Analyses of Factors Affecting Deaths Associated with COVID-19 in Ontario

Jie Huang

University of Toronto, 27 King's College Cir, Toronto, Ontario, M5S 1A1
jieh.huang@mail.utoronto.ca

Abstract

Since the outbreak of the COVID-19 in 2019, it has been a great challenge for the whole world. When the epidemic is serious and the vaccine will play a role, the statistic is an effective tool. It can help the government collect various data and conduct modelling analysis, so that it can face the actual situation and issue appropriate policies. This paper aims to analyse the factors that could affect the death rates among all COVID-19 confirmed cases in Ontario. Specifically, Seasonal ARIMA is used to fit past one-year data to predict short-term trend of confirmed case. An overall upward slope is predicted by selected time series model. Logistic regression is then used to determine how age group and vaccination could affect the mortality risk quantitatively. According to the information as of November 6, 2021, the forecast trend in the short term is expected to show an upward trend. In addition, age group and vaccination status significantly affect the probability of death of confirmed cases. The mortality increased with age. It has also been proved that the mortality of fully vaccinated patients is lower than that of partially vaccinated patients, followed by unvaccinated patients.

Keywords: *COVID-19; Mortality; Time Series; Logistic Regression; Vaccination*

1. Introduction

Since 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected a large proportion of the population worldwide. As a result, deaths associated with COVID-19 are observed as a new category on the mortality table. As of October 31, 2021, the confirmed cases are 247 million, and related deaths are 5 million worldwide [1]. Many potential factors could affect COVID mortality, such as individual age, ethnicity, gender, medical condition, and vaccination status.

Since Israel launched mass vaccination of new crown vaccine booster in July 2021, more than a dozen countries around the world have started similar plans. It is expected that more countries will join, causing more attention to vaccine hybrid vaccination. More and more signs show that the immune protection provided by the new crown vaccine is significantly weakened 4-6 months after the completion of vaccination, and the cases of breakthrough infection and antibody attenuation are increasing. There are also more data indicating that the vaccine enhancer plays an obvious role in improving immune protection [2].

On November 6, 2021, the Ontario government announced that certain public groups are eligible to book appointments for their vaccine if they received second dose at least six months ago. There are several eligible population groups, including residents over 70 years old [3].

This paper will talk about the factors affecting deaths with COVID-19 in Ontario. Most of the analysis online is about whether the vaccine is effective or the trend of the next epidemic situation. Because the changes in the epidemic are difficult to predict, there are relatively few studies of this kind. In addition, because the epidemic has not been well controlled since 2019, it is hard to make a complete analysis with official data. Therefore, this paper chooses only part of official data and creates a staged comment to predict the development of the epidemic in the future.

This paper preliminarily explores the short-term trend of confirmed cases in Ontario for the next 15 days. It then investigates how the age group and vaccination status of confirmed cases in Ontario could affect mortality risk among all confirmed cases. Based on the historical trend of confirmed cases, seasonal Autoregressive Integrated Moving Average (ARIMA) is employed to predict future

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 1294–1300, 2023.
https://doi.org/10.2991/978-94-6463-042-8_186

trends [4]. Regression analysis is a predictive modeling technique used to find the relationship between dependent variables and any independent variables. In this paper, the logistic regression model compares the odds ratio of death versus alive between different age groups and vaccination status [5]. Furthermore, the age group is divided up by each twenty-year interval. Vaccination status is based on the dose taken, which can also be known as not yet vaccinated, partially vaccinated, and fully vaccinated. It is a global responsibility that each country should take actions to prevent the spread of COVID-19 and lower the mortality risk. This thesis discovers population in Ontario that exposes higher mortality risk after COVID-19 infection. The result in this report is applicable to other provinces in Canada even worldwide. That is, to help control the mortality among confirmed cases, government should consider the fact that age and vaccination status could significantly affect COVID-19 mortality.

2. Methodology

2.1. Data Source

The main object is in Ontario province, and hence the Ontario COVID-19 Data Tool [6] is considered a reliable source. The Ontario COVID-19 Data Tool provides epidemiological information on COVID-19 activity in Ontario to date. It allows to explore the most recent COVID-19 data, including daily case counts by hospitalizations and deaths, vaccine uptake by age, and public health units.

2.2. Data Treatment

This paper mainly uses past one-year data for time series analysis and past two-week data for mortality risk investigation as of November 06, 2021. Past two-week data regarding confirmed cases and death counts obtained from the Data Tools mentioned in section 2.1 are translated to binary variables to fit logistic regression later. Specifically, 0 represents the survival of a confirmed case, whereas 1 indicates the death. In addition, vaccination status is an ordinal variable with 0, 1, or 2 to indicate the number of doses taken among the confirmed cases. Another interpretation for the ordinal variable would be no vaccination taken, partially vaccinated, and fully vaccinated.

2.3. Seasonal ARIMA Model for Future Case Prediction

Autoregressive Integrated Moving Average (ARIMA) [7] is a forecasting method for time series data. χ_t is an autoregressive-integrated-moving average process of order (p, d, q) (ARIMA (p, d, q) process) if $\Delta^d \chi_t$ is a stationary ARMA (p, q) process. A zero-mean autoregressive moving average process of order (p, q) (ARMA (p, q)) is defined by (1).

$$x_t = \sum_{s=1}^{p} \emptyset_s x_{t-s} + \varepsilon_t + \sum_{s=1}^{q} \beta_s \varepsilon_{t-s} \qquad (1)$$

where ε_t is white noise process [8].

However, a problem with ARIMA is that it does not support seasonal data. It expects that the data is either not seasonal or has the seasonal components removed, eg., seasonally adjusted via methods such as seasonal differencing. Seasonality is a regular pattern of changes that repeats over m-time periods, where m is the period until the pattern repeats. For example, for monthly data, m = 12 is the span of the periodic seasonal behavior, and m = 4 times per year for quarterly data. Seasonality usually causes non stationarity of the time series data since the average values at sometime within the seasonal span may be different from the average values at other time. Thus, the seasonal differencing is used to make the time series stationary.

Since the ARIMA process cannot support seasonal data, then the Seasonal Autoregressive Integrated Moving Average (SARIMA or Seasonal ARIMA) is used to replace. The seasonal ARIMA model contains both non-seasonal and seasonal components. It can be denoted as (2).

$$\text{ARIMA} \quad \underbrace{(p, d, q)}_{\text{Non-seasonal part}} \quad \underbrace{(P, D, Q)_m}_{\text{Seasonal part}} \qquad (2)$$

Three trend parameters are the same as the ARIMA model:

- p: Trend autoregression order
- d: Trend differencing order
- q: Trend moving average order

Four new hyperparameters are added to ARIMA to capture a seasonal component:

- P: Seasonal autoregressive order
- D: Seasonal differencing order
- Q: Seasonal moving average order
- m: Period of repeating seasonal pattern

The trend parameters can be chosen from the lags of the ACF and PACF. Similarly, the seasonal parameters can be chosen from the seasonal lags of the ACF and PACF, and m can be chosen from the frequency of data [9]. Detailed results in respect of the parameter chosen based on ACF and PACF is included in section 3.1.

2.4. Logistic Regression

A logistic regression model [10] is used to estimate the probability of death amongst all confirmed cases in

Ontario based on their age group and vaccination condition. The logistic regression model is with four features finally, and the model formula can be written as (3):

$$ln\frac{p(death)}{p(alive)} = \alpha + \beta X$$

$$\Rightarrow ln\frac{p(death)}{1 - p(death)} = \alpha + \beta X$$

$$\Rightarrow p(death) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \qquad (3)$$

where $\beta\mathrm{X} = \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \beta_4\chi_4$

Specifically, α represents age group above 60 years old, $\beta_1$represents age group under 19, $\beta_2$represents age group between 20 to 39, β_3 represents age group between 40-59, and β_4 represents vaccination dose taken. Age group above 60 years old is included as constant to remove multicollinearity in regression. Model parameters and probability results are shown in section 3.2.

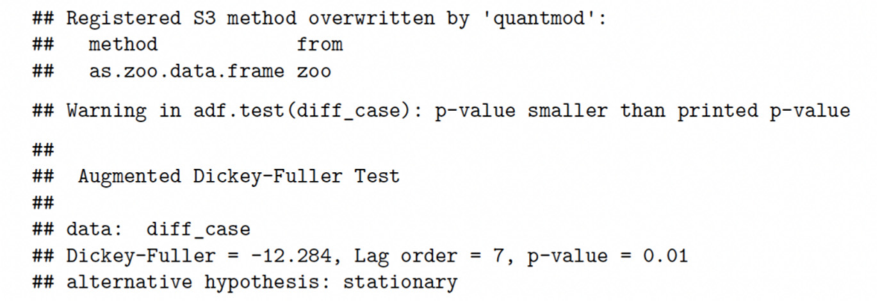
A ROC curve (receiver operating characteristic curve) is plotted after fitting the model [11], for which a ROC curve shows the performance of a classification model at all classification thresholds. This curve plots true positive rate and false positive rate can be determined, from which the classification accuracy. Moreover, Area under the ROC Curve (AUC) is also calculated to provide an aggregate measure of model performance [12]. AUC ranges in value from 0 to 1. A model whose predictions are all wrong has an AUC of 0, whereas all correct predictions have an AUC of 1.

3. Result

3.1. Time Series Results

3.1.1. Stationary Test

The ADF test is going to see whether the time series is stationary [13]. The p-value of the ADF is 0.3166, which is greater than 0.05. Thus, the data is not stationary. This might be because of some trend or seasonal components, so the time series pattern can be used to verify. (as shown in figure1 and figure 2).



```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Warning in adf.test(diff_case): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  diff_case
## Dickey-Fuller = -12.284, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

Fig. 1 Summary of Stationary Test

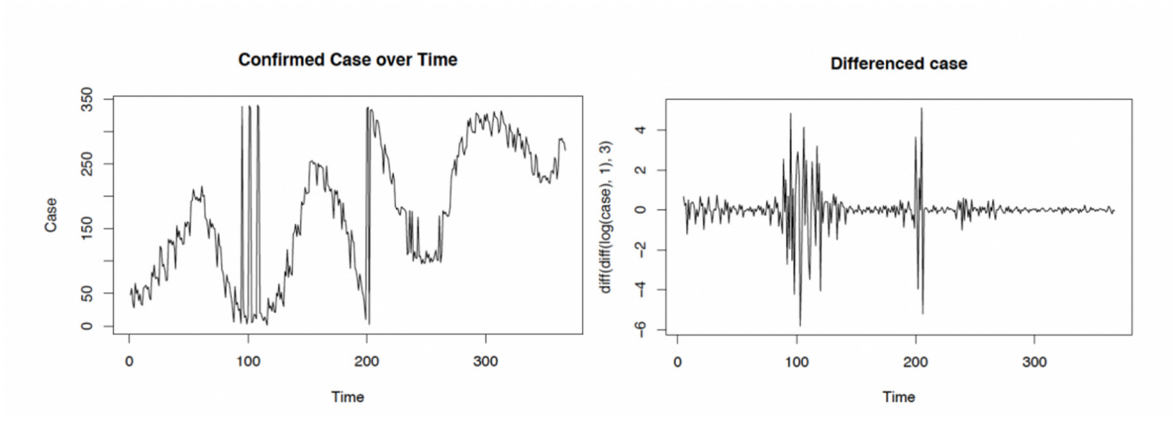


Fig. 2 Time series plot of original data and first-differenced data

It is obviously that the time series shows a long-term trend and roughly three-month-based periodicity as shown in figure 1. Thus, to remove these components by differencing. From the time series plot for differenced personal claim frequency, the data is up and down around zero. To conduct the ADF test for the differenced frequency, and obtain the p-value of 0.01, which is less than 0.05. Therefore, the differenced frequency is stationary, and it can be used for further time series analysis.

3.1.2. Seasonal ARIMA Model

The ACF and partial ACF for the differenced frequency data are drawn to give some hints for the choice of order of the ARIMA model. Due to the seasonal component and periodicity, the paper chooses the Seasonal ARIMA model [14]. From the plots as shown in figure 3, the order for the AR component is 6, since that the partial correlogram is not significant after the 6th lag. The four models are fitted by selecting different sequences for Ma components and finding the best model. Since this is roughly a three-month periodic data, the frequency 3 is the most appropriate. From R output, the absolute value of AIC is largest when the MA order is three. Thus, the best model is ARIMA (6,1,3) (0,1,0) [3] (as shown in formula 2).

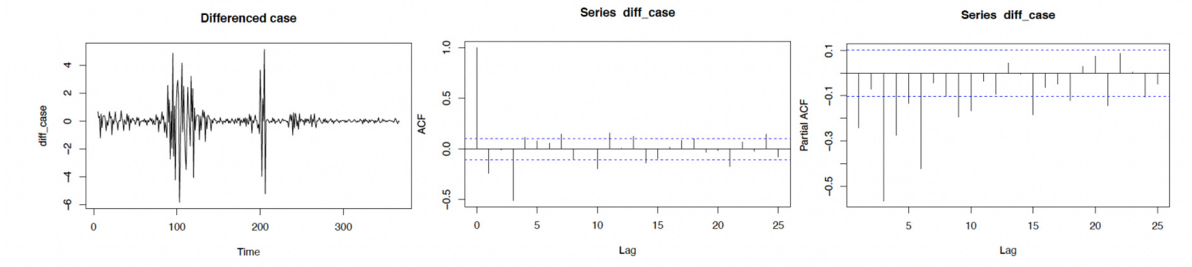
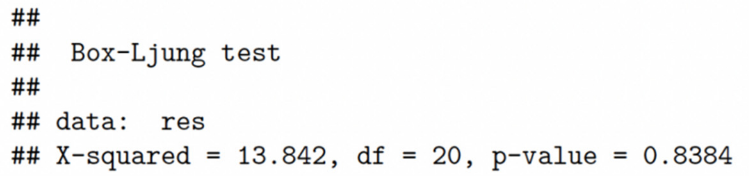


Fig. 3 Auto Correlation Function and Partial Auto Correlation Functions Plot

(1) Residual White Noise Test

Based on these models, the time series plot of residuals of the best model, all the residuals [15] are around zero. All the correlograms are within the blue dashed lines. Moreover, the White Noise Test for the residuals of the best model needs to conduct. The p-value of the Box-Ljung test is 0.8384, which is greater than 0.05 [16]. Thus, the result does not reject the null hypothesis and conclude that the residuals of the best model are close to White Noise.



```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 13.842, df = 20, p-value = 0.8384
```

Fig 4 Summary of Box-Ljung Test

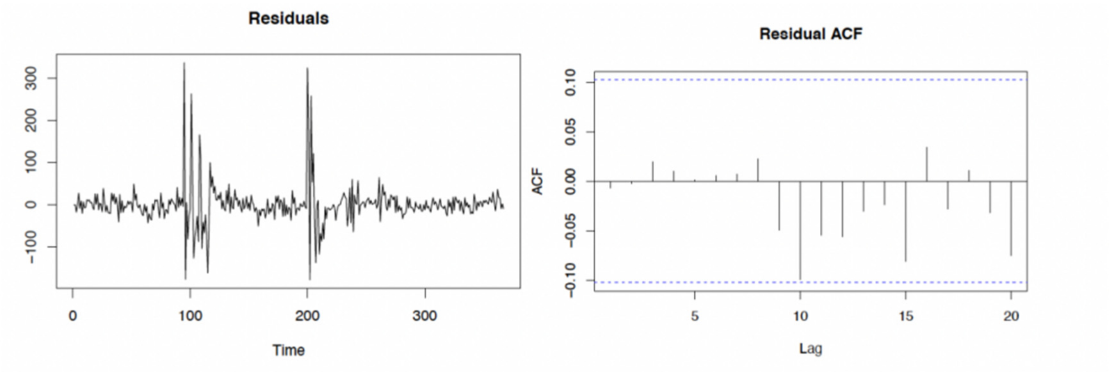


Fig. 5 Residual of data and Auto Correlation Function Plot

(2) ARCH Effect Test

This paper conducts a McLeod-Li test [17] on residuals to see whether the ARCH/GARCH model for the data. The p-values for all lags are less than zero, which implies an ARCH effect. To handle the ARCH effect, the ARCH or GARCH model for non-constant variance can be used in the future [18], which depends on the PACF of the squared residuals of the SARIMA model above. Through the preliminary investigation of short-term trends, the ARCH effect can be ignored, so this is the next step in predicting the long-term change of confirmed cases in Ontario.

(3) Forecast

Then, the best model can make short-term predictions and the trend of the confirmed case in the next 15 days. The expected result is shown in figure 6, there is a continuous upward trend for the following 15 days.

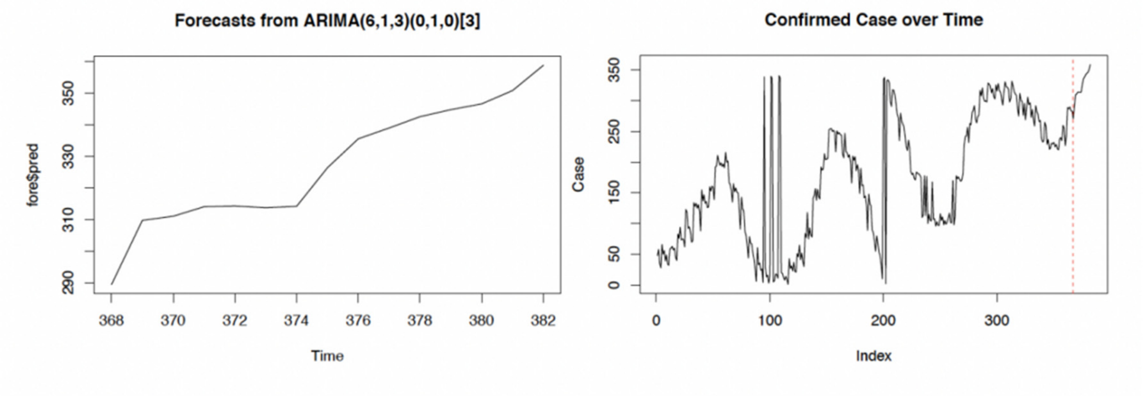


Fig. 6 Short-term Prediction and Overall Time Series Plot with Prediction included

3.2. Logistic Regression Results

3.2.1. Model Parameters

As shown in Result 1 and Result 2, all parameters are significantly based on p-value, whereas age group parameters present a higher significance level compared with vaccination. Therefore, the vaccination parameter as an ordinal parameter can be interpreted as negatively affecting the mortality rate as shown in table 1. On the other hand, with more doses taken by a COVID patient, less mortality is observed on the individual.

Table 1: Mortality of Confirmed COVID Cases among Age and Vaccination Condition

Age Group	*Not*	*Partially*	*Fully*
vs Vaccination	*Vaccinated*	*Vaccinated*	*Vaccinated*
19-	0.01%	0.05%	0.02%
20-39	0.09%	0.04%	0.02%
40-59	0.6%	0.03%	0.01%
60+	4.9%	2.24%	1.02%

Log of odds ratio can be directly computed by using the fitted parameters of logistic regression. The probability of death is further derived and calculated based on the odds ratio. With the same vaccination condition, people with older age tend to have substantial mortality once they are infected with COVID-19. Moreover, comparing vaccination status for the same age group proves that vaccination helps reduce mortality risk among all confirmed COVID-19 cases in Ontario.

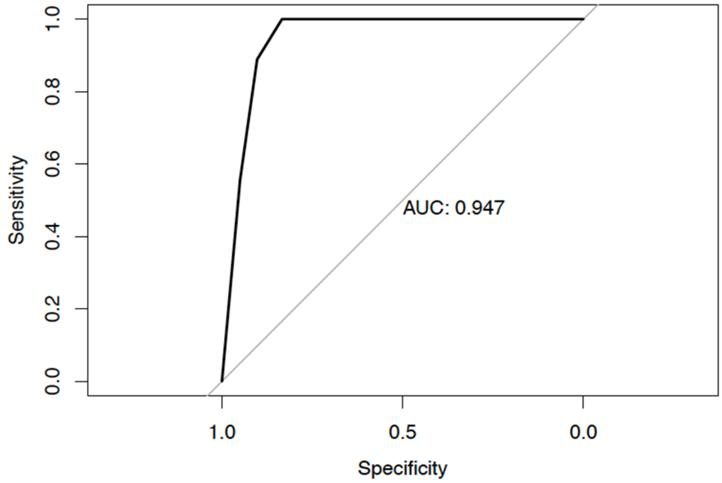


Fig. 7 Receiver Operating Characteristic Curve

The Figure 7 shows that the area under the ROC curve is 0.947, which is close to 1. Therefore, the logistic regression would present a reliable classification result.

4. Conclusion

This paper will fit past one-year data of confirmed cases in Ontario on the Seasonal ARIMA model by predicting the COVID trend in Ontario for the next 15 days preliminarily. The predicted trend in the short run is expected to be upward sloping based on information as of November 06, 2021. Furthermore, age group and vaccination status significantly impact the probability of death amongst confirmed cases. The mortality gradually increases as age develops. It has also been proved that fully vaccinated patients witness lower mortality than partially vaccinated patients, followed by no vaccination taken. All surveys may be related to the new vaccination policy that came into effect on November 6, 2021, which recommends that people over the age of 70 be vaccinated with the third covid vaccine. Third vaccination could help to reduce mortality risk for the aged population. The logistic regression is further verified as a reliable model by showing a high AUC score.

The next goal is to fit case trends in a more accurate model capable of capturing the volatile and dynamic nature of confirmed cases in Ontario. The current model gives an overview of what to expect shortly. In contrast, a more complicated model should provide the long-term trend and, meanwhile, consider volatility exhibited in the previous path. In addition, since the logistic model for mortality, more factors can be included in the model later to investigate further how other factors could potentially impact mortality risk amongst all confirmed cases.

References

[1] 2021, 30 Oct. "Epidemiological Update: Coronavirus Disease (COVID-19) - 30 October 2021." PAHO/WHO | Pan American Health Organization. December 4, 2021. DOI: https://www.paho.org/en/documents/epidemiological-update-coronavirus-disease-covid-19-30-october-2021.

[2] World Health Organization. (n.d.). How do vaccines work? World Health Organization. Retrieved December 12, 2021, from https://www.who.int/news-room/feature-stories/detail/how-do-vaccines work?adgroupsurvey=%7Badgroupsurvey%7D&gclid=CjwKCAiAtdGNBhAmEiwAWxGcUlUJYGr6J_5phymkamPznLJvUhmjOfw6tTye3YO0zIg2gL1D8J4xmBoCMMoQAvD_BwE.

[3] Tsekouras, Phil. "Ontario Expands Eligibility for Third Doses of COVID-19 Vaccine." Toronto. CTV News, November 3, 2021. DOI: https://toronto.ctvnews.ca/ontario-expands-eligibility-for-third-doses-of-covid-19-vaccine-1.5650095.

[4] "Forecasting: Principles&nbsp; and&nbsp; Practice (2nd Ed)." 8.9 Seasonal ARIMA models. December 4, 2021. DOI: https://otexts.com/fpp2/seasonal-arima.html.

[5] Molnar, Christoph. "Interpretable Machine Learning." 5.2 Logistic Regression, November 11, 2021. DOI: https://christophm.github.io/interpretable-ml-book/logistic.html.

[6] "Ontario COVID-19 Data Tool." Public Health Ontario. December 4, 2021. DOI:https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool?tab=trends.

[7] Forecasting: Principles and practice (2nd ed). 8.2 Backshift notation. (n.d.). Retrieved December 23, 2021, from https://otexts.com/fpp2/backshift.html

[8] Brownlee, J. (2020, August 14). White Noise Time Series with python. Machine Learning Mastery. Retrieved December 27, 2021, from https://machinelearningmastery.com/white-noise-time-series-python/

[9] Schoonjans, F. (2021, October 12). ROC curve analysis. MedCalc. Retrieved December 23, 2021, from https://www.medcalc.org/manual/roc-curves.php

[10] Molnar, C. (2021, December 19). Interpretable machine learning. 5.2 Logistic Regression. Retrieved December 23, 2021, from https://christophm.github.io/interpretable-ml-book/logistic.html

[11] Schoonjans, F. (2021, October 12). ROC curve analysis. MedCalc. Retrieved December 23, 2021, from https://www.medcalc.org/manual/roc-curves.php

[12] Narkhede, S. (2021, June 15). Understanding AUC - roc curve. Medium. Retrieved December 23, 2021, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[13] Stephanie. (2020, September 17). ADF -- augmented dickey fuller test. Statistics How To. Retrieved December 23, 2021, from https://www.statisticshowto.com/adf-augmented-dickey-fuller-test/

[14] 5.2 modeling seasonal time series. December 4, 2021. DOI: http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/xegbohtmlnode44.html.

[15] Forecasting: Principles and practice (2nd ed). 3.3 Residual diagnostics. (n.d.). Retrieved December 23, 2021, from https://otexts.com/fpp2/residuals.html

[16] Zach. (2020, October 15). Ljung-box test: Definition + example. Statology. Retrieved December 27, 2021, from https://www.statology.org/ljung-box-test/

[17] Ryu Dae SickRyu Dae Sick 1311 silver badge33 bronze badges, & Richard HardyRichard Hardy 52.4k1010 gold badges9393 silver badges213213 bronze badges. (1966, January 1). What is the null hypothesis of the McLeod and li test? Cross Validated. Retrieved December 27, 2021, from https://stats.stackexchange.com/questions/317556/what-is-the-null-hypothesis-of-the-mcleod-and-li-test

[18] Kenton, Will. “Autoregressive Conditional Heteroskedasticity (Arch).” Investopedia. Investopedia, December 3, 2021. DOI: https://www.investopedia.com/terms/a/autoregressive-conditional-heteroskedasticity.asp.