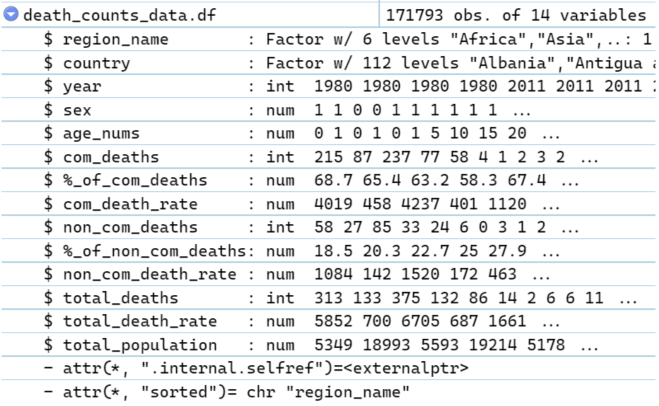
Predictive Analytics MGS 616 Group Project

Team FP_8: Sai Kartheek Mahankali, Nihareeka Suneel Gote, Viraj Vhatkar, Aditya Manjunath Naik

Dataset

The dataset we are going to use in this part of the project is “death_counts_data.df”. It has 171k rows and 14 columns, and below is how it looks –



```
death_counts_data.df                 171793 obs. of 14 variables
    $ region_name         : Factor w/ 6 levels "Africa","Asia",..: 1
    $ country             : Factor w/ 112 levels "Albania","Antigua
    $ year                : int  1980 1980 1980 1980 2011 2011 2011
    $ sex                 : num  1 1 0 0 1 1 1 1 1 1  ...
    $ age_nums            : num  0 1 0 1 0 1 5 10 15 20  ...
    $ com_deaths          : int  215 87 237 77 58 4 1 2 3 2  ...
    $ %_of_com_deaths     : num  68.7 65.4 63.2 58.3 67.4  ...
    $ com_death_rate      : num  4019 458 4237 401 1120  ...
    $ non_com_deaths      : int  58 27 85 33 24 6 0 3 1 2  ...
    $ %_of_non_com_deaths: num  18.5 20.3 22.7 25 27.9  ...
    $ non_com_death_rate : num  1084 142 1520 172 463  ...
    $ total_deaths        : int  313 133 375 132 86 14 2 6 6 11  ...
    $ total_death_rate    : num  5852 700 6705 687 1661  ...
    $ total_population    : num  5349 18993 5593 19214 5178  ...
    - attr(*, ".internal.selfref")=<externalptr>
    - attr(*, "sorted")= chr "region_name"
```

Choosing a Model

We are going to try various Regression models such as Linear Regression, Ridge Regression, Lasso Regression, Principal Component Regression and various Time Series models such as Naïve, Drift, Mean, ETS, and ARIMA on our data. We use 80% of our data to train each of the above-mentioned models, and then validate it on the remaining 20% of the data. We then compute and compare the accuracy metrics for each model and select the best choice.

Regression Analysis

From the original dataset “death_counts_data.df”, we have filtered out data belonging to USA and Brazil. We then applied different regression models onto the data separately and computed the accuracy metrics.

Below are the accuracy metrics of different regression methods on USA data –

Model	RMSE	MAE	MAPE	R Squared
Linear Regression - USA	82.62	72.63	3.38	0.96
Ridge Regression - USA	83.2	73.4	3.42	0.96
Lasso Regression - USA	89.7	81.2	3.77	0.95
PCR - USA	95.02	86.74	4.02	0.95

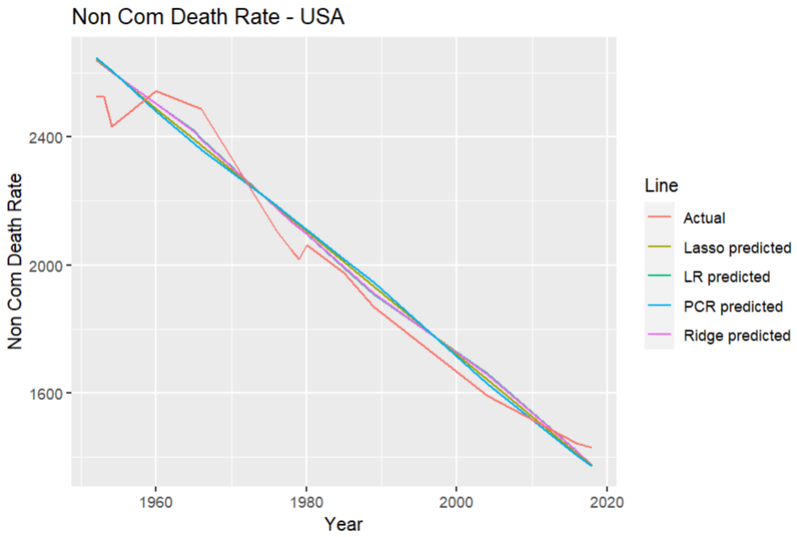
Below are the accuracy metrics of different regression methods on Brazil data –

Model	RMSE	MAE	MAPE	R Squared
Linear Regression - Brazil	40.37	29.58	1.78	0.54
Ridge Regression - Brazil	40.37	29.58	1.78	0.54
Lasso Regression - Brazil	39.72	28.7	1.73	0.54
PCR - Brazil	61.5	56.19	3.3	0.065

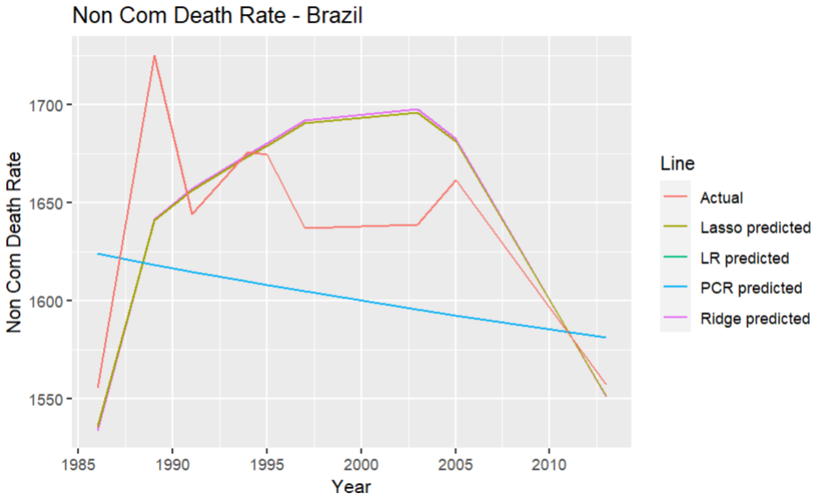
Observing both the tables, we see that the Lasso Regression gives the best overall metrics. Hence, for this data, Lasso Regression is our best choice.

Visualizations

Below are the visualizations of actual vs predicted values for different regression methods on USA and Brazil datasets –



Non Com Death Rate - USA



Non Com Death Rate - Brazil

Time Series Analysis

For Time-Series Analysis, instead of considering entire data from 1950s to 2020, we consider only the data from 1990-2020, as the data from previous decades doesn't necessarily influence/impact the recent data. We analyzed the time series data from 1990-2015 and try to forecast the data from 2016-2019 using the NAÏVE, DRIFT, and MEAN models. Below are the test accuracy metrics of these models on USA data –

Model	RMSE	MAE	MAPE	RMSSE	ACF1
Drift	8.80	6.32	0.440	0.280	-0.134
Mean	227	226	15.8	7.22	0.185
Naïve	45.2	42.9	3.01	1.44	0.185
ETS	45.2	42.9	3.01	1.44	0.185
Arima	8.80	6.32	0.43	NA	NA

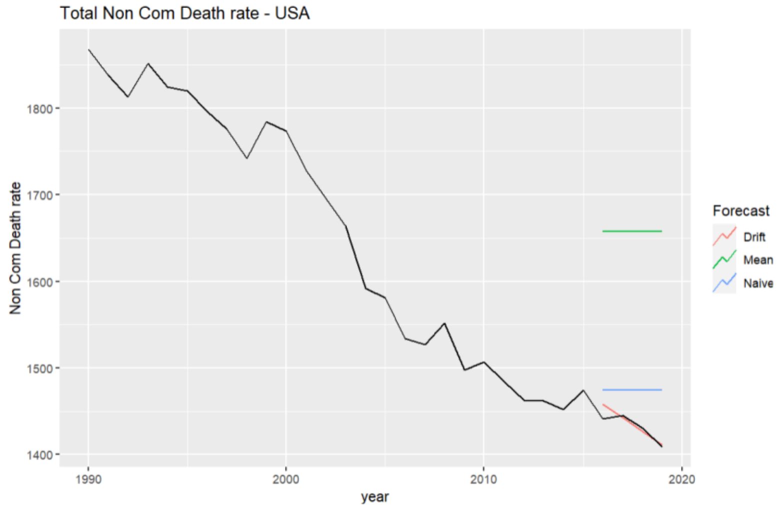
Below are the test accuracy metrics of these models on Brazil data –

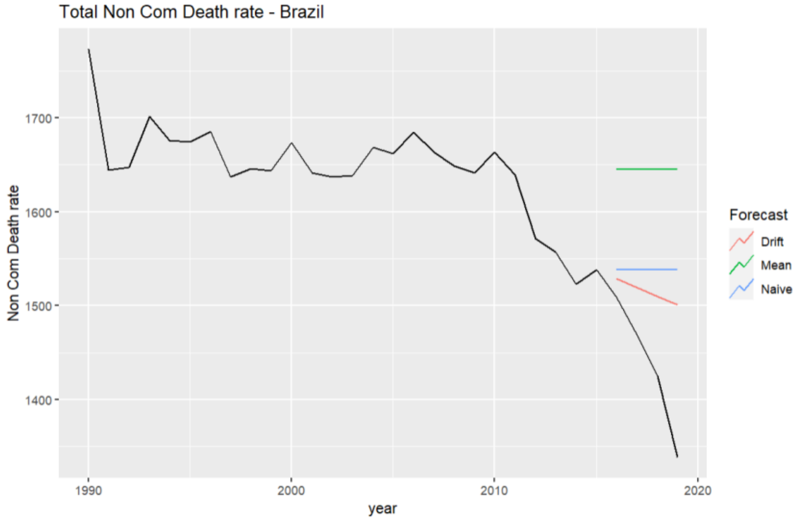
Model	RMSE	MAE	MAPE	RMSSE	ACF1
Drift	95.7	79.6	5.72	2.59	0.18
Mean	220	210	14.9	5.95	0.195
Naïve	121	103	7.4	3.28	0.195
ETS	120	102	7.3	3.25	0.195
Arima	121	103.05	7.39	NA	NA

Observing both the tables, we see that the Drift model gives the best accuracy metrics. Hence, for this data, Drift Model is our best choice.

Visualizations

Below are the visualizations of actual vs predicted values for different time-series methods on USA and Brazil datasets –





Conclusion

Based on above results, we can successfully predict the death rate per 100,000 population for a country, given the year, gender, and total population using either Regression Analysis or Time-Series Analysis. From the test accuracy metrics, we identify that the Drift method is the best choice for this problem at hand. Using the Drift Method, the predicted Non-Com Death rates per 100k population in USA for coming years (2023-2027) are - 1348,1333,1317,1301,1285. Using the same method, the predicted Non-Com Death rates per 100k population in USA for coming years (2023-2027) are – 1463,1454,1444,1435,1425.