Simple and Multiple Linear Regression in R

Claudio Mazzi
Department of Computer Science, University of Pisa
MeS Laboratory, Sant'Anna School for Advanced Studies, Pisa
claudio.mazzi@santannapisa.it

Applied Statistical Modelling 1
A.A. 2024-2025

Contents

1 Overview on R-programming language

R is a programming language used for statistical computing and data visualization. It is widely adopted in fields such as data mining, bioinformatics, and data analysis. The core R language is enhanced by a large number of extension packages, which include reusable code, documentation, and example data. R software is open-source and freely available. It is licensed under the GNU General Public License and is part of the GNU Project. It is primarily written in C, Fortran, and R itself. Precompiled executables are available for various operating systems. As an interpreted language, R has a native command-line interface. Additionally, several third-party graphical user interfaces are available, such as RStudio, an integrated development environment, which is the one we adopted in this course.

1.1 Download R and RStudio

Both R and the RStudio environment are free to download at the following links:

- R version 4.4.1: `https://www.r-project.org`
- RStudio: `https://posit.co/downloads/`

Both are supported by different platforms: Windows, Macintosh, Linux and other UNIX versions.

1.2 Syntax tips

The help() function in R is used to access the documentation and detailed information about specific functions, datasets, or packages. It allows users to look up descriptions, usage, arguments, examples, and other relevant details about an R function or object directly within the R environment.

Listing 1: *help* function

```
# Obtain information about cbind()
help(cbind)
```

This will open the help page for cbind(), displaying details such as usage, arguments, return value, and examples. As shown in 1.2, in the R environment, use the symbol # to add a comment. Well-commented code is essential as it enhances both shareability and comprehensibility, especially in collaborative work environments.

2 Introduction

In this exercise, we will conduct a comprehensive regression analysis using the R programming language. The dataset we will use contains information on CO_2 emissions and technical features of a set of vehicles. The goal is to model the relationship between emissions and different independent variables such as fuel consumption, vehicle class, engine size, etc, using both simple and multiple linear regression. The database is open source, and it is available for free download at `https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles`.

3 Installation of Packages

Before proceeding with the analysis, ensure the required R packages are installed. If not, run the following commands to install the necessary libraries.

Listing 2: Installation of Required Packages

```
# Install the necessary packages
install.packages("ggplot2")     # For plotting
install.packages("dplyr")       # For data manipulation
install.packages("summarytools") # For data overview
install.packages("corrplot")     # For data correlation
install.packages("car") # For regression diagnostics
install.packages("tidyverse") # Data processing
install.packages("psych") # Pairing variables
```

3.1 Loading the Libraries

Once the packages are installed, you can load them as follows:

Listing 3: Loading Libraries

```
# Load the necessary libraries
library(ggplot2)
library(dplyr)
library(summarytools)
library(corrplot)
library(car)
library(tidyverse)
library(psych)
```

4 Loading and Exploring the Dataset

In this section, we will load the dataset into RStudio, perform a quick exploratory analysis, and clean the data if necessary. Exploratory data analysis (EDA) is crucial, as it helps uncover underlying patterns, detect anomalies, and make sense of the dataset's structure. Even more critical is the data preparation

and cleaning process, which ensures the dataset is free from inconsistencies, missing values, and errors. A well-prepared dataset is the foundation of any reliable statistical analysis or predictive model, as it directly impacts the accuracy and validity of the results.

4.1 Loading the Dataset

We load the dataset using the `read.csv` function. Make sure the file path is correct.

Listing 4: Loading the Dataset

```
# Load the dataset
dataset <- read.csv("path_to/CO2_Emissions_Canada.csv")

# View the first few rows of the dataset and its dimension RxC
head(dataset)
dim(dataset)
[1] 7386   13
# Show the summary of each column
summary(dataset)
# Show the internal structure of each column
str(dataset)
```

	Make	Model	Vehicle.Class	Engine.Size.L.	Cyl...	Transm...	Fuel....	Fuel.Con.(L.100.km)	Fuel.Con.(L.100.km)	Fuel.Con.(L.100.km)	Fuel.Con.(mpg)	CO2.Emi(g.km)	CO2.Emi(kg.km)
1	ACURA	ILX	COMPACT	2.00	4	AS5	Z	9.90	6.7	8.50	33	196	0,196
2	ACURA	ILX	COMPACT	2.40	4	M6	Z	11.20	7.7	9.60	29	221	0,221
3	ACURA	ILX HYBRID	COMPACT	1.50	4	AV7	Z	6.00	5.8	5.90	48	136	0,136
4	ACURA	MDX 4WD	SUV - SMALL	3.50	6	AS6	Z	12.70	9.1	11.10	25	255	0,255
5			SUV - SMALL		0				xxxx				0
6	ACURA	RDX AWD	SUV - SMALL	3.50	6	AS6	Z	12.10	8.7	10.60	27	244	0,244

Table 1: Table of the first six rows of the database, provided by the command `head(dataset)`.

The dataset, as shown in Table 4.1, provides detailed information on how a vehicle's CO_2 emissions vary based on different features. The dataset was sourced from the official open data website of the Government of Canada and represents a compiled version spanning seven years. It contains a total of 7,387 rows and 13 columns.

Vehicle Features:

- 4WD/4X4: Four-wheel drive
- AWD: All-wheel drive
- FFV: Flexible-fuel vehicle
- SWB: Short wheelbase
- LWB: Long wheelbase
- EWB: Extended wheelbase

Transmission Types:

- A: Automatic
- AM: Automated manual
- AS: Automatic with select shift
- AV: Continuously variable
- M: Manual
- 3 - 10: Number of gears

Fuel Types:

- X: Regular gasoline
- Z: Premium gasoline

- D: Diesel
- E: Ethanol (E85)
- N: Natural gas

Dataset Variables:

- **Make**: Manufacturer of the vehicle
- **Model**: Specific car model
- **Vehicle Class**: Class of the vehicle, based on utility, capacity, and weight
- **Engine Size (L)**: Engine size in liters
- **Cylinders**: Number of cylinders in the engine
- **Transmission**: Transmission type and number of gears
- **Fuel Type**: Type of fuel used
- **Fuel Consumption (City)**: Fuel consumption on city roads (L/100 km)
- **Fuel Consumption (Hwy)**: Fuel consumption on highway roads (L/100 km)
- **Fuel Consumption (Comb)**: Combined fuel consumption (55% city, 45% highway) in L/100 km
- **Fuel Consumption (Comb g/km)**: Combined fuel consumption in both city and highway, expressed in grams per kilometres (g/km)
- **Fuel Consumption (Comb kg/km)**: Combined fuel consumption in both city and highway, expressed in kg per km (km/kg)
- **CO_2 Emissions (g/km)**: RESPONSE VARIABLE

4.2 Pre-processing data: cleaning and preparation

We need to check for missing values, redundant columns, and duplicate rows to ensure the data is ready for analysis. Our database presents a few missing values, and then we have to remove the associated rows. Moreover, the column "Fuel Consumption (Comb kg/km)" gives no further information than "Fuel Consumption (Comb g/km)", therefore we delete it to avoid *multicollinearity*. Finally, we remove duplicated rows, which are redundant for the analysis.

Listing 5: Pre-processing dataset

```
# Check for missing values in the dataset
sum(is.na(dataset))
[1] 5
# Remove rows with missing values
dataset_clean <- na.omit(dataset)
# Remove columns that are not useful --> NA's rows and redundant columns
dataset_clean <- dataset_clean %>% select(-c(CO2.Emissions.kg.km.))
# Removing duplicate rows
dataset_clean <- dataset_clean[!duplicated(dataset_clean), ]

# Rename the dataset
db_co2 <- dataset_clean
```

4.3 Exploratory Data Analysis

Exploratory data analysis helps us understand the structure of the data and spot any potential issues. It is useful to make some a-priori hypothesis on the behavior and the relations concerning the response variable (outcome) "CO_2 Emissions (g/km)". We perform a graphic correlation analysis using the R functions `pairs.panels()` and `cor()`, embedded in the corresponding libraries `psych`, and `corrplot`.

Listing 6: EDA: Correlation Plot

```
# scatter plot for quantitative variables
pairs.panels(db_co2[,colnames(qt_var)],
             method = "pearson", # Correlation function
             hist.col = "red",
             density = TRUE, # Show density plots
             ellipses = TRUE # Correletion allipses
             )
```

The correlation plot in Figure 1 provides a comprehensive visual representation of the relationships between the numerical variables in the dataset. The plot includes scatterplots, correlation coefficients, histograms, and ellipses that illustrate the nature and strength of the relationships between the variables. The scatterplots in the lower-left part of the matrix show the pairwise relationships between the variables, with each point representing a data observation. A red regression line is also included in each scatterplot to highlight the general trend. In the upper-right section of the plot, we see the Pearson correlation coefficients, which provide a numeric measure of the linear relationship between the variables. A correlation value close to 1 suggests a strong positive correlation, meaning that as one variable increases, so does the other. Conversely, a correlation close to -1 indicates a strong negative relationship. The histograms display the distribution of each variable, showing the spread and shape of the data. Overlaid on each histogram is a density curve, which measures whether the data is normally distributed or skewed. Additionally, the scatterplots include correlation ellipses, which visually convey the strength and direction of the correlation between two variables. Narrow ellipses correspond to stronger correlations, while wider ellipses suggest weaker relationships. Ellipses that slope upwards indicate positive correlations, while downward-sloping ellipses show negative correlations. This correlation plot is a valuable tool for understanding the interdependencies within the dataset and provides a solid foundation before further analysis.

Now we are interested in studying the correlation between variables. By default `cor()` computes correlation using the Pearson coefficient, but it allows also the use of Spearman or Kendall correlation definitions.

Listing 7: Correlation matrix

```
# Correltion matrix between quant. variables
corr_matrix <- cor(db_co2[, colnames(qt_var)], use = "complete.obs")
corrplot(corr_matrix, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```

5 Simple Linear Regression

In this section, we provide the proper Linear Regression analysis. We will focus on the relation between *CO_2 emissions* and *Combined Fuel Consumption in both city and highway* as they seem to be highly positively correlated, as shown in Figure 1.

Before proceeding with regression (or other statistical models), normalization is important to meet model assumptions, particularly the normality of the residuals and the linearity of the relationship between variables. For our specific database, this step is not crucial due to the good behavior of the response variable, as shown in the histogram in Figure 3. Otherwise one can use the function `bestNormalize()`, in `library(bestNormalize)`, for normalization transformations. It is important, indeed, to standardize all the numerical variables. It allows for the comparison of variables that are on different scales, ensuring that each variable has an equal impact on models. Additionally, it improves the performance of algorithms that are sensitive to the scale of the data, such as regression or clustering. In R, we use:

Listing 8: Standardization

```
# Standardization of numeric variables
db_co2 <- as.data.frame(scale(db_co2[, colnames(qt_var)]))
```

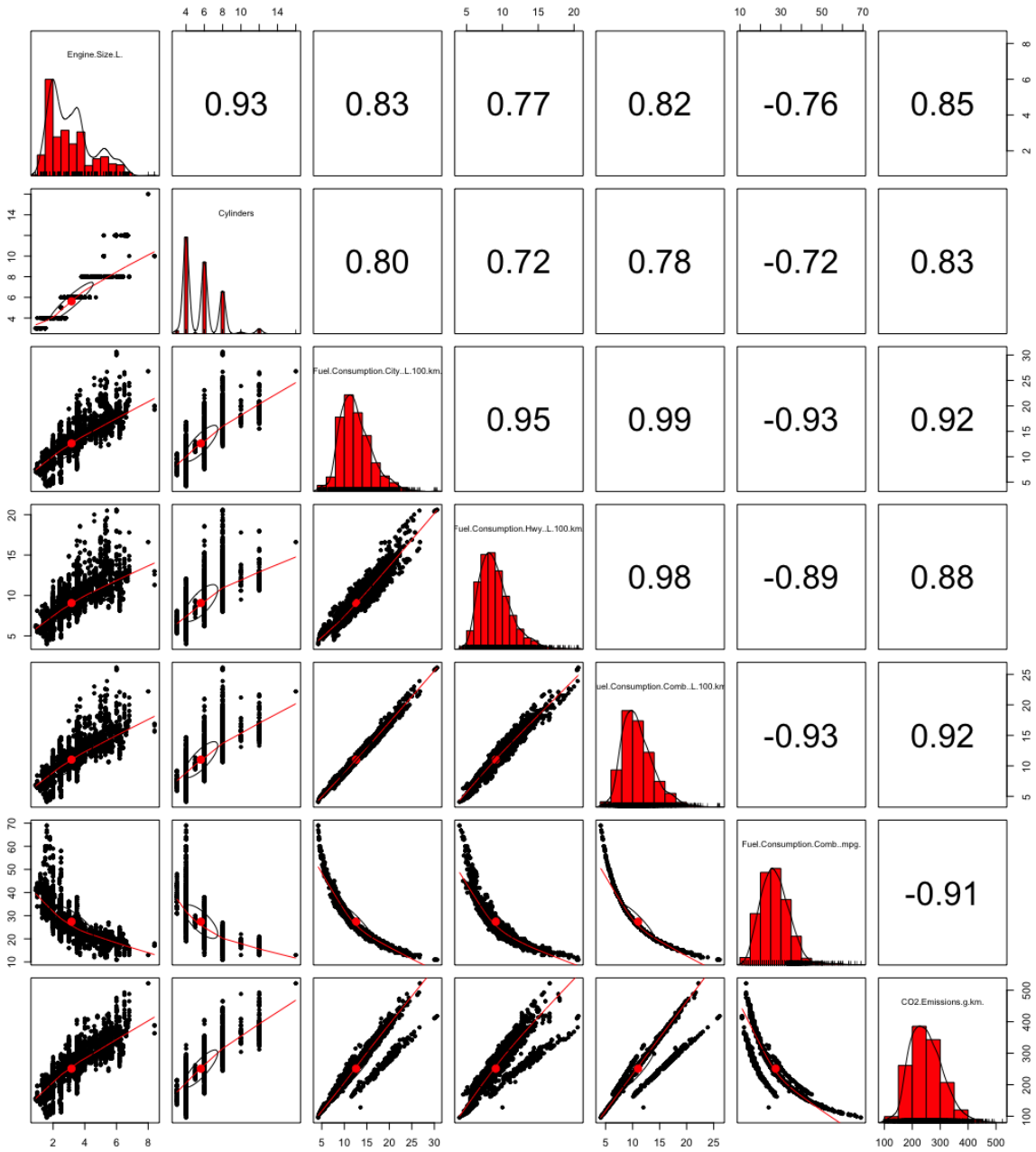

Figure 1: Pair Plot of the database quantitative variables.

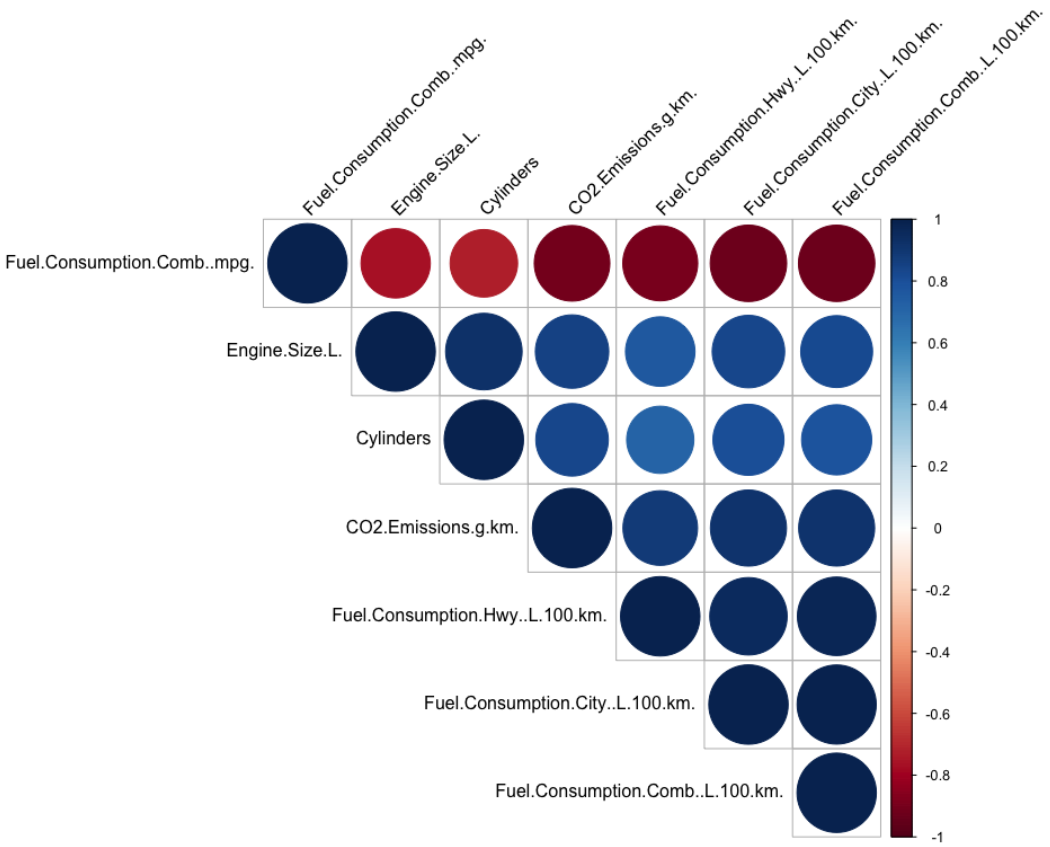

Figure 2: Correlation matrix with Pearson coefficient.

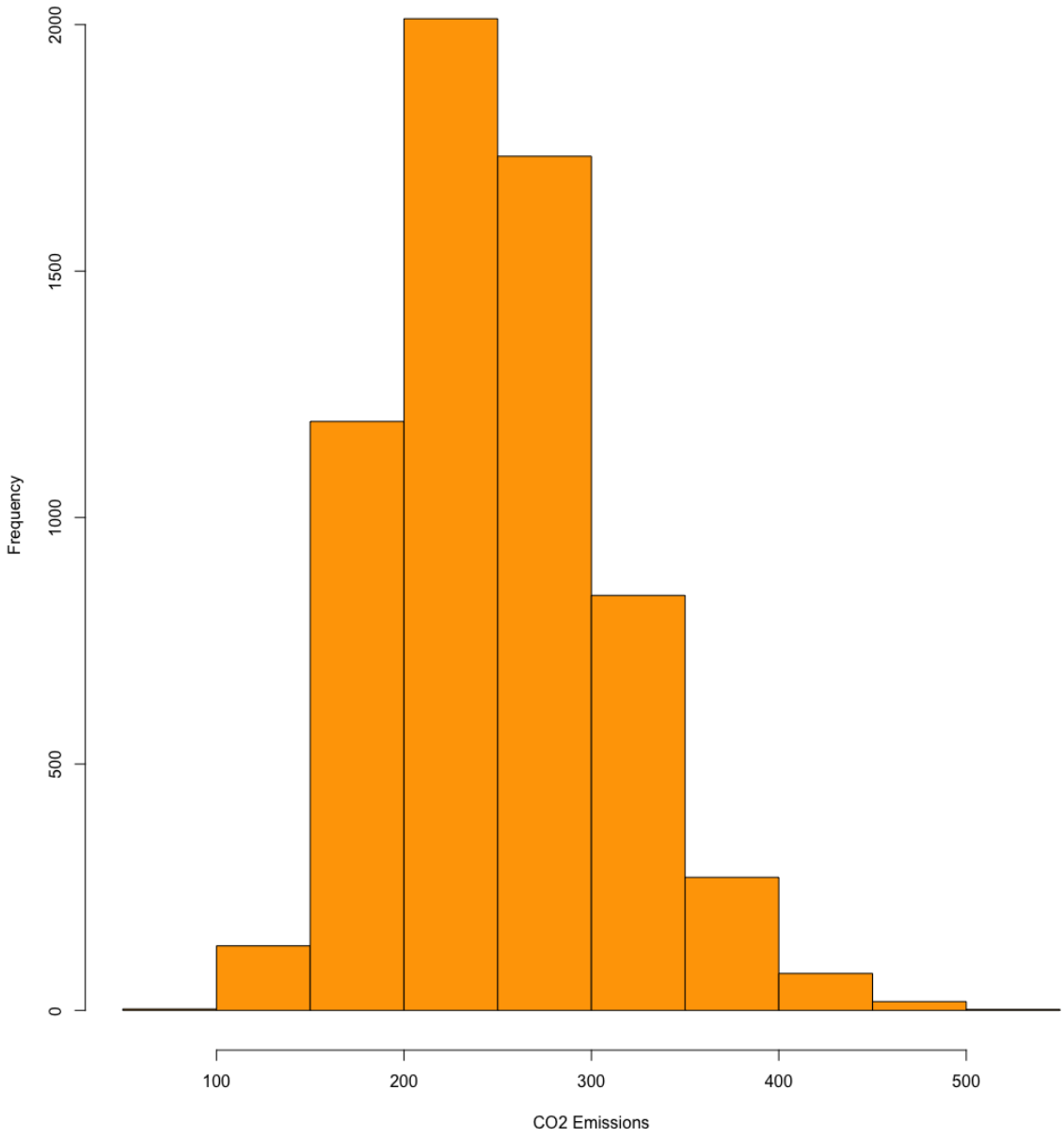

Figure 3: CO_2 Emission distribution.

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Fuel.Consumption.Comb..L.100.km.,
    data = db_co2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.38268 -0.10797  0.03515  0.20299  1.08903

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     6.634e-15  5.038e-03       0        1
Fuel.Consumption.Comb..L.100.km. 9.168e-01  5.039e-03     182   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3993 on 6279 degrees of freedom
Multiple R-squared:  0.8406,    Adjusted R-squared:  0.8406
F-statistic: 3.311e+04 on 1 and 6279 DF,  p-value: < 2.2e-16
```

Figure 4: Summary of the Simple Linear Regression between `CO2.Emissions.g.km.` and `Fuel.Consumption.Comb..L.100.km.`.

We are now ready to perform the Simple Linear Regression to study the relation between the response variable "CO_2 Emissions (g/km)" and the independent variable "Fuel Consumption (Comb L/100km)". The objective is to assess how the CO_2 emissions (g/km) change with respect to the vehicle fuel consumption (L/100km) and to determine whether there is a statistically significant linear relationship between these two variables. The following code performs the regression analysis and visualizes the results using a regression line.

Listing 9: Simple Linear Regression

```r
# Model: CO2 (Resp Var) as a function of Fuel Consumption (L/100km)
model_SLR <- lm(db_co2$CO2.Emissions.g.km. ~ db_co2$Fuel.Consumption.Comb..L
    .100.km., data=db_co2)

# View the summary of the model
summary(model_SLR)

# Plot the regression line
ggplot(db_co2, aes(x=Fuel.Consumption.Comb..L.100.km., y=CO2.Emissions.g.km.)) +
  geom_point() +
  geom_smooth(method="lm", col="red") +
  labs(title="", x="Fuel Consumption (L/100km)", y="CO2 Emission (g/km)")

# Inspected the model
# Generates diagnostic plots for the linear regression
# Residual vs Fitted / Normal Q-Q / Scale-Location / Residuals vs Leverage
par(mfrow = c(2, 2))
plot(model_SLR, sub = "")
```

The regression plot in Figure 5 displays the relationship between Fuel Consumption (L/100 km) on the x-axis and CO2 Emissions (g/km) on the y-axis. The scatterplot shows that most data points follow a positive linear trend, indicating that as fuel consumption increases, CO2 emissions also tend to rise. The red regression line fitted to the data further reinforces this trend, as it slopes upwards, demonstrating a clear linear relationship between the two variables. Although the majority of the data points cluster

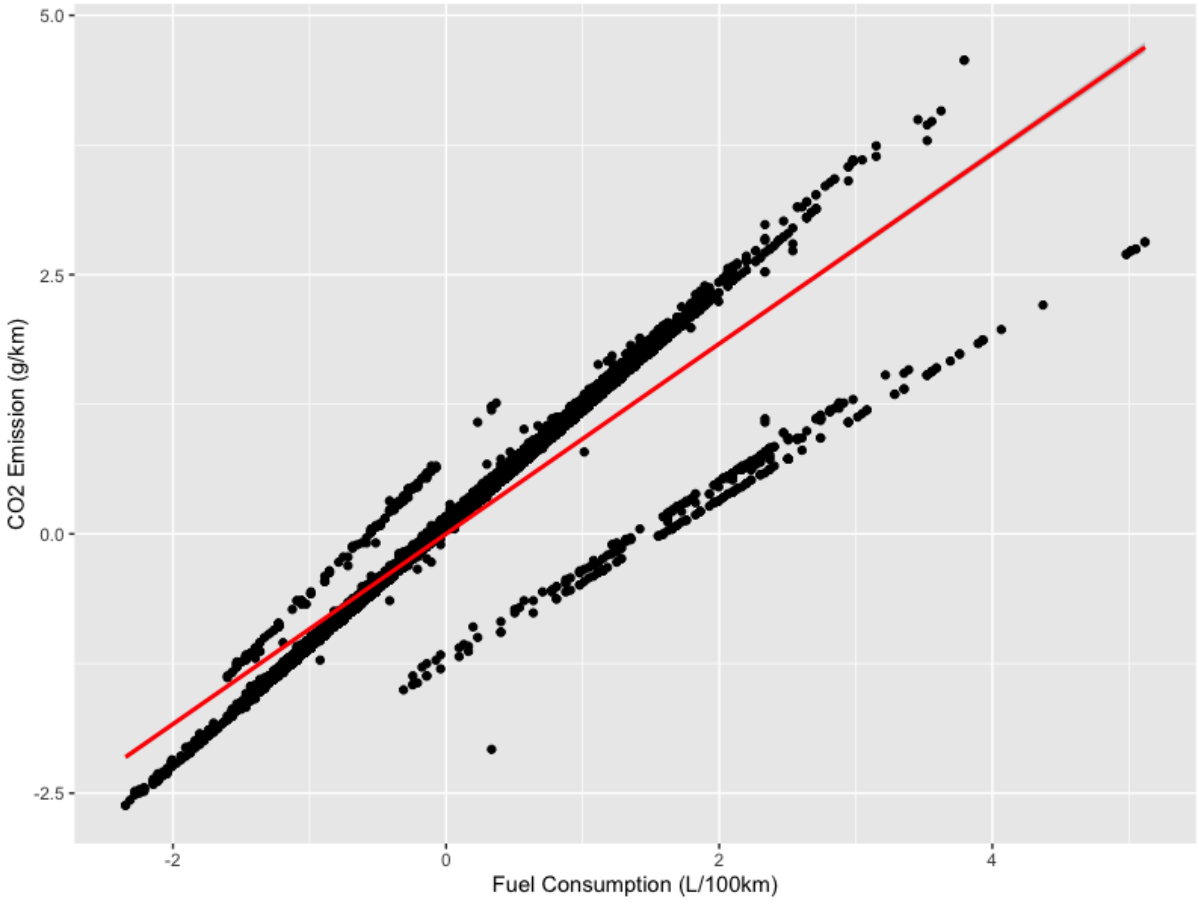

Figure 5: Scatter plot showing the relationship between Fuel Consumption (L/100 km) and CO2 Emissions (g/km). The red line represents the fitted linear regression model, indicating a strong positive relationship between the two variables.

closely around the regression line, suggesting a good fit, there are a few points that deviate significantly. These outliers may indicate instances where other factors—such as vehicle type and engine efficiency deviations in CO2 emissions beyond what can be explained by fuel consumption alone. These deviations could be managed by looking for a Multiple Linear Regression, which will be performed later. Anyway, the overall pattern strongly supports the hypothesis that fuel consumption is a strong predictor of CO2 emissions, as confirmed by the regression diagnostic provided by `summary(model_SLR)`. The results of the regression model are shown in Figure 4, which analyzes the key component of the model as follows:

- Coefficients: The intercept of the model represents the estimated CO2 emissions when fuel consumption is zero, and it is near zero due to standardization. While this value lacks a practical interpretation (as zero fuel consumption is unrealistic), it establishes the baseline level of CO2 emissions. The slope coefficient of 0.92 indicates that for each additional liter per 100 km of fuel consumed, CO2 emissions increase by approximately $0.92\,g/km$. This coefficient is highly significant, with a p-value of $< 2 \times 10^{-16}$, suggesting that fuel consumption is a strong and reliable predictor of CO2 emissions.

- Residuals: The residuals provide a measure of how well the model fits the data. The median residual is close to zero (0.04), indicating that the model does not systematically over- or under-estimate CO2 emissions. However, the range of residuals (from -2.38 to 1.09) suggests the presence of some outliers, which deviate from the model's predictions.

- Model Fit (R-squared): The R-squared value of 0.8406 indicates that 84.06% of the variability in CO2 emissions is explained by fuel consumption. This is a high R-squared value, suggesting that the model provides a strong fit to the data and captures the majority of the variance in CO2 emissions. The Adjusted R-squared, which accounts for the number of predictors in the model, is also 0.8406, reinforcing the robustness of the model.

- Residual Standard Error (RSE): The Residual Standard Error (RSE) is 0.40, representing the average deviation of the observed values from the regression line. While this value is relatively low, indicating a good fit, the existence of some large residuals suggests that there are additional factors influencing CO2 emissions that are not accounted for by fuel consumption alone.

- F-statistic: The model's F-statistic is extremely large (3.311×10^4), with a p-value of $< 2.2 \times 10^{-16}$. This indicates that the model as a whole is highly significant and that fuel consumption significantly contributes to explaining the variation in CO2 emissions.

5.1 Checking Linear regression assumptions

After performing the linear regression, it is important to check if the assumptions of the regression model are met. A set of diagnostic plots can be used to evaluate whether the model is well-fitted to the data and whether the assumptions such as linearity, normality of residuals, and homoscedasticity hold.

The command `plot(model_SLR)` generates four diagnostic plots (Figure 6), which are shown in Figure, that help assess the validity of the regression model. Specifically, the plot shown in Figure 6 provides a comprehensive view of potential issues in the regression model, including violations of key assumptions such as linearity, normality, and homoscedasticity, as well as the identification of outliers and influential points. Below is a summary of the key observations:

- Residuals vs Fitted: The residuals display a noticeable pattern rather than random scatter, suggesting that the model may not fully capture the linear relationship for all observations, particularly in the presence of outliers.

- Normal Q-Q Plot: Significant deviations from the diagonal line, especially in the tails, indicate a violation of the normality assumption for the residuals, which could impact the reliability of statistical inferences.

- Scale-Location Plot: The fan-shaped pattern suggests heteroscedasticity, indicating that the variance of the residuals is not constant across the range of fitted values, thus violating another key

assumption of the regression model.

- Residuals vs Leverage: A few points exhibit both high leverage and large residuals, suggesting the presence of influential outliers that may disproportionately affect the model's estimates.

The diagnostic plots highlight potential issues with non-linearity, heteroscedasticity, non-normality of residuals, and influential outliers. These problems suggest that the model could benefit from variable transformations or more robust modeling techniques to improve the accuracy and validity of the estimates.

Opposite to this not completely satisfier example, we provide the plots of the Simple Linear Regression between `CO2.Emissions.g.km.` and `Engine.Size.L.`. From the matrix plot 1 we expect a good result from the regression model. The results are shown in Figure 7 - 9.

The diagnostic plots for the new regression model show overall improvement compared to the previous analysis. While some issues such as non-linearity and heteroscedasticity persist, they are less pronounced. The model appears more stable, but the presence of influential outliers suggests that further refinement may still be necessary.

6 Multiple Linear Regression

After covering simple linear regression, we now move on to multiple linear regression, which allows us to model the relationship between a dependent variable and multiple independent variables. This technique extends the concepts we have already explored by incorporating additional predictors, enabling us to analyze more complex relationships within the data. In this section, we will demonstrate how to develop a multiple linear regression model using R, exploring the step-by-step process of fitting the model, interpreting the coefficients, and assessing its performance.

We perform a multiple linear regression embedding all the numerical predictors in the forecasting model for CO_2 emissions. The R implementation is the following:

Listing 10: Multiple Linear Regression

```
# Multiple Linear Regression
# ----------------------------------
# Choose the independent variables to include -> complete set
model_MLR <- lm(CO2.Emissions.g.km.  ~ Engine.Size.L. + Cylinders +
                  Fuel.Consumption.City..L.100.km. + Fuel.Consumption.Hwy..L
                      .100.km. +
                  Fuel.Consumption.Comb..L.100.km. + Fuel.Consumption.Comb..mpg
                      .,
                  data = db_co2)

summary((model_MLR))

# Model Diagnostics
# ----------------------------------
par(mfrow = c(2, 2))
plot(model_MLR, sub = "")
```

The summary, shown in Figure **??**, provides detailed information about the multiple linear regression model, which predicts CO_2 emissions based on all the independent numerical variables of the database. The key findings from the model are as follows:

- **Intercept**: The intercept is approximately zero (1.938×10^{-15}), indicating that when all independent variables are zero, the model predicts near-zero CO2 emissions. This value is theoretically expected given the nature of the variables.
- **Engine Size (L)**: The coefficient for engine size is positive (0.1233) and highly significant (p-value $< 2 \times 10^{-16}$). This suggests that larger engine sizes lead to a significant increase in CO2 emissions.
- **Cylinders**: The number of cylinders also has a positive and highly significant effect (0.2332, p-value $< 2 \times 10^{-16}$), indicating that vehicles with more cylinders emit more CO2.
- **Fuel Consumption (City)**: Although the coefficient is positive (0.03559), the variable is not statistically significant (p-value = 0.848), suggesting that city fuel consumption does not significantly contribute to the prediction of CO2 emissions in this model.

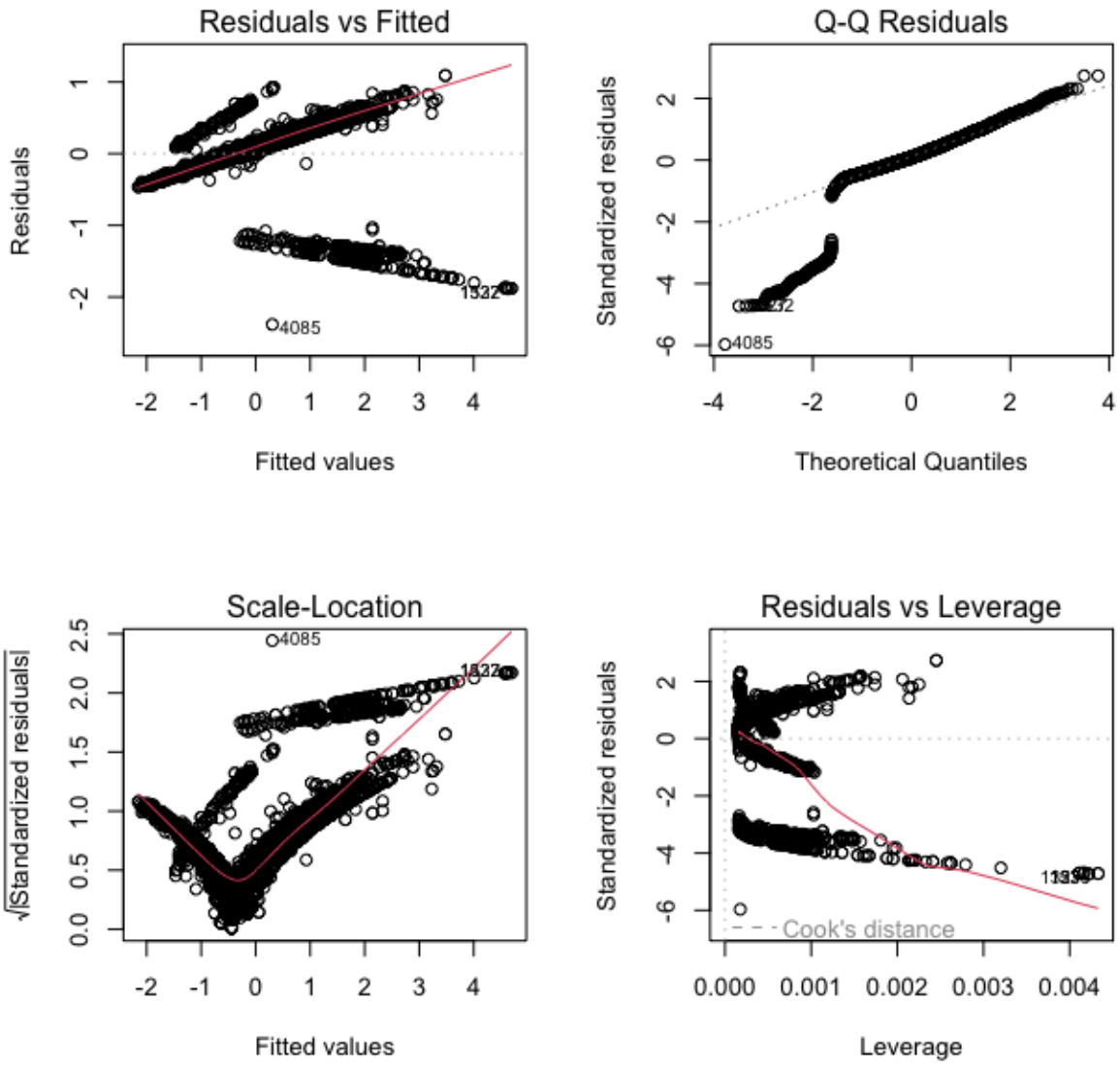

Figure 6: Diagnostic plots for the linear regression model. The top-left plot (Residuals vs Fitted) checks for linearity and homoscedasticity, while the top-right Q-Q plot tests the normality of residuals. The bottom-left Scale-Location plot helps assess the spread of residuals across fitted values for homoscedasticity, and the bottom-right plot (Residuals vs Leverage) identifies influential data points that could disproportionately affect the regression model.

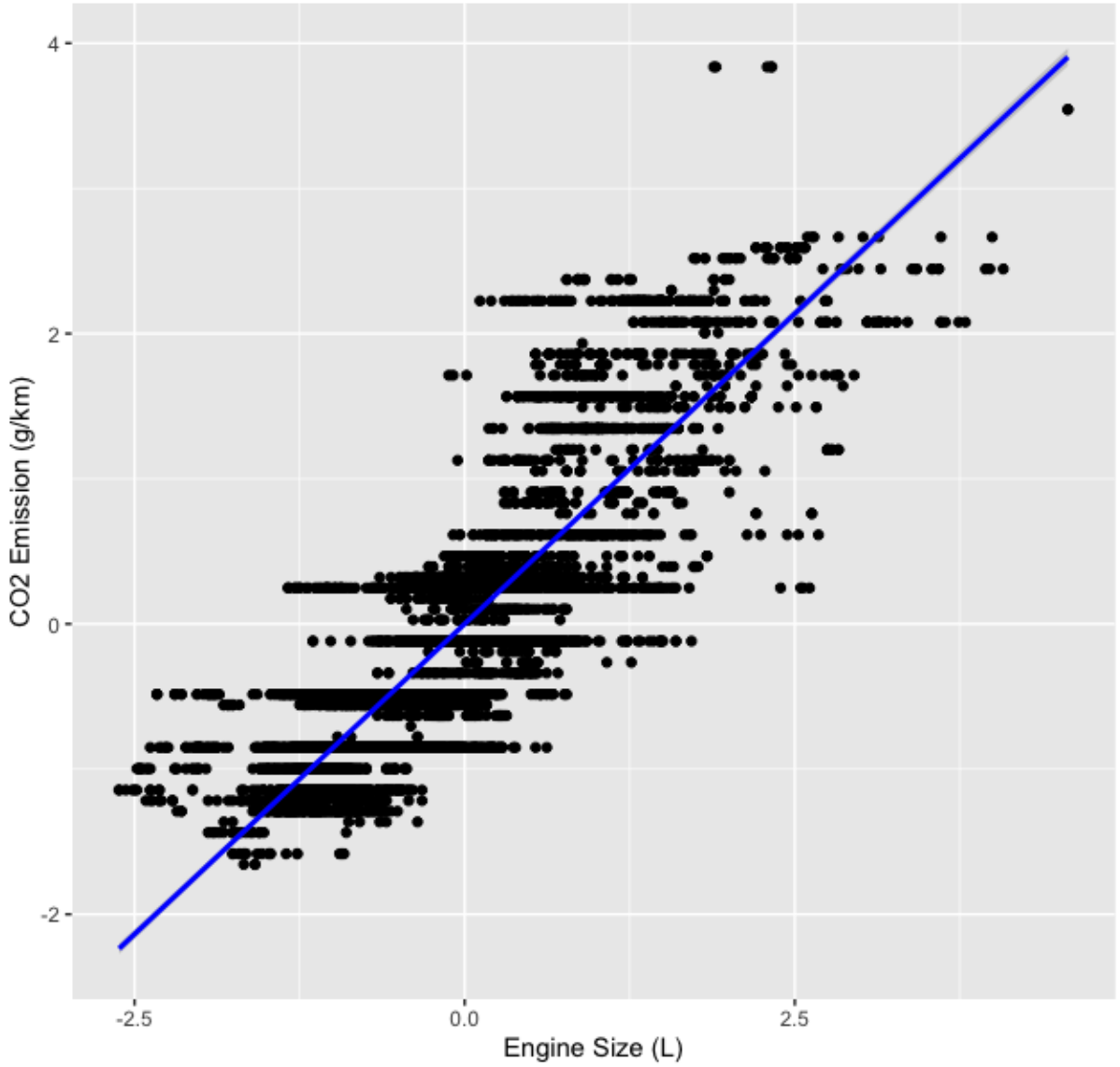

Figure 7: Scatter plot showing the relationship between Fuel Consumption (L/100 km) and Engine Size (L). The blue line represents the fitted linear regression model, indicating a strong positive relationship between the two variables.

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Engine.Size.L., data = db_co2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9158 -0.3015 -0.0215  0.3175  2.3998

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.147e-14  6.549e-03     0.0        1
Engine.Size.L.  8.548e-01  6.549e-03   130.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.519 on 6279 degrees of freedom
Multiple R-squared:  0.7307,    Adjusted R-squared:  0.7306
F-statistic: 1.703e+04 on 1 and 6279 DF,  p-value: < 2.2e-16
```

Figure 8: Detail of the `summary(model_SLR_new` regarding simple linear regression between CO_2 and Engine Size (L).

- **Fuel Consumption (Highway)**: The coefficient for highway fuel consumption is 0.01938 with a p-value of 0.0487, making it marginally significant.
- **Fuel Consumption (Combined L/100 km)**: This variable is not significant, with a p-value of 0.9408, and the coefficient is relatively small (0.02076).
- **Fuel Consumption (Combined mpg)**: The coefficient for fuel consumption in miles per gallon is negative and highly significant (-0.4189, p-value $< 2 \times 10^{-16}$), indicating that better fuel efficiency (measured in mpg) is strongly associated with lower CO2 emissions.

R-squared and Adjusted R-squared values are both around 0.90, meaning that about 90% of the variability in CO_2 emissions is explained by the model. The small difference between these two values suggests that the model is well-fitted without overfitting due to unnecessary predictors. The residual standard error is 0.3119, indicating a good fit, and the F-statistic (9711) with a p-value less than 2.2×10^{-16} confirms that the model is statistically significant overall.

The model shows that variables such as "Engine Size", "Cylinders", and "Combined Fuel Consumption (mpg)" have a significant impact on CO2 emissions. In contrast, other predictors contribute less significantly. The model performs well with an R-squared of over 90%, though some predictors may be reevaluated for simplification. Diagnostic plots for the multiple linear regression are shown in Figure 11.

6.1 Testing for Interaction Effects in Multiple Linear Regression

In the context of multiple linear regression, interaction terms are introduced to investigate whether the relationship between an independent variable and the dependent variable is affected by the presence of another independent variable. Specifically, interaction effects allow us to model situations where the influence of one predictor on the outcome is conditional on the level of another predictor. When analyzing the factors that influence *CO2 emissions*, we hypothesized that the effect of *Engine Size (L)* on emissions might vary depending on the vehicle's *Fuel Consumption in City Driving (L/100 km)*. For example, larger engines may have a disproportionately higher impact on CO2 emissions for vehicles with higher

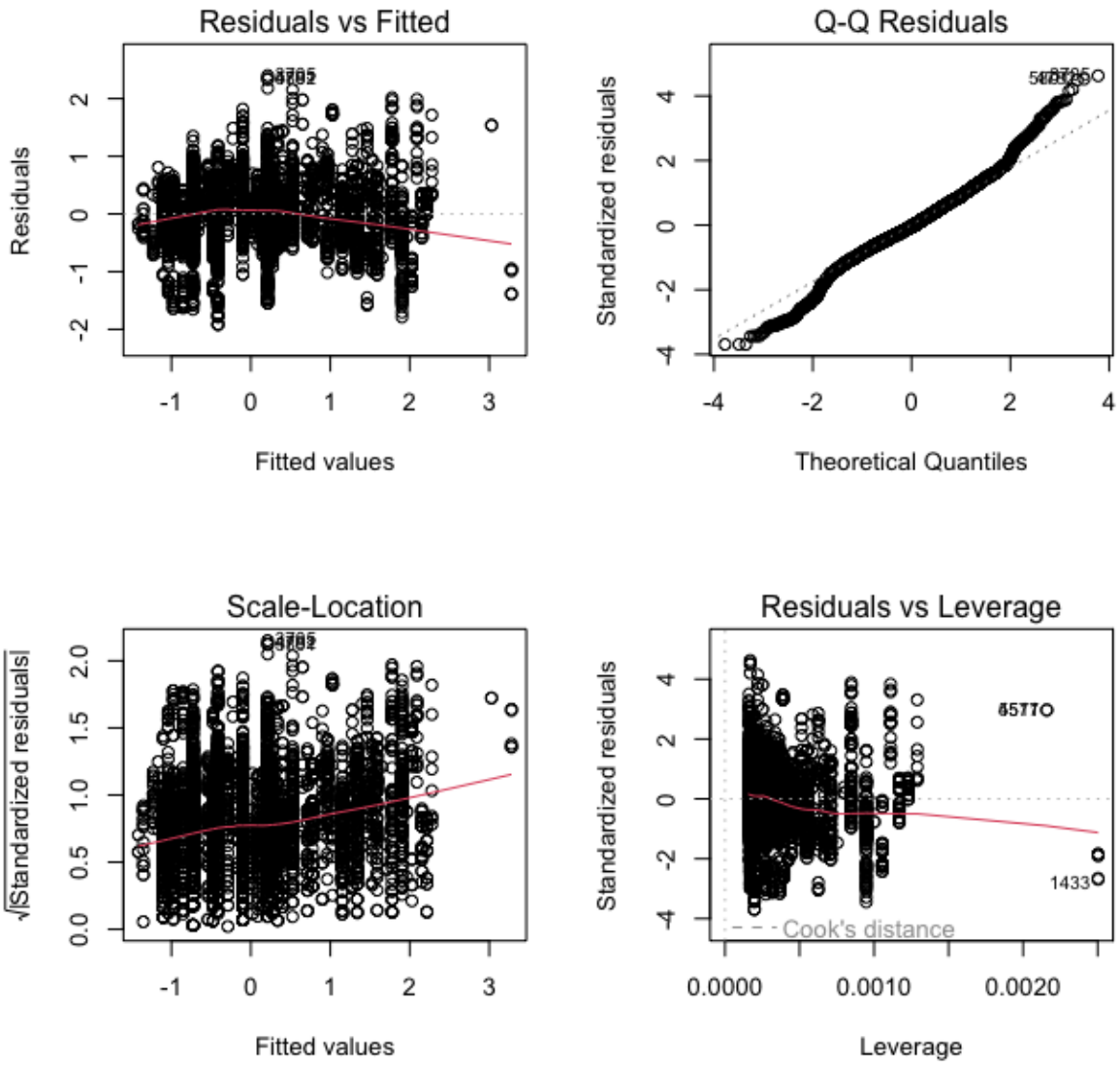

Figure 9: Diagnostic plots for the linear regression model between CO_2 and Engine Size (L).. The top-left plot (Residuals vs Fitted) checks for linearity and homoscedasticity, while the top-right Q-Q plot tests the normality of residuals. The bottom-left Scale-Location plot helps assess the spread of residuals across fitted values for homoscedasticity, and the bottom-right plot (Residuals vs Leverage) identifies influential data points that could disproportionately affect the regression model.

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Engine.Size.L. + Cylinders +
    Fuel.Consumption.City..L.100.km. + Fuel.Consumption.Hwy..L.100.km. +
    Fuel.Consumption.Comb..L.100.km. + Fuel.Consumption.Comb..mpg.,
    data = db_co2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06076 -0.09740 -0.00413  0.13020  1.56647

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.938e-15  3.936e-03   0.000   1.0000
Engine.Size.L.                   1.233e-01  1.165e-02  10.582   <2e-16 ***
Cylinders                        2.332e-01  1.094e-02  21.322   <2e-16 ***
Fuel.Consumption.City..L.100.km. 3.559e-02  1.857e-01   0.192   0.8480
Fuel.Consumption.Hwy..L.100.km.  1.938e-01  9.830e-02   1.971   0.0487 *
Fuel.Consumption.Comb..L.100.km. 2.076e-02  2.796e-01   0.074   0.9408
Fuel.Consumption.Comb..mpg.      -4.189e-01  1.062e-02 -39.454   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3119 on 6274 degrees of freedom
Multiple R-squared:  0.9028,    Adjusted R-squared:  0.9027
F-statistic:  9711 on 6 and 6274 DF,  p-value: < 2.2e-16
```

Figure 10: Summary of the multiple linear regression model predicting CO2 emissions (g/km) based on engine size, cylinders, and various fuel consumption metrics.

city fuel consumption, due to the inefficiencies that arise in urban driving conditions. To explore this possibility, we introduce an interaction term between *Engine Size (L)* and *City Fuel Consumption (L/100 km)*. The analysis in R is performed as follows:

Listing 11: Multiple Linear Regression with interaction

```r
# Switch on the interaction between Engine.Size.L. and Fuel.Consumption.City..L
    .100.km.
model_MLR_int <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. * Fuel.Consumption.
    City..L.100.km. +
                                Cylinders + Fuel.Consumption.Hwy..L.100.km. +
                                Fuel.Consumption.Comb..L.100.km. + Fuel.
                                    Consumption.Comb..mpg.,
                                data = db_co2)

# Results of the Interacting Model
summary(model_MLR_int)
```

By including this interaction term, we can determine whether the impact of engine size on CO2 emissions changes as fuel consumption in the city increases. A significant interaction term would suggest that the relationship between these two variables is not simply additive but instead dependent on their combined levels. In this way, interaction terms provide a more nuanced understanding of how variables jointly influence the outcome, which may lead to more accurate predictions and better model performance. The goodness of fit, from `summary(model_MLR_int` is shown in Figure 12.

From Figure 12 we see that the results of the model with interaction confirm the significance of the relationship between *Engine Size (L)* and *City Fuel Consumption (L/100 km)* on *CO2 emissions*. The interaction term has a positive and highly significant coefficient (0.092451, p-value $< 2 \times 10^{-16}$),

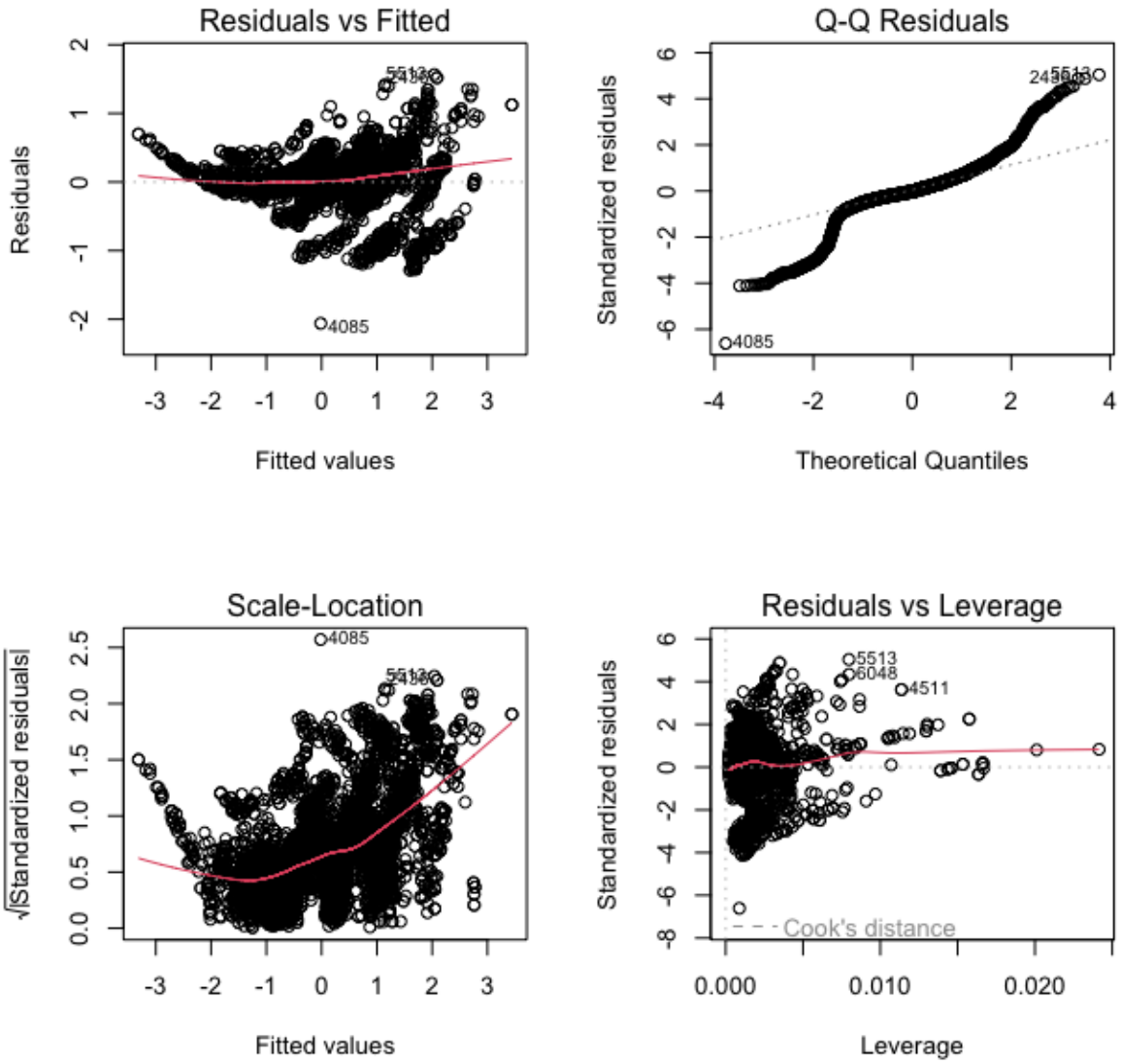

Figure 11: Diagnostic plot for the multiple linear regression on the complete CO_2 database.

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Engine.Size.L. * Fuel.Consumption.City..L.100.km. +
    Cylinders + Fuel.Consumption.Hwy..L.100.km. + Fuel.Consumption.Comb..L.100.km. +
    Fuel.Consumption.Comb..mpg., data = db_co2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01279 -0.11588 -0.01776  0.13109  1.63265

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    -0.077111   0.006015 -12.821   <2e-16 ***
Engine.Size.L.                                  0.103900   0.011458   9.068   <2e-16 ***
Fuel.Consumption.City..L.100.km.               -0.048497   0.181770  -0.267   0.7896
Cylinders                                       0.226286   0.010713  21.123   <2e-16 ***
Fuel.Consumption.Hwy..L.100.km.                 0.233093   0.096221   2.422   0.0154 *
Fuel.Consumption.Comb..L.100.km.               -0.148537   0.273783  -0.543   0.5875
Fuel.Consumption.Comb..mpg.                    -0.614875   0.015678 -39.218   <2e-16 ***
Engine.Size.L.:Fuel.Consumption.City..L.100.km.  0.092451   0.005538  16.692   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3053 on 6273 degrees of freedom
Multiple R-squared:  0.9069,    Adjusted R-squared:  0.9068
F-statistic:  8732 on 7 and 6273 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of the Multiple Linear Regression Model with Interaction between *Engine Size (L)* and *City Fuel Consumption (L/100 km)*.

indicating that the effect of engine size on emissions increases with higher city fuel consumption. This suggests that larger engines have a more pronounced impact on CO2 emissions for vehicles that are less fuel-efficient in urban conditions. Moreover, it is interesting that *Fuel Consumption in City Driving* alone is not significant, the interaction with engine size reveals its importance in determining CO2 output. The model explains approximately 90.7% of the variance in emissions ($R^2 = 0.9069$), highlighting the utility of incorporating interaction terms to capture complex, real-world relationships between variables.

7 Conclusion

In this exercise, we demonstrated how to perform a simple and multiple linear regression in R. We covered the steps of data cleaning, exploratory analysis, simple and multiple regression modeling, and diagnostics. The next step is to practice these techniques on different datasets to strengthen your understanding.