Academic Spree Day and Fall Fest

Academic Spree Day 2020

May 17th, 12:00 AM - 11:00 PM

Analyzing COVID 19 Data in USA using Autocorrelation

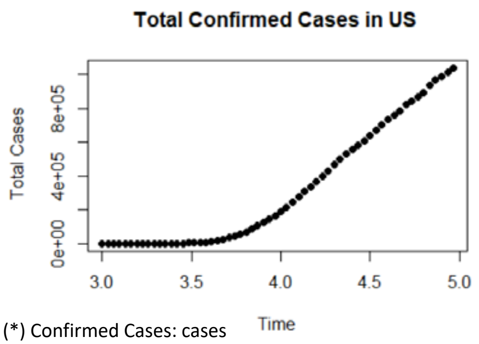
Trang Hoang
Clark University, trhoang@clarku.edu

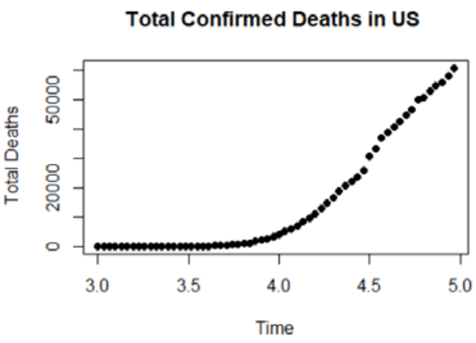
Analyzing COVID 19 Data in USA using Autocorrelation

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





(*) Confirmed Cases: cases have been documented

→ The 2 models might behave like an exponential model

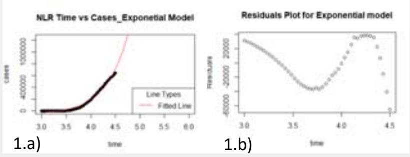
Through Nonlinear Regression Model, analysis focus on time vs confirmed cases

1) Exponential Regression

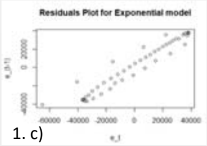
$$Y_i = f(X,\gamma) = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i) + e_i$$

2) Logistic Regression

$$Y_i = f(X,\gamma) = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X_i)} + e_i$$

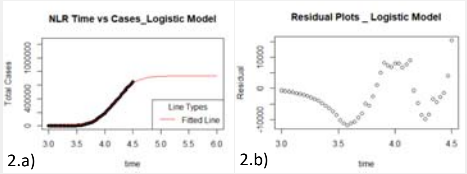


1.a) 1.b)



1. c)

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



2.a) 2.b)

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

Eliminate correlated error using

First Order Autoregressive Error Model (AR(1))

Autocorrelation

- Error terms correlated over time are said to be *autocorrelated* or *serially correlated*
- Problem of autocorrelation:
  - Confidence interval and test using t and F distribution are not applicable
  - Underestimate regression coefficient

Simple Linear Regression: (when the random error terms follow AR(1)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_t$$
$$\varepsilon_t = \rho e_{t-1} + u_t$$

Where:

ρ is a parameter, $|\rho| < 1$
u_t are independent $N(0, \sigma^2)$

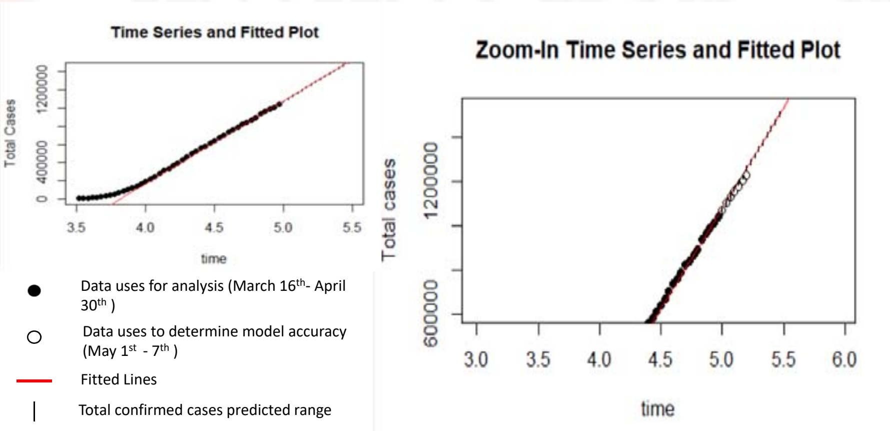
- To test for autocorrelation, we use Durbin Watson (DW) Test. The test alternatives:

$H_0 : \rho = 0$ _ error terms are independent
$H_a : \rho > 0$ _ error terms are correlated

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

- Kutner, M. H., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models*.
- European Centre for Disease Prevention and Control. "*today's data on the geographic distribution of COVID-19 cases worldwide*"