

Linear Regression Using Python

Dr. Hakeem–Ur–Rehman
CSAS–PU

What is Regression?

Method of determining the statistical relationship between a *response (or output) and one or more predictor (or input) variables*.

$$Y = f(X_1, X_2, \dots, X_n)$$

Where 'Y' is the RESPONSE and X_1 to X_n are the PREDICTORS

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable

Simple & Multivariate (Multiple) Regression

Simple Linear Regression...

Is when the dependent variable is linearly proportional to just ONE independent variable.

Multivariate (Multiple) Regression...

May be viewed as an extension of *simple regression analysis (where only one predictor is involved)* to the situation where there is more than ONE predictor to be considered.

Simple Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

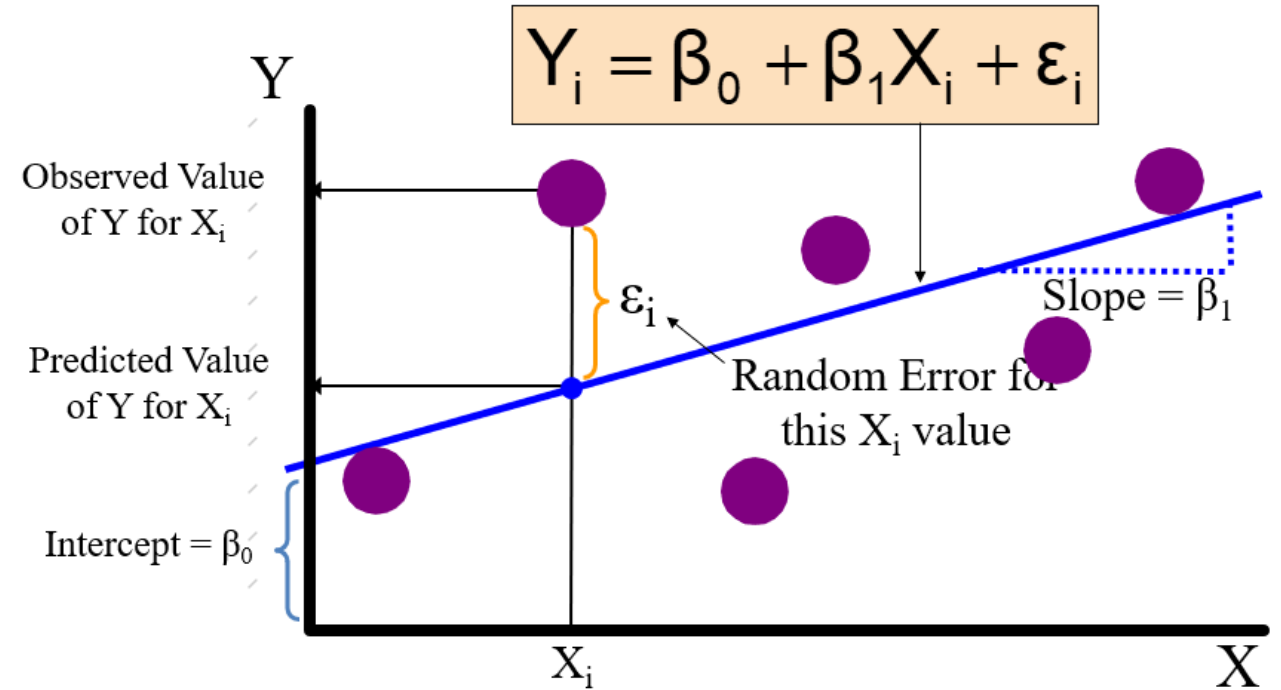
Independent Variable

Random Error term

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

Random Error component



Simple Linear Regression Equation (Prediction Line) – Estimated Regression Model

- The simple linear regression equation provides an **estimate** of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one-unit increase in X

Multivariate (Multiple) Linear Regression

The Multiple Regression Model

- Idea: Examine the linear relationship between
- 1 dependent (Y) & *2 or more independent variables (X_i)*

Multiple Regression Model with k Independent Variables:

The diagram shows the equation $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$ with arrows pointing from labels to its parts: 'Y-intercept' points to β_0 , 'Population slopes' points to the β coefficients, and 'Random Error' points to ε_i .

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

The coefficients of the multiple regression model are estimated using sample data

The diagram shows the estimated equation $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}$ with arrows pointing from labels to its parts: 'Estimated (or predicted) value of Y' points to \hat{Y}_i , 'Estimated intercept' points to b_0 , and 'Estimated slope coefficients' points to the b coefficients.

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}$$

Multivariate (Multiple) Linear Regression

Album Sales 2: Data

Imagine that the record company executive was now interested in extending the model of album sales to incorporate other variables. He decides to measure two new variables:

- 1. the number of times songs from the album are played on Radio 1 during the week prior to release (**airplay**); and
- 2. the attractiveness of the band (**attract**)

Before an album is released, the executive notes the amount spent on advertising, the number of times songs from the album are played on radio the week before release, and the attractiveness of the band. He does this for 200 different albums (each made by a different band).

Attractiveness was measured by asking a random sample of the target audience to rate the attractiveness of each band on a scale from 0 (hideous potato-heads) to 10 (gorgeous objects). The mode attractiveness given by the sample was used in the regression (because he was interested in what the majority of people thought, rather than the average of people’s opinions).

The data are in a file called **Album Sales 2.csv**.

	A	B	C	D	
1	adverts	sales	airplay	attract	
2	10.256	330	43	10	
3	985.685	120	28	7	
4	1445.563	360	35	7	
5	1188.193	270	33	7	
6	574.513	220	44	5	
7	568.954	170	19	5	
8	471.814	70	20	1	
9	537.352	210	22	9	
10	514.068	200	21	7	
11	174.093	300	40	7	
12	1720.806	290	32	7	
13	611.479	70	20	2	
14	251.192	150	24	8	
15	97.972	190	38	6	
16	406.814	240	24	7	
17	265.398	100	25	5	
18	1323.287	250	35	5	
19	196.65	210	36	8	
20	1326.598	280	27	8	
21	1380.689	230	33	8	
22	792.345	210	33	7	
23	957.167	230	28	6	
24	1789.659	320	30	9	
25	656.137	210	34	7	
26	613.697	230	49	7	
27	313.362	250	40	8	
28	336.51	60	20	4	
29	1544.899	330	42	7	
Album Sales 2					+

QUESTIONS

