LINEAR REGRESSION – LEAST SQUARES

Submitted by

Rahul Vyas – RA2211047010096 – **TL**
Harsha Vardhan – RA2211047010095 – **TM**
Jayasurya – RA22110470100123 – **TM**
Maneesh Kumar – RA2211047010094 – **TM**
Harsha Vardhan – RA2211047010113 – **TM**
Nikhil – RA2211047010102 – **TM**
Vedhashree – RA2211047010088 – **TM**
Kellen – RA2211047010137 – **TM**
Rishik – RA2211047010112 – **TM**
Charitesh – RA2211047010085 – **TM**
Ajay – RA2211047010087 – **TM**

Under the Guidance of
Dr. Sheryl Oliver A
Associate Professor, Department of Computational Intelligence

In partial satisfaction of the requirements for the degree of
BACHELORS OF TECHNOLOGY
in
ARTIFICIAL INTELLIGENCE

SCHOOL OF COMPUTING
COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603203

APRIL 2023

SRM INSTITUTION OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR-603203

BONAFIDE CERTIFICATE

Certified that this Course Project Report titled **"LINEAR REGRESSION - LEAST SQUARES MODEL"** is the bonafide work done by **Rahul Vyas Team** who carried out under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other work.

SIGNATURE
Faculty In-Charge
Dr. Sheryl Oliver A
Associate Professor
Department of Computational Intelligence
SRM Institute of Science and Technology

HEAD OF THE DEPARTMENT
Dr. Annie Uthra R.
Professor and Head,
Department of Computational Intelligence
SRM Institute of Science and Technology

LINEAR REGRESSION – LEAST SQUARES MODEL

21AIC101J – Foundation of Data Analysis

Mini Project Report

Submitted by

Rahul Vyas – RA2211047010096 – **TL**
Harsha Vardhan – RA2211047010095 – **TM**
Jayasurya – RA22110470100123 – **TM**
Maneesh Kumar – RA2211047010094 – **TM**
Harsha Vardhan – RA2211047010113 – **TM**
Nikhil – RA2211047010102 – **TM**
Vedhashree – RA2211047010088 – **TM**
Kellen – RA2211047010137 – **TM**
Rishik – RA2211047010112 – **TM**
Charitesh – RA2211047010085 – **TM**
Ajay – RA2211047010087 – **TM**

Under the Guidance of
Dr. Sheryl Oliver A
Associate Professor, Department of Computational Intelligence

In partial satisfaction of the requirements for the degree of
BACHELORS OF TECHNOLOGY
in
ARTIFICIAL INTELLIGENCE

SCHOOL OF COMPUTING
COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603203

APRIL 2023

Table of Contents

Linear Regression – Least Squares

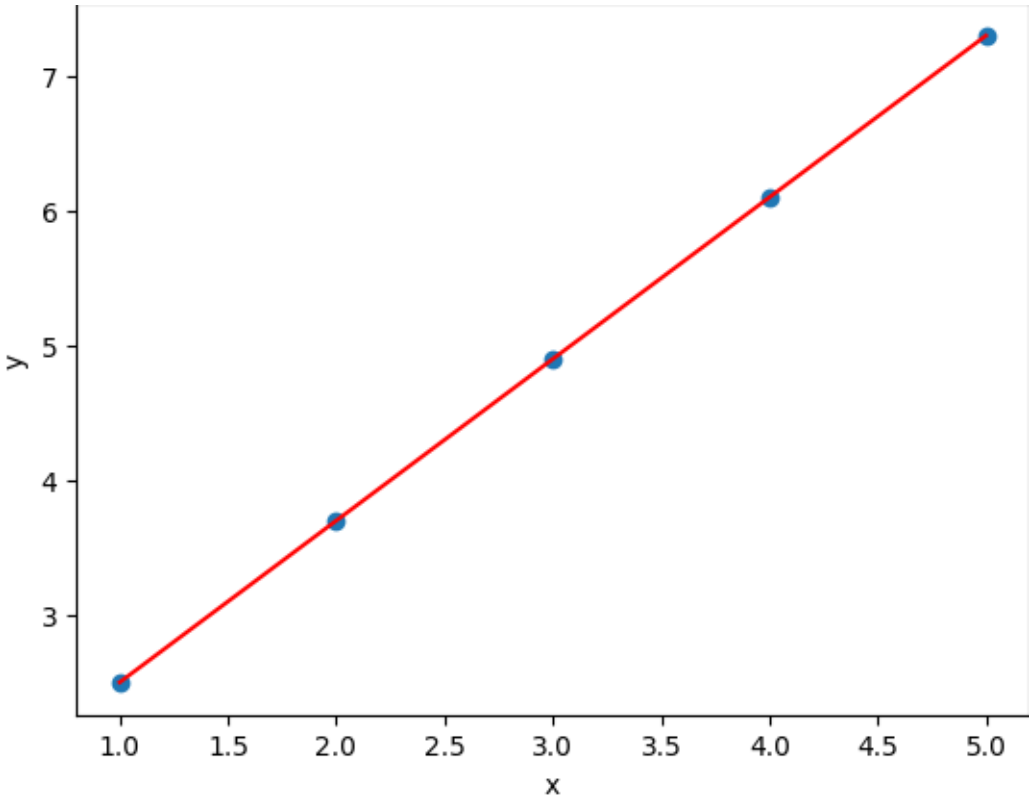
Linear regression is a statistical method that is commonly used in machine learning to model the relationship between a dependent variable and one or more independent variables. In its simplest form, linear regression involves fitting a straight line to a set of data points, with the goal of minimizing the distance between the line and the data points.

The **least squares** method is one common way to estimate the parameters of a linear regression model. In the least squares method, we minimize the sum of the squared differences between the observed values of the dependent variable and the predicted values from the linear regression model.

Here's how to perform linear regression using the least squares method in Python:

```
import numpy as np
import matplotlib.pyplot as plt

# Define the data
x = np.array([1, 2, 3, 4, 5])
y = np.array([2.5, 3.7, 4.9, 6.1, 7.3])

# Calculate the means of x and y
x_mean = np.mean(x)
y_mean = np.mean(y)

# Calculate the slope and y-intercept using the least squares method
numerator = np.sum((x - x_mean) * (y - y_mean))
denominator = np.sum((x - x_mean) ** 2)
b1 = numerator / denominator
b0 = y_mean - b1 * x_mean

# Print the equation of the line
print(f"The equation of the line is y = {b0:.2f} + {b1:.2f}x")

# Make predictions for new values of x
new_x = np.array([6, 7, 8, 9, 10])
new_y = b0 + b1 * new_x
print(f"Predicted values of y for new values of x: {new_y}")

# Plot the data and the line of best fit
plt.scatter(x, y)
plt.plot(x, b0 + b1 * x, color='red')
plt.xlabel('x')
plt.ylabel('y')
plt.show()
```

Output:



Problem Statement

You work as a data scientist for a company that sells houses. Your boss has asked you to analyse a dataset of house prices and square footage to see if there is a relationship between the two variables. Specifically, your boss wants you to perform linear regression using the least squares method and use the resulting model to predict the price of a house with a given square footage.

The dataset is provided in a CSV file named **house_data.csv**. The first column contains the square footage of each house, and the second column contains the price of each house.

house_data.csv:

```
square_feet,price
1400,245000
1600,312000
1700,279000
1875,308000
1100,199000
1550,219000
2350,405000
2450,324000
1425,319000
1700,255000
```

Methodology or Procedure

The methodology for the problem of using linear regression and least squares to analyse a dataset of house prices and square footage and make predictions can be broken down into the following steps:

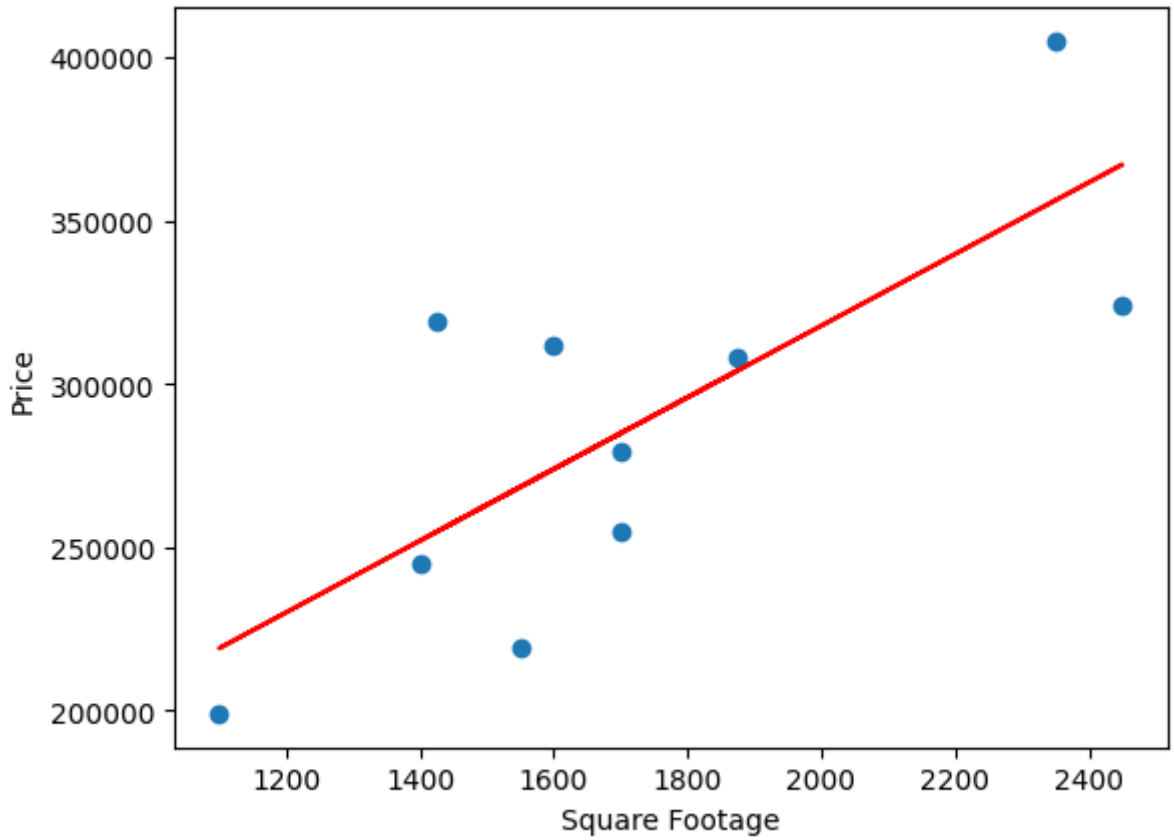
1. Load the dataset: The first step is to load the dataset from a CSV file into Python using a library like Pandas. In this case, we are interested in two columns of the dataset: the square footage and the price.
2. Visualize the data: Before performing linear regression, it is helpful to visualize the data to see if there is a relationship between the two variables. A scatter plot is a common way to do this, which can be created using a library like Matplotlib.
3. Calculate the means: Calculate the means of the square footage and the price. These will be used later in the least squares formula to calculate the slope and y-intercept.
4. Calculate the slope and y-intercept: Use the least squares method to calculate the slope and y-intercept of the line of best fit. This can be done using NumPy and the formula **b1 = sum((x - x_mean) * (y - y_mean)) / sum((x - x_mean) ** 2)** for the slope and **b0 = y_mean - b1 * x_mean** for the y-intercept.
5. Print the equation of the line: Once the slope and y-intercept have been calculated, print the equation of the line of best fit in the form y = b0 + b1*x.
6. Plot the line of best fit: Plot the line of best fit on the scatter plot from step 2. This can be done by calculating the predicted values of y for each value of x using the equation of the line, and then plotting those values as a line on the scatter plot.
7. Making predictions: Finally, using the equation of the line to make predictions for new values of x. In this case, we are interested in predicting the price of a house with a square footage of 2000. This can be done by plugging the value of x into the equation of the line and solving for y.

Coding

```
# Importing necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Load the data from the CSV file
data = pd.read_csv('house_data.csv')
x = data['square_feet'].values
y = data['price'].values

# Calculate the means of x and y
x_mean = np.mean(x)
y_mean = np.mean(y)

# Calculate the slope and y-intercept using the least squares
method
numerator = np.sum((x - x_mean) * (y - y_mean))
denominator = np.sum((x - x_mean) ** 2)
b1 = numerator / denominator
b0 = y_mean - b1 * x_mean

# Print the equation of the line
print(f"The equation of the line is y = {b0:.2f} + {b1:.2f}x")

# Plot the data and the line of best fit
plt.scatter(x, y)
plt.plot(x, b0 + b1 * x, color='red')
plt.xlabel('Square Footage')
plt.ylabel('Price')
plt.show()

# Predict the price of a house with a square footage of 2000
new_x = 2000
new_y = b0 + b1 * new_x
print(f"The predicted price of a house with a square footage of
2000 is {new_y:.2f}")
```

Output

```
The equation of the line is y = 98248.33 + 109.77x
```




```
The predicted price of a house with a square footage of 2000 is 317783.81
```

Conclusion

In this problem, we used linear regression and the least squares method to analyze a dataset of house prices and square footage and make predictions. We first loaded the dataset into Python using Pandas and visualized the data using a scatter plot. We then used the least squares method to calculate the slope and y-intercept of the line of best fit and printed the equation of the line. We plotted the line of best fit on the scatter plot and made predictions for new values of x using the equation of the line.

From the analysis, we can conclude that there is a positive correlation between square footage and house price, as evidenced by the line of best fit. We can use this equation to make predictions for the price of a house given its square footage. However, it is important to note that this model is based on a small sample size and may not be representative of the entire population of houses. Therefore, caution should be taken when making predictions based on this model.

*****Thank You *****