A Statistical Learning Regression Model Utilized to Determine Predictive Factors of Social Distancing During COVID-19 Pandemic

Matthew V. Chin, Timothy A. Smith, Albert J. Boquet
Embry-Riddle Aeronautical University, Daytona Beach, FL.

Issue

Predictive regression modeling 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.

Research Objective

The research goal is to determine what are the most deterministic predictor variables of the regression model.

Mobility GPS Data

Mobility data was collected from Unacast in Figure 1 that analyzes real-time data in finding the change in average distance traveled from confirmed cases along with ping frequency of people's movements.

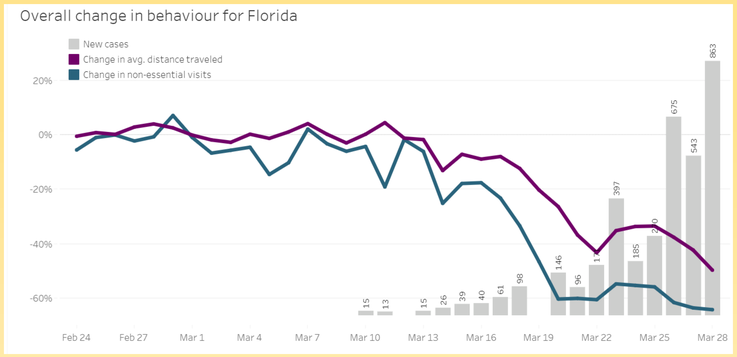

Fig. 1. Unacast Social Distancing Scoreboard [1]

Methodology

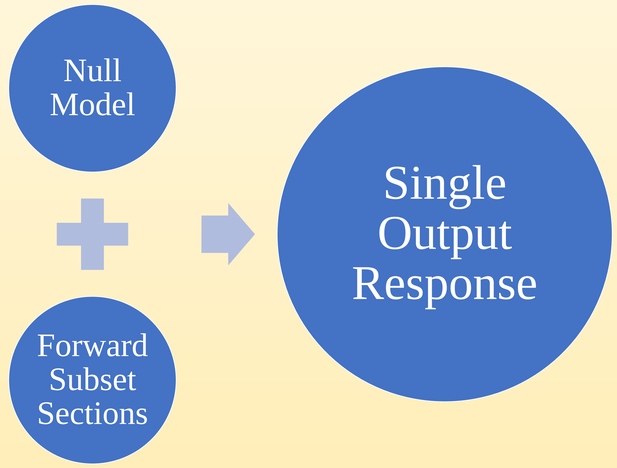

Fig. 2. Multiple Regression Model [2]

1. For each of the "Forward Subset Sections" in Figure 2, the coefficient of determination (R^2), the F Statistic and the P-Value were computed for analysis.
2. As part of regression analysis, the Variance Inflation Factor (VIF) showed that some predictors were moderately correlated.
3. Predictors with a T Statistic lower than 1.96 or had multicollinearity concerns were excluded.

Key Findings

The best resulting model was a three-variable model:

1. County median income (β_1 = 0.00061)
2. County population (β_2 = 1.10E-05)
3. City or County lockdown measures (β_3 = 0.37681)

The resulting multiple regression model is:

$$y = 0.00061x_1 + 1.10 \times 10^{-5}x_2 + 0.37681x_3$$

Discussion

	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-06	6.0266592

Fig. 3. Regression analysis showing the T-Statistic of the predictor variables [3]

From the T Statistic, median income (5.90571) and county population (3.53724) were the most deterministic.

Future Research

1. Subdivide the State of Florida into subsets by counties that contain certain levels of the predictor variables from the current data to further investigate.
2. Conduct a new similar study nation-wide.
3. Study the effectiveness of social distancing in correlation to the number of SARS-CoV-2 infections or deaths or both.

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] T. A. Smith, A. J. Boquet, M. Chin. "*A Statistical Learning Regression Model utilized to determine predictive factors of social distancing during COVID-19 pandemic*," International Journal of Mathematics Trends and Technology, Vol. 66, no. 11, pp. 63, Nov. 2020, doi: 10.14445/22315373/IJMTT-V66I11P504