De La Salle University - Manila
Term 1, A.Y. **2022 - 2023**

In partial fulfillment
of the course
NUMMETS - ER1

Term Project:
Analysis of Covid Cases and Vaccine in the Philippines using Regression Techniques

Submitted by:
Chua, Marcus Luis M.
Hernandez, Miro Manuel L.
Orga, Matthew R.
Ramos, Lorenzo Agustin B.

Submitted to:
Engr. Gerald P. Arada

Submitted on:
Dec 15, 2022

I. Project Description and Objectives

Covid-19 has affected our lives for the past few years and it has affected many people in the world. With the rapid development of covid vaccines by different companies from all over the world and its quick deployment to affected areas, the severity of Covid has been drastically reduced. But even with the application of the vaccine, there are still many people who still get infected with the virus, regardless of inoculation. The number of Covid cases also spikes in certain areas and time periods; It becomes difficult to pinpoint when Covid will be completely eradicated.

This project aims to find out about that change through data analysis and graphs with the utilization of least squares regression. The project will focus on the number of active cases in the Philippines per month from the year 2022 while comparing the number of vaccine doses per month of the same year. By comparing the graphs, the researchers can visualize the trend of covid active cases by checking if the regression of the two graphs shows an inverse relationship.

The objectives of this project are to utilize the least squares regression, compare the trend between the vaccine doses and active cases, and predict the possible trend for the year. By the end of this project, the researchers will find out if there is no relationship between the two graphs of vaccine doses and active cases. Finding out future trends for both would help in predicting the number of vaccine doses that the country will continue to apply and the possible rise of infection rates to avoid another case of quarantine.

II. Program/Codes and Program Optimization

The program used for our project is python which helped us in creating a linear regression with a certain number of data points.

Analysis of Covid Cases and Vaccine in the Philippines using Regression Techniques

by Group 2: Chua, Hernandez, Orga, Ramos

1. Importing modules and Excel Sheet

In [ ]:
```
import pandas as pd
import numpy as np
import datetime as dt
import os
import warnings
from collections import UserDict
from glob import glob
import matplotlib.pyplot as plt
from IPython.display import Image
%matplotlib inline
import matplotlib.dates as mpl_dates
import math
```

In [ ]:
```
df_cases = pd.read_excel("./Nummets Project.xlsx" , sheet_name="LinearRegression")
```

In [ ]:
```
df_cases.shape
```

In [ ]:
```
df_cases
```

Figure 1. Importing data into the Jupyter Notebook

In the python program, the researchers used data analysis libraries such as pandas, NumPy, and matplotlib to aid in the project.

2. Data Cleaning

In [ ]:
```
#df_cases["Numerical Mapping"] = "2022-01-" + df_cases["Months Numerical Mapping"].astype(int).astype(str)
df_cases["Months Numerical Mapping"] = df_cases["Months Numerical Mapping"].astype(int).astype(str)
df_cases
```

In [ ]:
```
df_cases["Numerical Mapping"] = "null"
for index in df_cases.index:
    value = df_cases["Months Numerical Mapping"][index]
    if len(value) == 1:

        df_cases["Numerical Mapping"][index] = "2022-01-0" + value

    else:
        df_cases["Numerical Mapping"][index] = "2022-01-" + value

df_cases
```

In [ ]:
```
df_cases["Numerical Mapping"]
```

In [ ]:
```
df_cases["Date"] =   pd.to_datetime(df_cases["Numerical Mapping"])
```

In [ ]:
```
df_cases["Date"]
```

In [ ]:
```
df_cases
```

In [ ]:
```
df_cases
df_cases.drop(["Months", "Numerical Mapping"], axis = 1, inplace = True)
```

In [ ]:
```
df_cases["Months Numerical Mapping"] = df_cases["Months Numerical Mapping"].astype(int)
df_cases
```

Figure 2. Creation of Variables and Data Setting

This part of the program initializes the table and the plot by mapping the different axes, and rows and columns values. x

3. Data Visualization

In [ ]:
```
# This program was taken from ALEKSANDR MOROZOV123 in the Kagel Website
# https://www.kaggle.com/code/aleksandrmorozov123/time-series-forecasting-with-python/notebook
# Modified at 15/12/2022 for the project purposes of Group 2

pd.options.display.float_format = "{:, .2f}".format
np.set_printoptions (precision = 2)
warnings.filterwarnings ("ignore")

# import lag_plot function
# Lag plots are used to check if a time series is random: random data should not exhibit any structure in the lag plot
from pandas.plotting import lag_plot
sns.set ()

# pass the lag argument and plot the values
# when lag = 1 the plot is essentially data [:-1] vs. data [1:]


# Creates the Lag Plot for the column Covid Cases 2022
print("Covid Cases Lag Plot")
lag_plot (df_cases ["Covid Cases 2022"], lag = 1)
```

In [ ]:
```
# by increasing the lag, we are checking for the seasonality of the data
print("Covid Cases Lag Plot")
lag_plot (df_cases ["Covid Cases 2022"], lag = 3)
```

In [ ]:
```
# Creates the Lag Plot for the column Vax Doses 2022
print("Vaccine Doses Lag Plot")
lag_plot (df_cases ["Vax Doses"])
```

```
def main():
    # observations / data
    Months = df_cases["Months Numerical Mapping"]
    covid_cases = df_cases["Covid Cases 2022"]
    vax_doses = df_cases["Vax Doses"]


    # estimating coefficients
    b = estimate_coef(Months, covid_cases)
    print("Estimated coefficients:\nb_0 = {}  \
          \nb_1 = {}".format(b[0], b[1]))

    # plotting regression line
    print("Linear Regression Plot of Covid Cases")
    plot_regression_line(Months, covid_cases, b)


    # estimating coefficients
    b = estimate_coef(Months, vax_doses)
    print("Estimated coefficients:\nb_0 = {}  \
          \nb_1 = {}".format(b[0], b[1]))

    # plotting regression line
    print("Linear Regression Plot of Vaccine Doses")
    plot_regression_line(Months, vax_doses, b)

if __name__ == "__main__":
    main()
```

Figure 3. Creation of the Regression Line Plots

The program will output a lag plot in showcasing the graphs. A lag plot is a different type of scatter plot with the variables being “lagged”. According to Glen, ”A “lag” is a fixed amount of passing time” (2018). Since one of our variables is time-related, the lag plot helps in plotting the data values to create a linear regression.

Another program and function used for the analysis and visualization of the data is Google Sheets. The function “FORECAST” was used to aid the constructed linear regression. The FORECAST function is used to predict a future value given a set of known values.

III. Considerations for Alternative Design

An alternative method that the researchers have thought of is the polynomial regression to properly graph the data that the researchers have retrieved. Polynomial regression allows us to gain a graph with a lot more curvature in order to properly see which part of the graph increased or decreased. However, having a curvature will result in a higher degree polynomial which has a chance of overfitting and underfitting the data (Selvi, 2021). This is a major reason as to why the researchers did not choose to use polynomial regression since the overfitting and under-fitting of the graph can hinder us in achieving one of our objectives. According to Abhigyan, ”A model that has been over-fitted, has poor predictive performances, as it overreacts to minor fluctuations in the training data” (2020). This project requires the predictive performance of the model to create a forecast graph of the two variables that the researchers have decided on.

IV. Analysis and Discussion

The researchers have gathered data from different sources, collecting values including the total number of covid cases in the country during the year 2022 and the total number of vaccine doses given during that year. By using these values, the researchers can predict the number of average daily covid cases by December 1, 2023, through the use of linear regression. The regression will be presented using Google Sheets Spreadsheet and Python codes.

Months	Active Covid Cases 2022	Vax Doses
January 2022	17378	575625
February 2022	200284	351428.5714

March 2022	51636	222580.6452
April 2022	38317	153142.8571
May 2022	51636	117200
June 2022	2320	139677.4194
July 2022	9112	124285.7143
August 2022	37060	133448.2759
September 2022	22245	112058.8235
October 2022	32748	74285.71429
November 2022	20935	33214.28571
December 2022	18298	28571.42857

Table 1. Covid Cases and Vaccination Doses

Sources: Mathieu et. al. March 5, 2020. . *Coronavirus (COVID-19) vaccinations*. Our World in Data. https://ourworldindata.org/covid-vaccinations?country=PHL&fbclid=IwAR0Y9AAx-7SQUhtRYMGjntjIpYFIorPtSSSM_e29vAXeR5UELwUEU5WEC-Q

and Worldometer. n.d. https://www.worldometers.info/coronavirus/country/philippines/?fbclid=IwAR21GYUZPw4wZKMZ79z9NXGQ-MEqsUvJP6izA2UvnOMec2FvUJyBYCSoHOA

The table presents the active covid cases per month and vaccination doses per month throughout 2022. These data will be the main basis for all visualization, and for all numerical techniques to be applied in the analysis.

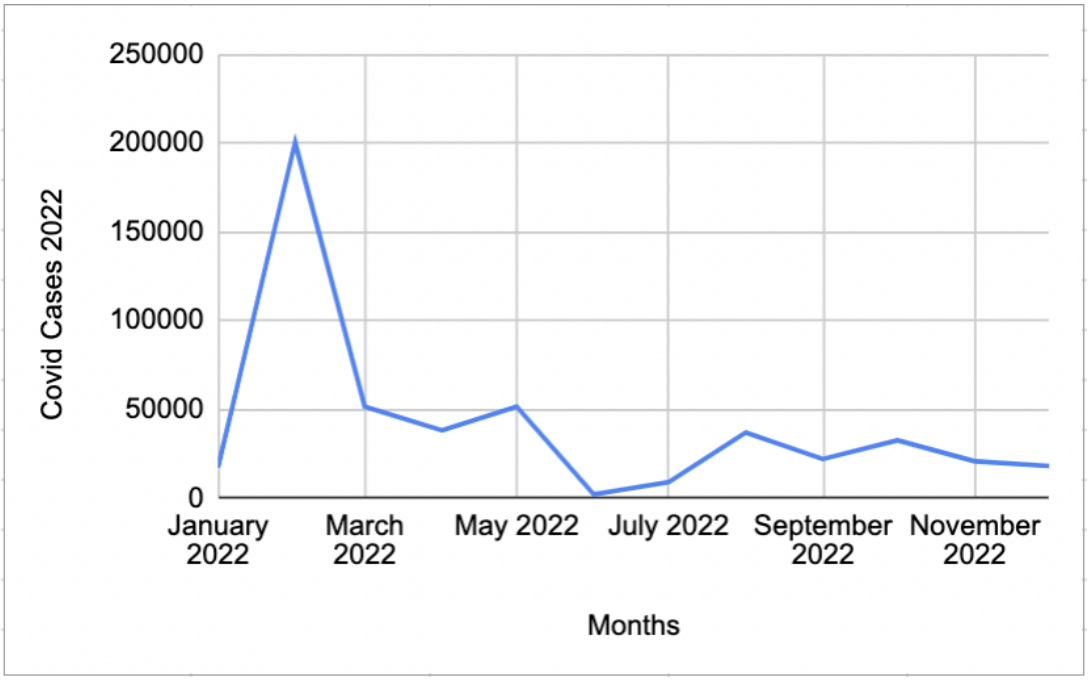

Figure 4. Active Covid Cases throughout 2022

The graph above shows the spreadsheet-generated projection of the active cases throughout 2022. The data shows the spike in February, which steeply declined afterward along with the administration of the vaccination.

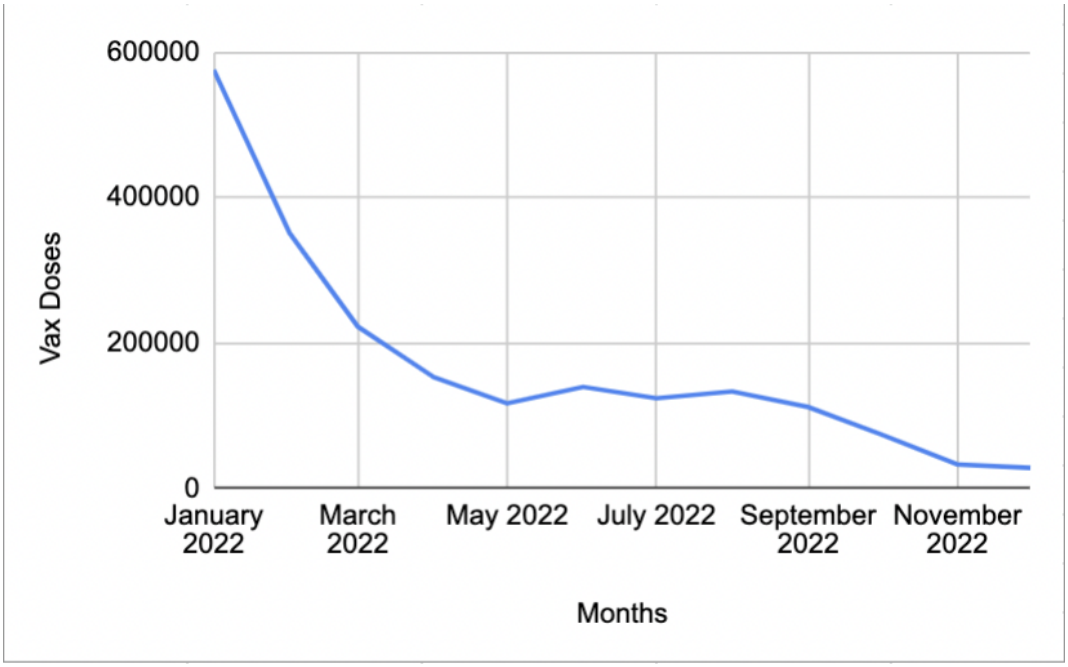

Figure 5. Vaccine Doses Administered throughout 2022

The graph shows the number of vaccination doses administered per month throughout 2022. From the range of the gathered data, January 2022 to the present, shows a downward trend

in the vaccination administration. This means that the peak vaccination doses were achieved at the start of January 2022.

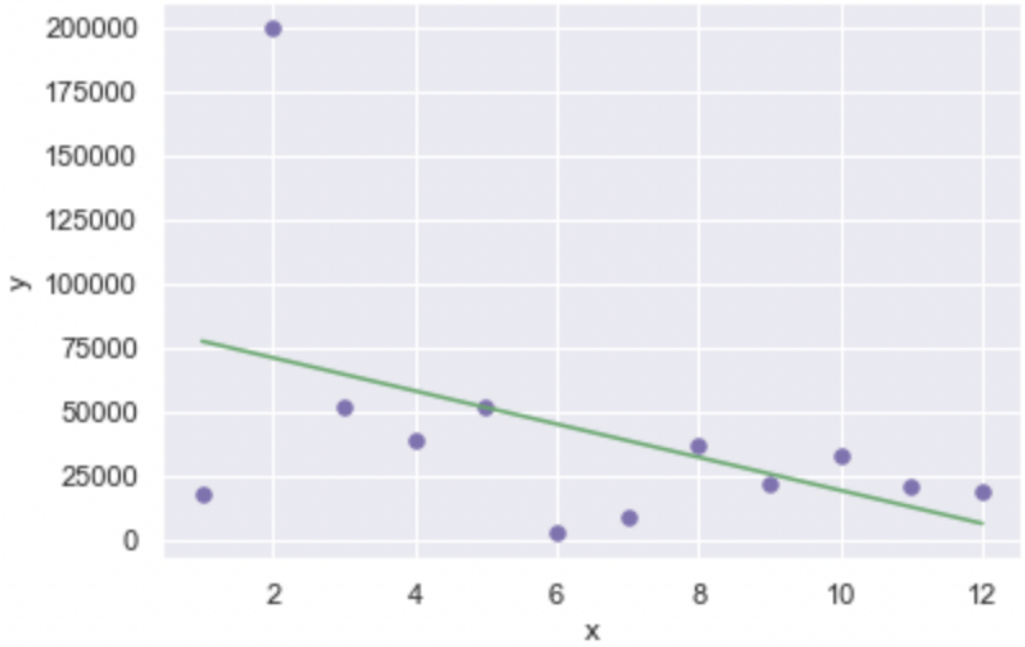

Figure 6. Linear Regression of Active Covid Cases

The linear regression that the researchers obtained from the data of active covid cases results in a downward trend. However, the researchers believe that this was the result due to the single spike of covid cases during February 2022, which is also called an outlier. According to Lemonaki, ”an outlier is an extremely high or extremely low data point relative to the nearest data point” (2021). Since the value of active cases during February is 200284, it is in the hundred-thousands place while the other values are only in the ten-thousands or lower which makes it an extremely high data point compared to the rest. This greatly affected our linear regression since, according to Choi, it can dramatically change the magnitude of regression coefficients (2009). This shows that the regression could have a flatter linearity instead of a sloped one. The presence of an outlier also affected our program and it concluded that the data values of the Active Covid Cases are randomly distributed which makes it impossible to create a forecast.

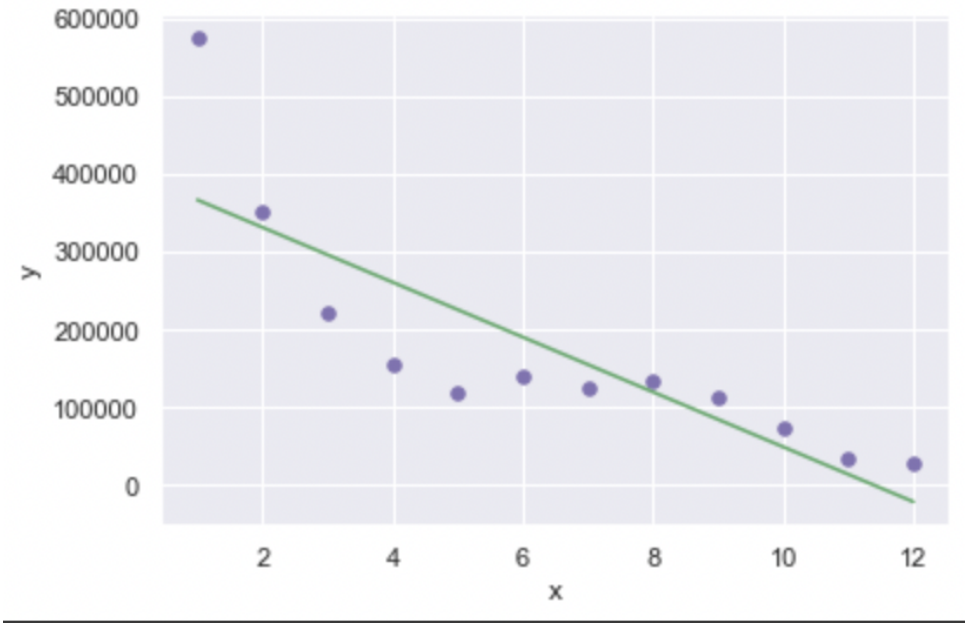

Figure 7. Linear Regression of Vaccine Doses

For the case of the linear regression of vaccine doses, it shows a downward trend, and the data points support this result as all the values steadily decrease in the number of doses. Moreover, even with a presence of a spike at the start, the difference between the high data point is not that large compared to the data point present in the values of covid cases.

Months	Covid Cases 2022	Vax Doses
January 2022	17378	575625
February 2022	200284	351428.5714
March 2022	51636	222580.6452
April 2022	38317	153142.8571
May 2022	51636	117200
June 2022	2320	139677.4194
July 2022	9112	124285.7143
August 2022	37060	133448.2759
September 2022	22245	112058.8235
October 2022	32748	74285.71429
November 2022	20935	33214.28571
December 2022	18298	28571.42857
January 2023	-169.7468129	-57388.18854
February 2023	-19359.40827	-48582.03734
March 2023	-5180.873651	-53328.10391

Figure 8. FORECAST Values

Upon using the FORECAST function to predict the Covid Monthly Cases and Vaccination Cases for January to March 2023, the result proved to be abnormal. Obviously, there cannot be a negative amount of active covid cases, and there cannot be a negative amount of vaccine doses. However, this does not mean that the linear regression was performed incorrectly, but the aforementioned limitations (not enough data points, outliers) were evident. The predicted values aligned with the trend of the linear regression which is a downward trend.

V. Conclusion

Through data analysis and the use of concepts we've learned in NUMMETS, we can safely conclude that there is little to no correlation between active Covid cases and the number of administered vaccines. Due to the sudden spike in active Covid cases in February 2022, it becomes an outlier in our data and prevents us from creating a forecast of active cases in future months. However, we discovered that there was a linear trend to vaccination, starting with peak administration in January 2022, and has been forecasted to continue to decrease for the following months.

With these findings, even with the high number of vaccinated individuals in the country, we still have to be careful going outdoors and socializing as Covid-19 is still present in the country. However, with vaccines, the severity of this virus for a person has been drastically reduced, and most individuals do not have to worry about experiencing intense symptoms if they contract the disease. We are now in the late stages of the pandemic. Little by little, through the efforts of medical professionals, the virus will soon be completely eradicated.

VI. Contribution of each member

Members:	**Contributions**
Chua, Marcus Luis M.	**Project Description and Objectives; Considerations for Alternative Design; Analysis and Discussion**
Hernandez, Miro Manuel L.	**Program Codes; Data Processing and Computation, Analysis and Discussion**
Orga, Matthew R.	**Program Codes, Data Processing, and Computation, Analysis and Discussion**
Ramos, Lorenzo Agustin B.	**Introduction, Project Description and Objectives, Data Gathering, Analysis and Discussion, Conclusion**

References:

Abhigyan. (2020, May 17). *Over-fitted and under-fitted models.* Medium. Retrieved December 15, 2022, from https://medium.com/analytics-vidhya/over-fitted-and-under-fitted-models-f5c96e9ac581

Choi, S.-W. (2009, January 3). *The effect of outliers on regression analysis: Regime type and foreign ...* Retrieved December 15, 2022, from https://whanchoi.people.uic.edu/TheEffectofOutliersonRegressionAnalysis.pdf

CFI Team. (2022, December 12). *FORECAST Function*. CFI. Retrieved from: https://corporatefinanceinstitute.com/resources/excel/forecast-function/

GeeksforGeeks. (2022, August 22). *Linear Regression (python implementation)*. Retrieved December 15, 2022, from https://www.geeksforgeeks.org/linear-regression-python-implementation/

Glen, S. (2018, June 7). *Lag plot: Definition, examples*. Elementary Statistics for the rest of us! Retrieved December 15, 2022, from https://www.statisticshowto.com/lag-plot/

Lemonaki, D. (2021, August 24). *What is an outlier? definition and how to find outliers in statistics*. freeCodeCamp. Retrieved December 15, 2022, from https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020, March 5). *Coronavirus (COVID-19) vaccinations*. Our World in Data. Retrieved December 15, 2022, from https://ourworldindata.org/covid-vaccinations?country=PHL&fbclid=IwAR0Y9AAx-7SQUhtRYMGjntjIpYFIorPtSSSM_e29vAXeR5UELwUEU5WEC-Q

Morozov123, A. (2022, September 9). *Time Series forecasting with python*. Kaggle. Retrieved December 15, 2022, from https://www.kaggle.com/code/aleksandrmorozov123/time-series-forecasting-with-python/notebook

Philippines. Worldometer. (n.d.). Retrieved December 15, 2022, from https://www.worldometers.info/coronavirus/country/philippines/?fbclid=IwAR21GYUZPw4wZKMZ79z9NXGQ-MEqsUvJP6izA2UvnOMec2FvUJyBYCSoHOA

Selvi, T. (2021, February 9). *Why -polynomial regression and not linear regression?* Numpy Ninja. Retrieved December 15, 2022, from https://www.numpyninja.com/post/why-polynomial-regression-and-not-linear-regression