ) " "           "*"           " "          " "           " "
9  ( 1 ) " "           "*"           " "          " "           " "
          age60plus:hhsize_cMid age60plus:hhsize_cSmall
1  ( 1 ) " "                   " "
2  ( 1 ) " "                   " "
3  ( 1 ) " "                   " "
4  ( 1 ) " "                   " "
5  ( 1 ) " "                   " "
6  ( 1 ) " "                   " "
7  ( 1 ) " "                   " "
8  ( 1 ) " "                   " "
9  ( 1 ) " "                   " "
```

Initial MLR: Backward Selection

Predictors	Coefficient	Standard Error	t-value	p-value
constant	10.89	0.477	23.4	< 2e-16
temp_mar20	-0.013	0.003	-3.8	0.00017
dist	-0.42	0.035	-12.3	< 2e-16
age60plus	0.205	0.29	0.72	0.474
black_pct	0.265	0.029	8.93	< 2e-16
hisp_pct	0.261	0.037	7.12	2.04e-12
wfh_share	0.137	0.195	0.70	0.485
italy-born	0.021	0.027	0.77	0.442
transit_modeshare	0.106	0.027	4	6.79e-05
hhsize_cMid	-0.94	0.533	-1.765	0.0778
hhsize_cSmall	-0.91	0.558	-1.629	0.1036
Age60plus X hhsize_c Mid	-0.557	0.347	-1.61	0.108
Age60plus X hhsize_c Small	-0.311	0.3857	-0.81	0.419

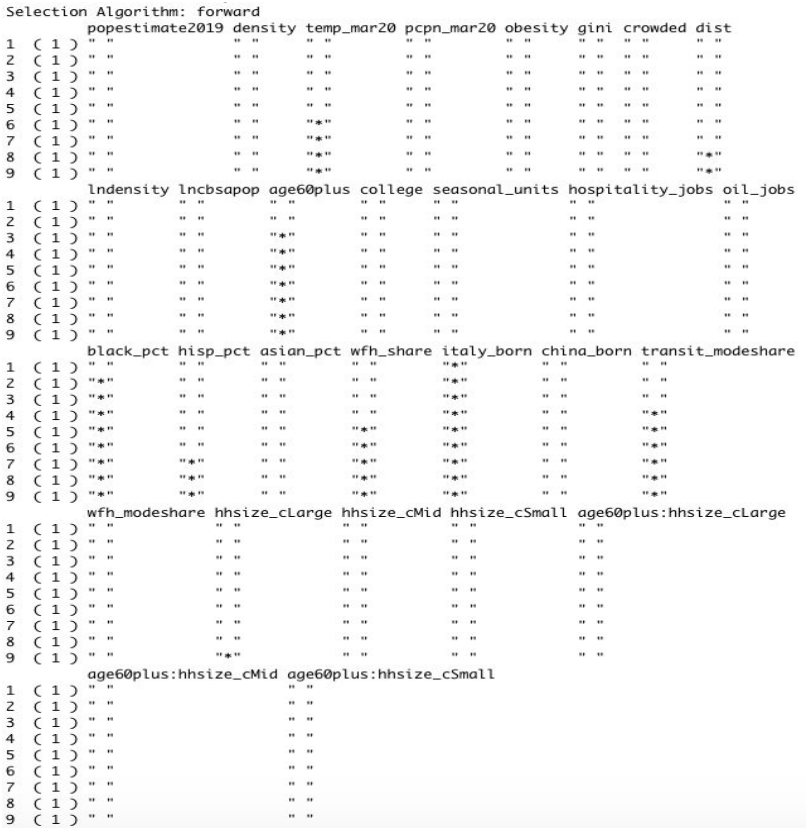
F value = 81.58,
R^2 = 0.4361,
Adj. R^2 = 0.4307

Interaction not significant

Exploratory Data Analysis - Continued

Forward Selection Model:

We used forwards selection to test what variables might be statistically significant for our MLR.



```
Selection Algorithm: forward
         popestimate2019 density temp_mar20 pcpn_mar20 obesity gini crowded dist
1  ( 1 ) " "             " "     " "        " "        " "     " "  " "     " "
2  ( 1 ) " "             " "     " "        " "        " "     " "  " "     " "
3  ( 1 ) " "             " "     " "        " "        " "     " "  " "     " "
4  ( 1 ) " "             " "     " "        " "        " "     " "  " "     " "
5  ( 1 ) " "             " "     " "        " "        " "     " "  " "     " "
6  ( 1 ) " "             " "     "*"        " "        " "     " "  " "     " "
7  ( 1 ) " "             " "     "*"        " "        " "     " "  " "     " "
8  ( 1 ) " "             " "     "*"        " "        " "     " "  " "     "*"
9  ( 1 ) " "             " "     "*"        " "        " "     " "  " "     "*"
         lndensity lncbsapop age60plus college seasonal_units hospitality_jobs oil_jobs
1  ( 1 ) " "       " "       " "       " "     " "           " "             " "
2  ( 1 ) " "       " "       " "       " "     " "           " "             " "
3  ( 1 ) " "       " "       "*"       " "     " "           " "             " "
4  ( 1 ) " "       " "       "*"       " "     " "           " "             " "
5  ( 1 ) " "       " "       "*"       " "     " "           " "             " "
6  ( 1 ) " "       " "       "*"       " "     " "           " "             " "
7  ( 1 ) " "       " "       "*"       " "     " "           " "             " "
8  ( 1 ) " "       " "       "*"       " "     " "           " "             " "
9  ( 1 ) " "       " "       "*"       " "     " "           " "             " "
         black_pct hisp_pct asian_pct wfh_share italy_born china_born transit_modeshare
1  ( 1 ) " "       " "      " "       " "       "*"        " "        " "
2  ( 1 ) "*"       " "      " "       " "       "*"        " "        " "
3  ( 1 ) "*"       " "      " "       " "       "*"        " "        " "
4  ( 1 ) "*"       " "      " "       " "       "*"        " "        "*"
5  ( 1 ) "*"       " "      " "       "*"       "*"        " "        "*"
6  ( 1 ) "*"       " "      " "       "*"       "*"        " "        "*"
7  ( 1 ) "*"       "*"      " "       "*"       "*"        " "        "*"
8  ( 1 ) "*"       "*"      " "       "*"       "*"        " "        "*"
9  ( 1 ) "*"       "*"      " "       "*"       "*"        " "        "*"
         wfh_modeshare hhsize_cLarge hhsize_cMid hhsize_cSmall age60plus:hhsize_cLarge
1  ( 1 ) " "           " "           " "         " "          " "
2  ( 1 ) " "           " "           " "         " "          " "
3  ( 1 ) " "           " "           " "         " "          " "
4  ( 1 ) " "           " "           " "         " "          " "
5  ( 1 ) " "           " "           " "         " "          " "
6  ( 1 ) " "           " "           " "         " "          " "
7  ( 1 ) " "           " "           " "         " "          " "
8  ( 1 ) " "           " "           " "         " "          " "
9  ( 1 ) " "           "*"           " "         " "          " "
         age60plus:hhsize_cMid age60plus:hhsize_cSmall
1  ( 1 ) " "                   " "
2  ( 1 ) " "                   " "
3  ( 1 ) " "                   " "
4  ( 1 ) " "                   " "
5  ( 1 ) " "                   " "
6  ( 1 ) " "                   " "
7  ( 1 ) " "                   " "
8  ( 1 ) " "                   " "
9  ( 1 ) " "                   " "
```

Initial MLR: Forward Selection

Predictors	coefficient	Standard Error	t-value	p-value
constant	10.6459	0.44379	23.988	< 2e-16
temp_mar20	-0.0134	0.0035	-3.869	0.000116
dist	-0.4259	0.0347	-12.270	< 2e-16
age60plus	-0.1288	0.1643	-0.784	0.4332
black_pct	0.2701	0.0296	9.133	< 2e-16
hisp_pct	0.2587	0.0366	7.076	2.71e-12
wfh_share	0.1614	0.1952	0.827	0.4083
italy_born	0.0192	0.0273	0.704	0.4813
transit_modeshare	0.1077	0.0266	4.053	5.42e-05
hhsize_cMid	-0.0950	0.0721	-1.318	0.187835
hhsize_cSmall	-0.4026	0.1027	-3.919	9.46e-05

F value: 81.24, R-squared: 0.4351, Adjusted R-squared: 0.4297

Correlation Matrix b/w Forward, Backward

	Backward	Forward
Backward	1.0000000	0.9825484
Forward	0.9825484	1.0000000

Correlation b/w backwards and forward model is 98.25%, thus they are almost identical and either model would work fine.

Checking performance of our models

	Adjusted R^2	AIC	AICc	BIC
Forward	0.4301	-345.59	-345.42	-305.82
Backward	0.439	-266.75	-266.51	-242.86

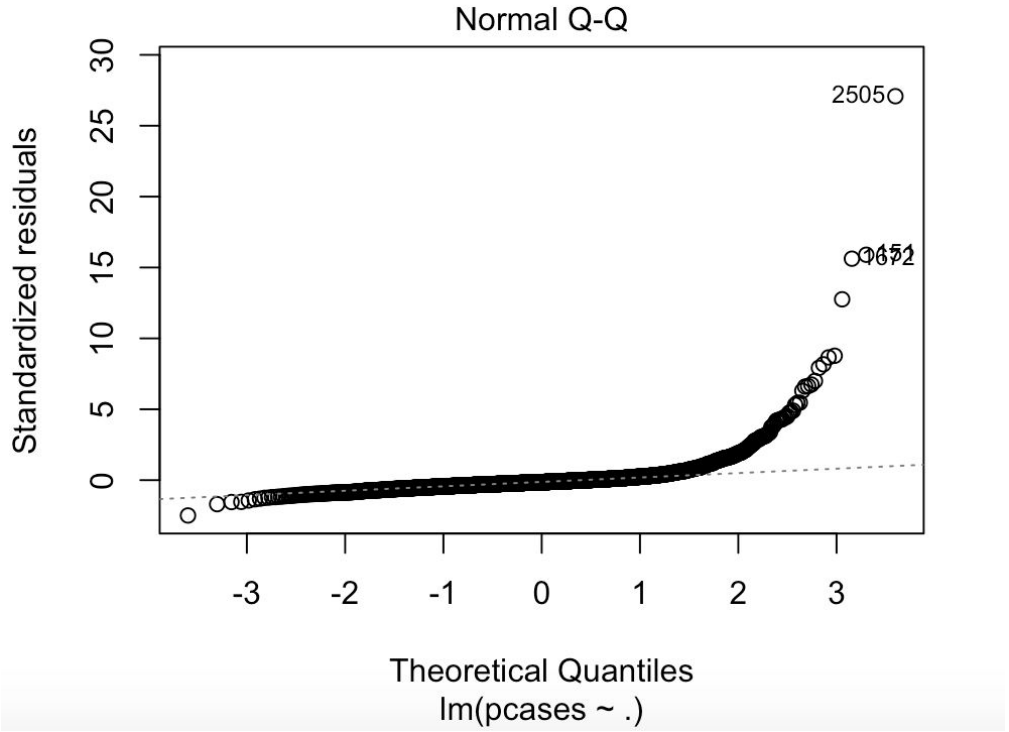
The model chosen from forward selection has lower AIC, AICc, and BIC values, so we picked this model as a final one.

Final Model Code Book

Variable Name	Definition	Variable Type	Outcome or Predictor?	Number of Levels	Transformations
pcases	per-capita case rate of COVID-19 for each county in the U.S.	Numeric	Outcome	0 to 12,088 cases per 100,000 residents	Log transformed
temp_mar20	actual temperature averages for each county in March 2020	Numeric	Predictor	-4.4 to 76.4 degrees Fahrenheit	Log transformed
dist	distance from the population-weighted county centroid to the population-weighted county centroid of Manhattan	Numeric	Predictor	5.204315 to 4382.810300 miles	Log transformed
black_pct	percentage of county residents who identify as Black	Numeric	Predictor	0 to 100%	Log transformed
hisp_pct	percentage of county residents who identify as Hispanic	Numeric	Predictor	0 to 100%	Log transformed
transit_modeshare	percentage of county residents who commute by public transportation	Numeric	Predictor	0 to 100%	Log transformed
hhsize	average household size of county residents	Categorical	Predictor	3 Levels hhsize <= 2.35 : Small Household 2.35 < hhsize < 2.63 : Medium Household hhsize >= 2.63 : Large Household	Log transformed

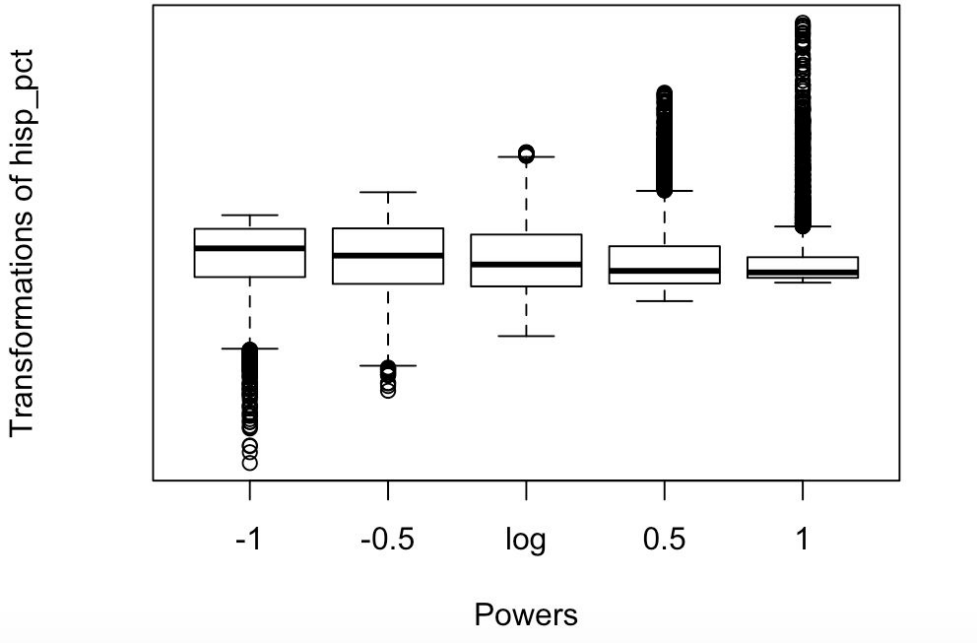
MLR Transformations: Pre-Transformation QQPlot

The exponential nature of the QQplot shows a violation in the assumption of normality, highlighting the potential need to transform the data.

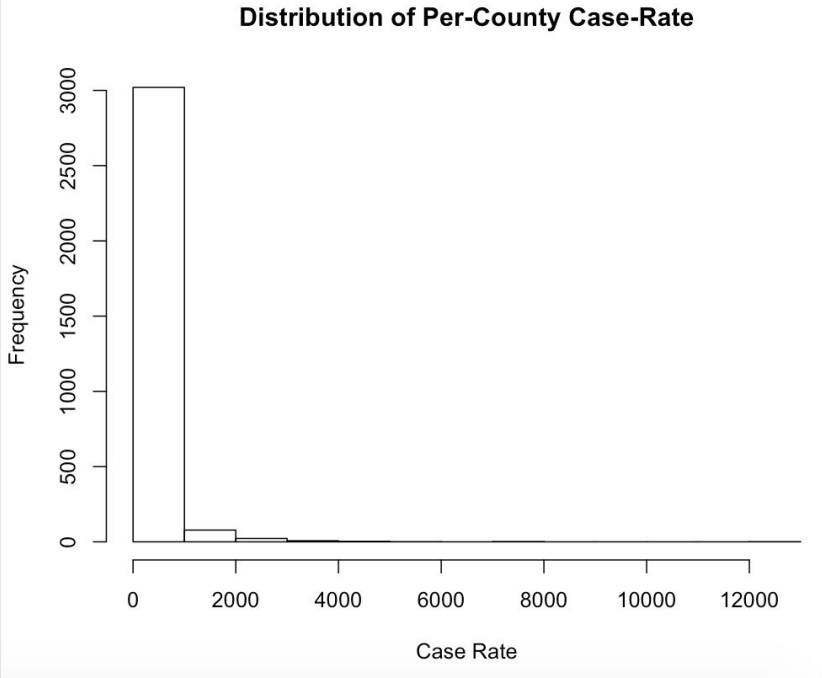


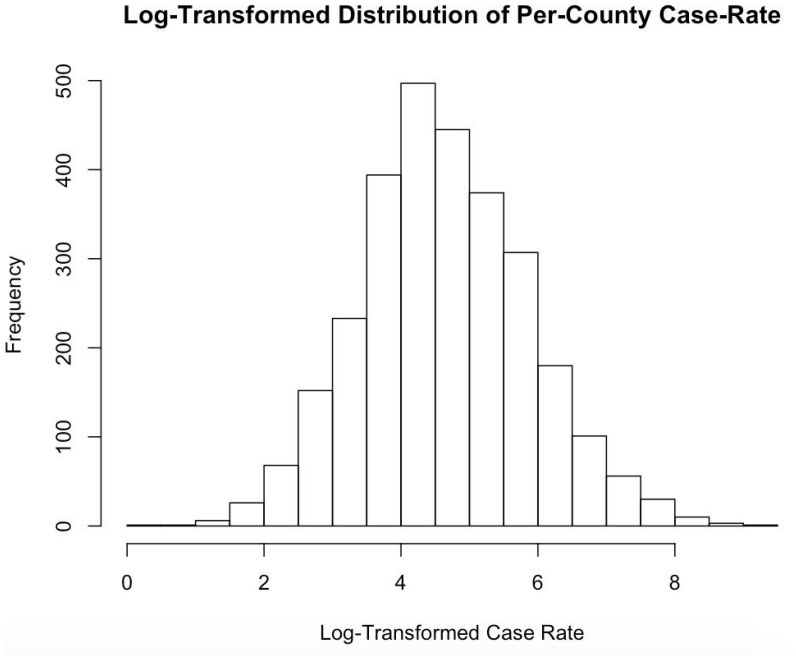
MLR Transformations: Using the Symbox Function

We used the symbox() function in order to check if we should transform variables and if so, which method would work the best. We found that log transforming our data worked the best for most variables.



MLR Transformations: pcases





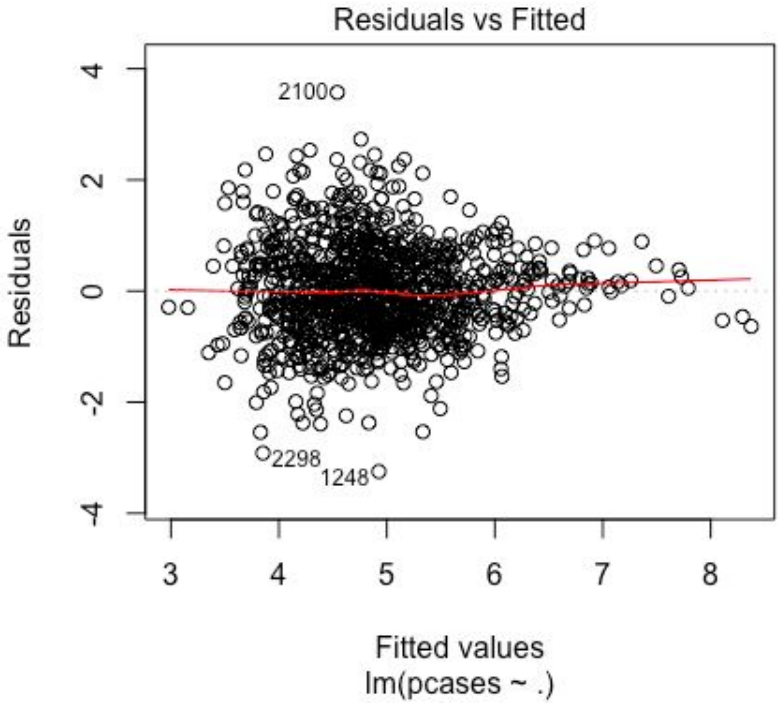
A visual representation of the variable pcases before and after log transformations

Final MLR

$$pcases = \beta_0 + (temp\_mar20)\beta_1 + (dist)\beta_2 + (black\_pct)\beta_3 + (hisp\_pct)\beta_4 +$$

$$(transit\_modeshare)\beta_5 + (hhsize_c Mid)\beta_6 + (hhsize_c Small)\beta_7$$

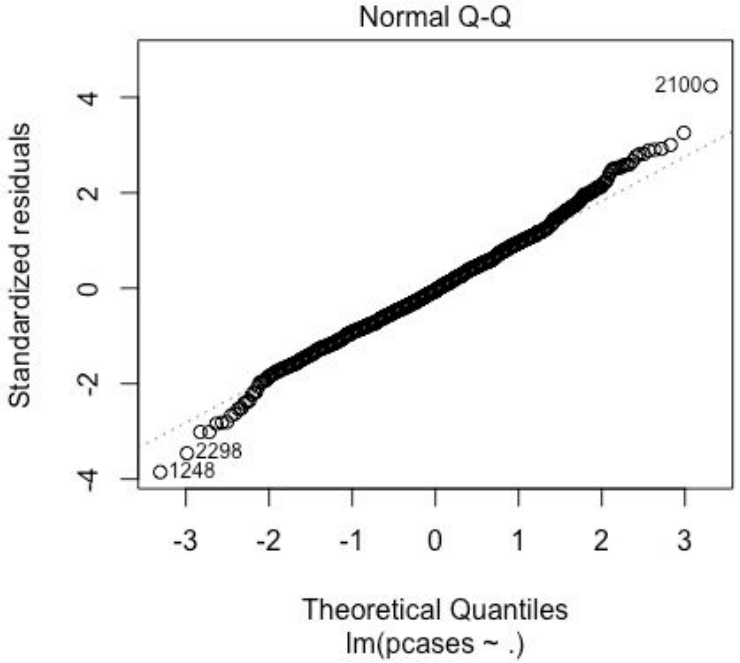
Checking Assumptions of Linearity and Error Constancy



There is no trend in the residuals and the red line looks almost horizontal. Therefore, the assumption of linearity is satisfied.

There is no fan shape as fitted values increase, so the assumption of constant error variance is met.

Checking Assumptions of Normality



Almost all the points lie on the qq line, which shows that the assumption of normality is also satisfied.

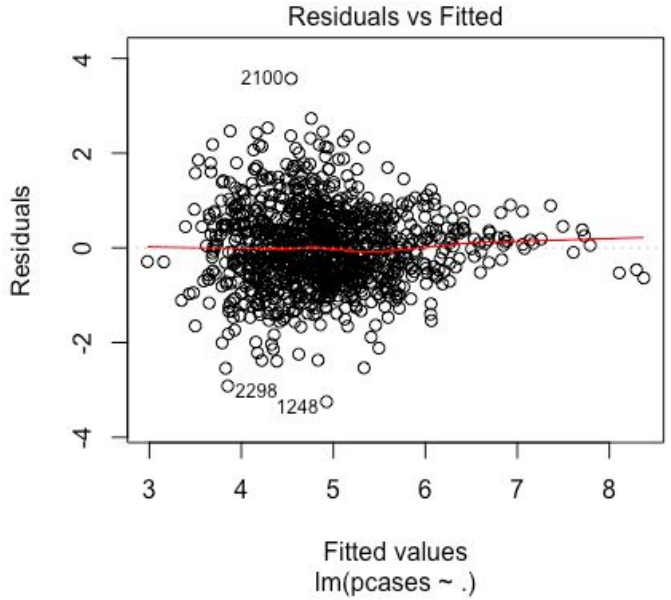
Correlation Matrix

	temp_mar20	dist	black_pct	hisp_pct	transit_mode share	pcases
temp_mar20	1.0000	0.15189	0.57607	0.30487	-0.11203	0.09174
dist		1.0000	-0.19816	0.24721	-0.23453	-0.42004
black_pct			1.0000	0.17981	0.25850	0.4300
hisp_pct				1.0000	0.31511	0.25706
transit_mode share					1.0000	0.41054
pcases						1.0000

All values are less than 0.7, so there is no multicollinearity.

Regression assumptions & NCV test

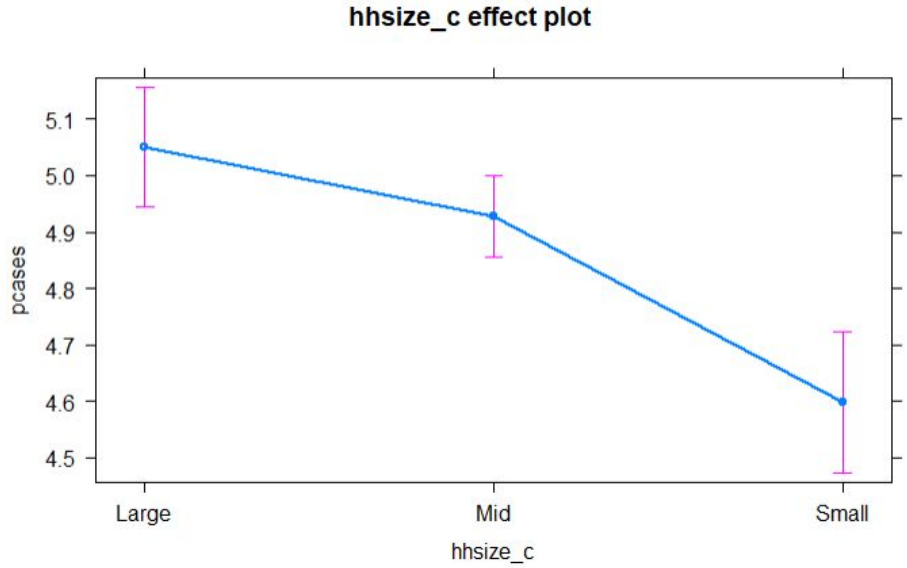
Non-constant Variance Score Test:
Variance formula: ~ fitted.values
Chisquare = 42.71812, Df = 1, p = 6.3224e-11



From NCV test, the p-value is almost equal zero, so we reject the null that the constant variance of error is satisfied.

Although we rejected the null hypothesis from ncv, the residuals plot suggests that residuals are evenly distributed, so we conclude that the residuals follow a constant variance.

MLR effect plot for the categorical variable



As average household size gets larger, COVID-19 cases per capita increases.

Checking VIF for Multicollinearity

	VIF	Df
temp_mar20	2.054606	1
dist	1.293628	1
black_pct	1.972575	1
hisp_pct	1.748292	1
transit_modeshare	1.507300	1
hhsize_c	1.403021	2

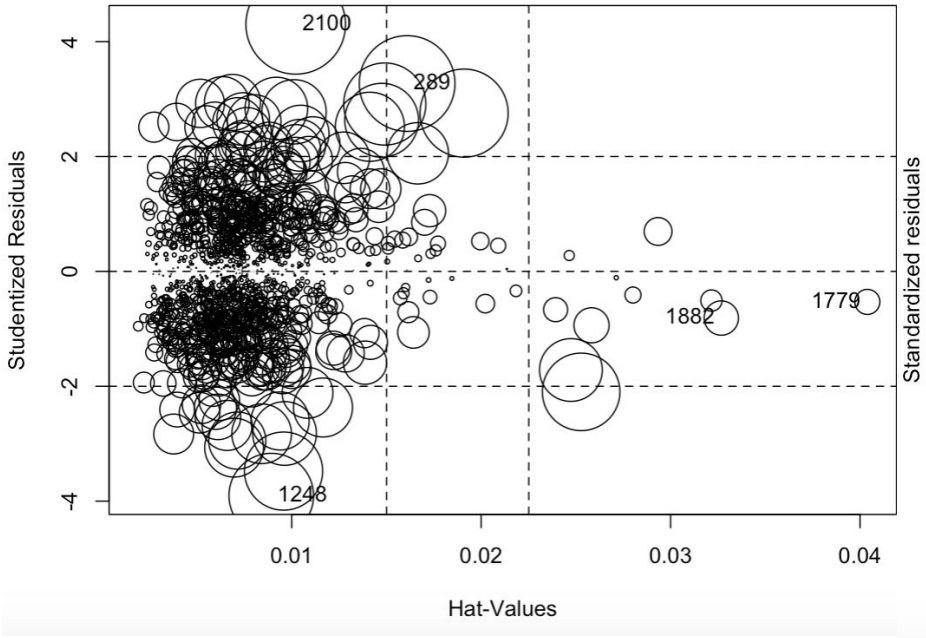
All VIF values are less than 10, so we are not overfitting the data.

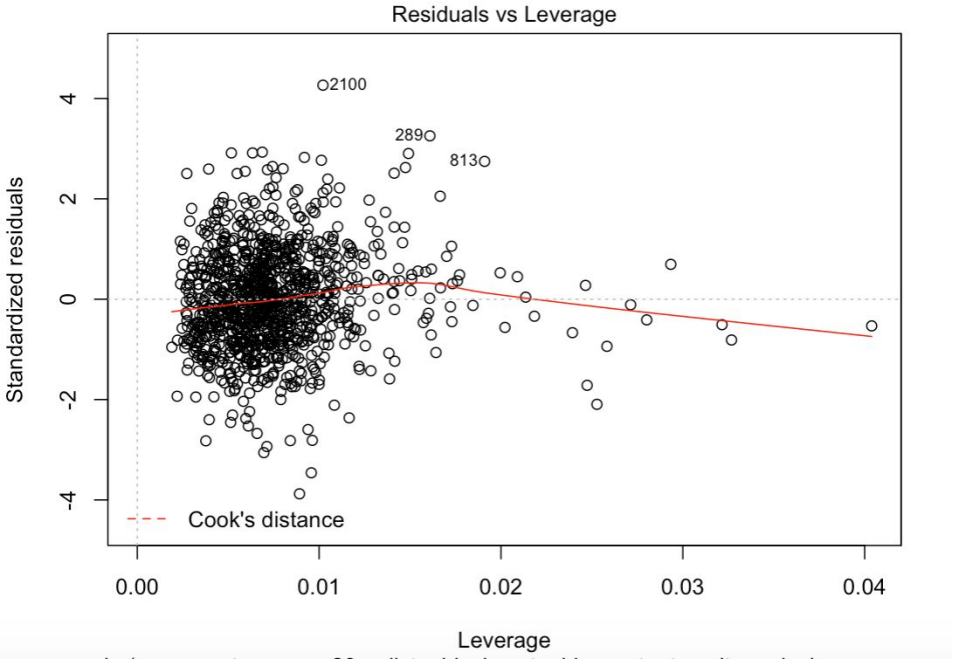
Cross Validation (6 Fold)

Variables	R^2 (w/out CV)	R^2 (CV)	RMSE (CV)
temp_mar20	0.007484	0.00609	1.084929
temp_mar20 + black_pct	0.2196	0.2133	0.9534859
temp_mar20 + black_pct + hisp_pct	0.2757	0.2792	0.9353828
temp_mar20 + black_pct + hisp_pct + transit_modeshare	0.3081	0.3111	0.9138431
temp_mar20 + black_pct + hisp_pct + transit_modeshare + hhsize_c	0.3252	0.3236	0.8956731
temp_mar20 + black_pct + hisp_pct + transit_modeshare + hhsize_c + dist	0.4301	0.4324	0.829038

RMSE consistently declines and log(pcases) ranges from 1 to 8, so a final RMSE of 0.8 is not alarming. The R^2 with and without cross-validation are very similar.

Performance of Model Leverage





Points 2100 and 1248 has a high studentized residuals, so we need to inspect in detail later.

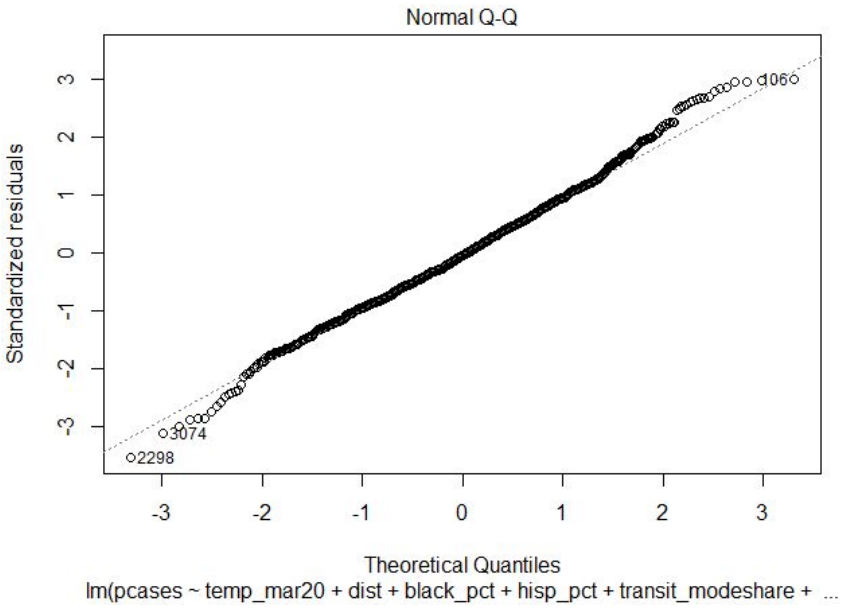
Influential plots

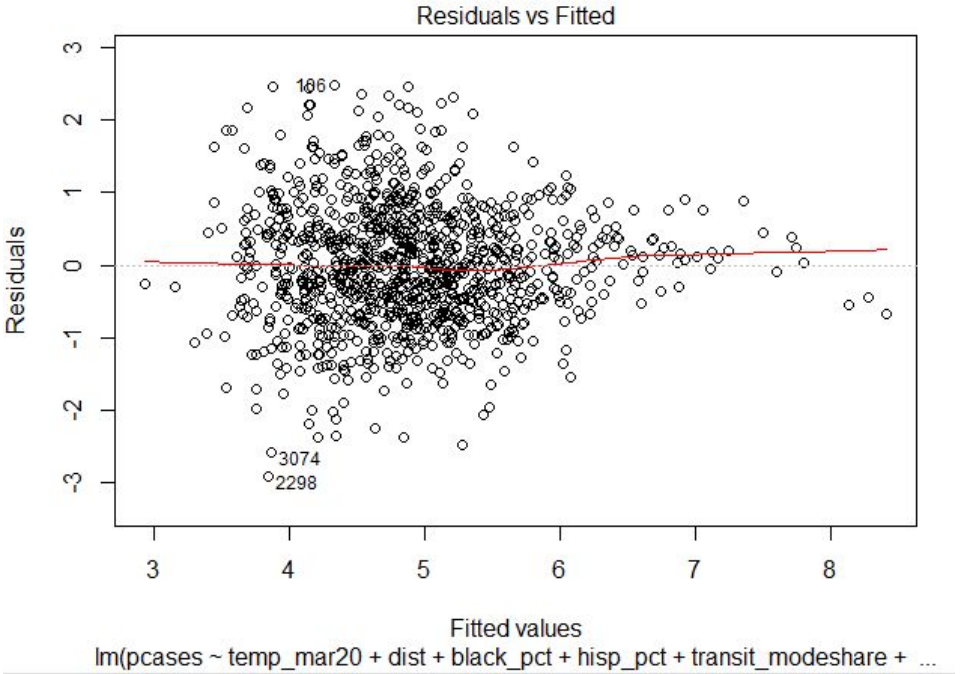
	Studentized residuals	Hat
289	3.2709	0.0161
1248	-3.904	0.0089
1779	-0.5312	0.0403
1882	-0.8123	0.0327
2100	4.3008	0.0102

h = 4 / 1066 = 0.003752345

Points 289, 1248, and 2100 are outliers obviously because their hat values are greater than 0.0038 and studentized residuals outside the -3 and +3 interval.

Final MLR (outliers dropped)





We have noticed that the points at the edges become closer to the qq-line after removing outliers, and the assumptions are still satisfied.

Final MLR (outliers dropped)

After removing outliers, we can observe that the R^2 value has improved. (0.4339 to 0.4458)

This model still meets the assumptions of regression.

Predictors	Coefficient	Standard Error	t_value	p-value
constant	10.718	0.266	40.29	< 2e-16
temp_mar20	-0.01	0.003	-4.251	2.32e-05
dist	-0.427	0.030	-14.161	< 2e-16
black_pct	0.279	0.027	10.321	< 2e-16
hisp_pct	0.260	0.035	7.376	3.29e-13
transit_modeshare	0.130	0.024	5.509	4.54e-08
hhsize_c Mid	-0.112	0.065	-1.713	0.0869
hhsize_c Small	-0.444	0.088	-5.067	4.76e-07

F value = 121.3 , R-squared = 0.4458, Adjusted R-squared = 0.4422

Interpretation of coefficients

Keeping all other predictors constant, on average...

1. a 1% increase in the distance from Manhattan leads to 0.43% decrease in the infection rates
2. a 1% increase in the percentage of black residents and hispanic residents leads to 0.28% and 0.26% increase in the infection rates respectively .
3. a 1% increase in the percentage of county residents who commute by transit will lead to 0.13% increase in the infection rate in a county.
4. a 1% increase in the temperature will lead to a 0.01% decrease in the infection rates.

Predictors	Coefficient	Standard Error	t_value	p-value
constant	10.718	0.266	40.29	< 2e-16
temp_mar20	-0.01	0.003	-4.251	2.32e-05
dist	-0.427	0.030	-14.161	< 2e-16
black_pct	0.279	0.027	10.321	< 2e-16
hisp_pct	0.260	0.035	7.376	3.29e-13
transit_modeshare	0.130	0.024	5.509	4.54e-08
hhsize_cMid	-0.112	0.065	-1.713	0.0869
hhsize_cSmall	-0.444	0.088	-5.067	4.76e-07

F value = 121.3 , R-squared = 0.4458, Adjusted R-squared = 0.4422

Interpretation of coefficients (continued)

Keeping all else constant…

1. On average, the infection rate in counties with median average household size is 0.11% lower than those with large average household size.
2. On average, the case rate in counties with small average household size is 0.44% lower than those with large average household size.
3. Based on the model, approximately 44.6% of the variation in infection rates is explained by the temp, distance, black pct, hispanic pct, transit_modeshare, and hhsize.

Predictors	Coefficient	Standard Error	t_value	p-value
constant	10.718	0.266	40.29	< 2e-16
temp_mar20	-0.01	0.003	-4.251	2.32e-05
dist	-0.427	0.030	-14.161	< 2e-16
black_pct	0.279	0.027	10.321	< 2e-16
hisp_pct	0.260	0.035	7.376	3.29e-13
transit_modeshare	0.130	0.024	5.509	4.54e-08
hhsize_cMid	-0.112	0.065	-1.713	0.0869
hhsize_cSmall	-0.444	0.088	-5.067	4.76e-07

F value = 121.3 , R-squared = 0.4458, Adjusted R-squared = 0.4422

Conclusions

Based on our in-depth exploratory data analysis, we eventually found the variables temp_mar20, dist, black_pct, hisp_pct, transit_modeshare, and hhsize to be the most statistically significant variables that contribute to the prediction of the county-level infection rates of COVID-19 (pcases) across the United States. Statistically, this is true because after we do the forward selection, these variables have p-value less than 0.05 so are statistically significant. Logically, these variables also make sense. If a county had higher average temperatures for March, county residents may not have socially distanced as well as states in which the average temperature was much colder. Counties with higher minority populations may not have as ample affordable healthcare as richer counties with lower minority populations. If a county has a higher percentage of its population that uses public transportation, there is a higher chance of COVID-19 being spread amongst patrons in crowded buses and trains. If a county has a higher average household size, there is more likely to be an increased spread of COVID-19 within individual households than counties with lower average household sizes.

Conclusions (Continued)

Among all the predictors that we use in the final model, the **dist variable (distance from the centroid of a given county to the centroid Manhattan)** is the most important factor in predicting the infection rate. This makes sense since as of the latest update to the data, May 13th, New York was the largest hotspot for COVID-19 infections. Living in a county closer to the largest hotspot in the U.S. would increase your chance of infection as compared to living in a county much further away. However, if we had current data, the dist variable may no longer be the strongest predictor, as the COVID-19 pandemic has changed quite a bit since May 13th. There is also a chance that the dist variable might remain as the strongest predictor, but the variable might need to be modified to instead highlight the distance from a different centroid, like Los Angeles or Phoenix, that is a much more current COVID-19 hotspot.

Conclusions (Continued)

While the age60plus, wfh_share, and italy_born variables were seen as potentially statistically significant in our backwards elimination model, these variables are actually not good predictors because they turned out to be statistically insignificant in our final multiple linear regression model and could lead to the possible overfitting of the model. This could be due to the fact that the variable's value is already represented by other predictors in the model (multicollinearity). For example, the percentage of those who work from home could already be represented in the percentage of county residents who do not take public transportation as their main method of transit since they do not need to travel for work or since jobs that offer the option to work from home tend to be higher paying, these county residents may use a different mode of transportation such as their own car.

Logistic Model

Question: Are the odds of a county being severely impacted by COVID-19 related to its average household size and its distance from NYC areas?

In the logistic model, the output variable “pcases” was made into a binary variable, which equals to 1 if the value is greater than or equal to the median, and equals to 0 if the value is smaller than the median.

For the independent variables, the numerical variable “dist” and the categorical variable “hhsize” are utilized. Instead of three levels, hhsize is categorized into two levels this time, “small” and “large”, with the median being the cutoff point, because that can make hhsize variable become statistically significant in the logistics model.

Logistics model

Predictors	Coefficients	Standard Error	Z value	p-value (>|z|)
constant	1.395	0.309e-02	14.98	<2e-16
hhsizeSmall	-8.570e-01	7.624e-02	-11.24	<2e-16
dist	-9.280e-04	6.814e-05	-13.62	<2e-16

Null deviance: 4341.9 on 3131 degrees of freedom
Residual deviance: 4030.8 on 3129 degrees of freedom
AIC: 4036.8
Number of Fisher Scoring iterations: 4

Logistic Model (Continued)

Exponentiated Coefficients with Confidence Intervals

Predictors	Coefficients	2.5%	97.5%
constant	4.0344	3.3669	4.8500
hhsizeSmall	0.4245	0.3653	0.4926
dist	0.9991	0.9989	0.9992

Keeping all else constant…

1. The odds of a county being severely-impacted by COVID-19 is 58% less likely for counties with small average household size than those with large household size and this is statistically significant.
2. Since exp(-9.28e-4*100) = 0.9113, for every 100 unit increases in the distance to centroid Manhattan, the odds of a county being severely affected by COVID-19 decreases by 8.9%.
3. The difference between Null deviance and Residual Deviance tells us that the model is a good fit.

Shortcomings & Improvements

The COVID-19 dataset was last updated May 13th. With improved testing and knowledge about COVID-19 since May 13th, we might see different variables predict case rate better than our current model and we would have much more accurate case rates at the county level. The data was also quite skewed because of the lack of nationwide testing at the time of the latest update. In addition, the dataset included variables such as italy_born that no longer reflect the current state of the pandemic, leaving room for the addition of new variables.

The COVID-19 dataset also made use of 2018 U.S. Census data. While this is the most up-to-date county-level data there is, a lot can happen in two years and we could potentially see major transformations and shifts within the data, affecting their fit as a good, statistically significant predictor. For example, the 2018 census

Shortcomings & Improvements (Continued)

data for the percentage of county-residents who work from home (wfh_share) does not match current percentage of county-residents who work from home, which has changed quite drastically in the last few months due to the pandemic.

To improve the dataset, we would want to include further demographic data that would give further insight into infection and death rates. For example, including data about if/when a county implemented a lockdown, the average percentage of county residents that regularly wear facemasks, and what percentage of the population is immunocompromised would all be medical factors that could greatly improve the data and give a more holistic approach to studying predictors for county-level COVID-19 case and death rates.

Appendix - A (Creating Categorical Var)

Creating Categorical Variable

```
#creating categorical variable
summary(covid$hhsize)
hhsize_c <- character(nrow(covid))

for (i in 1:nrow(covid)) {
  if (!is.na(covid$hhsize[i])) {
    if (covid$hhsize[i] >= 2.63) {
      hhsize_c[i] <- "Large"
    } else if (covid$hhsize[i] > 2.35 && covid$hhsize[i] < 2.63) {
      hhsize_c[i] <- "Mid"
    }
    else if (covid$hhsize[i] <= 2.35 ) {
      hhsize_c[i] <- "Small"
    }
  }
}

hhsize_c <- as.factor(hhsize_c)
table(hhsize_c)
covid$hhsize_c <- hhsize_c
```

Appendix - B

Backwards and Forwards Model Process:

```
#backwards elimination
#install.packages('leaps')
library(leaps)
b <- regsubsets(x = pcases ~ . + age60plus*hhsize_c, data = covid, y = covid$pcases, method= "backward")
summary(b)

#forward model
f <- regsubsets(x = pcases ~ . + age60plus*hhsize_c, data = covid, y = covid$pcases, method= "forward")
summary(f)
```

Model Creation (after looking at symbox and moving forward with log-transformed data)

```
#creating backwards-elim mlr model
modelb = lm(formula = pcases ~ . + age60plus*hhsize_c, data = covidb)
#need to remove all non statistically-significant data
modelb = lm(formula = pcases ~ temp_mar20 + dist + black_pct + hisp_pct + transit_modeshare + age60plus*hhsize_c, data = covidb)
```

```
#creating forwards-selection mlr model
modelf = lm(formula = pcases ~ . + age60plus*hhsize_c, data = covidf)
#need to remove all non-statistically significant data
modelf = lm(formula = pcases ~ temp_mar20 + dist + black_pct + hisp_pct + transit_modeshare + hhsize_c, data = covidf)
```

Appendix - C (Logistic Model code)

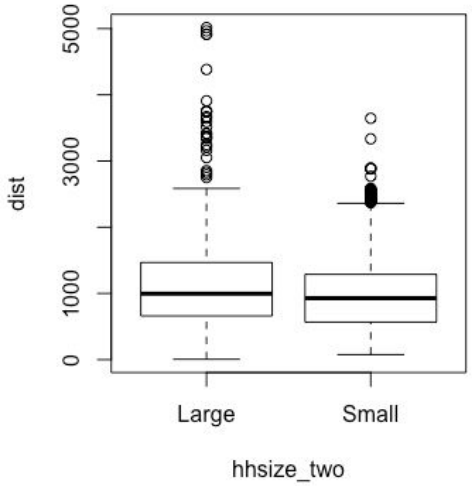
```r
#Split hhsize into "Small" and "large"
covid$hhsize_two <- NA
for (i in 1:nrow(covid)) {
  if (!is.na(covid$hhsize[i])) {
    if (covid$hhsize[i] <= 2.484) {
      covid$hhsize_two[i] <- "Small"
    } else if (covid$hhsize[i] > 2.484) {
      covid$hhsize_two[i] <- "Large"
    }

  }
}
logit_final <- as.data.frame(pcases_data)
logit_final$hhsize_two <- covid$hhsize_two
logit_final$dist <- covid$dist
colnames(logit_final) <- c("pcases","hhsize","dist")
logit_final <- na.omit(logit_final)

m1<-glm(pcases~hhsize+dist,data= logit_final,family='binomial')
summary(m1)
```

```r
#Boxplot of Distance VS hhsize
boxplot(dist~hhsize_two,data= covid, main = "Boxplot of Distance VS hhsize")

#Exponentiate Coefficients of the Logistic Model
round(exp(cbind(Estimate = coef(m1),confint(m1))),4)
```



Appendix D

Cross- Validation (covidf is dataframe with only significant predictors)

```
#CROSS VALIDATION STEP:
#CREATING TRAINING SET AND TEST SET
######
# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
split = sample.split(covidf$pcases, SplitRatio = 0.7)
training_set = subset(covidf, split == TRUE)
test_set = subset(covidf, split == FALSE)

modelf = lm(formula = pcases ~ ., data = training_set)
#revised modelb on just the training_set
y_pred = predict(modelf, newdata = test_set)
cor(y_pred, test_set$pcases)
#Correlation between pcases computed on the base of the training model and the actual pcases in the testing sample is pretty high
test_set$pred <- y_pred
test_set$real <- testset$pcases
#final plot
plot(modelf)
```

Appendix E

Interaction Variable: Code and Plots
(11, 12, 25) are the column #'s of pcases, age60plus and hhsize_c

```
interaction <- covid[, c(11, 12, 25)]
interaction$pcases <- log(interaction$pcases)
interaction$age60plus <- log(interaction$age60plus)
library(IDPmisc)
interaction <- NaRV.omit(interaction)
mint = lm(pcases~age60plus+hhsize_c+age60plus*hhsize_c, data = interaction)

library(effects)
plot(allEffects(mint), ask=FALSE,
     main = "age60plus X hhsize (categorical) Interaction Plot")

interact = aov(pcases~age60plus+hhsize_c+age60plus*hhsize_c, data = covid)
summary(interact)
```