prac1-2

October 18, 2024

Aim: Linear regression using linear least squares fit method

Theory: ### Linear Regression Using Linear Least Squares Fit Method (Theoretical Overview)

Overview

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The primary goal is to find a linear equation that best predicts the dependent variable based on the independent variables. The linear least squares fit method is a widely-used approach for fitting this linear model.

Purpose

The main purpose of linear regression is to understand how the independent variables influence the dependent variable and to make predictions. For example, in a real estate context, you might want to predict house prices (dependent variable) based on features like size, number of bedrooms, and location (independent variables).

How It Works

1. **Modeling the Relationship**: Linear regression assumes a linear relationship between the independent and dependent variables. This means that as the independent variable changes, the dependent variable changes in a consistent way. For example, a small increase in size might lead to a predictable increase in price.

2. **Finding the Best Fit**: The linear least squares method focuses on finding the best-fitting line (or hyperplane in multiple dimensions) that minimizes the difference between the actual values of the dependent variable and the values predicted by the model. This difference is measured in terms of the "least squares," which emphasizes minimizing the sum of the squared differences (residuals).

3. **Estimating Parameters**: The parameters of the linear model (like the slope and intercept) are estimated based on the data. This involves analyzing how the independent variables relate to the dependent variable and calculating the best-fitting coefficients that describe this relationship.

4. **Making Predictions**: Once the model is trained and parameters are estimated, it can be used to make predictions. By inputting new values for the independent variables, you can obtain predicted values for the dependent variable.

Assumptions

Linear regression relies on several key assumptions to ensure the validity of the results:

- **Linearity**: The relationship between the dependent and independent variables is linear.

- **Independence**: The residuals (errors) should be independent of each other, meaning that one observation does not affect another.
- **Homoscedasticity**: The residuals should have constant variance across all levels of the independent variable. This means that the spread of the errors should remain consistent.
- **Normality**: The residuals should be approximately normally distributed, especially when making inferences about the coefficients.

Applications

Linear regression is used across various fields, including economics, biology, engineering, and social sciences. It is commonly applied for:

- Predicting outcomes based on historical data.
- Identifying trends and relationships between variables.
- Informing decision-making processes based on data analysis.

[1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:

```
X = np.random.rand(100, 1)
y = 4 + 3 * X + np.random.randn(100, 1)
```

[3]:

```
data = np.hstack((X, y))
df = pd.DataFrame(data, columns = ["x", "y"])
```

[4]:

```
df["xy"] = df["x"] * df["y"]
df["x2"] = df["x"]**2
```

[5]:

```
df.head()
```

[5]:

```
          x         y        xy        x2
0  0.309017  6.145334  1.899015  0.095492
1  0.946136  5.067804  4.794833  0.895174
2  0.619322  5.364781  3.322529  0.383560
3  0.990558  6.734716  6.671125  0.981204
4  0.666165  6.554390  4.366302  0.443775
```

[6]:

```
sum_xy = df["xy"].sum()
sum_x = df["x"].sum()
sum_y = df["y"].sum()
sum_x2 = df["x2"].sum()
```

[7]:

```
n = len(df)
m = ((n*sum_xy) - (sum_x*sum_y))/(n*sum_x2 - sum_x**2)
b = (sum_y - m*sum_x)/n
```

[8]:

```
df["y_pred"] = m*df["x"] + b
```

[9]:

```
df.head()
```

[9]:

```
          x         y        xy        x2    y_pred
0  0.309017  6.145334  1.899015  0.095492  4.927643
1  0.946136  5.067804  4.794833  0.895174  6.722344
2  0.619322  5.364781  3.322529  0.383560  5.801742
3  0.990558  6.734716  6.671125  0.981204  6.847475
4  0.666165  6.554390  4.366302  0.443775  5.933692
```

[10]:

```
df["error"] = abs(df["y"] - df["y_pred"])
```

[11]:

```
error = df["error"].sum()
```

[12]:

```
print(error)
```

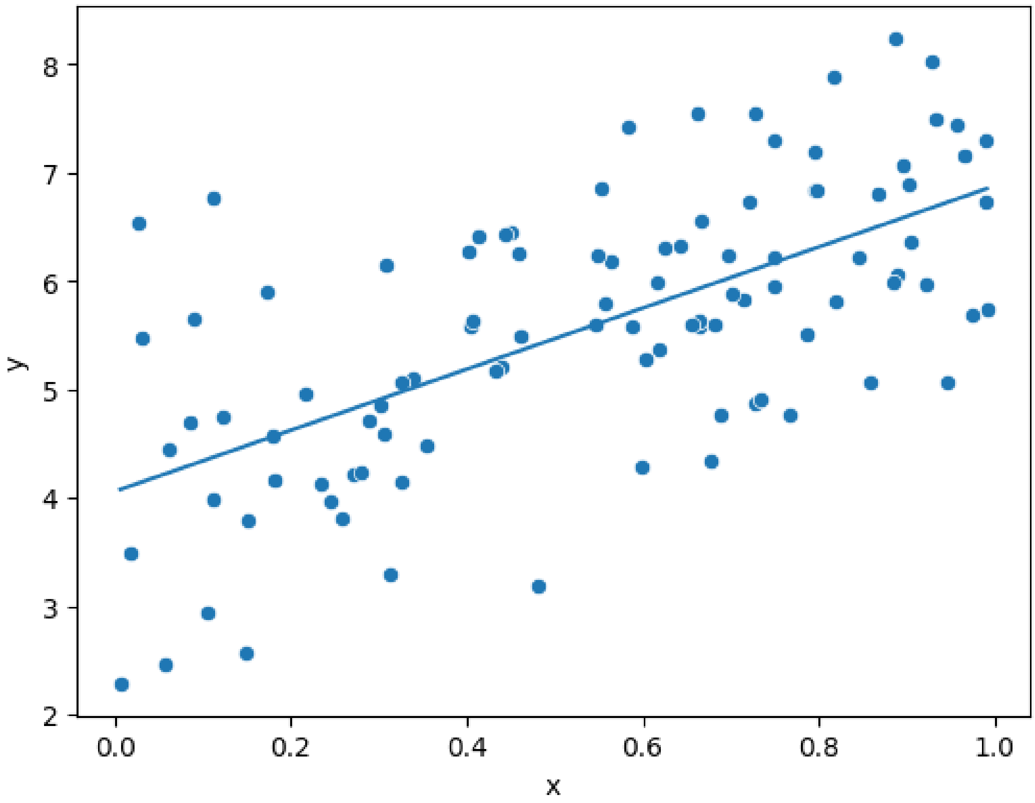
```
77.68452502972075
```

[13]:

```
sns.scatterplot(x = "x", y = "y", data = df)
sns.lineplot(x = "x", y = "y_pred", data = df)
```

[13]:

```
<Axes: xlabel='x', ylabel='y'>
```



Conclusion

The linear least squares fit method is a foundational technique in statistical modeling that helps uncover relationships between variables and make predictions. While it is powerful and versatile, users must be mindful of its assumptions and ensure that the data meets these criteria for effective modeling and accurate results.