Introduction to Linear Regression

Linear regression model

Definition: A statistical method to model the relationship between a dependent variable and one or more independent variables. Example: Predicting house prices based on square footage.

Assumptions

- Linearity: The relationship between independent and dependent variables should be linear.
- Independence: Observations should be independent of each other.
- Homoscedasticity: The variance of the residuals should remain constant.
- Normality: The residuals should be normally distributed.

Applications

- Economics: Predicting GDP based on different factors.
- Business: Forecasting sales based on advertising spend.
- Medicine: Predicting patient outcomes based on treatment variables.

How to calculate simple linear regression line

Formula (kindly note that there's an alternative formula to calculate 'b' the slope of the regression line)

- **Regression line y= a + bx**

Where

y = is the predicted value of the dependent variable

a = is the y-intercept.

b = is the slope of the regression line.

x = is the value of the independent variable

- **b=** $\frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{(n(\Sigma x^2)-(\Sigma x)^2)}$

- **a = $\bar{y}$ - b $\bar{x}$** (note $\bar{y}$ = mean of y, and $\bar{x}$ = mean of x)

Example

A school coach is interested in predicting a student's 1600m (1 mile) race finish time based on the number of laps they run per week during training. The coach collects data from 5 track team members:

Student	**Laps/week (x)**	**1600m Time (minutes) (y)**
P	10	7.5
Q	15	7.0
R	20	6.5
S	12	7.3
T	18	6.7

Using simple linear regression analysis, determine the equation of the regression line and predict the 1600m time for a student who trains for 22 laps per week.

Solution

Student				
P	10	7.5	75	100
Q	15	7.0	105	225
R	20	6.5	130	400
S	12	7.3	87.6	144
T	18	6.7	120.6	324

- $b = \frac{5(518.2)-(75)(35)}{(5(1193)-(75)^2)} = \frac{-34}{340} = -0.1$

$\bar{y} = \frac{35}{5} = 7; \bar{x} = \frac{75}{5} = 15$

a = 7-(-0.1)15= 8.5

Regression line y= 8.5 - 0.1x

predicting the 1600m time for a student who trains for 22 laps per week.

X = 22

y= 8.5 - 0.1(22)= **5.8**

Activity 1

A researcher is interested in predicting an employee's daily productivity (tasks completed) based on the number of cups of coffee they consume during the day. The researcher gathers data from 6 employees:

Employee	Coffee cups/day (x)	Tasks Completed (y)
A1	1	5
B1	2	8
C1	3	11
D1	1.5	7
E1	2.5	10
F1	3.5	12

Using simple linear regression analysis, determine the equation of the regression line and predict the number of tasks completed by an employee who consumes 4 cups of coffee.

Activity 2

A teacher wants to predict a student's test score based on the number of hours they study the night before the test. The teacher records data from 6 students:

Student	Study hours/night (x)	Test scores (y)
John	1	60
Dan	2	70
Bob	3	85
Taina	1.5	65
Sally	2.5	78
Jane	3.5	90

Using simple linear regression analysis, determine the equation of the regression line and predict the test score for a student who studies for 4 hours.