Report on analysis of the cars dataset - Linear regression

“Milica Djurkovic”

2023-08-27

Introduction

In this report, we delve into the utilization of linear regression for data modelling pertaining to a car’s stopping distance. Our objective is to develop a model capable of predicting stopping distances based on a car’s speed. We employ the “Cars” dataset containing information about car speeds and their corresponding stopping distances. Through the application of linear regression, we will analyze statistical relationships between speed and stopping distance, with the aim of constructing a model that can efficiently predict stopping distances based on predefined speeds.

The subsequent sections of this report will encompass data collection, analysis, the development and evaluation of the linear regression model, as well as a discussion of the results and their significance.

Data Exploration

After loading the dataset, speed values are converted from miles per hour (mph) to kilometres per hour (km/h) using a conversion factor of 1.60934. Similarly, stopping distance values are converted from feet to meters using a conversion factor of 0.3048.

The resultant cars_data data frame contains the preprocessed data with appropriate units and selected columns. This preprocessed dataset will serve as the foundation for our subsequent analysis.

Correlation analysis

A key aspect of our analysis is understanding the correlation between speed and stopping distance. Correlation quantifies the strength and direction of the linear relationship between two variables. We used the cor() function to calculate the correlation coefficient between speed (in km/h) and stopping distance (in meters).

The correlation coefficient is approximately 0.807. This positive value suggests a strong positive linear correlation between speed and stopping distance.

```
## [1] 0.8068949
```

Development of Linear Regression Model: Model Fitting and Interpretation

Coefficients and Significance: The coefficients section provides crucial insights into the relationships between the variables. The coefficient for the intercept (Intercept) is estimated to be approximately -5.3581, with a standard error of 2.0600. The coefficient for the variable speed_kmh is estimated to be approximately 0.7448, with a standard error of 0.0787.

Model Fit and Predictive Power: The model’s goodness of fit is assessed through various metrics. The Multiple R-squared value of 0.6511 indicates that around 65.11% of the variability in the stopping distance can be explained by the linear relationship with speed. The Adjusted R-squared value of 0.6438 is a version of the R-squared metric that considers the number of predictors in the model.

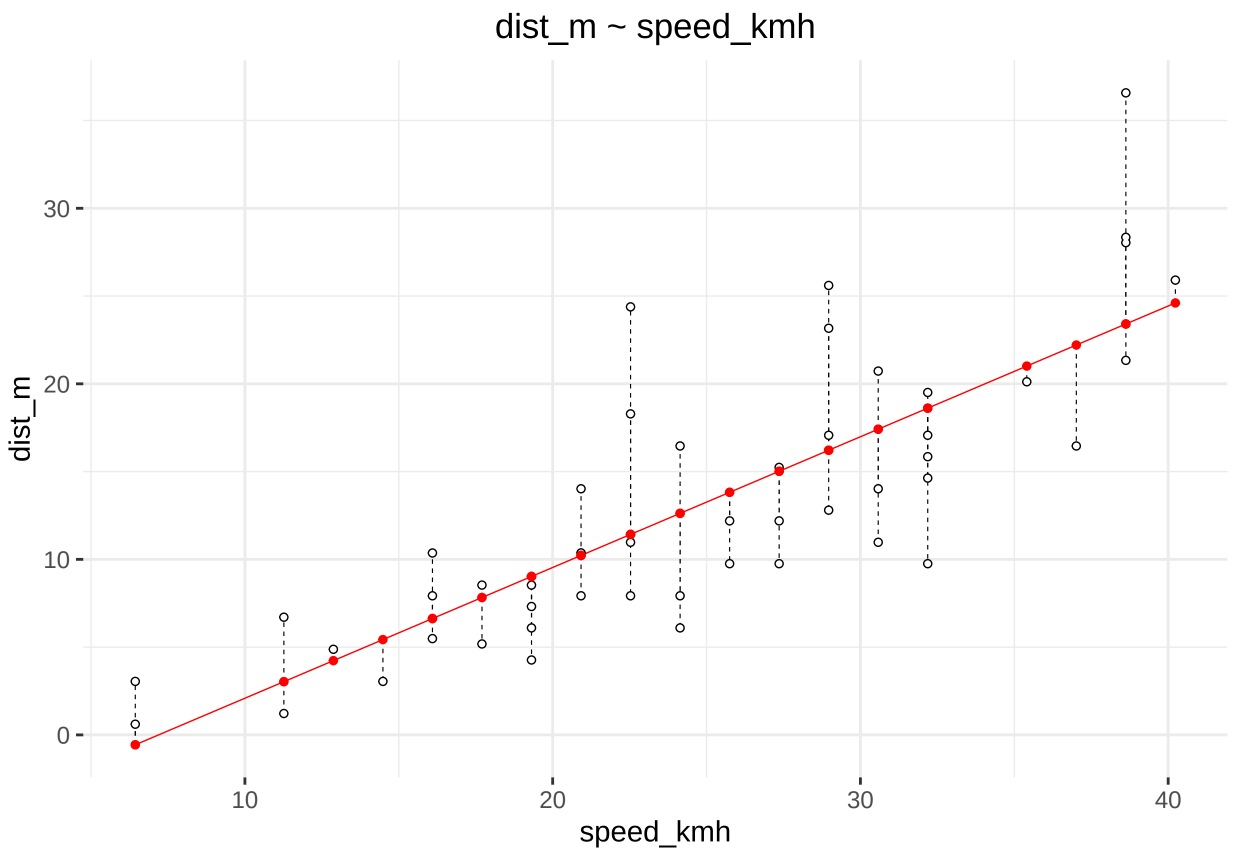
Residual Analysis: The residuals, representing the differences between observed and predicted values, exhibit a distribution with a minimum value of -8.8603 and a maximum value of 13.1678. The interquartile range (IQR) is between -2.9033 and 2.8086. These values provide an insight into the distribution of errors in our model's predictions.

Model Significance: The F-statistic, with a value of 89.57, assesses the overall significance of the model. The corresponding p-value (1.49e-12) indicates that the model is highly significant in explaining the variability in stopping distances.

```
##
## Call:
## lm(formula = dist_m ~ speed_kmh, data = cars_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8603 -2.9033 -0.6925  2.8086 13.1678
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.3581     2.0600  -2.601   0.0123 *
## speed_kmh     0.7448     0.0787   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.688 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Visualizing the Linear Regression Model

The visualization of the linear regression model underscores the correlation between speed and stopping distance.



Pearson’s product-moment correlation

Pearson’s correlation coefficient is a statistical measure that measures the linear correlation between two numerical variables. In our case, Pearson’s correlation coefficient is 0.8068949, indicating a strong positive linear correlation between car speed and stopping distance. A higher value of the coefficient indicates a stronger linear relationship between these variables.

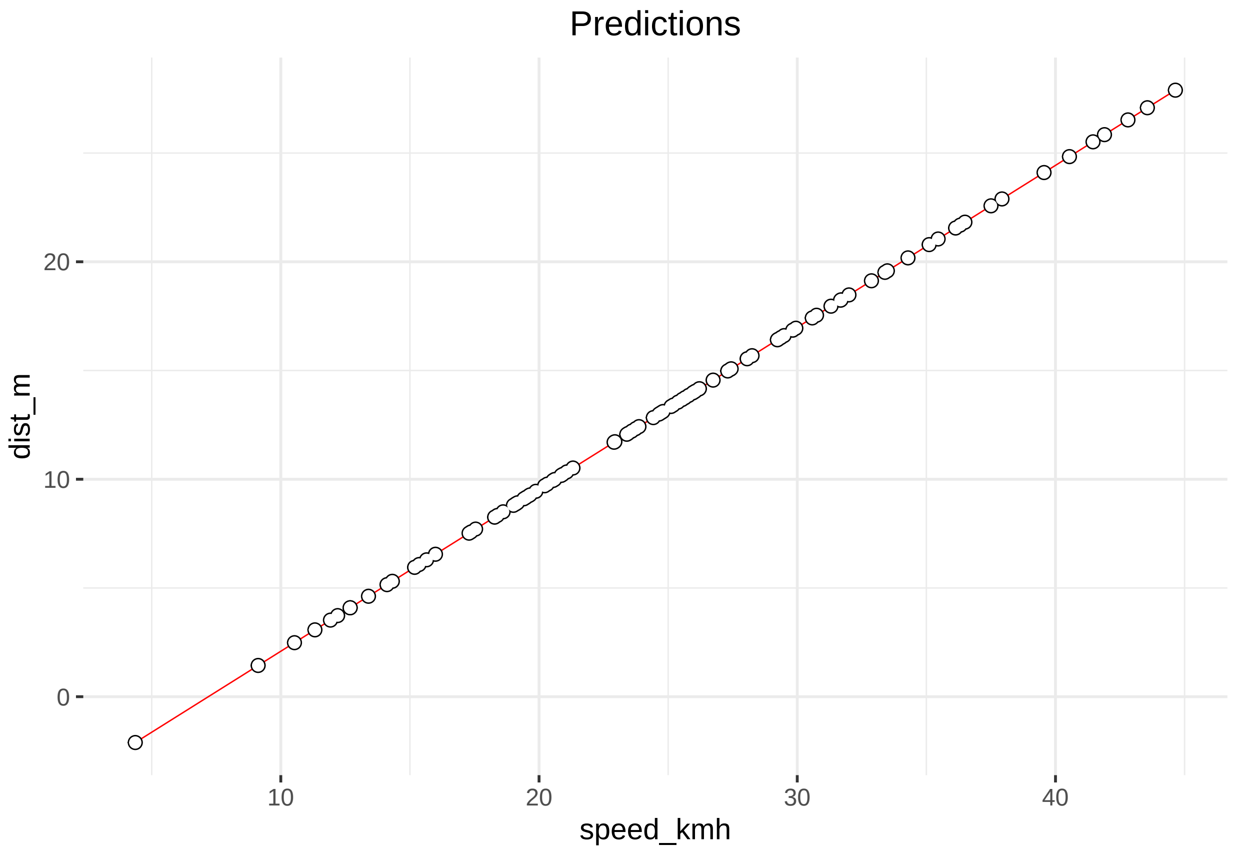
```
## [[1]]
##
##  Pearson's product-moment correlation
##
## data:  data$speed_kmh and data$dist_m
## t = 9.464, df = 48, p-value = 1.49e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6816422 0.8862036
## sample estimates:
##       cor
## 0.8068949
##
##
## [[2]]
##       cor
## 0.8068949
##
```

```
## [[3]]
##       cor
## 0.6510794
```

Predicting New Data Using the Linear Model

We make predictions based on new velocity data that is generated using a normal distribution with mean and standard deviation, this gives us insight into how the new data would behave according to the model.

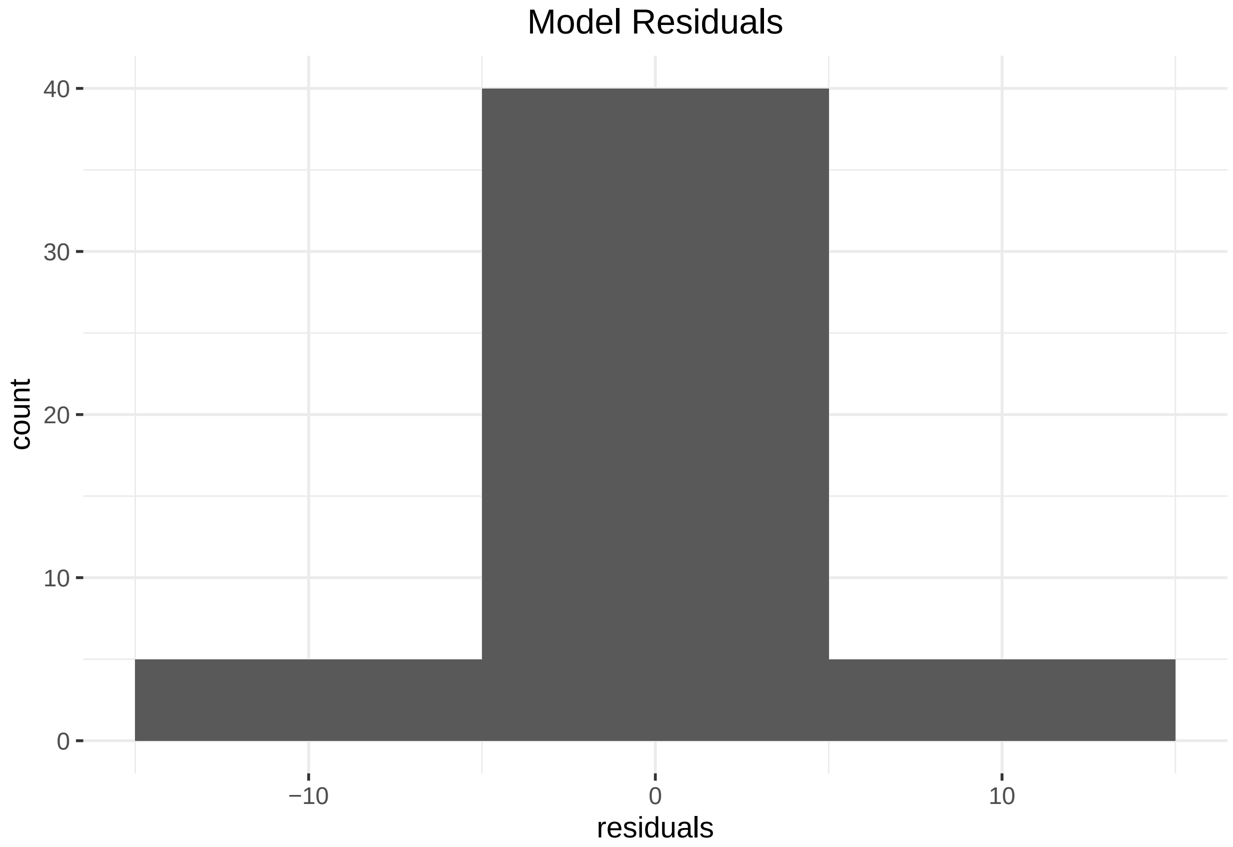
So, we can conclude that the previously constructed linear regression model predicts stopping distances based on vehicle speed relatively well. The red line of the model fit follows the general trend of the data, suggesting that the model provides a reasonably good approximation of the actual relationships between speeds and stopping distances.



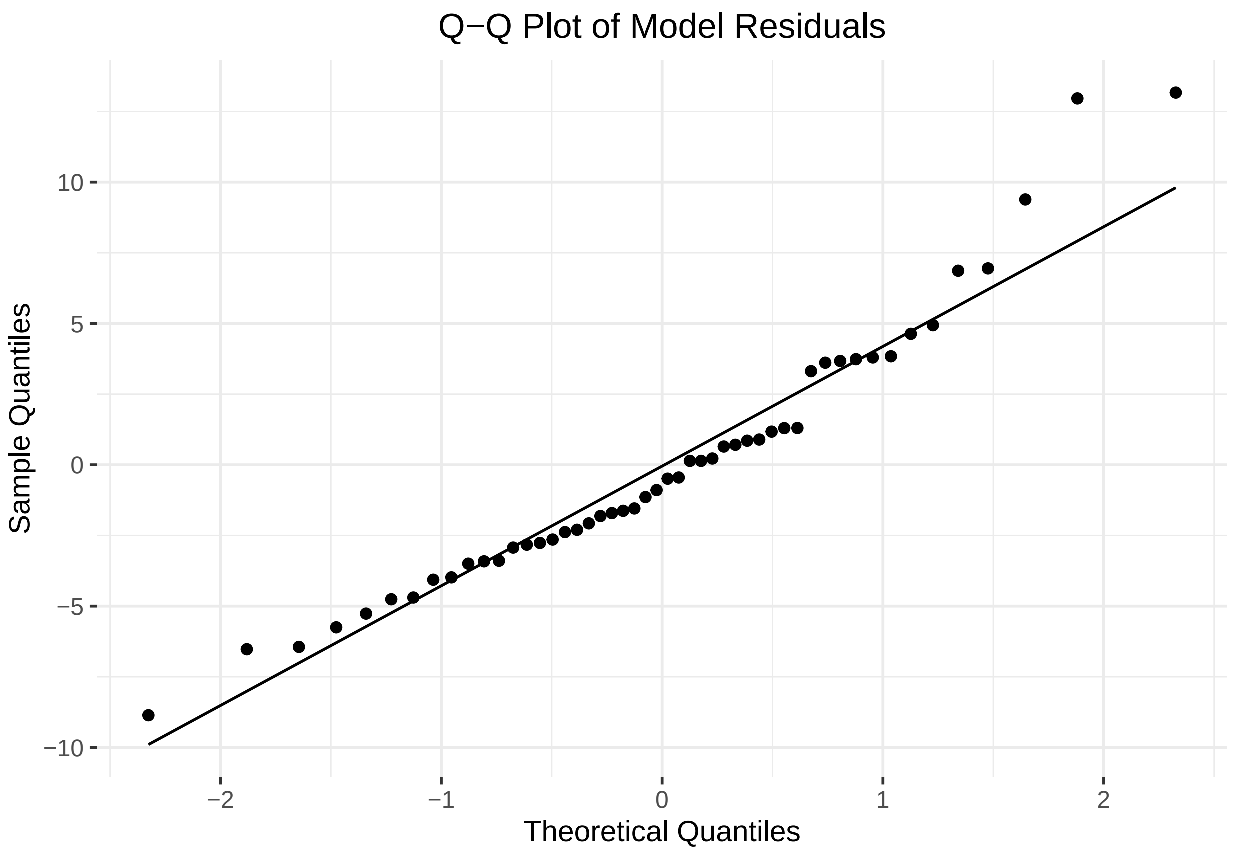
Analysis of the Distribution of Residuals and QQ Plots

Based on these graphs, we can conclude that:

The distribution of residuals appears fairly symmetric and centred around zero, suggesting that the model describes the variation in the data well.



The QQ plot shows a straight line shape, suggesting that the residuals follow a normal distribution. This is a positive sign because it means that the model makes a good assumption about the normality of the residuals.

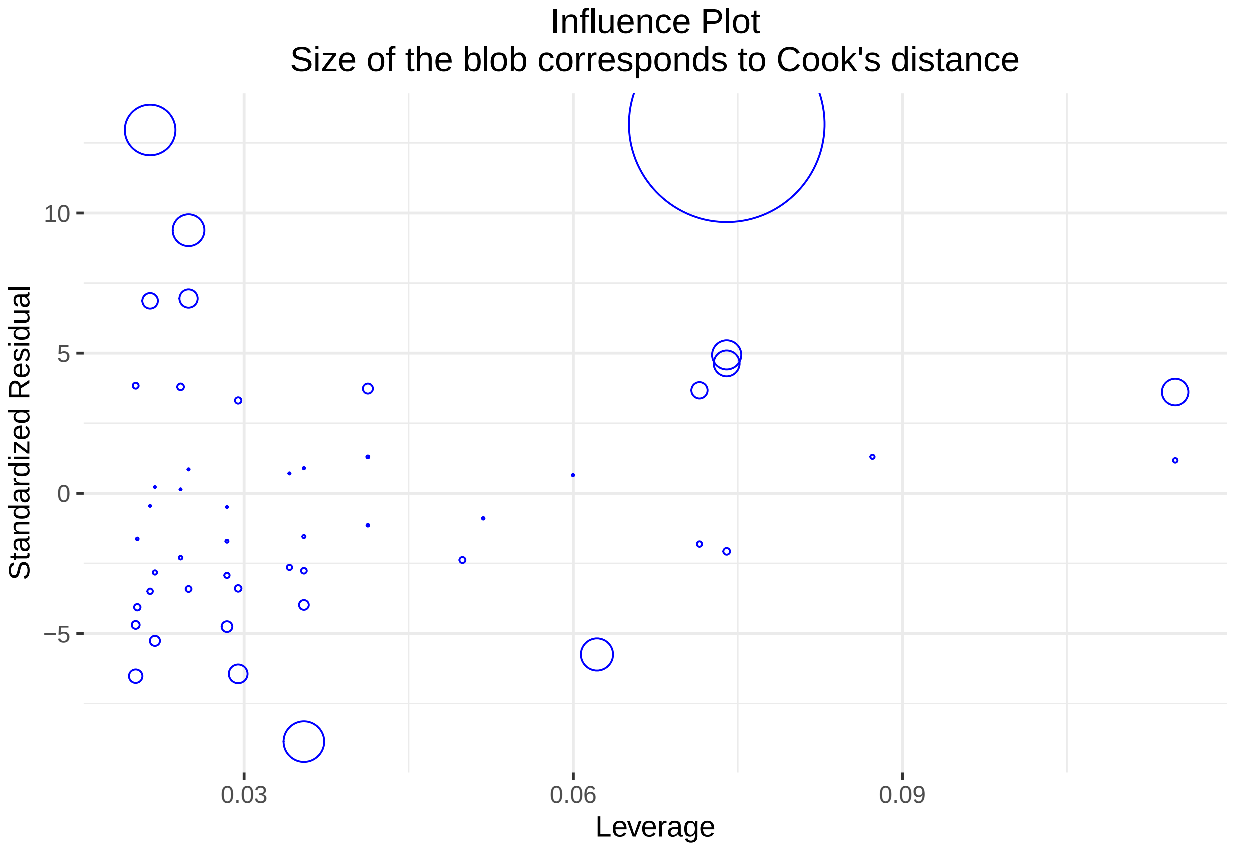


Identify influential cases using Cook’s distance

According to Cook’s distance, there are no extreme points of influence, but there are certain points with high leverage, which plot_influence shows.

```
## integer(0)
```

Plot influence



Conclusion

The overall conclusion is that the linear model provides a relatively good basis for predicting the stopping distance based on vehicle speed. However, there are certain influence points that may require additional analysis.