Efficacy of Mask Usage Adherence

I. Introduction

More than one year into the SARS-CoV-2 pandemic, research on the association between mask usage adherence and COVID-19 cases is scarce. Due to ethical reasons, researchers could not carry out COVID-19 studies using randomized controlled trials, so research was limited to observational data (Howard et al., 2020). Previous studies have investigated the effects of statewide face mask mandates on the development of COVID-19 cases in the respective states. A study by Lyu and Wehby (2020) found significant declines in COVID-19 case growth rates from 1 to 20 days after state mandates to wear masks in public are signed into law. The researchers used an event study from April to May of 2020, but were limited by their inability to measure public compliance to the orders. Another study by Fischer et al. (2021) also explored the role of public mask mandates in reducing the spread of COVID-19. This study found that states with the lowest mask adherence rates yielded the greatest percentages of high COVID-19 rates, and vice versa. The researchers concluded that states in which at least 75% of their population adhered to mask mandates tended to be associated with lower COVID-19 rates in the subsequent month.

The previously mentioned studies, while providing insight into the relationship between state law and COVID-19 contraction rates, did not account for human interest and behavioral decisions. Over the course of the COVID-19 pandemic, the issue of mask usage has become an increasingly politicized issue, with some viewing it as a tradeoff between public health and personal freedom (Fischer et al., 2021). This study serves to investigate the effect of actual, not simply theoretical or desirable, human behavior regarding mask adherence on cumulative COVID-19 cases in the United States at the county level.

II. Statistical Question

Is there significant evidence of a nonzero linear association between consistent mask usage adherence and COVID-19 cases in US counties?

Hypotheses	Definitions
H_0: $\beta = 0$	β: the slope of the fitted linear regression line relating consistent
H_a: $\beta \neq 0$	mask usage and COVID-19 cases in US counties

III. Data Collection

This study uses two open access datasets approved by the New York Times. The first dataset contains the results of online interviews conducted online by the survey firm Dynata. Surveyors asked the question, “How often do you wear a mask in public when you expect to be within six feet of another person?”, and respondents answered with one of the following categories: “never,” “rarely,” “sometimes,” “frequently,” and “always.” The survey firm then calculated the proportions of respondents in each US county that answered each category. The survey was conducted between July 2 and July 14, 2020, during which 250,000 responses were collected. The data was also weighted by age and gender, and respondents were categorized to counties according to their residency zip codes. The survey firm calculated the weighted average of the 200 nearest responses for each county, with a greater weight for closer responses.

The second dataset utilized in this study contains cumulative COVID-19 case counts for each county in the United States and is updated daily. For the purposes of this study, we extracted the cumulative case counts reported for July 2, 2020, since this would align with the first dataset. The reason for choosing this particular date was because cumulative COVID-19 cases could

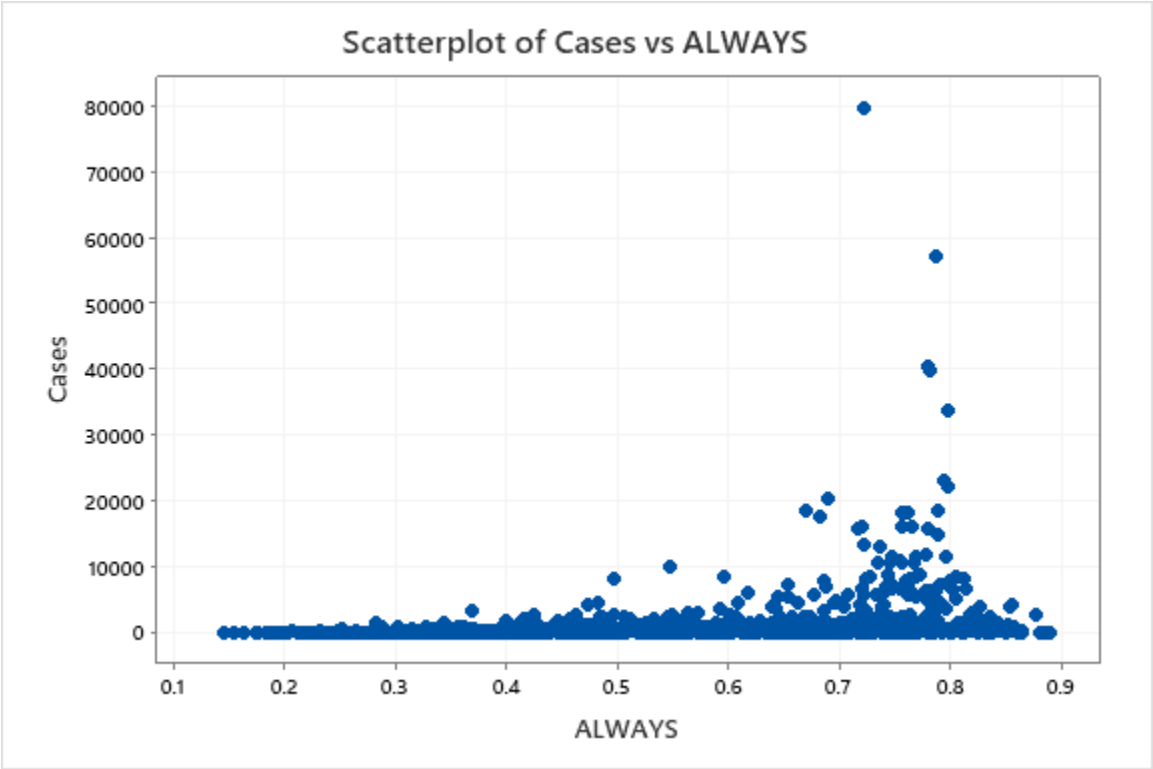
increase substantially over a 12-day period, while mask-wearing behavior was not likely to change much during this time. The number of COVID-19 cases include both confirmed cases, individuals whose COVID-19 contractions were tested in a laboratory and reported by an official government agency, and probable cases, individuals who did not have a confirmed test but were likely infected according to public health official evaluation. The database noted that this data could contain imperfections due to various sources of possible error: delays and faults in the American healthcare system and inaccurate reports.

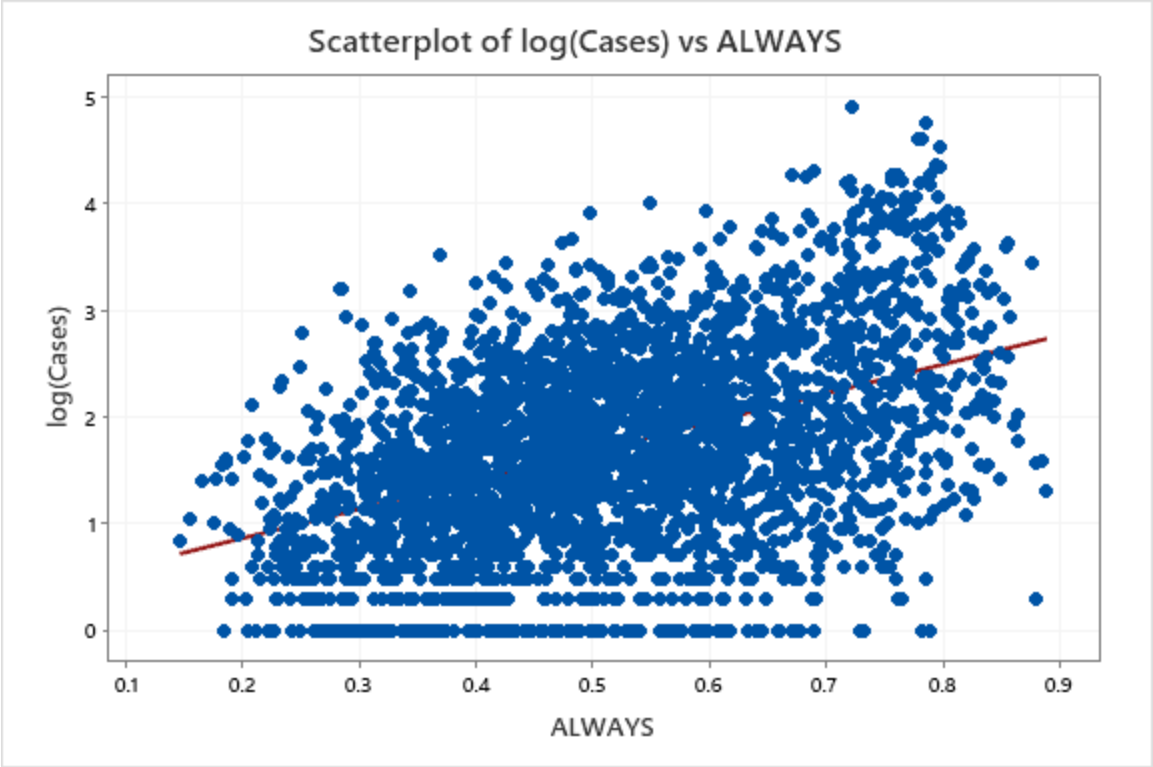
In order to fully utilize both datasets to understand the relation between estimated prevalence of mask-wearing and the number of COVID-19 cases in US counties, we used the R programming language to merge the two datasets together by county FIPS codes. After evaluating patterns in the raw data, we decided to calculate the logarithm of COVID-19 cases in order to perform a logarithmic transformation of our data.

IV. Data Display

	FIPS	ï..Date	County	State	Cases	Deaths	NEVER	RARELY	SOMETIMES	FREQUENTLY	ALWAYS	log(Cases)
1	1001	6/2/2020	Autauga	Alabama	240	5	0.053	0.074	0.134	0.295	0.444	2.38021
2	1003	6/2/2020	Baldwin	Alabama	308	9	0.083	0.059	0.098	0.323	0.436	2.48855
3	1005	6/2/2020	Barbour	Alabama	176	1	0.067	0.121	0.120	0.201	0.491	2.24551
4	1007	6/2/2020	Bibb	Alabama	79	1	0.020	0.034	0.096	0.278	0.572	1.89763
5	1009	6/2/2020	Blount	Alabama	65	1	0.053	0.114	0.180	0.194	0.459	1.81291
6	1011	6/2/2020	Bullock	Alabama	214	6	0.031	0.040	0.144	0.286	0.500	2.33041
7	1013	6/2/2020	Butler	Alabama	418	18	0.102	0.053	0.257	0.137	0.451	2.62118
8	1015	6/2/2020	Calhoun	Alabama	173	3	0.152	0.108	0.130	0.167	0.442	2.23805
9	1017	6/2/2020	Chambers	Alabama	370	26	0.117	0.037	0.150	0.136	0.560	2.56820
10	1019	6/2/2020	Cherokee	Alabama	39	3	0.135	0.027	0.161	0.158	0.520	1.59106
11	1021	6/2/2020	Chilton	Alabama	108	1	0.060	0.070	0.058	0.194	0.618	2.03342
12	1023	6/2/2020	Choctaw	Alabama	154	10	0.049	0.038	0.126	0.219	0.568	2.18752
13	1025	6/2/2020	Clarke	Alabama	149	2	0.049	0.088	0.164	0.268	0.430	2.17319
14	1027	6/2/2020	Clay	Alabama	28	2	0.148	0.158	0.195	0.169	0.329	1.44716
15	1029	6/2/2020	Cleburne	Alabama	15	1	0.151	0.125	0.138	0.217	0.368	1.17609
16	1031	6/2/2020	Coffee	Alabama	241	1	0.101	0.152	0.094	0.186	0.466	2.38202
17	1033	6/2/2020	Colbert	Alabama	190	2	0.082	0.096	0.152	0.159	0.510	2.27875
18	1035	6/2/2020	Conecuh	Alabama	41	1	0.099	0.052	0.259	0.192	0.399	1.61278

	FIPS	ï..Date	County	State	Cases	Deaths	NEVER	RARELY	SOMETIMES	FREQUENTLY	ALWAYS	log(Cases)
2952	56013	6/2/2020	Fremont	Wyoming	283	8	0.160	0.062	0.060	0.294	0.424	2.45179
2953	56015	6/2/2020	Goshen	Wyoming	5	0	0.201	0.169	0.111	0.223	0.296	0.69897
2954	56017	6/2/2020	Hot Sp>>	Wyoming	13	0	0.208	0.093	0.068	0.307	0.324	1.11394
2955	56019	6/2/2020	Johnson	Wyoming	18	1	0.183	0.208	0.103	0.216	0.290	1.25527
2956	56021	6/2/2020	Laramie	Wyoming	188	2	0.143	0.127	0.100	0.221	0.409	2.27416
2957	56023	6/2/2020	Lincoln	Wyoming	15	0	0.131	0.183	0.128	0.235	0.324	1.17609
2958	56025	6/2/2020	Natrona	Wyoming	79	1	0.100	0.084	0.094	0.325	0.398	1.89763
2959	56027	6/2/2020	Niobrara	Wyoming	2	0	0.169	0.191	0.177	0.222	0.241	0.30103
2960	56029	6/2/2020	Park	Wyoming	2	0	0.189	0.153	0.191	0.205	0.262	0.30103
2961	56031	6/2/2020	Platte	Wyoming	1	0	0.149	0.149	0.123	0.160	0.418	0.00000
2962	56033	6/2/2020	Sheridan	Wyoming	16	0	0.170	0.251	0.099	0.203	0.278	1.20412
2963	56035	6/2/2020	Sublette	Wyoming	3	0	0.223	0.111	0.061	0.231	0.374	0.47712
2964	56037	6/2/2020	Sweetw>>	Wyoming	30	0	0.061	0.295	0.230	0.146	0.268	1.47712
2965	56039	6/2/2020	Teton	Wyoming	100	1	0.095	0.157	0.160	0.247	0.340	2.00000
2966	56041	6/2/2020	Uinta	Wyoming	13	0	0.098	0.278	0.154	0.207	0.264	1.11394
2967	56043	6/2/2020	Washakie	Wyoming	36	3	0.204	0.155	0.069	0.285	0.287	1.55630
2968	56045	6/2/2020	Weston	Wyoming	1	0	0.142	0.129	0.148	0.207	0.374	0.00000





Regression Equation

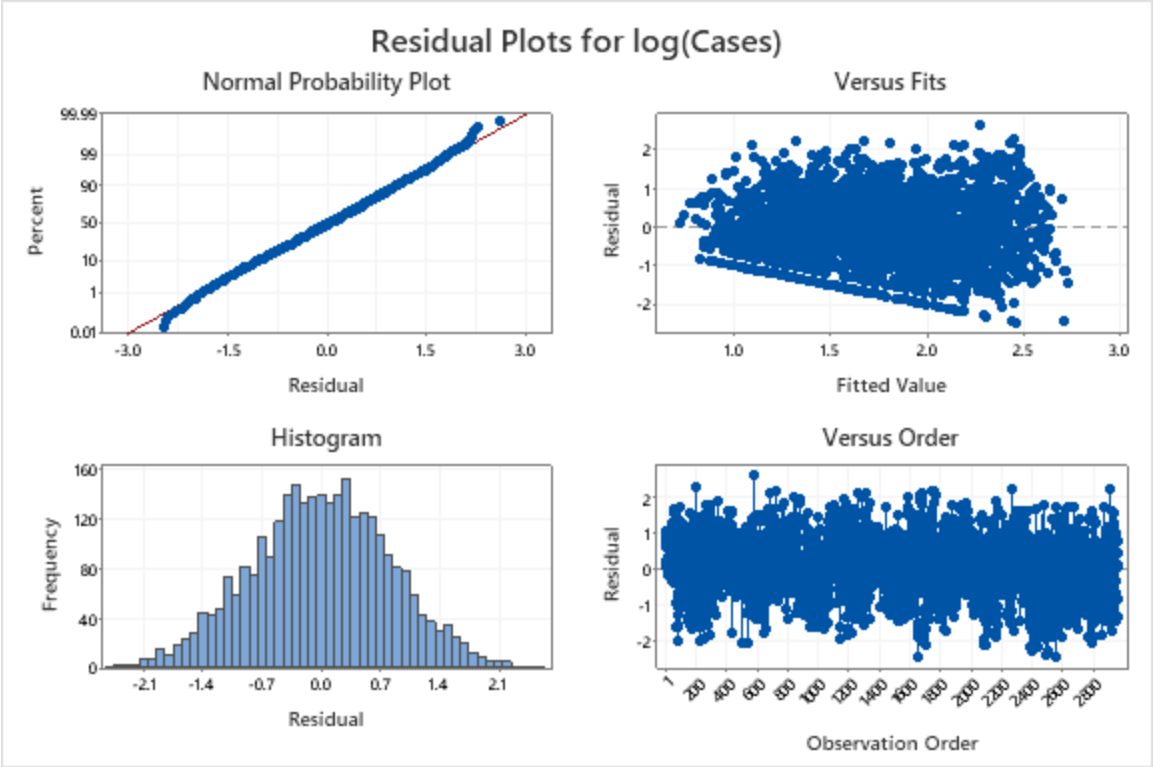
log(Cases) = 0.3240 + 2.7118 ALWAYS

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.814688	19.93%	19.91%	19.82%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0.3240	0.0534	(0.2192, 0.4287)	6.07	0.000	
ALWAYS	2.7118	0.0998	(2.5161, 2.9075)	27.17	0.000	1.00



V. Data Analysis

The scatter plot of cumulative COVID-19 cases count against the proportion of respondents that answered “ALWAYS” showed very strong curvature. Therefore, we performed a logarithmic transformation so that we could appropriately apply inference on linear regression. After taking the logarithm of cumulative case counts and plotting those values against consistent mask use proportions, the new scatter plot indicated better adherence to a linear model. The new scatter plot displays a moderate positive linear association between the two variables (r=0.4464). The R^2 value suggests that 19.93% of the variation in the logarithm of cumulative cases count can be explained by the consistent mask usage proportion. According to the linear regression equation, a one unit increase in the proportion of those who responded “ALWAYS” corresponds to a 2.7118 unit increase in the logarithm of cumulative cases count, on average. A few points

can be considered outliers due to their abnormally large residuals, specifically in the bottom right region of the plot. If the conditions are met, we can conduct a t-test for the slope of a population regression line.

Conditions for Inference:

1. Linear: The scatter plot of the logarithm of cumulative cases count against the proportion of respondents that answered "ALWAYS" roughly follows a linear pattern. The residuals vs. fitted values plot shows no obvious pattern, and the points are randomly scattered around the residual=0 line.
2. Independent: Individual observations are independent of each other. One county's proportions and cumulative cases count do not affect another county's proportions and cumulative cases count.
3. Normal: The normal probability plot of residuals follows approximately a straight line. In addition, the histogram of frequency vs. residuals is single-peaked and mound shaped.
4. Equal variance: The residuals vs. fitted values plot shows slight heteroscedasticity, although not severe.
5. Random: No information about random sampling for the survey was published, but data was collected for each county in the US so random sampling on this front is not necessary. We will proceed with caution and take these factors into consideration when determining our final conclusion.

Inference Summary Values:

$\alpha = 0.05$	$df = n-1 = 2267$	$R^2 = 19.93\%$
$t = 27.17$	$p = 0$	$r = 0.4464$
$b = 2.7188$	$SE_b = 0.0998$	$S = 0.814688$

VI. Conclusion

Since p=0 which is less than α=0.05, we reject the null hypothesis. We have sufficient evidence to conclude that there is a nonzero linear relationship between the consistent mask usage proportion and the logarithm of COVID-19 cases in US counties.

Confidence intervals provide more information about this relationship. We are 95% confident that the interval from 2.5161 to 2.9075 captures the true slope of the population regression line relating the consistent mask usage proportion and the logarithm of COVID-19 cases in US counties. This interval is entirely greater than 0, which suggests a positive linear relationship between these variables. Since association does not imply causation, we cannot conclude that a greater proportion of consistent mask adherence causes more COVID-19 cases.

VII. Reflection

The results found in this study were very surprising, since a positive association is counterintuitive when discussing mask usage adherence and COVID-19 case counts. This could have occurred due to a few reasons. Firstly, surveys do not guarantee full honesty, and the data source did not indicate whether the surveys ensured anonymity. If more people answered "always" than the true proportion, the data would contain bias and could therefore lead to a false positive association. Secondly, despite our best efforts to align the two datasets, there could be

slight discrepancies within and between the mask usage data and the COVID-19 cumulative cases count data. For the mask usage data, the surveyors calculated proportion estimates for each US county, but these estimates might not be accurate. The COVID-19 cases data could include false positive tests as well as miscounted cases, which would lead to further error. Finally, we matched up the two datasets using the common date of July 2, 2020, but this choice might not have been truly representative of the overall trends and daily variations of mask usage and COVID-19 cases count.

Overall, the study went well, and we were able to effectively analyze the data available. One issue we ran into was the heteroscedasticity in the residual plot. We experimented with further transformations, but there did not seem to be a combination of transformations that substantially improved the residual variance. We were also unable to find information on random sampling for the mask usage survey, but we do know that the researchers calculated county estimates for mask usage proportions, so hopefully they were representative of each county. We did have COVID-19 cumulative case counts for each county in the US, so we did not have to worry about random sampling regarding counties. From this study, we learned an immense amount on statistical techniques and the importance of evaluating conclusions in context.

For further study, we would recommend employing random sampling so that conclusions could be generalized to the general population. We did not have the resources or means to carry out the study by collecting our own data, but data firms should give respondents the assurance of anonymity to ensure honest responses and minimize bias. Moreover, future research should utilize more advanced technology to acquire a more appropriate transformation of data to eliminate the heteroscedasticity apparent in the residual plots.

VIII. Works Cited

Fischer, C.B., Adrien, N., Silguero, J.J., Hopper, J.J., Chowdhury, A.I., & Werler, M.M. (2021). Mask adherence and rate of COVID-19 across the United States. *PLoS ONE 16*(4): e0249891. https://doi.org/10.1371/journal.pone.0249891

Howard, J., Huang, A., Li, Z., Tufekci, Z., Zdimal, V., van der Westhuizen, H., von Delft, A., Price, A., Fridman, L., Tang, L., Tang, V., Watson, G.L., Bax, C.E., Shaikh, R., Questier, F., Hernandez, D., Chu, L.F., Ramirez, C.M., & Rimoin, A.W. (2020). An evidence review of face masks against COVID-19. *Proceedings of the National Academy of Sciences 118*(4) e2014564118; DOI: 10.1073/pnas.2014564118

Lyu, W., & Wehby, G.L. (2020). Community use of face masks and COVID-19: Evidence from a natural experiment of state mandates in the US. *Health Affairs 39*(8), 1419-1425. doi: 10.1377/hlthaff.2020.00818