In the Name of God

دانشگاه شیراز

Shiraz University

Navid Zare

40435071

Statistical Pattern Recognition
Bonus Section: Regularized Linear Regression with SGD

1. Introduction and Methodology

1.1 Objective

The objective of this bonus section is to extend the linear regression implementation by incorporating Ridge Regression (L2 regularization) trained using Stochastic Gradient Descent (SGD). This analysis aims to:

- Analyze multicollinearity in the feature set using correlation analysis and VIF
- Implement Ridge Regression with SGD for multiple regularization parameters
- Compare model stability and performance across different λ values
- Analyze the bias-variance tradeoff in regularized linear models

1.2 Dataset Description

The Auto MPG dataset contains 392 samples with three input features (weight, horsepower, displacement) and one target variable (mpg - miles per gallon). The data was split into training (314 samples, 80%) and testing (78 samples, 20%) sets using random_state=42.

Feature	Mean	Std Dev	Description
weight	2977.58	848.32	Vehicle weight in lbs
horsepower	104.47	38.44	Engine power
displacement	194.41	104.51	Engine displacement
mpg	23.45	7.81	Fuel efficiency (target)

1.3 Mathematical Foundations

1.3.1 Linear Regression Model

The linear regression model with multiple features is defined as:

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = \theta^\top X$$

where θ_0 is the bias term and $\theta_1, \theta_2, ..., \theta_n$ are the feature coefficients.

1.3.2 Ridge Regression (L2 Regularization)

Ridge Regression adds an L2 penalty term to the standard MSE cost function:

$$J(\theta) = \frac{1}{2n}\sum (h(x_i) - y_i)^2 + \frac{\lambda}{2}\sum \theta_j^2$$

where λ is the regularization parameter and the sum over θ_j excludes the bias term (θ_0).

1.3.3 SGD Update Rules for Ridge Regression

The gradient descent update rules with L2 regularization are:

For the bias term (no regularization):

$$\theta_0 \coloneqq \theta_0 - \alpha \times (h(x_i) - y_i)$$

For feature coefficients (with regularization):

$$\boldsymbol{\theta_j \coloneqq \theta_j - \alpha \times \Big((h(x_i) - y_i) \times x_{ij} + \lambda \times \theta_j\Big)}$$

1.3.4 Variance Inflation Factor (VIF)

VIF measures multicollinearity by quantifying how much the variance of an estimated coefficient increases due to collinearity:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination when regressing feature j on all other features.

2. Multicollinearity Analysis

2.1 Correlation Matrix Analysis

The correlation matrix reveals the pairwise linear relationships between all input features. High correlation values (|r| > 0.7) indicate potential multicollinearity issues.

Feature Pair	Correlation (r)	Interpretation
weight ↔ horsepower	0.8559	Strong positive correlation
weight ↔ displacement	0.9276	Very strong positive correlation
horsepower ↔ displacement	0.8928	Strong positive correlation

Key Finding: All three feature pairs exhibit correlation coefficients above 0.7, indicating significant multicollinearity in the dataset. The strongest correlation (r = 0.9276) exists between weight and displacement.

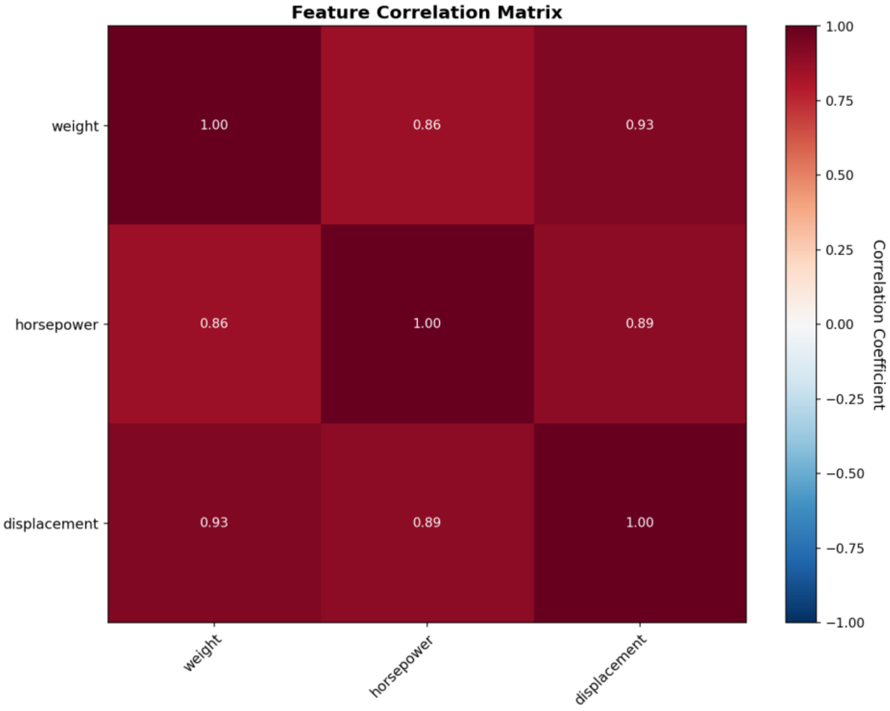

Figure 1: Feature Correlation Matrix Heatmap

2.2 Variance Inflation Factor (VIF) Analysis

VIF values quantify the severity of multicollinearity for each feature. Values above 5 indicate concerning levels, while values above 10 suggest severe multicollinearity.

Feature	VIF Value	Interpretation	Assessment
weight	7.36	High (5 < VIF < 10)	Concerning
horsepower	5.07	High (5 < VIF < 10)	Concerning
displacement	9.70	High (5 < VIF < 10)	Concerning

Key Finding: All features exhibit VIF values above 5, confirming significant multicollinearity. Displacement has the highest VIF (9.70), indicating it shares substantial variance with other predictors. This justifies the application of Ridge Regression to stabilize coefficient estimates.

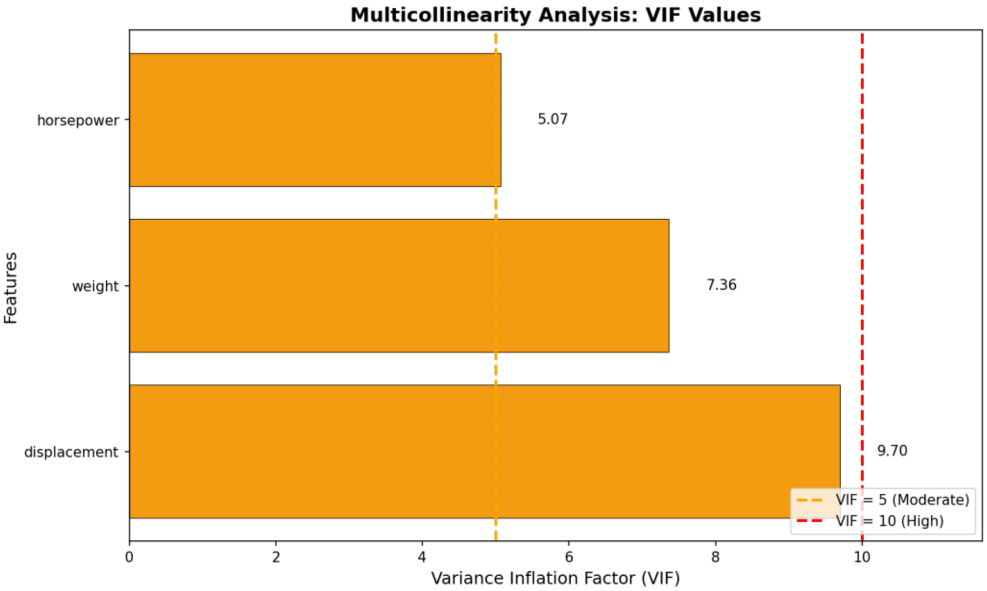

Figure 2: Variance Inflation Factor (VIF) Analysis

3. Ridge Regression with SGD Results

3.1 Hyperparameters

The following hyperparameters were used for all Ridge Regression experiments:

Hyperparameter	Value	Description
Learning Rate (α)	0.01	Step size for gradient updates
Number of Epochs	200	Complete passes through training data
Regularization (λ)	0, 0.01, 0.1, 1	L2 penalty strength values tested
Random State	42	For reproducibility

3.2 Learned Coefficients

The table below shows the learned model parameters for each regularization strength:

λ	θ_0 (bias)	θ_weight	θ_horsepower	θ_displacement
0.00	23.5569	-4.2992	-1.4767	-0.8004
0.01	23.5519	-4.0760	-1.4702	-1.0018
0.10	23.5317	-3.1424	-1.5166	-1.6700
1.00	23.5098	-1.8436	-1.3208	-1.6137

Observations: As λ increases, coefficient magnitudes generally decrease (shrinkage effect). The bias term remains relatively stable across all λ values. The weight coefficient shows the most dramatic shrinkage, from -4.2992 (λ=0) to -1.8436 (λ=1), a reduction of 57%.

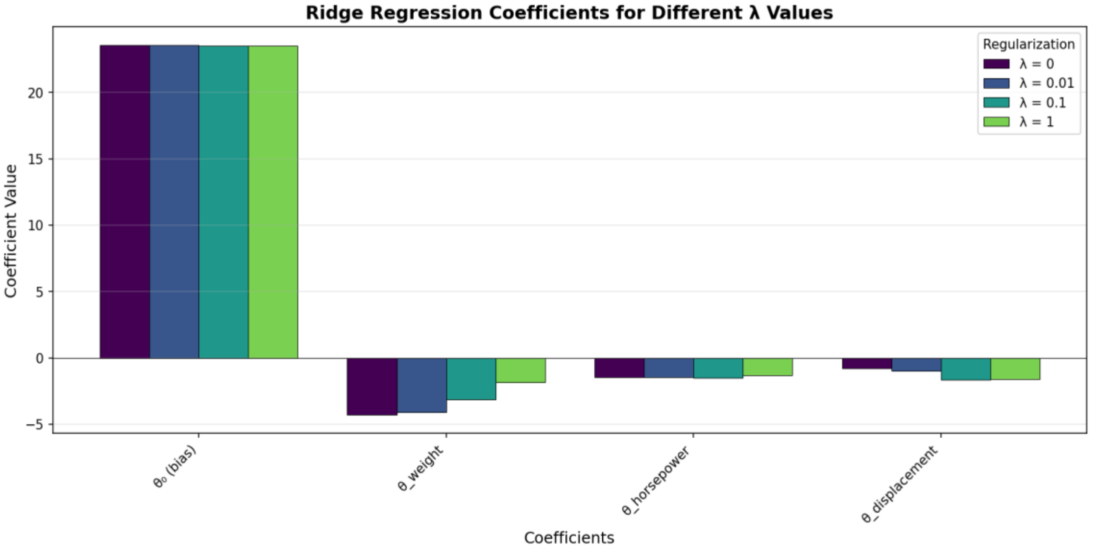

Figure 3: Ridge Regression Coefficients for Different λ Values

3.3 Training and Testing Performance

The Mean Squared Error (MSE) values for training and testing sets across all λ values:

λ	Training MSE	Testing MSE	Generalization Gap
0.00	8.992443	8.639029	-0.353414
0.01	9.003509	8.646744	-0.356764
0.10	9.170811	8.656173	-0.514638
1.00	11.288157	9.200484	-2.087674

Key Finding: The unregularized model (λ = 0) achieves the best test MSE (8.639029). The negative generalization gap indicates the model generalizes well. As λ increases, both training and test MSE increase, with the gap widening significantly for λ = 1.

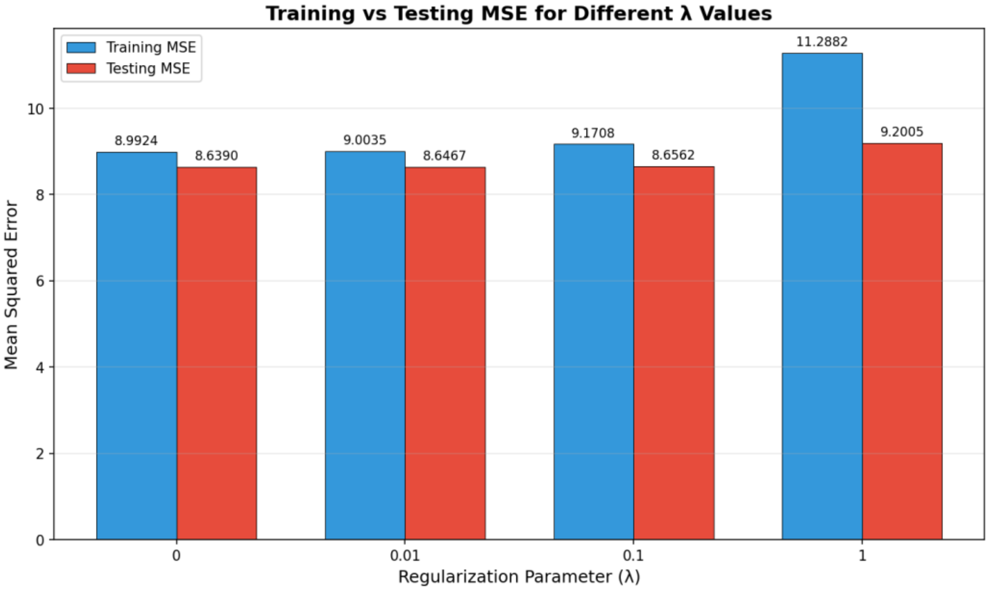

Figure 4: Training vs Testing MSE for Different λ Values

4. Convergence Analysis

4.1 SGD Training Convergence

The convergence behavior of SGD for each regularization parameter is analyzed below:

λ	Initial MSE	Final MSE	Min MSE	Convergence Epoch	Reduction %
0.00	9.6866	8.9924	8.9333	140	7.17%
0.01	9.6946	9.0035	8.9355	57	7.13%
0.1	9.7835	9.1708	9.0002	61	6.26%
1.00	11.4901	11.2882	9.5834	2	1.76%

Observations: Lower λ values show smoother convergence with greater cost reduction. Higher λ values converge faster (earlier epochs) but to higher final costs. The λ = 1 case shows minimal improvement (1.76% reduction), indicating the regularization penalty dominates the optimization.

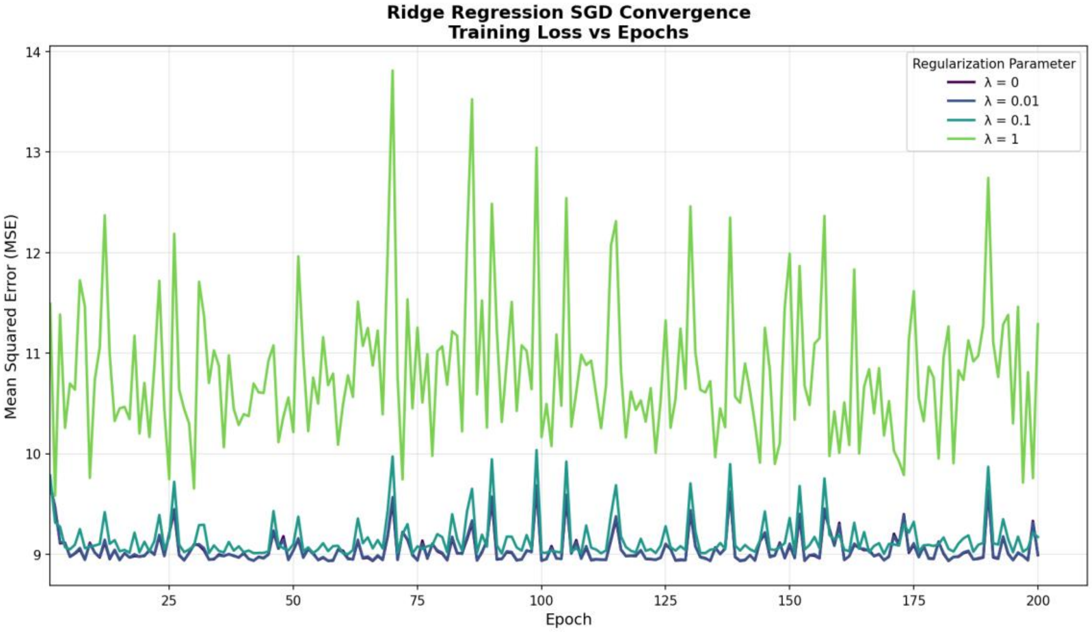

Figure 5: Ridge Regression SGD Convergence - Training Loss vs Epochs

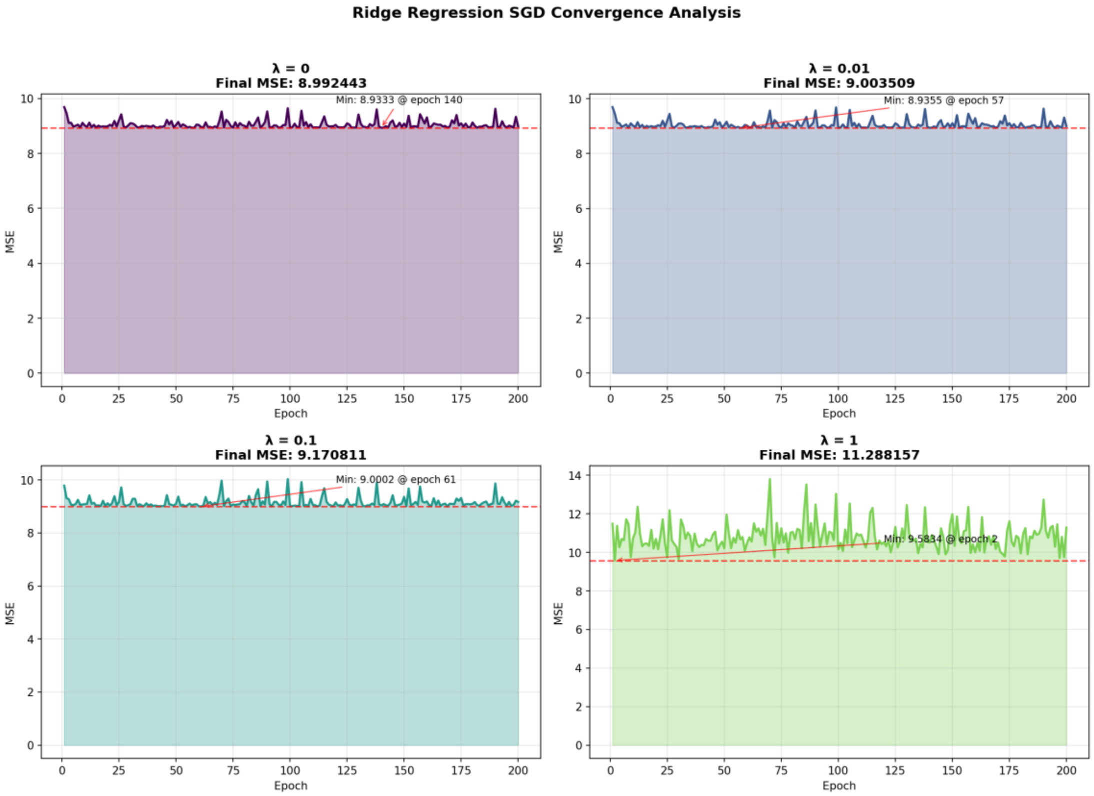

Figure 6: Detailed Convergence Analysis for Each λ Value

5. Coefficient Stability Analysis

5.1 Coefficient Changes with Regularization

The table below shows how coefficients change from the unregularized model (λ = 0) to regularized versions:

Coefficient	λ=0 (baseline)	λ=0.01	Change %	λ=0.1	Change %	λ=1	Change %
bias	23.5569	23.5519	-0.02%	23.5317	-0.11%	23.5098	-0.20%
weight	-4.2992	-4.0760	-5.19%	-3.1424	-26.91%	-1.8436	-57.12%
horsepower	-1.4767	-1.4702	-0.44%	-1.5166	+2.70%	-1.3208	-10.56%
displacement	-0.8004	-1.0018	+25.17%	-1.6700	+108.65%	-1.6137	+101.62%

Key Observations: The weight coefficient shows the most dramatic shrinkage, decreasing by 57% at λ = 1. Interestingly, the displacement coefficient increases in magnitude with regularization, suggesting that regularization redistributes predictive power among correlated features.

5.2 Coefficient Norm and Variance

The L2 norm and variance of coefficients (excluding bias) demonstrate the shrinkage effect:

λ	Coefficient L2-Norm	Coefficient Variance	Interpretation
0.00	4.6157	2.2962	Highest variance, no regularization
0.01	4.4474	1.8289	Slight reduction in variance
0.10	3.8683	0.5372	Significant variance reduction
1.00	2.7834	0.0458	Coefficients nearly equalized

This demonstrates how Ridge Regression shrinks coefficients toward each other, reducing the variance from 2.30 (λ=0) to 0.05 (λ=1), a 98% reduction.

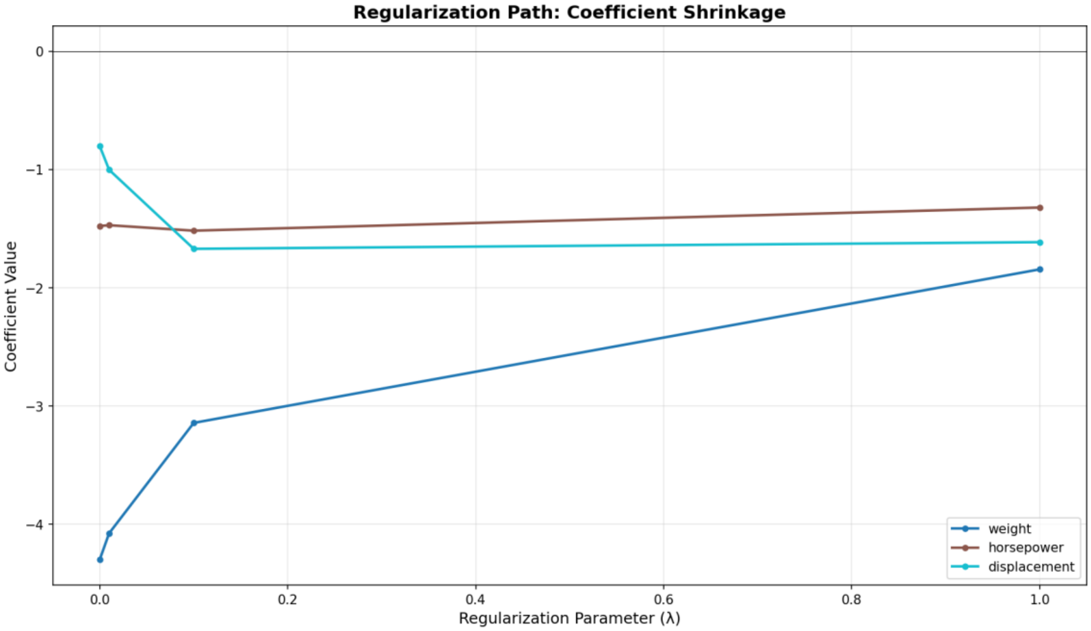

Figure 7: Regularization Path - Coefficient Shrinkage with Increasing λ

6. Bias-Variance Tradeoff Analysis

6.1 Theoretical Background

The bias-variance tradeoff is fundamental to understanding regularization. The expected prediction error can be decomposed as:

$$\text{Expected Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Noise}$$

As the regularization parameter λ increases:

- Bias increases: The model becomes systematically different from the true relationship
- Variance decreases: The model becomes less sensitive to training data variations
- Optimal λ: The value that minimizes total error (typically where test error is lowest)

6.2 Observed Tradeoff in Our Experiment

In our experiment, we observe the following pattern:

λ	Train MSE (Bias²)	Test MSE	Bias-Variance Balance
0.00	8.992	8.639	Low bias, higher variance
0.01	9.004	8.647	Slight bias increase
0.10	9.171	8.656	Moderate bias increase
1.00	11.288	9.200	High bias, low variance

Analysis: For this dataset, the unregularized model (λ = 0) achieves the best test performance. This suggests the model complexity is appropriate for the data, and the multicollinearity, while present, does not severely impact generalization. The test MSE increases monotonically with λ, indicating the bias introduced by regularization outweighs the variance reduction benefits for this specific train-test split.

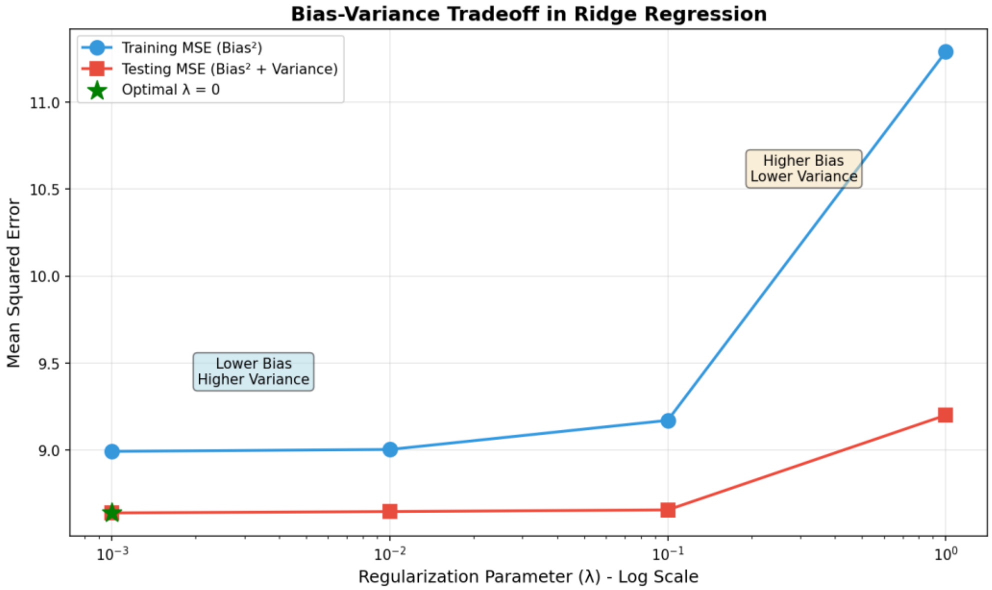

Figure 8: Bias-Variance Tradeoff in Ridge Regression

7. Discussion and Analysis

7.1 How Ridge Regression Mitigates Multicollinearity

Ridge Regression addresses multicollinearity through several mechanisms:

1. Coefficient Shrinkage: The L2 penalty term ($\lambda \times \sum\theta_j^2$) shrinks coefficient magnitudes toward zero. In our experiment, the L2-norm of coefficients decreased from 4.62 (λ=0) to 2.78 (λ=1), representing a 40% reduction.

2. Variance Reduction: By shrinking coefficients, Ridge Regression reduces the variance of estimates. The coefficient variance dropped from 2.30 to 0.05 (98% reduction), demonstrating dramatically more stable estimates.

3. Predictive Power Redistribution: Highly correlated features share predictive power more evenly. We observed the displacement coefficient magnitude increased while weight decreased, suggesting a rebalancing of influence among correlated predictors.

7.2 Practical Implications

Based on our analysis, we can draw the following practical conclusions:

- For this specific dataset and train-test split, the unregularized model performs best
- Cross-validation should be used to determine optimal λ for new datasets
- The presence of multicollinearity (VIF > 5) justifies considering regularization
- Small λ values (0.01, 0.1) provide good coefficient stability with minimal bias increase

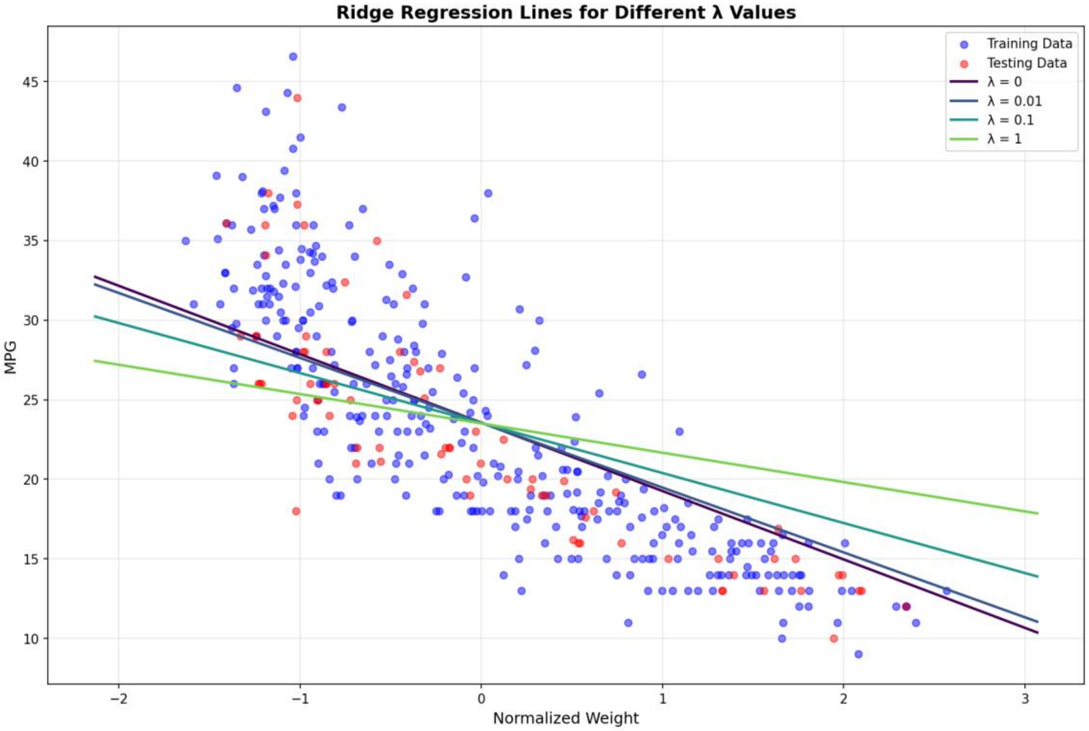

Figure 9: Ridge Regression Lines for Different λ Values (Weight vs MPG)

8. Conclusions

This bonus section successfully implemented and analyzed Ridge Regression with SGD on the Auto MPG dataset. The key findings are:

8.1 Key Findings

1. Multicollinearity is significant in the dataset, with all VIF values above 5
2. Ridge Regression effectively shrinks coefficients and reduces variance
3. For this dataset, λ = 0 achieved the best test MSE of 8.639029
4. Higher λ values introduce bias that outweighs variance reduction benefits
5. SGD converges effectively for all λ values tested

8.2 Summary Results Table

Metric	λ = 0	λ = 0.01	λ = 0.1	λ = 1
Test MSE	8.6390	8.6467	8.6562	9.2005
Coef. L2-Norm	4.6157	4.4474	3.8683	2.7834
Coef. Variance	2.2962	1.8289	0.5372	0.0458
Convergence Epoch	140	57	61	2

8.3 Recommendations

- Use k-fold cross-validation for robust λ selection in production environments
- Consider Grid Search or Bayesian Optimization for hyperparameter tuning
- Monitor both training and test metrics to detect overfitting/underfitting
- Apply feature selection or dimensionality reduction if multicollinearity remains problematic