ANALYZING COVID 19 DATA USING AUTOCORRELATION

TRANG HOANG

GOAL

- Understand Autocorrelation and Autoregressive Error Model
- Apply the autocorrelation concepts and methods to solve real life problems
- How would autocorrelation improve model and time series analysis

AUTOCORRELATION

I. What is autocorrelation?

Error terms correlated over time are called *autocorrelated* or *serially correlated.* The causes for positively correlated error terms happen when your models are missing key variables which includes time-ordered effects.

Eg: the regression of annual sales of a product against average yearly price over a period. If population is an important factor of the model, its omission from the model may lead to the error terms being autocorrelated.

AUTOCORRELATION

II. Problems with autocorrelation

- Mean Squared Error may underestimate the variance of the error terms
- The standard deviation of the regression coefficient may not accurately calculate
- Confidence intervals and tests using t and F distribution are not applicable (this affects time series analysis)

FIRST ORDER AUTOREGRESSIVE ERROR MODEL

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad (1)$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \qquad (2)$$

Where ρ is a autocorrelation parameter, $|\rho| < 1$

$u_t \sim N(0, \sigma^2)$

II. Remedial for autocorrelation

- Add one or more variable to the regression model
- Use transformed model

$$Y_t' = Y_t - \rho Y_{t-1}$$

$$Y_t' = (\beta_0 + \beta_1 X_t + \varepsilon_t) - \rho(\beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1})$$

$$\Rightarrow Y_t' = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1}$$

$$\Rightarrow Y_t' = \beta_0' + \beta_1' X_t' + u_t$$

where

$$Y_t' = Y_t - \rho Y_{t-1\rho\rho}$$

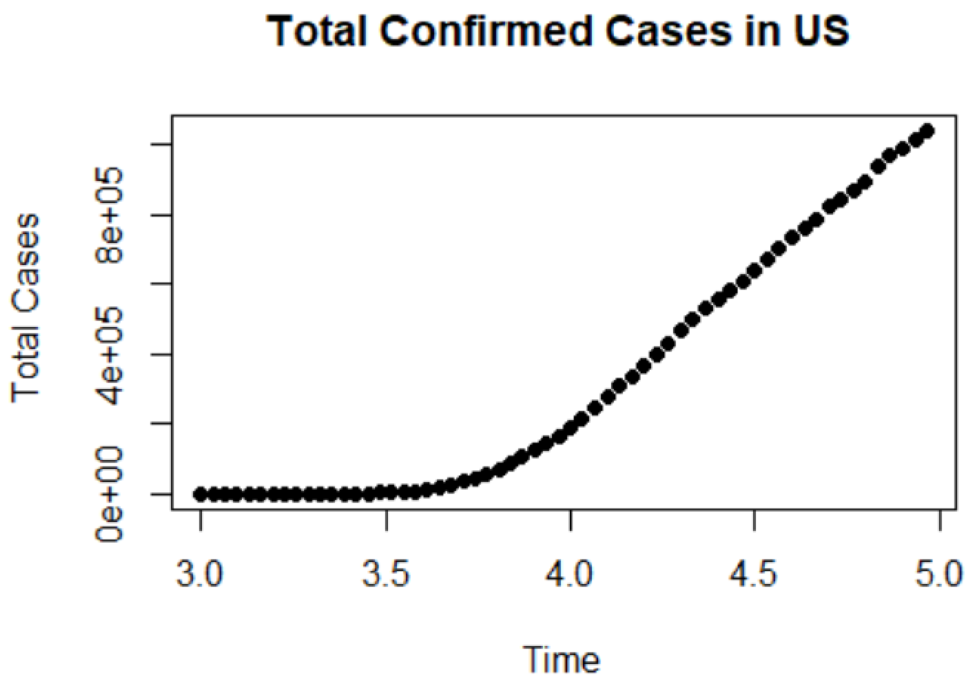
$$X_t' = X_t - \rho X_{t-1}$$

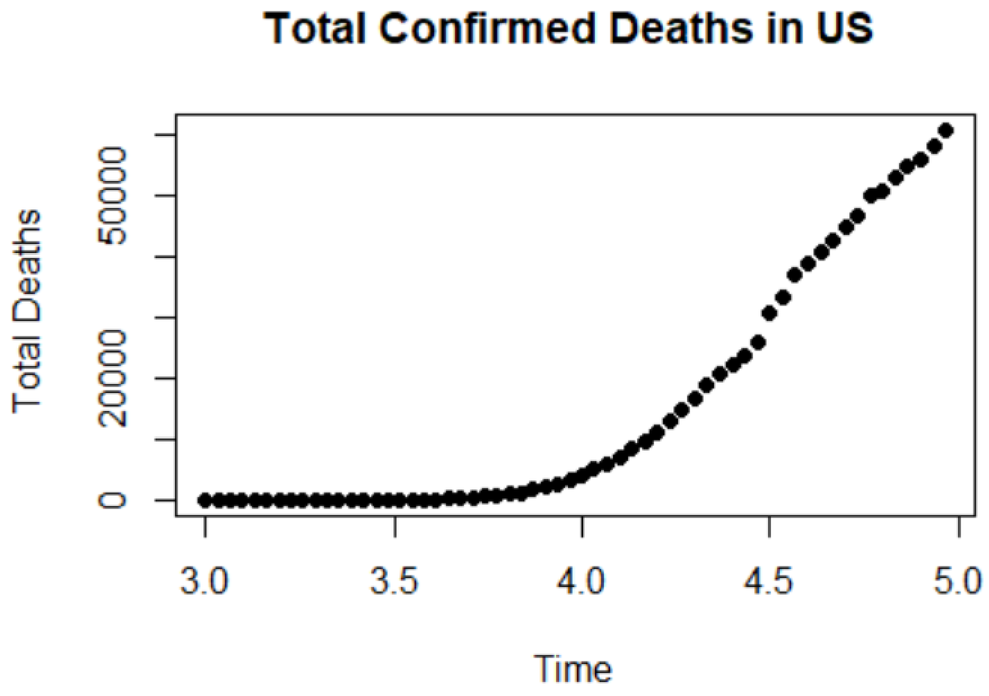
$$\beta_0' = \beta_0(1-\rho) \ and \ \beta_1' = \beta_1$$

$$\varepsilon_t - \rho\varepsilon_{t-1} = u_t$$

- Hildreth Lu Procedure
  - Estimates the autocorrelation parameter by finding ρ that minimizes SSE ($SSE = \sum(Y_t' - \hat{Y}_t')^2$)

COVID 19 DATA

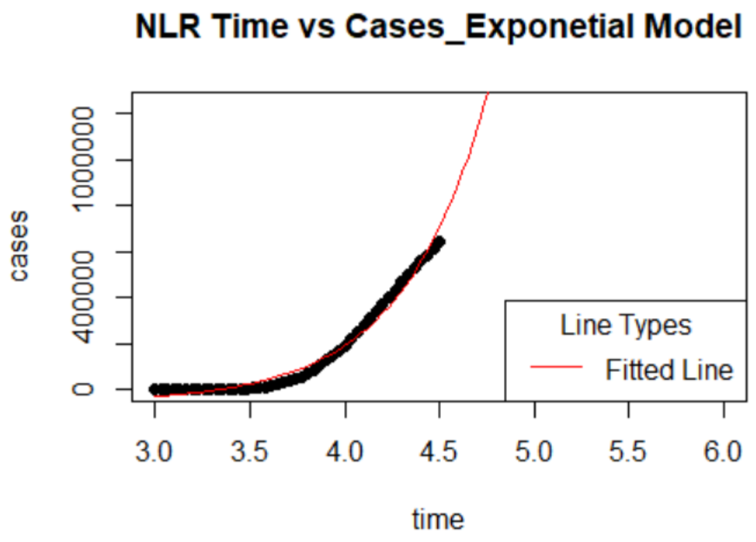


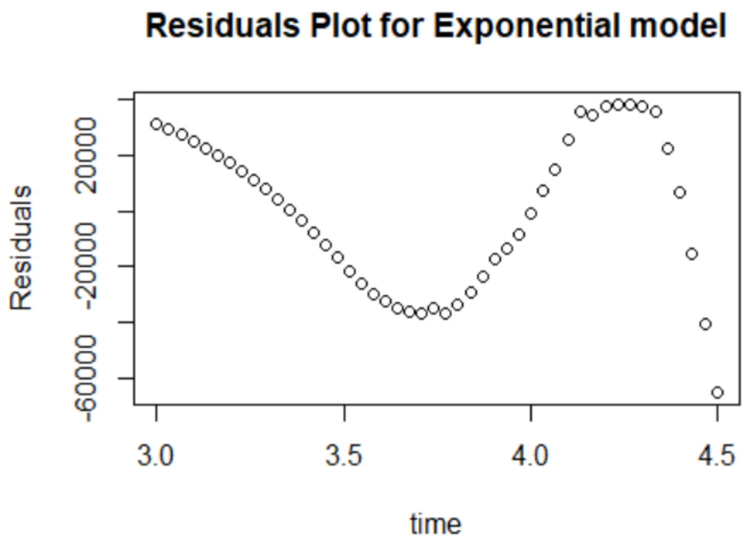


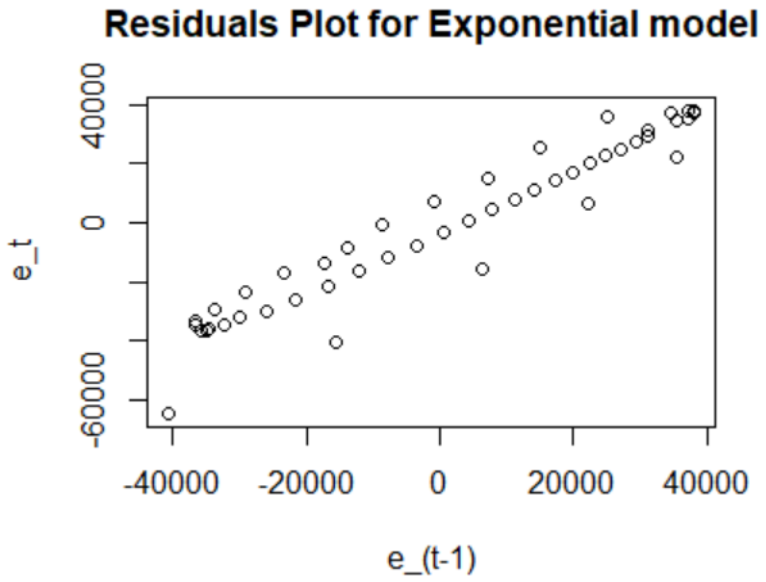
CONFIRMED CASES VS TIME ANALYSIS

- Guess that Confirmed Cases vs Time Graph is a Nonlinear Regression model.

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







The error terms are autocorrelated

DURBIN WATSON TEST FOR AUTOCORRELATION

- The test determines whether ρ is zero.
- Test alternatives

$$H_0 : \rho = 0$$
$$H_a : \rho > 0$$

- Calculating test statistic: $D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$

- - Decision Rule: if $D > d_U$, conclude H_0
  if $D < d_L$, conclude H_a
  if $\mathrm{d_L} \leq D \leq d_U$, the test is inconclusive

COVID 19 DATA

Data, Regression Result (before remedial measures) Error and Durbin Watson Test Calculations

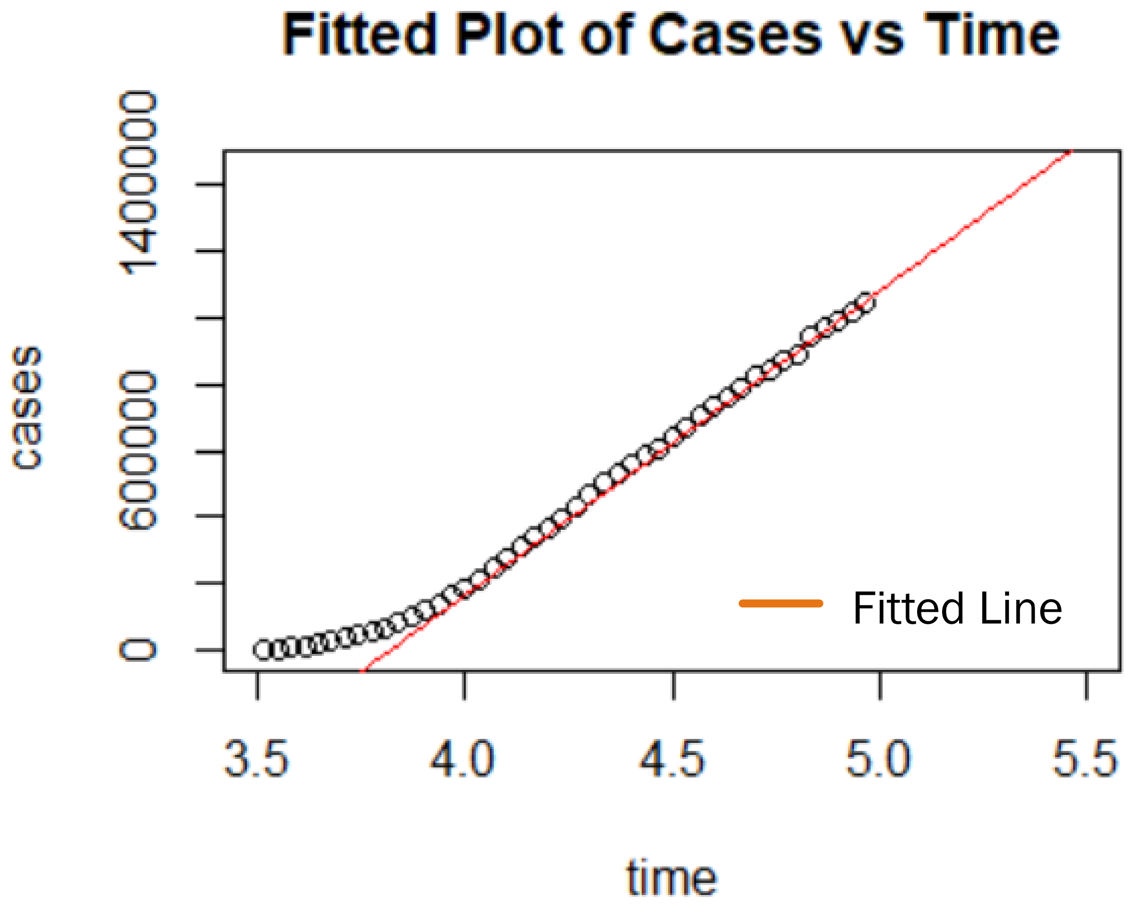
Time (month)	Cases	e_t	e(t-1)
3.516	4,661	132763.763	---
3.548	6,427	109479.314	132763.763
3.581	9,415	86634.038	109479.314
…			
4.9	988,451	33121.839	36675.057
4.9333	1,012,583	31159.621	33121.839
4.9667	1,039,909	32391.404	31159.621

$$Fitted\ Lines:\ \widehat{Y}_t = -2880520.9 + 782826.5X_t$$

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{4517921549}{88080319905} = 0.051293$$

$d_U\ and\ d_L$ for 45 data points: $d_L = 1.48\ and\ d_U = 1.57$

$\rightarrow$ Conclude $\mathrm{H_0}$



Before Remedial Measures:

$$\widehat{Y}_t = \text{-}2880520.9 + 782826.5X_t$$

After Remedial Measures

$$\widehat{Y}_t = \text{-}3472567 + 910165X_t$$

FORECASTING + TIME SERIES ANALYSIS

- Forecast for period n+1 is:

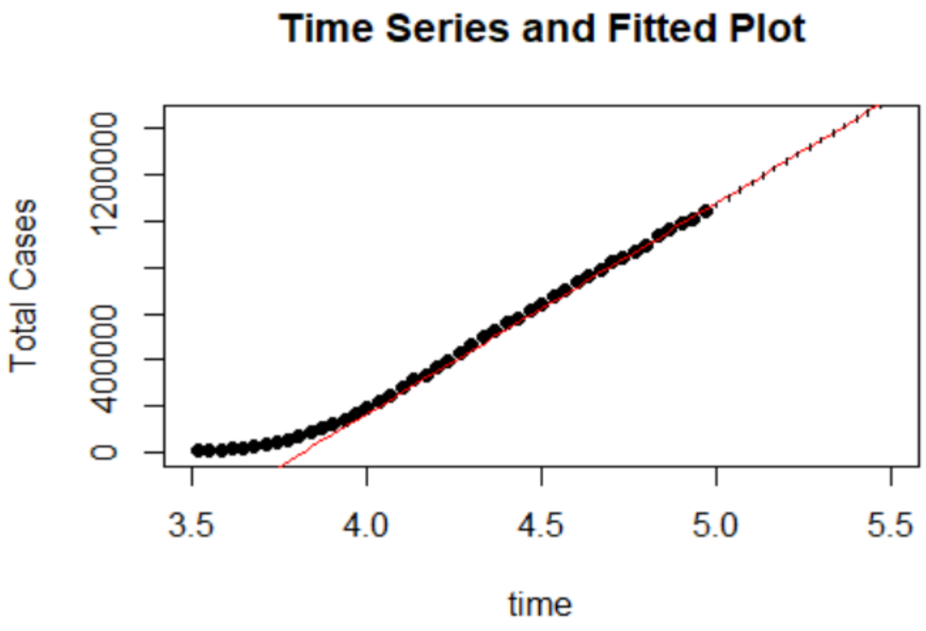
$$F_{n+1} = \widehat{Y_{n+1}} + re_n$$

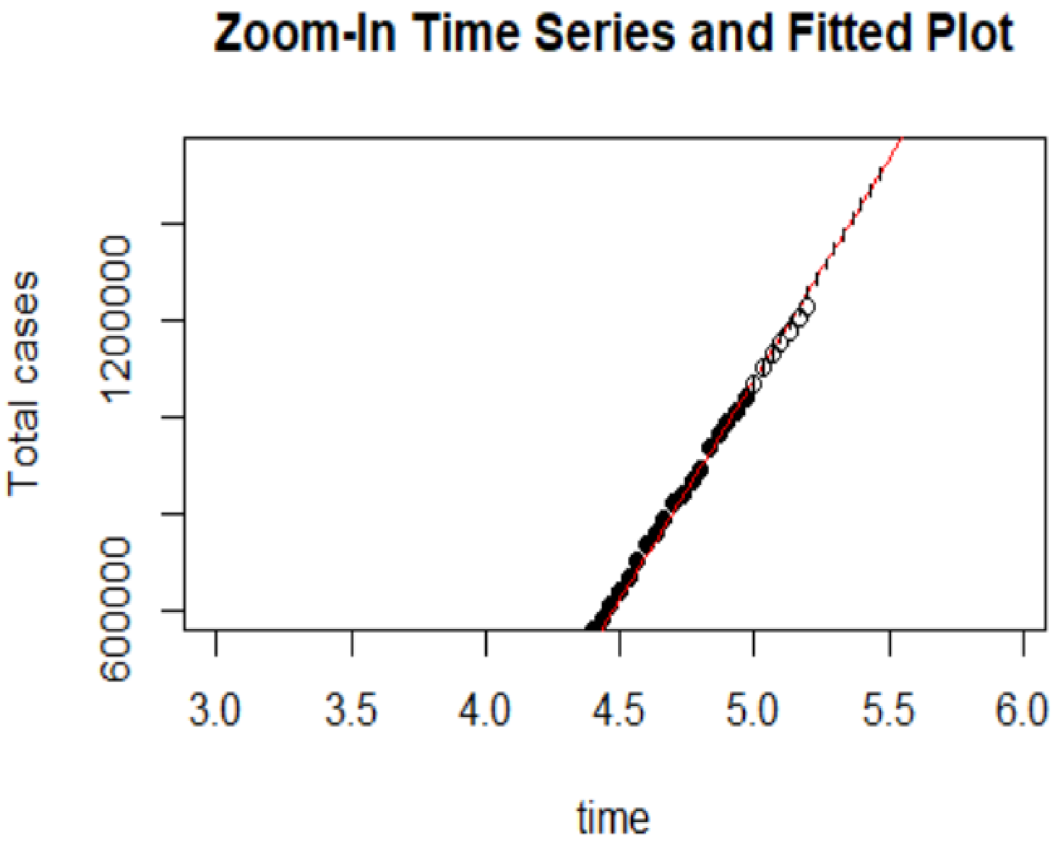
  - The confidence intervals for F_{n+1} (or the prediction limits) are $F_{n+1} \pm t\left(1-\frac{\alpha}{2}; n-3\right) s\{pred\}$

Where $s\{pred\} = \sqrt{\frac{SSE}{n-3}\left(1+\frac{1}{n}+\frac{(X_{n+1}-\bar{\bar{X}})^2}{\sum(X_i-\bar{\bar{X}})^2}\right)}$

- For the Confirmed Cases vs Time analysis, I use data from March 16th to April 30th to build the model.

- Forecasting from May 1st to May 15th then comparing the forecasted data with the real time data





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

Date	Period	Predicted Intervals	Actual Value
May 1st , 2020	1	$1{,}060{,}145 \le Y_1 \le 1{,}082{,}273$	1,069,826
May 2nd,. 2020	2	$1{,}091{,}298 \le Y \le 1{,}113{,}489$	1,103,781
May 3rd, 2020	3	$1{,}122{,}349 \le Y \le 1{,}144{,}605$	1,133,069
May 4th , 2020	4	$1{,}153{,}308 \le Y \le 1{,}175{,}633$	1,158,041
May 5th, 2020	5	$1{,}184{,}188 \le Y \le 1{,}206{,}584$	1,180,634
May 6th , 2020	6	$1{,}214{,}997 \le Y \le 1{,}237{,}467$	1,204,475
May 7th , 2020	7	$1{,}245{,}745 \le Y \le 1{,}268{,}290$	1,228,603

Table: Predicted Confirmed Cases from Time Series Analysis and Actual Real Time Confirmed Cases

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*”

- Analysis: This is not a complete prediction model. My goal for this research is to use autocorrelation and time series analysis to see how effective is it. The confirmed cases deaths increases through out depends on various variable such as how serious people adhere to social distancing