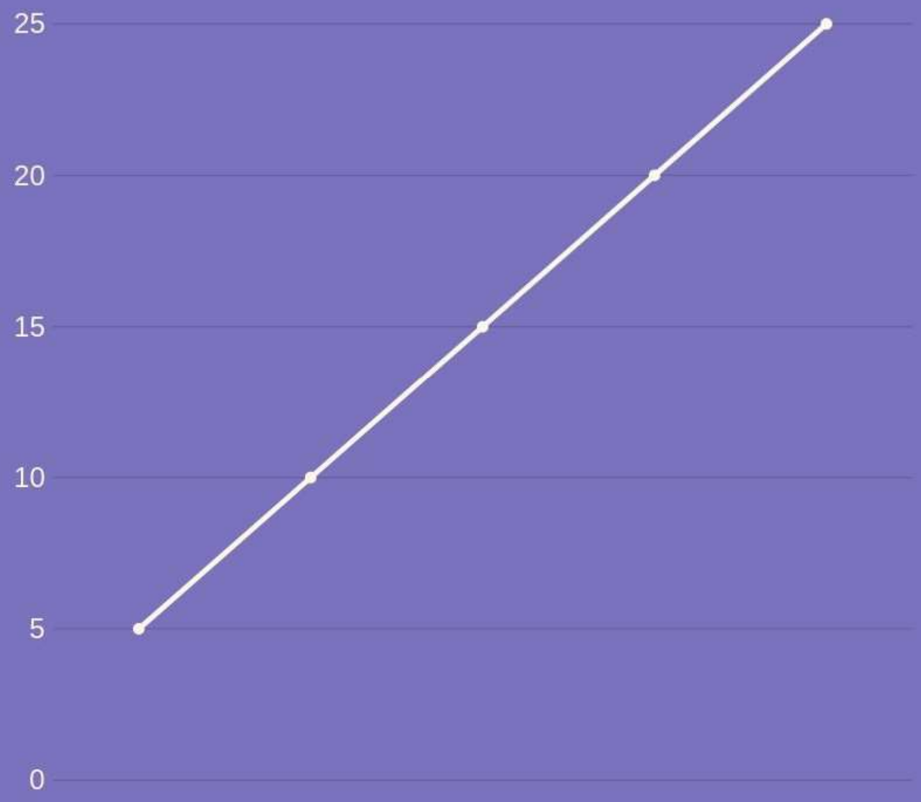


Linear Regression in Machine Learning

By Anil Kumar
APSSDC

Linear Regression

Linear Regression is one of the most simple yet widely used statistical Machine Learning technique. The linear regression machine learning algorithm tries to map one or more independent variable (features) to a dependent variable (scalar output). In this session, you will be learning about:

- Different types of linear regression in machine learning.
- A bit of statistics/mathematics behind it.
- And what kind of problems you can solve with linear regression.

What is Regression?

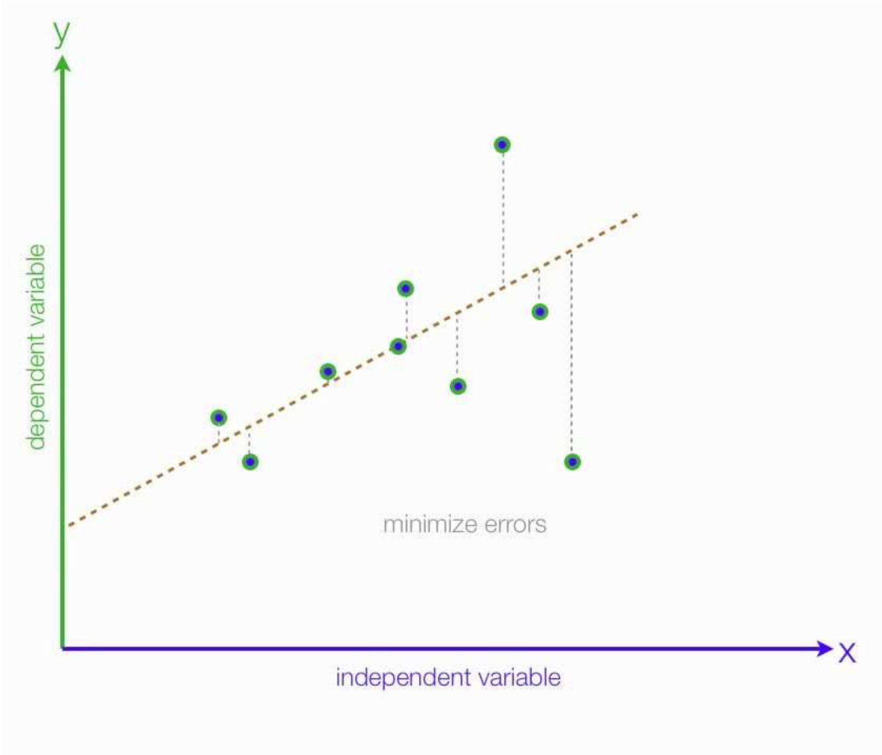
- Function: a mathematical relationship enabling us to predict what values of one variable (Y) correspond to given values of another variable (X).
- Y: is referred to as the *dependent variable*, the *response variable* or the *predicted variable*.
- X: is referred to as the *independent variable*, the *explanatory variable* or the *predictor variable*.

Thus Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable.

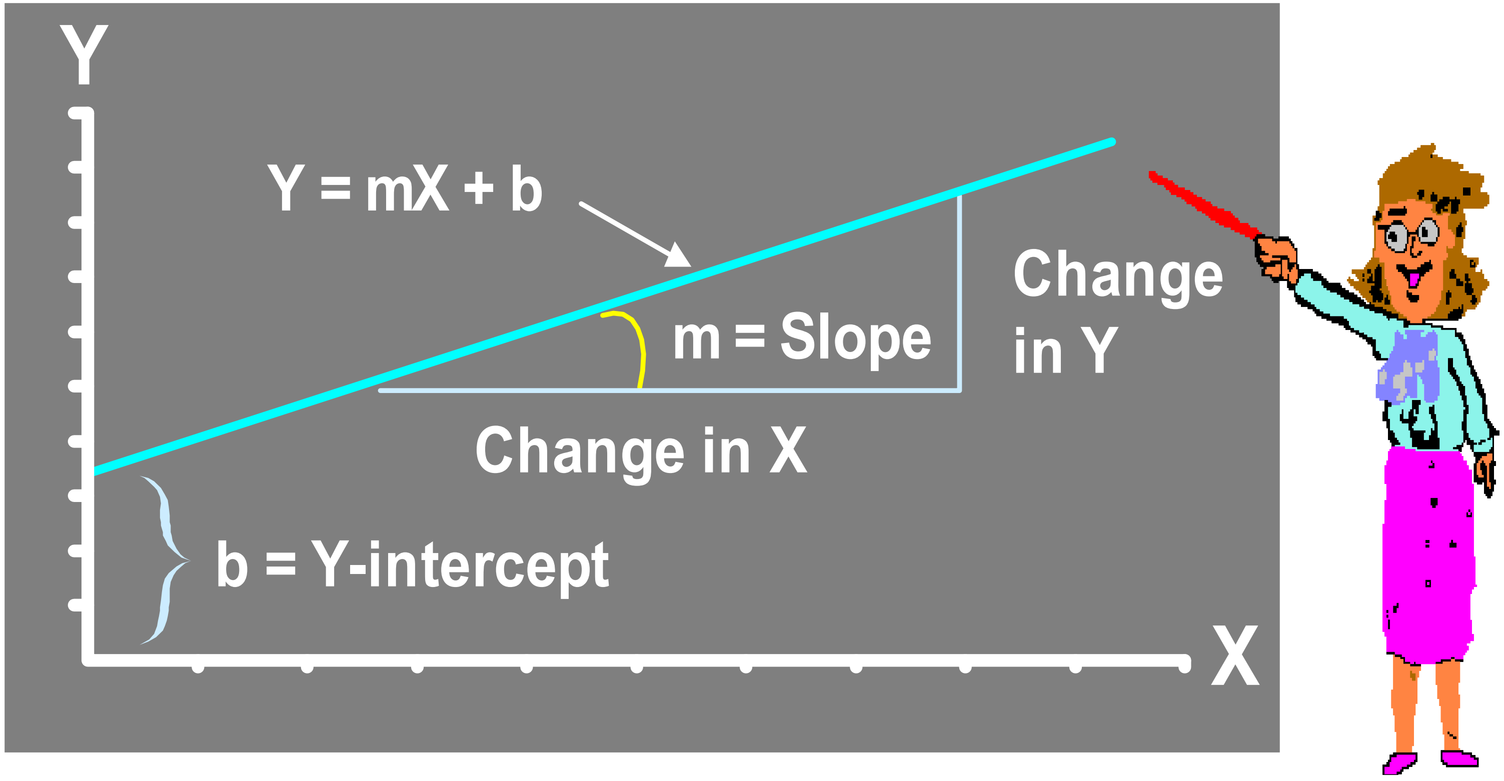
Contd.

A typical Linear Regression model can be represented in the form :

- y = b1x + b0 where b1 is slope and b0 is the intercept.



Linear Equations



Linear Regression Model

Relationship Between Variables Is a Linear Function

Y-Intercept

Slope

Random Error

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

Dependent (Response) Variable

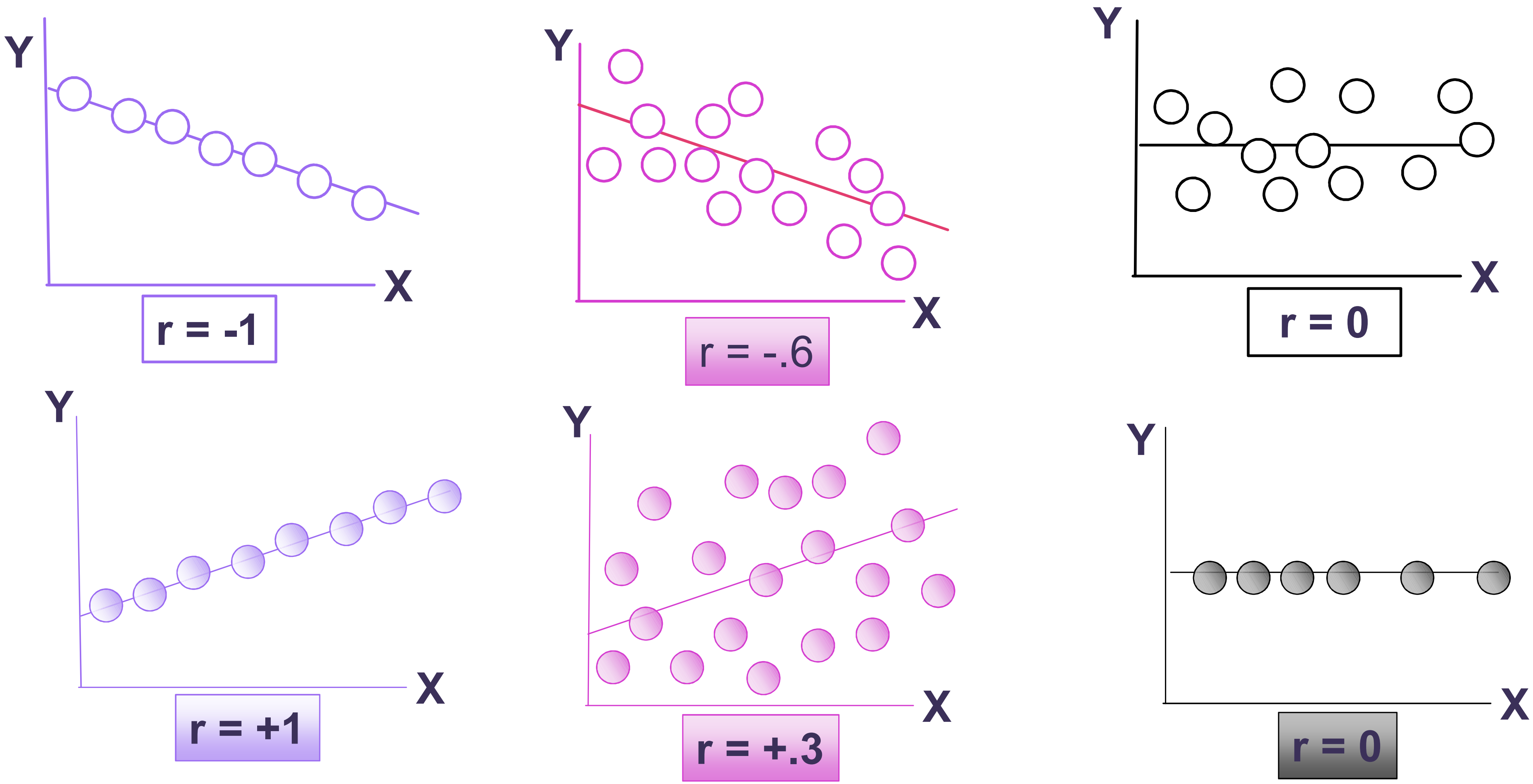
Independent (Explanatory) Variable

- $\beta_1 > 0 \Rightarrow$ **Positive Association**
- $\beta_1 < 0 \Rightarrow$ **Negative Association**
- $\beta_1 = 0 \Rightarrow$ **No Association**

Correlation

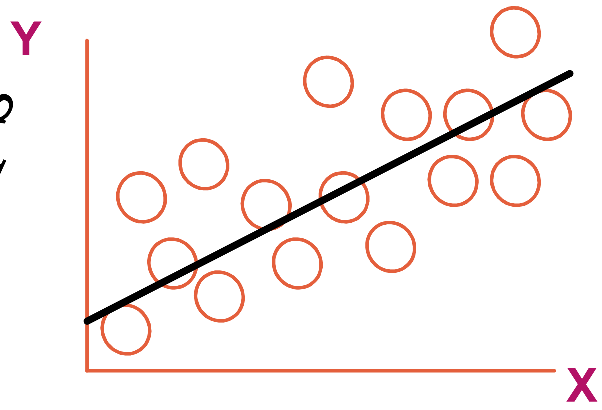
- Measures the relative strength of the *linear* relationship between two variables Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

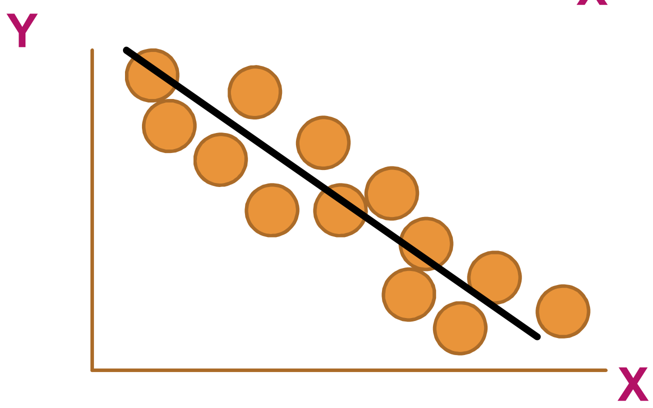
Scatter Plots of Data with Various Correlation Coefficients



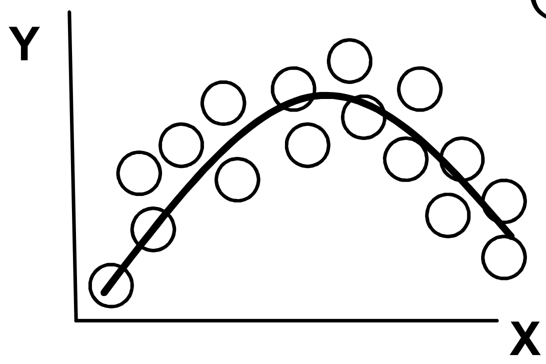
Linear Correlation

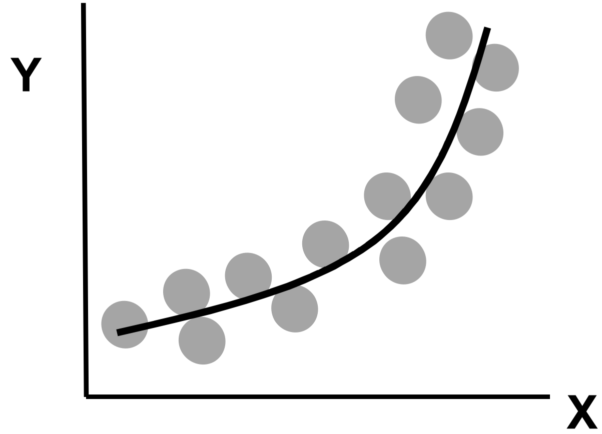
Linear relationships



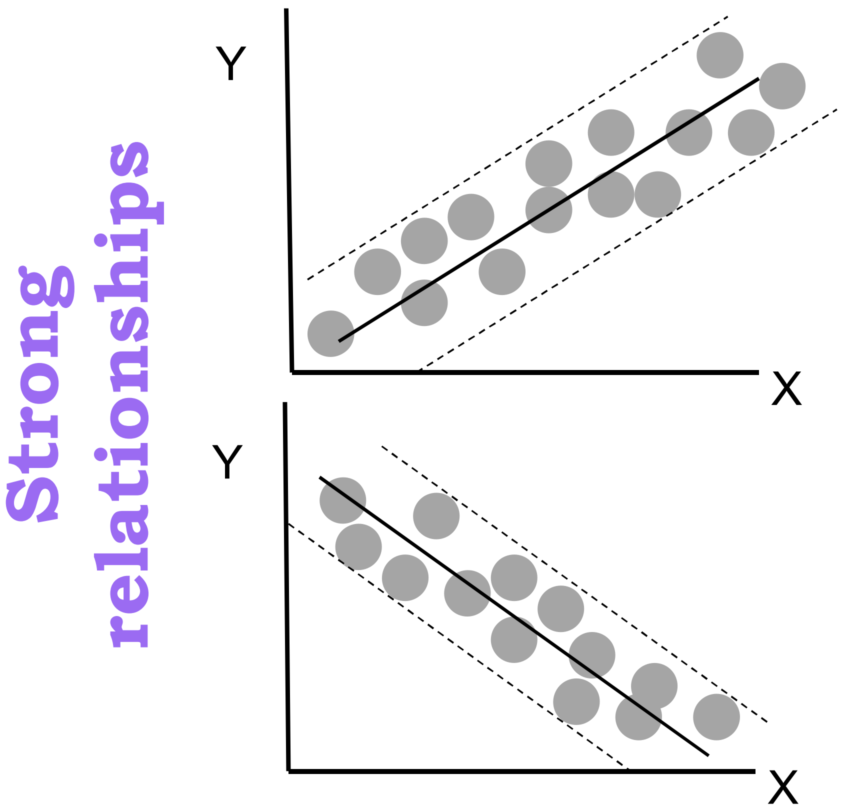


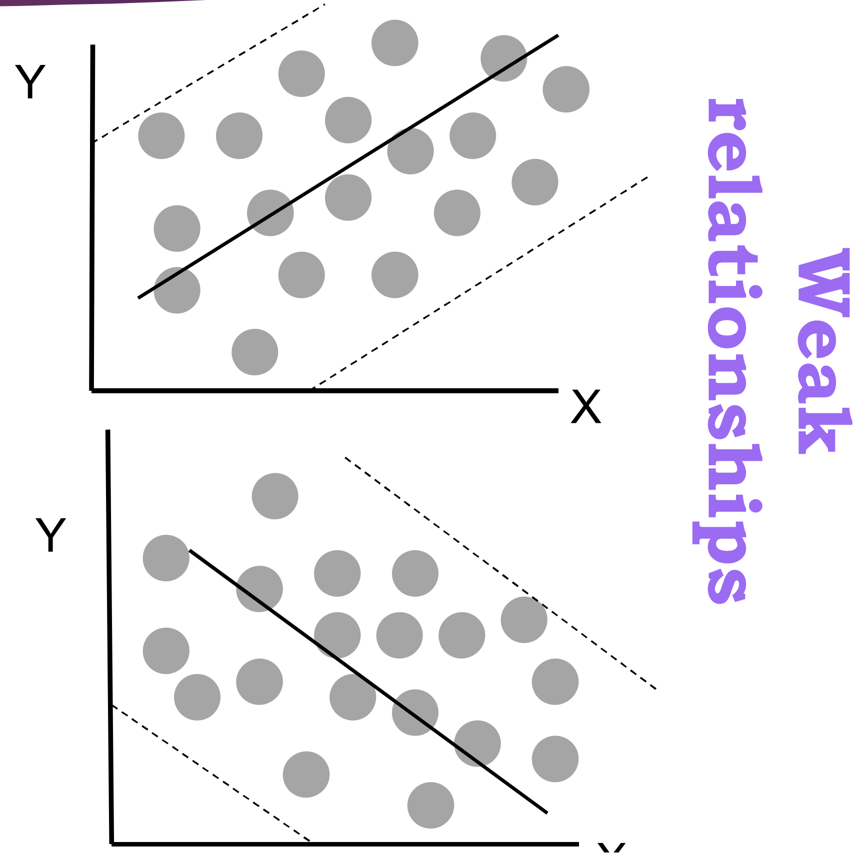
Curvilinear relationships





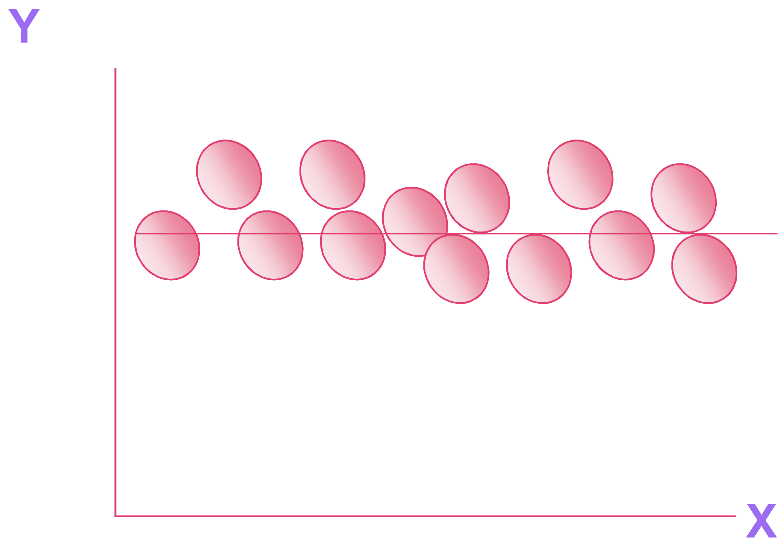
Linear Correlation

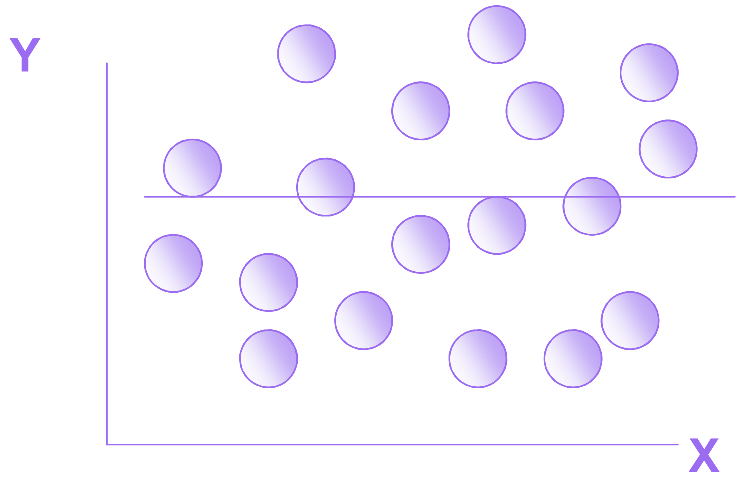




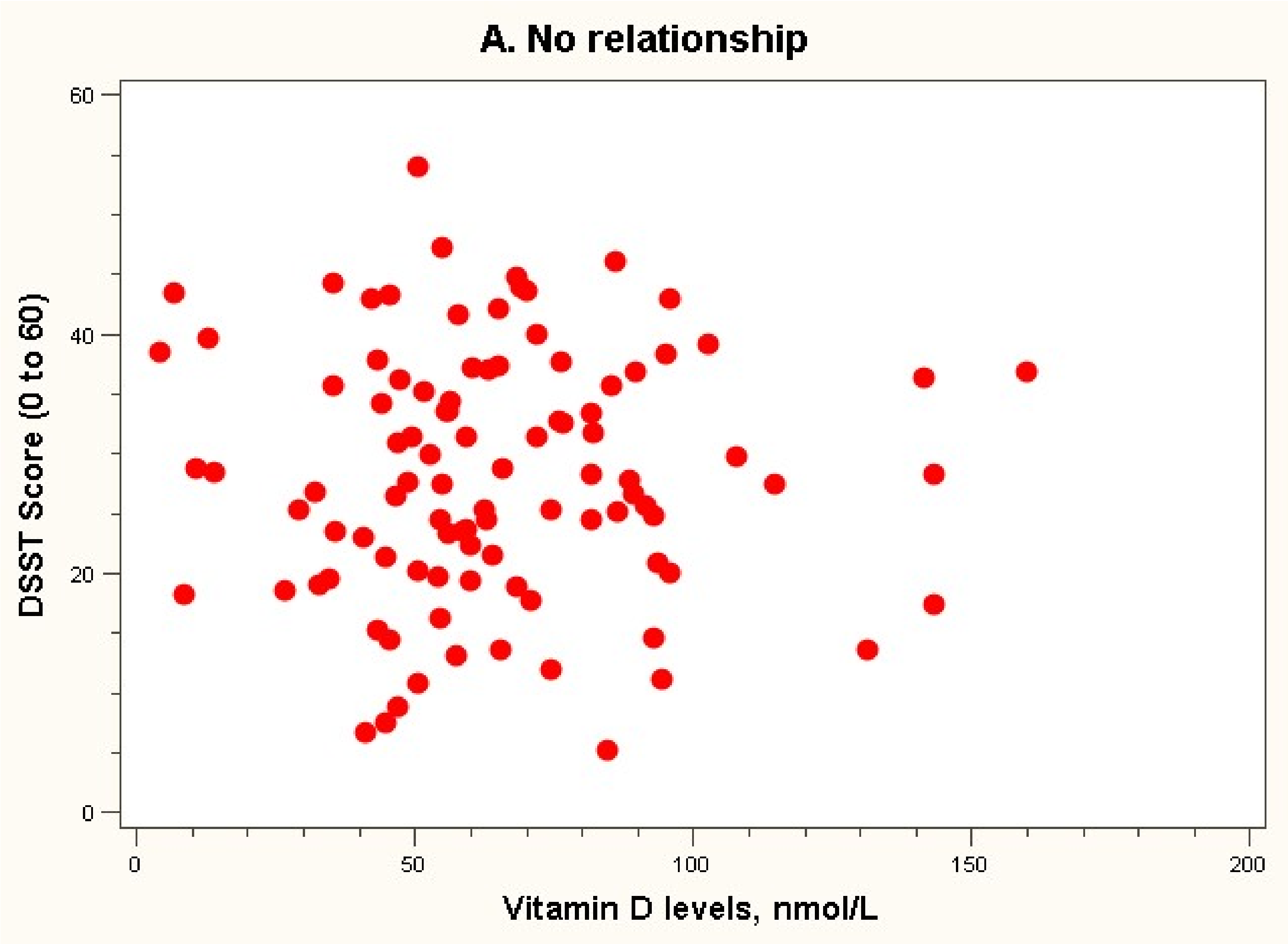
Linear Correlation

No relationship

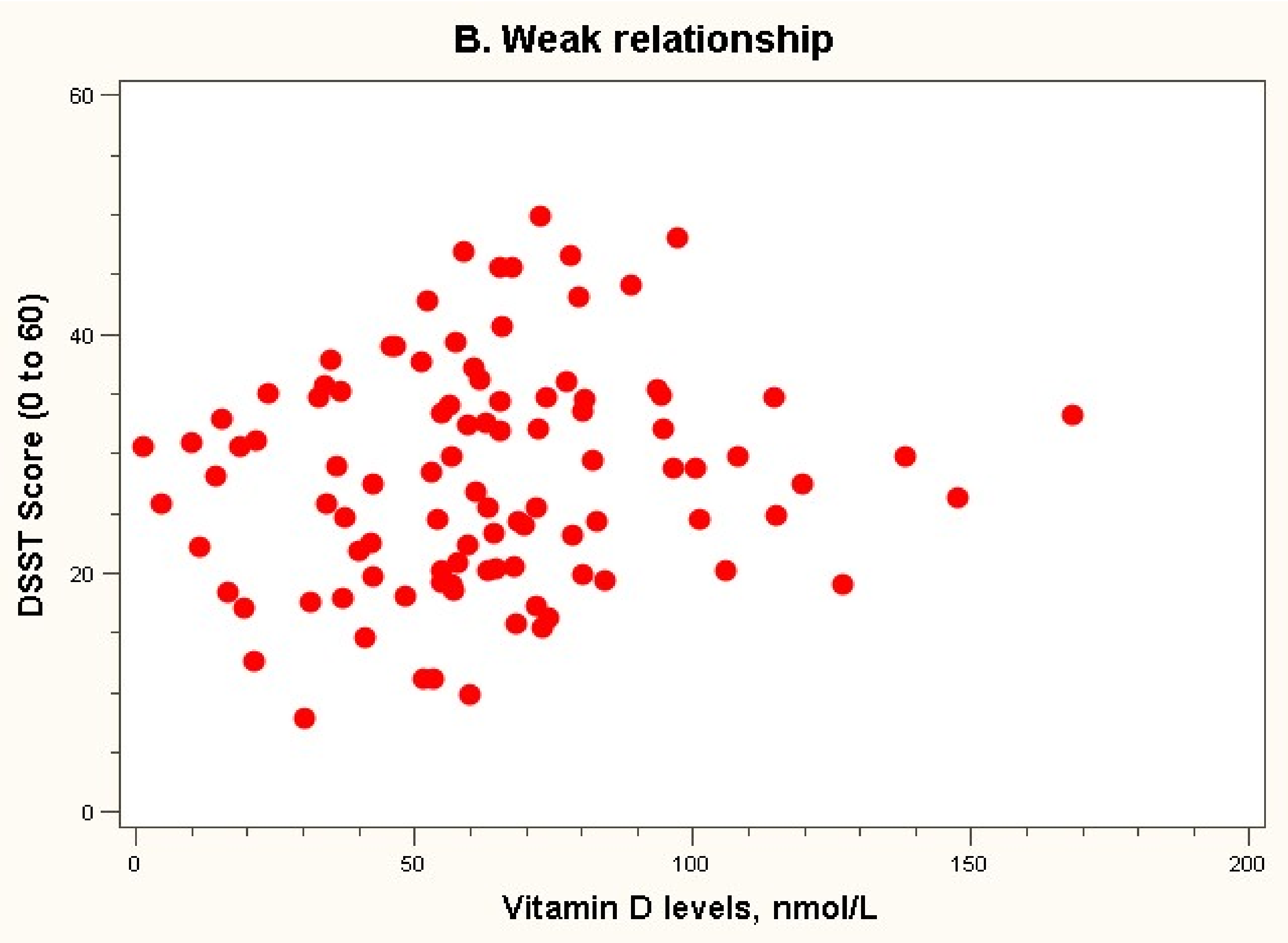




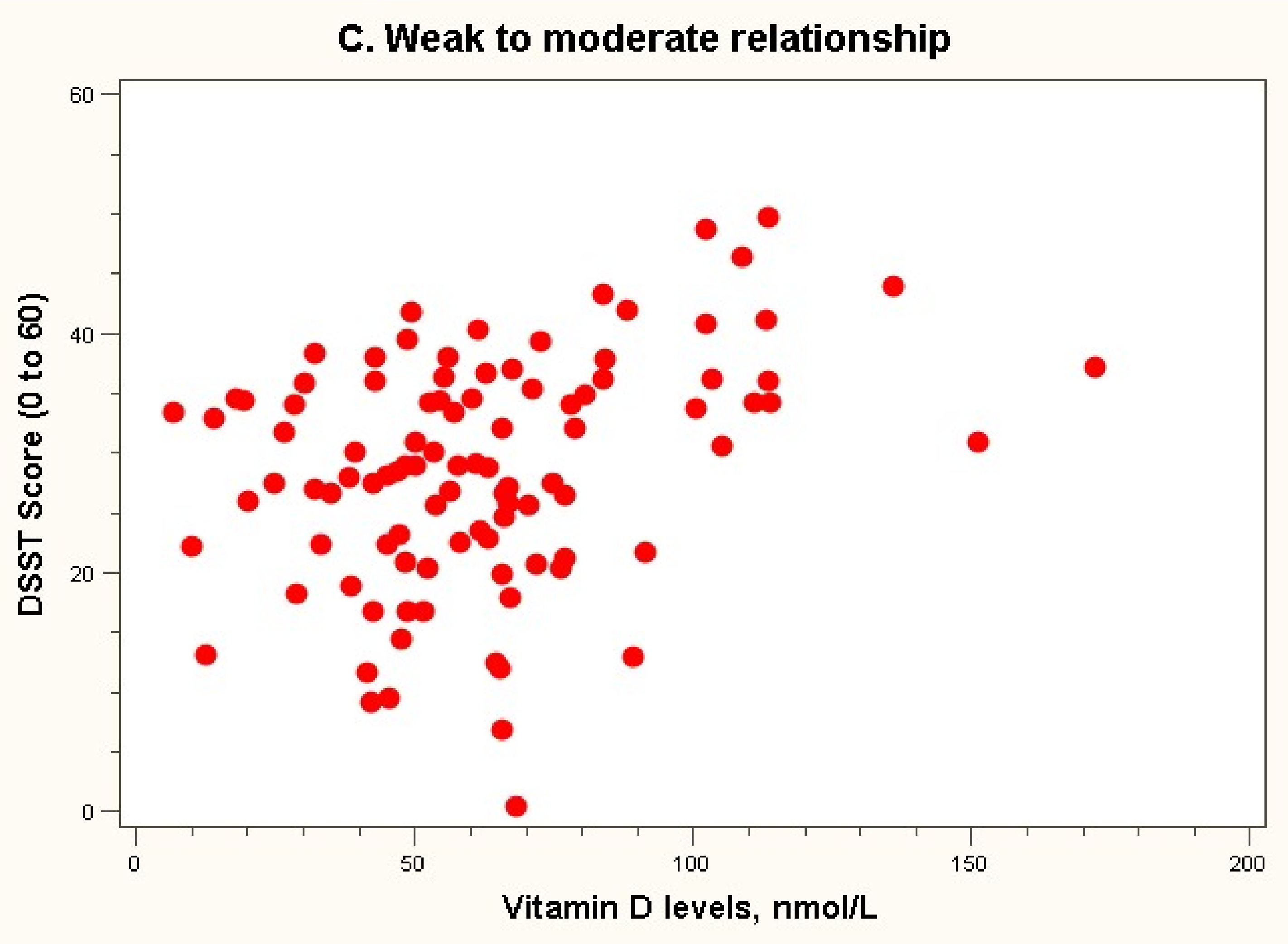
Dataset 1: No Relationship



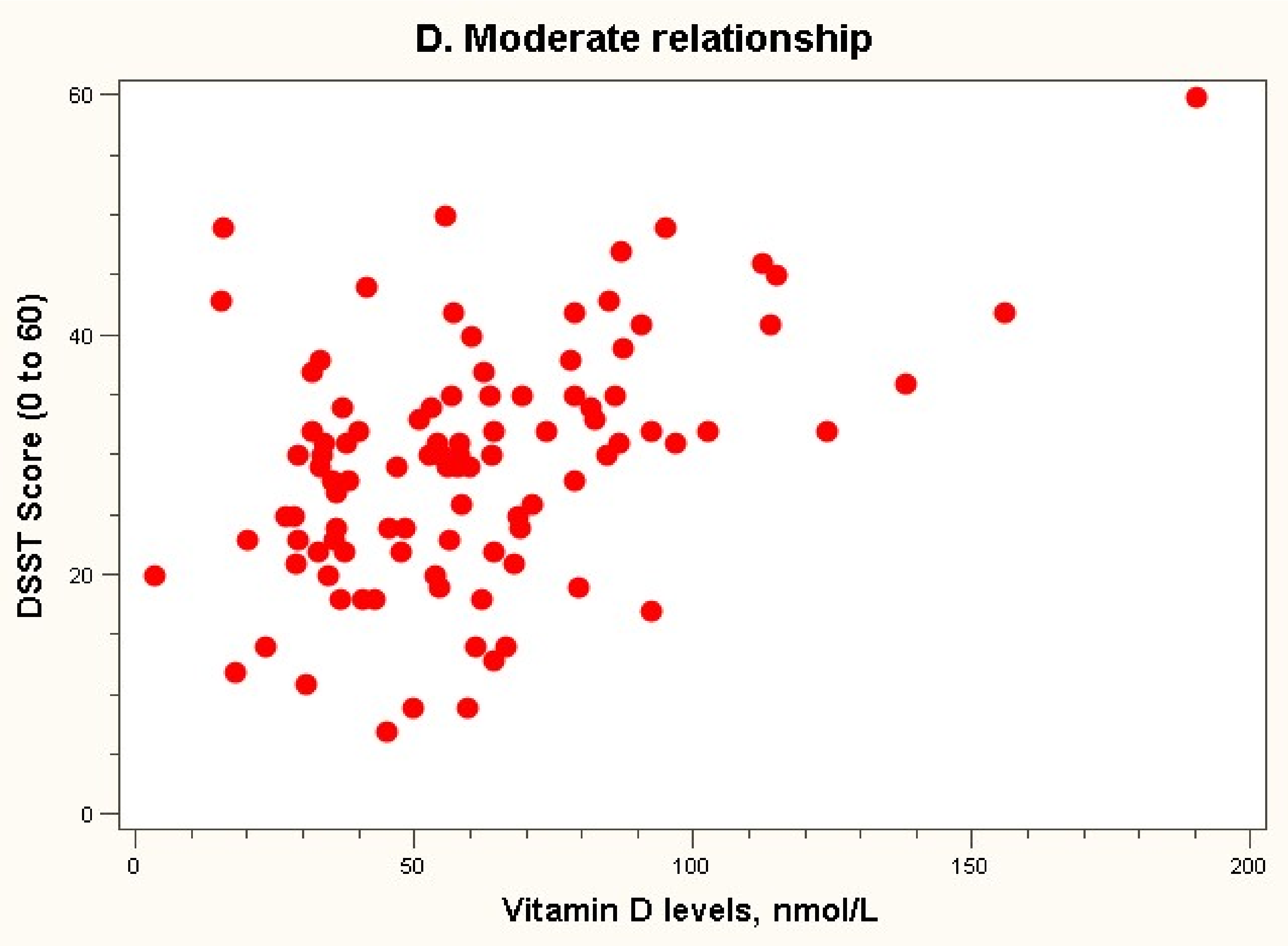
Dataset 2: weak relationship



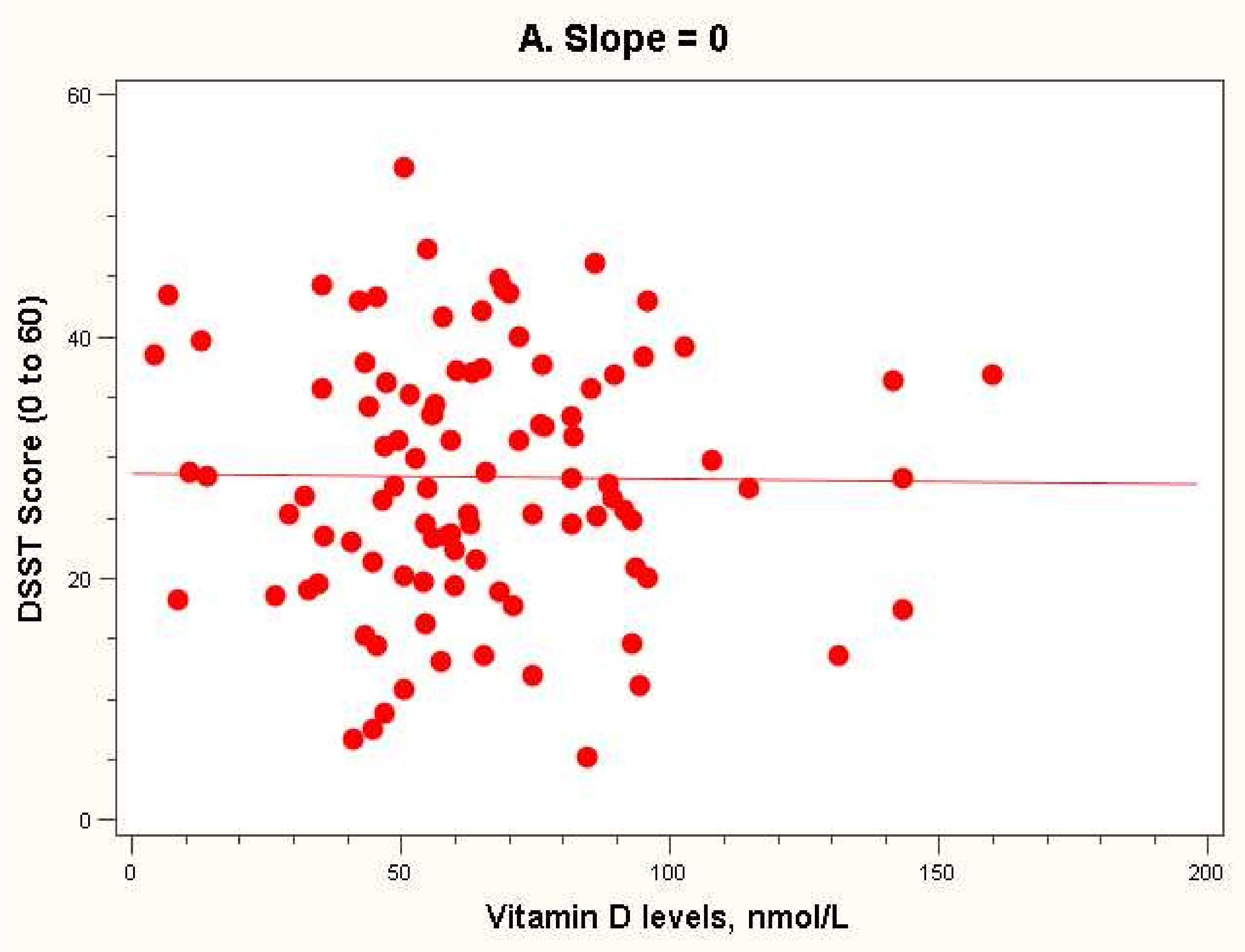
Dataset 3: weak to moderate relationship



Dataset 4: moderate relationship



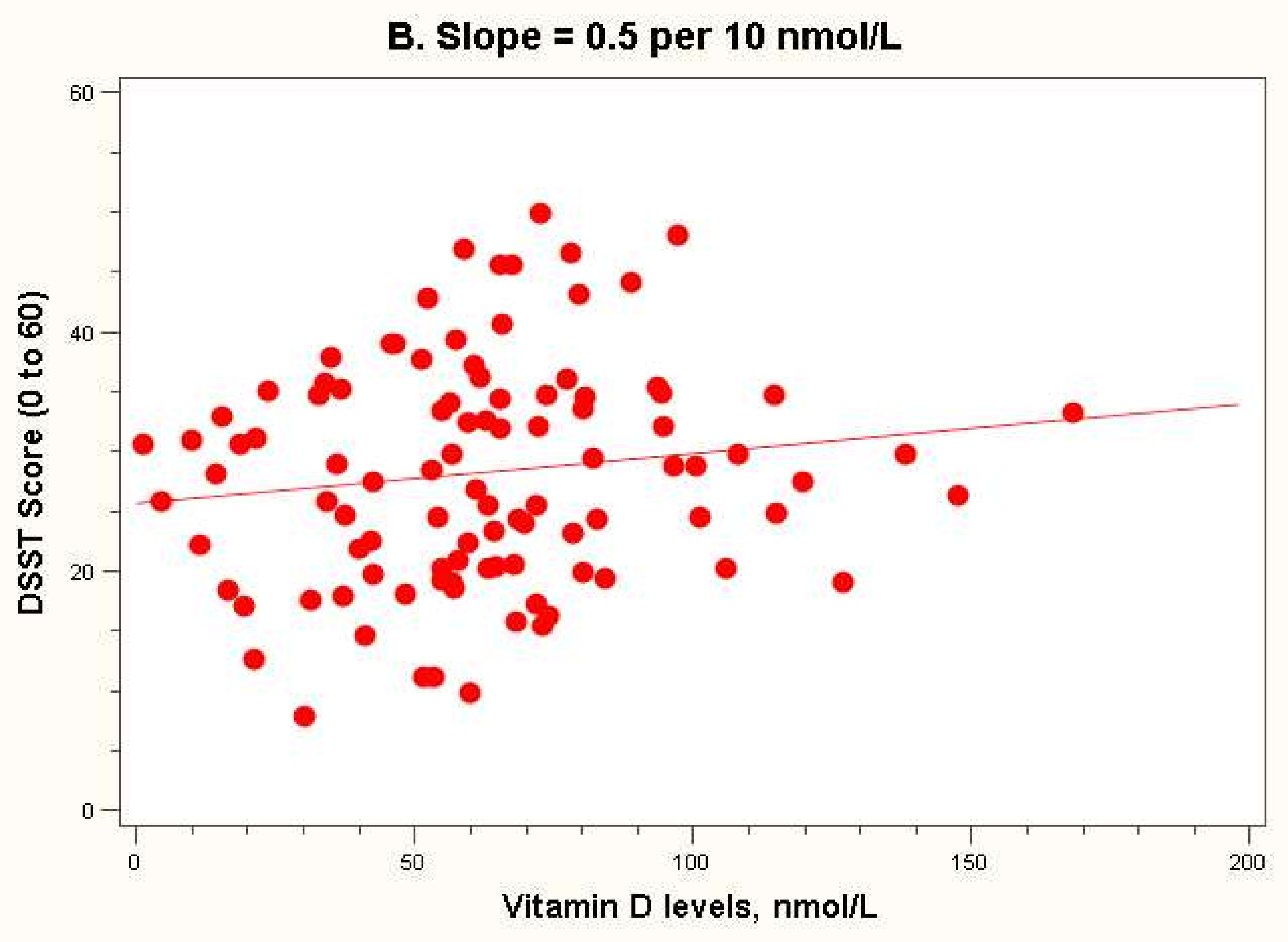
The “Best fit” line



Regression equation:

E(Yi) = 28 + 0*vit Di (in 10 nmol/L)

The "Best fit" line

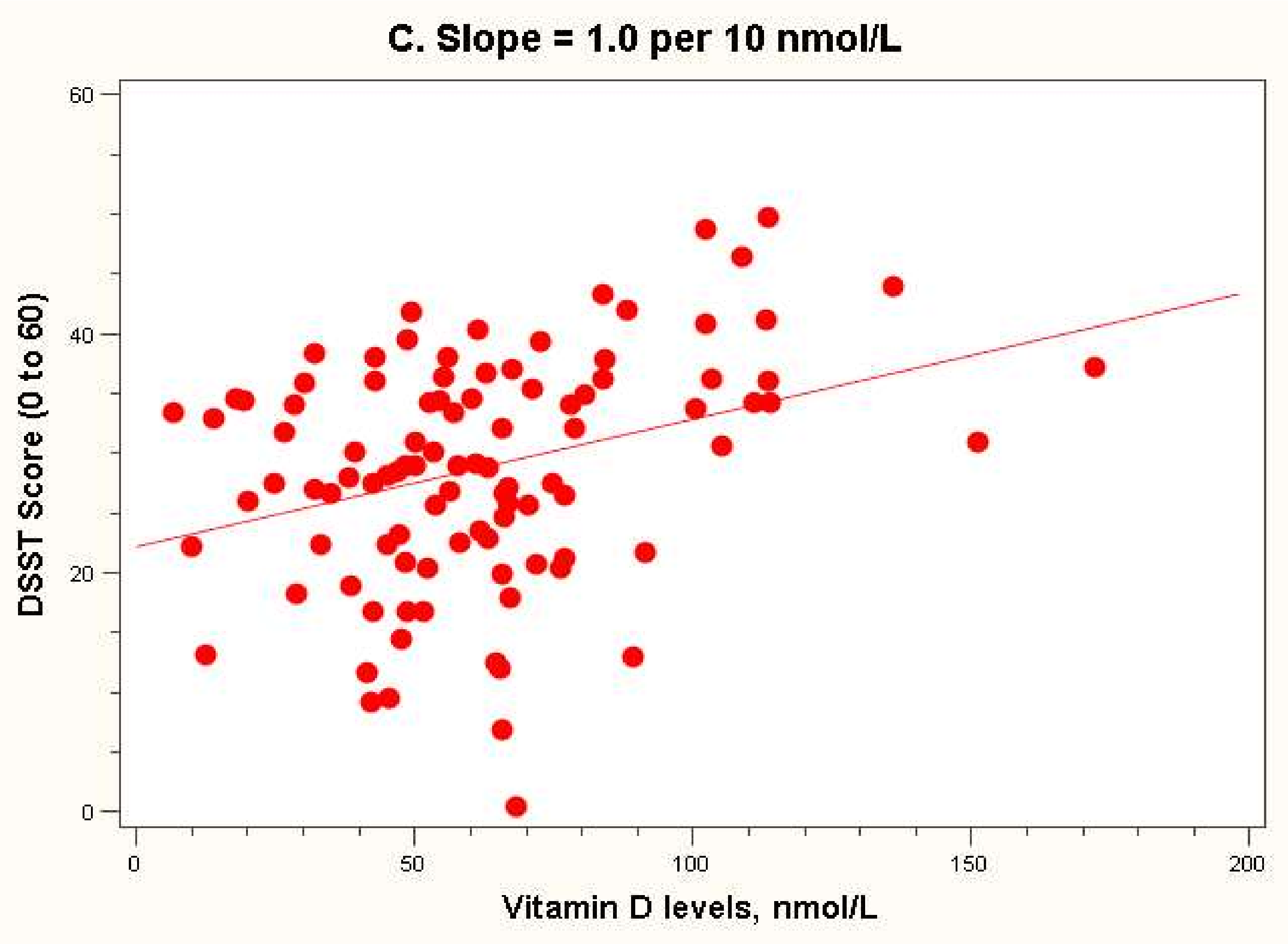


Note how the line is a little deceptive; it draws your eye, making the relationship appear stronger than it really is!

Regression equation:

E(Yi) = 26 + 0.5*vit Di (in 10 nmol/L)

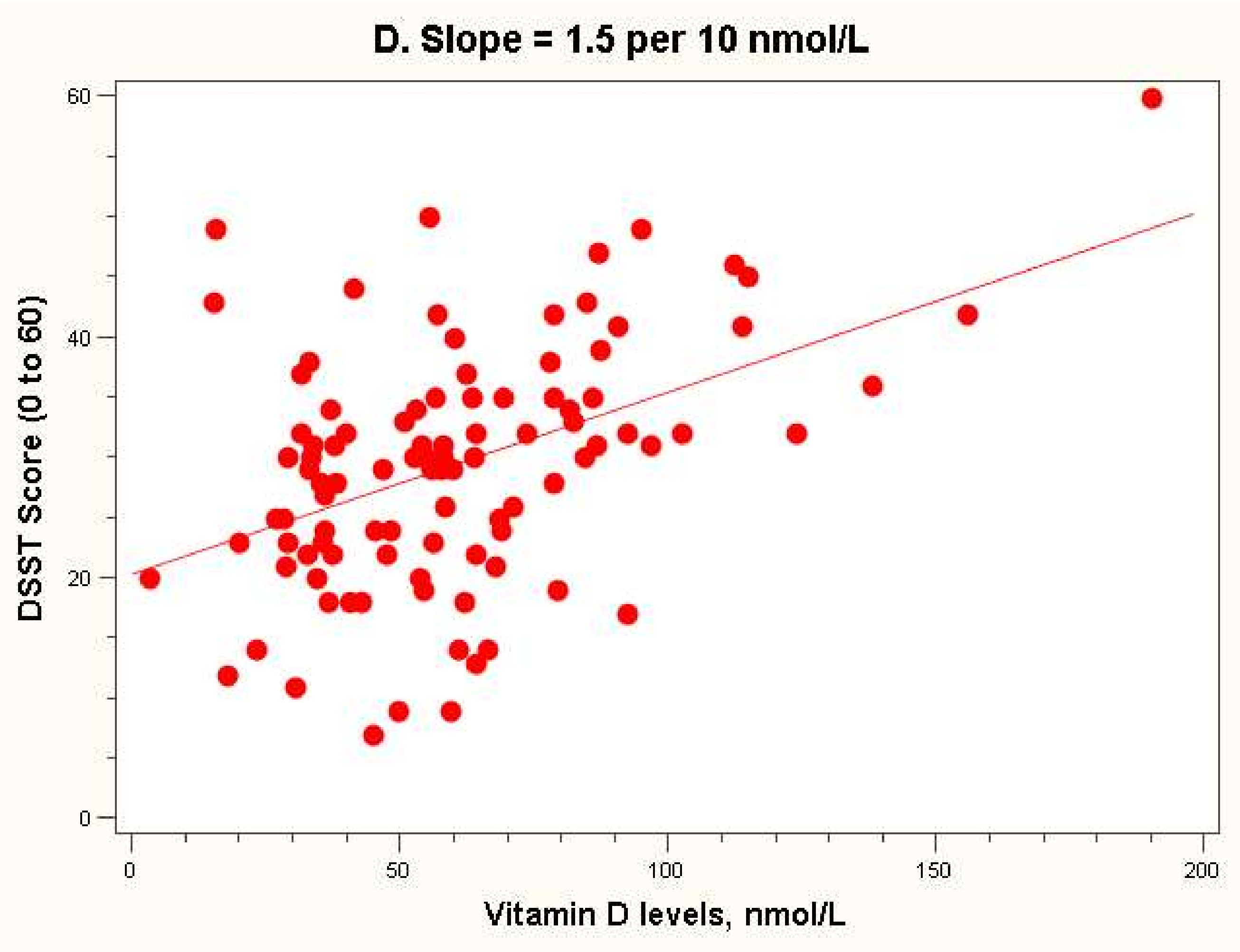
The "Best fit" line



Regression equation:

E(Yi) = 22 + 1.0*vit Di (in 10 nmol/L)

The “Best fit” line



Regression equation:

E(Yi) = 20 + 1.5*vit Di (in 10 nmol/L)

Note: all the lines go through the point (63, 28)!

Steps in Regression Analysis

- Examine the scatterplot of the data.
  I. Does the relationship look linear?
  II. Are there points in locations they shouldn't be?
  III. Do we need a transformation?
- Assuming a linear function looks appropriate, estimate the regression parameters.
  I. How do we do this? (Method of Least Squares)
- If there is a significant linear relationship, estimate the response, Y, for the given values of X, and compute the residuals

Regression Analysis

- Thus we have the regression formula as :

  Y = MX + C + error(e).

Initially we calculate the value for slope and predict the values of Y for any given X values we have.

$$\text{Slope(M)} = \sum_{i=0}^{len(X)} \frac{(X_i - X_{mean}) * (Y_i - Y_{mean})}{(X_i - X_{mean})^2}$$

Thus we calculate the C value and find out the "Line of Regression".

Regression Analysis

- Next our job is to reduce the distance between the actual value and the predicted value or in other words reduce the error between the actual and predicted value. Thus the line with least error will be the **"Best Fit Line"**.
- In order to check it out we calculate the "Coefficient of Determination".

$$Mean\ Squared\ value\ (R^2) = \sum_{i=0}^{len(X)} \frac{(Y_{pred} - Y_{mean})^2}{(Y - Y_{mean})^2}$$

- Thus our ultimate aim is to reduce the error i.e. distance between the actual and predicted values.