Final Project: Linear Regression Analysis of the "Auto-mpg" Dataset

Omar A. Mohammed, Omar M. Ahmed, Amgad A. Shaban, Mahmoud M. Ahmed

Department of Systems and Biomedical Engineering, Faculty of Engineering, Cairo University

KEYWORDS

Statistics
Machine Learning
Data Science
Linear Regression
Prediction
Inference
Fuel Consumption

ABSTRACT

This project conducted a linear regression analysis on the Auto-mpg dataset to investigate the relationship between various attributes and fuel efficiency. The study identified significant factors such as cylinders, displacement, horsepower, weight, and model year that impact mpg. Outliers in acceleration and horsepower were addressed, and limitations and assumptions were discussed. Suggestions for future research included exploring additional variables and employing advanced regression techniques. The insights gained from this analysis provide valuable information for vehicle manufacturers, policymakers, and consumers in optimizing fuel efficiency and making informed decisions.

1. Introduction

Objective: The objective of this project is to use linear regression analysis on the Auto-mpg dataset to predict the fuel efficiency (mpg) of a vehicle based on its other characteristics.

Dataset: The Auto-mpg dataset is a well-known automotive dataset available on Kaggle. It contains information about various cars from the late 1970s and early 1980s, including attributes such as miles per gallon (mpg), cylinders, displacement, horsepower, weight, acceleration, model year, and origin.

The purpose of this dataset is to predict the fuel efficiency (mpg) of a vehicle based on its other characteristics. It is often used as a benchmark dataset in regression tasks and machine learning algorithms.

the data is available here Auto-mpg-dataset

Linear regression: Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find a linear equation that best describes this relationship.

The dependent variable is the variable to be predicted or explained, which is – in our case –: mpg, while the other variables are the independent variables used to make predictions. The linear equation represents the relationship between the dependent and independent variables Is expressed as $Y= \beta_0+\beta_1 X_1+\beta_1 X_2+\cdots+\beta_n X_n$

where Y is the dependent variable, β0 is the y-intercept, β1, β2, ..., βn are the coefficients for the independent variables X1, X2, ..., Xn, respectively.

The coefficients are estimated using the Ordinary Least Squares (OLS) method, which minimizes the squared differences between observed and predicted values. Linear regression relies on assumptions such as linearity, independence, normality of residuals, and constant variance. Various evaluation metrics, such as MSE and R-squared, are used to assess the performance of the regression model. Linear regression is widely used in fields like economics, finance, and social sciences for data analysis and prediction.

Studying the relationship between the mpg variable and other variables in the Auto-mpg dataset is significant for several reasons. It helps understand the factors that impact fuel efficiency, enabling better vehicle design and optimization. By analyzing these relationships, predictive models can be built to estimate fuel efficiency for new vehicles. This information is valuable for manufacturers, consumers, and policymakers to make informed decisions. Additionally, studying the relationship informs efforts to improve fuel efficiency, reduce emissions, and shape policies and regulations related to vehicle standards and environmental impact.

2. Methods

First, identify the column type (Quantitative or Categorial):

Quantitative: mpg, cylinders, displacement, horsepower, weight and acceleration.

Categorial: model year (not suitable for mathematical calculations), origin and car name.

Notice that origin column value: 1 is a car made in America, 2 in Europe and 3 in other parts of the world

The steps are divided into Exploratory Data Analysis (EDA) and Inference.

2.1. Exploratory Data Analysis (EDA):

2.1.1. Data Cleaning: this step includes:

2.1.1.1. Remove null values by dropping the rows which have missing values. Horsepower is the only column that has missing values (delimited by "?").

2.1.1.2. Change the data type of the "horsepower" column to numeric by the "to_numeric()" method in Pandas.

2.1.1.3. Drop the "car name" column (has nearly no effect on our analysis) by the "drop" method in Pandas.

2.1.2. Data Visualization and Identifying Outliers: to deal with outliers, there are two options depending on the distribution of data:

- Normal Distribution: outliers are the points that have values that fall more than three standard deviations from the mean
- Other Distribution: the outlier data points are the ones falling below Q1–1.5 IQR or above Q3 + 1.5 IQR. Q1 is the 25th percentile and Q3 is the 75th percentile of the dataset, and IQR represents the interquartile range calculated by Q3 minus Q1 (Q3–Q1),

So, Knowing the distribution of each feature is necessary to objectively select the appropriate method of outlier detection through visualization (boxplot). The power of the boxplot appears in showing us outliers without the need for any calculations.

Based on the data visualization, it can be observed that the "acceleration" feature/column appears to follow a normal distribution. Furthermore, the "acceleration" and "horsepower" features/columns exhibit outliers. To objectively assess the normality of a feature/column, the Kolmogorov-Smirnov test can be conducted. This test can be implemented from scratch by utilizing the algorithm or by utilizing the "kstest()" built-in method provided in Scipy.stats Python library. Both methods output the same results.

By the visualization and the "Kolmogorov-Smirnov test", the "acceleration" column has outliers and is normally distributed and the "horsepower" column has outliers and is not normally distributed. So, the "acceleration" outliers are the points that have values that fall more than three standard deviations from the mean and the "horsepower" outliers are the points falling below Q1–1.5 IQR or above Q3 + 1.5 IQR. The outliers are calculated, and the corresponding records are removed using the "dropna()" method in the Pandas library.

2.1.3. Descriptive statistics: the method "describe()" in Pandas shows the descriptive statistics for each column in the dataset: Count of records, mean, standard deviation, minimum, 25%, 50%, 75% and maximum values.

The measures of central tendency are the mean and the median (50%). the measure of dispersion is the standard deviation. The Correlation Coefficient between 'mpg' (target) and other features is calculated from scratch:

$$r = \frac{cov(xy)}{\sqrt{Var(x)Var(y)}} = \frac{\frac{\Sigma(x_i-\bar{X})(y_i-\bar{Y})}{n-1}}{\sqrt{\frac{\Sigma(x_i \ \ X)^2}{n-1}}\sqrt{\frac{\Sigma(x_i \ \ X)^2}{n-1}}}$$

and using "corr()" the built-in method provided in Pandas. The results from both methods are very close.

A scatter plot is used to visually display the correlation between two continuous variables, target and continuous feature data. A bar plot is used to visually display the correlation between the target and discrete feature data. Matplotlib – which is a plotting library in python – provides the scatter and bar plot.

Features Standardization: range of values for each feature is different among different features, one way to solve this problem is to do features standardization.

$$Z = (X - \mu)/\sigma$$

After standardization, the data is ready for prediction.

2.2. Inference:

Linear regression coefficients (m, b) are calculated for each predictor individually.

m and b are calculated from scratch by the following equations:

$$\hat{Y} = m \times X + b_{\bar{Y}}$$

$$. m = \frac{cov(xy)}{Var(x)} = \frac{\frac{\Sigma(x_i-\bar{X})(y_i-\bar{Y})}{n-1}}{\frac{\Sigma(x_i-\bar{X})^2}{n-1}}$$

$$. b = \bar{Y} - m \times \bar{X}$$

. where

$$\bar{X} = \frac{\Sigma x_i}{n} \ \bar{Y} = \frac{\Sigma y_i}{n}$$

and by the “LinearRegression()” method from “sklearn” python library. The results from both methods are almost the same.

Y predicted is calculated by the above equation, the chosen metric to calculate the error between true and predicted data is (MAE) Mean Absolute Error.

The MAE calculated from Y predicted based on coefficients calculated in step ‘a’ almost equals 141.96.

The Baseline Model error is come from calculating the MAE of the product of the mean of the target multiplied by its length. It almost equals 6.42. The MAE value of the model must be less than this value to can be reliable.

Multivariable Linear Regression analysis:

$$\hat{Y}i = m \times Xi + b$$

$$y = m1 * X1 + m2 * X2 + m3 * X3 + m4 * X4 + m5 * X5 + m6 * X6 + m7 * X7 + b$$

Where:

y: is the dependent variable.

X1 - X7: the independent variable.

m1 - m7: are the regression coefficients.

b: is the intercept

Y predicted is recalculated by Multivariable Linear Regression analysis. The MAE of this calculation almost equals 2.46. That means this model performs with a lower error than that of the baseline model.

3. Results and Discussion

3.1. Values of Descriptive statistics (before standardization): see *Table 3.1*.

3.2. Correlation coefficients between the target and each feature: see *Table 3.2*.

3.3. Results of different linear regression methods: see *Table 3.2*.

3.4. Scatter plot of features of continuous data: see Figures 3.1 – 3.4.

3.5. Bar plot of features of discrete data: see Figures 3.5 – 3.7.

3.6. Histogram plot and box plot of each feature: see Figures 3.8 – 3.15.

3.7. Type of Distribution:

- mpg: continuous the truncated normal distribution.

- cylinders: discrete.

- displacement: continuous double log-normal distribution.

- horsepower: continuous right-skewed distribution.

- weight: continuous frequency distribution.

- acceleration: continuous normal distribution.

- model year: discrete.

- origin: discrete.

4. Conclusion

In this linear regression project on the Auto-mpg dataset, we analyzed the relationship between various attributes and the fuel efficiency (mpg) of vehicles. Here are the key findings and insights obtained from our analysis:

1. The linear regression model revealed significant relationships between the independent variables and the dependent variable (mpg). Features such as cylinders, displacement, horsepower, weight, and model year showed varying degrees of impact on fuel efficiency.

2. We identified that the number of cylinders, weight, and horsepower had a negative correlation with mpg, indicating that as these factors increased, the fuel efficiency decreased. On the other hand, the model year showed a positive correlation, suggesting that newer vehicles tend to have better fuel efficiency.

3. The scatter plot visualizations allowed us to observe the relationships between the variables and detect outliers in the acceleration and horsepower attributes. These outliers were treated by removing the corresponding records using the "dropna()" method in Pandas.

4. The limitations of our analysis include the assumption of linearity between the independent variables and mpg, as well as the potential influence of unmeasured factors that could impact fuel efficiency. Additionally, the presence of missing values in the horsepower attribute required handling with caution.

5. Further research could explore additional variables or interactions between attributes to enhance the accuracy of the linear regression model. Additionally, incorporating non-linear regression techniques or applying more advanced machine learning algorithms may provide better predictive performance.

Overall, our analysis on the Auto-mpg dataset provided valuable insights into the factors influencing fuel efficiency. These findings can be utilized by vehicle manufacturers, policymakers, and consumers to make informed decisions regarding vehicle design, optimization, and energy consumption.

5. Tables and figures:

Table 3.1

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
count	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000
mean	23.659474	5.413158	188.856579	101.686842	2942.426316	15.619211	76.084211	1.589474
std	7.657241	1.678778	99.149436	34.122383	826.694251	2.592484	3.631911	0.812108
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	70.000000	1.000000
25%	17.600000	4.000000	103.250000	75.000000	2220.000000	14.000000	73.000000	1.000000
50%	23.000000	4.000000	145.500000	92.000000	2764.500000	15.500000	76.000000	1.000000
75%	29.000000	6.000000	258.000000	120.000000	3542.000000	17.125000	79.000000	2.000000
max	46.600000	8.000000	429.000000	200.000000	5140.000000	23.700000	82.000000	3.000000

Table 3.2

Feature	**Correlation coefficient**	**Relationship/ Independency**
Cylinders	-0.7678	Strong negative relationship
Displacement	-0.8028	Strong negative relationship
Horsepower	-0.7881	Strong negative relationship
Weight	-0.8281	Strong negative relationship
Acceleration	0.3797	Moderate positive relationship
Model year	0.5565	Moderate positive relationship
Origin	0.5589	Moderate positive relationship

Table 3.3

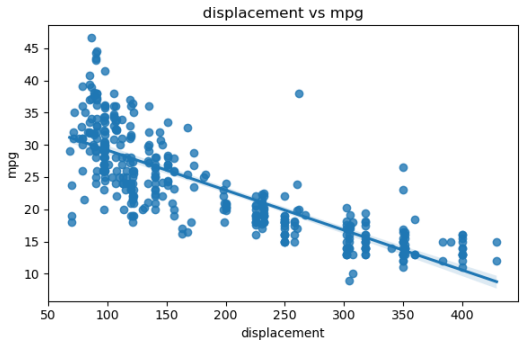
	LR from scratch / Built-in LR method	
Cylinders	-5.87944	23.65947
Displacement	-6.14739	23.65947
Horsepower	-6.03440	23.65947
Weight	-6.34122	23.65947
Acceleration	2.90759	23.65947
Model year	4.26089	23.65947

	Multivariable LR
$\mathbf{m_1}$	-0.381164
$\mathbf{m_2}$	1.494207
$\mathbf{m_3}$	-1.327211
$\mathbf{m_4}$	-4.907946
$\mathbf{m_5}$	-0.120290
$\mathbf{m_6}$	2.668320

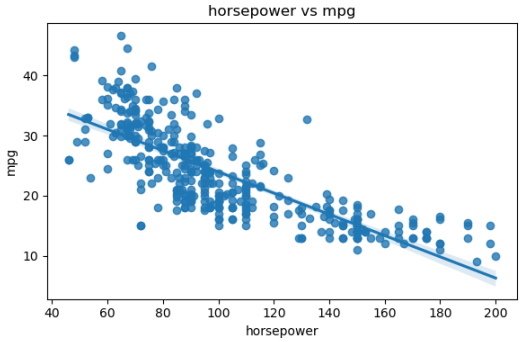
Origin	4.27980	23.65947

m_7	1.124575
b	23.659473684210525

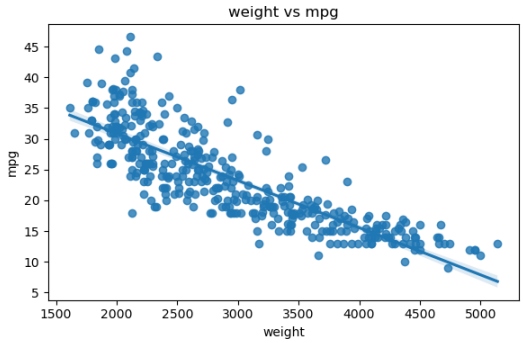
- Displacement: Figure 3.1



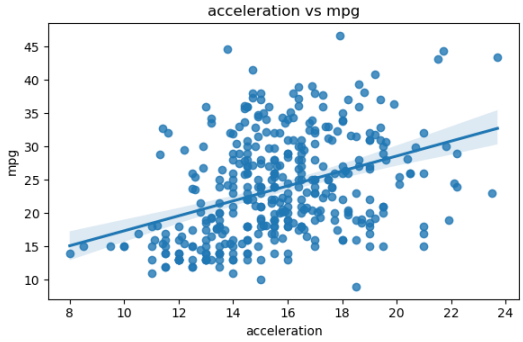
- Horsepower: Figure 3.2



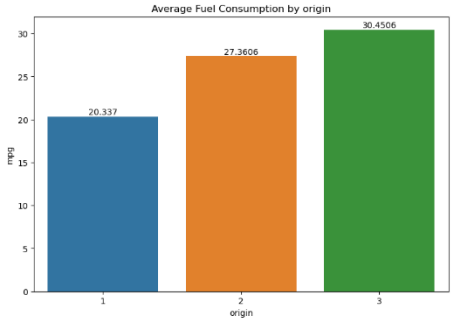
- Weight: Figure 3.3



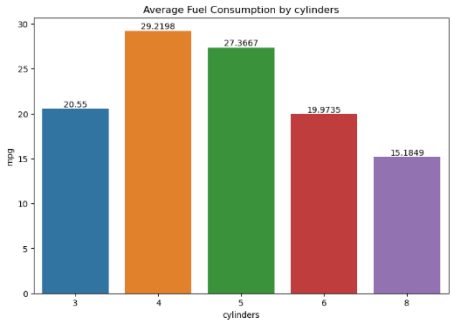
- Acceleration: Figure 3.4



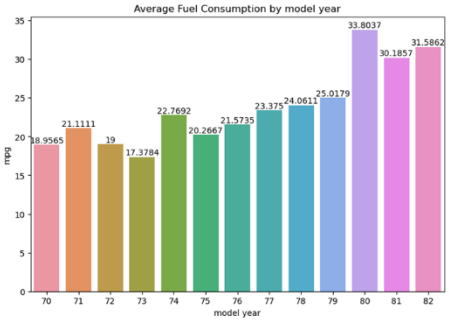
- Origin: Figure 3.5



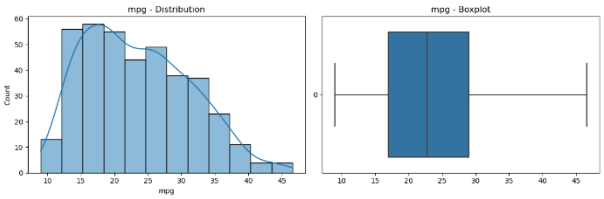
- Cylinders: Figure 3.6



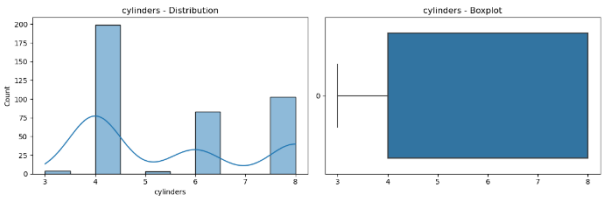
- Model year: Figure 3.7



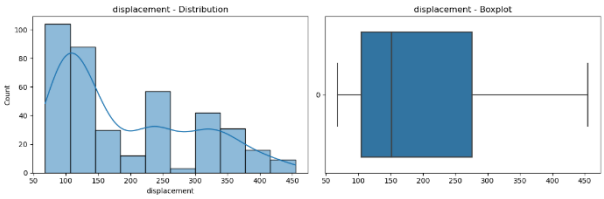
- Mpg: Figure 3.8



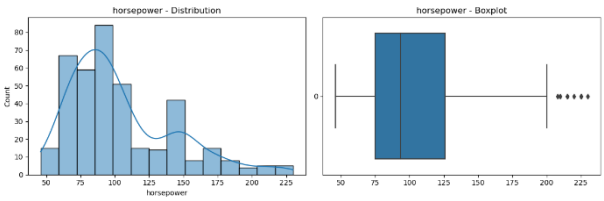
- Cylinders: Figure 3.9



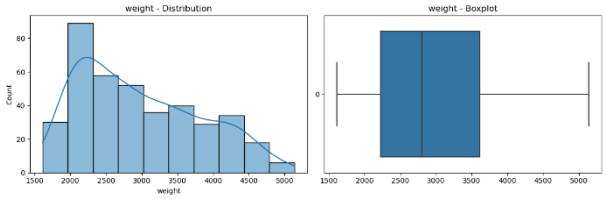
- Displacement: Figure 3.10



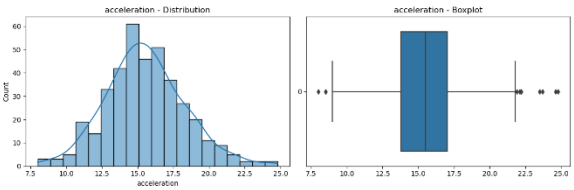
- Horsepower: Figure 3.11



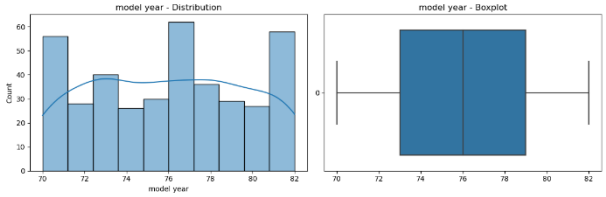
- Weight: Figure 3.12



- Acceleration: Figure 3.13



- Model year (not suitable for mathematical calculations): Figure 3.14



- Origin: Figure 3.15