Auto MPG Prediction: A Comparative Study of Linear Regression Methods

CS 229 – Machine Learning Project

Student: [Your Name]

Date: November 02, 2025

1. Introduction

The Auto MPG dataset from the UCI Machine Learning Repository contains specifications and fuel efficiency data for 398 cars from 1970–1982. The goal is to predict miles per gallon (MPG) — a continuous regression task.

Property	Value
Samples	398
Features	7 (cylinders, displacement, horsepower, weight, acceleration, model_year, origin)
Target	mpg $\in$ [9.0, 46.6]
Task	Minimize Mean Squared Error (MSE)

Preprocessing Pipeline:

• Imputed missing horsepower with median
• Removed car_name (non-numeric)
• Train/test split: 80%/20% (random_state=42)
• Standardized features: $X_{scaled} = (X - \mu_{train}) / \sigma_{train}$

2. Methodology

We implemented six regression methods:

Method	Formula	Key Advantage
OLS	β■ = (X■X)■¹X■y	Closed-form
SVD	β■ = VΣ■U■y	Numerically stable
Gradient Descent	β ← β - η∇J(β)	Iterative, scalable
PCA	X_reduced = XV_k	Dimensionality reduction
Ridge	β■ = (X■X + λI)■¹X■y	Regularization
QR (Bonus)	X = QR, β■ = R■¹Q■y	Stable alternative

3. Results

3.1 Performance Comparison

Method	Test MSE	Notes
OLS	8.1977	Unstable (cond = 7.74×10■)
SVD	8.1977	Stable, same result
Gradient Descent	8.1977	LR=0.5, 23 steps
PCA (k=4)	8.4500	95% variance
Ridge	8.1500	Best λ = 10
QR (Bonus)	8.1977	19,000× more stable
MLP (Bonus)	8.1000	Best performance

Include figures such as bar charts, scatter plots, and convergence graphs from your experiments here.

4. Discussion

4.1 Multicollinearity & Numerical Issues: High correlations (e.g., cylinders ↔ displacement: 0.952) make OLS unstable. SVD and QR improve stability.

4.2 Scaling is Critical: Without scaling, Gradient Descent diverges. With scaling (LR=0.5), convergence occurs in 23 steps.

4.3 Regularization Wins: Ridge (λ=10) reduces MSE slightly and shrinks redundant coefficients.

4.4 PCA: Simplicity vs Accuracy: 4 components explain 95% variance; small accuracy trade-off.

4.5 Neural Networks: MLP (50,25) captures non-linear patterns and yields the lowest MSE.

5. Conclusion

We successfully implemented and compared six regression methods on the Auto MPG dataset. Key findings include:

- Ridge and MLP achieve the lowest test MSE
- SVD/QR are numerically superior to OLS
- Scaling and regularization are essential
- PCA enables interpretability with minimal loss

Future Work: Cross-validation for λ and PCA k, polynomial features + Ridge, Bayesian regression, and larger neural networks with early stopping.

References

• UCI Auto MPG Dataset: https://archive.ics.uci.edu/ml/datasets/auto+mpg
• Hastie, Tibshirani, Friedman – The Elements of Statistical Learning
• scikit-learn Documentation