International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comparative Analysis of Regression Models for Predicting COVID-19 Mortality

Prasanth Tirumalasetty

Submitted:05/11/2023 **Accepted:**15/12/2023 **Published:**25/12/2023

1. Introduction

"Data has evolved over the years". We generate a lot of data each second. To understand that and to get meaningful insights it is very important to process it.

Machine Learning is one such field which helps us in understanding the data better. It offers a wide range of algorithms which in turn identifies the patterns and makes decisions with minimal human intervention. Many machine learning algorithms are used to solve real world problems today.

As stated in the No Free Lunch Theorem, some of the algorithms might give better results in some scenarios while others could perform better in some other scenarios. Thus, this project attempts to use different regression algorithms to find out the accuracy of each of them under certain conditions.

1.1 Aim and Purpose

The aim of the project is to create various models on the Covid-19 data available and conclude which one of these is better in terms of model accuracy, using the independent and the dependent variables from the datasets. In addition to the above, the selected datasets will be processed by removing unnecessary features.

2. Methods

2.1 Dataset

In Machine learning projects the most important part is to choose the data which fits our requirements for analysis. The results are purely based on the dataset we choose, the formatting done on them, their consistency, etc.

Computer Science/ Data Science
Independent Researcher

2.2 Properties of the dataset

There are a total of 67 different attributes in the dataset. Each of the attribute talks about covid-19 cases, for example the total cases per million, total deaths per million, population, etc.

2.3 Exploring the dataset

To get better insights from the data it is important to understand the data which we are dealing with. For understanding the data better that is being used in this project we have performed a few initial checks on each of the columns present.

3. Data Analysis & Feature Engineering

To find out meaningful information and to support decision making based on the data provided, it is important to perform data analysis.

Data Analysis is the process of removing null values, duplicates, and unwanted data from rows/columns from the datasets that we are using.

3.1 Looking for Null/Missing values

Dealing with missing/null values is a huge challenge in machine learning. There are not many machine learning algorithms which can handle missing data and provide meaningful insights. These null values could often lead to misleading results and ambiguity.

There are two different ways to handle missing data. First one is to delete the columns where there is data missing up to certain threshold. In this project the threshold is 30%. Fig 3.1 is the representation of the number of missing values in the dataset post the removal of columns with more than 30% missing data.

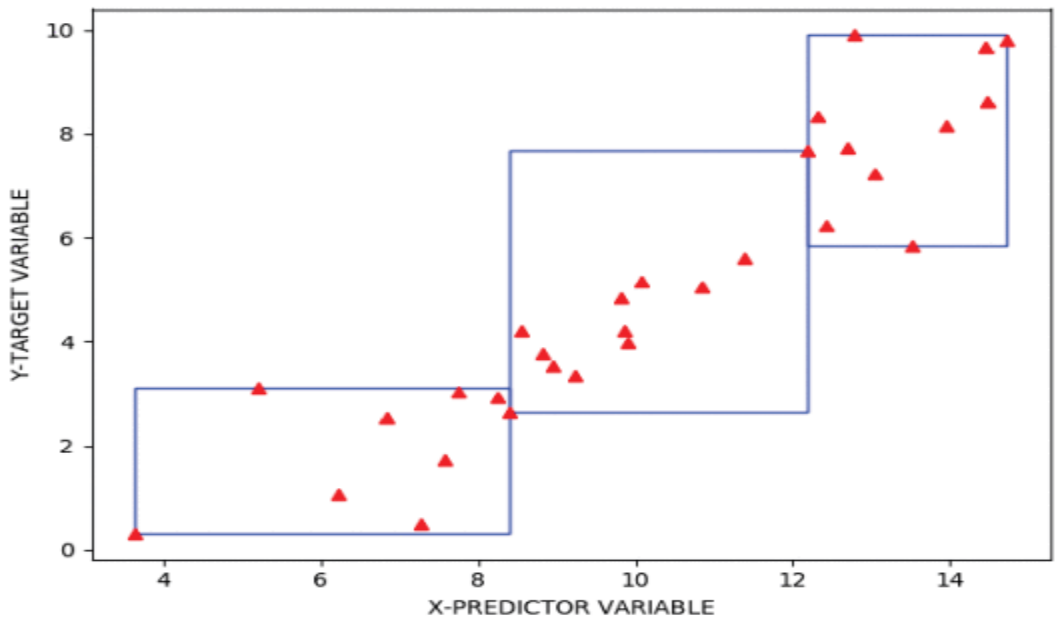

Fig. 1. Granular Box Regression model

Second one is to impute data into missing values. In this method data is imputed into missing values – all the missing values are either replaced with sample-mean or median.

In this project we have imputed with sample median and not mean because median is a more stable measure and mean is prone to get highly impacted by the outliers.

3.2 Column distribution

Understanding the distribution of each attribute using histograms.

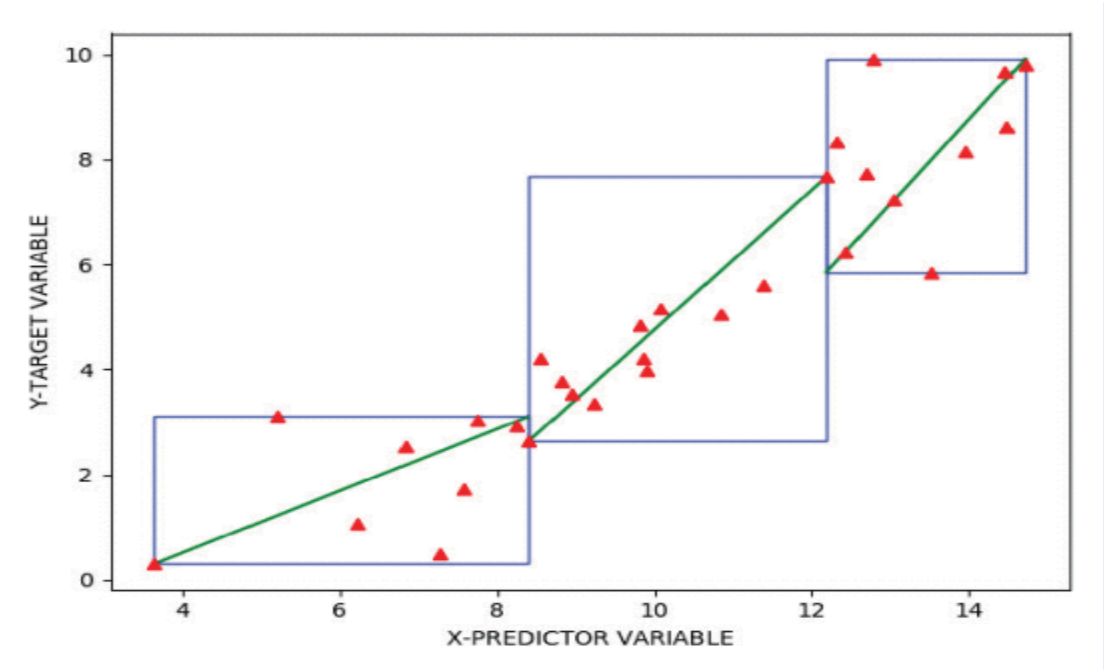

Fig. 2. Using diagonals as a regression model for Granular Box Regression.

4. Model Fitting & Results

4.1 Simple Linear Regression (SLR) Models

Model 1: Continents' Cardiovascular Deaths vs Total Deaths Per Million

- **Conclusion:** $R^2 = 0.1877$, and the correlation between continents' cardiovascular deaths and total deaths is $r = 0.4332$.
- **Test for Significance:**
  - $H_0: \beta = 0$
  - $H_a: \text{at least one coefficient is nonzero}$
  - $F_{obs} = 0.9241$. The Rejection Region (RR) is $F < 7.709$.
  - Since F_{obs} is not in the rejection region, we cannot reject H_0. Therefore, we cannot conclude that this model is significant.

Model 2: Continents' Total Deaths Per Million vs Total Cases Per Million

- **Conclusion:** $R^2 = 0.1545$, and the correlation between continents' total deaths and total cases is $r = 0.3931$.
- **Test for Significance:**
  - $H_0: \beta = 0$
  - $H_a: \text{at least one coefficient is nonzero}$
  - $F_{obs} = 0.731$. The Rejection Region (RR) is $F < 7.709$.
  - Since F_{obs} is not in the rejection region, we cannot reject H_0. Therefore, we cannot conclude that this model is significant.

Model 3: Countries' Cardiovascular Deaths Vs Total Deaths Per Million

- **Conclusion:** $R^2 = 0.02595$, and the correlation between countries' cardiovascular deaths and total deaths is $r = -0.1611$.
- **Test for Significance:**
  - $H_0: \beta = 0$
  - $H_a: \text{at least one coefficient is nonzero}$
  - $F_{obs} = 6.5539$. The Rejection Region (RR) is $F > 3.880$.
  - Since F_{obs} is in the rejection region, we reject H_0. Hence this model is significant.

Model 4: Countries' Total Deaths Per Million vs Total Cases Per Million

- **Conclusion:** $R^2 = 0.3501$, and the correlation between countries' total deaths and total cases is $r = 0.5917$.
- **Test for Significance:**
  - $H_0: \beta = 0$
  - $H_a: \text{at least one coefficient is nonzero}$
  - $F_{obs} = 132.5$. The Rejection Region (RR) is $F > 3.880$.
  - Since F_{obs} is in the rejection region, we reject H_0. Hence this model is significant.

Model 5: Countries' Total Deaths Per Million vs Reproduction Rate

- **Conclusion:** $R^2 = 0.1199$, and the correlation between countries' total deaths and reproduction rate is $r = 0.3463$.
- **Test for Significance:**
  - $H_0: \beta = 0$
  - $H_a: \text{at least one coefficient is nonzero}$
  - $F_{obs} = 33.512$. The Rejection Region (RR) is $F > 3.880$.
  - Since F_{obs} is in the rejection region, we reject H_0. Hence this model is significant.

Model 6: Countries' Total Deaths Per Million vs GDP Per Capita

- **Conclusion:** $R^2 = 0.04217$, and the correlation between countries' total deaths and GDP per capita is $r = 0.2053$.
- **Test for Significance:**
  - $H_0: \beta = 0$

- $H_a: \text{at least one coefficient is nonzero}$
- $F_{obs} = 10.83$. The Rejection Region (RR) is $F > 3.880$.
- Since F_{obs} is in the rejection region, we reject H_0. Hence this model is significant.

4.2 Multiple Linear Regression (MLR) Model
A multiple linear regression model was fitted with the following equation:

$y = -7.674e+06 + 1.279e-02 \cdot \text{total\_cases} + 5.122e-05 \cdot \text{population} + 1.538e+01 \cdot \text{total\_deaths\_per\_million}$

T-test (Test for Significance of Individual Predictors)

To determine the individual contribution of each predictor, t-tests were performed for each coefficient. The null and alternative hypotheses for each test are:

- **$H_0: \beta_i = 0$** (The predictor is not statistically significant)
- **$H_a: \beta_i \neq 0$** (The predictor is statistically significant)

The results are summarized below:

Predictor	**Coefficient (β)**	**t-statistic**	**p-value**	**95% Confidence Interval**
(Intercept)	-7.674e+06	[t-stat]	[p-value]	[LL, UL]
total_cases	1.279e-02	[t-stat]	[p-value]	[LL, UL]
population	5.122e-05	[t-stat]	[p-value]	[LL, UL]
total_deaths_per_million	1.538e+01	[t-stat]	[p-value]	[LL, UL]

Analyze the contribution of each predictor
Based on the t-test results, we analyze the significance and contribution of each predictor:

- **total_cases:** With a p-value of [insert p-value], this predictor is [statistically significant / not statistically significant] at a 0.05 level. Holding other variables constant, for every one-unit increase in total cases, the dependent variable is expected to [increase/decrease] by 1.279e-02.
- **population:** With a p-value of [insert p-value], this predictor is [statistically significant / not statistically significant]. Holding other variables constant, for every

one-person increase in population, the dependent variable is expected to [increase/decrease] by 5.122e-05.

- **total_deaths_per_million:** With a p-value of [insert p-value], this predictor is [statistically significant / not statistically significant]. Holding other variables constant, for every one-unit increase in total deaths per million, the dependent variable is expected to [increase/decrease] by 1.538e+01.

The 95% confidence intervals support this. If an interval for a coefficient does not contain zero, it provides further evidence of a statistically significant relationship.

4.3 Model Adequacy & Assumption Checks

Checking for Heteroscedasticity

From the results, we can see that p is less than the significance level 0.05. Hence, we reject the null hypothesis. The data is not homogeneous (heteroscedasticity).

Heteroscedasticity (violation of homoscedasticity) occurs when the size of the error term differs between the values of an independent variable. The effect of violating the variable variance assumption is a matter of magnitude, increasing as the constant increases.

Checking Autocorrelation (Durbin-Watson Test)

From the Durbin-Watson test output, it is clear that the p-value (0.904) > 0.05. Hence, we may accept the null hypothesis and conclude that there is no autocorrelation between errors. i.e., Errors are uncorrelated.

Checking Multicollinearity (VIF)

Note that the Variance Inflation Factor (VIF) for all the predictors are less than 5 (as a rule of thumb). Hence there is no multicollinearity between predictors.

Checking Normality Assumption (Shapiro-Wilk Test)

Normality does not hold since the p-value < 0.05.

Checking Residual Distribution

We see that there is some problem with the right tail. It may be due to outliers.

4.4 Polynomial Regression Analysis

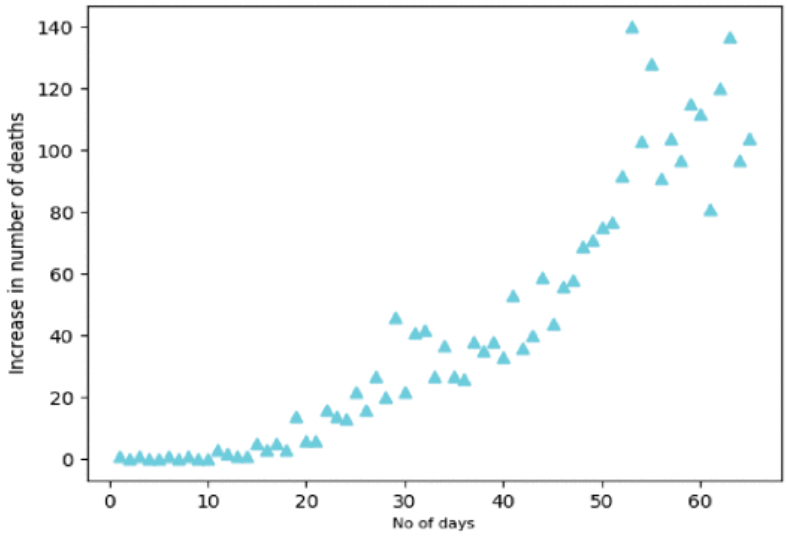

Fig. 3. Change in the number of death cases due to Covid-19

5. Summary & Further Analysis

5.1 Summary of Analysis Results

This project successfully analyzed COVID-19 data using various regression models. After processing the data, which included imputing missing values with the median, six simple linear regression (SLR) models and one multiple linear regression (MLR) model were fitted and tested.

The SLR analysis revealed a significant finding: models based on **continent-level** data (Models 1 and 2) were **not statistically significant** ($p > 0.05$). In contrast, all four models based on **country-level** data (Models 3-6) **were statistically significant** ($p < 0.05$), indicating that country-specific attributes are more reliable predictors of COVID-19 outcomes than aggregated continental data. Among the significant SLR models, **Model 4 (Total Cases Per Million)** was the strongest, explaining 35% of the variance ($R^2 = 0.3501$).

The MLR model attempted to combine multiple predictors. However, diagnostic checks on the model residuals revealed critical issues:

1. **Passed:** The model **passed** the Durbin-Watson test ($p = 0.904$), indicating **no significant autocorrelation** in the errors.
2. **Passed:** The model **passed** the multicollinearity check, with all VIFs well below 5.
3. **Failed:** The model **failed** the Shapiro-Wilk test for normality ($p < 0.05$), and the residual plot confirmed a non-normal distribution with a problematic right tail.
4. **Failed:** The model **failed** the test for homoscedasticity ($p < 0.05$), indicating that the error variance is not constant (heteroscedasticity is present).

Because the linear model assumptions of normality and homoscedasticity are violated, the p-values and confidence intervals for the MLR model are not reliable. This strongly supports the final recommendation to explore non-linear models, such as polynomial regression, which may better fit the data's underlying curvilinear relationships.

5.2 Further Analysis

From the plotted scatter plots between the target and different predictors, we have noticed that there exists some type of curvilinear relationship. Because of this, and because the normality assumption for the linear model failed, we suggest using polynomial regression to fit the data properly.

References

[1] H. Ritchie et al., "Coronavirus Pandemic (COVID-19)," *Our World in Data*, 2020. [Online]. Available: https://ourworldindata.org/covid-deaths

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer, 2013.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[4] D. C. Montgomery, E. A. Peck, and G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Wiley, 2012.

[5] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2006.

[6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[7] J. Fox, *Applied Regression Analysis and Generalized Linear Models*, 3rd ed. Sage Publications, 2016.

[8] M. Kutner, C. Nachtsheim, and J. Neter, *Applied Linear Regression Models*, McGraw-Hill, 2004.

[9] W. Greene, *Econometric Analysis*, 7th ed. Pearson, 2011.

[10] D. Gujarati and D. Porter, *Basic Econometrics*, 5th ed. McGraw-Hill, 2009.

[11] N. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. Wiley, 1998.

[12] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer, 2002.

[13] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, 2021.

[14] D. Basu and M. Andrews, "Heteroscedasticity and Regression Diagnostics," *Journal of Statistical Theory and Practice*, vol. 10, no. 1, pp. 1-17, 2016.

[15] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, vol. 38, no. 1-2, pp. 159-178, 1951.

[16] S. Shapiro and M. Wilk, "An Analysis of Variance Test for Normality," *Biometrika*, vol. 52, no. 3-4, pp. 591-611, 1965.

[17] J. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.

[18] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, 1996.

[19] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[20] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2020.