Title: Comparison of Linear & Nonlinear Models Using COVID-19 Time Series Dataset

---

Objective :-

The objective of this task is to:

- Compare **Generalized Linear Models (GLM)** and **Nonlinear Regression** techniques using real-world data.
- Analyze which model better captures the underlying patterns in the data.
- Justify the need for nonlinear models by understanding the limitations of linear assumptions.

Dataset Used :- Source: Johns Hopkins University – Center for Systems Science and Engineering (JHU CSSE)

- **Dataset:** Time Series of Confirmed COVID-19 Cases (Jan 2020 onward)
- **Link:** Click here
- **Selected Country:** India
- **Features:**
    - Date: Timeline of reported data
    - Cases: Cumulative COVID-19 cases
    - Days: Number of days since the first reported case in India

---

Background: Linear vs. Nonlinear Models :-

◆ **Linear Models (GLM)**

- Assumes a **straight-line** relationship between independent and dependent variables.
- Effective when the data shows a **constant rate of change**.
- GLM (Gaussian family) is a generalization of linear regression that can be extended to other distributions.

◆ **Nonlinear Models**

- Models that allow for **curvature** or **complex relationships**.
- Useful for **growth curves**, **exponential patterns**, and **saturation behavior**.
- In this task, a **logistic function** was used to represent the **S-curve** nature of pandemic growth.

Methodology :-

◆ **1. Data Preprocessing**

- Filtered India-specific data.
- Converted cumulative case counts into numeric arrays.
- Calculated Days as a feature for modeling.

◆ **2. Applied GLM**

- Model:

$Y=\beta 0+\beta 1 \cdot \text{Days}$

- • Fit using `statsmodels.GLM()` with Gaussian distribution.
- • Prediction: `GLM_Pred`

◆ **3. Applied Nonlinear Regression (Logistic)**

- Fit using scipy.optimize.curve_fit()
- Prediction: Nonlinear_Pred

Evaluation Metrics

Metric	**GLM (Linear)**	**Nonlinear (Logistic)**
RMSE	268,920.32	✅ **48,762.58**
R² Score	0.8987	✅ **0.9963**
Residual Pattern	Systematic	✅ Random

Prediction Trend	Underestimates saturation	✅ Accurately models curve

Interpretation:

- **GLM RMSE** is much higher, indicating **greater error**.
- **R^2 Score** is much closer to **1.0** for the logistic model, suggesting a better fit.
- Residuals for GLM **increase over time** (underfitting).
- Residuals for nonlinear model are **scattered and close to zero**, indicating accuracy.

Visual Analysis

1. Actual vs. Predicted Cases

- **GLM** follows the early phase of growth but fails to account for saturation.
- **Logistic Model** captures the typical **epidemic S-curve** — initial slow growth, exponential phase, and final plateau.

2. Residual Plots

- GLM residuals are **not randomly distributed**, indicating poor fit.
- Nonlinear model residuals are **evenly spread**, suggesting a well-fit model.

Limitations of Linear Models

- Cannot capture **non-linear behaviors**, such as exponential growth or logistic saturation.
- Often **underfit** when dealing with real-world phenomena like:
  - Epidemics
  - Population growth
  - Financial trends
- Predicts **unbounded values**, which is **not realistic** for many practical applications.

When to Use Nonlinear Models

- When data exhibits **non-constant change**, especially:
  - **Exponential or logistic growth**
  - **Periodic behavior** (e.g., sine waves, weather cycles)
  - **Threshold effects** (e.g., drug dosage vs. effect)
- Logistic models are widely used in:
  - **Epidemiology**
  - **Ecological modeling**
  - **Marketing and adoption curves**

Conclusion

Through this task, we conclude that:

- **Nonlinear models** (especially logistic models) are **superior** for modeling real-world phenomena like the COVID-19 pandemic.
- While **linear models** are simple and easy to interpret, they are **not suitable** for datasets that show **complex or bounded growth**.
- **Model evaluation using residuals and prediction accuracy** strongly supports the use of nonlinear regression for epidemic modeling.
- A sound understanding of data characteristics is essential before selecting the modeling technique.