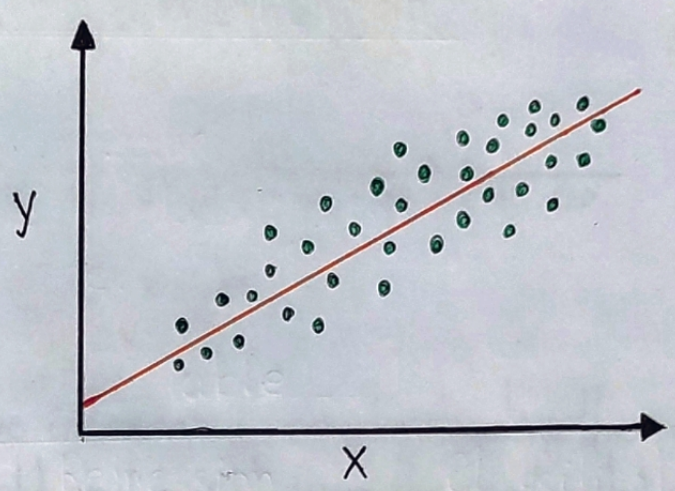
Simple Linear Regression

Simple linear regression is a statistical method you can use to understand the relationship between two variables, x and y.

One variable, x, is known as the predictor variable.

The other variable, y, is known as the response variable.

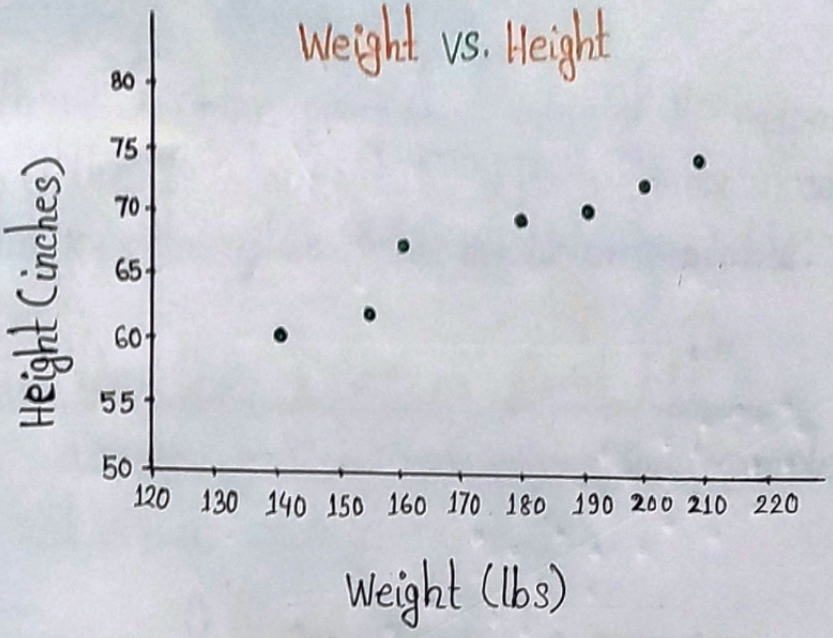


For example, suppose we have the following dataset with the weight and height of seven individuals:

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

Let weight be the predictor variable and let height be the response variable.

If we graph these two variables using a scatterplot, with weight on the x-axis and height on the y-axis, here's what it would look like:



Suppose we're interested in understanding the relationship between weight and height. From the scatterplot we can clearly see that as weight increases, height tends to increase as well, but to actually quantify this relationship between weight and height, we need to use linear regression.

Using linear regression, we can find the line that best "fits" our data. This line is known as the least squares regression line and it can be used to help us understand the relationships between weight and height.

Usually you would use software like Microsoft Excel, SPSS, or a graphing calculator to actually find the equation for this line.

The formula for the line of best fit is written as:

$$\hat{y} = b_0 + b_1 x$$

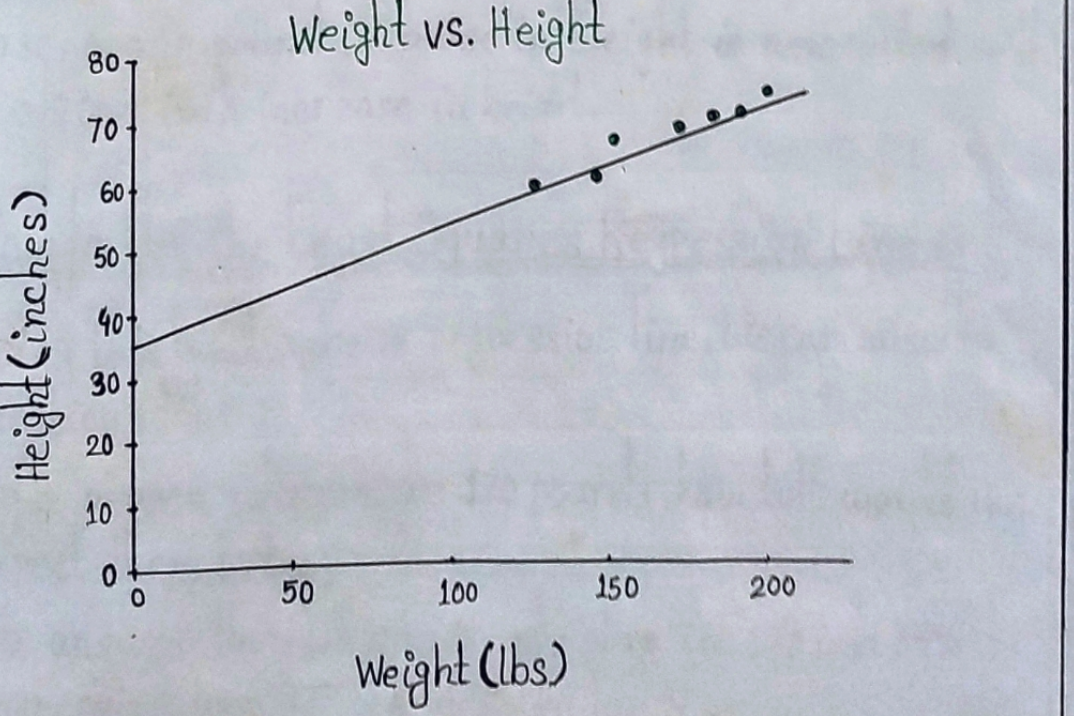
Where $\hat{y}$ is the predicted value of the response variable, b_0 is the y-intercept, b_1 is the regression coefficient, and x is the value of the predictor variable.

Finding the "Line of Best Fit"

We assume a fictitious values for example of the least squares regression line.

$$\hat{y} = 32.7830 + 0.2001x$$

If we zoom out on our scatterplot from earlier and added this line to the chart, here's what it would look like:



Notice how our data points are scattered closely around this line. That's because this least squares regression lines is the best fitting line for our data out of all the possible lines we could draw.

How to Interpret a Least Squares Regression Line

Here is how to interpret this least squares regression line: $\hat{y} = 32.7830 + 0.2001x$

$b_0 = 32.7830$. This means when the predictor variable weight is zero pounds, the predicted height is 32.7830 inches. Sometimes the value for b_0 can be useful to know, but in this specific example it doesn't actually make sense to interpret b_0 since a person can't weight zero pounds.

$b_1 = 0.2001$. This means that a one unit increase in x is associated with a 0.2001 unit increase in y. In this case, a one pound increase in weight is associated with a 0.2001 inch increase in height.

How to Use the Least Squares Regression Line

Using this least squares regression line, we can answer questions like:

For a person who weighs 170 pounds, how tall would we expect them to be?

To answer this, we can simply plug in 170 into our regression line for x and solve for y:

$$\hat{y} = 32.7830 + 0.2001(170) = 66.8 \text{ inches}$$

The Coefficient of Determination

One way to measure how well the least squares regression line "fits" the data is using the coefficient of determination, denoted as R^2.

The coefficient of determination is the proportion of the variance in the response variable that can be explained by the predictor variable.

The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

An R^2 between 0 and 1 indicates just how well the response variable can be explained by the predictor variable. For example, an R^2 of 0.2 indicates that 20% of the variance in the response variable can be explained by the predictor variable; an R^2 of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable.

Notice in our output from earlier we got an R^2 of 0.9311, which indicates that 93.11% of the variability in height

can be explained by the predictor variable of weight.

Assumptions of Linear Regression

For the results of a linear regression model to be valid and reliable, we need to check that the following four assumptions are met :

1. Linear relationship : There exists a linear relationship between the independent variable, x, and the dependent variable, y.

2. Independence : The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

3. Homoscedasticity : The residuals have constant variance at every level of x.

4. Normality : The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.