An analysis of regression models for predicting cumulative confirmed Covid-19 cases in the United States of America over time

Natasia Fernandez
M.S. Data Science
Montclair State University
Montclair, NJ, USA
fernandezn9@montclair.edu

Marc Stern
M.S. Data Science
Montclair State University
Montclair, NJ, USA
sternm2@montclair.edu

Sergey Zyl
M.S. Computer Science
Montclair State University
Montclair, NJ, USA
zyls1montclair.edu

Abstract*— Since the start of the novel coronavirus pandemic, healthcare systems around the world have struggled to cope with the influx of hospitalizations, limited medical supplies, and implementing adequate sanitary protocols. The purpose of this study is to review and analyze different types of regression models using time-series techniques to predict confirmed cases of the Covid-19 virus in the United States of America (USA). These predictions can help identify future trends of the virus, which can help the world prepare for further changes to control the pandemic. The results of the study suggest that Support Vector Regression (SVR) was the best fitting regression model using both Sklearn and the WEKA tool with a correlation coefficient of 0.99. With further research, the results can be a foundation to controlling the pandemic and creating more specialized health measures for the future.

Keywords*— Covid-19, Regression, Prediction, Time-Series, SVR, ARMA

I. Introduction

A. Background

The COVID-19 pandemic dictates its own rules around the world, causing many countries to incur enormous losses in many areas. The healthcare sector is experiencing a special burden since the start of emergence of the virus at the end of 2019. During the COVID-19 pandemic, the main difficulty is to prevent a critical overload of the healthcare system due to an unpredictable increase in hospitalization of patients, and as a result, a large-scale increase in requests for the use of limited medical resources. Therefore, predictive data mining can have a significant impact in helping predict the trends of the cases in efforts to prevent major outbreaks and future lockdowns [10].

B. Project Definition and Goals

Nowadays there are several main problems caused by COVID-19, on which most scientific research is focused: the development of an antiviral vaccine, the prognosis of the number of patients with COVID-19, the forecast of the spread of the virus. This study focuses on predicting the number of COVID-19 cases in the United States. The main task is to develop a regressive model that allows predicting COVID-19 cases over time. There are different types of regression analysis that we research, discuss, and implement in this paper. We focus on using time-series data with cumulative confirmed case counts as features.

C. Data Overview

This data set [1] was downloaded on December 1, 2021, and is created from the New York Times' Github page. The data ranges from January 21, 2020, through December 1, 2021, and is specific to the USA. It displays the US COVID-19 cases and deaths during this time period. There are 1,972,603 total rows in the dataset. The data set includes the following features:

- date: specific day
- county: name of county
- state: specific US state
- fips: a code for counties that may differ for metro areas
- cases: confirmed COVID-19 cases
- deaths: confirmed COVID-19 deaths

II. Related Works

Since the novel Covid-19 virus has entered the forefront of global society, it has also entered the forefront of the scientific community. Barría-Sandoval, et al., [2] implements and analyzes four types of time-series Regression models to predict Covid-19 case counts in Chile. It looks into ARIMA, Poisson, GLARMA, and Holt's Local Trend and Damped Trend models for predicting case counts and death counts. It does not consider cumulative case counts. Borchering, et al, [3] uses Ensemble learning to predict Covid-19 case counts with various features, including vaccination status, and hospitalizations, among others. Their research is implemented in two different interactive dashboards called the Covid-19 Scenario Modeling Hub, and the Covid-19 Forecast Hub.

S. Chan et al., [10] conducted a similar analysis in which the paper explores different count regression models to make short term predictions COVID-19 data from eighteen countries. The results showed that count regression models, such as Poisson and negative binomial regressions, are efficient for predicting cases because they are relatively simple to gather results and graph trends. Negative binomial distribution with log link function proved to be the most effective with the p-values of most countries being less than 0.05, indicating

statistical significance. This is similar to our study as it also analyzes regression models to predict COVID cases. However, our models use a different variety of regression models and do not only focus on count-based regression models.

III. Approach and Implementation

A. Linear Regression Approach

In this paper, predictive data mining is conducted using four different regression models, which are Linear Regression, ARIMA, Poisson, and Support Vector Regression. The first model we will look at is a Linear Regression model. Since the data show that case counts have consistently trended upward since the virus was first detected in the USA, it makes sense to use linear regression as a starting place for comparing regression models. Linear Regression is a simple model, using the formula y = mx + b to fit the model. This model will be implemented using a combination of scikit-learn's Linear Regression function [4], which implements an ordinary least squares Linear Regression, and statsmodels' Deterministic Process function [5] to create a "time dummy" to work with time-series data. The model is also created using the WEKA tool for comparison of how creating the model with two different platforms can possibly impact the results of the predictions.

B. ARMA Approach

The second model for comparison is called ARIMA, which stands for Auto-Regression Integrated Moving-Average. In Barría-Sandoval, et al., [2] it is established that the ARIMA model is a standard starting point, and determined to be an accurate one, when fitting Regression models with time-series data. The ARIMA model includes p, d, and q values, representing the AR, I, and MA models respectively. The implementation in this uses the AutoARIMA function from pmdarima [6], which automatically determines the ideal p, d, and q values for fitting the model. In this case, the AutoARIMA function determines p = 3, d = 0, and q = 2, which is confirmed by the ADFTest function [7] of the pmdarima module. This function outputs a value of 0.01, which is less than the default alpha value of 0.05, showing that the data is stationary, and therefore does not need differencing, which is represented by integration. When d is set to 0, the "I" in ARIMA is removed, and what is left is considered an ARMA model. The formula for the ARMA model is:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

C. Poisson Approach

The third model that will be implemented is the Poisson Regression model, which is often used for count-based data. Since this data deals with cumulative cases and deaths, it is considered count-based data. The model works by fitting the observed counts of the data to the regression matrix, which is done by expressing the rate vector as a function of the regression coefficients and matrix [8]. The Poisson formula is as follows:

$$PMF(y_i \mid x_i) = \frac{e^{-\lambda} * \lambda^x}{x!}$$

Where, λ is the expected number of times the event occurs, and x is the number of times it is predicted to happen in the future [8]. Using the Generalized Linear Models (GLM) class in the Python statsmodels package, the model is created to fit the training set and predictions are made on the testing set.

D. SVR Approach

The fourth model that is used to analyze the dataset and make predictions is Support Vector Regression (SVR). The idea behind SVR is to find the hyperplane within the decision boundary that includes the maximum amount of data points. The decision boundary line marks positive examples from negative ones and are each a certain distance away from the hyperplane. The model is created by taking these points within the boundary line that give the least error rate [9]. The model is created using both Scikit-learn and WEKA in order to compare how the different tools may generate different results.

IV. Experiments and Observations

A. Linear Regression Model

After the data set is split into a training and test set with about an 0.85 to 0.15 ratio, respectively, the Linear Regression model reports an r2 value (the coefficient of determination) equal to 0.56, rounded to two decimal places, on the test set. This scores only slightly above the random model threshold of 0.50, and below the scientific standard minimum threshold of 0.60 for valid machine learning models. Therefore we would not accept this Linear Regression model as a valid model for prediction. It is also worth noting that the nature of Linear Regression is for the model to predict case counts to increase perpetually. We accept this as a starting point since cumulative case counts will never decrease, but the linear fashion of the model also does not account for the possibility of new case counts decreasing, which would require the model to predict a flattening of the predictor line. The results are shown in Figure 1.

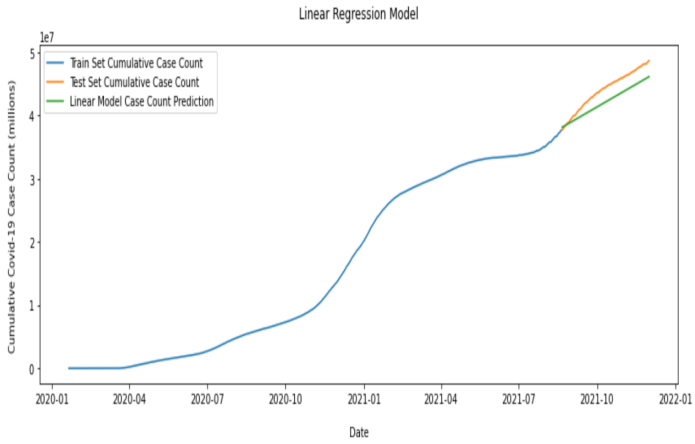

Figure 1. Linear Regression Model

Using the WEKA tool, the linear regression model is also created with the Linear Regression function in classifiers. For this model, a 10-fold cross validation is performed, which means that the model is fitted to the data ten times using a training to testing data ratio of 0.90 to 0.10. The correlation coefficient of this model is 0.9795, which is close to 1 indicating a high

correlation. According to the correlation coefficient, this linear regression is a valid model and can possibly be useful in predicting future cases. The model also only takes 52.9 seconds to build, which means it is a quick model to create. The results of the model are displayed in Figure 2.

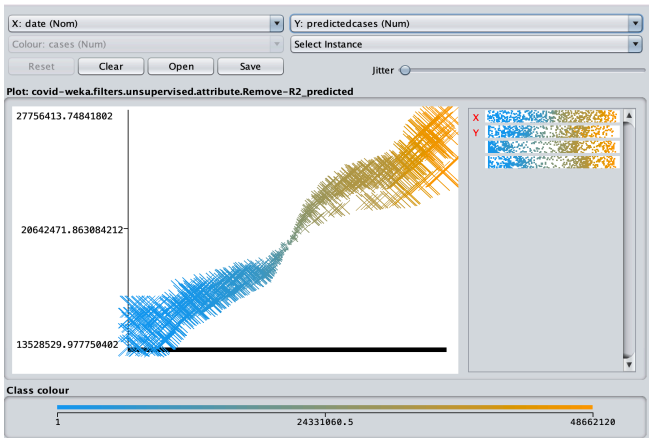

Figure 2. COVID-19 WEKA Linear Regression Model

B. *ARMA Model*

The ARMA model is implemented with an r2 value of 0.63, rounded to two decimal places, on the test set. This is a much more promising model than the Linear Regression, which we do not consider viable. While 0.63 is not as accurate as would be desired in predictive modeling, it is a viable model that is above the scientific minimum threshold of 0.60. Being that it is above 0.60, but not by much, this ARMA model is considered the baseline for predictive modeling in this study. The results are visualized below in Figure 3.

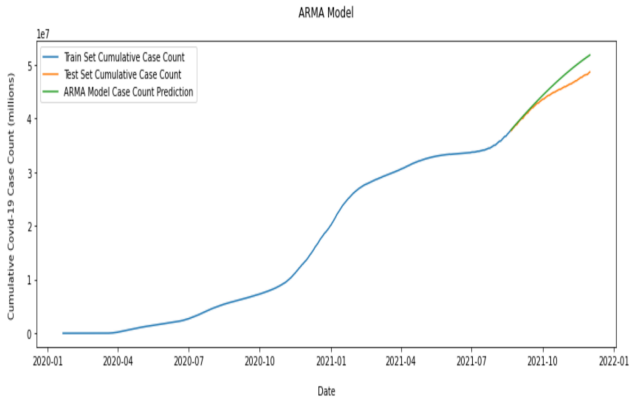

Figure 3. COVID-19 ARMA Model

C. *Poisson Regression Model*

For the implementation of the Poisson Regression model, the data is first split into a training set and testing set with a ratio of about 0.80 to 0.20. Then, using the Generalized Linear Models (GLM) class in Python statsmodels package, the regression model is configured to the training data and the Poisson model shown in Figure 4 is created [11]. The model is then tested to predict the test data. According to the results, the deviance and Pearson chi2 statistic is very large with 1.0213e+09 and 8.26e+08, respectively. The GLM class does not have a correlation of determination statistics in its summary table. Therefore, by looking at these large values, it seems like the Poisson model may not be the best fit for the dataset. However, the model does follow the same relative trend even though the numerical prediction may not be as accurate.

The predictions seem to be under-calculated in the middle and over calculated after around August 2021. Overall, the model follows the general trend in its predictions, but it is not very accurate for this dataset.

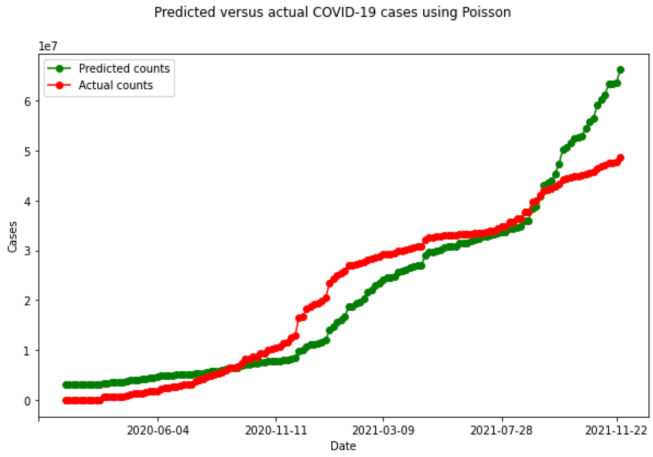

Figure 4. COVID-19 Poisson Regression Model

D. *Support Vector Regression Model*

For the SVR model, it is created with two different tools, which are Sklearn and WEKA. For the model created in Sklearn, feature scaling is performed to standardize the scales used and normalize the data [12]. This step is important for this type of regression because it is not a class type that is typically used. The default kernel, “rbf”, is used in this model. The coefficient of determination for this model was about 0.995, which means that features were highly correlated, and this model could be a good fit for this data. The graph of the SVR model created by Scikit-learn is shown in Figure 5.

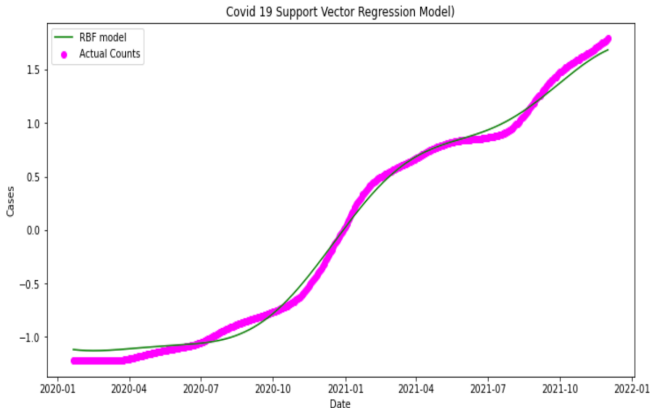

Figure 5. COVID-19 Support Vector Regression Model

The model follows the actual data very closely, and with further research, could potentially be a suitable model for predicting COVID-19 cases. For the SVR model created in WEKA, the SMOreg function was used in classifiers. The test mode used was a 10-fold cross validation and the model took about 0.58 seconds to create. The kernel used for the SVR model in WEKA is Polynomial Kernel, which fits the data within a curved line, instead of using a linear kernel. The correlation coefficient of this model was 0.9928, which also indicates a high correlation. The WEKA tool [13] generated a similar model to the Sklearn model as seen in Figure 6.

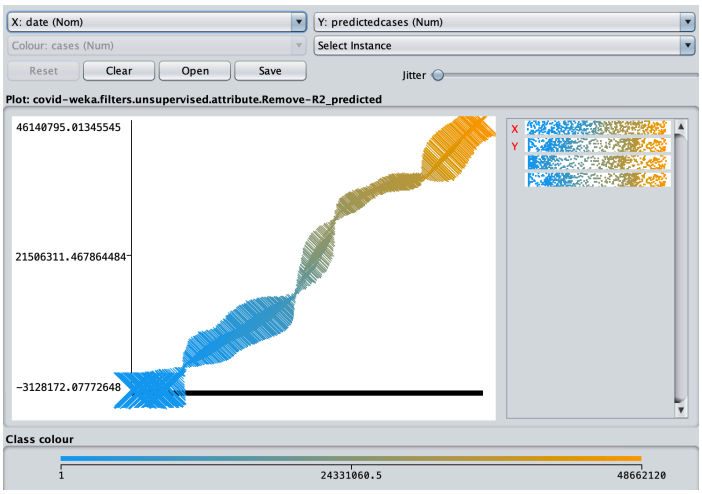

Figure 6. COVID-19 WEKA Support Vector Regression Model

The graph shows a similar trend to the Sklearn SVR model and has a similar correlation coefficient. The WEKA tool may be more advantageous since the tool facilitates the creation process of the model. The WEKA tool is quick and easy to use and generates a SVR model that portrays a good fit.

V. Conclusion

The main motivation of this study is to review and analyze different types of regression models using time-series techniques in order to predict confirmed cases of the Covid-19 virus in the United States of America. In this paper, we have analyzed four different time series models in predictive data mining. In particular, Linear Regression, ARIMA model, Poisson model, and Support Vector Regression have been proposed.

One of the important conditions for analyzing regression models is the availability of reliable data. If we analyze the intermediate results of the construction of mathematical models, then none of the simulated scenarios allows us to hope that it will be possible to obtain more accurate forecast data. That is why we used a big dataset that includes the information about confirmed cases and deaths from January 21, 2020, through December 1, 2021.

Linear regression was the first type of regression in our analysis. We implemented this model using two different approaches. The first one used a combination of scikit-learn's Linear Regression function and statsmodels' Deterministic Process function. This model reported an r2 value (the coefficient of determination) equal to 0.56 that is below the scientific standard minimum threshold of 0.60 for valid machine learning models. The second way for implementing Linear regression was WEKA. This linear regression is a valid model and can possibly be useful in predicting future cases. The ARMA model showed an r2 value of 0.63 on the test set. This model isn't as accurate as desired but can be used for predictive modelling. Poisson model predictions seem to be under-calculated in the middle of the time period and over calculated after around August 2021. However, the model follows the general trend in its predictions, but it is not very accurate for this particular data. The best model for predicting confirmed COVID-19 cases corresponds to the SVR model using both Sklearn and WEKA implementations. The correlation coefficient of this model was 0.9928 that indicated a high correlation.

In conclusion, we have presented several regression models that can be used to predict confirmed cases of Covid-19. But it is difficult to predict the number of confirmed cases of Covid-19 accurately from one type of data. However, it is important to understand that the main purpose of building mathematical models is to provide certain guidelines for determining a further strategy to be able to control trends. This paper will help to generate novel ideas of using data mining methods for accurate prediction of the number of confirmed cases of Covid-19. We strongly believe that this paper will be helpful to the researchers in developing an in-depth understanding of the research area.

References

[1] MyrnaMFL. (2021). *US counties COVID 19 dataset. NYT's github CSV on COVID19 per US Counties (Version 298).* [Online]. Available: https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset

[2] C. Barría-Sandoval, G. Ferreira, K. Benz-Parra, P. López-Flores. (2021). *Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques: A comparative study.* PLoS ONE 16(4): e0245414. DOI: https://doi.org/10.1371/journal.pone.0245414

[3] R.K. Borchering, C. Viboud, E. Howerton, et al. Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios — United States, April–September 2021. MMWR Morb Mortal Wkly Rep 2021;70:719–724. DOI: http://dx.doi.org/10.15585/mmwr.mm7019e3

[4] scikit-learn developers. (2021). *sklearn.linear_model.LinearRegression (1.0.1)* [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression

[5] J. Perktold, et al. (2021, December 10). *statsmodel.tsa.deterministic.DeterministicProcess (v0.14.0dev0 (+131))* [Online]. Available: https://www.statsmodels.org/dev/generated/statsmodels.tsa.deterministic.DeterministicProcess.html#statsmodels.tsa.deterministic.DeterministicProcess.out_of_sample

[6] T.G. Smith. (2021). *API Reference » pmdarima.arima.AutoARIMA (1.8.4)* [Online]. Available: https://alkaline-

ml.com/pmdarima/modules/generated/pmdarima.arima.AutoARIMA.html#pmdarima.arima.AutoARIMA

[7] T.G. Smith. (2021). *API Reference » pmdarima.arima.ADFTest (1.8.4)* [Online]. Available: https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.ADFTest.html

[8] "The Poisson regression model," *Time Series Analysis, Regression and Forecasting*, 29-Sep-2021. [Online]. Available: https://timeseriesreasoning.com/contents/poisson-regression-model/. [Accessed: 10-Dec-2021].

[9] "Support vector regression in machine learning," *Analytics Vidhya*, 01-Apr-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/. [Accessed: 11-Dec-2021].

[10] S. Chan, J. Chu, Y. Zhang, and S. Nadarajah, "Count regression models for covid-19," *Physica A: Statistical Mechanics and its Applications*, vol. 563, p. 125460, 2021.

[11] "Statsmodels.genmod.families.family.poisson¶," *statsmodels*. [Online]. Available: https://www.statsmodels.org/devel/generated/statsmodels.genmod.families.family.Poisson.html#statsmodels.genmod.families.family.Poisson. [Accessed: 11-Dec-2021].

[12] "Sklearn.svm.SVR," *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html. [Accessed: 11-Dec-2021].

[13] "Class SMOreg," *SMOreg (Weka-Dev 3.9.5 API)*, 21-Dec-2020. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/SMOreg.html. [Accessed: 11-Dec-2021].