Analyzing COVID 19 Data in USA using Autocorrelation

Trang Hoang
Clark University, trhoang@clarku.edu

Follow this and additional works at: https://commons.clarku.edu/asdff

https://commons.clarku.edu/asdff/ASD2020/Posters/6

This Open Access Event is brought to you for free and open access by the Conference Proceedings at Clark Digital Commons. It has been accepted for inclusion in Academic Spree Day and Fall Fest by an authorized administrator of Clark Digital Commons. For more information, please contact mkrikonis@clarku.edu, jodolan@clarku.edu, dltuz@clarku.edu.
Analyzing COVID 19 Data in USA using Autocorrelation

Trang Hoang Thu (Sponsor: Professor Michael Satz)
CS/Math Department

COVID 19 (Coronavirus)
A highly contagious virus that can be spread from person to person. This new virus is an outbreak of respiratory illness

Through Nonlinear Regression Model, analysis focus on time vs confirmed cases

1) Exponential Regression
\[ Y_t = f(X_t) = \beta_0 + \beta_1 X_t + \epsilon_t \]

2) Logistic Regression
\[ Y_t = \frac{\beta_0}{1 + \exp(-\beta_1 X_t)} + \epsilon_t \]

In 1.b) + 2.b) the residuals (the distance from the data to the fitted lines) show time dependent patterns. This mean that the error of the model is correlated over time

From 2.a) graph, the fitted line reaches a plateau halfway through April. This model is not suitable because the total confirmed cases doesn’t portray reality

First Order Autoregressive Error Model (AR(1))

Eliminate correlated error using

\[ (\text{AR}(1)) \]

\[ Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \]
\[ \epsilon_t = \rho \epsilon_{t-1} + u_t \]

Where:
\[ \rho \text{ is a parameter, } |\rho| < 1 \]
\[ u_t \text{ are independent } N(0, \sigma^2) \]

Research goal:
- Using time series analysis to predict confirmed cases from May 1st to May 15th using data from March 16th to April 30th (test model reliability with data from May 1st to May 7th)

Forecasting with autocorrelated terms

Data uses for analysis (March 16th - April 30th)
Data uses to determine model accuracy (May 1st - 7th)
Fitted Lines
Total confirmed cases predicted range

Analysis: Majority of the actual confirmed cases (open dots) are in range of the predicted confirmed cases. However, the forecasting loses its accuracy overtime. A forecasting model is best to analyze a few additional period. My next step for this research is to use predicted confirmed cases to figure a confidence range of the total deaths. Due to the lack of significant predictors, my prediction for confirmed cases are limited to 5 days. The model can produce a more accurate and precise overtime dependent prediction by adding more significant predictors.

Conclusion: Autoregressive error model is a strong time series analysis tool for close time step analysis due to its ability to predict within reasonable errors.

Reference:
- European Centre for Disease Prevention and Control. “today’s data on the geographic distribution of COVID-19 cases worldwide”