IAC 621 Project

Stage IV (Basic Machine Learning)

Reetika Sarkar

NOTE: For details of the team task, please refer to the team notebook and the team report posted in the GitHub repository for Team 4.

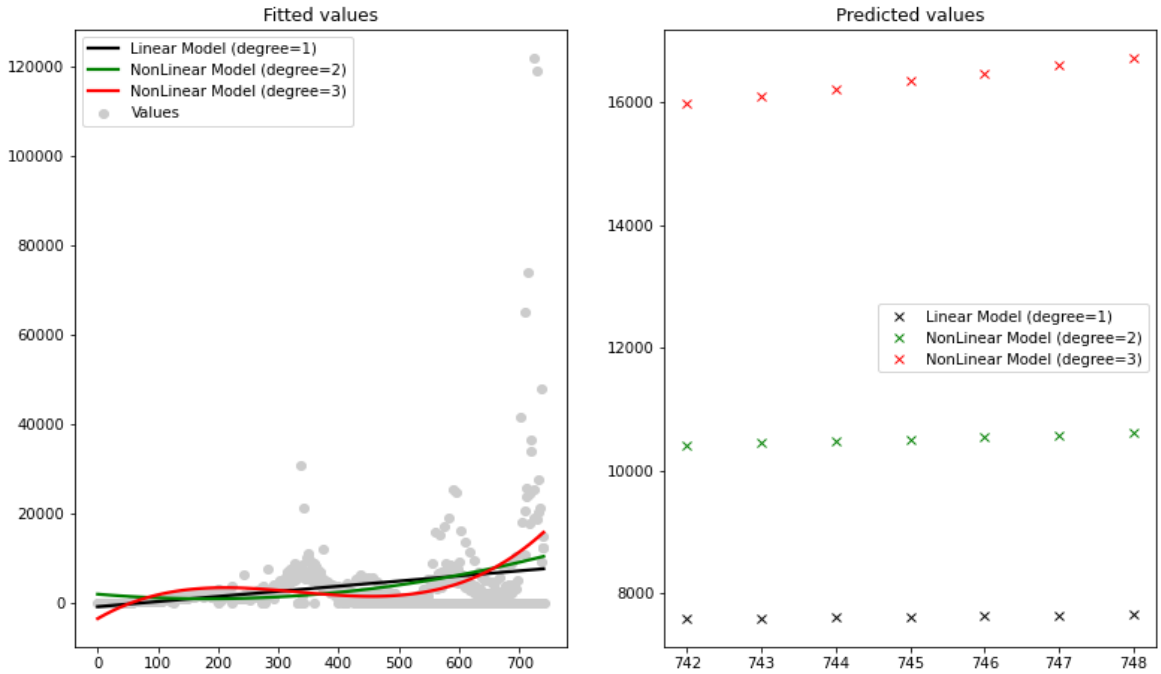
Member Tasks:

Task 1: Utilize Linear and Non-Linear (polynomial) regression models to compare trends for a single state and its counties (top 5 with highest number of cases). Start your data from the first day of infections. For each of the analysis plot trend line, confidence intervals (error in prediction) and prediction path (forecast).

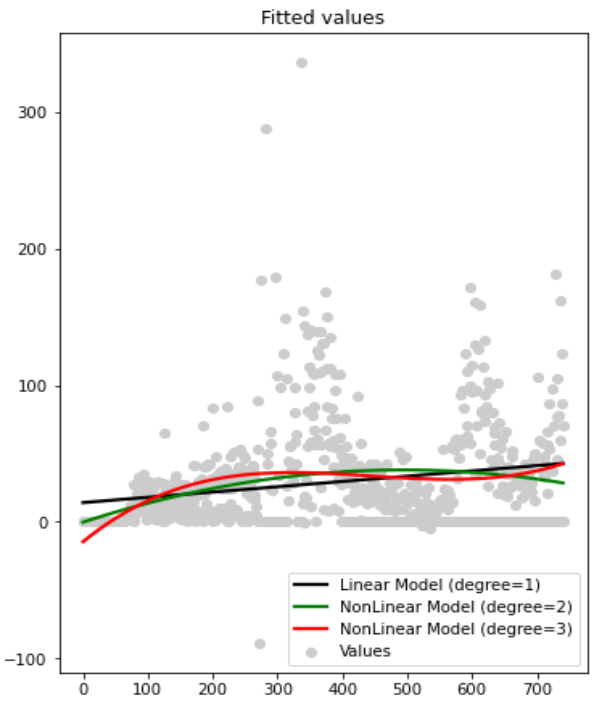
- I have computed linear and polynomial regression models and generated trend and forecast plots for the state of North Carolina (NC) and its top five counties with highest number of cases, namely, Wake County, Mecklenburg County, Guilford County, Forsyth County, Cumberland County.
- For fitting data I have defined a function `model()` that generates the estimated coefficients and RMSE based on linear regression and polynomial regression models of order 2 and 3.
- It also generates fitted (trend) plots and predicted (forecast) plot for all the models. The fitted plot shows the trend line and 95% confidence interval for each of the three models. The predicted plot shows the predicted counts for the next seven days after February 5, 2022.

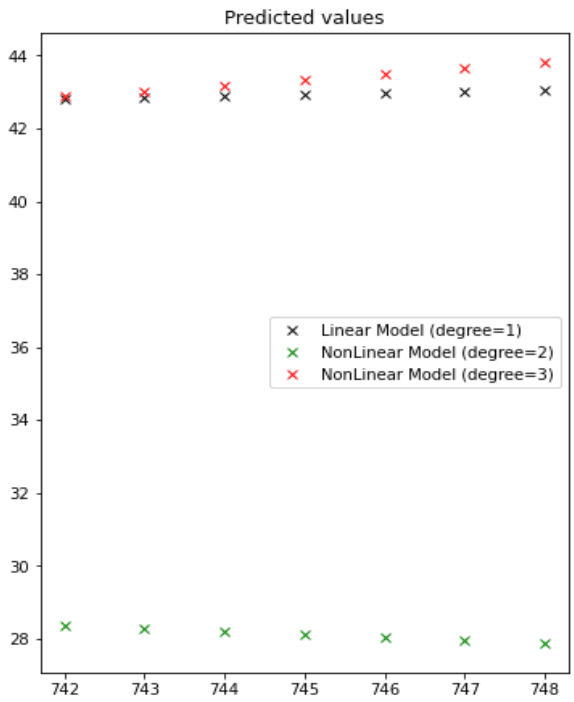
Modelling cases and deaths for the state of North Carolina

A. Cases



B. Deaths





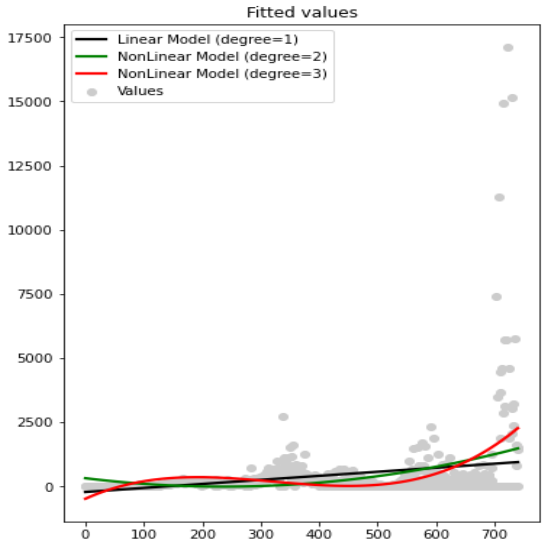
Interpretations:

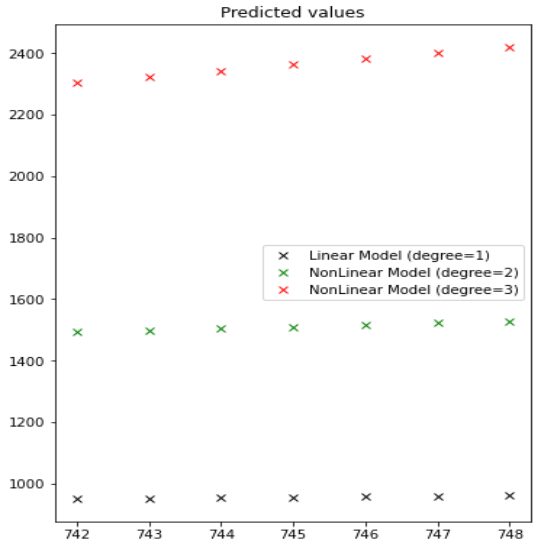
- Cases in NC have an upward trend as days progress, as do deaths (for linear model and nonlinear model of degree 3).
- The polynomial regression of degree 3 indicates an increase in cases followed by a decrease and then a peak which is the closest to the pattern of the data points.
- The forecasted cases and deaths are higher for the polynomial models than that for the linear model.
- Cases forecasted by linear model for the week ahead are the lowest while the nonlinear model of degree 2 predicts a decline in deaths.

Now we look at the top five counties with highest number of new COVID-19 cases.

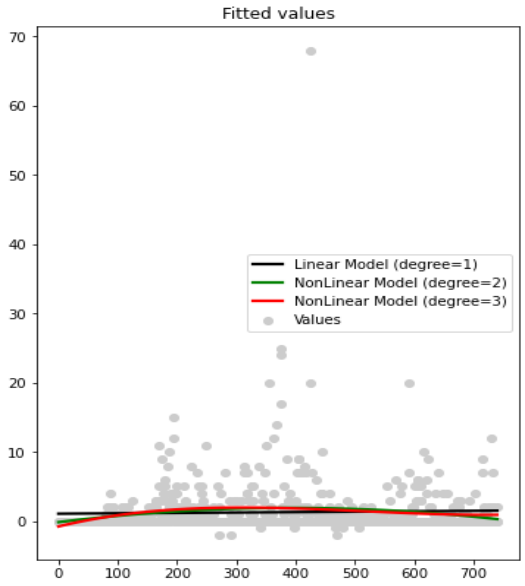
I. Wake County

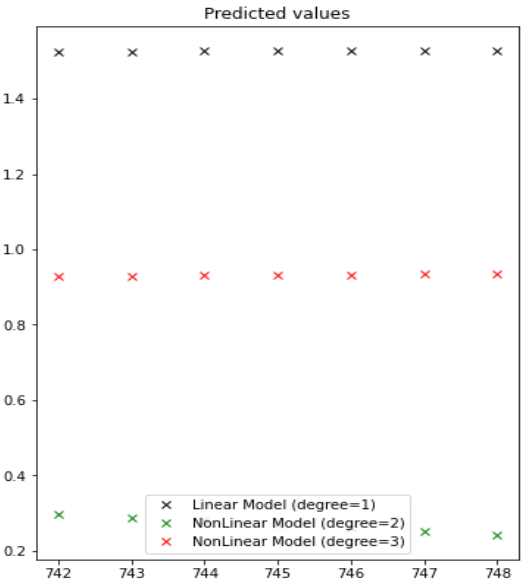
A. Cases





B. Deaths



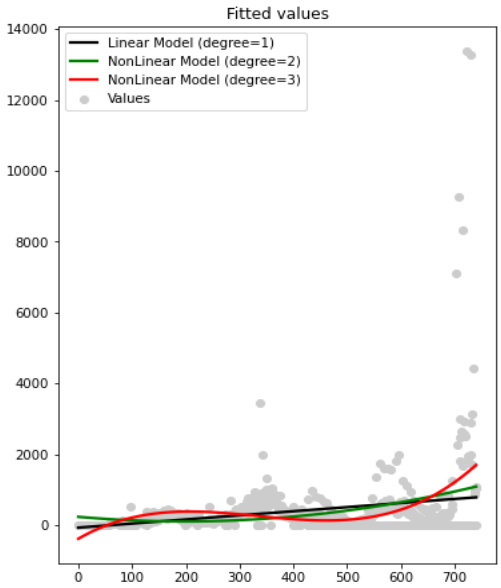


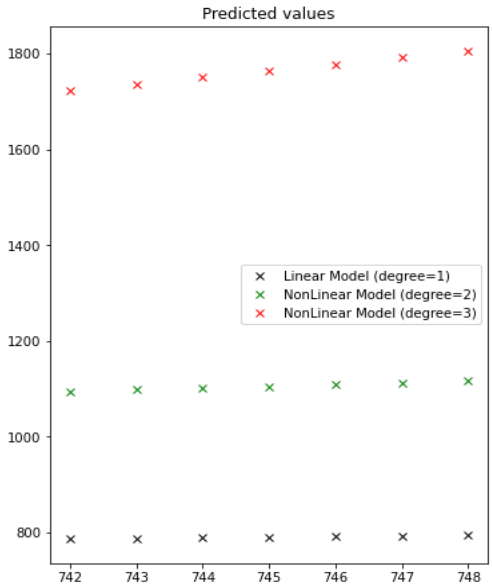
Interpretations:

- Cases have an upward trend as days progress, but deaths do not.
- The fitted death curves are very close for all the three models.
- The polynomial regression of degree 3 indicates an increase in cases followed by a decrease and then a peak which is the closest to the pattern of the data points.
- The forecasted cases higher for the polynomial models than that for the linear model. Forecasted deaths are lowest for nonlinear regression model of degree 2.

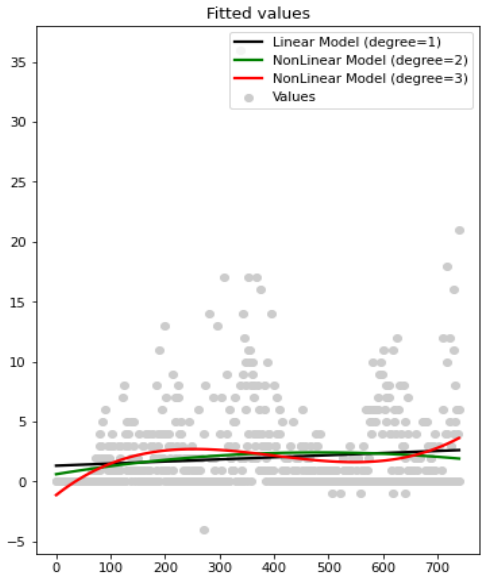
II. Mecklenberg County

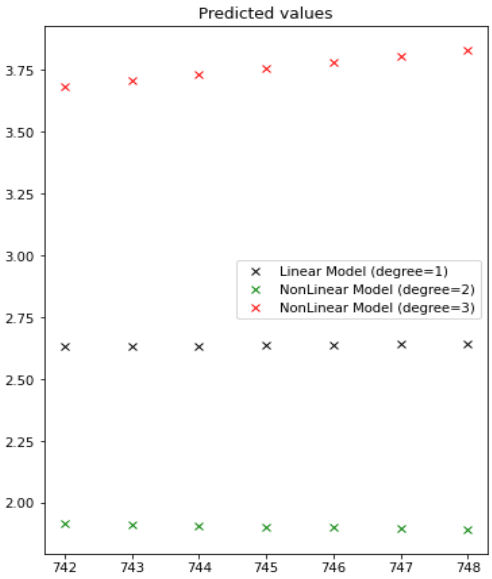
A. Cases





B. Deaths



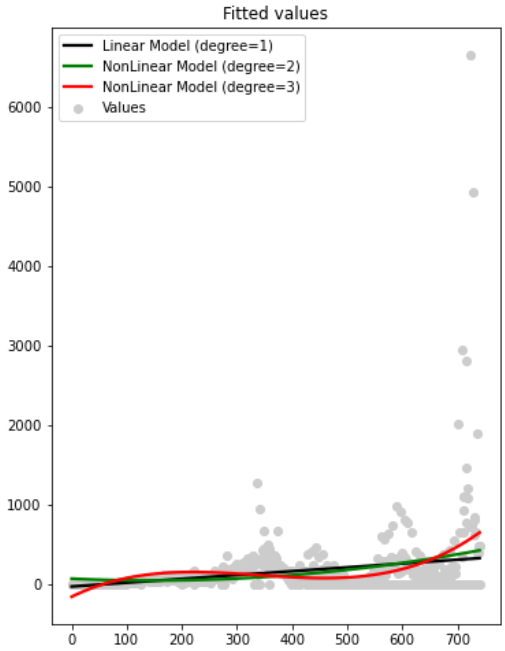


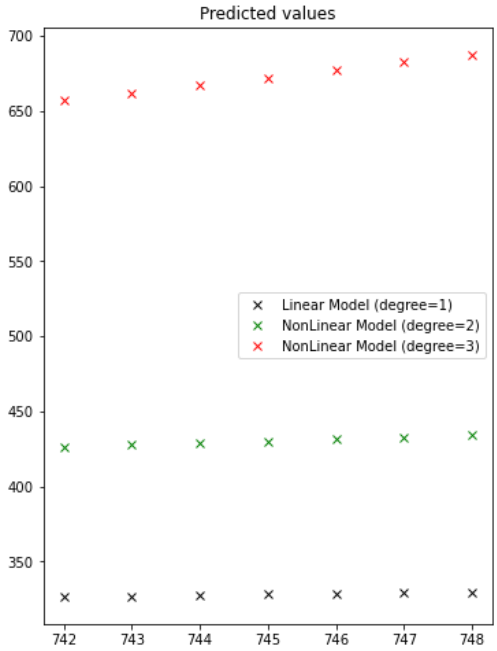
Interpretations:

- Cases have an upward trend as days progress, but deaths do not (except polynomial regression of degree 3).
- The polynomial regression of degree 3 indicates an increase in cases followed by a decrease and then a peak which is the closest to the pattern of the data points.
- The forecasted cases higher for the polynomial models than that for the linear model. Forecasted deaths are lowest for nonlinear regression model of degree 2.

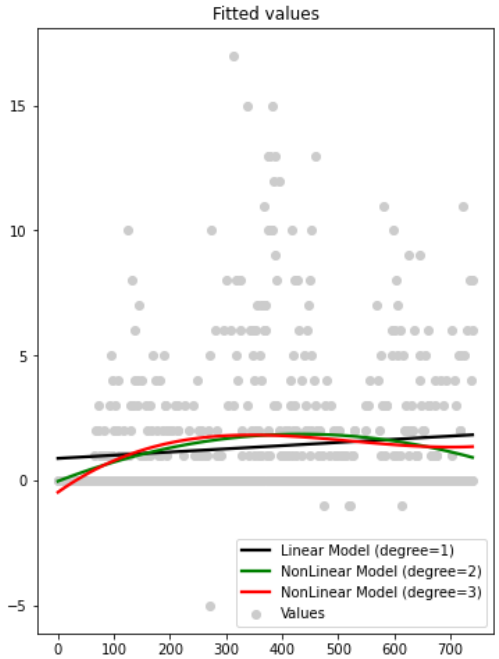
III. Guilford County

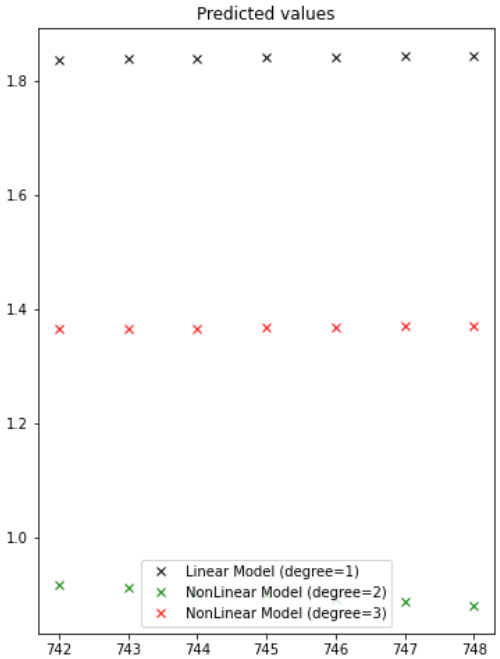
A. Cases





B. Deaths



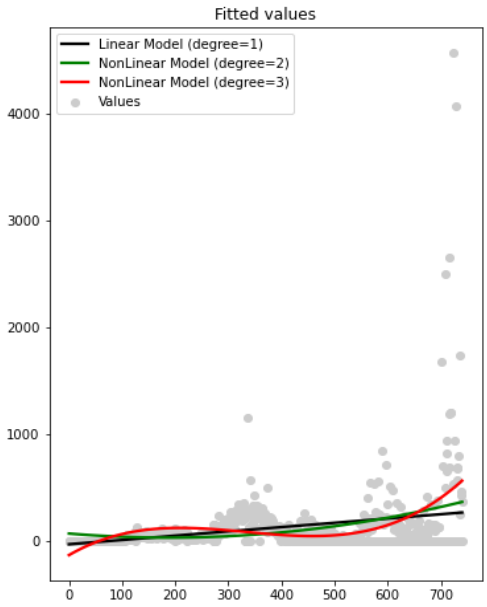


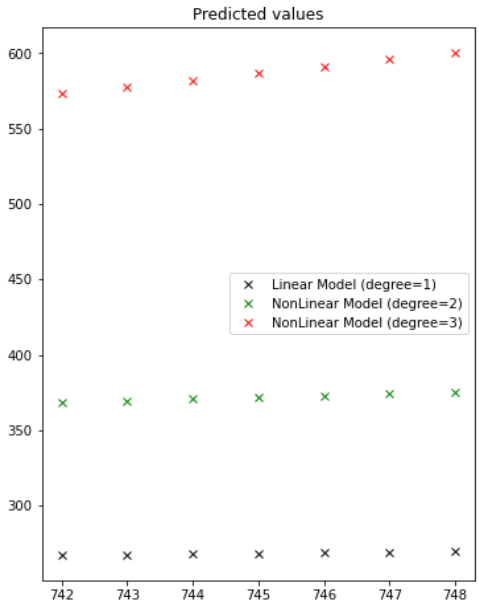
Interpretations:

- Both the cases and deaths plots are similar to the ones for Wake County, so similar interpreations can be made.

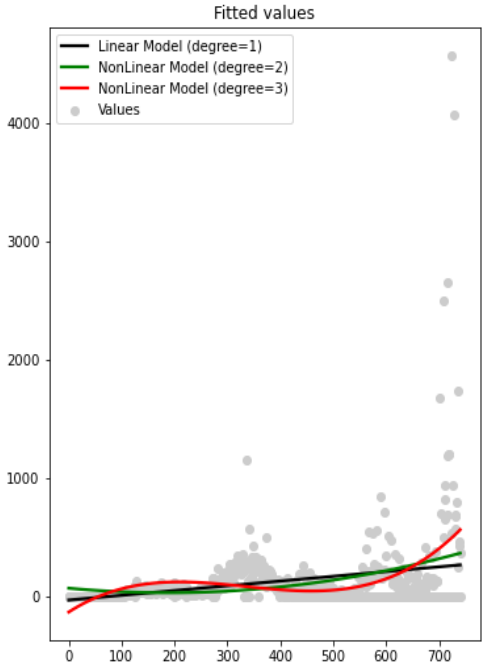
IV. Forsyth County

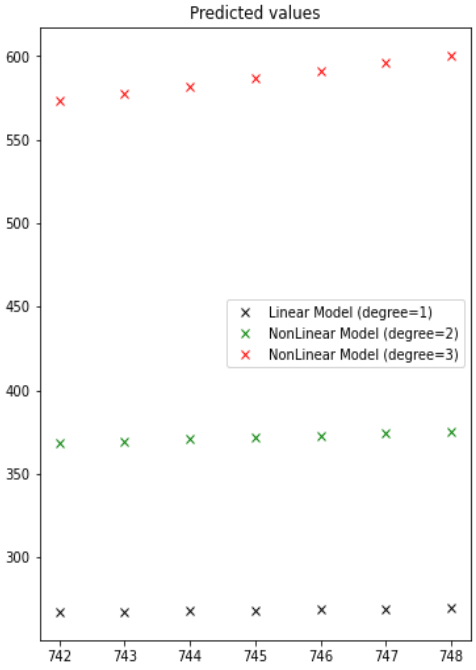
A. Cases





B. Deaths



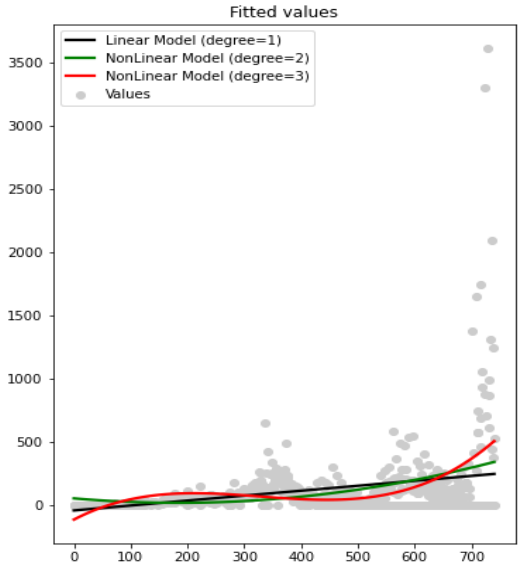


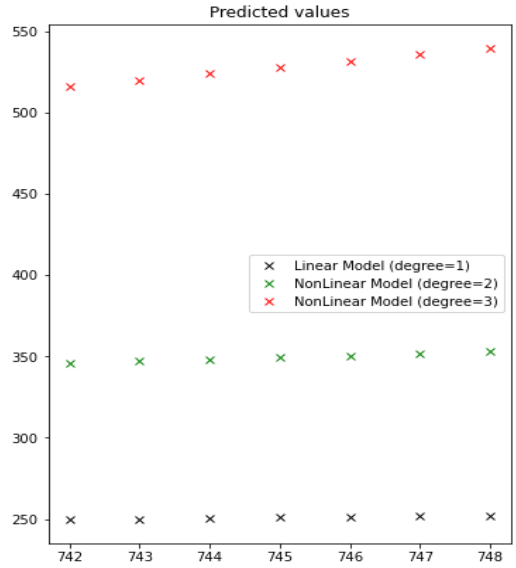
Interpretations:

- Both the cases and deaths plots are similar to the ones for Mecklenberg County, so similar interpreations can be made, the only visible difference is that the death forecasts by linear model are the lowest.

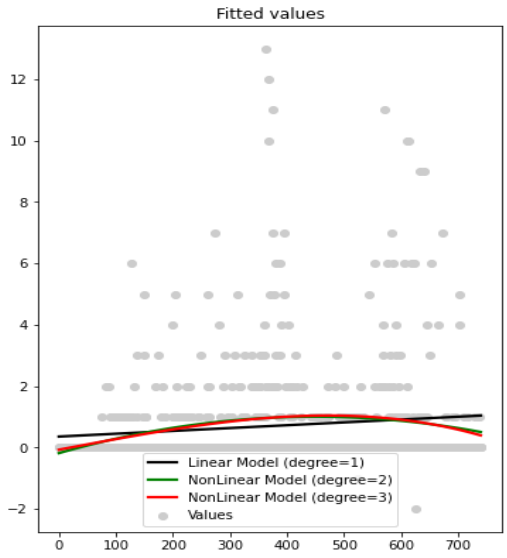
V. Cumberland County

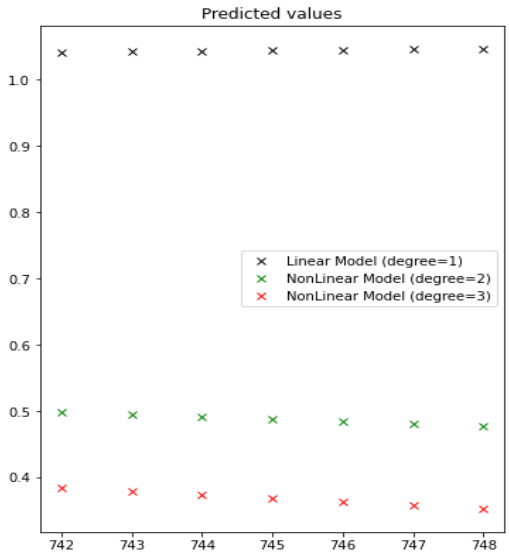
A. Cases





B. Deaths



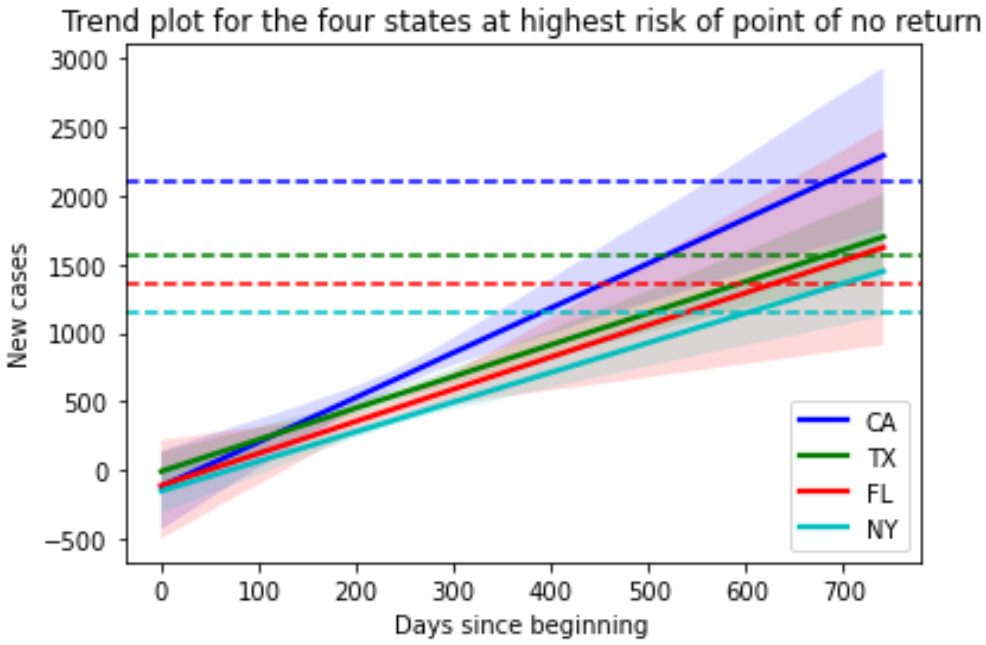


Interpretations:

- Both the cases and deaths plots are similar to the ones for Wake County and Guilford County, so similar interpreations can be made.
- The only visible difference is in the plot of forecasted deaths where nonlinear regression model of degree 3 has the lowest predicted death counts.

Task 2: Utilize the hospital data to calculate the point of no return for a state. Use percentage occupancy / utilization to see which states are close and what their trend looks like.

- Assuming that roughly 10% of all COVID-19 cases need hospitalization, we can say that if there are not sufficient hospital beds in the state for 10% of the average of predicted/forecasted cases in the week ahead, i.e., if 10% of the number of cases predicted from 02/06/2022 to 02/12/2022 exceed the total beds available in the state, it is a point of no return.
- During the week from 02/06/2022 to 02/12/2022, the mean predicted cases in NC that might have needed hospitalization (about 10% of total cases) was 761 while the total number of hospital beds in NC is 235. This indicates that there was a shortage of beds for COVID-19 patients in the state who needed hospitalization in this period.
- To look for states which were possibly at the point of no return, we could look at the difference between the average predicted COVID-19 cases needing hospitalization in the next week (estimated to be about 10% of total cases) and the average number of total hospital beds in the state. If the difference is **positive**, meaning that possible hospitalizations in the state in the next week are higher than the bed availability at hospitals there, it indicates a **shortage of beds** in that state.
- Four states with highest shortage of hospital beds for COVID-19 patients needing hospitalization are California, Texas, Florida and New York. These states have the highest shortage of beds available for COVID-19 patients in need of hospitalization.
- Looking at the trend of new cases in these states will give a clearer picture. The trend plot below shows that the cases in these states are indeed rising rapidly. The dotted lines show the point of no return for each state and each state crosses its point of no return.
- It may also be noted that February 2022 onwards, COVID-19 deaths were declining due to increased rollout of the vaccine and booster doses, so a point of no return may technically not have existed for any of the states.



Task 3: Perform hypothesis tests.

In stage 2 of the project, I proposed to perform hypothesis test based on the political leanings dataset. However, since two of the datasets I chose in stage 2 have few data points, I will be changing my questions as.

My updated hypothesis tests are:

(a) whether political leaning to Democrat and Republican party has any effect on new COVID-19 cases, and

(b) whether having a shortage of hospital beds has an effect on COVID-19 deaths.

(a) To test the first hypothesis, I used the dataset `diff_president_votes_states.csv` from stage 3 which indicates the difference in votes won by Democrats and Republicans. I recoded the differences as 1 or 0 based on whether the Democratic party won or the Republican party, respectively.

- The null hypothesis is that political affiliation has no association with the cases in the state. The alternate hypothesis is that there is an association.
- A decision regarding this is made by taking mean number of cases (per 100,000 persons in the state) as the **response variable** and the state wise information on presidential election 2020 results as the **(categorical) explanatory variable**. Linear regression is used to fit the explanatory variable (political leaning) on the response variable (COVID-19 cases) and the model summary is used to make a decision whether to reject or not to reject null hypothesis.
- Interpretations:
    - The linear regression coefficient estimate of the winning party is negative (-5.58) with a p-value = 0 (which is less than 0.05). Thus, it can be said with 95% level of statistical confidence that there is sufficient evidence to reject the null hypothesis.
    - We can conclude that political leaning has an effect on the number of COVID-19 cases across the states in the US.
    - Further, since the coefficient estimate is negative, it can be inferred that being affiliated to the Democratic party (party_won=1) is associated with a decrease in new COVID-19 cases per 100,000 persons (on an average) compared to states affiliated with the Republican party.
    - This result resonates with the intuitive idea that being a Democrat is seen to coincide with a liberal view on vaccination and following of COVID-19 related safety protocols.

(b) To test the second hypothesis, I used the hospital beds dataset used in Task 2 above which denotes whether or not a state has a shortage of hospital beds to handle the forecasted new COVID-19 cases (in the week ahead) that would need hospitalization.

- The null hypothesis is that shortage of hospital beds has no association with COVID-19 deaths in a state. The alternate hypothesis is that there is an association.
- A decision is made by taking mean number of deaths (per 100,000 persons in each state) as the **response variable** and the state wise information on shortage of hospital beds (yes=1, no=0) as the **(categorical) explanatory variable**. Linear regression is used to fit the explanatory variable (hospital bed shortage) on the response variable (COVID-19 deaths) and the model summary is used to make a decision whether to reject or not to reject null hypothesis.

- Interpretations:
    - The linear regression coefficient estimate of the winning party is positive (0.0901) with a p-value = 0.010 (which is less than 0.05). Thus, it can be said with 95% level of statistical confidence that there is sufficient evidence to reject the null hypothesis.
    - We can conclude that shortage of hospital beds has a significant effect on the number of COVID-19 deaths across the states in the US.
    - Further, since the coefficient estimate is positive, it can be inferred that having shortage of hospital beds (shortage=1) is associated with increase in the number of COVID-19 deaths per 100,000 persons (on an average), compared to states that do not have hospital beds shortage.
    - This result thus statistically substantiates the belief that shortage of hospital beds would lead to unavailability of suitable medical treatment for patients needing so, leading to increase in deaths due to the disease.