Stage 04 - Member Report

Aman Tej Vidapu, Graduate student

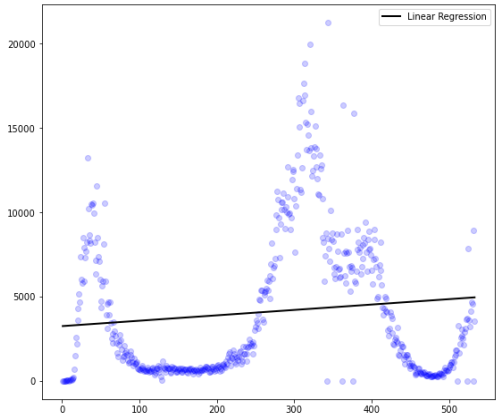
In this stage we are taking cases/deaths of a specific state or county starting from first registered case or death. I have selected New york for getting trend line and confidence intervals. Also selected new york top 5 counties to do same. So, we are mapping cases to a particular day like day 1 - 1, day 2 - 3, day 4 - 5 … so on. So, we are taking specific cases/deaths per specific day. And training a linear and non - linear regression model to predict further trends of this virus.

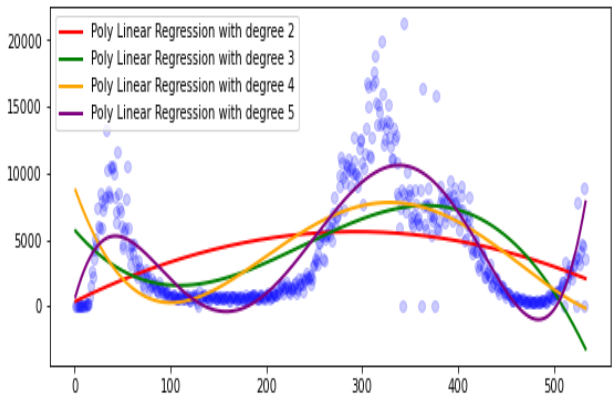
	Date	cases	on_particula_day
40	2020-03-02	1.0	1
41	2020-03-03	1.0	2
42	2020-03-04	9.0	3
43	2020-03-05	12.0	4
44	2020-03-06	2.0	5
...	...	...	...
568	2021-08-12	4701.0	529
569	2021-08-13	4591.0	530
570	2021-08-14	0.0	531
571	2021-08-15	8953.0	532
572	2021-08-16	3575.0	533

533 rows × 3 columns

In order to fit a linear model we are taking sklearn library ro get model. As we have data in terms of cases per day, inorder to get non linearity we will be using poly features and train the non-linear model. And, testing is done for all the data as we are working with limited or less training observations it is not suggested to split the data furthermore.

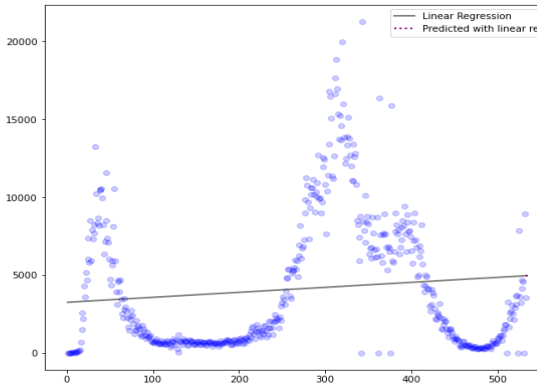
Trend Line: Developed Linear and Non-Linear (polynomial) regression models for predicting cases and deaths in terms of functions for reusability. We can call this function with changing certain attributes (mentioned details in notebook) to get appropriate outputs. So above data is passed to both models.

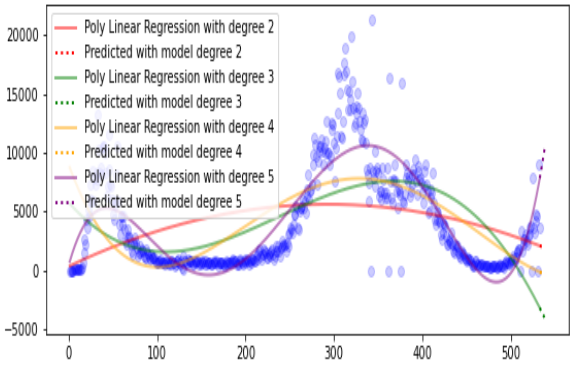




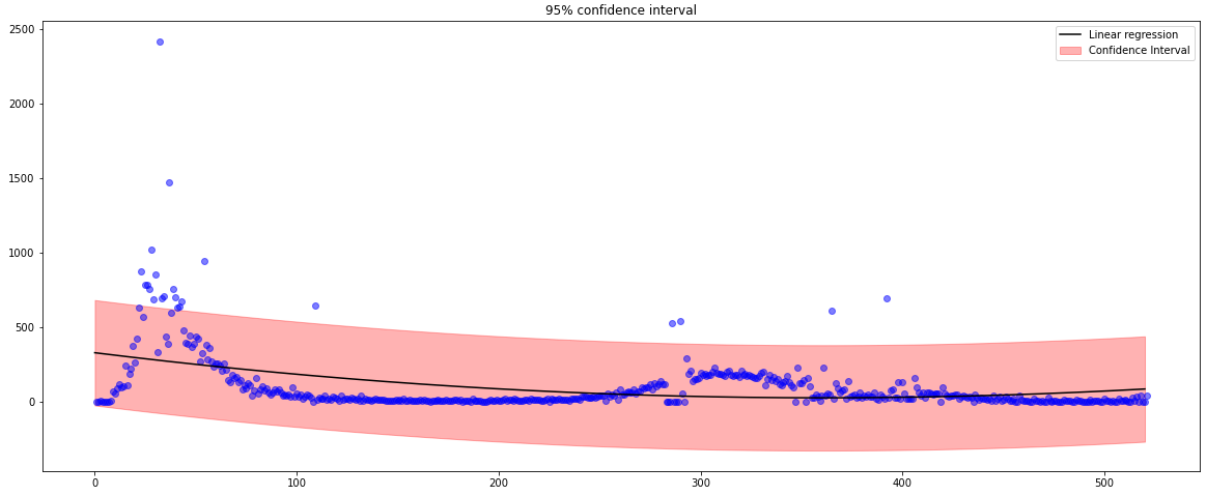
Calcluate error using RMSE:
The root mean square error (RMSE) is the residuals' standard deviation (prediction errors). Residuals are a measure of how far the data points are from the regression line; RMSE is a measure of how spread out these residuals are. In other words, it indicates how tightly the data is clustered around the line of best fit. We calculate this by calculating error with respect to actual y values to predicted ‘y’ values and applying a square root to it. By checking this we can estimate how good is our model.

Prediction path (forecast) and Confidence intervals (error in prediction):
We already trained the model with the data that is present, so we can predict the further values of cases/deaths count on a specific day by passing the new days to the model and plotted the predicts with actual trends for both models. Refer to dotted line for predicted values(forecast) in plot for both linear and non-linear models.





A 95 percent confidence interval is a range of values within which our prediction has a 95 percent chance of occurring. The standard deviation and a gaussian curve are used to compute this. We'll write a function to generate our confidence interval for a single sample, and then use it to calculate all of our predictions.



Point of no return:
It is a stage at where there are less resources than actual users, we are taking hospital data from (https://healthdata.gov/dataset/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/6xf2-c3ie) to calculate ICU beds utilization to number of deaths per a state. Also check trend line(predictions) may cross this stage.

Hypothesis testing: We are actually evaluating the hypothesis by conducting one - two tail tests which are considered in stage 03.

references :
https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/#:~:text=Root%20Mean%20Square%20Error%20(RMSE)%20is%20the%20standard%20deviation%20of,the%20line%20of%20best%20fit.