**IAC 621 Project Team Task**

**Stage IV (Basic Machine Learning)**

**Team Task: Develop Linear and Non-Linear (polynomial) regression models for predicting cases and deaths in the US. Start your data from the first day of infections in the US. X-Axis - number of days since the first case, Y-Axis - number of new cases and deaths. Calculate and report Root Mean Square Error (RMSE) for your models (linear and non-linear). Discuss bias versus variance tradeoff. Plot trend line for the data along with the forecast of 1 week ahead. Describe the trends as compared to other countries.**

For the team task, we have used a combination of codes developed by the team members Kyle and Reetika individually for our member tasks. Kyle prepared the data for analysis and drew interpretations, while Reetika designed the team task code for plotting the trends and forecast.

For COVID-19 cases and deaths in USA, the RMSE for the three models are:

|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 137828.36 | 957.01 |
| Polynomial (degree = 2) | 131662.24 | 955.52 |
| Polynomial (degree =3) | 112982.22 | 934.77 |

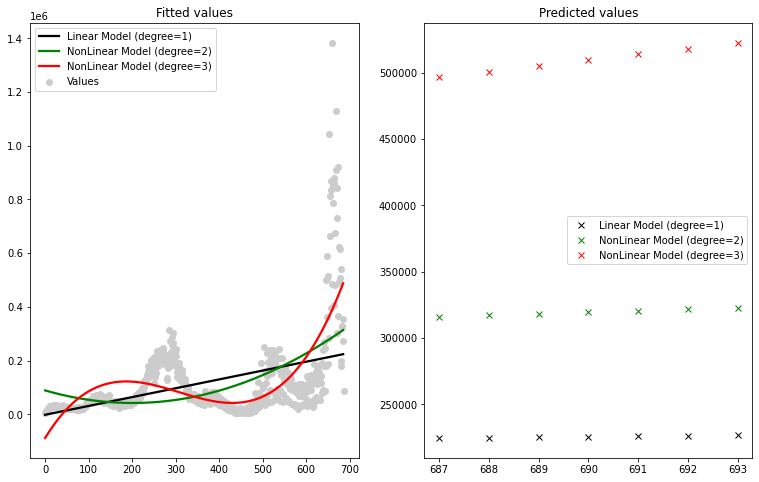
**Bias vs Variance Tradeoff**

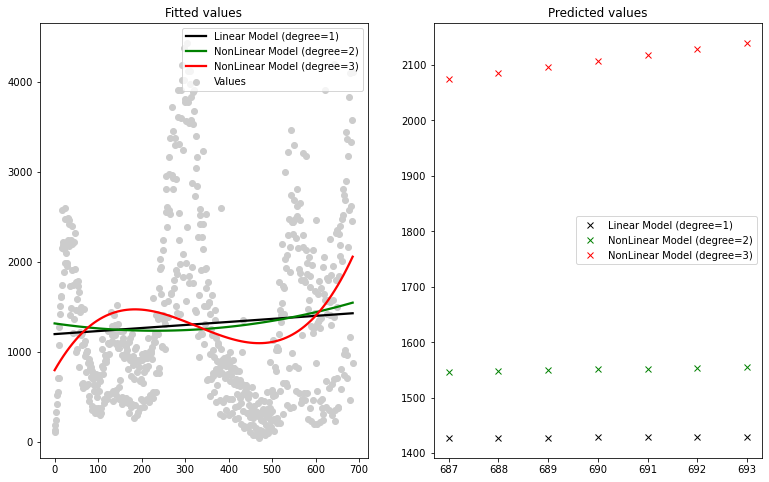
In regression, RMSE is a function of both bias and variance, and both are forms of error. Bias is the extent to which the model makes incorrect assumptions regarding the data, and can cause the resulting model to underfit the data. Variance is how much the model reacts to small changes in the data, and can cause the resulting model to overfit the data. Bias and Variance are inversely related; as one decreases, the other increases. Linear models are considered high Bias and low Variance, while non-linear models decrease in Bias and increase in Variance as their complexity rises. The decision as to which model to use based on Bias and Variance is the Bias-Variance tradeoff.

RMSE is calculated using both errors, and decreases with increased model complexity. For this task, our decision as to which model is best will be based on how much RMSE decreases as the model increases in complexity. For example, we can see with the US COVID cases that the linear model and the nonlinear model with two degrees have similar RMSE, indicating that they're equally as useful models for the data. In contrast, the nonlinear model with three degrees has a much lower RMSE. This would indicate that perhaps the nonlinear model with three degrees is the best model for the data, despite its increased variance.

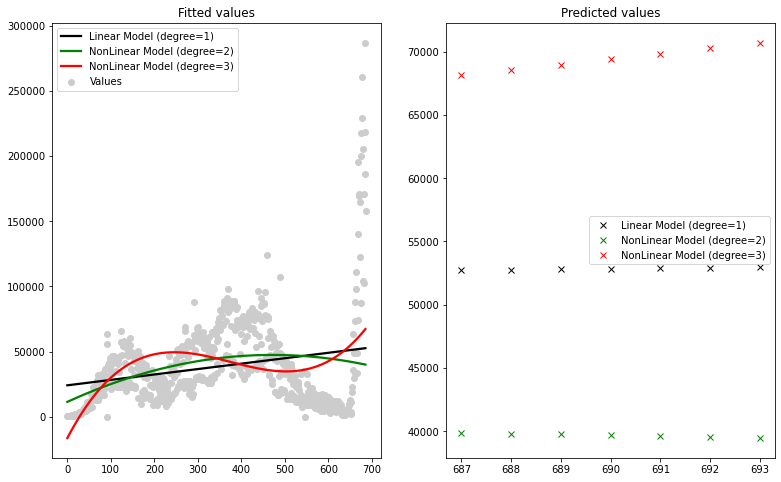
Following are the trend and forecast plots of the USA and other countries:

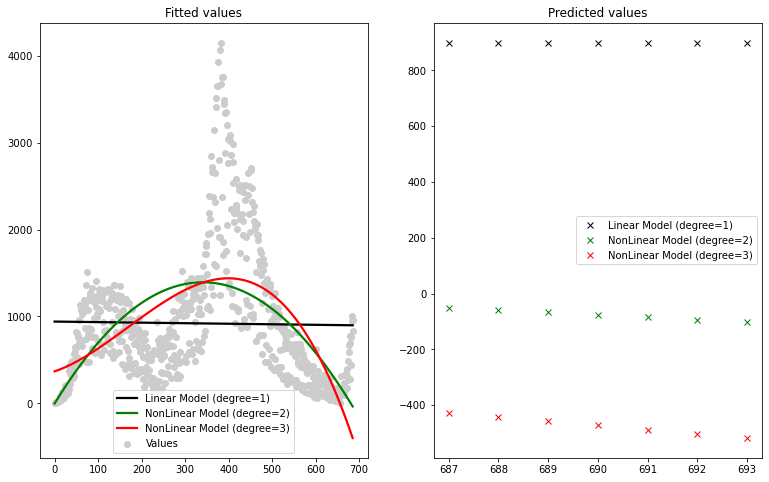
**Modeling cases and deaths for USA**

****



**Modeling cases and deaths for Brazil**

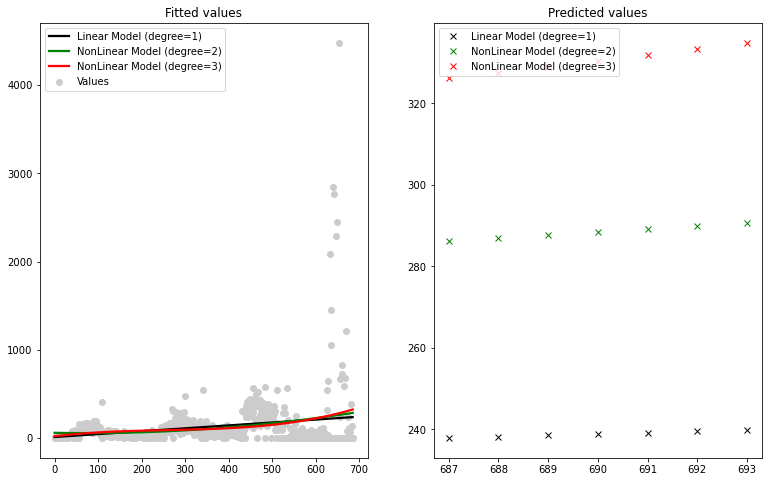


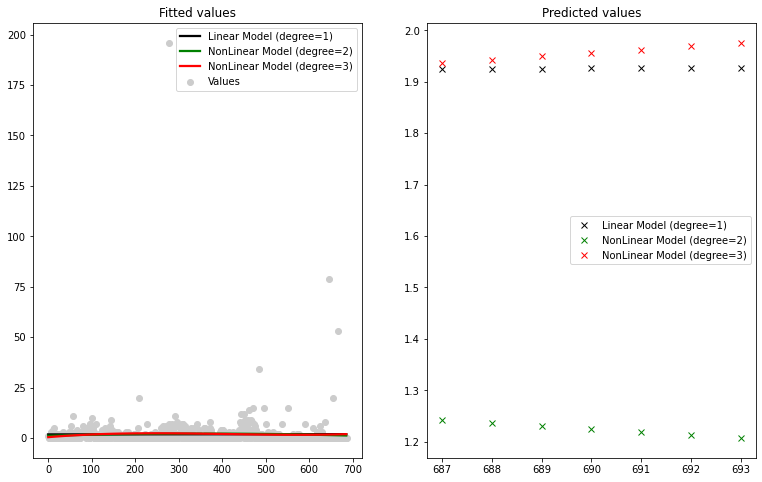


|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 33925.86 | 772.49 |
| Polynomial (degree = 2) | 33441.96 | 646.39 |
| Polynomial (degree =3) | 31723.17 | 630.85 |

Regarding COVID cases, Brazil appears to have a flatter curve than the US, and its nonlinear model with 2 degrees is actually showing a downward curve to the data, as opposed to the US' upward curve for all models. Brazil differs even more with COVID deaths. All 3 models predict a downward trend in COVID deaths, while the US predicts an upward trend with all models.

**Modeling cases and deaths for Demographic Republic of Congo**

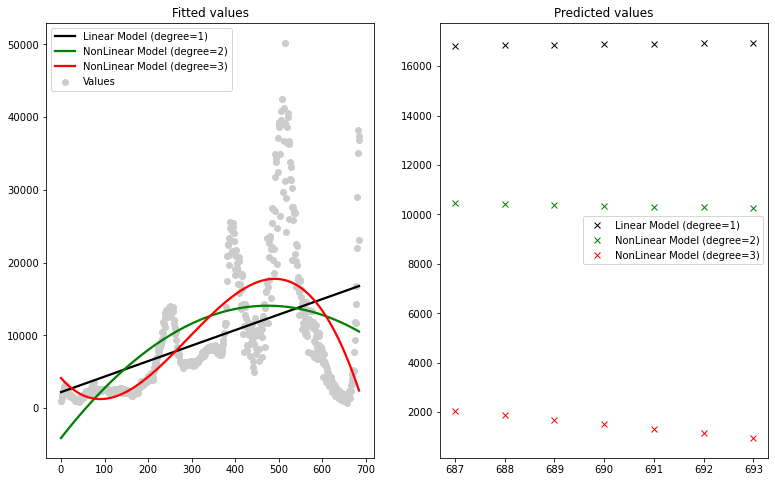


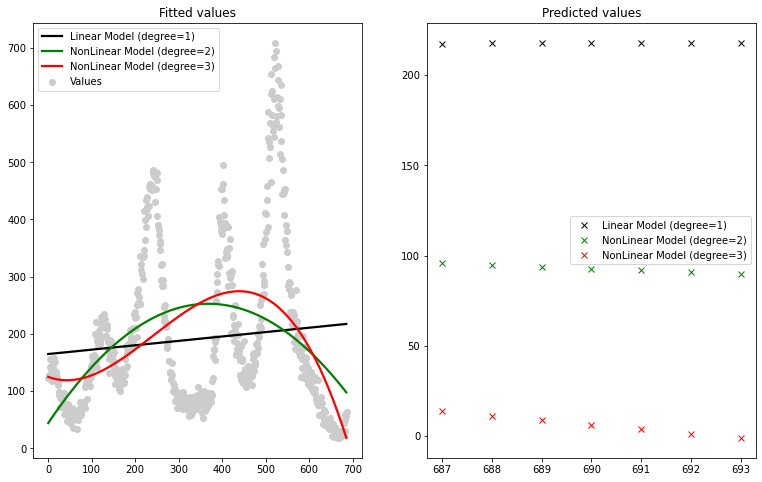


|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 293.36 | 8.70 |
| Polynomial (degree = 2) | 292.57 | 8.69 |
| Polynomial (degree =3) | 292.18 | 8.69 |

It appears that Congo has been relatively consistent with COVID cases throughout the pandemic, as opposed to the peaks and valleys we see in the US. For deaths, this flat trend continues, and is even more pronounced than with the cases. All 3 models seem to be equivalent, to the point that both nonlinear models are nearly linear themselves.

**Modeling cases and deaths for Iran**

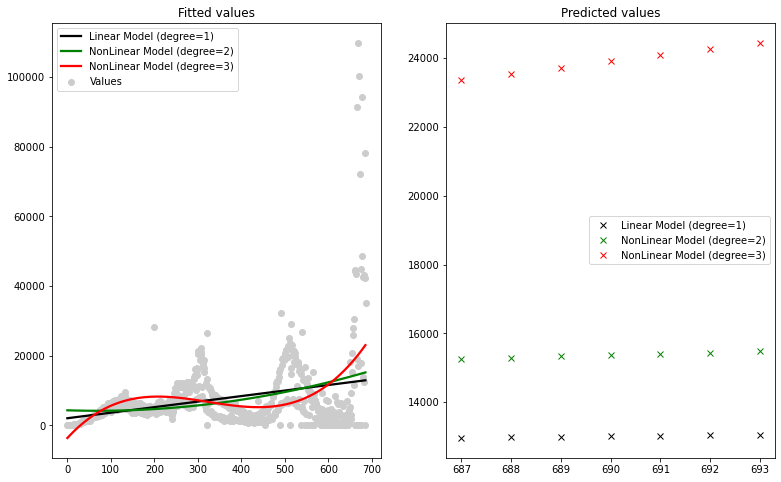


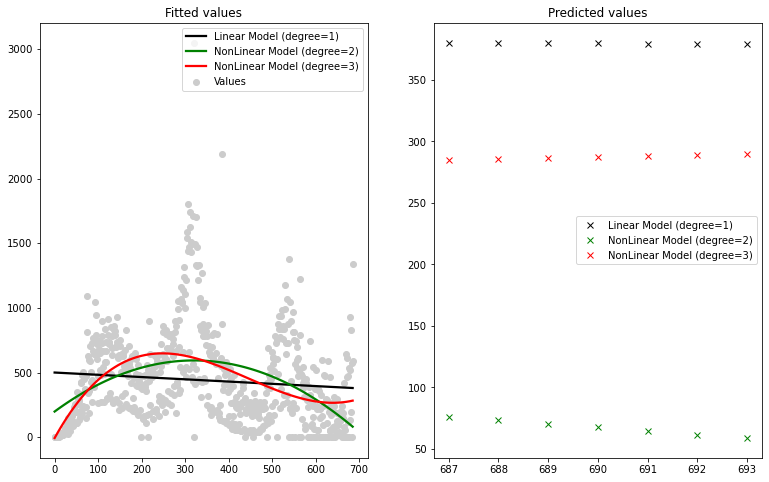


|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 8101.86 | 144.02 |
| Polynomial (degree = 2) | 7589.73 | 133.44 |
| Polynomial (degree =3) | 6906.81 | 129.84 |

For the nonlinear models of COVID cases, Iran peaks later than the US, and is now currently in a downward trend. Their linear models are similar, however. The deaths for Iran have quite similar trends to the new cases, and again differ in their nonlinear models by having a later peak and a downward trend at the end.

**Modeling cases and deaths for Mexico**

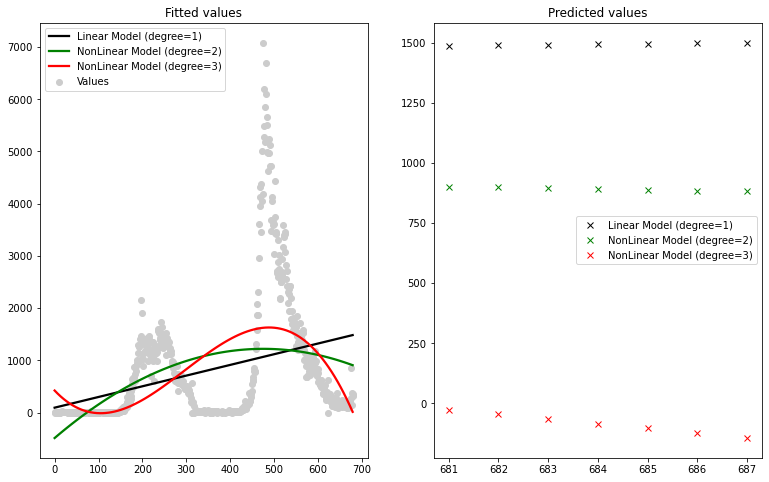


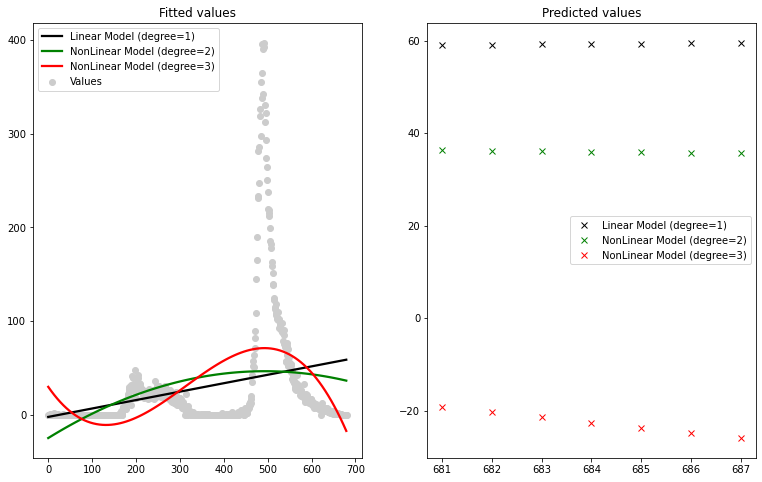


|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 10074.92 | 359.83 |
| Polynomial (degree = 2) | 10022.69 | 333.38 |
| Polynomial (degree =3) | 9550.90 | 324.10 |

Mexico's COVID cases have similar trends in terms of when they peak and valley to the US, but are flatter. By contrast, deaths differ from the US quite a bit. Both the linear and nonlinear with 2 degrees predict a downward trend, versus the US' upward trend for all 3 models.

**Modeling cases and deaths for Myanmar**

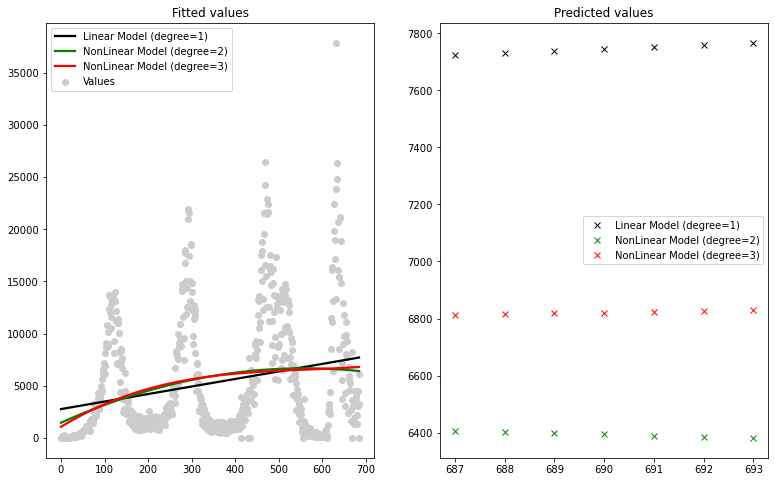
****

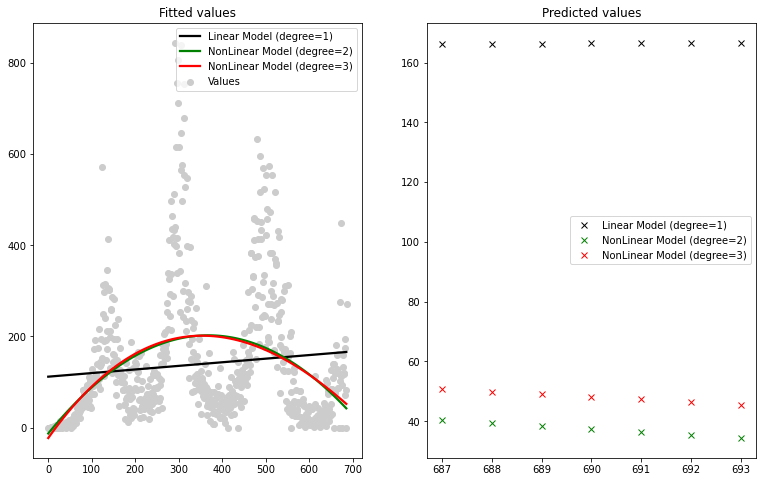


|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 1144.43 | 62.19 |
| Polynomial (degree = 2) | 1114.22 | 61.36 |
| Polynomial (degree =3) | 1058.82 | 57.72 |

Myanmar predicts a downward trend in cases in both nonlinear models, versus the US's upward trends. Myanmar has already hit its peak with those 2 models, while the US is currently experiencing its peak. The deaths are nearly identical in overall trend to the cases, and differ from the US in the same ways.

**Modeling cases and deaths for South Africa**

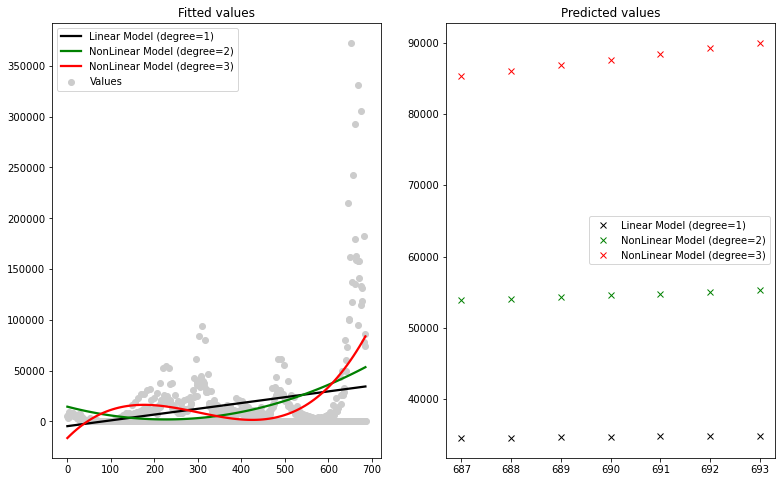
****

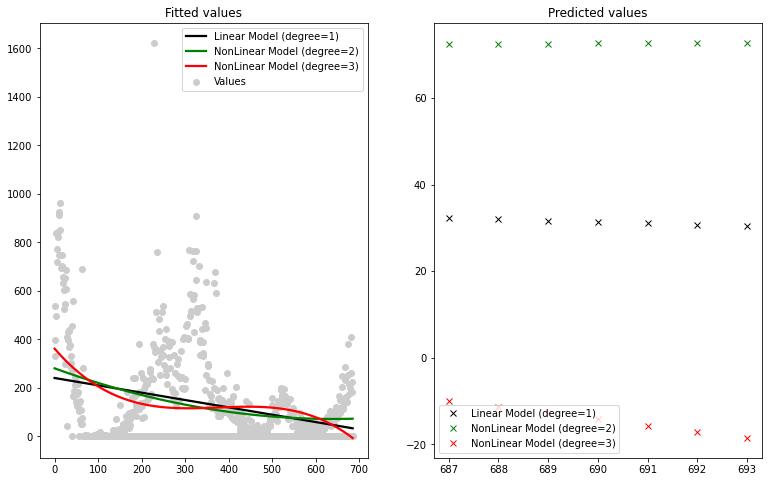


|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 5562.83 | 146.43 |
| Polynomial (degree = 2) | 5531.84 | 135.32 |
| Polynomial (degree =3) | 5529.75 | 135.27 |

Similar to Congo, COVID cases appear relatively consistent based on the predicted trends of the 3 models, versus the US having more distinct peaks and valleys. By contrast, the 2 nonlinear models for deaths have a more pronounced curve. Unlike the US, they appear to almost follow a normal bell curve, now trending downward.

**Modeling cases and deaths for Spain**





|  |  |  |
| --- | --- | --- |
| Model | RMSE | |
| Cases | Deaths |
| Linear (degree = 1) | 34714.33 | 192.33 |
| Polynomial (degree = 2) | 33630.22 | 191.50 |
| Polynomial (degree =3) | 31506.00 | 188.99 |

COVID cases for Spain have a nearly identical trend to the US for the 2 nonlinear models, though the trend is flatter in nature. Deaths show a different picture. Of all the countries we looked at, Spain has the most consistent downward curve as the days since first infection increase. It and Myanmar are the only two countries with an overall downward trend in the linear model.

For more information about the individual member tasks, refer to the member task reports in the GitHub repository for Team 4.