Indian Institute of Technology Kanpur

MTH686 - Non-linear Regression - PROJECT

Prof. Debasis Kundu

Student: Panuganti Neha

Roll No: 220751

Introduction

In this project, we will work with three non-linear regression models and estimate their parameters using the method of least squares. The models are as follows:

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

Given assumption: $\eta(t)$ is a sequence of i.i.d. normal random variables with mean zero and variance $\sigma^2, i.e., \eta(t) \sim \mathcal{N}(0, \sigma^2)$.
I was given a dataset 'data-39.txt' containing 65 data points, $(t, y(t))$, say.
The entire process with the corresponding code can be found in the following Colab notebook: link.
Let's answer the questions given:

1) Find the least squares estimators of the unknown parameters under three different model assumptions.

After finding initial guesses and applying the Gauss-Newton Method to each model, we obtain the least squares estimators (LSE) for the unknown parameters. The final results are:

- **LSE1 for Model 1:**

$$\alpha_0 = 1.649718,\ \alpha_1 = 0.9712231,\ \beta_1 = 1.53519,\ \alpha_2 = -0.3352953,\ \beta_2 = -1.705869$$

- **LSE2 for Model 2:**

$$\alpha_0 = 4.38258,\ \alpha_1 = 1.771753,\ \beta_0 = 1.916959,\ \beta_1 = -0.9093837$$

- **LSE3 for Model 3:**

$$\beta_0 = 2.286051,\ \beta_1 = 2.04803,\ \beta_2 = 0.7712057,\ \beta_3 = 0.5524677,\ \beta_4 = 0.4391256$$

2) How did you find the least squares estimators? What kind of initial guesses did you choose?

For each model, the least squares estimators were found using the following approaches:

- **Model 1:** I simplified the model by reducing the number of variables (i.e., set $\alpha_2 = 0$) and then applied the linear model (lm) method to find the other coefficients.(Also $\beta_0 = 1.3$, I tried with 0,1,2,1.5 first as I observed Guass Method was not converging and according to data I just choose randomly. Mathematically maybe this can be considered under random search for β_0.)
- **Model 2:** I approximated the model by transforming the terms into reciprocals and assumed $\beta_0 = 1$, then applied the linear model (lm).
- **Model 3:** Since the model is a polynomial regression, I directly applied the linear model (lm) to estimate the coefficients.

3) Find the best fitted model.

I used *AIC(Akaike Information Criterion)* to analyse between the models. *AIC=2k – 2ln(L)* Where:
k is the number of estimated parameters in the model, L is the likelihood of the model
A lower AIC value indicates a better model, as it suggests a model that has a good fit with fewer parameters. After implementing, I got
AIC for Model 1: -218.1109 , Model 2: -219.641 , Model 3: -218.1792 Hence,
Best model based on AIC is **Model 2**
(NOTE == I think Model 1(particularly because Model 2 for high values of t y(t) tends to $\alpha 1/\beta 1$ which is negative and I think it might have overfitted too) would be a better fit as it has less residue sum than Model 2 and the AIC doesn't differ much from Model 2.)

4) Find the estimate of σ^2.

We estimate the variance of the residuals as:

$$\sigma^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\eta}_i^2 = RSS/degreesoffreedom$$

where $\hat{\eta}_i$ are the residuals, n is the number of data points, RSS is the residue square sum and p is the number of parameters estimated. And I got:
$\sigma^2(Model1) = 0.001891561$
$(Model2) = 0.001876198$
$(Model3) = 0.001889575$

5) Find the associated confidence intervals based on the Fisher information matrix.

Involves computing the inverse of the matrix of second derivatives of the sum of squared residuals with respect to the parameters.

$$\text{Confidence Interval for } \hat{\theta}_i: \quad \left(\hat{\theta}_i - z_{\alpha/2}\cdot\sqrt{\left(I(\hat{\theta})^{-1}\right)_{ii}}, \hat{\theta}_i + z_{\alpha/2}\cdot\sqrt{\left(I(\hat{\theta})^{-1}\right)_{ii}}\right)$$

- $\hat{\theta}_i$ is the estimated parameter.
- $z_{\alpha/2}$ is the critical value for the desired confidence level (e.g., $z_{0.025} \approx 1.96$ for a 95% confidence interval).
- $I(\hat{\theta})$ is the Fisher Information Matrix.

$$I_{ij}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln L(X;\theta)}{\partial\theta_i\,\partial\theta_j}\right]$$

- $\left(I(\hat{\theta})^{-1}\right)_{ii}$ is the i-th diagonal element of the inverse Fisher Information Matrix.

My final result:
Confidence Intervals for Model 1:
alpha0 = [1.049113 2.250323]
alpha1 = [0.6831917 1.259255]
beta1 = [1.164576 1.905805]
alpha2 = [-0.5254107 -0.1451799]
beta2 = [-2.33617 -1.075568
Confidence Intervals for Model 2:
alpha0 = [3.551376 5.213784]
alpha1 = [0.9283312 2.615175]
beta0 = [1.430207 2.403711]
beta1 = [-1.632635 -0.1861321]
Confidence Intervals for Model 3:
beta0 = [2.249004 2.323097]
beta1 = [1.79486 2.301199]
beta2 = [0.2263617 1.31605]
beta3 = [-0.1045485 1.209484]
beta4 = [0.01897369 0.8592776]

6) Plot the residuals.

(plots are in the code) I used Quantile plot. (blue line is y=x line(reference) , we can just see if the points line on it ,i.e, theoretical quantities and samples are same) A Q-Q plot compares a dataset's quantiles to those of a theoretical distribution (usually normal). If the points align along a straight line, it suggests the data follows that distribution. Deviations from the line may indicate skewness, outliers, or non-normality, and can be used to check if residuals in regression models meet normality assumptions.

7) Test whether it satisfies the normality assumption or not?

Simply $\eta(t)$ in model equations is expected to be residual and which is expected to be i.i.d,normally distributed so we use 3 Residual plots to examine:

(i) Residuals vs Observation Number(or Time):
The purpose is to check for any patterns in the residuals, which could indicate a problem with the model or the data.
Interpretation in the Context of Normality Assumption:
Assumption: Residuals should be independent. This means that no correlation should exist between residuals across observations or over time.
Expected Pattern: The plot should appear random with no obvious pattern. Ideally, it should look like a "cloud" of points spread around zero without any trend (e.g., no increasing or decreasing spread).
If there's a pattern, such as residuals increasing or decreasing over time, this could suggest autocorrelation, which violates the assumption of independence.

(ii) Residuals vs Fitted Values
The purpose of this plot is to check for non-constant variance, which means that the residuals should have constant variance across all levels of the predicted values.
Interpretation in the Context of Normality Assumption:
Assumption: Residuals should have constant variance (homoscedasticity). This means that the spread of the residuals should be approximately the same for all fitted values

Expected Pattern: There should be no discernible pattern in the spread of residuals. The residuals should look like a random scatter of points with no change in spread as fitted values increase or decrease. It should form a "cloud" or band around the zero line.

(iii) Residuals vs Predictor

The purpose of this plot is to check for any non-linear relationships between the residuals and the predictors, which may indicate that the model is mis-specified.

Interpretation in the Context of Normality Assumption:

Assumption: Residuals should be independent of predictors. This means that there should be no relationship between the residuals and the predictors, as the model should have captured all the relevant structure in the data.

Expected Pattern: Residuals should be randomly scattered around zero. There should be no systematic patterns or structure in the plot. If the model is correctly specified, residuals should not show any relationship with the predictor.

and also a histogram is plotted which is supposed to be in bell shape simply similar to normal distribution curve.

Final Observation:

We can observe all the plots came out as desired i.e cloud/random and histogram shaped bell. This indicates that the residues found do satisfy normal assumptions.

8) Plot the observed data points and fitted curve.

(plots are in the code) If we expect our data to be exponentially growing we can consider Model 1 as the best fit or if we expect it to be a polynomial then Model 3 is the best fit, but if we don't know how the testing data is gonna be we should probably choose Model 2, despite having a slightly higher RSS, Model 2 likely has fewer parameters and fits the data well enough without overfitting. This is reflected in its lower AIC.

link for code -

https://colab.research.google.com/drive/1d9i2sYSJfL5PpNtueHVmpEuXGK1$_4rE8?usp = sharing$