International Journal of Mathematics Trends and Technology
ISSN: 2231 – 5373 /doi:10.14445/22315373/IJMTT-V66I11P504
Volume 66 Issue 11, 60-66, November 2020

A Statistical Learning Regression Model utilized to determine predictive factors of social distancing during COVID-19 pandemic

Timothy A. Smith[#1], Albert J. Boquet[*2], Matthew Chin[#3]

[1]*Professor & Program Coordinator of Data Science, Department of Mathematics*
[2]*Professor, Department of Human Factors*
[3]*Student, Computational Mathematics*

Embry Riddle Aeronautical University
[1] *Aerospace Boulevard, Daytona Beach FL 32114, United States of America*

Abstract *— In an application of the mathematical theory of statistics, predictive regression modelling can be used to determine if there is a trend to predict the response variable of social distancing in terms of multiple predictor input "predictor" variables. In this study the social distancing is measured as the percentage reduction in average mobility by GPS records, and the mathematical results obtained are interpreted to determine what factors drive that response. This study was done on county level data from the state of Florida during the COVID-19 pandemic, and it is found that the most deterministic predictors are county population density along with median income.*

Keywords *— Regression, statistical machine learning, theory of infectious disease transmission & control. AMS classification: 62P99.*

I. INTRODUCTION

Social distancing is a term applied to certain actions that are taken by Public Health officials to stop or slow down the spread of a contagion. In this study we will be focusing on the SARS-CoV-2 novel coronavirus "COVID-19" pandemic of 2020. Social distancing aids in slowing the spread by reducing the close-contact rate of a population. The application of social distancing would be to take measures to include limiting large groups of people coming together, closing buildings and cancelling events. And, it is well accepted that social distancing measures is the most effective way to stop an actively spreading virus; however, not much research has been conducted on what actually causes people to social distance. This is primarily due to the fact that a real world event occurring which requires the implementation of Social Distancing is so rare that previously very little data was available, especially less in modern day format with real time data collection.

In this research we develop a multivariable regression model with several input predictor variables $x_1,x_2,\ldots,x_n$ applied with the desire to predict a single output response variable, y. The response variable here will be a value which is used to measure how well individual counties within the state of Florida are socially distancing during the early stages of the 2020 Pandemic, namely the time period from late March to April 2020. This response variable recorded is the taken on the average mobility reduction, which is based on average daily distance travelled, obtained from GPS measurements of cell phone movement that is collected and organized by a private company [1]. Namely, the larger the percentage reduction a county has the better that county is at partaking in social distancing. Then, for input variables various factors of the county

were included with data available from public records such as: population and/or population density of the county, percentage of the population in various demographic, median income of the county, how many days early government offices business were closed prior to the state-wide order and/or local county level orders, how many days county or city orders were implemented prior to the sate-wide order, and a few other variables.

II. INITIAL STATISTICAL REGRESSION MODEL

Initially several predictor variables were subjectively selected to conduct a multivariable regression model which has several input predictor variables $x_1,x_{2,\ldots},x_n$ to predict the single output response variable y as the percentage average mobility reduction which we will take as our measurement of the amount of social distancing (e.g. the county that has the largest reduction in average distance travelled from the GPS measurements will be interpreted as the one that is practicing social distancing the most). The input variables were initially selected by a subjective process in relying on expertise and common sense to find what factors are thought of to steer the social distancing; the predictor variables utilized were the various demographics of the county, the population density of the county, the median salary of the county, and number of days early county level closures and/or orders were put in effect prior to the April 3rd state wide stay at home order. All of this data is publicly available data [2], and is available upon request. From this data a multiple regression model was constructed.

To develop this scheme we will outline the general procedure. To begin we assume the initial data set is kept with all possible predictors variables, but many additional data sets could be added into the predictor data matrix that "may" correlate and/or predict the value of the Social Distancing. Thus, a data matrix of $X = [x_1,x_2,x_{3,\ldots},x_n]$ is obtained. To begin we define the null model $\{S_0\}$ as the model which only contains the intercept, namely

$$y = \beta_0 .$$

We then apply a statistical learning process to seek the most mathematical pure model with the optimal set of predictor variables included. Namely, we will construct the first level set of models $\{S_1\}$ as the set of n models where each individual model is a simple linear regression model of the form,

$$y = \beta_0 + \beta_1 x_1 .$$

In this first level set of models only one predictor is utilized; moreover, these models are computed n times separately (i.e. x_1 is not our first predictor, rather just a default label used for notation). Each model individually takes from the predictor data matrix one column predictor variable, running through 1 to n. Then, for each model the R^2 along with the F Statistic is recorded along with all test stats for each predictor variable. Now, at this stage, depending on the number of possible

predictor variables available, one may only want to keep predictor variables that are showing some significance (a common cut off is to keep a predictor variable if its correlation to Y individually is, $R^2 > 0.6$) to be used for the next level set of models.

We will construct the second level set of models $\{S_{12}\}$ as the set of n*(n-1) models were each individual model is a simple linear regression model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where the model is computed n*(n-1) times separately. Each model individually takes into the predictor data matrix only two predictor variable, firstly running through 1 to n through the variable used in S_1, then removing that column from the matrix. Then adding the second variable running through 1 to n-1 through the variables remaining in the reduced matrix, and this is done for all possible combinations of the two variables. Again for each model the R^2 along with F statistic is recorded. And, we then continue this process until the n^{th} level set of models is obtained as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

For each of these models the R^2 value along with the F statistic & P value has been recorded, and we desire to find the best model that is not only mathematically sound (R^2), but is also strongly deterministic of our predicted data. Various methods can be applied to select the optimal regression model, but one must use caution if just looking to select the model with the highest R squared value as mathematically adding any variable to the data will increase the R squared but may not actually be adding much value to the model (e.g. it would really be just noise in the model for our interpretations, as our desire is primarily to determine which factors are driving the social distancing so really our goal is to find the most deterministic predictors rather than usual modelling to explain the most amount of variance in the response). Thus, for this scheme here the value looked to optimize was a linear combination of the model's R squared value plus it's P value (with a correcting factor to address concerns such as VIF). After obtaining a solution set for the best models from each level, a subroutine was added to exclude any predictors which have T Stats lower than 1.96. From this logic a code can be routinely be created, and the only remaining question would be the details of the linear optimization function utilized, and for simplification one can just think to use the function of the model's coefficient of determination plus a constant times the model's P value. In regards to coding, an even simpler approach could be taken to just use the models R squared value, but those results may lead to some bias. Those details aside, a scheme is now developed which can be applied on alternate data to obtain the best models (e.g. an initialization or IVC for the machine learning algorithm)

For the analysis on our data the resulting best model was found to be a three variable model which included only the predictor variable of county median income, county population and the variable describing city/county lockdown measures, as seen below in Table 1

TABLE I

Regression Statistics	
Multiple R	0.797811
R Square	0.636502

	df	*F*
Regression	3	36.772
Residual	63	
Total	66	
	Coefficients	*t Stat*
Intercept	-5.69897	-1.10881
county median income	0.00061	5.90571
county total population	1.10E-05	3.53724
city/county lockdown	0.37681	0.89383

Thus, we have determined, from analysing the individual predictor variable's T Stats, that the most deterministic factor is the median income while population (likewise population density) is not too far behind. And, we have obtained the most optimal model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where the predictor variables x_1,x_2 & x_3 are county median income, county population and variable describing city/county lockdown measures respectively.

Mathematically, and technically speaking, the variable "city/county" could be removed from this model due to the fact that its test stat is significantly lower than 1.96. Doing so, and as shown below in Table 2, we could have obtained the improved reduced optimal model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where the predictor variables, x_1 & x_2, are the county median income and population respectively.

TABLE2

Regression Statistics	
Multiple R	0.794916
R Square	0.631892

	df	*F*
Regression	2	54.931
Residual	64	
Total	66	
	Coefficients	*t Stat*

Intercept	-5.84512	-1.13961
county median income	0.000608	5.907079
county total population	1.31E-05	6.266592

This alternate two variable model maybe more mathematically pure, and more computationally efficient; however, it is important to recall that the original goal of our research study here was to obtain the desired result of obtaining the information as to what predictor variables are the most deterministic. Thus, it is useful to look at the process leading up to this final model, as opposed to just accepting the program's output; namely, it is very inserting to compare both the optimal three and two predictor variable models, and analyse the statistical results in the later steps of the backwards stepwise regression leading up to our final results. Specifically it is important to note which variables were removed early in the process as opposed to those removed later, and also the magnitude of predictor variables coefficient test statistics as the information contained within can be quite useful for interpretations.

III. CONCLUSIONS

In this concluding section, we will end by interpreting our mathematically results obtained in the prior sections, and demonstrating how they can be applied in real time. Now, the most significant, and somewhat unexpected, result from this data analysis is that the driving factors are county median income and county population while expected factors such as county demographics did not remain significant. This is a very interesting result, but it does make sense as those with financial security are more able to choose to stay at home while individuals living in scenarios such as pay check to pay check do not have such an ability.

Another very interesting result is of the various predictor variables originally inputted into the model to address either stay at home orders or business shut downs, it is found that the variable "city/county lockdown" was the more deterministic. This is perhaps the most interesting take away obtained here. In addition to the expected results of income, we determined the most deterministic of the predictor variables entered into the modelling which were designed to measure the effectiveness of how well "lockdown" measures worked. The two predictors were differing as while one basically described the timeline of county level laws which were intended to enforce lockdown, the other one described when business in the county actually began to close. For example the latter variable measured how far in advanced official business, including government offices, closed prior to state lockdown orders. It is expected that our intuition would lead us to believe that the first variable would be the more deterministic, but that was not found to be the case. The other variable, the one that basically described when business actually closed, was found to have a coefficient test stat almost three times as large.

A significant takeaway here is that in the real world application the more deterministic of these two predictors is a variable that was describing when things actually started to close. Hence, one can gain the understanding that, within our population studied, people really began to lockdown when they found out that things such as their local business actually closed, as opposed to the threat of a county level "lockdown rule" enforcement through measures such as imposing financial penalties through fines. Obviously there could be some subjectivity here as to if the counties actually enforced their "lockdown rule," or if they were just rules which were announced yet never truly enforced. However, a clear interpretation of these results is that an effective way for a government to get a population to practice social distancing, by home lockdown procedures, is to actually communicate this information to the consumer at the direct level of local business to consumer, as opposed to threating measures such as imposing financial penalties for noncompliance. This is an extremely useful information for local governments to be made aware of in the event that a future sudden lockdown is needed, either due to an expected second wave of the current pandemic and/or a potential future spreading virus

IV. SUGGESTED FURTHER STUDIES

A natural continuation of this research would be to subdivide the state of Florida into subsets organized by counties which fall into certain levels of the predictor variables (e.g. call one set the list of counties with higher income and another subset the list of counties with lower median income), and then perform statistical analysis on the difference of their SARS-CoV-2 infections and/or deaths. Furthermore, one may desire to study the effectiveness of social distancing and/or correlation to the number of SARS-CoV-2 infections and/or deaths, but that was never the intent of this study so those related applications will not be mentioned here. However, it is extremely clear from just observational data analysis that during the time period when social distancing was strongly enforced is extremely different than the time period when it was relaxed. For example looking at the two counties in the state of Florida which are the most populous (Dade) and the most densely populated (Pinellas) then comparing their case numbers during the month of April (when the social distancing was applied) to the months of July (when the social distancing was initially relaxed) the result is beyond clear. Namely, in for the state of Florida overall the numbers in April were around the values of 900 to 1000 new cases daily, and in July the values were around the values of 10,000 to 15,000. Furthermore, in Dade County the daily numbers in April were around the value of 300 to 400 new cases daily, and in July the values were around the values of 2,000 to 3,000. In addition, in Pinellas County there was almost no cases in April while in July the values jumped up to averaging around 400. Again, these results are beyond conclusive by just observation, and while it is not the intent of this study to prove that social distancing is effectiveness, it is worthy to note for anyone who desires to continue any study involving the number of cases that it is important to also consider the sample size tested as it can cause bias. Putting aside external threats to validity, such as determining why individuals were seeking and/or given testing, a good measure to resolve this matter

mathematically is to not study the raw numbers rather study the ratio of positive to tested; this is illustrated within the Miami-Dade COVID project data site [5]

In addition, a very interesting follow on study would be to repeat this analysis on different states and/or countries to see if similar results are obtained in different regions or if these results were specific to the state of Florida. The majority of the data utilized within this study is publically available and it is expected that similar information would be available for other states. For example, it was found that the encyclopaedia website, Wikipedia, had very detailed and easily available information about county level population data for Florida [6] along with a large amount of other related data. It is expected that similar data should be available for other counties and/or countries. Likewise, data such as county median income and demographics should be available by routine web searches; the current data was obtained from a county survey [7] and a local university study [8] respectively

REFERENCES

[1] Unacast Social Distancing Scoreboard https://www.unacast.com/covid19/social-distancing-scoreboard
[2] T. Smith, *MA Self-Contained Course in the Mathematical theory of Statistics for Scientists & Engineers with an emphasis on predictive regression modelling & financial applications." Embry Riddle Aeronautical University Creative Commons, 2019*.
[3] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*." Springer, 2013.
[4] T. Smith, "A Statistical Learning Model utilized to validate a well-know market hypothesis of the moving average "death cross.", 2019 international journal of mathematical trends and technology, International Journal of Mathematics Trends and Technology 65.11 (2019):72-82.
[5] Miami Dade County COVID project data site https://rwilli5.github.io/MiamiCovidProject/Trajectory/
[6] Wikipedia list of Florida counties https://en.wikipedia.org/wiki/List_of_counties_in_Florida
[7] Florida Survey of Income https://www.countyhealthrankings.org/app/florida/2020/measure/factors/63/data
[8] Florida University Study of Demographics http://edr.state.fl.us/Content/population-demographics/data/PopulationEstimates2019.pdf