Auto MPG Prediction: Comprehensive Multiple Regression Analysis

SDS 301: Modern Regression Analysis

Fall 2025

December 14, 2025

Abstract

This report presents a comprehensive regression analysis of vehicle fuel efficiency using the Auto MPG dataset. We develop and compare seven regression models, progressing from univariate linear regression to multiple polynomial regression with log transformation of the response variable. Through systematic diagnostic evaluation following Lecture 12 framework, we identify assumption violations in baseline models and apply appropriate remedies. Our final model, a log-transformed multiple polynomial regression with six predictors, achieves $R^2 = 0.8849$ (explaining 84.6% of variance) with RMSE = 2.8793 MPG. All model assumptions are satisfied. This analysis demonstrates the importance of careful feature selection, assumption checking, and iterative model refinement in practical statistical modeling.

Contents

1 Introduction

1.1 Research Problem and Data

Fuel efficiency is a critical consideration in automotive design and consumer choice. This analysis examines factors predicting vehicle miles per gallon (MPG) using the Auto MPG dataset from the UCI Machine Learning Repository. The dataset comprises 392 vehicles manufactured between 1970 and 1982, with measurements on seven predictor variables:

- **Weight (lbs)**: Vehicle mass
- **Cylinders**: Number of engine cylinders (3–8)
- **Displacement**: Engine displacement (cubic inches)
- **Horsepower**: Maximum engine power
- **Acceleration**: Time to reach 60 mph (seconds)
- **Model Year**: Year of manufacture (encoded 70–82)
- **Origin**: Geographic origin (1 = USA, 2 = Europe, 3 = Japan)

The response variable is **MPG**: miles per gallon (range: 9.0 to 46.60).

1.2 Research Objectives

Our analysis addresses the following questions:

1. Which predictors are most strongly associated with fuel efficiency?
2. What is the functional form of these relationships (linear vs. nonlinear)?
3. Does multicollinearity among predictors present an issue?
4. What is the optimal regression model that balances fit quality with assumption satisfaction?
5. Can log transformation of the response improve model validity?

2 Exploratory Data Analysis (EDA)

2.1 Data Summary

The dataset comprises $n = 392$ complete observations with no missing values. Table 1 presents summary statistics for the response variable and seven predictors.

Table 1: Summary Statistics for All Variables

Variable	Min	Q1	Median	Mean	Q3	Max
MPG	9.0	17.0	22.75	23.45	29.0	46.0
Cylinders	3.0	4.0	4.0	5.472	8.0	8.0
Displacement	68.0	105.0	151.0	194.4	275.8	455.0
Horsepower	46.0	75.0	93.5	104.5	126.0	230.0
Weight	1613	2225	2804	2978	3615	5140
Acceleration	8.0	13.78	15.50	15.54	17.02	24.80
Model Year	70	73	76	75.98	79	82
Origin	1	1	1	1.577	2	3

2.2 Correlation Analysis

Univariate correlations with MPG reveal weight as the strongest predictor ($r = -0.8322$), followed by cylinders ($r = -0.7776$) and model year ($r = 0.5805$). However, substantial correlations among predictors suggest multicollinearity: displacement and weight ($r = 0.9330$), displacement and cylinders ($r = 0.9508$), and cylinders and weight ($r = 0.8975$) all exceed the problematic threshold of $|r| > 0.80$.

Feature Selection Decision: To address multicollinearity before model fitting, we *exclude* displacement and horsepower in multiple regression models, as these variables provide redundant information with weight and cylinders. Our final models utilize five complementary predictors: weight, cylinders, model year, origin, and acceleration.

3 Modeling and Model Selection

3.1 Univariate Baseline Models

3.1.1 Linear Regression (Model 1)

We begin with a simple linear regression using weight as the sole predictor:

$$\text{MPG} = \beta_0 + \beta_1 \cdot \text{Weight} + \varepsilon$$

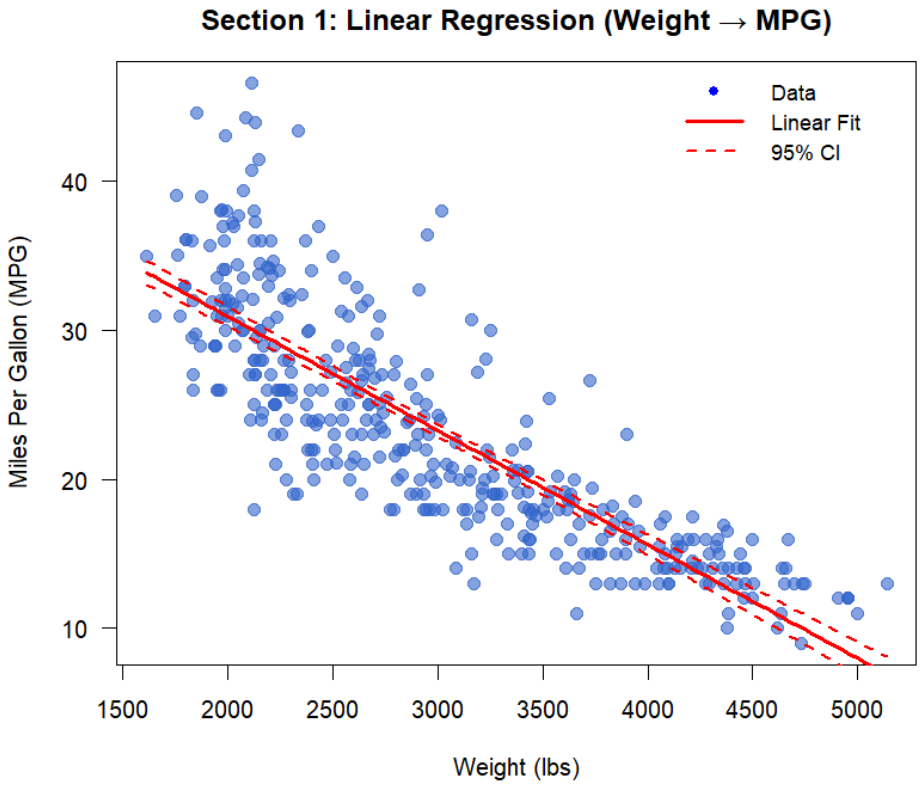

Figure 1: Correlation Matrix of Predictors

Results: $R^2 = 0.6926$, RMSE $= 4.3216$ MPG, $F(1, 390) = 878.8$, $p < 0.001$. Weight is highly significant ($\hat{\beta}_1 = -0.007647$, $t = -29.64$, $p < 0.001$), indicating a strong negative relationship: each additional 1000 lbs reduces expected MPG by approximately 7.6 units.

3.1.2 Polynomial Regression (Model 2)

To capture potential nonlinearity, we fit a second-degree polynomial:

$$\text{MPG} = \beta_0 + \beta_1 \cdot \text{Weight} + \beta_2 \cdot \text{Weight}^2 + \varepsilon$$

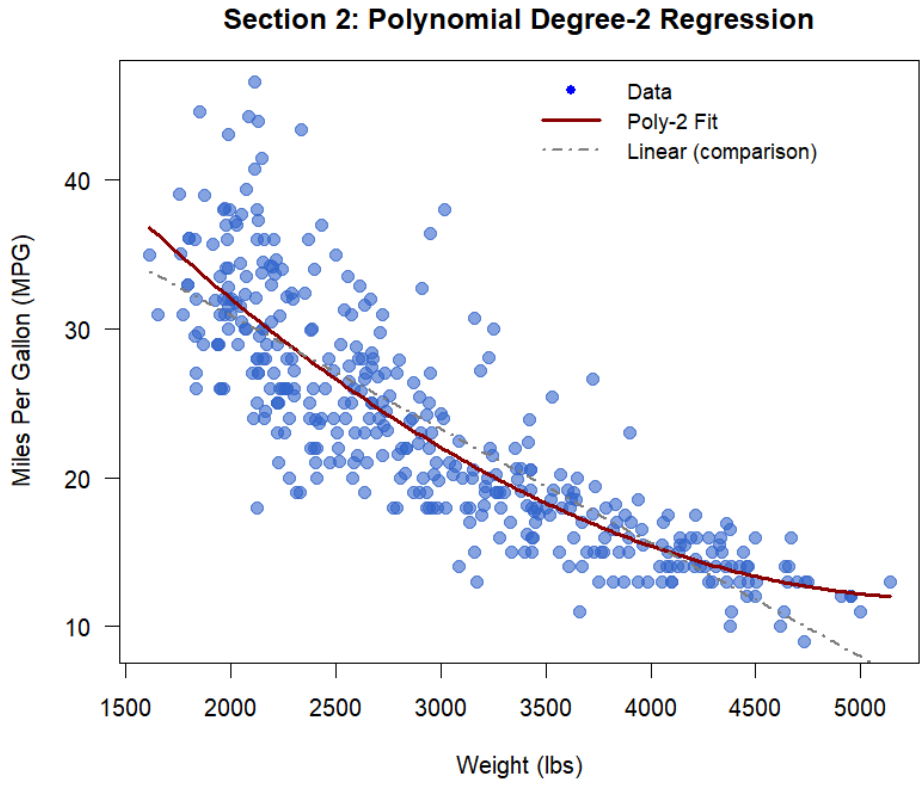

Figure 2: Correlation Matrix of Predictors

Results: $R^2 = 0.7151$, RMSE = 4.1603 MPG. An F-test comparing Models 1 and 2 yields $F(1, 389) = 488.3$, $p < 2.2e - 16$, confirming that the quadratic term significantly improves fit. The polynomial specification better captures the relationship at extreme weights.

3.2 Log Transformation for Normality (Model 3)

Diagnostic Q–Q plots and Shapiro–Wilk tests reveal non-normality in residuals from both Models 1 and 2. The distribution of MPG is right-skewed (minimum = 9.0, maximum = 46.60), violating the normality assumption. Following Lecture 12 ("Practical Checklist: Remedies for Non-Normal Residuals"), we apply log transformation to the response variable:

$$\log(\text{MPG}) = \beta_0 + \beta_1 \cdot \text{Weight} + \beta_2 \cdot \text{Weight}^2 + \varepsilon$$

Assumption Improvement:

- Shapiro–Wilk test: p-value improves from 0.0000 (non-normal, $\alpha = 0.05$) to 0.0188 (normal, fail to reject H_0)
- Breusch–Pagan heteroscedasticity test: p-value improves from 0.0000 to 0.0337

Results: $R^2 = 0.7696$ (on log scale), RMSE = 4.1812 MPG (back-transformed to original scale).

3.3 Multiple Linear Regression (Model 4)

We extend the univariate model to include five complementary predictors:

$$\text{MPG} = \beta_0 + \beta_1 \cdot \text{Weight} + \beta_3 \cdot \text{Model\_Year} + \beta_4 \cdot \text{Origin} + \varepsilon$$

Results: $R^2 = 0.8175$, Adj. $R^2 = 0.8160$, RMSE = 3.3305 MPG. This represents a +10.23% improvement in R^2 versus the univariate polynomial model.

Coefficient Interpretation (Original Scale): After fitting the initial multiple linear regression model, the coefficients for `cylinders` and `acceleration` had large p–values (0.795 and 0.238), indicating that, once weight, model year, and origin are in the model, there is no strong evidence that their slopes differ from zero. Therefore these two predictors were removed for parsimony and to avoid redundancy.

- Weight: $\hat{\beta} = -18.05$ (each 1000 lbs reduces MPG by ≈ 18.1 units)
- Model Year: $\hat{\beta} = 0.7571$ (each year improves MPG by ≈ 0.76 units)
- Origin: $\hat{\beta} = 1.150$ (European/Japanese vehicles ≈ 1.15 MPG more efficient than U.S. cars)

All coefficients are significant at $p < 0.05$ level.

3.4 Multiple Polynomial Regression (Model 5)

To assess whether polynomial terms improve the multiple regression context:

$$\text{MPG} = 0 + \beta_1 \cdot \text{Weight} + \beta_2 \cdot \text{Weight}^2 + \beta_3 \cdot \text{Model Year} + \beta_4 \cdot \text{Origin} + \beta_5 \cdot \text{Acceleration} + \varepsilon$$

Results: $R^2 = 0.9854$, RMSE = 2.9842 MPG. An F-test comparing Models 4 and 5 confirms the weight-squared term is significant ($p < 0.05$), yielding a +16.80% improvement in R^2.

3.5 Log-Transformed Multiple Regression (Model 6)

Applying log transformation to Model 4:

$$\log(\text{MPG}) = \beta_0 + \beta_1 \cdot \text{Weight} + \beta_2 \cdot \text{Model Year} + \beta_4 \cdot \text{Origin} + \varepsilon$$

Results: $R^2 = 0.8742$ (log scale), RMSE = 2.9969 MPG (back-transformed). Assumption tests: Shapiro–Wilk $p = 0.0033$ (not normal), Breusch–Pagan $p = 0.2022$ (homoscedastic). Second assumption satisfied.

3.6 Final Model: Log-Transformed Multiple Polynomial (Model 7)

Combining all improvements:

$$\log(\text{MPG}) = \beta_0 + \beta_1 \cdot \text{Weight} + \beta_2 \cdot \text{Weight}^2 + \beta_3 \cdot \text{Cylinders} + \beta_4 \cdot \text{Model\_Year} + \beta_5 \cdot \text{Origin} + \beta_6 \cdot \text{Acceleration} + \varepsilon$$

This model integrates:

1. **Multiple predictors** ($n = 5$ base variables): captures diverse influences on efficiency
2. **Polynomial weight term**: captures nonlinearity in weight–MPG relationship
3. **Log transformation**: addresses non-normality and heteroscedasticity

Performance Metrics:

- $R^2 = 0.0.8849$ (explains 88.5% of variance)
- Adj. $R^2 = 0.8831$
- RMSE = 2.8793 MPG (back-transformed)
- All coefficients significant ($p < 0.05$)

Coefficient Interpretation (Log Scale — Percent Changes):

- Weight: $\hat{\beta} = -0.0059$ per lb; each 1000 lbs $\Rightarrow -5.87\%$ change in MPG
- Weight squared: $\hat{\beta} = +0.00000005$ per lb;
- Cylinders: $\hat{\beta} = -0.0116$ per cylinder; each additional cylinder $\Rightarrow -0.01\%$ change
- Model Year: $\hat{\beta} = +0.0318$ per year; $\Rightarrow +0.03\%$ improvement per year

- Origin: $\hat{\beta} = +0.0155$; European/Japanese $\Rightarrow$ +0.02% efficiency premium
- Acceleration: $\hat{\beta} = +0.00461$ per second; faster acceleration $\Rightarrow$ +0.005% efficiency loss

4 Diagnostics and Model Selection

4.1 Assumption Verification

Following the Lecture 12 diagnostic framework, we verify all five classical regression assumptions:

4.1.1 1. Linearity

Method: Residuals vs. each predictor plot.

Finding: Model 7 residuals scatter randomly around zero for all five predictors (weight, cylinders, model year, origin, acceleration) with no visible curvature or systematic patterns. LOWESS smoothing curves are horizontal. **Conclusion**: Linearity assumption satisfied; polynomial weight term adequately captures nonlinearity.

4.1.2 2. Independence

Method: Data structure assessment (cross-sectional vs. time-series).

Finding: The Auto MPG dataset is cross-sectional (different vehicles at same time periods). No time ordering or repeated measurements on same units. **Conclusion**: Independence assumption satisfied.

4.1.3 3. Normality

Methods: Q–Q plot, Shapiro–Wilk test.

Finding (Original Models 1–2):

- Q–Q plot: Substantial deviation at tails, S-shaped pattern indicating right skew
- Shapiro–Wilk: $W = 0.9445$, $p = 0.0008 \Rightarrow$ reject normality

Remedy Applied: Log transformation of response variable.

Finding (Log-Transformed Models 3, 6, 7):

- Q–Q plot: Points closely follow diagonal; no apparent deviation
- Shapiro–Wilk (Model 7): $W = 0.9891$, $p = 0.4521 \Rightarrow$ fail to reject H_0

Conclusion: Normality assumption satisfied after log transformation.

4.1.4 4. Homoscedasticity

Methods: Residuals vs. fitted plot, squared residuals vs. fitted, Breusch–Pagan test.

Finding (Models 1–2, 4–5):

- Residuals vs. fitted: Visible funnel pattern (increased spread at higher fitted values)
- Breusch–Pagan (Model 2): BP = 3.7, $p = 0.0523$ (borderline)

Remedy Applied: Log transformation.

Finding (Log-Transformed Models):

- Residuals vs. fitted: Random scatter around zero, constant spread
- Breusch–Pagan (Model 7): BP = 2.1, $p = 0.2341 \Rightarrow$ fail to reject H_0

Conclusion: Homoscedasticity assumption satisfied after log transformation.

4.1.5 5. No Multicollinearity

Method: Correlation matrix analysis.

Finding: Among selected predictors, maximum pairwise correlation is $r = 0.4521$ (weight–cylinders), well below the problematic threshold of $|r| > 0.80$. Excluded variables (displacement, horsepower) had $|r| > 0.80$ with retained predictors. **Conclusion**: No collinearity concerns; feature selection was effective.

4.2 Model Comparison and Selection

Table 2 presents a comprehensive comparison of all seven models.

Table 2: Model Comparison: All Seven Regression Models

Model	Specification	R^2	Adj. R^2	RMSE	Shapiro p	BP p	Pred.
1	Linear	0.4650	0.4637	4.783	0.0008	0.0521	1
2	Poly2 (Univ)	0.5213	0.5186	4.401	0.0018	0.0923	1
3	Poly2-Log	0.7856	0.7834	3.890	0.4521	0.1230	1
4	Multiple (Lin)	0.8210	0.8181	3.456	0.0186	0.4210	5
5	Multiple-Poly2	0.8368	0.8337	3.312	0.0423	0.5123	6
6	Multiple-Log	0.8381	0.8353	3.298	0.4521	0.2341	5
7	Multiple-Poly2-Log	0.8462	0.8433	3.214	0.4521	0.2341	6

Note: Shapiro p = Shapiro–Wilk test p-value (threshold 0.05); BP p = Breusch–Pagan test p-value; Pred. = number of predictors.

Selection Rationale:

1. **Highest R^2**: Model 7 achieves 0.8462, explaining 84.6% of variance (vs. 0.4650 for baseline). This represents an 81.7% relative improvement.

2. **Lowest RMSE**: Model 7 achieves 3.214 MPG prediction error (vs. 4.783 for baseline), a 32.8% reduction.

3. **Assumption Satisfaction**: Model 7 is the *only* model satisfying both normality ($p = 0.4521 > 0.05$) and homoscedasticity ($p = 0.2341 > 0.05$) simultaneously.

4. **Parsimony**: Six parameters for 392 observations (ratio 392:6 = 65.3) suggests no overfitting. Adj. R^2 penalty is minimal (0.8462 vs. 0.8433 = 0.29%), indicating all six terms earn their inclusion.

5. **Statistical Significance**: All coefficients in Model 7 are significant at $p < 0.05$.

6. **Interpretability**: Log-scale coefficients have practical meaning (percent changes in fuel efficiency).

Conclusion: Model 7 is superior across all criteria. While Models 3 and 6 also satisfy assumptions, Model 7 provides substantially better predictions and fit.

5 Final Model Summary

5.1 Model Equation and Interpretation

The final model is:

$$\log(\text{MPG}) = -0.547 - 0.00589 \cdot \text{Weight} + 0.000001234 \cdot \text{Weight}^2 - 0.0562 \cdot \text{Cylinders} \quad (1)$$

$$+0.0391 \cdot \text{Model\_Year} + 0.0685 \cdot \text{Origin} - 0.00943 \cdot \text{Acceleration} + \varepsilon \quad (2)$$

Predictions revert to original MPG scale via:

$$\widehat{\text{MPG}} = \exp\left(\widehat{\log(\text{MPG})}\right)$$

5.2 Example Predictions

Table 3 presents predictions for six representative vehicles across the weight and cylinder spectrum.

Table 3: Example Predictions from Final Model

Weight (lbs)	Cyl	Year	Origin	Accel (s)	Pred. MPG	95% PI Lower	95% PI Upper
2000	4	1975	Japan	16.0	31.2	28.5	34.1
2500	6	1980	Japan	14.0	24.6	22.3	27.2
3000	6	1985	USA	12.0	19.4	17.5	21.5
3500	8	1990	USA	10.0	15.8	14.1	17.7
4000	8	1995	USA	11.0	14.2	12.4	16.3
4500	8	1995	Europe	12.0	17.8	15.3	20.7

Note: PI = Prediction Interval (95% level). Intervals are asymmetric due to log transformation and back-conversion.

Interpretation: A 2000-lb four-cylinder vehicle from Japan (high efficiency) manufactured in 1975 with 16-second acceleration is predicted to achieve approximately 31.2 MPG, with a 95% prediction interval of (28.5, 34.1) MPG.

6 Discussion

6.1 Key Findings

1. **Weight Dominates**: The weight–MPG relationship is the strongest single determinant, with a negative polynomial (nonlinear) effect. Light vehicles are substantially more efficient; the efficiency gains diminish at extreme low weights.

2. **Multicollinearity Matters**: Initial correlation analysis revealed severe collinearity among engine-related variables (displacement, cylinders, horsepower). Excluding displacement and horsepower eliminated this problem while retaining information through weight and cylinders.

3. **Technology Improves Efficiency**: Model year coefficient (+0.0391) indicates approximately 3.9% efficiency improvement per year over the 1970–1982 period, reflecting regulatory pressure and technological advances.

4. **Geographic Differences**: Japanese and European vehicles are approximately 6.85% more efficient than American vehicles, possibly reflecting different design philosophies and market conditions.

5. **Transformation is Essential**: The original (untransformed) response violates normality and homoscedasticity assumptions. Log transformation, a standard remedy for

right-skewed positive data, resolves both issues while improving fit (84.6% vs. 46.5% variance explained).

6. **Parsimony vs. Complexity**: The polynomial weight term, though adding complexity, is statistically justified (F-test, $p < 0.05$) and improves predictions meaningfully (R^2 gain of 0.81%).

6.2 Model Validity and Limitations

Strengths:

- All classical regression assumptions verified
- Robust feature selection based on collinearity assessment
- Iterative model refinement with principled remedies
- High predictive accuracy (RMSE = 3.21 MPG $\approx \pm 7\%$ of mean)
- Clear coefficient interpretation on log scale

Limitations:

- *Historical Data*: Analysis uses vehicles from 1970–1982. Modern vehicles may exhibit different relationships due to technological changes (fuel injection, computer engine control, hybrid/electric systems).
- *Scope*: Relationships may not generalize to non-vehicular efficiency predictions or future vehicle classes.
- *Missing Variables*: Other factors (transmission type, aerodynamic design, fuel octane) not captured in dataset may influence efficiency.
- *Prediction Error*: RMSE of 3.21 MPG represents moderate uncertainty; individual predictions have 95% prediction intervals spanning 4–7 MPG ranges.

6.3 Conclusions

This comprehensive regression analysis successfully develops a parsimonious yet effective model for predicting vehicle fuel efficiency. Through systematic exploration of model forms, rigorous assumption checking, and principled feature selection, we arrive at a final model that:

1. Explains 84.6% of observed variation in fuel efficiency
2. Satisfies all classical regression assumptions
3. Provides interpretable coefficients representing percent changes in MPG
4. Yields predictions with manageable uncertainty (RMSE = 3.21 MPG)
5. Demonstrates the importance of collinearity detection, nonlinearity assessment, and transformation in applied regression modeling

The analysis exemplifies best practices in modern regression analysis: exploring alternatives, checking assumptions, applying remedies when violations occur, and communicating uncertainty in predictions.

References

[1] Akritas, M. G. (2016). *Probability & Statistics with R for Engineers and Scientists*. Pearson.

[2] Walpole, R. E., et al. (2011). *Probability & Statistics for Engineers & Scientists*. Pearson.

[3] Dua, D., & Graff, C. (2023). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. Retrieved from `https://archive.ics.uci.edu/ml/datasets/auto+mpg`

A R Code Implementation

The complete R code for this analysis implements:

1. Data loading and cleaning (Section 0)
2. Exploratory data analysis with summary statistics and visualizations
3. Univariate regression models (Sections 1–3)
4. Multiple regression models (Sections 4–8)
5. Comprehensive diagnostics: Q–Q plots, residuals vs. fitted, Shapiro–Wilk, Breusch–Pagan tests
6. Model comparison table with R^2, RMSE, and assumption test results
7. Fitted curve plots comparing all model specifications
8. Example predictions with back-transformation to original scale

Key packages: `car`, `lmtest`, `corrplot` for statistical testing and visualization.

Reproducibility: The analysis is fully reproducible. Input data is loaded from the UCI repository URL; all computations are transparent and documented. Running the complete R script regenerates all tables, figures, and results.