Handling Multicollinearity in Linear Models: An Empirical Evaluation of PCA and Ridge Regression

Regression and Time Series Models
MA60280
Spring Semester 2025
IIT Kharagpur

Submitted by: Sumanth Javvaji

Abstract

Multicollinearity is a common challenge in Multiple Linear Regression (MLR) models, often leading to numerically unstable parameter estimates and reduced predictability. In this study, we explore various techniques to detect and mitigate multicollinearity within a dataset. We begin by diagnosing multicollinearity using Variance Inflation Factor (VIF), correlation analysis, and Eigen System Analysis. A baseline least squares MLR model is fitted to establish performance benchmarks.

To address multicollinearity, we implement Principal Component Analysis (PCA) and Ridge Regression. For Ridge Regression, we investigate multiple approaches for selecting the Biasing Factor, comparing their effectiveness through R^2-score reduction and the variance reduction of the regression coefficients. Finally, we provide a comparative analysis of the methods in terms of predictive performance and model stability. Our results highlight the trade-offs involved in dimensionality reduction and regularization, offering insights into effective model selection under multicollinearity.

1 Introduction

1.1 Multicollinearity

In the case of Multiple Linear Regression, a major issue is **Multicollinearity**, which makes the model unstable in nature. The model is given by:

$$\hat{y} = \hat{\beta}X \quad \text{where,} \quad \hat{\beta} = (X^T X)^{-1} X^T y$$

In the above expression, we can see that $\hat{\beta}$ depends on the inverse of the matrix $X^T X$. Let us define:

$$C = X^T X$$

Now, if this matrix C is non-invertible or singular, then the inverse $(X^T X)^{-1}$ cannot be computed. The matrix $X^T X$ is given by:

$$X^T X = \begin{bmatrix} \sum_{k=1}^{n} x_{k1}^2 & \sum_{k=1}^{n} x_{k1}x_{k2} & \cdots & \sum_{k=1}^{n} x_{k1}x_{kp} \\ \sum_{k=1}^{n} x_{k2}x_{k1} & \sum_{k=1}^{n} x_{k2}^2 & \cdots & \sum_{k=1}^{n} x_{k2}x_{kp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} x_{kp}x_{k1} & \sum_{k=1}^{n} x_{kp}x_{k2} & \cdots & \sum_{k=1}^{n} x_{kp}^2 \end{bmatrix}$$

In the presence of multicollinearity, if some rows or columns of X are linear combinations of others, then $\det(C)$ becomes 0 or very close to 0. This causes:

$$\frac{1}{\det(C)} \to \infty$$

leading to unstable calculations when estimating $\hat{\beta}$. This makes the model highly sensitive to small changes in the data, resulting in unreliable predictions.

Essentially, multicollinearity implies that there exists a non-trivial linear relationship among the regressors. If we have a vector of predictors $X = (X_1, X_2, \ldots, X_p)^T$, then there exists a set of coefficients t_j such that:

$$\sum_{j=1}^{p} t_j X_j = 0$$

1.2 Detecting Multicollinearity

To diagnose multicollinearity in a dataset, several approaches can be utilized:

- **Correlation matrix** of the predictors
- **Determinant** of $X^T X$
- **Eigenvalues** of $X^T X$
- **Variance Inflation Factor (VIF)**

In particular, eigenvalue analysis leads to the calculation of the **Condition Number** and **Condition Indices**, which help detect near-linear dependencies among the predictors:

$$\kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}, \quad \kappa_j = \frac{\lambda_{\text{max}}}{\lambda_j}$$

where λ_j is the j-th eigenvalue of $X^T X$. A condition number greater than 100 is considered indicative of moderate multicollinearity, and values exceeding 1000 indicate severe multicollinearity.

The **Variance Inflation Factor (VIF)** for each regressor is defined as:

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

where R_j^2 is the coefficient of determination obtained from regressing X_j on the remaining predictors. A high R_j^2 implies that X_j is well explained by the other variables, resulting in a high VIF_j, and consequently, a high variance for the estimated $\hat{\beta}_j$.

To address multicollinearity, one widely accepted approach is **Ridge Regression**, introduced by Hoerl and Kennard [HK70]. Ridge Regression stabilizes the estimates by adding a biasing term, but introduces the challenge of selecting an optimal penalty parameter k. Multiple strategies have been proposed for this purpose, including those by Hoerl and Kennard [HK70], McDonald and Galarneau [MG75], Lawless and Wang [LW76], among others.

In this study, we systematically analyze multicollinearity in a dataset using VIF, correlation matrices, and eigen system analysis. We first fit a baseline multiple linear regression model and then address collinearity using Principal Component Analysis (PCA) and Ridge Regression. Additionally, we evaluate several strategies for selecting the ridge penalty parameter and compare their performance using R^2-score and the variance of the estimated coefficients.

2 Handling Multicollinearity

2.1 Ridge Regression

In the presence of multicollinearity, where predictor variables are highly correlated, the variance of the ordinary least squares (OLS) estimator becomes large, making predictions unstable and sensitive to small changes in the data. Ridge regression, first introduced by Hoerl and Kennard [HK70], is a prominent remedy that introduces a small bias to obtain estimators with significantly reduced variance.

Ridge regression can be naturally derived from the following constrained optimization problem:

$$\min_{\beta} \left(\beta - \hat{\beta}\right)^\top X^\top X \left(\beta - \hat{\beta}\right) \quad \text{subject to} \quad \beta^\top \beta \leq d^2$$

This formulation seeks a vector β that remains close to the OLS estimator $\hat{\beta}$ under the geometry induced by $X^\top X$, while also constraining its Euclidean norm to lie within a ball of radius d. Constructing the Lagrangian:

$$L(\beta, \lambda) = (\beta - \hat{\beta})^\top X^\top X(\beta - \hat{\beta}) + \lambda\left(\beta^\top \beta - d^2\right),$$

and taking the gradient with respect to β yields:

$$\nabla_\beta L = 2X^\top X(\beta - \hat{\beta}) + 2\lambda\beta = 0.$$

Solving this equation gives:

$$(X^\top X + \lambda I)\beta = X^\top X\hat{\beta},$$

which implies:

$$\beta = (X^\top X + \lambda I)^{-1} X^\top X\hat{\beta}.$$

Substituting $\hat{\beta} = (X^\top X)^{-1} X^\top y$, we obtain the closed-form ridge regression estimator:

$$\hat{\beta}_R = (X^\top X + \lambda I)^{-1} X^\top y.$$

Thus, ridge regression emerges as the projection of the OLS estimate onto an ℓ_2-ball under the metric defined by $X^\top X$.

While the OLS estimator is unbiased, its variance can be large in the presence of multicollinearity. Ridge regression introduces bias to reduce variance, often leading to a lower mean squared error (MSE). The MSE of the ridge estimator is given by:

$$\text{MSE}(\hat{\beta}_R) = \text{Var}(\hat{\beta}_R) + \text{Bias}^2.$$

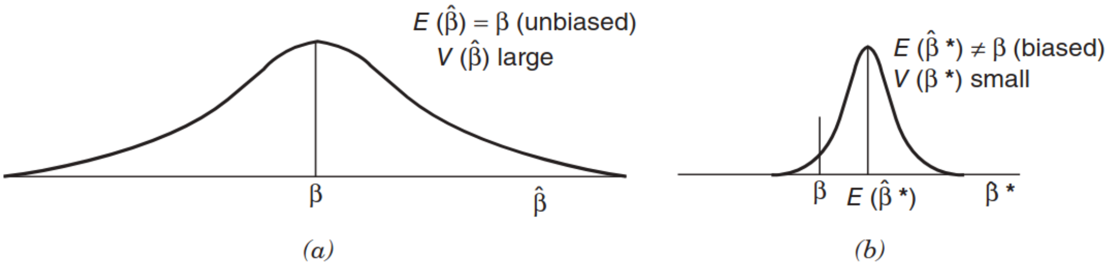

Figure 1: Sampling distribution of (a) nnbiased and (b) biased estimators of β [MPV12]

Given $\hat{\beta}_R = Z\hat{\beta}$, where $Z = (X^\top X + \lambda I)^{-1} X^\top X$, we have:

$$E[\hat{\beta}_R] = Z\beta, \quad \text{and} \quad \text{Var}(\hat{\beta}_R) = \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}.$$

In comparison, the MSE of the OLS estimator is:

$$\mathrm{MSE}(\hat{\beta}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j},$$

where λ_j are the eigenvalues of $X^\top X$. For the ridge estimator, the MSE becomes:

$$\mathrm{MSE}(\hat{\beta}_R) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + \lambda)^2} + \lambda^2 \hat{\beta}^\top (X^\top X + \lambda I)^{-2} \hat{\beta}.$$

This expression highlights the bias-variance trade-off: the first term, corresponding to variance, decreases with increasing λ, while the second term, the squared bias, increases with λ. The optimal choice of λ thus balances these two effects to minimize the total MSE.

Furthermore, the residual sum of squares for ridge regression can be decomposed as:

$$SSE(\hat{\beta}_R) = (y - X\hat{\beta}_R)^\top (y - X\hat{\beta}_R) = (y - X\hat{\beta})^\top (y - X\hat{\beta}) + (\hat{\beta}_R - \hat{\beta})^\top X^\top X (\hat{\beta}_R - \hat{\beta}),$$

where the second term quantifies the additional error due to regularization. This further reinforces the interpretation of ridge regression as a compromise between data fidelity and model complexity.

2.2 Principal Component Analysis

The principal-component analysis (PCA) approach is designed to handle multicollinearity in regression models by reducing the dimensionality of the predictor space. The main idea is to transform the original regressors into a new set of orthogonal variables, known as the principal components. These components are derived from the eigenvalues and eigenvectors of the matrix $X^\top X$.

The model is given by:

$$y = X\beta + \epsilon$$

where y is the response vector, X is the matrix of regressors, β is the vector of regression coefficients, and ϵ is the error term.

To address multicollinearity, PCA uses the eigenvalue decomposition of $X^\top X$. Suppose the eigenvalues of $X^\top X$ are ordered as:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$$

If the last s eigenvalues are near zero, these correspond to the components that are almost linearly dependent and contribute little to the explanation of variance in the data. Thus, PCA removes these components from the analysis and applies least squares to the remaining components.

$$\hat{a} = (Z^\top Z)^{-1} Z^\top y$$

The principal-component estimator $\hat{\alpha}_{\mathrm{PC}}$ can be written as:

$$\hat{\alpha}_{\mathrm{PC}} = B\hat{\alpha}$$

where B is a vector with components $b_1 = b_2 = \cdots = b_{p-s} = 1$ and $b_{p-s+1} = b_{p-s+2} = \cdots = b_p = 0$. Thus, the principal-component estimator is:

$$\hat{\alpha}_{\mathrm{PC}} = [\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{p-s}, 0, 0, \ldots, 0]^T$$

where the first $p - s$ components are the non-zero coefficients, and the remaining s components are set to zero. This effectively removes the influence of components associated with near-zero eigenvalues.

The regression model prediction is given by:

$$\hat{y} = X\hat{\beta}_{\mathrm{PC}} \quad \text{where,} \quad \hat{\beta}_{\mathrm{PC}} = T\hat{\alpha}_{\mathrm{PC}}$$

This equation shows how the modified regression coefficients $\hat{\beta}_{\mathrm{PC}}$ are used to predict the response variable.

The covariance matrix of the regression coefficients is:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 T^\top T$$

where T is the matrix of eigenvectors, and the variance of each regression coefficient is inversely proportional to the corresponding eigenvalue. Therefore, small eigenvalues increase the variance of the corresponding coefficients, which is why removing these components helps in improving the precision of the regression estimates.

3 Methodology

3.1 Overview of Approach

The primary objective of this study is to analyze and mitigate the effects of multicollinearity in linear regression models. We begin by diagnosing multicollinearity using three diagnostic tools: the **Variance Inflation Factor (VIF)**, **correlation heatmap**, and **eigensystem analysis** of the matrix X^TX. Based on the severity of multicollinearity observed, we then apply **Ridge Regression** and **Principal Component Analysis (PCA)** to stabilize coefficient estimates and reduce redundancy among predictors. Finally, we evaluate and compare the models based on their performance and coefficient stability.

3.2 Data Preprocessing

All predictor variables were **standardized** to have zero mean and unit variance. This step ensures comparability across features and is essential for both Ridge Regression and PCA, which are sensitive to feature scaling. No train-test split or cross-validation was employed, as the focus of the study was not on generalization performance but rather on **minimizing coefficient variance** and improving **model stability under multicollinearity**.

3.3 Diagnosing Multicollinearity

To identify the presence and extent of multicollinearity, we applied the following diagnostic techniques:

- **Correlation Matrix**: Computed and visualized as a heatmap to detect strong linear associations between predictors.
- **Variance Inflation Factor (VIF)**: Calculated for each predictor to measure the degree of multicollinearity.
- **Eigenvalue Analysis of X^TX**: Used to compute **condition numbers** and **condition indices**, which indicate the sensitivity of the solution to numerical instability. A large condition number suggests high multicollinearity.

3.4 Ridge Regression

Ridge Regression was applied to stabilize coefficient estimates by shrinking them toward zero. The estimator used is given by:

$$\hat{\beta}_R = (X^TX + \lambda I)^{-1}X^Ty$$

Unlike traditional approaches that use cross-validation to determine the optimal penalty parameter λ, we evaluated a **range of λ values** and observed their impact on coefficient variance and model performance. This direct comparison allows us to study the trade-off between bias and variance induced by regularization.

3.5 Principal Component Regression

Principal Component Analysis (PCA) was used to transform the predictor matrix into a set of orthogonal components derived from the eigenvectors of $X^T X$. Components associated with small eigenvalues—typically indicative of redundancy and instability—were discarded.

Regression was then performed using the retained principal components, and predictions were mapped back to the original feature space using the corresponding transformations.

3.6 Selection of λ for Ridge Regression

To further explore the effect of regularization in Ridge Regression, we considered multiple strategies for selecting the biasing parameter λ. These include:

- **Fixed λ values** for comparison against the baseline MLR and PCA models.
- **Heuristic-based methods** from literature, such as those proposed by Hoerl and Kennard, McDonald and Galarneau, and Lawless and Wang.

This allowed us to assess how different choices of λ influenced model stability and predictive power under high multicollinearity.

3.7 Evaluation Strategy

The effectiveness of each approach was evaluated using the following metrics:

- **R^2-score**: To measure the proportion of variance in the response variable explained by the model.
- **Trace of the Dispersion Matrix** ($\mathrm{tr}(\mathrm{Var}(\hat{\beta}))$): To quantify the overall variability of the coefficient estimates.
- **Frobenius Norm of the Dispersion Matrix**: To measure the overall size or magnitude of the coefficient variance in matrix form.

These metrics provide a comprehensive view of both model fit and coefficient stability, enabling a well-rounded comparison between methods.

4 Experimental Evaluation

4.1 Dataset Description

The dataset consists of atmospheric and chemical sensor measurements aimed at predicting **Absolute Humidity (AH)**. The predictor variables include several chemical gas concentrations, sensor responses, and a meteorological variable. A summary of the key features, their means, and variances before standardization is provided below:

Table 1: Mean and Variance of Each Feature (Before Standardization)

Feature	**Mean** ($\mathbb{E}[X]$)	**Variance** ($\mathrm{Var}[X]$)
CO(GT)	2.354	1.987
PT08.S1(CO)	1207.742	58480.178
NMHC(GT)	231.025	43456.369
C6H6(GT)	10.772	55.014
PT08.S2(NMHC)	965.984	70975.960
NOx(GT)	143.502	6696.103
PT08.S3(NOx)	963.178	70706.082
NO2(GT)	100.260	991.861
PT08.S4(NO2)	1600.507	91379.266
PT08.S5(O3)	1045.691	160104.239
T	15.600	23.286
AH (Target)	0.832	0.0319

- **Chemical gas concentrations**: `CO(GT)`, `NOx(GT)`, `NO2(GT)`, `C6H6(GT)`, `NMHC(GT)`
- **Sensor responses**: `PT08.S1(CO)`, `PT08.S2(NMHC)`, `PT08.S3(NOx)`, `PT08.S4(NO2)`, `PT08.S5(O3)`
- **Meteorological variable**: `T` (Temperature)

A distributional analysis was carried out for all features, including the target. Several predictors exhibited skewed distributions. These observations informed our choice to standardize all predictor variables prior to modeling.

The dataset did not contain any categorical variables, and no missing value imputation was necessary. All predictors were scaled to zero mean and unit variance to ensure compatibility with the assumptions of both Ridge Regression and Principal Component Analysis.

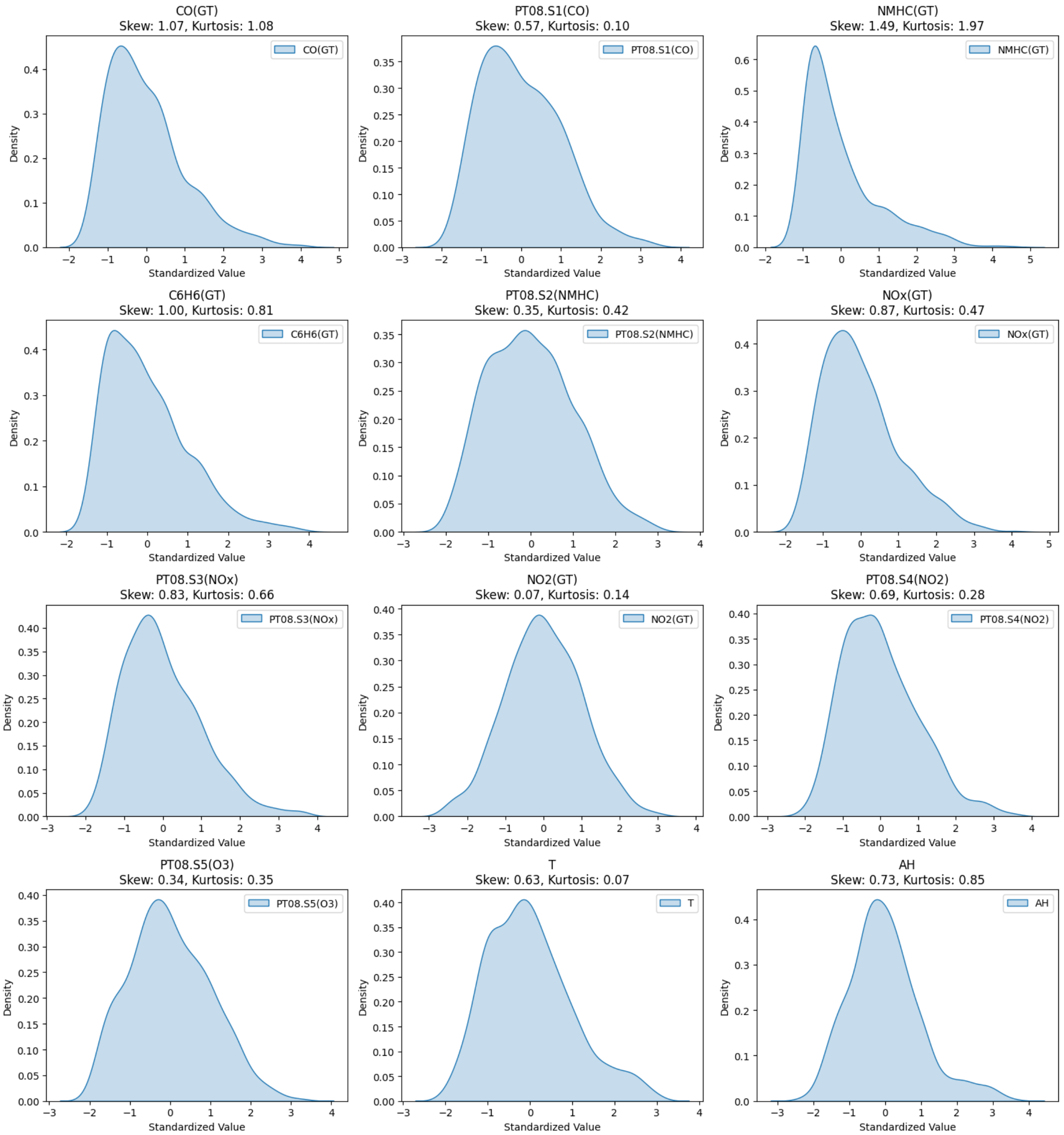

Figure 2: Distributions of the features

4.2 Multicollinearity Diagnosis

To ensure the robustness and interpretability of our regression model, it is crucial to assess multicollinearity among the predictor variables. Multicollinearity occurs when two or more predictors are highly correlated, leading to unstable coefficient estimates and inflated standard errors. We employed three complementary methods to detect multicollinearity: **Variance Inflation Factor (VIF)**, **correlation matrix inspection**, and **condition index analysis**.

4.2.1 Variance Inflation Factor (VIF)

The Variance Inflation Factor quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value exceeding 10 is typically considered a threshold beyond which multicollinearity may distort model estimates. The computed VIF values for all predictors are summarized in Table 2.

Table 2: VIF values for predictor variables

Feature	**VIF**
CO(GT)	39.12
PT08.S1(CO)	22.01
NMHC(GT)	7.12
C6H6(GT)	155.94
PT08.S2(NMHC)	189.88
NOx(GT)	15.50
PT08.S3(NOx)	19.74
NO2(GT)	7.78
PT08.S4(NO2)	71.34
PT08.S5(O3)	11.55
T	2.31
AH	5.42

Significantly high VIF values are observed for *PT08.S2(NMHC)*, *C6H6(GT)*, and *PT08.S4(NO2)*, indicating strong linear dependencies with other predictors. These inflated values suggest potential instability in the model's coefficient estimates if these features are retained without proper regularization or dimensionality reduction.

4.2.2 Correlation Matrix Inspection

To gain insight into the pairwise relationships among features, we computed the Pearson correlation coefficients. Several predictor pairs exhibit extremely high correlations ($|r| > 0.9$), suggesting the presence of multicollinearity. Key observations include:

- *PT08.S2(NMHC)* and *C6H6(GT)*: $r = 0.98$
- *CO(GT)* and *C6H6(GT)*: $r = 0.97$
- *PT08.S1(CO)* and *PT08.S4(NO2)*: $r = 0.95$
- *PT08.S2(NMHC)* and *PT08.S4(NO2)*: $r = 0.96$
- *PT08.S5(O3)* and *PT08.S4(NO2)*: $r = 0.92$
- Strong negative correlation: *PT08.S2(NMHC)* and *PT08.S3(NOx)*: $r = -0.91$

These high correlations reaffirm the findings from the VIF analysis and point to a network of strongly interdependent features, particularly among the gas concentration values and their corresponding sensor readings.

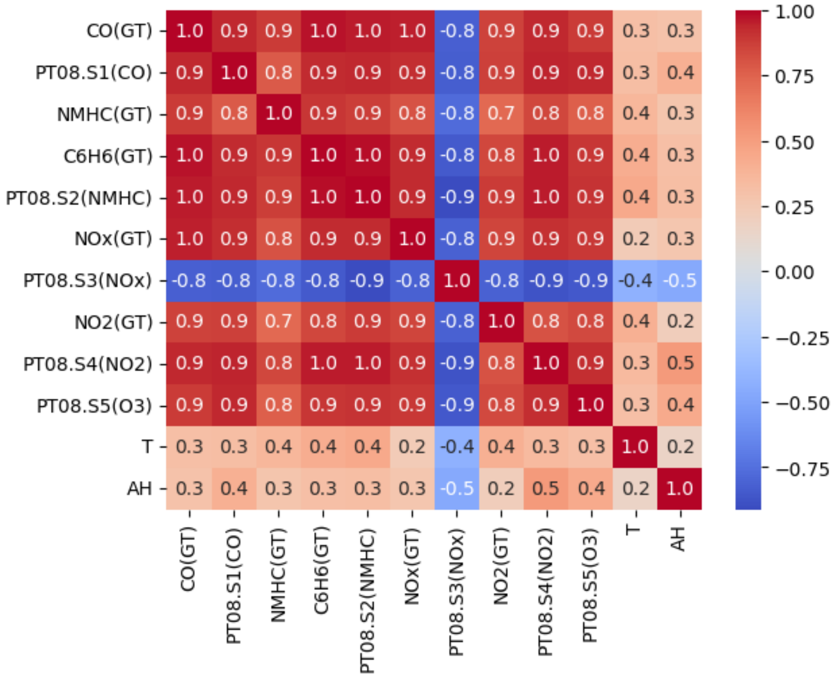

Figure 3: Correlation Heatmap

4.2.3 Condition Index and Eigenvalue Decomposition

Finally, we performed a condition number analysis to evaluate the collinearity structure across all features. The condition number of the design matrix was found to be **2829.77**, far exceeding the threshold of 100, suggesting severe multicollinearity.

Table 3: Condition indices of predictor variables

Feature	**Condition Index**
CO(GT)	1.00
PT08.S1(CO)	10.16
NMHC(GT)	27.50
C6H6(GT)	40.94
PT08.S2(NMHC)	48.74
NOx(GT)	97.66
PT08.S3(NOx)	138.22
NO2(GT)	2829.77
PT08.S4(NO2)	460.55
PT08.S5(O3)	335.09
T	237.46

Condition indices greater than 30 generally suggest moderate to strong multicollinearity, while values above 100 indicate serious issues. Here, multiple features (e.g., *NO2(GT)*, *PT08.S4(NO2)*, *PT08.S5(O3)*) have indices that vastly exceed this limit, confirming the presence of severe multicollinearity.

4.3 Baseline Linear Regression Model

The general form of the Multiple Linear Regression (MLR) model is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Where:

- y is the dependent variable.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the independent variables $x_1, x_2, \ldots, x_p$.
- ϵ is the error term.

For this model, we performed Ordinary Least Squares (OLS) regression, which estimates the coefficients that minimize the sum of squared residuals. The estimated coefficients for the model are as follows:

Feature	**Beta**
CO(GT)	0.19496052
PT08.S1(CO)	0.12548033
NMHC(GT)	0.09499804
C6H6(GT)	−0.86397065
PT08.S2(NMHC)	−2.31300323
NOx(GT)	−0.41590781
PT08.S3(NOx)	−0.56178816
NO2(GT)	0.07160975
PT08.S4(NO2)	2.78262635
PT08.S5(O3)	0.15278455
T	0.2417588

4.3.1 Model Performance Metrics

R² Score

The R^2 score, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that can be explained by the independent variables in the model. In this case, the R^2 score is 0.8156, which indicates that approximately 81.56% of the variance in the dependent variable is explained by the model.

ANOVA Table

The Analysis of Variance (ANOVA) table provides information about the sources of variance in the model. It helps us understand whether the model as a whole is statistically significant. Below is the ANOVA table for the regression:

Source	**SS**	**DF**	**MS**	**F**
Regression	673.6550	10	67.3655	360.8273
Residual	152.3450	816	0.1867	
Total	826.0000	826		

Where:

- SS is the sum of squares,
- DF is the degrees of freedom,
- MS is the mean square (SS/DF),
- F is the F-statistic for testing the overall significance of the model.

The high F-statistic (360.8273) and the small p-value associated with it suggest that the regression model is statistically significant.

Frobenius Norm

The Frobenius norm of the residual matrix is a measure of the magnitude of the residuals. It is calculated as:

$$\|\mathbf{E}\|_F = \sqrt{\sum_{i,j} e_{ij}^2}$$

Where e_{ij} are the elements of the residual matrix. The Frobenius norm for this model is:

$$0.38502272529883347$$

This value suggests the residuals are of relatively small magnitude, indicating a good fit of the model to the data.

Trace of the Variance-Covariance Matrix

The trace of the variance-covariance matrix of the estimated coefficients provides a summary of the variance of the model parameters. It is given by the sum of the diagonal elements of the variance-covariance matrix, which represent the variances of each coefficient. The trace for this model is:

$$0.5608176808744026$$

This value suggests that the overall variability in the coefficients is low, contributing to a stable model with relatively low variance.

4.3.2 Observations on Instability and High Variance in Coefficients

While the model appears to explain a significant amount of variance in the dependent variable ($\mathrm{R}^2 = 0.8156$), it is important to note that some of the coefficients have relatively high values in magnitude, such as for 'PT08.S2(NMHC)' (-2.31300323) and 'C6H6(GT)' (-0.86397065). These large magnitudes could be indicative of potential instability or high variance in the model, which might arise from multicollinearity. In particular, the high correlation between some of the predictors in the dataset (as seen in the correlation matrix) could result in inflated standard errors for the coefficients, making them less reliable.

However, the low Frobenius norm and the relatively low trace of the variance-covariance matrix suggest that the model's overall stability is still quite good.

4.4 Ridge Regression Analysis

Ridge Regression is an extension of Multiple Linear Regression (MLR) that includes a regularization term controlled by λ. In this analysis, we used $\lambda = 0.7588693182551097$ to balance the penalty on the coefficients and the model's fit. The Ridge regression model provides the following results:

Beta Coefficients:

The estimated coefficients for the Ridge regression model are:

Feature	Beta
CO(GT)	0.1785706
PT08.S1(CO)	0.13520787
NMHC(GT)	0.09788709
C6H6(GT)	−0.93203297
PT08.S2(NMHC)	−2.12722245
NOx(GT)	−0.41855228
PT08.S3(NOx)	−0.5293408
NO2(GT)	0.04385014
PT08.S4(NO2)	2.71683455
PT08.S5(O3)	0.16474356
T	0.23275705

4.4.1 Model Performance Metrics

R² Score

For the Ridge regression model, the R^2 score is 0.8151, indicating that approximately 81.51% of the variance in the dependent variable is explained by the model. This is nearly identical to the R^2 score from the Linear Regression model, reflecting a minor reduction in explanatory power due to the regularization.

Frobenius Norm

The Frobenius norm of the residual matrix for the Ridge regression model is:

$$0.2407633635118524$$

This is smaller than the Frobenius norm in the MLR model, indicating a better fit with smaller residuals.

Trace of the Variance-Covariance Matrix

The trace of the variance-covariance matrix for the Ridge regression model is:

$$0.4021659091544845$$

This value is lower compared to the MLR model, suggesting reduced variability in the estimated coefficients due to regularization.

4.4.2 Observations on Model Stability

The Ridge regression model shows improved stability compared to the standard MLR model. The regularization, controlled by λ, has reduced the magnitude of the coefficients, leading to a lower trace and smaller Frobenius norm. While there is a slight decrease in the R^2 score, this reduction is minimal, indicating that the regularization has not significantly affected the model's explanatory power.

The reduction in variance and the stability in the coefficients reflect the effectiveness of Ridge regression in controlling model complexity. Although a small increase in bias is introduced due to coefficient shrinkage, the overall stability of the model improves, leading to better generalization to unseen data.

4.5 Principal Component Analysis

Principal Component Regression (PCR) combines Principal Component Analysis (PCA) with linear regression to handle multicollinearity by reducing dimensionality. By transforming the original features into uncorrelated principal components, PCR mitigates overfitting risks and improves model stability. This approach helps when dealing with high-dimensional data, making the model more interpretable.

4.5.1 Regression Coefficients

In PCR, the regression coefficients (β_{pc}) correspond to the principal components. These are mapped back to the original features to show their contribution to the response variable. PCR reduces multicollinearity and improves model stability. Below are the regression coefficients (β_{pc}) for the actual regressors (original features):

Feature	β_{pc}
CO(GT)	0.1786
PT08.S1(CO)	0.1352
NMHC(GT)	0.0979
C6H6(GT)	−0.9320
PT08.S2(NMHC)	−2.1272
NOx(GT)	−0.4186
PT08.S3(NOx)	−0.5293
NO2(GT)	0.0439
PT08.S4(NO2)	2.7168
PT08.S5(O3)	0.1647
T	0.2328

4.5.2 Principal Components

The first 10 principal components are selected based on a threshold eigenvalue/sum(eigenvalue) of 1×10^{-3}, as shown below (head of principal components):

Component	PC0	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Row 0	-0.603	0.619	-0.618	-0.410	0.472	0.142	-0.174	0.163	-0.113	0.036	-0.106
Row 1	0.850	0.421	-0.228	-0.335	0.555	0.087	-0.536	0.170	-0.238	-0.123	-0.020
Row 2	0.342	0.817	-0.828	-0.653	0.317	0.205	-0.557	0.085	-0.219	-0.104	0.226

Table 4: Principal Component Scores

4.5.3 Model Performance Metrics

R² Score

The R^2 score for the PCR model is 0.8136, indicating that 81.36% of the variance in the dependent variable is explained by the model.

4.5.4 Comparison with Ridge Regression and Baseline MLR

The PCR model performs similarly to Ridge Regression, with an $\mathbf{R^2}$ score of 0.8136. However, PCR reduces multicollinearity by transforming the features into uncorrelated components, whereas Ridge Regression penalizes the coefficients. PCR offers better coefficient stability compared to the baseline MLR model, making it an effective alternative for high-dimensional data.

5 Selection of λ for Ridge Regression

One of the central challenges in ridge regression is the appropriate selection of the biasing parameter, λ (also denoted as k). While traditional approaches such as ridge trace plots rely on visual inspection and subjective judgment, analytical techniques aim to determine λ in a principled and data-driven manner. In this section, we evaluate two influential analytical methods proposed in 1975 by Hoerl, Kennard, and Baldwin; and by McDonald and Galarneau.

5.1 Hoerl, Kennard, and Baldwin Method

Hoerl et al. suggested choosing λ according to the expression:

$$\lambda = \frac{p \cdot \sigma^2}{\boldsymbol{\beta}^\top \boldsymbol{\beta}}$$

where:

- p is the number of predictors,
- $\boldsymbol{\beta}$ and σ^2 are derived from the ordinary least squares (OLS) solution.

This formula balances the trade-off between variance and bias using the ratio of noise variance to signal strength. It offers a closed-form estimation of λ, making it computationally efficient.

Empirical Results:

- Biasing Factor (λ): 0.7589
- R^2 Score: 0.8151
- Frobenius Norm ($\|\boldsymbol{\beta}_\lambda\|_F$): 0.2408
- Trace of Dispersion Matrix: 0.4022

5.2 McDonald and Galarneau Method

This method proposes selecting λ such that:

$$\boldsymbol{\beta}_\lambda^\top \boldsymbol{\beta}_\lambda = \boldsymbol{\beta}^\top \boldsymbol{\beta} - \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

where:

- $\boldsymbol{\beta}_\lambda$ is the ridge estimator,
- λ_j are the eigenvalues of $\mathbf{X}^\top \mathbf{X}$,
- σ^2 is the OLS estimate of noise variance.

As this equation does not permit a closed-form solution for λ, we solve it numerically using the Newton-Raphson method.

Newton-Raphson for Optimal λ

Let

$$f(\lambda) = \boldsymbol{\beta}_\lambda^\top \boldsymbol{\beta}_\lambda - \left(\boldsymbol{\beta}^\top \boldsymbol{\beta} - \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \right)$$

Then the update rule is:

$$\lambda^{(t+1)} = \lambda^{(t)} - \frac{f(\lambda^{(t)})}{f'(\lambda^{(t)})}$$

The iteration continues until convergence, i.e., when:

$$|\lambda^{(t+1)} - \lambda^{(t)}| < \epsilon$$

for some small threshold ϵ (here, 10^{-6}).
This method ensures a data-driven selection of the biasing parameter λ by directly minimizing the difference in coefficient norms based on the model's shrinkage behavior.

Empirical Results:

- Biasing Factor (λ): 0.3617
- R^2 Score: 0.8154
- Frobenius Norm: 0.3024
- Trace of Dispersion Matrix: 0.4712

5.3 Comparison and Insights

Metric	Hoerl et al.	McDonald & Galarneau
Biasing Factor (λ)	0.7589	0.3617
R^2 Score	0.8151	0.8154
Frobenius Norm	0.2408	0.3024
Trace of Dispersion Matrix	0.4022	0.4712

Table 5: Comparison of analytical methods for selecting λ

Both methods provide comparable R^2 scores, suggesting similar predictive power. However, Hoerl et al.'s method results in a smaller trace and Frobenius norm, indicating potentially better model stability and smaller deviation from the OLS coefficients. McDonald and Galarneau's method yields a smaller λ, leading to less shrinkage and coefficients closer to the OLS solution, albeit with slightly higher variance.

These findings highlight the trade-offs between bias and variance inherent in ridge regression and reinforce the importance of context-specific tuning when selecting λ.

6 Results

To evaluate the effectiveness of ridge regression using analytically chosen biasing parameters, we compare the performance of the two proposed methods—Hoerl et al. and McDonald & Galarneau with the baseline Ordinary Least Squares (OLS) solution. The comparison is made across three key metrics: the coefficient of determination (R^2 score), the Frobenius norm between ridge and OLS coefficients, and the trace of the dispersion (covariance) matrix of the ridge estimator.

Quantitative Comparison

Method	λ	R^2 Score	Frobenius Norm	Trace of Dispersion Matrix
OLS	0	0.8156	0.3850	0.5608
Ridge (Hoerl et al.)	0.7589	0.8151	0.2408	0.4022
Ridge (McDonald & Galarneau)	0.3617	0.8154	0.3024	0.4712

Table 6: Comparison of OLS and Ridge Regression using analytically selected λ values

From the results above, it is evident that both ridge regression methods achieve substantial variance reduction at the cost of a marginal drop in R^2 performance. The Frobenius norm between ridge and OLS coefficients is lowest for the Hoerl method, indicating a stronger shrinkage effect. However, the McDonald & Galarneau method achieves a slightly better R^2 score while still offering significant variance control compared to OLS.

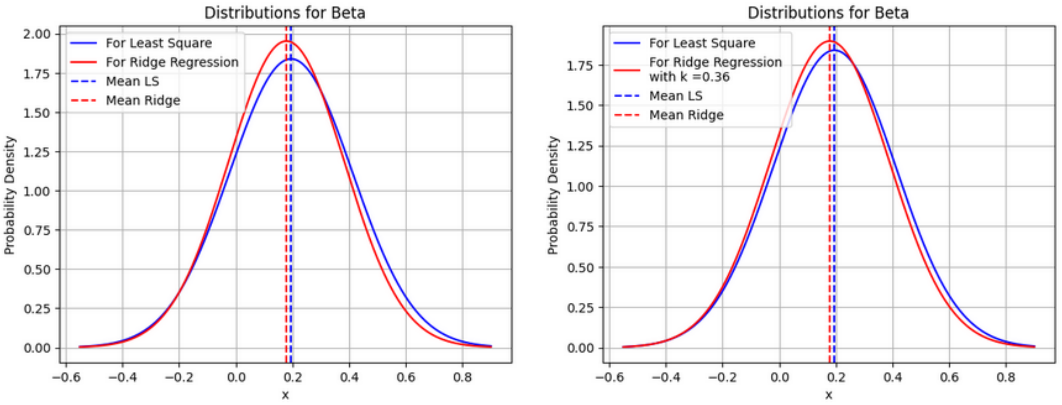

Figure 4: Distribution of β_1 for Hoerl et al. and McDonald & Galarneau

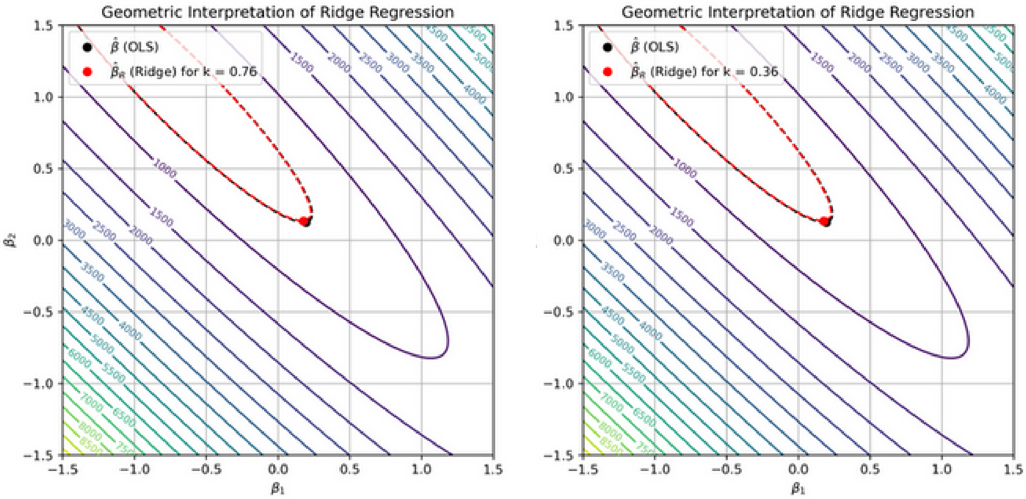

Figure 5: Geometric Interpretation of Ridge Regression

These findings underscore the core benefit of ridge regression: reducing estimator variance and improving model stability without sacrificing much predictive accuracy.

7 Conclusion

In this study, we investigated the challenges posed by multicollinearity in multiple linear regression and evaluated two popular methods—Ridge Regression and Principal Component Analysis (PCA)—to mitigate its effects. Our analysis revealed that multicollinearity was indeed severe in the dataset, as indicated by extremely high VIF values, strong pairwise correlations, and large condition indices.

Ridge Regression proved effective in stabilizing coefficient estimates by introducing regularization, leading to a significant reduction in coefficient variance without compromising model accuracy. The analytical methods used to select the biasing parameter λ—namely, those proposed by Hoerl et al. and McDonald & Galarneau—offered a principled approach to tuning and demonstrated distinct trade-offs between bias and variance.

Principal Component Regression, on the other hand, addressed multicollinearity by transforming the predictors into orthogonal components. Although its R^2 score was slightly lower than that of the Ridge models, PCR provided comparable coefficient stability and eliminated redundant information effectively.

Overall, both Ridge Regression and PCA emerged as viable strategies for handling multicollinearity, each with its strengths. Ridge Regression is particularly useful when interpretability of original features is desired, whereas PCA is better suited for scenarios requiring dimension reduction and noise elimination. Our findings highlight the importance of diagnosing multicollinearity and adopting tailored solutions to enhance model robustness and reliability.

References

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[LW76] J. F. Lawless and P. Wang. A simulation study of ridge and principal components regression. *Communications in Statistics - Simulation and Computation*, 5(4):307–323, 1976.

[MG75] G. C. McDonald and D. I. Galarneau. A monte carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350):407–416, 1975.

[MPV12] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, Hoboken, New Jersey, 5th edition, 2012.