DA5401 Assignment #2

Roll no : DA24C021
Name : Venkatesh Duraiarasan

Task 3-0 : Simple Linear Regression Model for Harmonic Oscillator Data

Assignment Overview

The objective of this task was to model the given harmonic oscillator data using a simple linear regression approach. This model aims to provide a **baseline** for comparison with more complex models, if necessary, and to understand how well a straightforward linear fit performs on the dataset.

Data Description

The dataset provided consists of amplitude measurements (y) of a harmonic oscillator at various time points (x).

Approach

1. **Feature Transformation:**
   - The feature matrix (X) was transformed using polynomial features with a degree of 1. This effectively creates a feature matrix suitable for linear regression by including both the original time variable and a bias term.
2. **Model Training:**
   - Applied Linear Regression to the transformed feature matrix (X_{poly}) and the amplitude data (y).

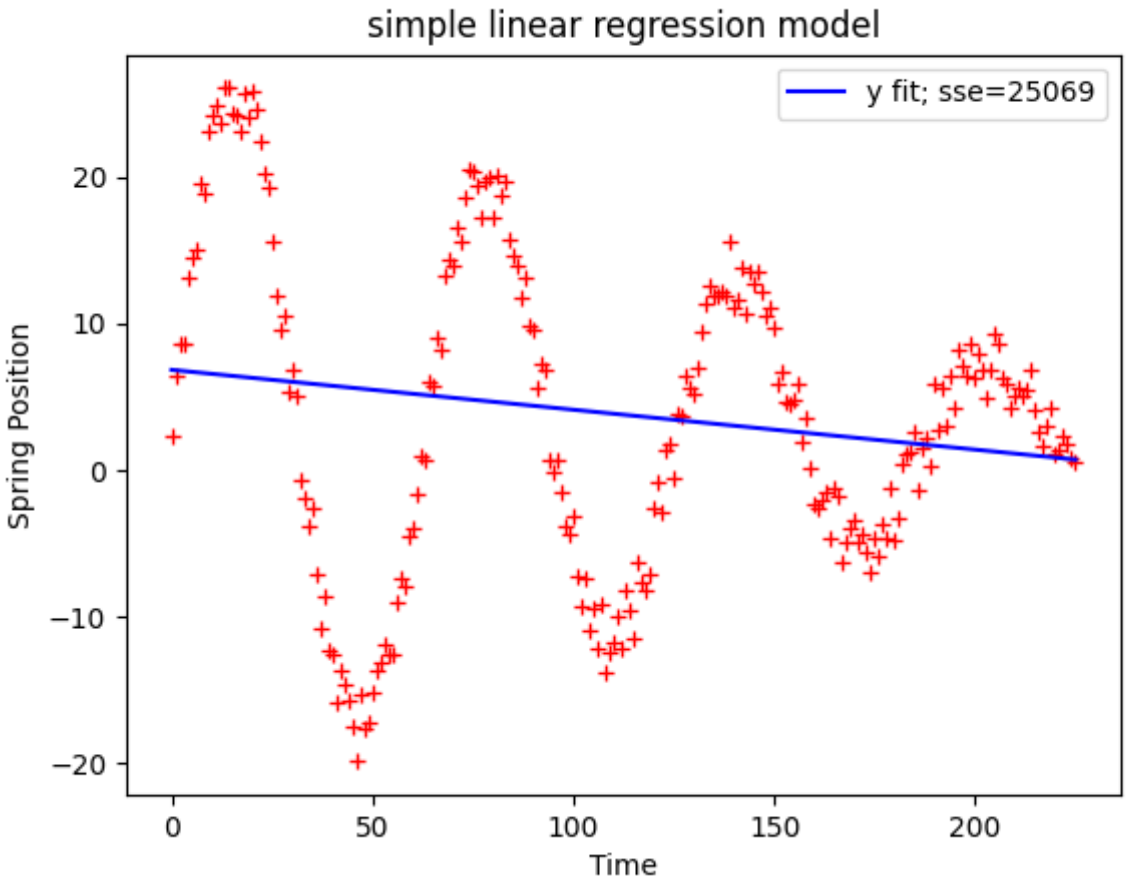
Results

- **Model Parameters:**
  - **Intercept**= 6.8505
  - **Beta** = -0.0272
- **SSE (Sum of Squared Errors):**
  - The SSE of the simple linear regression model was calculated as (**SSE** = **25069**).

Visualization

The plot displays:

- **Raw Data:** Represented with red plus signs ('r+').
- **Regression Fit:** Shown as a blue line with the label indicating the SSE value.



Task 3-2

Regression Model for Harmonic Oscillator Dataset

Task Overview

In this task, the objective was to implement a regression model using the provided harmonic oscillator dataset, with the goal of achieving a lower Sum of Squared Errors (SSE) compared to the previous task. The model was to be implemented with appropriate feature transformation to enhance performance.

Data Description

The dataset consists of amplitude measurements of a damped harmonic oscillator. The independent variable x represents the time index, while the dependent variable (y) represents the amplitude of the oscillator at each time index.

Feature Transformation

Given the nature of the data from a damped harmonic oscillator, directly applying a linear model was challenging due to the non-linearity of the underlying physical model. The original model $y = A\exp(-dx)\sin(\sqrt{1-d^2}wx + p)$ was difficult to linearize. Therefore, an alternative approximation was used:

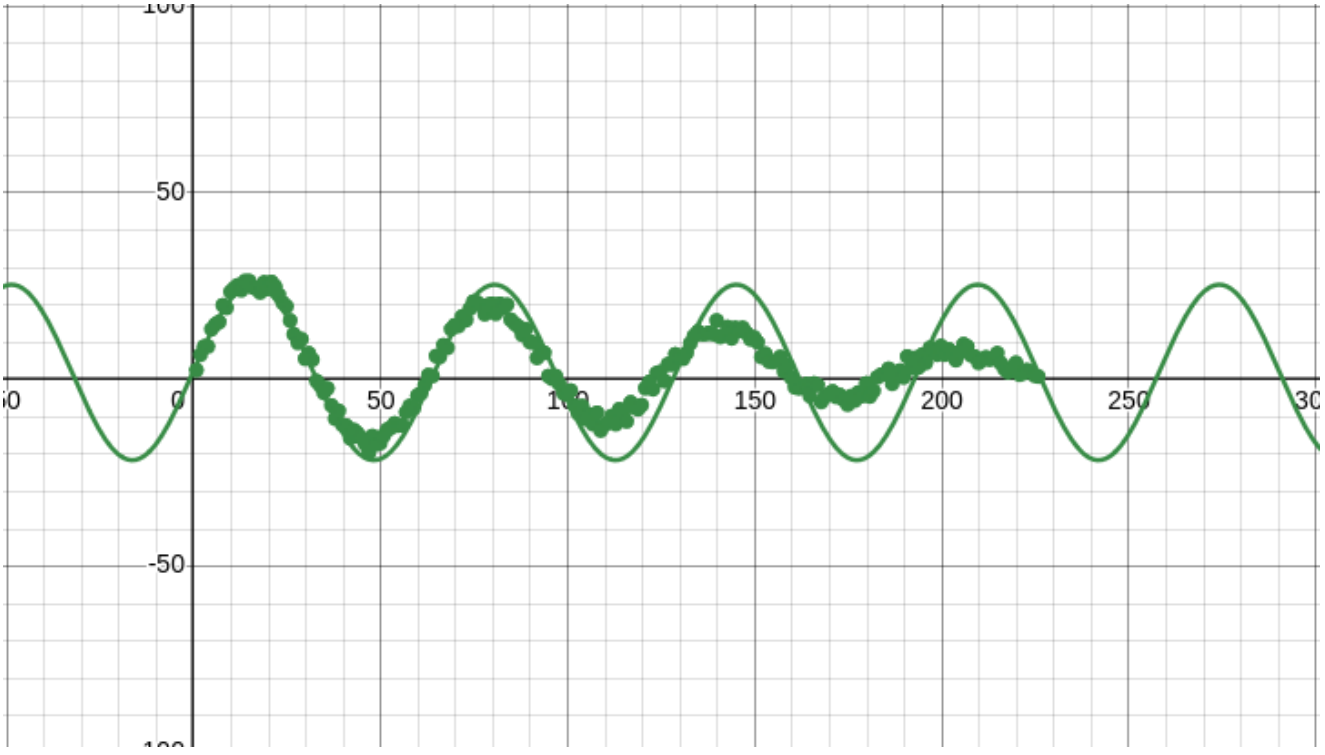
$y = c + m_1x + m_2\sin(x) + m_3x\sin(x)$

Here, (x) is the time index transformed into radians to capture the oscillatory nature of the data.

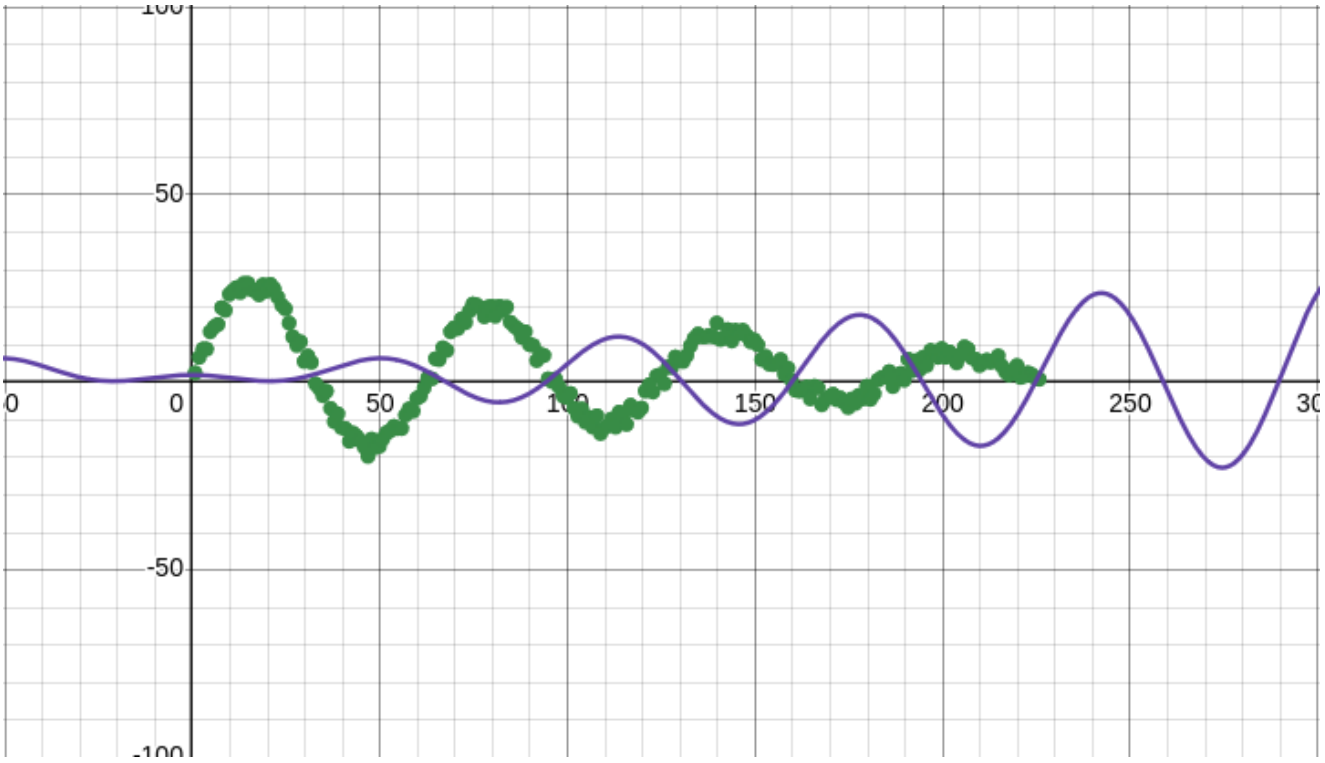
Why this model $(y = c + m_1x + m_2\sin(x) + m_3x\sin(x))$ was considered?

$y = c + m_1x + m_2\sin(x) + m_3x\sin(x)$ approximates very well for the damped oscillation for the given limited time interval of the dataset

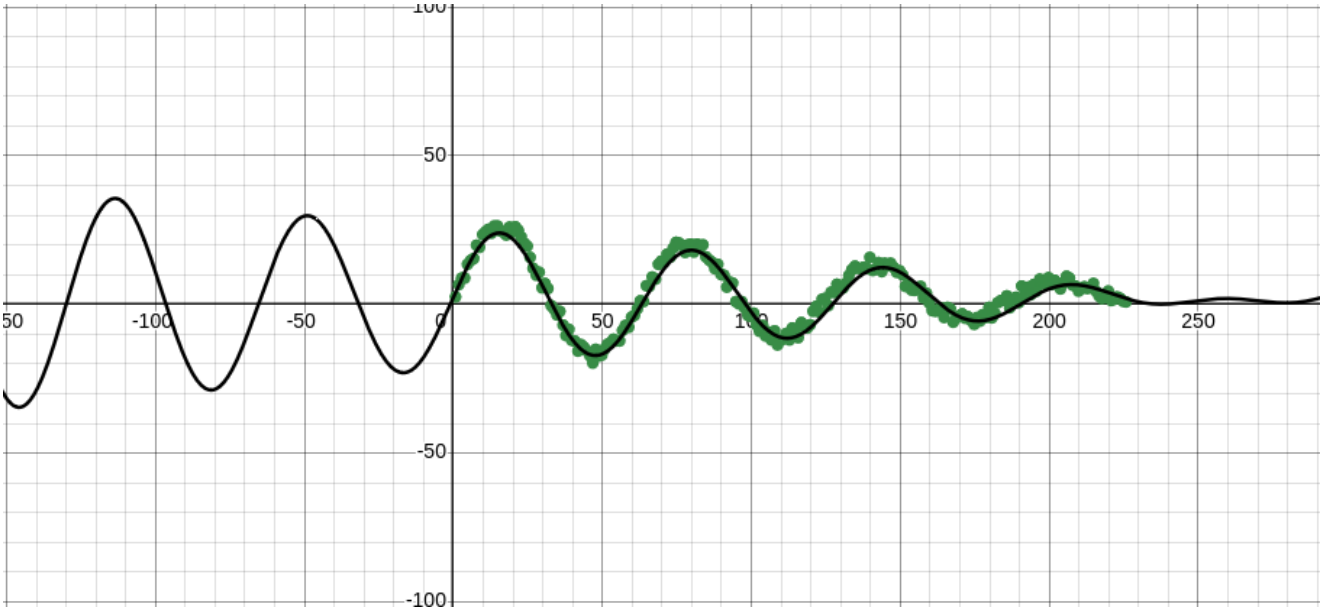
$m_2\sin(x)$ - part models the periodicity contained in the dataset as shown below(dots are the given y values)



$m_3x\sin(x)$ - part models the decaying part by cancelling the $m_2\sin(x)$ part as x or θ time increases



so, combining the two graphs $y = c + m_2 \sin(x) + m_3 x \sin(x)$ (**black curve**) approximates the amplitude data and **can be used for both right and left extrapolation**



Steps Taken

1. **Data Preparation:**
   - Loaded and organized the data into a DataFrame with columns for time (x) and amplitude (y).
   - Found zero crossings in the amplitude data to estimate the number of cycles and angular frequency.
   - Used a visual inspection to confirm that there were approximately 7 valid zero crossings.
2. **Feature Transformation:**
   - Scaled the time variable (x) to radians based on the estimated number of cycles.
   - Created new features for the regression model:

- $x_\theta = \text{angular\_freq} \times x$
- $x2 = \sin(x_\theta)$
- $x3 = x_\theta \times \sin(x_\theta)$

- Added a bias term to the features for the linear regression model.

3. **Model Training:**
   - Applied Linear Regression using the transformed features.
   - Predicted amplitude values using the trained model.
4. **Evaluation:**
   - Calculated the SSE of the model's predictions.
   - Compared the SSE to that from Task 1 to ensure improvement.

Results

- **Model Parameters:**
  - Intercept (c): 2.529
  - Coefficients (m_1), (m_2), (m_3): 0.00398, 24.925, -1.054, respectively.
- **SSE (Sum of Squared Errors):**
  - The SSE achieved with the new model was 1038.86.

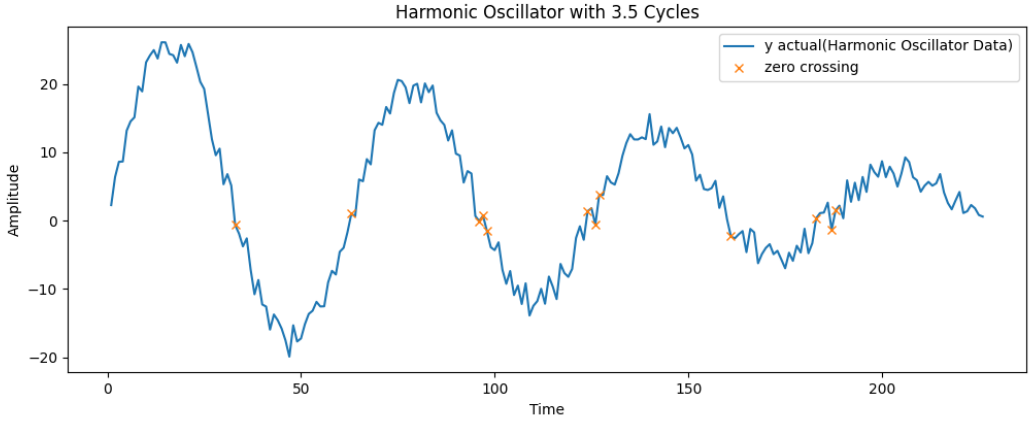
Inference

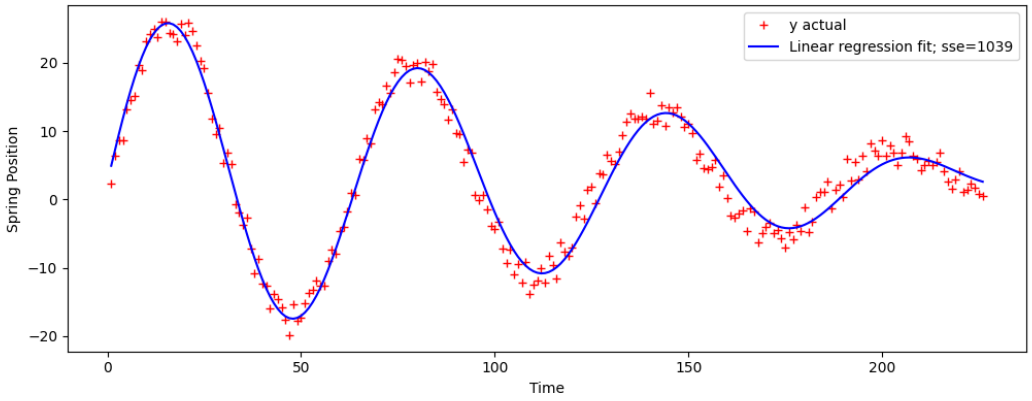
- From coefficients we can infer that
  - weight $m_1 = 0.003$ is relatively smaller than other components. This means factor $m_1 x$ can be dropped from model
  - $m_2 = 24.925$ is effectively the (A), amplitude of the damped oscillator model
  - $m_3 = -1.054$ is the factor for damping variable $x \sin(x)$ which cancels out $m_2 \sin(x)$ as x increases

Visualization

Two plots were generated:

1. **Harmonic Oscillator Data (y):**
   - Displays the amplitude data along with zero crossing points.
2. **Regression Fit:**
   - Shows the actual amplitude data(y) and the regression line($\hat{y}$), indicating the fit quality and the SSE value.





Conclusion

The regression model with the specified features provided an SSE of 1038.86, showing an improvement in fitting the data compared to Task 1 SSE=25069. The chosen model ($y = c + m_1x + m_2\sin(x) + m_3x\sin(x)$) effectively captures the damped oscillatory nature of the data, with parameter values indicating a reasonable fit.

Task 3-3

Regression Model for Interpolation and Evaluation

Assignment Overview

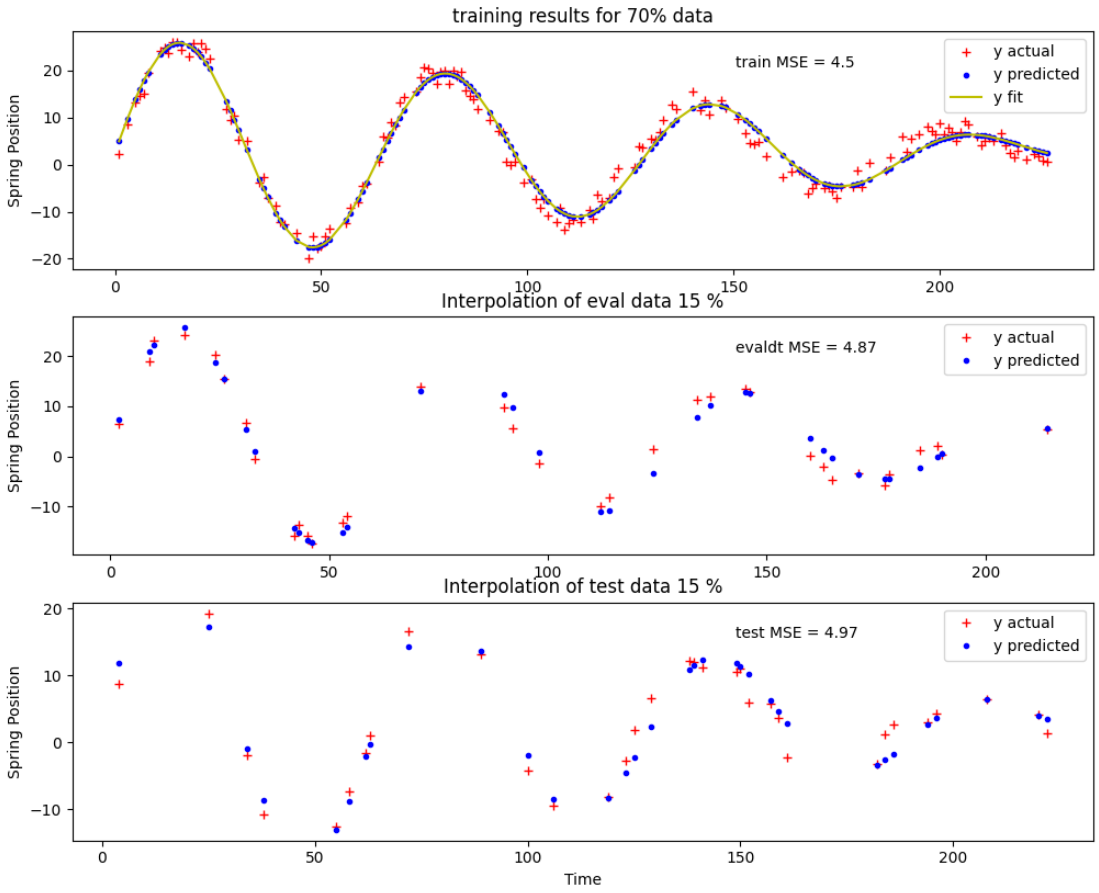
In this task, the goal was to train a regression model on the harmonic oscillator dataset with a focus on interpolation. The model's performance was evaluated using the Mean Squared Error (MSE) on both training and test datasets. The process involved feature transformation, data splitting, model training, prediction, and evaluation.

1. **Data Splitting:**

- Utilized the `split_data_for_interpolation` a custom function to divide the data into training (70%), eval (15%) and test (15%) subsets.

2. **Model Training:**
   - Trained a Linear Regression model using the training dataset.
3. **Prediction and Evaluation:**
   - Generated predictions on the training and test datasets.
   - Calculated the Mean Squared Error (MSE) for both datasets to evaluate model performance.

Visualization



Model Performance

The Linear Regression model was trained and evaluated with the following outcomes:

- **Training MSE:** 4.5
- **Eval Dataset MSE**: 4.87
- **Test MSE:** 4.97

The model performed reasonably well on both training and test datasets, with the test MSE slightly higher than the training MSE, indicating some degree of **overfitting**. The visual plots confirm that the model captures the general periodicity of the data, although there is room for improvement in fitting the data more precisely.

Task 3-4

Regression Model for Extrapolation and Evaluation

Assignment Overview

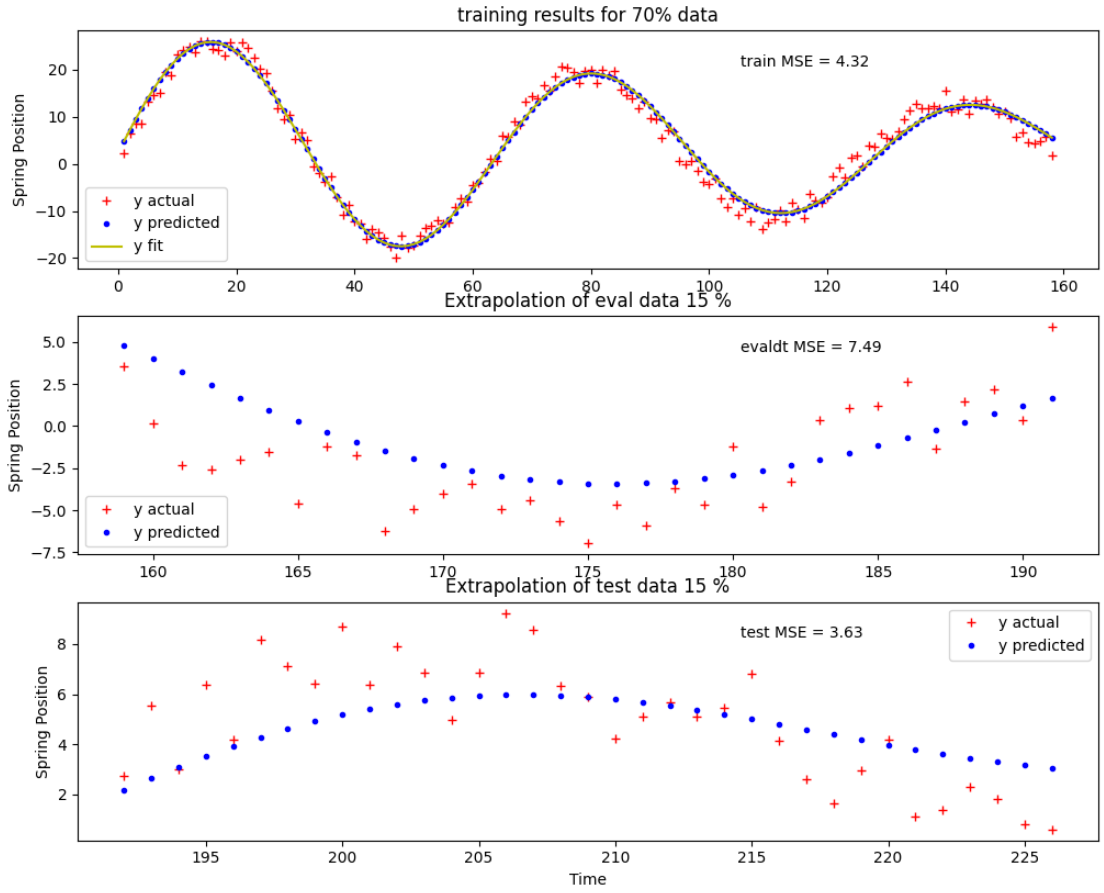
The objective of Task 3-4 was to train a regression model for extrapolation using the harmonic oscillator dataset and evaluate its performance based on the Sum of Squared Errors (SSE). Unlike interpolation, extrapolation involves making predictions beyond the range of the training data, which often poses additional challenges.

Data Description

The dataset comprises amplitude measurements ((y)) of a harmonic oscillator at various time points ((x)). The dataset was initially loaded and organized into a DataFrame for analysis.

Approach

1. **Data Splitting:**
   - Divided the data into training (70%), eval (15%) and test (15%) subsets sequentially as extrapolation typically requires careful splitting to ensure the test set contains data outside the training range.
2. **Feature Matrix Transformation:**
   - same as in previous task
3. **Model Training:**
   - same as in previous task
4. **Prediction and Evaluation:**
   - Generated predictions on the training and test datasets.
   - Calculated the Mean Squared Error (MSE) for both datasets to evaluate model performance.



Model Performance

- **Training Data MSE:** 4.32
- **Eval Data MSE:** 7.49
- **Test Data MSE:** 3.63

The model showed reasonable performance on the training data but exhibited a higher MSE on the eval data, indicating difficulties with extrapolation. This outcome is expected as linear models often struggle with extrapolating beyond the observed data range.