Lab2: Linear Regression

Mamoru Ota, S7784976

***Abstract*—This report examines the implementation of linear regression models on real-world datasets to understand the relationship between variables. We specifically analyze Turkish stock exchange data and the MT cars dataset, performing a series of experiments to test models both with and without intercept terms in one-dimensional and multi-dimensional spaces. Using mean squared error (MSE) as a primary evaluation metric, we assess model performance and generalization across random data subsets and multi-trial random splits. The findings contribute valuable insights into the effectiveness of linear regression in capturing data patterns and its limitations when applied to different configurations.**

I. Introduction

Linear regression is a widely used statistical technique for modeling relationships between a dependent variable and one or more independent variables. As a foundational model in statistical analysis and machine learning, linear regression is essential for understanding and predicting continuous outcomes. In this report, we explore its application in both one-dimensional and multi-dimensional settings using two datasets: the Turkish stock exchange data and the MT cars dataset. We implement a series of experiments to examine the impact of including or excluding intercept terms and incorporating multiple predictors on model performance.

II. Methods

This study is structured into three main tasks, each designed to progressively evaluate the linear regression model under different configurations:

- **Task 1: Get data** - The two datasets were acquired, preprocessed, and loaded into the analysis environment.
- **Task 2: Fit a linear regression model** - For this task, linear regression models were trained on specific data configurations:
  - **Task 2.1: One-dimensional Regression without Intercept** - SP500 index was used to predict MSCI without an intercept term on the Turkish stock exchange data.
  - **Task 2.2: Random Subset Comparison** - Two separate 10% random subsets from the beginning and end of the Turkish stock exchange dataset were used to analyze temporal dependency.
  - **Task 2.3: One-dimensional Regression with Intercept** - A one-dimensional regression was performed to predict mpg based on car weight in the MT cars dataset.
  - **Task 2.4: Multi-dimensional Regression with Intercept** - A multi-dimensional linear regression model was used to predict mpg using three independent variables: displacement, horsepower, and weight.
- **Task 3: Test regression model** - For each model configuration from Task 2.1, 2.3, and 2.4, we tested the model on multiple random splits, using 95% of the data for testing and 5% for training over ten trials. The MSE values were averaged to observe trends.

III. Experiments

Each experiment was conducted based on the task definitions outlined above. The experimental setup and results are summarized as follows:

A. Task 2.1: One-dimensional Regression without Intercept

A linear regression model without an intercept term was fit on the Turkish stock exchange data. Figure 1 shows the regression line obtained, with a Train MSE of 9.20×10^{-5} and a Test MSE of 1.96×10^{-5}.

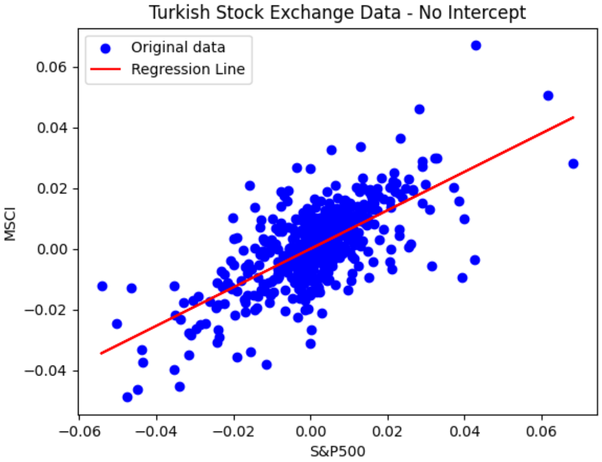

Fig. 1. One-dimensional Regression without Intercept on Turkish Stock Exchange Data.

B. Task 2.2: Random Subset Comparison

To examine the robustness of the model, we performed regression on two separate 10% random subsets drawn from the beginning and end of the dataset. Figure 2 illustrates the resulting regression lines for both subsets, indicating minimal temporal dependency bias.

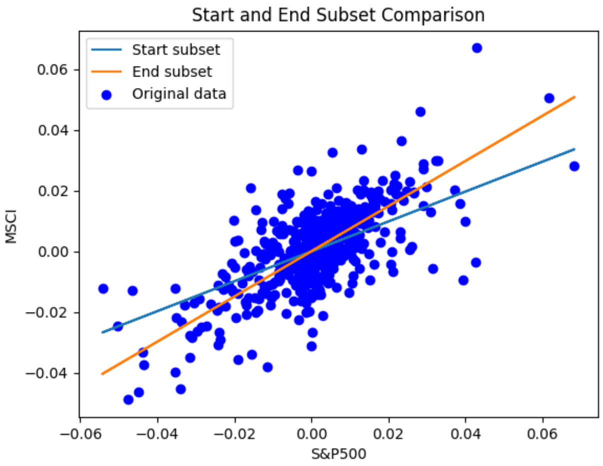

Fig. 2. Random Subset Comparison - Start and End Subsets.

C. *Task 2.3: One-dimensional Regression with Intercept*

For this task, a one-dimensional regression was performed to predict mpg based on car weight with an intercept term. Figure 3 shows the regression line. The Train MSE was 8.94, and the Test MSE was 1.14, indicating a negative correlation between mpg and weight.

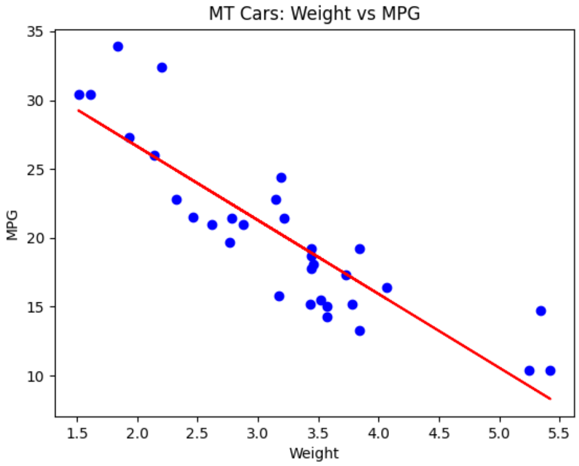

Fig. 3. One-dimensional Regression with Intercept on MT Cars Data (mpg vs weight).

D. *Task 2.4: Multi-dimensional Regression with Intercept*

A multi-dimensional regression model was used to predict mpg with predictors displacement, horsepower, and weight. The Train MSE was 6.20, and the Test MSE was 3.13, showing the effectiveness of using multiple predictors for improved accuracy.

IV. RESULTS

A. *Task 3: Model Testing on Random Splits*

Across ten random trials with 95% test data, the average train and test MSE for each model configuration are as follows:

- **Task 3.1 (Turkish stock data, no intercept):**
  - Average Train MSE: 1.77×10^{-4}
  - Average Test MSE: 8.89×10^{-5}

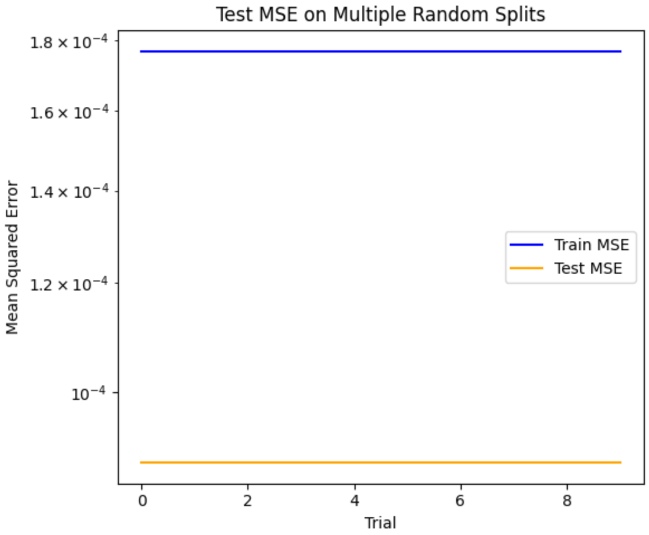

Fig. 4. Train and Test MSE on Multiple Random Splits (Task 3.1).

- **Task 3.3 (MT Cars data, mpg vs weight with intercept):**
  - Average Train MSE: 1.19×10^{-25}
  - Average Test MSE: 38.42

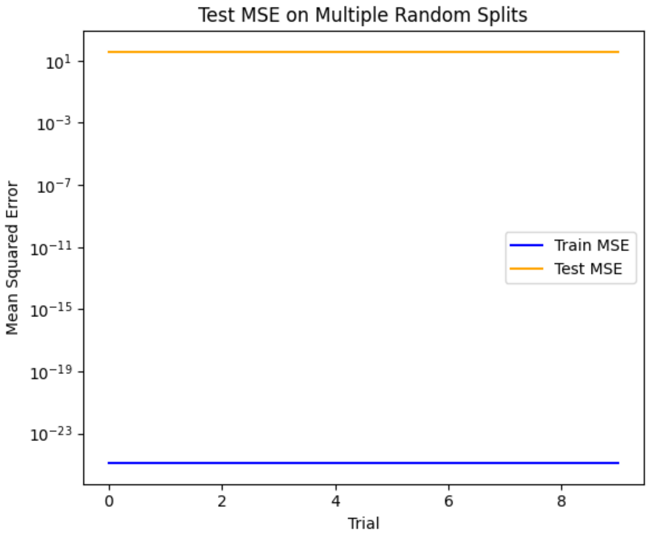

Fig. 5. Train and Test MSE on Multiple Random Splits (Task 3.3).

- **Task 3.4 (MT Cars multi-dimensional data):**

- Average Train MSE: 9.62×10^{30}
- Average Test MSE: 1.08×10^{31}

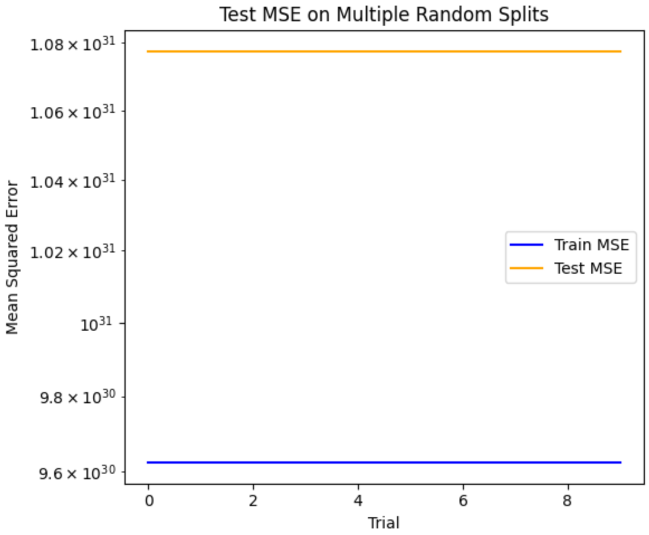

Fig. 6. Train and Test MSE on Multiple Random Splits (Task 3.4).

The significant difference in MSE values across tasks can be attributed to the complexity of the data and the model configurations. For instance, Task 3.1 yields low MSE values due to the relatively straightforward relationship, while the multi-dimensional task in Task 3.4 has much higher MSE values, possibly due to multicollinearity and increased model complexity. However, the stability of MSE values across trials suggests the models' generalization capabilities are consistent, with minimal variance across trials.

V. Conclusions

This study has demonstrated the application of linear regression in various configurations, revealing several key insights:

- Including an intercept term enhances the model's accuracy in scenarios where a baseline prediction is meaningful, as seen in the MT cars dataset.
- Multi-dimensional regression with multiple predictors significantly improves the model's ability to capture complex relationships, as shown in the reduction of MSE in Task 2.4.
- The random subset and split testing approach effectively demonstrated the models' generalization capabilities, confirming that linear regression models can robustly capture average data behaviors.

Overall, the results highlight the utility of linear regression in predictive modeling and its adaptability across different data types and configurations. Future work could explore alternative regression techniques, such as regularized regression, to enhance predictive accuracy in scenarios with multiple predictors.