SPEED-DISTANCE REGRESSION

Overview

This report presents a machine learning regression model that predicts the car's braking distance (dependent variable) given the speed (dependent variable). The model was trained on a dataset of 50 cases. The model was then evaluated using Mean Squared Error with value of 238.87 and R^2 value of 62.3% .

Data

The data comprises the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s.

Results

The model was not able to capture the variability of the data and should be rejected from application.

Conclusion

The results of this study suggest that regression cannot be used to effectively predict the braking distance of a car. The regression model should be used with more recent data with more features.

Exploratory Data Analysis

Raw data consists of 50 samples and 2 features with no missing values.

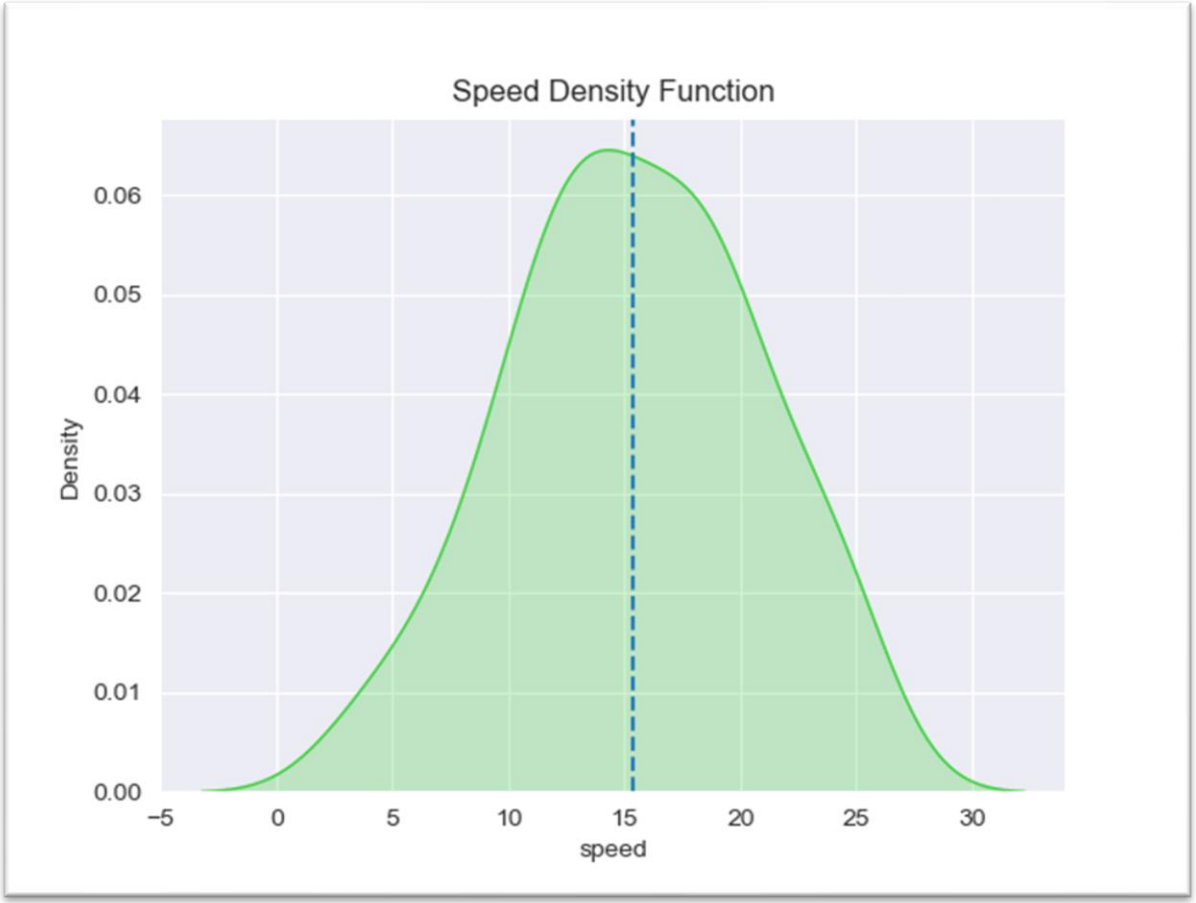
Features

Feature name	Description	Data type
speed	Car Speed	Decimal
dist	Stopping distance	Decimal

Univariate analysis

Descriptive statistics

MEASURE	SPEED	DIST
COUNT	50	50
STANDARD DEVIATION	5.28	25.76
FIRST QUARTILE	12	26
MEDIAN	15	36
MEAN	15.4	42.98
THIRD QUARTILE	19	56
MIN	4	2
MAX	25	120



Speed density function has skewness of -0.11 and kurtosis of -0.57.

Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images. A distribution can have right (or positive), left (or negative) skewness, or no skewness (also known as symmetry).

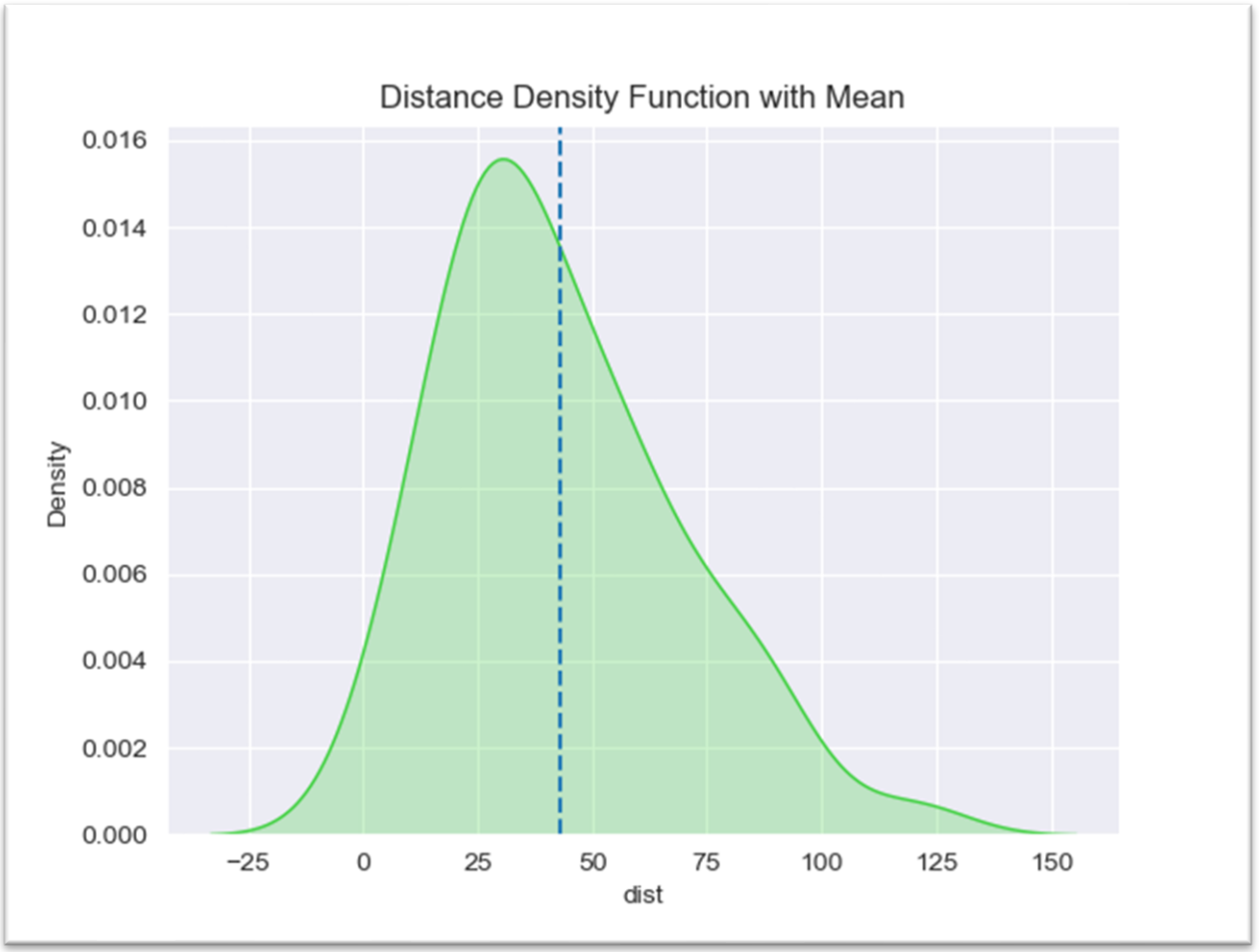
Here, skewness is negative, meaning density is slightly curved to the left, possible outliers with small values.

Kurtosis is a measure of the "tailedness" of a distribution. A distribution with a high kurtosis has a sharp peak and fat tails. A distribution with a low flat peak and thin tails. A distribution with a kurtosis of zero is said to have peak of medium height and tails of medium thickness**. Kurtosis of -0.57 indicates small tails and close to no outliers.**

Distance density function has skewness of 0.78 and kurtosis of 0.24.

Here, skewness is highly positive, meaning curved to the right. It has outliers of great values.

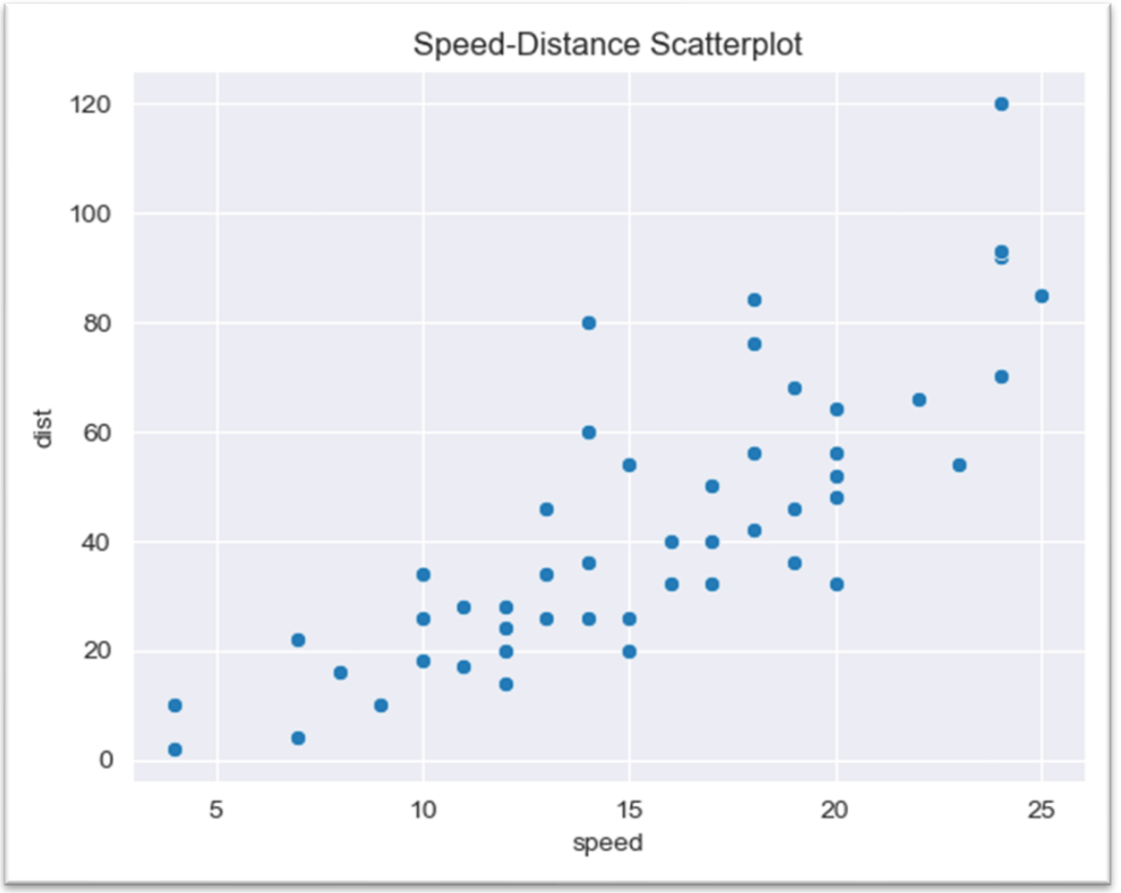
Kurtosis of 0.24 indicates tails on both sides and presence of outliers.



Bivariate Analysis

By looking at scatterplot, there is positive clearly correlation between variables. **Pearson's coefficient** indeed shows **0.81**, suggesting there is strong positive correlation.

However, as speed increases, distance variability grows and becomes unpredictable. It is possible that this fact can obstruct regression line from capturing actual pattern in the data,



Covariance	*speed*	*dist*
speed	27.95	109.94
dist	109.94	664.06

Methodology

Regressor

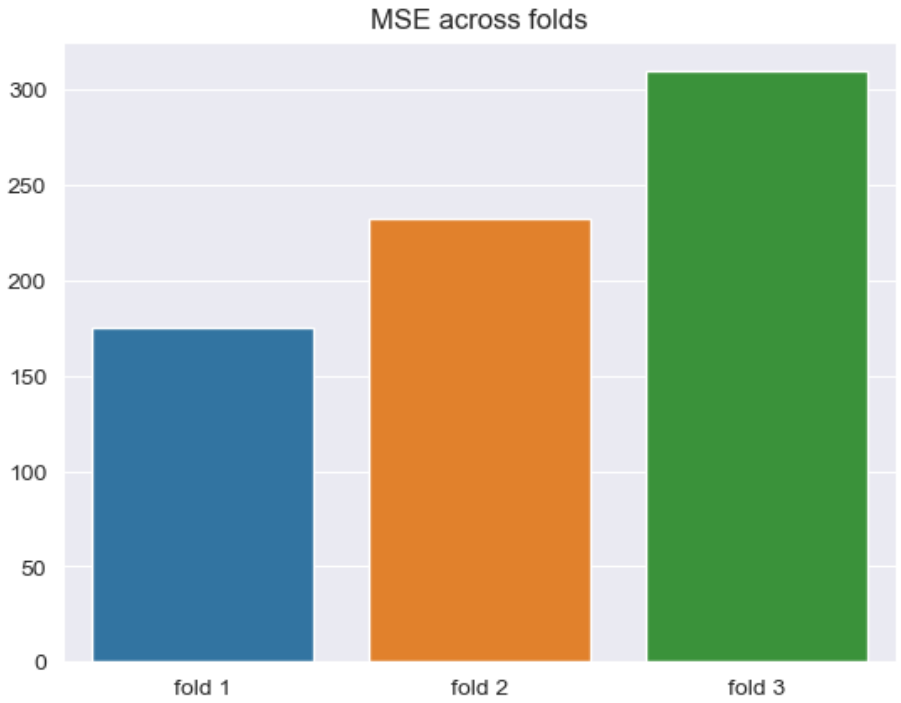
For this task, a **Linear Regression** regressor is used. Data points follow a straight line, and Linear Regressor plots a line that fits the data, which is what is needed for this problem.

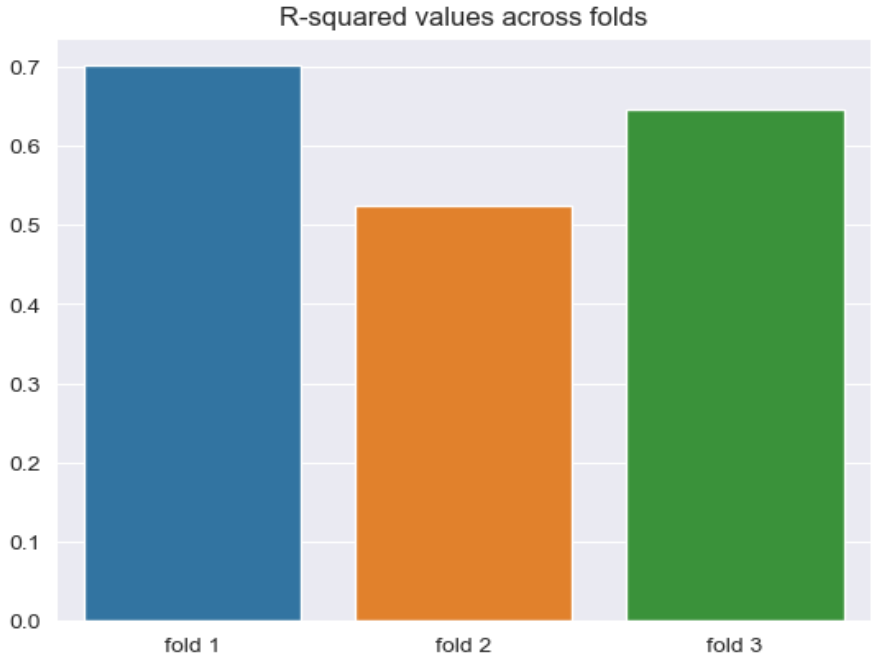
Results

The model was cross validated 3 fold. Cross validation splits the data into folds, or partitions. The model is then trained on two of the folds and tested on the third fold. This process is repeated three times, with each fold being used as the test set once. Then, the average of the results from the three iterations is used to estimate the performance of the model.

After performing 3 fold cross validation, these were the results:

3 folds	Mean sqared error	R^2
Average	238.87	0.62
Standard Deviation	55.12	0.073





Result interpretation

Results we re-fitted to the whole dataset for the sake of interpretation.

Simple linear regression line formula:

$$y = \alpha + \beta x$$

y – dependent variable Distance

x -independent variable speed

a – intercept, where regression line meets the dependent variable

β – coefficient, if x moves by one point, y moves by this value

Our model estimated the best fitted line to have β coefficient of 3.923 and intercept of -17.579

Regression formula:

$$\boldsymbol{y = -17.579 + 3.923x}$$

Residuals

For each individual observation i that deviates from regression line there is an error value ε_i, called ***residuals***. The formula for actual values becomes:

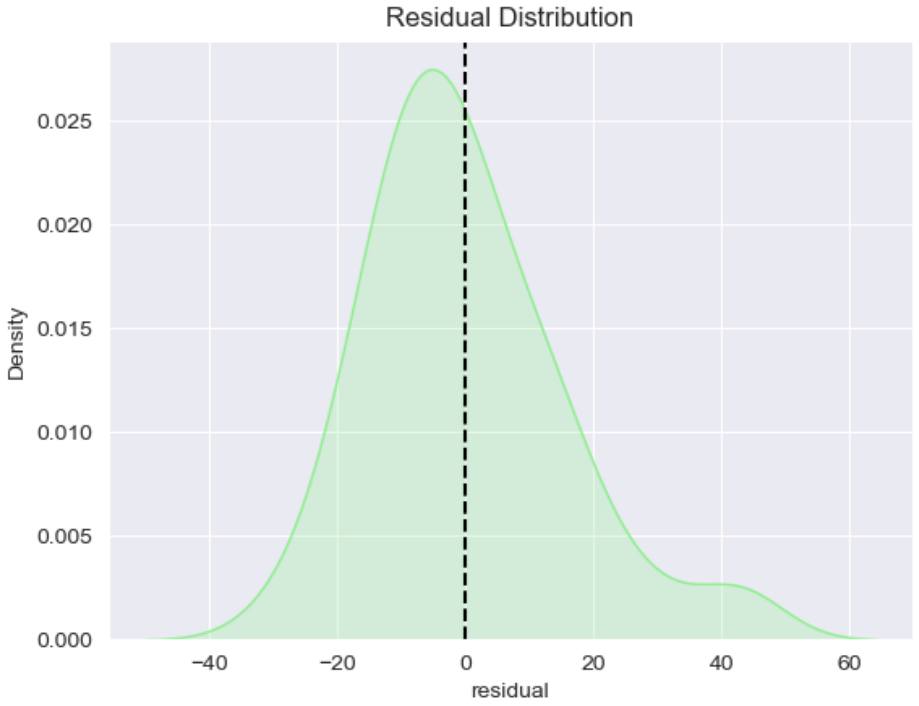
$$y = \alpha + \beta x + \varepsilon_i$$

The goal is to find such a line that gives the lowest residual values. Function for residuals is called Least Squares:

$$\sum_{i=1}^{n} \varepsilon_i^2$$

The goal of linear regression is to fit the line in a way that Least Squares is as low as possible. Residuals should ideally be close and centered around 0 without much variation.
Applying regression to the entirety of data gives 50 residuals that we can analyze.

The residual distribution function has skewness of 0.88 and kurtosis 0.89.



Residuals are heavily skewed to the right,
implying data had too much variation to be captured with simple linear regression.

However, right skewness shows the model more frequently predicted Braking Distance at lower value than it is (hence the positive residuals). Regarding the problem of *predicting braking distance*, predicting lower value than real value has worse consequence than predicting braking distance to be larger.

Based on that, model should be rejected for application

Model summary

Dep. Variable:	y	R-squared:	0.651
Model:	OLS	Adj. R-squared:	0.644
Method:	Least Squares	F-statistic:	89.57
	Prob (F-statistic):	1.49e-12	
	Log-Likelihood:	-206.58	
	No. Observations:	50	
	Df Residuals:	48	

R^2

Total sum of squares (TSS) is a value representing variance of the dependent variable. $\sum_{i=1}^{n}(y-\bar{y})^2$ where $\bar{y}$ is mean value of y. Residual sum of squares (RSS) is the same as Least Squares function

R^2 is represented by value of $1 - \frac{RSS}{TSS}$. It ranges from (-inf, 1] but is usually in range [0,1]. R squared value 1 of means that all the movements of a dependent variable are completely explained by movements of independent variable.

We have a value of 0.651, R squared value of 0.651 means that Speed variable explains about 65.1% of variability of Distance variable.

F-statistic

F-statistics is a statistic used to test the significance of regression coefficients. It is calculated as

$$\frac{STANDAR\ ERROR}{RESIDUAL\ STANDAR\ ERROR},$$

where standard error is $\frac{TSS}{dF}$ and residual standard error is $\frac{RSS}{df}$. DF is number of independent variables and df is sample size – number of independent variables – 1. F-statistic is used for hypothesis testing.

- H0: there is no relationship between any of the independent variables,
- H1: there is a relationship between at least one independent variable.

p – value of this statistic is 1.49e-12. If we chose significance level of 1%, we reject null hypothesis. **There is a relationship between dependent variable and at least one independent variable.** F – statistic value is also used to compare regression models and feature selection. Model with larger F – statistic is better (not p -value).

Intercept and coefficient t-test

Statistic test to determine whether the independent variable (and intercept) has significant impact on independent variable.

	Value	std err.	t	P > | t |
intercept	-17.5791	6.758	-2.601	0.012
x1	3.9324	0.416	9.464	0.000

- H0: there is no relationship between this variable and dependent variable,
- H1: there is a relationship between this variable and dependent variable.

$$std\ err. = \frac{std}{\sqrt{n-2}},\ \text{n = 50,}$$

$$t = \frac{value}{std\ err.}$$

We perform 2-tailed t-test for 48 degrees of freedom. If the p – value of t-statistic is lower than significance level, H0 is rejected. If we choose significance level 2%, meaning there is 2% chance we made a mistake. P value of intercept is 1.2% and for speed it is < 0.0%

We can reject null hypothesis for both. **There is a relationship between both independent variables and the dependent variable.**

Implementation

The model was built in Python programming language.

Libraries used:

- Pandas – data manipulation and analysis,
- Numpy – linear transformation and matrices,
- Seaborn, matplotlib – data visualization,
- Scikit-learn – machine learning,
- Statsmodels for model summary