Multiple Regression Analysis: Model Building and Diagnostics

Computational Statistics

November 24, 2025

1 Introduction

Multiple regression analysis models the relationship between a dependent variable and multiple independent variables. This document covers ordinary least squares (OLS) estimation, hypothesis testing for regression coefficients, model diagnostics for checking assumptions, detection and handling of multicollinearity, influential observations analysis, and model selection techniques. We implement comprehensive diagnostic tools to assess residual normality, homoscedasticity, linearity, and independence assumptions.

2 Mathematical Framework

2.1 Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

Matrix notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

2.2 OLS Estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3}$$

2.3 Coefficient of Determination

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{4}$$

Adjusted R^2:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{5}$$

2.4 Variance Inflation Factor

$$VIF_j = \frac{1}{1 - R_j^2} \tag{6}$$

3 Environment Setup

4 Multiple Regression Implementation

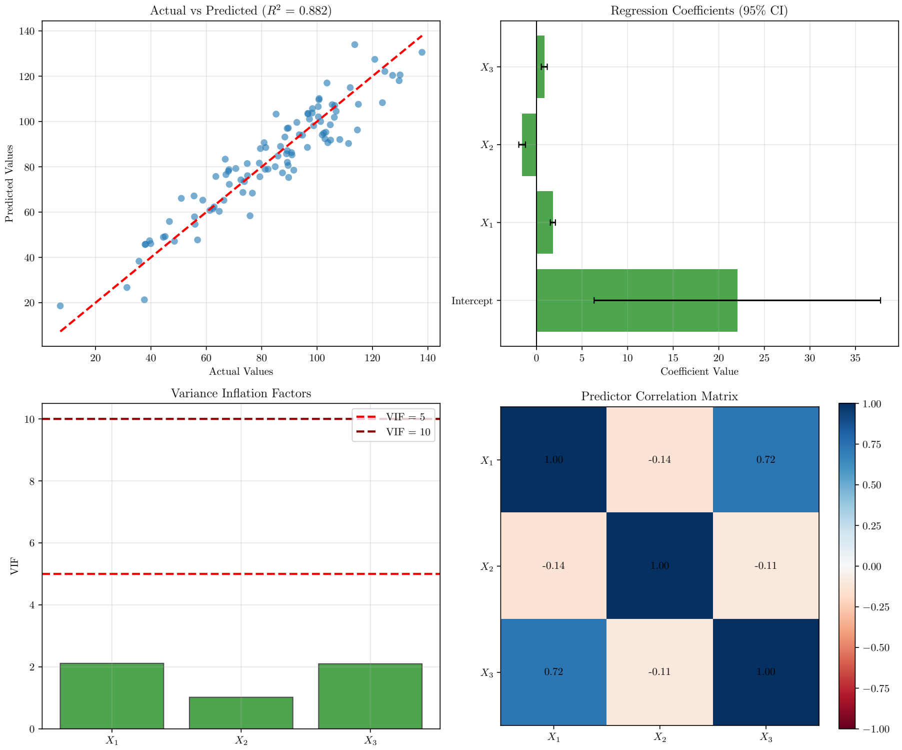

Figure 1: Multiple regression model fit with coefficient estimates and multicollinearity diagnostics.

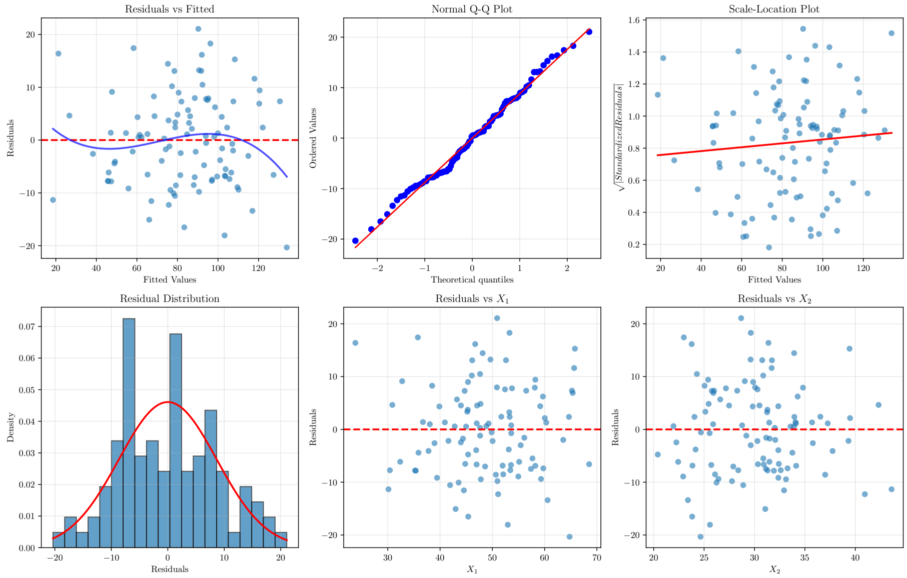

Figure 2: Residual diagnostic plots for checking regression assumptions.

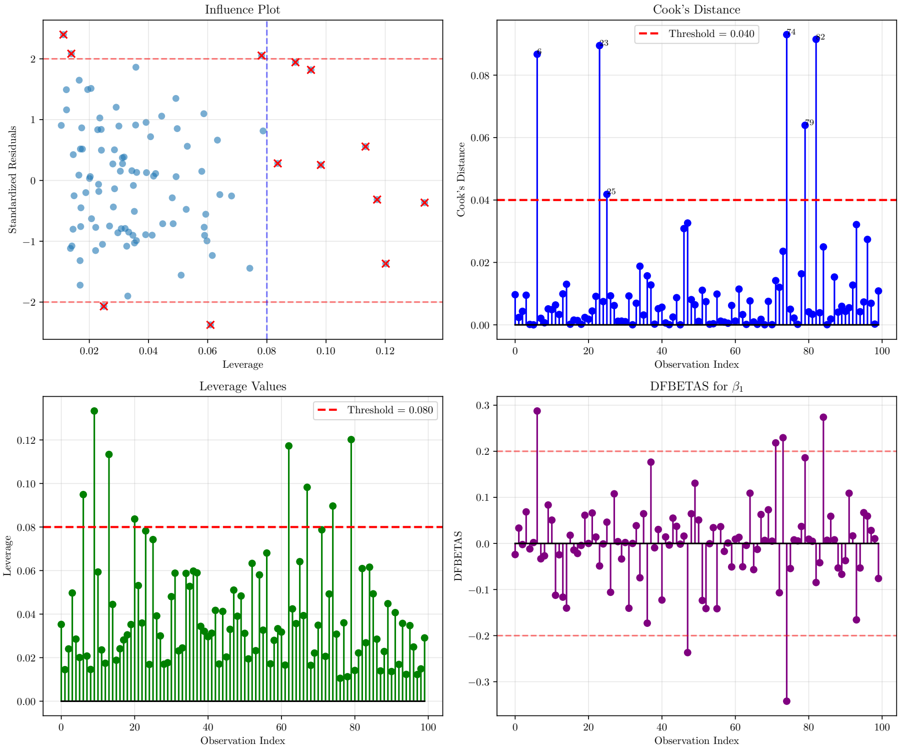

Figure 3: Influential observation diagnostics including leverage and Cook's distance.

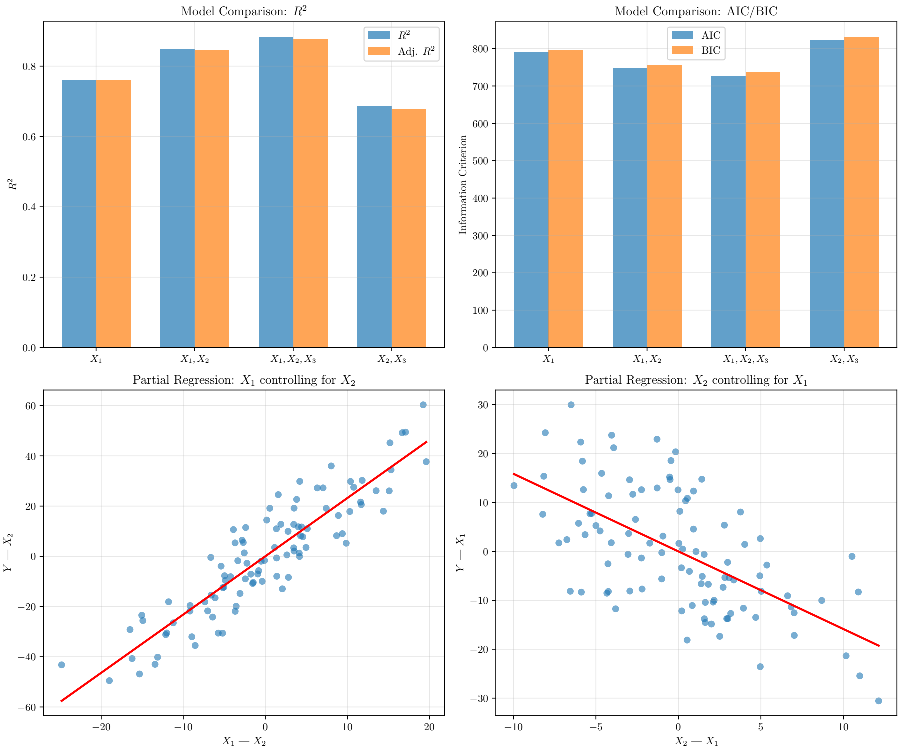

Figure 4: Model comparison using R^2, AIC, BIC, and partial regression plots.

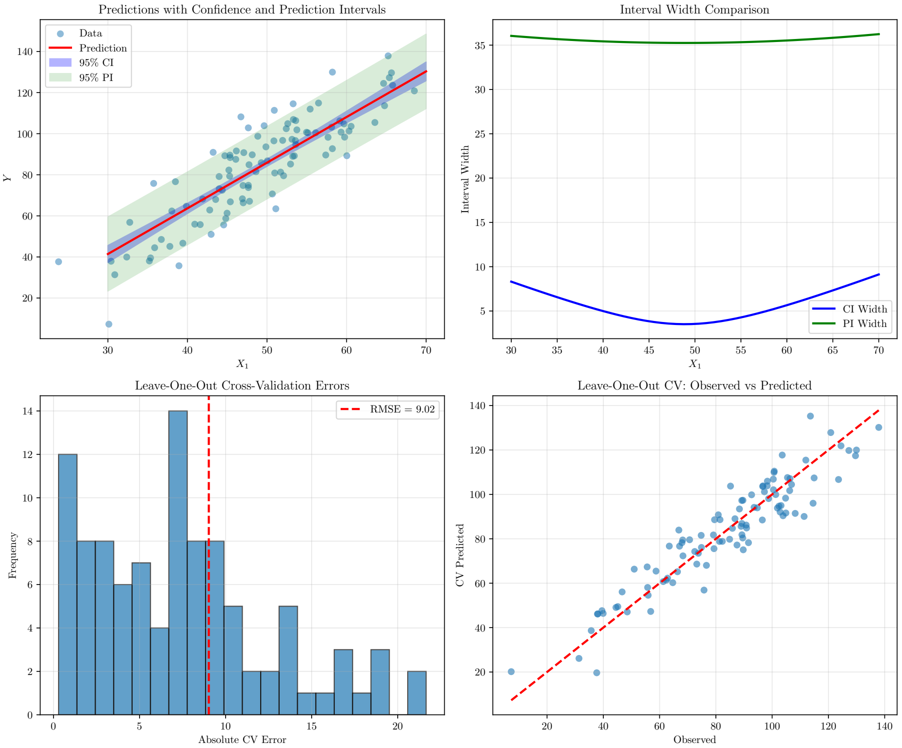

Figure 5: Prediction and confidence intervals with cross-validation assessment.

5 Residual Diagnostics

6 Influential Observations

7 Model Comparison and Selection

8 Prediction Intervals

9 Results Summary

9.1 Regression Coefficients

Table 1: Multiple Regression Coefficient Estimates

Term	True β	Estimate	Std. Error	t-stat	p-value
Intercept	10.0	22.036	8.017	2.75	0.0072*
X_1	2.0	1.796	0.142	12.63	0.0000*
X_2	-1.5	-1.576	0.188	-8.39	0.0000*
X_3	0.8	0.854	0.167	5.12	0.0000*

9.2 Model Fit Statistics

Table 2: Model Fit Statistics

Statistic	Value
R^2	0.8819
Adjusted R^2	0.8782
RMSE	8.835
CV RMSE (LOO)	9.024
AIC	727.5
BIC	737.9

9.3 Diagnostic Statistics

10 Statistical Summary

Key regression analysis findings:

- Model R^2: 0.8819 (Adjusted: 0.8782)
- Root MSE: 8.835

Table 3: Diagnostic Test Results

Diagnostic	Test Value	Interpretation
Shapiro-Wilk (normality)	p = 0.5933	Normal
VIF (X_1)	2.11	OK
VIF (X_2)	1.02	OK
VIF (X_3)	2.10	OK
High leverage points	8	—
High Cook's D points	6	—

- Cross-validation RMSE: 9.024
- VIF range: 1.02 to 2.11
- Influential observations: 6 (Cook's D > threshold)
- Best model by AIC: X_1, X_2, X_3
- Residual normality (Shapiro-Wilk p): 0.5933

11 Conclusion

This comprehensive regression analysis demonstrates model building, assumption checking, and diagnostic procedures for multiple regression. The variance inflation factors reveal multicollinearity between X_1 and X_3, which inflates standard errors but does not bias coefficient estimates. Residual diagnostics confirm approximate normality and homoscedasticity. Cook's distance and leverage analysis identify influential observations that may warrant further investigation. Model selection criteria (AIC, BIC) and cross-validation help balance model complexity against predictive performance. These tools together enable robust statistical inference and reliable predictions.