Işık Topçu
Prof. D. Carlson
CSSM 502
29 January 2022

How variant tweets affect mobility in the US?
PROJECT REPORT

In this project, I wanted to know if the number of tweets with the keywords in the United States: ["omicron", "omicron ny", "booster jab", "omicron variant", "booster", "omicron symptoms","third dose","3rd dose", "pfeizer", "omicron news","omicron deadly"] had an effect on mobility -especially walking- patterns of people. I have created a dataset between November 15,2021 - January 21, 2022 containing US keyword tweet counts for each day, cumulative death, administered booster shots and covid cases counts that I acquired from github[1]. I acquired the mobility data from Apple mobility[2], which contains daily mobility (walking, transit and driving) (base=100, Jan13,2020) for every state daily. I have selected all US states mobility data and took the mean of them, resulting in an average US walking, transit and driving data which I also plotted in my data notebook. I saw a downward slope in overall mobility. I have concatenated mobility, cases and vaccine data. I also added weekends dummy but later when I compared the tweet counts and walking mobility data (plotted in the model notebook) I realized that there is a pattern happening between mobility and tweet counts, so I thought it could be better if I made dummies for each day of the week. I performed box-cox on "walking" to conditionally normalize it, and I did a log transformation on all predictors and appended them to the original dataset.

Because my data wasn't a good fit for any GLM model, I performed OLS, trying out many different combinations of variables. I found an extremely low R-squared value and none of the variables were significant on the outcome (walking_box_cox). The condition number was large,

[1] https://raw.githubusercontent.com/nytimes/covid-19-data/master/us.csv
[2] https://covid19.apple.com/mobility

(1.11e+09) which might indicate that there is strong multicollinearity. Autocorrelation is relatively one of the most important problems of my data. The Durbin-Watson statistic of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 point to positive autocorrelation. In my case, the DW was 1.492 which meant positive autocorrelation which I think can happen in time series data. Adding weekday daily dummy variables might help with this issue and there are also other remedial procedures like Cochrane-Orcutt Procedure. For autocorrelation, I might also add as an independent variable data from the last week for past-dependence. Secondly, there was no linearity because expectation of residuals are 0 but in my case it was -3.01029. For normality, Jacque-Bera p-value was 0.4677. The closer to zero is this test, it can be hypothesized that the data is more normally distributed. I didn't really look at the Shapiro-Wilk and K- Smirnov tests because they are less and less used and more used in non-parametric tests. My data was not normal according to the Andrew-Darling (0.9302, 5% critical value: 0.7440) but for such a small sample size, I thought non-normality could be ignored for OLS. Normality isn't really the biggest issue for this dataset, although I performed Box-Cox on the outcome to fix some of the issue. For homoscedasticity, Breusch-pagan test : reject null. (p-value = 0.01) There is heteroscedasticity problem. Goldfield quandt: do not reject null. (P-value = 0.96) There is no heteroscedasticity. These two tests are contradictory, but in our case, Breusch-pagan seems more reliable. Therefore, we also have a homoscedasticity problem. For multicollinearity, variation inflation factors were : [1.03551809, 10.31509241, 21.34613018, 12.2374227 , 1.24206276] we can either drop the ones above 10 or perform PCA. I performed PCA at the end which showed one component as ideal. By using just the first principal component, we can explain 57.45% of the variation in the response variable. We can see that the test RMSE turns out to be 41.6460. This is the average deviation between the predicted value for hp and the observed value for hp for the observations in the testing set. I also performed Poisson at the end but there wasn't any significance either. As advised, I know that my estimator is unstable, and the linear model needs to be improved. This is a time-series data, so maybe another relevant model could be used. My cases, death and booster data are cumulative. I think that non-cumulative data could provide better results because people's behavior can be affected by daily numbers better. All and all, an R-squared of

100% means that all movements of a security (or another dependent variable) are completely explained by movements in the index (or the independent variable(s) we are interested in. In our case, the R-squared was 0.109 which is so low. I'm not even sure of my model's fit to be sure about my R-squared significance. I hope to improve my model after submitting my project to better make use of this data. I think it would be better if we could go back in time to collect more tweets. There could also be other variables that affect human mobility behavior, my independent variables were limited. There could also be a reciprocal relationship between tweet count and mobility, people who see tweets might decrease movement but also people who decreased mobility can also increase tweeting.

OLS Regression Results

Dep. Variable:	walking_box_cox	**R-squared:**	0.109
Model:	OLS	**Adj. R-squared:**	0.030
Method:	Least Squares	**F-statistic:**	1.389
Date:	Sat, 29 Jan 2022	**Prob (F-statistic):**	0.242
Time:	13:42:04	**Log-Likelihood:**	-315.78
No. Observations:	63	**AIC:**	643.6
Df Residuals:	57	**BIC:**	656.4
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
weekend	-8.4754	11.040	-0.768	0.446	-30.582	13.631
cases_log	78.2352	433.279	0.181	0.857	-789.392	945.862
deaths_log	-1403.4397	3197.212	-0.439	0.662	-7805.746	4998.867
boosters_log	115.4248	256.451	0.450	0.654	-398.109	628.959
tweets	-1.742e-05	9.01e-05	-0.193	0.847	-0.000	0.000
constant	1.584e+04	3.15e+04	0.503	0.617	-4.72e+04	7.89e+04

Omnibus:	2.149	**Durbin-Watson:**	1.492
Prob(Omnibus):	0.341	**Jarque-Bera (JB):**	1.520
Skew:	0.035	**Prob(JB):**	0.468
Kurtosis:	3.758	**Cond. No.**	1.11e+09