Regression and Logistic Regression Analysis Using the Cars Dataset

Ane Novrup Larsen, Nathasja Skov Fink Nielsen

October 24, 2025

Abstract

This report applies two supervised learning models—multiple linear regression and logistic regression—to the `cars.csv` dataset. The linear regression model predicts a car's acceleration time based on engine and weight features, while the logistic regression model classifies whether a car is US-made or foreign-made. Both models were trained and evaluated using standard metrics. Linear regression achieved an R^2 of 0.59, showing moderate predictive power. Logistic regression reached ROC AUC scores above 0.94 on both validation and test sets, indicating strong classification performance. The results demonstrate how regression techniques can be used to model both continuous and binary outcomes from structured data.

1 Introduction

Regression models are widely used in machine learning to predict outcomes based on input features. This project applies two types of regression—linear and logistic—to a dataset of car specifications.

The goal is to demonstrate how regression techniques can model both continuous and binary outcomes. Linear regression is used to predict a car's acceleration time, while logistic regression classifies whether a car is US-made. Both are based on features like *horsepower*, *displacement*, *weight*, and *cylinders*.

The comparison highlights both their similarities in preprocessing and their differences in evaluation and interpretation.

1.1 Dataset Overview

The dataset contains 398 cars with technical specifications such as `cylinders`, `displacement`, `horsepower`, `weight`, and `acceleration`, along with metadata like `model year`, `origin`, and `car name`. This structure allows for both regression and classification tasks using numeric and categorical features.

2 Linear Regression Analysis

2.1 Methodology

A multiple linear regression model was used to predict a car's acceleration time (0–60 mph) from four features: `weight`, `horsepower`, `displacement`, and `cylinders`. The workflow consisted of data preparation, model training, and evaluation.

2.1.1 Data Preparation

The `cars.csv` dataset was cleaned by converting non-numeric `horsepower` values and removing missing rows. The inputs (X) included the four predictors and `acceleration` was the target (y). Data was split into 80% training and 20% testing sets.

2.1.2 Model Training

The model was trained using `Linear Regression` from `scikit-learn`, estimating coefficients (β_i) that define how each feature influences acceleration.

2.1.3 Model Evaluation

Performance was assessed using Mean Squared Error (MSE) and the coefficient of determination (R^2). The model achieved an R^2 of approximately 0.59 and an MSE of 3.04, indicating that it explains about 59% of the variation in acceleration times.

2.2 Results

The fitted regression equation was expressed as:

$$\begin{aligned}\text{acceleration} = \beta_0 + \beta_1(\text{weight}) + \beta_2(\text{horsepower}) \\ + \beta_3(\text{displacement}) + \beta_4(\text{cylinders})\end{aligned} \tag{1}$$

The coefficients followed expected trends: `weight` had a positive coefficient, indicating that heavier cars accelerate more slowly, while `horsepower`,

`displacement`, and `cylinders` had negative coefficients, showing that stronger engines reduce acceleration time. The results indicate that at least one predictor variable significantly affects the acceleration time, supporting the alternative hypothesis (H_1) and leading to the rejection of the null hypothesis (H_0).

2.3 Conclusion

The linear regression model effectively captured how vehicle characteristics influence acceleration. Although its predictive accuracy was moderate, it provided interpretable insights consistent with real-world automotive dynamics. Future improvements could include polynomial or regularized regression methods to better capture nonlinear relationships.

3 Logistic Regression Analysis

3.1 Methodology

A logistic regression model was used to classify cars as American-made or foreign-made based on five features: `cylinders`, `displacement`, `horsepower`, `weight`, and `acceleration`. The process included cleaning the data, scaling the features, training the model, and checking how well it performed.

3.1.1 Data Preparation

The `cars.csv` dataset was cleaned by removing rows with missing `horsepower` values and converting them to numeric. A new binary target column was created: cars with `origin` = 1 were labeled as 1 (US/American-made), and all others as 0 (foreign-made). The data was split into 70% training, 15% validation, and 15% test sets. I scaled all features using `StandardScaler`, fitted only on the training data to prevent data leakage.

3.1.2 Model Training

The model was trained using `Logistic Regression` from `scikit-learn`, with no regularization (`penalty=None`) and the `lbfgs` solver. I set the tolerance to `1e-2`, which made the model stop early after 10 iterations. This configuration was selected after comparison with regularized alternatives as it has better log-loss performance.

3.1.3 Model Evaluation

Performance was assessed using accuracy, log-loss, and ROC AUC on the validation and test sets. The model achieved a validation ROC AUC of 0.9457 and a test ROC AUC of 0.9286. Additionally, 5-fold cross-validation on the training set gave a mean ROC AUC of 0.9435. This shows that the model performs consistently across different splits.

3.2 Results

The logistic regression equation looked like this:

$$\log\left(\frac{P(\text{US})}{1-P(\text{US})}\right) = \beta_0 + \beta_1(\text{displacement}) + \beta_2(\text{cylinders}) + \beta_3(\text{weight}) + \beta_4(\text{horsepower}) + \beta_5(\text{acceleration})$$

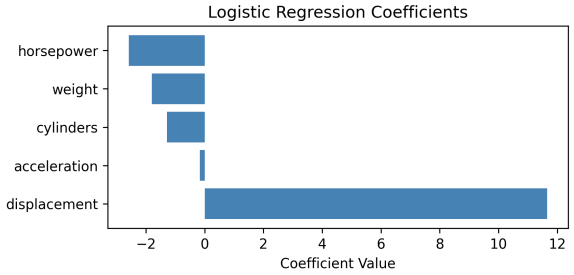

Figure 1: Logistic regression coefficients sorted by magnitude.

The coefficients revealed that `displacement` had the strongest positive influence on predicting American-made cars, while `horsepower`, `weight`, and `cylinders` had negative coefficients. These results support the alternative hypothesis (H_1), indicating that at least one feature significantly affects the classification. Therefore, the null hypothesis can be rejected.

3.3 Conclusion

The logistic regression model showed strong predictive performance. Its ROC AUC scores were consistent across validation and test sets, suggesting good generalization. Regularized versions were tested but gave worse log-loss, so the final setup used no penalty. Future work could explore alternative models to improve performance.

4 Conclusion

Both models performed well within their respective tasks. Linear regression provided interpretable insights into acceleration, while logistic regression achieved high classification accuracy. Future work could explore more complex models to improve performance further.