JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

TECHNOLOGY FOR DEVELOPMENT

Modeling and Forecasting of COVID-19 Cases in Kenya using ARIMA Model

Research Proposal Presented By:

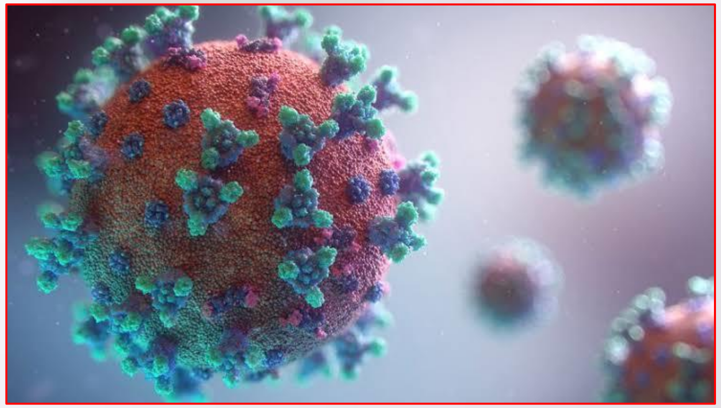
Benson Mwangi, Christopher Njimu, Collins Kipkirui, Faith Chepkoech

Outline

- INTRODUCTION
- LITERATURE REVIEW
- METHODOLOGY

Introduction

- Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus, whose family of viruses also caused SARS(2002) and MERS(2012).
- The first case of human infection was reported in Wuhan, China (Dec,2019).
- 11th March 2020, declared a pandemic by the WHO.
- 13th March 2020, first confirmed case in Kenya.
- 17 months into the pandemic, Kenya has had over 220,000 confirmed cases, 4 300 attributable deaths and currently cases are surging into the 4th wave.



The face of the pandemic

- The pandemic has also had serious implications on widening social inequalities, loss of livelihoods, reduced economic productivity, stretched the health sector, disrupted the school calendar.
- With still a long way to go in terms of increasing percentage of vaccinated population, and the virus property of constantly mutating into new variants, there's still need to reduce the spread by informing decisions using updated data, in order to cushion livelihoods while saving lives.

Problem statement

- Covid-19 pandemic has disrupted the economic and health sectors of the country. It has caused deaths, reduced productivity, business closures, trade disruptions, school calendar disruption and decimation of the tourism industry.
- Community transmission is happening within the country, making it hard to trace sources of exposure.
- Need to predict possible future surges in order to tighten measures in good time to avoid loss of lives and interrupt predicted surges.
- Need for regular updating of models to factor in new variants and vaccination progress.
- Government needs COVID-19 forecasts to inform its decisions.
- Our study focuses on developing a suitable ARIMA model that can be used to forecast the future trend of COVID-19 cases in Kenya.

Justification

- Through modeling and forecasting of the covid-19 cases, it will establish the likely Impact of the pandemic and help inform national and local government planning, budgeting, readiness, response and monitoring.
- Help the government in adjusting existing measures accordingly, by striking a fine balance in between saving lives and maintaining livelihoods.
- Help the health sector to proper plan through anticipating future possible surges to avoid a similar situation as that seen in India.

General and specific objectives

General objective.

To model and forecast COVID-19 cases in Kenya using ARIMA model.

Specific objectives.

- To analyze the stationarity of COVID-19 cases in Kenya.
- To develop an ARIMA model for COVID-19 cases in Kenya.
- To forecast future COVID-19 cases for the next eight weeks.

Research Questions and Scope

Research Questions

- Is there a trend in the covid-19 cases?
- What is the most suitable ARIMA model for the Covid-19 cases?
- What are the expected number of cases in the near future?

Scope of the study

Our study will analyze covid-19 cases in Kenya from March 2020 to October 2021, find a suitable ARIMA model, and project future covid-19 cases.

Literature Review

Previous related studies:

- India, (Gupta, 2020) conducted an analysis of COVID-19 cases using ARIMA model.
- Globally, (Chaurisia and Pal, 2020) applied ML in a time series analysis, comparing various models for prediction of the Covid-19 pandemic. ARIMA was the second best after Naïve method.
- Saudi Arabia,(Alzahrani et al.,2020) used the ARIMA model to forecast the spread of Covid-19 pandemic in Saudi Arabia considering the measures put in place.
- China, (Wang, 2018) modelled and compared the ARIMA and Grey Models for Hepatitis B monthly cases. ARIMA showing better performance than Grey Models.
- Kenya, (Sam el al., 2021) developed Otio-NARIMA model for forecasting seasonality of Covid-19 waves in Kenya.

Following this researched studies, we arrived at the conclusion that since the ARIMA model has been used before to model other susceptible diseases, including COVID-19 in other nations, the ARIMA model might be suitable to model COVID-19 cases in Kenya and to forecast the expected near-future tendencies.

Methodology

1. Research Design

Analytical research design.

2. Data

Weekly reported cases for the period *March 2020 – October 2021*.

R Statistical software, will be used for analysis of data.

3. Sources of data

Ministry of Health Kenya and Worldometer.

4. The model: ARIMA model Procedural Steps

i. Data validation.

ii. Testing the stationarity of the data.

iii. Estimation of Parameters and order selection.

iv. Model validation.

v. Forecasting.

Methodology

Time series is a chronological sequence of observations on a particular variable. Usually the observations are taken at regular intervals (days, week, months, years)

ARIMA Model

- The ARIMA(p, d, q) model is a mixture of the autoregressive model AR(p) , moving average model MA(q) and integrated through some differencing I(d).

- The autoregressive model uses past values of the dependent variable to predict the future values of the variable while moving average process uses past errors to forecast the dependent variable.

- The combining of AR and MA helps in keeping the number of parameters used small to achieve parsimony in parameterization.

The general form of the ARIMA (p, d, q) model is:

$$y_t = \beta_0 + \emptyset_1 y_{t-1} + \cdots\cdots + \emptyset_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots\cdots + \theta_q \varepsilon_{t-q}$$

Where,

y_t, is the dependent variable

y_{t-i}, $i = 1, \ldots, p$ lags of dependent variable.

$\emptyset_i$, $i = 1, \ldots, p$ coefficients of lags of dependent variables for the AR process

θ_i, $i = 1, \ldots, p$ coefficients of lags of dependent variables for the MA process

ϵ_j, j= 1, …, q lags of forecast error.

Methodology

1. Data validation

- Confirm whether we have the correct data by comparing data from our two data sources.
- Check for missing entries, inconsistency, data type errors, date format errors and rectify them.

2. Test for stationarity

- A time series is said to be stationary in weak sense if its mean and variance is constant and it's autocovariance function is independent in time but depends only on the distance between the observations.
- Augmented Dickey-Fuller (ADF) test will be used to test for stationarity at 5% level of significance.

Hypothesis

$\boldsymbol{H_0}$: The time series is non-stationary vs. $\boldsymbol{H_1}$: The time series is stationary.

Decision: If the P-value is less than 0.05 level of significance, we Reject the null hypothesis and conclude the time series is stationary otherwise Fail to Reject null hypothesis.

Methodology

3. Estimation of Parameters and order selection

We'll use the ACF and PACF to identify a suitable ARIMA(p, d, q) model.

- If ACF graph cut off after lag n and PACF decays exponentially we identify MA (q) resulting to ARIMA (0, d, n) model.
- If ACF decays exponentially and PACF cut off after lag n, we identify AR (p) resulting to ARIMA (n, d, 0) model.
- If ACF and PACF decay exponentially that is mixed ARIMA model, then differencing is needed.

The chosen model should have as few parameters as feasible while still being able to describe the series (parsimony).

We fit the identified model using MLE procedures which allows to produce the estimates and standard errors for the parameter coefficients

Minimizing the AIC or AICc, while maximizing log likelihood yields good models.

The corrected Akaike Information Criterion (AICc) is our preferred metric, and the parsimonious model with the lowest AICc and highest log likelihood will be chosen.

Methodology

4. Model checking

- Estimated models will be checked using ACF and PACF of estimated series, and model(s) with no significant lags will be chosen as the one that best explains the temporal dependency.
- To assess the goodness of fit, over-fitting and a battery of diagnostic tests will be used to test residuals of the fitted model to see whether they appear to be a white-noise innovations.
- The Ljung-Box test will also be used to test for serial autocorrelation in the time series at 5% level of significance.

- The hypothesis to be tested:

$\boldsymbol{H_0}$: No serial autocorrelation vs.

$\boldsymbol{H_1}$:There's serial autocorrelation

If the Ljung-Box statistic of the model is not significantly different from 0 we fail to reject the null hypothesis of no remaining significant autocorrelation in the residuals of the model and conclude that the model seems adequate in capturing the correlation information in the time series

Methodology

5. Forecasting

- The ARIMA model to be used in forecasting is of the form:

$$y_t=\beta_0+\emptyset_1 y_{t-1}+\cdots\cdots+\emptyset_p y_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots\cdots+\theta_q\varepsilon_{t-q}$$

- AIC, Root Mean Square Error (RMSE), Root Mean Square Percent Error (RMSPE), and Mean Absolute Error (MAE) will be used to examine the forecasting performance of the various types of ARIMA models.
- The smaller the statistics, the better the model.
- Conclusions will be reached based on the aforementioned selection and evaluation criteria.
- Finally, a diagnostic check will be carried out to ensure that the chosen model is the optimum fit. We will check the relative inaccuracy by comparing the predicted and true values of October cases.

Ljung Box-test

H_0: Model is adequate vs. H_1: Model is not adequate

If p-value of the test is less than 0.05 level of significance reject null hypothesis and conclude model is not adequate; otherwise fail to reject.

References

Alzahrani, S., Aljamaan, I., & Al-Fakih, E. (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using the ARIMA prediction model under current public health interventions. *Journal Of Infection And Public Health*, *13*(7), 914-919. https://doi.org/10.1016/j.jiph.2020.06.001

Chaurasia, V., & Pal, S. (2020). Application of machine learning time series analysis for prediction COVID-19 pandemic. *Research On Biomedical Engineering*. https://doi.org/10.1007/s42600-020-00105-4

Khan, F., & Gupta, R. (2020). ARIMA and NAR-based prediction model for time series analysis of COVID-19 cases in India. *Journal Of Safety Science And Resilience*, *1*(1), 12-18. https://doi.org/10.1016/j.jnlssr.2020.06.007

Sam, S. O., Pokhariyal, G. P., Rogo, K., & Ndhine, E. O. (2021). Otoi-NARIMA model for forecast seasonality of COVID-19 waves: Case of Kenya. International Journal of Statistics and Applied Mathematics 2021; 6(2): 48-58. https://doi.org/10.22271/maths.2021.v6.i2a.675

Wang, Y., Shen, Z., & Jiang, Y. (2018). Comparison of ARIMA and GM (1,1) models for prediction of hepatitis B in China. *PLOS ONE*, *13*(9), e0201987. https://doi.org/10.1371/journal.pone.0201987

THANK YOU-AHSANTE

Stay Safe