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

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 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. I acquired the mobility data from Apple mobility, 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 a mean US walking, transit and driving data which I also plotted in my data notebook. I saw a downward slope in overall mobility. I also added weekends dummy but later when I compared the tweet counts and walking mobility data(plotted in the model notebook) there is a routine happening, 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. Then 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, 1.11e+09. This might indicate that there are strong multicollinearity or other numerical problems. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value 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 weekly dummy variables might help with this issue and there are also other remedial procedures like Cochrane-Orcutt Procedure. Secondly, there was no linearity because expectation of residuals are 0 but

in my case it was 3.6200. My data was on the line of “normal” according to the Andrew-Darling (0.7838, 5% critical value: 0.7440) but for such a small sample size, I thought non-normality could be ignored for OLS. For homoscedasticity, Breusch-pagan test : reject null. (P-value is smaller) There is heteroscedasticity problem. Coldfield quandt: do not reject H_null. 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 : [36.31531822, 1.0867786 , 10.86870258, 28.73421675, 23.91275499, 1.2541409] 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 64.72% 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 is also cumulative, I think that a non-cumulative data could provide better results. 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. 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.