Simple Linear Regression

Rafiq Islam

2024-08-29

Table of contents

# Simple Linear Regression

A simple linear regression in multiple predictors/input variables/features/independent variables/ explanatory variables/regressors/ covariates (many names) often takes the form

where are regression parameters or constant values that we aim to estimate and is a normally distributed error term independent of or also called the white noise.

In this case, the model:

Therefore, in our model we need to estimate the parameters . The true relationship between the explanatory variables and the dependent variable is . But our model is . Here, this is the working model with the data. In other words, . Therefore, there should be some error in the model prediction which we are calling where is the true value and is the predicted value. This error term is normally distributed with mean 0 and variance 1. To get the best estimate of the parameters we can minimize the error term as much as possible. So, we define the residual sum of squares (RSS) as:

Using multivariate calculus we see

Setting the partial derivatives to zero we solve for as follows

and,

Therefore, we have the following

Simple Linear Regression slr is applicable for a single feature data set with contineous response variable.

import numpy as np   
import matplotlib.pyplot as plt   
from sklearn.linear\_model import LinearRegression

## Assumptions of Linear Regressions

* **Linearity:** The relationship between the feature set and the target variable has to be linear.
* **Homoscedasticity:** The variance of the residuals has to be constant.
* **Independence:** All the observations are independent of each other.
* **Normality:** The distribution of the dependent variable has to be normal.

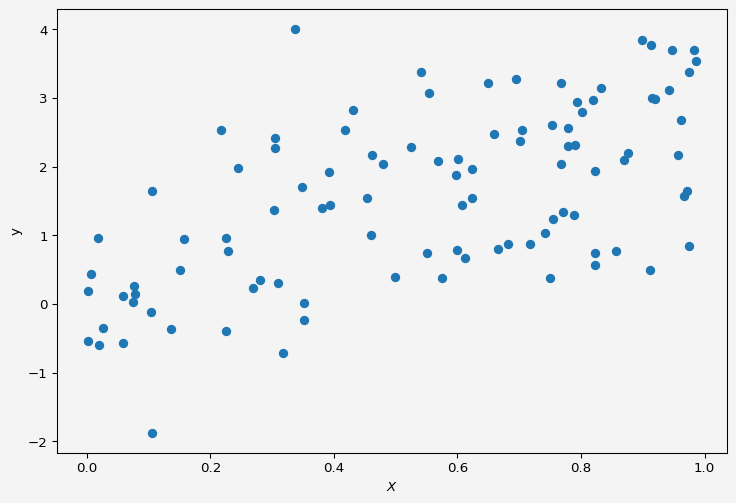
## Synthetic Data

To implement the algorithm, we need some synthetic data. To generate the synthetic data we use the linear equation where

X=np.random.random(100)  
y=2\*X+0.5+np.random.randn(100)

Note that we used two random number generators, np.random.random(n) and np.random.randn(n). The first one generates random numbers of values from the range (0,1) and the second one generates values from the standard normal distribution with mean 0 and variance or standard deviation 1.

plt.figure(figsize=(9,6))  
plt.scatter(X,y)  
plt.xlabel('$X$')  
plt.ylabel('y')  
plt.gca().set\_facecolor('#f4f4f4')   
plt.gcf().patch.set\_facecolor('#f4f4f4')  
plt.show()



## Model

We want to fit a simple linear regression to the above data.

slr=LinearRegression()

Now to fit our data and we need to reshape the input variable. Because if we look at ,

X

array([0.7883275 , 0.15761825, 0.07529538, 0.70055347, 0.82271156,  
 0.65872153, 0.77121189, 0.10567363, 0.00116792, 0.26927407,  
 0.97504326, 0.98671896, 0.64889315, 0.70436647, 0.30962975,  
 0.41835792, 0.54134812, 0.52455931, 0.10381667, 0.94813526,  
 0.56907126, 0.9569906 , 0.86956356, 0.75321578, 0.05913759,  
 0.57543484, 0.68247861, 0.35179219, 0.10571113, 0.75519707,  
 0.62345204, 0.91135784, 0.69485151, 0.55399232, 0.60150896,  
 0.6240679 , 0.45351372, 0.96696175, 0.94183314, 0.76838013,  
 0.83164992, 0.97410349, 0.02547007, 0.22914456, 0.02046789,  
 0.07749543, 0.59839183, 0.55040568, 0.74975822, 0.59862276,  
 0.35175982, 0.49912501, 0.76699467, 0.80107143, 0.00192689,  
 0.15118861, 0.21747373, 0.46008377, 0.27971315, 0.79073619,  
 0.81986377, 0.97131932, 0.91908873, 0.61244347, 0.39159103,  
 0.82296895, 0.39420905, 0.22578059, 0.22497153, 0.9140219 ,  
 0.38066391, 0.430839 , 0.00711581, 0.05771307, 0.74242321,  
 0.98321477, 0.89898564, 0.47967304, 0.13666663, 0.82269662,  
 0.34868445, 0.30352255, 0.01764527, 0.24483754, 0.96256478,  
 0.79284933, 0.07607341, 0.46113388, 0.71777806, 0.77922514,  
 0.30412777, 0.66553213, 0.77914009, 0.33651092, 0.31750325,  
 0.6075774 , 0.91423798, 0.85616323, 0.30484469, 0.87593732])

It is a one-dimensional array/vector but the slr object accepts input variable as matrix or two-dimensional format.

X=X.reshape(-1,1)  
X[:10]

array([[0.7883275 ],  
 [0.15761825],  
 [0.07529538],  
 [0.70055347],  
 [0.82271156],  
 [0.65872153],  
 [0.77121189],  
 [0.10567363],  
 [0.00116792],  
 [0.26927407]])

Now we fit the data to our model

slr.fit(X,y)  
slr.predict([[2],[3]])

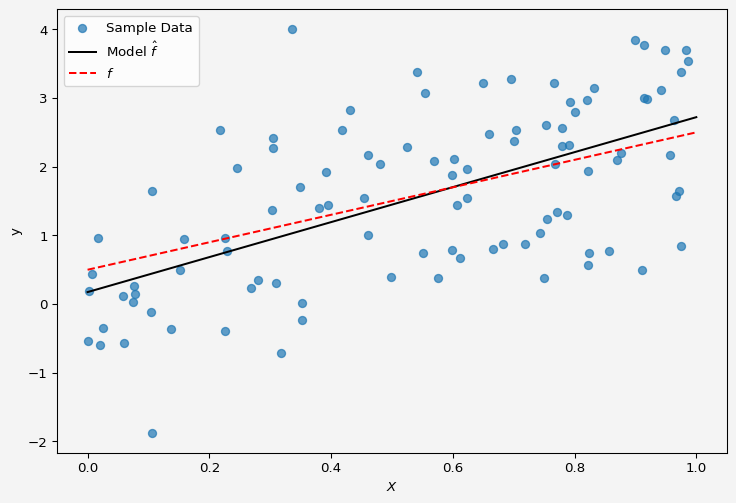
array([5.26711706, 7.81329375])

We have our and the corresponding values are from the above cell output, which are pretty close to the model .

intercept = round(slr.intercept\_,4)  
slope = slr.coef\_

Now our model parameters are: intercept np.float64(0.1748) and slope array([2.54617668]).

plt.figure(figsize=(9,6))  
plt.scatter(X,y, alpha=0.7,label="Sample Data")  
plt.plot(np.linspace(0,1,100),  
 slr.predict(np.linspace(0,1,100).reshape(-1,1)),  
 'k',  
 label='Model $\hat{f}$'  
)  
plt.plot(np.linspace(0,1,100),  
 2\*np.linspace(0,1,100)+0.5,  
 'r--',  
 label='$f$'  
)  
plt.xlabel('$X$')  
plt.ylabel('y')  
plt.legend(fontsize=10)  
plt.gca().set\_facecolor('#f4f4f4')   
plt.gcf().patch.set\_facecolor('#f4f4f4')  
plt.show()



So the model fits the data almost perfectly.

Up next [multiple linear regression](../../posts/multiplelinreg/index.qmd).

**Share on**

**You may also like**