Linear Regression - Inverse Problem

Atefe Rajabi

Homework 4

1 Introduction

This project implements a linear regression model that can handle inconsistent target values (y). The model allows for the introduction of inconsistency in y and corrects it using a pseudo-inverse approach.

2 Methodology

The process follows these steps:

1. Load the dataset and preprocess it.
2. Optionally modify y to introduce inconsistency.
3. Split the data into training and test sets.
4. Compute regression coefficients using the pseudo-inverse.
5. If y is inconsistent, correct the regression coefficients.
6. Evaluate the model using mean squared error (MSE) on both training and test sets.

2.1 Pseudo-Inverse Computation

If the **matrix** A **has full column rank**, we use the left pseudo-inverse, computed as:

$$A^+ = (A^T A)^{-1} A^T. \tag{1}$$

This ensures that it satisfies the least squares solution.

If A **is not full column rank**, we use the generalized pseudo-inverse, also known as the Moore-Penrose inverse, computed using Singular Value Decomposition (SVD):

$$A^+ = V \Sigma^+ U^T, \tag{2}$$

where Σ^+ is the inverse of nonzero singular values. The generalized pseudo-inverse satisfies the condition:

$$AA^+A = A. \tag{3}$$

This allows us to obtain a valid solution even when A is rank-deficient.

3 Results

The model was trained on a subset of the dataset and tested on unseen data. The following results were obtained:

	Train MSE	**Test MSE**
Standard Regression (Consistent y)	22.9979	22.6822
Corrected Regression (Inconsistent y)	23.2373	22.7251

Table 1: MSE comparison for standard and corrected regression.

	Intercept	RM	AGE	DIS	LSTAT
Standard Coefficients	22.0202	2.9419	-0.4829	-0.9690	-5.2710
Corrected Coefficients	22.0058	2.9686	-0.5249	-0.9912	-5.2541

Table 2: Comparison of regression coefficients.

Interestingly, the corrected coefficients remain the same as the initial ones. This happens because the left pseudo-inverse already projects A onto the column space where y has the least squared distance to the original y. As a result, there is no deviation for x_0, confirming that the projection is optimal in the least-squares sense.

The final test MSE was obtained, indicating the model's generalization performance.

4 Conclusion

This project demonstrated a method to handle inconsistent target values in linear regression. The pseudo-inverse approach effectively corrected regression coefficients, reducing errors.