Predicting Student Math Performance Using Linear Regression

Name: Shafeer Saidalavi

Course/Module: Data Science Foundations (Linear algebra, statistics)

Institution: Brototype

Submission Date: 02-07-2025

Abstract

This project applies statistical and linear algebra techniques to analyze and predict students' math performance using the "StudentsPerformance" dataset. Key methods include exploratory data analysis, linear regression, and principal component analysis (PCA). The model achieved an R^2 score of 0.88, indicating high predictive accuracy. Strongest influencing features included writing score. This analysis provides valuable insights into factors that impact student achievement and demonstrates the effective use of linear modelling in educational data science.

Keywords

Linear regression, PCA, student performance, educational analytics, statistical modeling

Contents

Introduction

Student academic performance is a key metric of education systems. Understanding the factors that influence performance can help design better policies and interventions. This project uses data science to analyze how various demographic and preparatory factors affect math scores. Linear regression is used for

prediction, supported by statistical testing and linear algebraic analysis such as PCA.

Problem Statement

Objective: Predict math scores of students based on demographic and academic-related features.

Question: What factors most strongly influence student math scores, and how accurately can we predict these using a linear regression model?

Impact: Enables data-driven educational strategies, resource allocation, and personalized support.

Dataset Description

- Source: Kaggle - StudentsPerformance.csv
- Records: 1000
- Features: Gender, race/ethnicity, parental education level, lunch type, test preparation course, reading score, writing score
- Target: Math score
- Tools: Python, pandas, NumPy, seaborn, scikit-learn

Methodology

Exploratory Data Analysis (EDA)

- Box plots, scatterplots and heatmaps
- Correlation matrix showed strong positive correlation among reading, writing, and math scores
- Outliers detected using boxplots for all score distributions

Statistical Techniques Applied

- Hypothesis testing for relationships between categorical variables and scores
- Linear regression for predictive modelling

Linear Algebra Concepts Applied

- Dataset represented as matrix X (features) and vector y (math score)
- PCA performed manually using covariance matrix and eigenvalues
- Explained variance ratio for each principal component calculated
- Linear transformation applied

Model Building & Evaluation

Model Selection

- Chosen model: Linear Regression

Implementation

- Scaled features
- One-hot encoded categorical variables
- Dropped target column
- Train-test split (80:20)

Performance Evaluation

- R^2 Score: 0.88

Strongest predictors: writing score, gender.

Results and Discussion

- Linear regression model performed well with high R^2
- PCA showed that 2 components explained ~34% variance

Conclusion

This project successfully predicted student math scores using a simple linear model. Statistical and linear algebra tools were effectively applied to preprocess data, reduce dimensionality, and interpret results. The approach highlights the importance of preparation and academic background in educational performance.