ROHIT RAGHUWANSHI 12341820 ASG9
COLAB LINK- https://colab.research.google.com/drive/1LptfMu5lEwF4rTyWqyLi3qROyaae6sX9#scrollTo=0l3yqolTm05c
QUESTION 1
Car Features and Fuel Efficiency Analysis Report

Introduction

This study investigates the relationship between vehicle characteristics and fuel efficiency. Understanding how features such as engine size, weight, and horsepower affect miles per gallon (MPG) is crucial for automotive engineers, manufacturers, and consumers interested in fuel economy. This analysis employs multiple linear regression to quantify these relationships and determine which factors have the most significant impact on fuel efficiency.

Data

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

- **Engine Size**: Measured in liters (L)
- **Weight**: Measured in kilograms (kg)
- **Horsepower**: Engine power output
- **MPG**: Miles per gallon, representing fuel efficiency

Summary statistics of the dataset:

Statistic	Engine Size (L)	Weight (kg)	Horsepower	MPG
Count	20	20	20	20
Mean	2.23	1355	138.8	28.9
Std Dev	0.64	208.5	30.8	6.2
Min	1.3	1020	98	18
Max	3.5	1700	200	39

Methodology

A multiple linear regression model was fitted to determine the relationship between the three predictor variables (Engine Size, Weight, and Horsepower) and the response variable (MPG). The model takes the form:

$MPG = \beta_0 + \beta_1(\text{Engine Size}) + \beta_2(\text{Weight}) + \beta_3(\text{Horsepower}) + \varepsilon$

Where:

- β_0 is the intercept
- β_1, β_2, β_3 are the coefficients for each predictor
- ε represents the error term

The statistical significance of each predictor was assessed using t-tests at different significance levels ($\alpha = 0.01$, 0.05, and 0.10). The overall model fit was evaluated using the coefficient of determination (R^2) and analysis of residuals.

Results

Regression Equation

The fitted regression model is:

MPG = 61.0782 + (-5.4978 × Engine Size) + (-0.0204 × Weight) + (0.0545 × Horsepower)

Model Statistics

Statistic	Value
R-squared	0.990
Adjusted R-squared	0.989
F-statistic	548.9
Prob (F-statistic)	2.44e-16

Coefficients and Statistical Significance

Predictor	Coefficient	Std Error	t-value	p-value	Significance at α=0.05
Intercept	61.O782	2.372	25.747	0.000	Significant
Engine Size	-5.4978	2.646	-2.078	0.054	Not Significant
Weight	-0.0204	0.003	-5.887	0.000	Significant
Horsepower	0.0545	0.058	0.935	0.364	Not Significant

At a significance level of α = 0.05:

- Engine Size is a significant predictor (p = 0.0542)
- Weight is a significant predictor (p = 0.000)
- Horsepower is not a significant predictor (p = 0.3637)

At significance level α = 0.1:
EngineSize: p-value = 0.0542, Significant
Weight: p-value = 0.0000, Significant
Horsepower: p-value = 0.3637, Not significant
At a significance level of α = 0.01:
EngineSize: p-value = 0.0542, Not significant
Weight: p-value = 0.0000, Significant
Horsepower: p-value = 0.3637, Not significant

Residual Analysis

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.

Multicollinearity Assessment

Variance Inflation Factors (VIF):

Variable	VIF
Engine Size	128.954543
Weight	23.865845
Horsepower	152.743866

The high VIF values suggest multicollinearity among predictors, particularly for Horsepower, which may affect the stability of coefficient estimates.

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 α= 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 1 kg 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 (128.954543), Weight (23.865845), and especially Horsepower (152.743866) 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.

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.

QUESTION 2
Parent-Child Height Relationship Analysis Report
Introduction
Understanding the genetic and environmental factors that influence human height has been a subject of scientific interest for generations. This study examines the relationship between parents' heights and their male offspring's height, with particular attention to the phenomenon of "regression toward the mean" - the tendency for children of unusually tall or short parents to be closer to average height than their parents. This analysis employs multiple linear regression to quantify parental influence on offspring height and test for evidence of regression toward the mean.
Data
The dataset contains height measurements (in inches) from 10 families, including:

- **Father's Height**: Height of the biological father
- **Mother's Height**: Height of the biological mother
- **Son's Height**: Height of the male offspring

Summary statistics of the dataset:

Statistic	Father's Height	Mother's Height	Son's Height
Count	10	10	10
Mean	66.8	65.2	67.0
Std Dev	4.3	2.5	2.2
Min	60	61	63.6
Max	74	69	70.1

Methodology
A multiple linear regression model was fitted to determine the relationship between the two predictor variables (Father's Height and Mother's Height) and the response variable (Son's Height). The model takes the form:
Son's Height = β_0 + β_1(Father's Height) + β_2(Mother's Height) + ε
Where:

- β_0 is the intercept
- β_1, β_2 are the coefficients for father's and mother's heights
- ε represents the error term

To test for regression toward the mean, one-sided hypothesis tests were conducted to determine if the regression coefficients are significantly less than 1. The null and alternative hypotheses are:
H_0: $\beta_1 = 1$ (or $\beta_2 = 1$) H_1: $\beta_1 < 1$ (or $\beta_2 < 1$)
If the coefficients are significantly less than 1, this provides evidence for regression toward the mean, suggesting that children of unusually tall or short parents tend to be closer to average height than their parents.
Results
Regression Equation
The fitted regression model is:
Regression Equation:
Son's Height = 30.3171 + (0.3497 × Father's Height) + (0.2045 × Mother's Height)
Model Statistics

Statistic	Value

R-squared	0.963
Adjusted R-squared	0.952
F-statistic	90.60
Prob (F-statistic)	9.93e-06

Coefficients and Statistical Significance

Predictor	Coefficient	Std Error	t-value	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

Testing for Regression Toward the Mean

Parameter	Father's Height	Mother's Height
Coefficient	0.3497	0.2045
H_0	$\beta_1 = 1$	$\beta_2 = 1$
H_1	$\beta_1 < 1$	$\beta_2 < 1$
t-statistic	-3.0355	-2.1135
p-value (one-sided)	0.9905	0.9638
Decision at α=0.05	Fail to reject H_0	Fail to reject H_0

Residual Analysis

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.

Discussion

Interpretation of Results:

1. The multiple regression model has an R-squared value of 0.9628, indicating that 96.3% of the variation in son's height is explained by parents' heights.
2. Father's height coefficient (0.3497): For each additional inch in father's height, son's height increases by approximately 0.3497 inches, holding mother's height constant.
3. Mother's height coefficient (0.2045): For each additional inch in mother's height, son's height increases by approximately 0.2045 inches, holding father's height constant.

Regression to the Mean Summary:

- Father's height coefficient is not significantly less than 1
- Mother's height coefficient is not significantly less than 1

The data does not provide strong evidence for regression toward the mean.

Implications

1. Genetic influence: Both parents' heights significantly contribute to predicting their son's height.
2. The sum of parental coefficients (0.5542) represents the total parental influence on a son's height.
3. The intercept represents other factors that influence height beyond parental genetics (e.g., nutrition, environment).
4. The regression model allows us to predict a son's expected height based on his parents' heights.

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 onto 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.