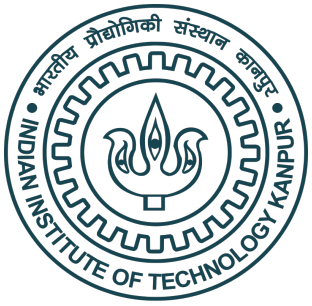

Indian Institute of Technology, Kanpur

Department of Mathematics and Statistics

Non-Linear Regression [MTH-686]

Course Project

Student:
Tushar Anand
Roll No: 221145
tushar22@iitk.ac.in

Supervisor:
Prof. Debasis Kundu
kundu@iitk.ac.in

DATA SET - 37
November 09, 2024

Project Report on Non-linear Regression Analysis

1 Introduction

In this project, we analyze a dataset containing paired values $(t, y(t))$ using three distinct nonlinear regression models. The main goal is to estimate the unknown parameters for each model using the least squares method. By finding the optimal parameter values, we aim to determine the best-fitting curve for each model and assess their accuracy in representing the data.

After parameter estimation, we will perform a comprehensive analysis that includes calculating confidence intervals, examining residuals, conducting normality tests, and estimating variance. This will allow us to compare the effectiveness of each model and evaluate the quality of fit, ultimately identifying the model that best captures the underlying patterns in the dataset.

2 Models

The three models we will fit to the data are:

- **Model 1:**
$$y(t) = \alpha_0 + \alpha_1 e^{\beta_1 t} + \alpha_2 e^{\beta_2 t} + \epsilon(t)$$
- **Model 2:**
$$y(t) = \frac{\alpha_0 + \alpha_1 t}{\beta_0 + \beta_1 t} + \epsilon(t)$$
- **Model 3:**
$$y(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \epsilon(t)$$

We assume that the error terms $\epsilon(t)$ are i.i.d. normal random variables with a mean of zero and a variance of 2.

3 Question 1

For each model, the least squares estimators (approximated) are as follows:

Model 1: The least squares estimators for Model 1 are given by the vector

$$\hat{\theta} = [\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2] = [\ 2.24822521,\ 0.75791348,\ -1.19816668,\ 0.63901615,\ -1.19752691]$$

Model 2: The least squares estimators for Model 2 are represented by the vector

$$\hat{\theta} = [\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_0, \hat{\beta}_1] = [\ 18.12240246,\ 5.58007048,\ 4.96483903,\ 3.9314642]$$

Model 3: The least squares estimators for Model 3 are represented by the vector

$$\hat{\theta} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4] = [\ 3.62516562,\ -1.36703762,\ -0.0918586,\ 0.98043964,\ -0.46943353]$$

4 Question 2

4.1 Finding the Least Squares Estimators

The least squares estimators were computed using the `curve_fit` function from `scipy.optimize`, which minimizes the sum of squared residuals (SSR):

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This function adjusts the parameters iteratively to minimize the residual sum of squares, ultimately providing the optimal fit for each model.

4.2 Choosing Initial Guesses p_0

The initial parameter guesses (p_0) were selected based on the structures of the models and the trends observed in the data.

- **Model 1**: A constant offset (α_0) of approximately 3 was chosen, reflecting the approximate central value of the data points. Moderate values (around 1) were used for the exponential terms.
- **Model 2**: Parameters were selected to stabilize the fraction, using a constant value of 3 for α_0 in the numerator and smaller values for the denominator to capture gradual changes in the data.
- **Model 3**: The polynomial terms required alternating positive and negative guesses to account for nonlinear trends in the data.

These initial guesses provided a strong foundation for the optimization process, helping it converge effectively.

The initial guesses used for each model are as follows:

- **Model 1**: p_0 `= [3, 1, -1, 1, -1]`
- **Model 2**: p_0 `= [3, 0.5, 1, -0.5]`
- **Model 3**: p_0 `= [3, -1, 1, -1, 1]`

5 Question 3: Best Fitted Model

Answer: ***Model 3*** was selected as the best fit for the data.
Method Used: We obtained the residual sum of squares(SSR) and then compared them for each model. The model with the smallest residual sum of squares was chosen as the best-fit model.

6 Question 4: Estimate of σ^2

To estimate the variance σ^2, we use the formula:

$$\sigma^2 = \frac{\text{Residual Sum of Squares (SSR)}}{n - p}$$

where:

- n is the number of data points,
- p is the number of parameters in the model (including the intercept),
- SSR $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares.

Answer:

- **Model 1:** Variance$_1$: 0.001821, SSR$_1$: 0.1457
- **Model 2:** Variance$_2$: 0.001826, SSR$_2$: 0.1480
- **Model 3:** Variance$_3$: 0.001789, SSR$_3$: 0.1432

7 Question 5: Confidence Intervals Based on the Fisher Information Matrix

The 95% confidence intervals for the parameters of the best-fitted curve are calculated using the method of the Fisher information matrix.

A 95% confidence interval means that you can say with 95% confidence that the true value of the parameter you're estimating lies within that interval.

The confidence intervals for the parameters are as follows:

- **95% Confidence Interval for** β_0: $(3.6733313343298004, 3.5769998984976343)$
- **95% Confidence Interval for** β_1: $(-0.7164062775492698, -2.0176689596053703)$
- **95% Confidence Interval for** β_2: $(2.501033072605175, -2.6847502715027325)$
- **95% Confidence Interval for** β_3: $(4.820122199125774, -2.859242927214341)$
- **95% Confidence Interval for** β_4: $(1.4135798009989946, -2.352446859263766)$

These intervals represent the range within which the true values of the parameters are likely to lie with 95% confidence.

8 Question 6: Plot the Residuals

The residuals plot is shown in **Figure 1**, which can be generated using the code provided in the appendix. For more details, please refer to Figure 1 on the next page.

9 Question 7: Test for Normality Assumption

To test whether the residuals satisfy the normality assumption, we employed the **Shapiro-Wilk test**, which is used to determine if a given sample follows a normal distribution.

- **P-value:** The p-value from the test indicates the likelihood that the sample comes from a normal distribution.
  - If the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that the sample is likely to come from a normal distribution.
  - If the p-value is less than or equal to 0.05, we reject the null hypothesis and conclude that the sample does not follow a normal distribution.
- **p-value for each model is given below:**
  - **Model 1:** p-value = [0.2070]
  - **Model 2:** p-value = [0.3132]
  - **Model 3:** p-value = [0.0819]

    Here, the p-value for each model is greater than 0.05, indicating that all models follow a normal distribution.

Additionally, we plotted the histograms with kernel density estimates for the residuals of each model. From these plots, we observe that the residuals approximately follow a Gaussian distribution. Therefore, we can conclude that the assumption of normally distributed errors $\epsilon(t)$, with a mean of zero and variance σ^2, holds true for the data.

10 Question 8: Plot the Observed Data Points and Fitted Curve

The provided dataset points are plotted in the **Figure 2**, along with the fitted curves according to three different models. Among them, the third model provides the best fit for the data.

11 Conclusion

In this project, we fitted three nonlinear regression models to the dataset and identified **Model 3** as the best fit based on the smallest residual sum of squares (SSR). We computed 95% confidence intervals for the parameters of the best-fitting model, confirming that the estimates are precise. Normality tests on the residuals showed that they follow a normal distribution, validating the assumption of normally distributed errors. Finally, visualizations of the residuals and fitted curves further supported the choice of **Model 3** as the most accurate representation of the data.

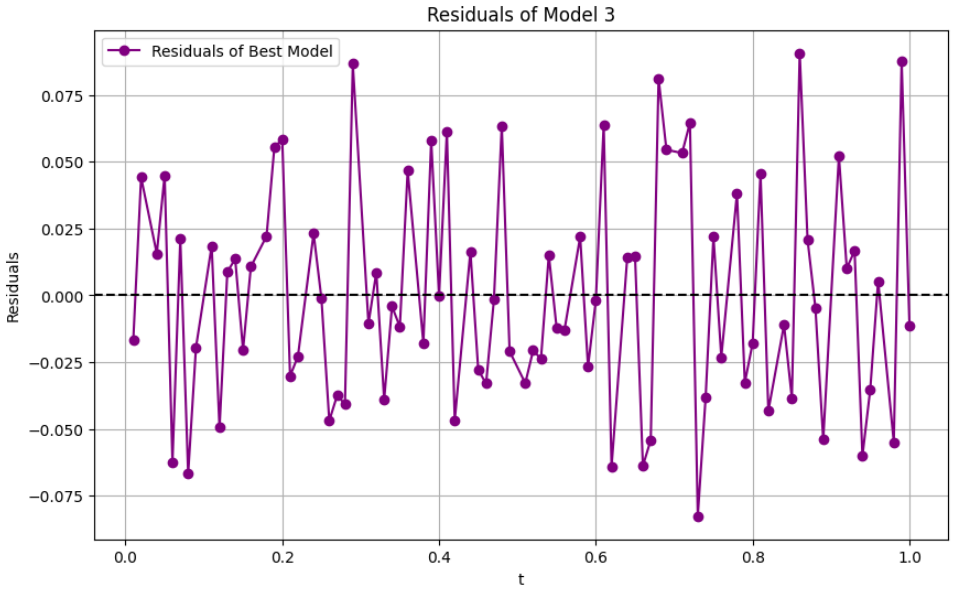

Figure 1: Residuals plot generated from the model.

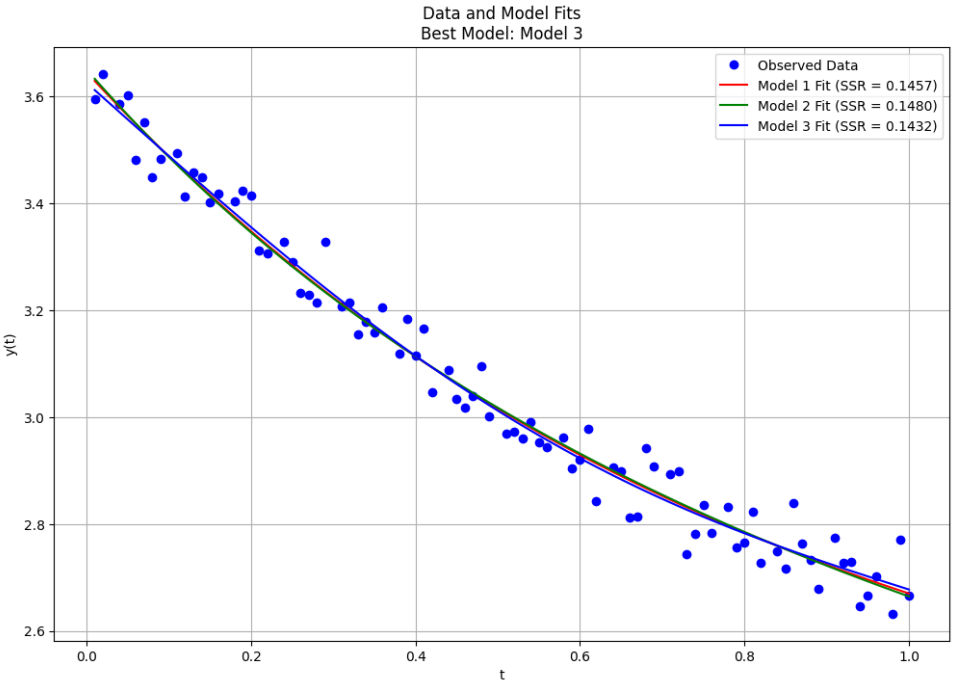

Figure 2: Observed data points with fitted curves for the three models. Model 3 is the best fit.