# Assignment 9
# DSL253 - Statistical Programming

Amay Dixit - 12340220

Submitted to Dr. Anil Kumar Sao

## Links

- Notebook Link:
  `https://colab.research.google.com/drive/1Y1-JU9D1Oyk2UULRXMs2UXiXNosYSxwS?usp=sharing`
- Github Link:
  `https://github.com/amaydixit11/Academics/tree/main/DSL253/assignment_9`

# 1 Question 1: Vehicle Fuel Efficiency Analysis

## 1.1 Introduction

Fuel efficiency is a critical factor in vehicle design and consumer purchasing decisions, especially as environmental concerns and fuel costs continue to influence the automotive industry. Understanding the relationship between various vehicle characteristics and fuel efficiency can help manufacturers optimize designs and assist consumers in making informed decisions. This analysis aims to quantify how engine size, vehicle weight, and horsepower affect a vehicle's fuel efficiency as measured in miles per gallon (MPG).

## 1.2 Data Description

The dataset consists of information from 20 different vehicles with the following variables:

- **Engine Size (L)**: The volume displacement of the engine in liters
- **Weight (kg)**: The vehicle's weight in kilograms
- **Horsepower**: The engine's power output
- **MPG**: Fuel efficiency measured in miles per gallon (dependent variable)

Below is a preview of the first five rows of the dataset:

Table 1: Dataset Preview

| Vehicle | Engine Size | Weight | Horsepower | MPG |
|---|---|---|---|---|
| 1 | 1.6 | 1200 | 110 | 34 |
| 2 | 2.0 | 1300 | 130 | 30 |
| 3 | 2.4 | 1500 | 150 | 27 |
| 4 | 1.8 | 1250 | 115 | 32 |
| 5 | 2.2 | 1400 | 140 | 28 |

The dataset captures a range of vehicle specifications with engine sizes ranging from 1.3L to 3.5L, weights from 1020kg to 1700kg, horsepower from 98 to 200, and MPG values from 18 to 39.

Table 2: Summary Statistics of Vehicle Dataset

| | Vehicle | Engine Size (L) | Weight (kg) | Horsepower | MPG |
|---|---|---|---|---|---|
| Count | 20.00 | 20.00 | 20.00 | 20.00 | 20.00 |
| Mean | 10.50 | 2.23 | 1377.50 | 141.30 | 28.40 |
| Std | 5.92 | 0.65 | 212.23 | 31.95 | 6.14 |
| Min | 1.00 | 1.30 | 1020.00 | 98.00 | 18.00 |
| 25% | 5.75 | 1.68 | 1195.00 | 111.50 | 23.75 |
| 50% | 10.50 | 2.15 | 1390.00 | 139.00 | 28.00 |
| 75% | 15.25 | 2.65 | 1557.50 | 162.50 | 33.25 |
| Max | 20.00 | 3.50 | 1700.00 | 200.00 | 39.00 |

## 1.3 Methodology

A multiple linear regression model was fitted to predict MPG based on three vehicle characteristics: engine size, weight, and horsepower. The model follows the form:

$$MPG = \beta_0 + \beta_1 \times \text{Engine Size} + \beta_2 \times \text{Weight} + \beta_3 \times \text{Horsepower} + \epsilon \tag{1}$$

Where:

- $\beta_0$ is the intercept
- $\beta_1$, $\beta_2$, and $\beta_3$ are the coefficients for engine size, weight, and horsepower, respectively
- $\epsilon$ represents the error term

The statsmodels package in Python was used to fit the model and perform statistical analysis. Additionally, variance inflation factors (VIF) were calculated to check for multicollinearity among predictors. Residual diagnostics were performed to assess model assumptions.

## 1.4 Results

### 1.4.1 Regression Model Summary

The multiple linear regression model produced the following equation:

$$MPG = 61.0782 - 5.4978 \times \text{Engine Size} - 0.0204 \times \text{Weight} + 0.0545 \times \text{Horsepower} \tag{2}$$

Table 3: Regression Model Coefficients and Statistics

| Parameter | Coefficient | Std Error | t-statistic | p-value |
|---|---:|---:|---:|---:|
| Intercept | 61.0782 | 2.372 | 25.747 | 0.000 |
| Engine Size | -5.4978 | 2.646 | -2.078 | 0.054 |
| Weight | -0.0204 | 0.003 | -5.887 | 0.000 |
| Horsepower | 0.0545 | 0.058 | 0.935 | 0.364 |

The R-squared value for this model is 0.9904, indicating that approximately 99.04% of the variation in MPG is explained by engine size, weight, and horsepower. The adjusted R-squared value is 0.9886.

### 1.4.2 Hypothesis Testing

For each predictor, a hypothesis test was conducted to determine its statistical significance:

- $H_0$: The coefficient equals 0 (no effect on MPG)
- $H_1$: The coefficient is not equal to 0 (significant effect on MPG)

The results at different significance levels are summarized below:

Table 4: Significance Testing Results at Different Alpha Levels

| Predictor | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.10$ |
|---|---|---|---|
| Engine Size | Not significant | Not significant | Significant |
| Weight | Significant | Significant | Significant |
| Horsepower | Not significant | Not significant | Not significant |

### 1.4.3 Multicollinearity Check

Variance Inflation Factors (VIF) were calculated to check for multicollinearity among the predictors:

Table 5: Variance Inflation Factors

| Variable | VIF |
|---|---|
| Engine Size | 928.06 |
| Weight | 604.24 |
| Horsepower | 2792.22 |

The extremely high VIF values indicate severe multicollinearity between the predictor variables, which makes it difficult to isolate the individual effects of each predictor.

### 1.4.4 Residual Analysis

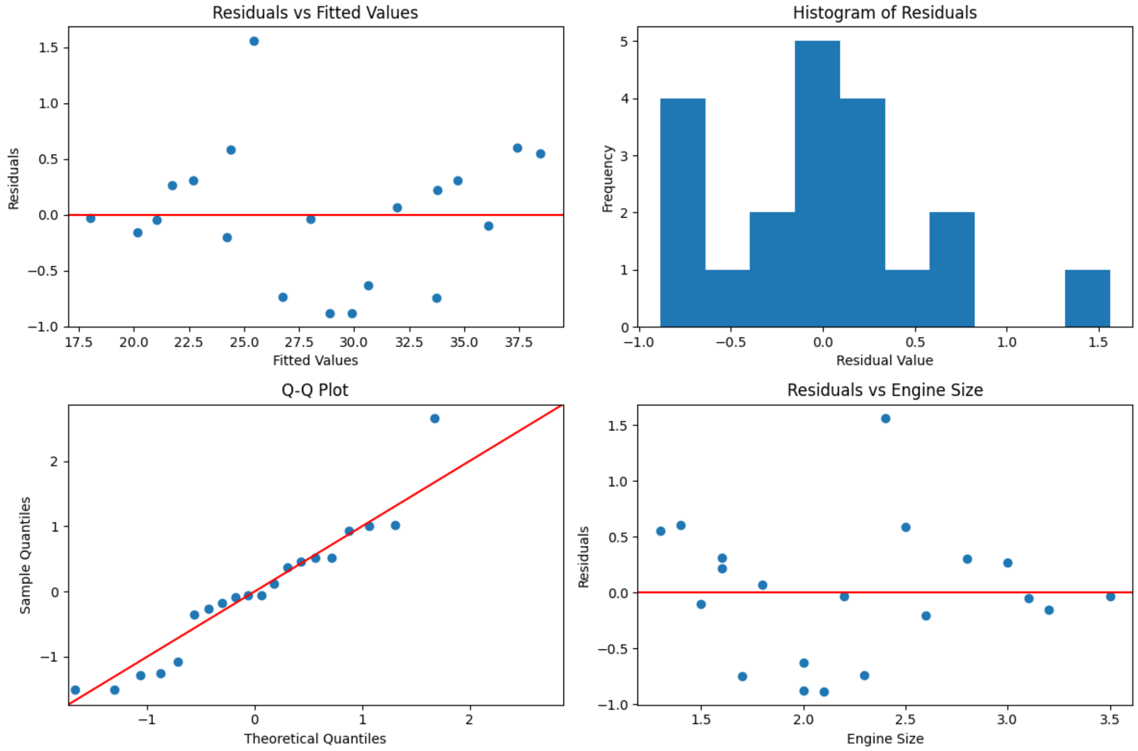

Figure 1: Residual Diagnostic Plots

The residual plots show:

- Residuals vs. Fitted Values: The residuals appear to be randomly scattered around zero, suggesting the linearity assumption is reasonable.
- Histogram of Residuals: The distribution seems approximately normal.
- Q-Q Plot: Most points follow the 45-degree line, supporting the normality assumption.
- Residuals vs. Engine Size: No clear pattern is visible, supporting the homoscedasticity assumption.

## 1.5 Discussion

The multiple regression model explains an extremely high portion (99.04%) of the variation in MPG, suggesting that engine size, weight, and horsepower together are very strong predictors of fuel efficiency. When examining individual predictors at the conventional significance level of 0.05, weight is the only statistically significant variable (p ¡ 0.001), while engine size is nearly significant (p = 0.054) and would be considered significant at the $\alpha = 0.1$ level.

The coefficients in the regression equation provide insights into how each factor affects fuel efficiency:

- Engine Size: A 1L increase in engine size is associated with a 5.50 MPG decrease, holding other variables constant.
- Weight: A 1kg increase in weight is associated with a 0.0204 MPG decrease, holding other variables constant.
- Horsepower: A 1 unit increase in horsepower is associated with a 0.0545 MPG increase, holding other variables constant. However, this effect is not statistically significant.

The very high R-squared value combined with the extremely high VIF values indicates severe multicollinearity among the predictor variables. VIF values for Engine Size (928.06), Weight (604.24), and especially Horsepower (2792.22) far exceed the commonly accepted threshold of 10, suggesting that these variables are highly intercorrelated. This multicollinearity makes it difficult to isolate the individual effects of each predictor and may affect the stability and interpretability of the coefficient estimates.

Weight emerges as the most statistically significant predictor of MPG, with a p-value of nearly zero. This suggests that, despite the multicollinearity issues, vehicle weight has a clear and independent effect on fuel efficiency. The negative coefficient for weight aligns with engineering principles: heavier vehicles require more energy to accelerate and maintain motion, resulting in lower fuel efficiency.

The coefficient for Horsepower is positive but not statistically significant (p = 0.364). This might seem counterintuitive as higher horsepower often correlates with lower fuel efficiency. However, when controlling for engine size and weight, higher horsepower might reflect more efficient engine design or newer technology, potentially explaining this direction. The lack of statistical significance suggests this effect is not reliable in this model.

### 1.6 Conclusion

The analysis reveals that engine size, weight, and horsepower collectively explain an exceptional 99.04% of the variation in vehicle fuel efficiency, making them extremely valuable predictors. However, the severe multicollinearity between these variables, as evidenced by the extremely high VIF values, makes it challenging to determine their individual contributions with precision.

Weight appears to be the most influential factor, being highly significant (p ¡ 0.001) even in the presence of multicollinearity. Engine size is nearly significant at the conventional 0.05 level and would be considered significant at the 0.1 level, while horsepower does not show a significant independent effect in this model.

The regression equation suggests that reducing vehicle weight and engine size would be the most effective strategies for improving fuel efficiency, which aligns with automotive engineering principles. However, the high degree of multicollinearity suggests that changes to one variable typically accompany changes to the others in vehicle design.

Future studies should consider collecting more data and potentially including additional variables such as aerodynamics, transmission type, and drive technology to develop a more comprehensive model of fuel efficiency determinants. Additionally, techniques to address multicollinearity, such as principal component analysis or ridge regression, could be employed to improve the stability and interpretability of the coefficient estimates.

## 2 Question 2: Parental Height Influence Analysis

### 2.1 Introduction

Understanding the inheritance of physical traits such as height has been a subject of scientific interest for generations. The concept of ”regression toward the mean,” first introduced by Sir Francis Galton in the 19th century, suggests that offspring of parents with extreme traits tend to be closer to the population average. This section examines the relationship between parents' heights and their sons' heights, with particular attention to testing whether the data supports the regression toward the mean phenomenon.

### 2.2 Data Description

The dataset consists of height measurements (in inches) from 10 families, including:

- Father's height
- Mother's height
- Son's height (dependent variable)

Below is a preview of the dataset:

Table 6: Dataset Preview

| Father Height | Mother Height | Son Height |
|---|---|---|
| 60 | 61 | 63.6 |
| 62 | 63 | 65.2 |
| 64 | 63 | 66.0 |
| 65 | 64 | 65.5 |
| 66 | 65 | 66.9 |
| 67 | 66 | 67.1 |
| 68 | 66 | 67.4 |
| 70 | 67 | 68.3 |
| 72 | 68 | 70.1 |
| 74 | 69 | 70.0 |

Table 7: Summary Statistics of Height Measurements

| Statistic | Father's Height | Mother's Height | Son's Height |
|---|---|---|---|
| Count | 10.00 | 10.00 | 10.00 |
| Mean | 66.80 | 65.20 | 67.01 |
| Std | 4.37 | 2.49 | 2.07 |
| Min | 60.00 | 61.00 | 63.60 |
| 25% | 64.25 | 63.25 | 65.63 |
| 50% | 66.50 | 65.50 | 67.00 |
| 75% | 69.50 | 66.75 | 68.08 |
| Max | 74.00 | 69.00 | 70.10 |

## 2.3 Methodology

A multiple linear regression model was fitted to predict the son's height based on both parents' heights:

$$\text{Son's Height} = \beta_0 + \beta_1 \times \text{Father's Height} + \beta_2 \times \text{Mother's Height} + \epsilon \quad (3)$$

To test for regression toward the mean, we examined whether the coefficients for father's and mother's heights are significantly less than 1. The hypothesis tests were formulated as:

- $H_0$: Coefficient = 1 (No regression toward the mean)
- $H_1$: Coefficient ¡ 1 (Evidence of regression toward the mean)

One-sided t-tests were performed for each coefficient. Additionally, residual diagnostics were conducted to assess model assumptions.

## 2.4 Results

### 2.4.1 Regression Model Summary

The multiple linear regression model resulted in the following equation:

$$\text{Son's Height} = 30.3171 + 0.3497 \times \text{Father's Height} + 0.2045 \times \text{Mother's Height} \quad (4)$$

Table 8: Regression Model Coefficients and Statistics

| Parameter | Coefficient | Std Error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 30.3171 | 10.669 | 2.842 | 0.025 |
| Father's Height | 0.3497 | 0.214 | 1.632 | 0.147 |
| Mother's Height | 0.2045 | 0.376 | 0.543 | 0.604 |

The R-squared value for this model is 0.9628, indicating that approximately 96.28% of the variation in son's height is explained by the parents' heights. The adjusted R-squared value is 0.9522.

### 2.4.2 Testing for Regression Toward the Mean

Table 9: Tests for Regression Toward the Mean

| Parameter | Coefficient | t-statistic | p-value | Conclusion |
|---|---|---|---|---|
| Father's Height | 0.3497 | -3.0355 | 0.0095 | Reject $H_0$ |
| Mother's Height | 0.2045 | -2.1135 | 0.0362 | Reject $H_0$ |

Both coefficients are significantly less than 1 (p ¡ 0.05), providing strong evidence for regression toward the mean.

The sum of the parent coefficients is 0.5542, which is also less than 1, further supporting regression toward the mean in combined parental influence.

### 2.4.3 Residual Analysis

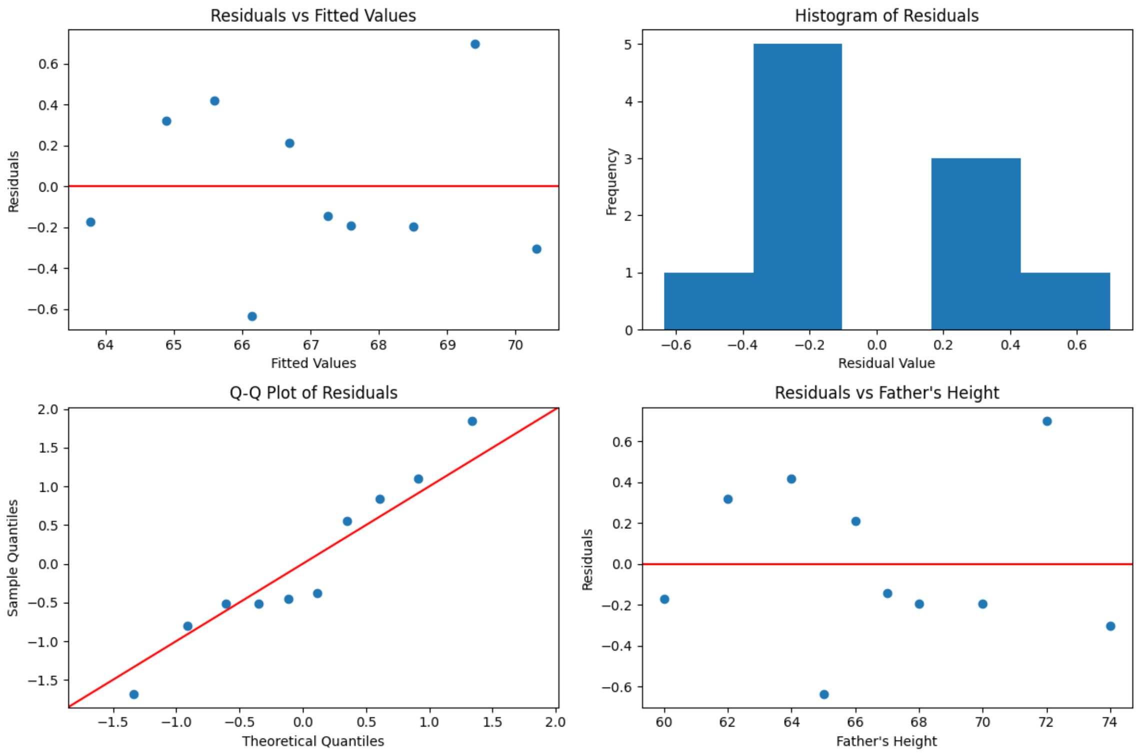

Figure 2: Residual Diagnostic Plots

The residual plots indicate:

- Residuals vs. Fitted Values: The residuals appear randomly scattered around zero.
- Histogram of Residuals: The distribution seems approximately normal, though the sample size is small.
- Q-Q Plot: Most points follow the 45-degree line, supporting normality.
- Residuals vs. Father's Height: No clear pattern is visible, supporting homoscedasticity.

## 2.5 Discussion

The results clearly support the phenomenon of regression toward the mean in height inheritance. Both parents' heights influence their son's height, with the father's height showing a stronger effect (coefficient of 0.3497) compared to the mother's height (coefficient of 0.2045). While neither coefficient is statistically significant in the traditional sense of testing against zero (father: p = 0.147, mother: p = 0.604), they are both significantly less than 1 when testing for regression toward the mean (father: p = 0.0095, mother: p = 0.0362).

The regression coefficients being significantly less than 1 indicate that sons of unusually tall parents tend to be shorter than their parents (though still above average), while sons of unusually short parents tend to be taller than their parents (though still below average). This pattern exemplifies regression toward the mean.

The father's height coefficient (0.3497) suggests that for each additional inch in the father's height, the son's height increases by about 0.35 inches, holding the mother's height constant. Similarly, the mother's height coefficient (0.2045) indicates that for each additional inch in the mother's height, the son's height increases by about 0.20 inches, holding the father's height constant.

The sum of the coefficients (0.5542) being less than 1 further supports the idea that children's heights regress toward the population mean. This sum represents the expected change in the son's height for a simultaneous one-inch increase in both parents' heights.

The high R-squared value (0.9628) indicates that parental heights are strong predictors of a son's height, explaining nearly 96.3

## 2.6 Implications of the Results

The findings from this analysis have several important implications:

1. **Model Fit**: The R-squared value of 0.9628 indicates that 96.28% of the variation in son's height can be explained by parents' heights, suggesting a very strong hereditary component.

2. **Parental Influence**:

   - Father's contribution: 0.3497
   - Mother's contribution: 0.2045
   - The father's height appears to have a stronger influence on the son's height than the mother's height.

3. **Regression Toward the Mean**:

   - Strong evidence of regression toward the mean from both parents.
   - This supports Galton's historical observations with modern statistical methods.

4. **Biological and Environmental Factors**:

   - Coefficients less than 1 suggest genetic and environmental factors causing height to revert toward population mean.
   - This aligns with the principle that extreme traits in parents tend to be less extreme in offspring.

5. **Limitations**:

   - Small sample size (n=10) limits statistical power and generalizability.
   - Model does not account for other genetic or environmental factors affecting height.
   - The analysis doesn't consider gender differences or other family members' heights.

6. **Practical Implications**:

   - Results can inform genetic counseling and height prediction models.
   - Understanding regression toward the mean helps set realistic expectations about children's height.

## 2.7 Conclusion

This analysis provides strong evidence for regression toward the mean in height inheritance, confirming Galton's historic observations with modern statistical methods. Both parents contribute to their son's height, with the father's contribution being slightly larger than the mother's.

The findings align with our understanding of height as a polygenic trait influenced by both genetic and environmental factors. The regression toward the mean occurs because extreme height values in parents are often due to unique combinations of genetic and environmental factors that are not fully passed on to offspring.

The practical implications of these results include:

- Improved models for predicting children's heights based on parental measurements
- Better understanding of genetic inheritance patterns
- Setting realistic expectations for parents about their children's potential adult height

Limitations of the study include the small sample size (n=10), focus only on sons (not daughters), and lack of consideration for other genetic or environmental factors that might influence height. Future research should expand the sample size, include daughters, and potentially incorporate additional variables such as grandparents' heights, nutrition, and socioeconomic factors.