Machine Learning

Introduction

What is Machine Learning

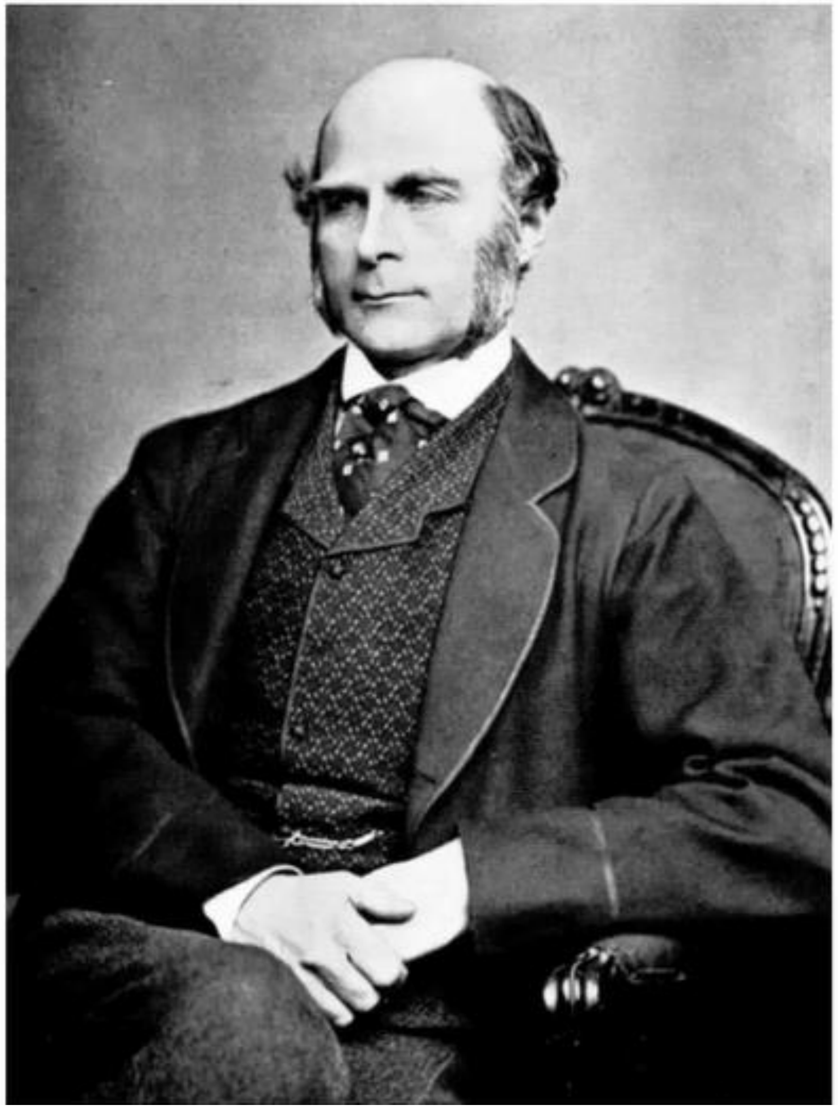
Machine learning is a method of data analysis that automates analytical model building.
Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

Why it is used for

- Fraud detection.
- Web search results.
- Real-time ads on web pages
- Credit scoring and next-best offers.
- Prediction of equipment failures.
- New pricing models.
- Network intrusion detection.
- Recommendation Engines
- Customer Segmentation
- Text Sentiment Analysis
- Predicting Customer Churn
- Pattern and image recognition.
- Email spam filtering.
- Financial Modeling

Introduction to Linear Regression

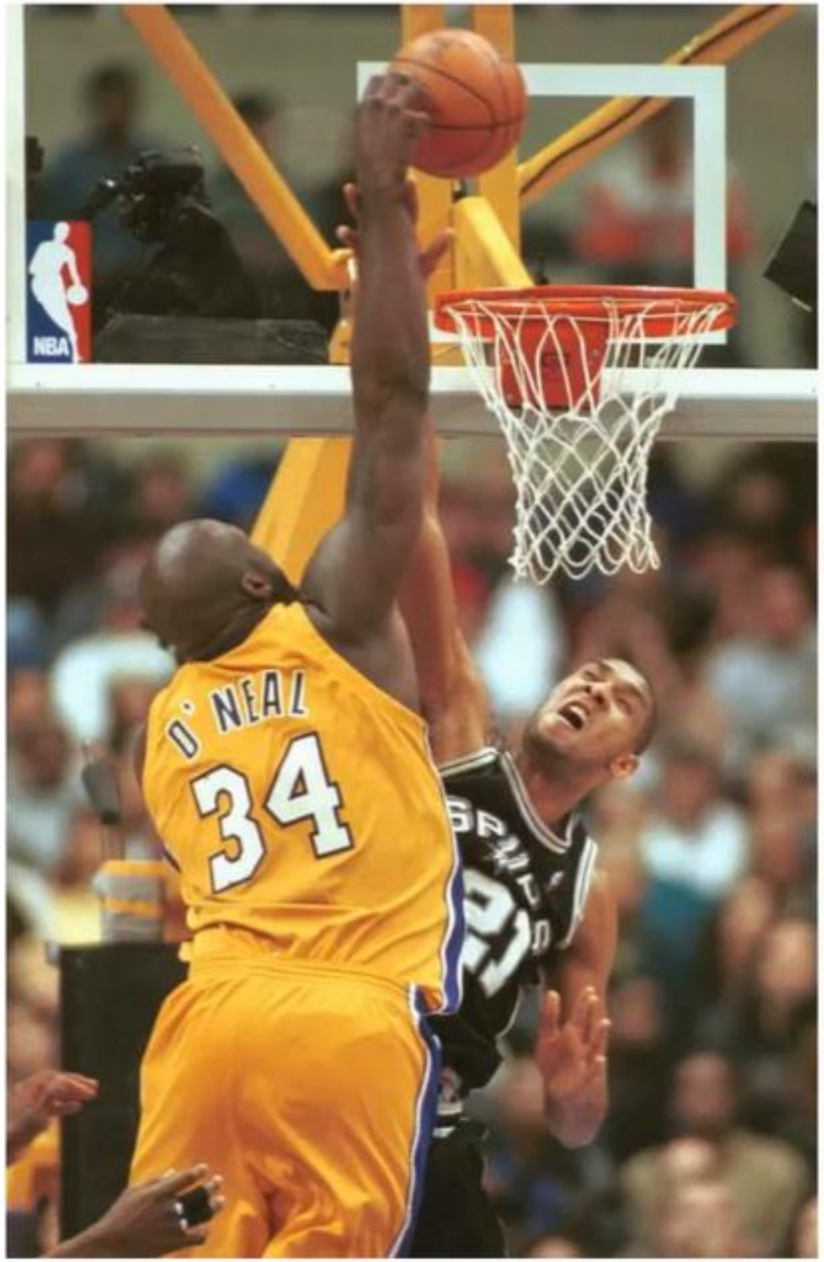
This all started in the 1800s with a guy named Francis Galton. Galton was studying the relationship between parents and their children. In particular, he investigated the relationship between the heights of fathers and their sons.



Example

Let's take Shaquille O'Neal as an example. Shaq is really tall:7ft 1in (2.2 meters).

If Shaq has a son, chances are he'll be pretty tall too. However, Shaq is such an anomaly that there is also a very good chance that his son will be **not be as tall as Shaq**.



What is Linear Regression?

- Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
- simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Linear Model

- Comparison of Two Models, and Consistent changes between x and y

Rate of Change or Slope

- Slope → Length and Steepness of the line
- The More Electricity → Increase the Bill
- Example 2: How much water you pour based Plant Grow
  - The Amount of Water based Plant grow
  - Here Height of Plant is output, and water pour is input
  - So Output always depends on input
  - Here Water is Input(Independent), Height of plant is output (Dependent)
- In Our case which is Independent and Dependent Variable

Formula for Linear

$Y = a + bX$,

where Y is the dependent variable (that's the variable that goes on the Y axis),

X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

How to Find Linear Regression

- From the above table, Σx = 247, Σy = 486, Σxy = 20485, Σx2 = 11409, Σy2 = 40022. n is the sample size (6, in our case).
- Below Formula for a and b

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

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

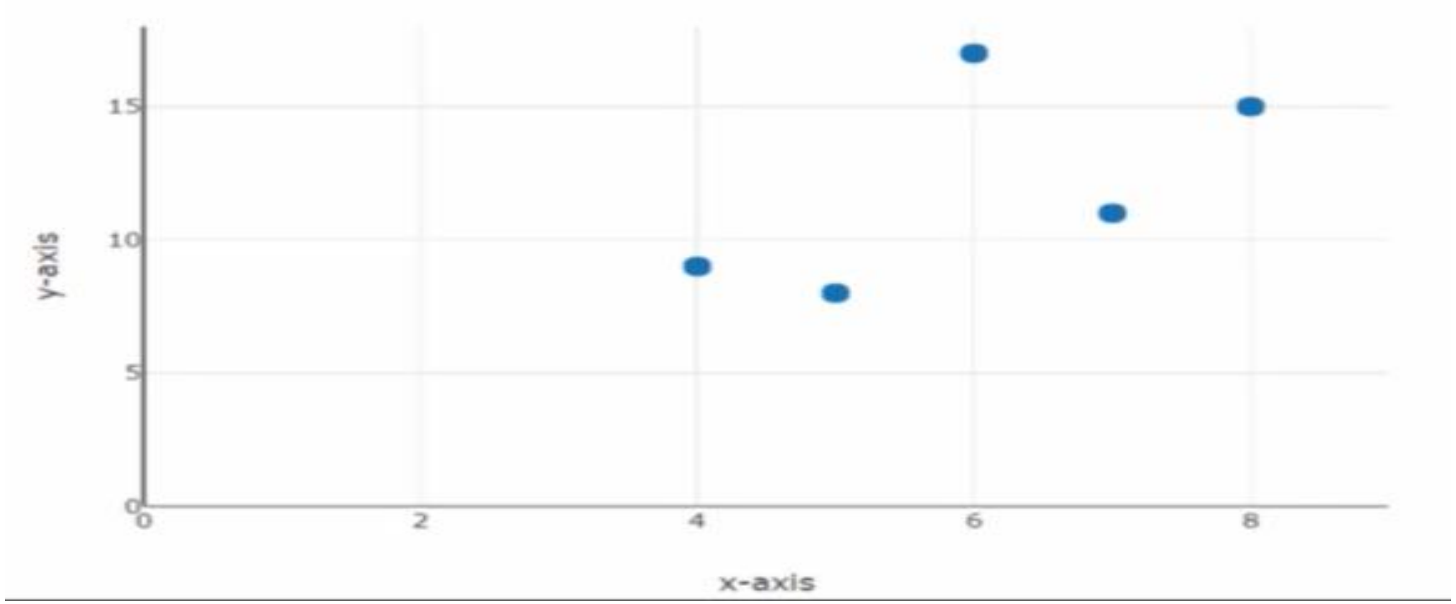
a = 65.1416
b = .385225

- y' = a + bx

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Linear Regression

The goal of **regression** is to develop an equation or formula that **best describes** the relationship between variables.

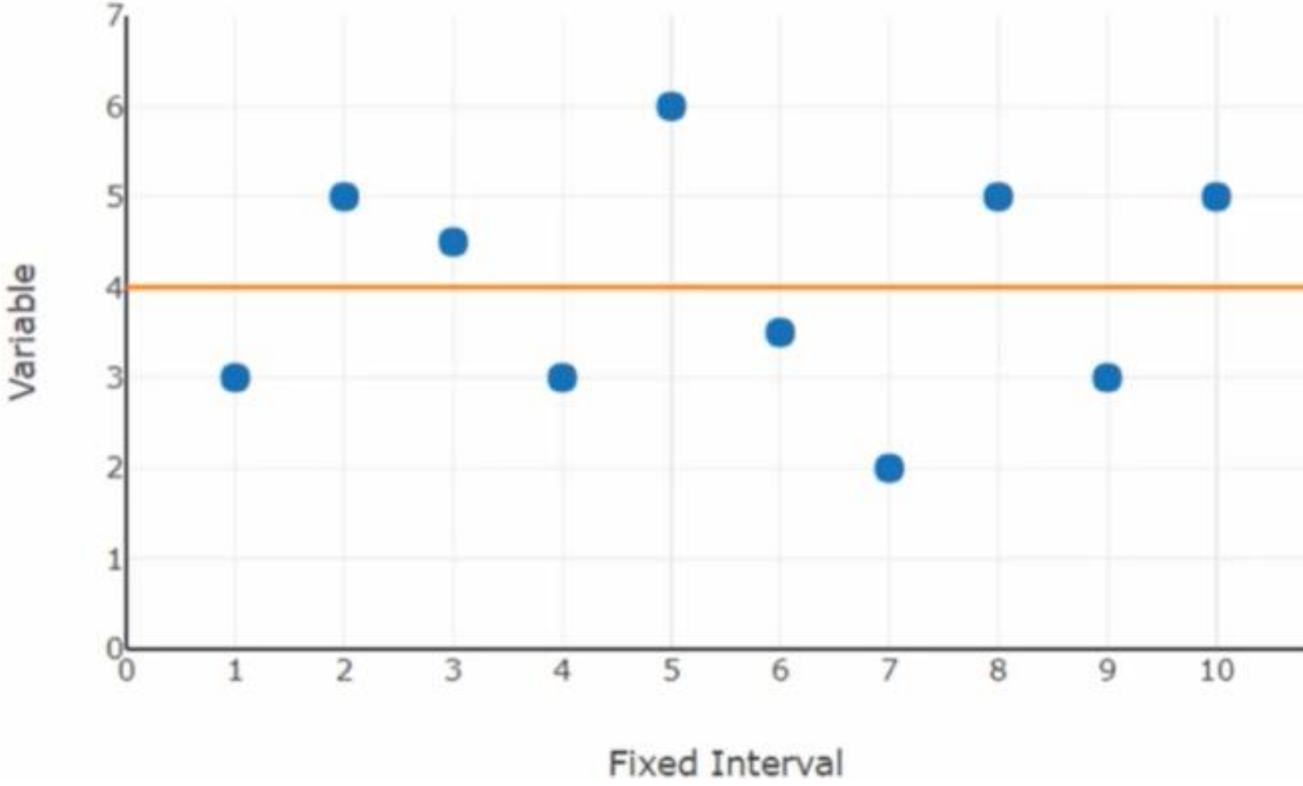


Linear Regression

How do we find a best-fit line?

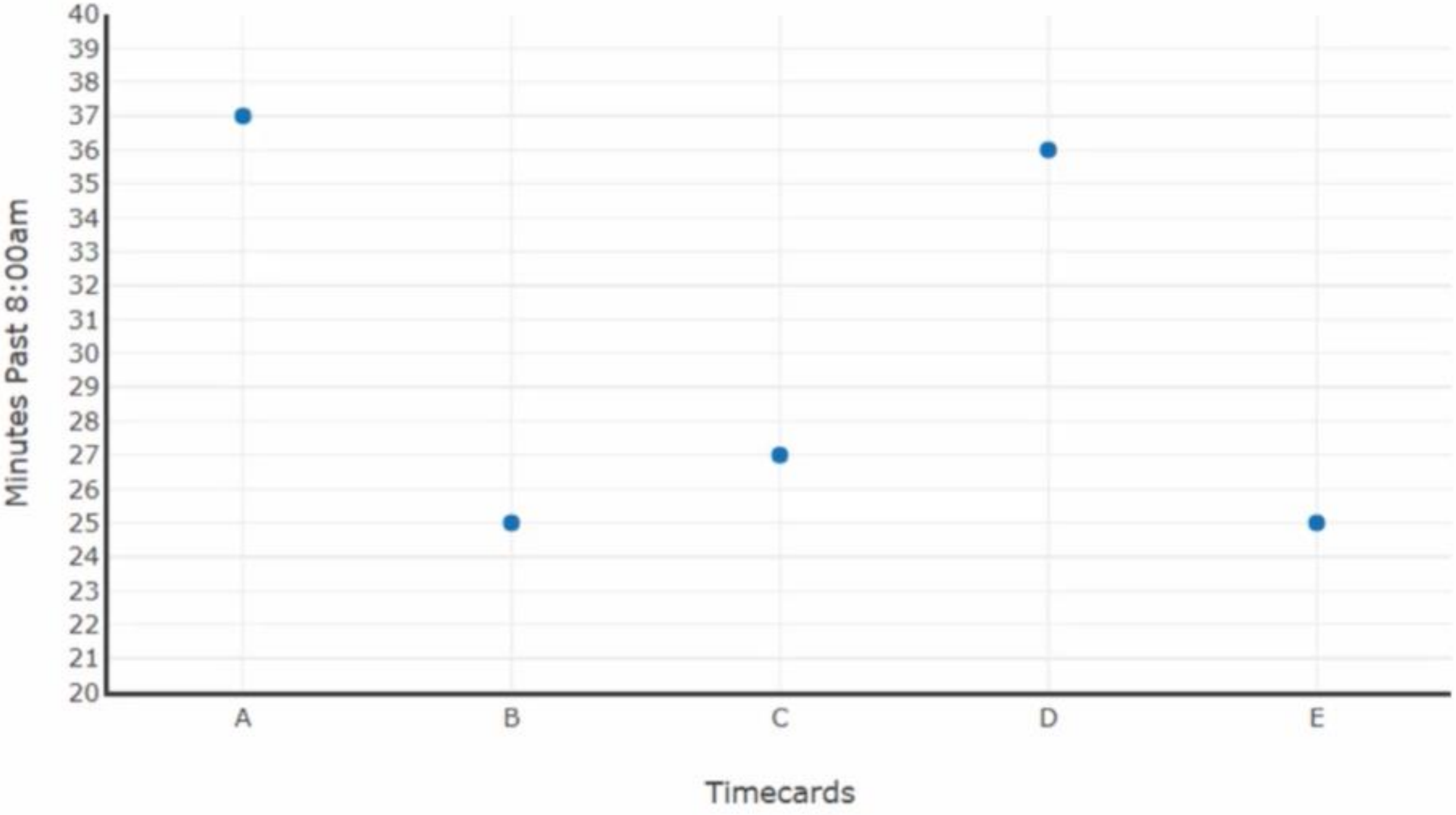
Consider a dataset with only one variable

The best-fit line is just the mean value of the data points



Linear Regression

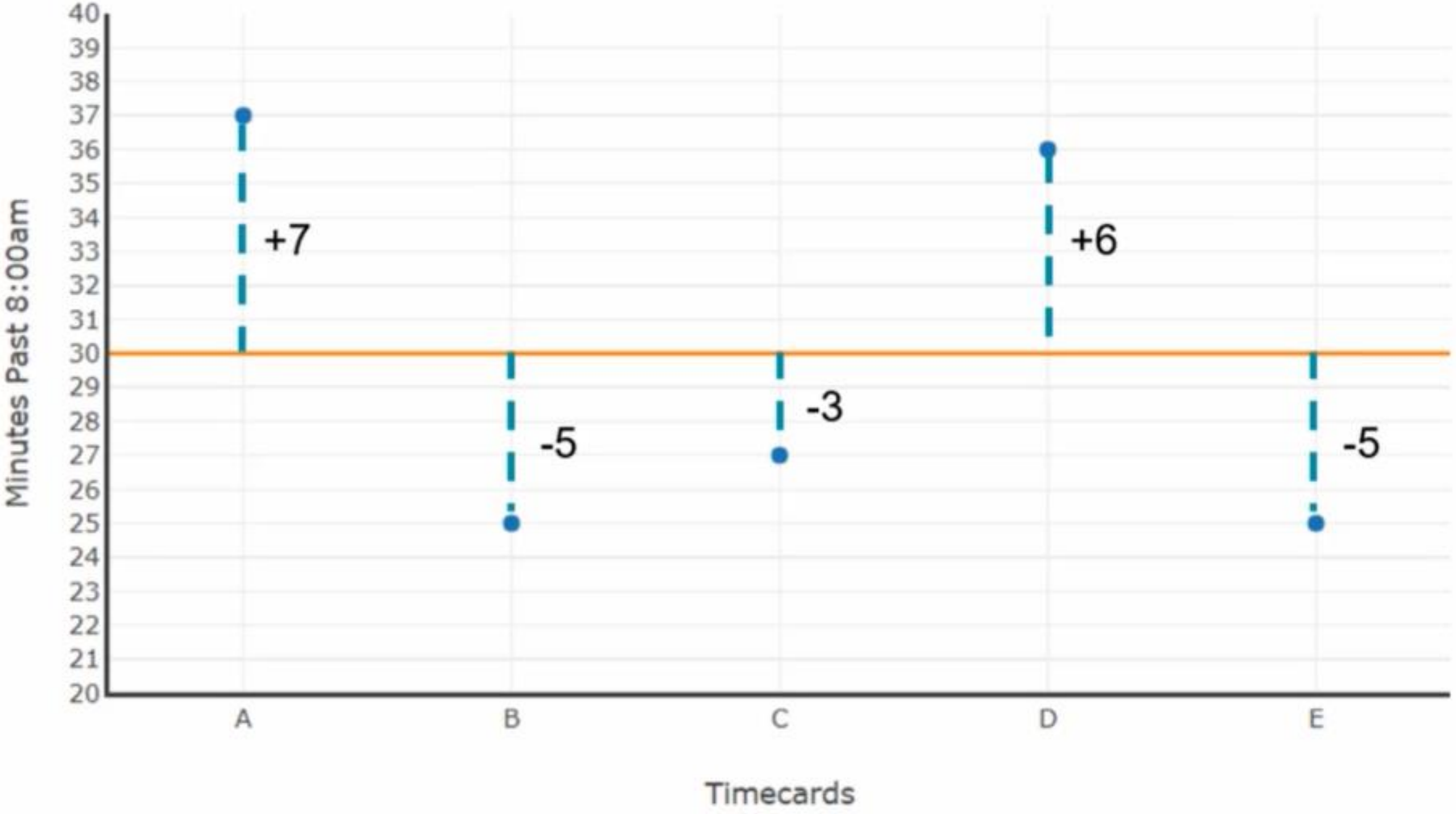
Timecard	Minutes past 8:00am
A	37
B	25
C	27
D	36
E	25
Total:	**150**
Mean	**30**



Linear Regression

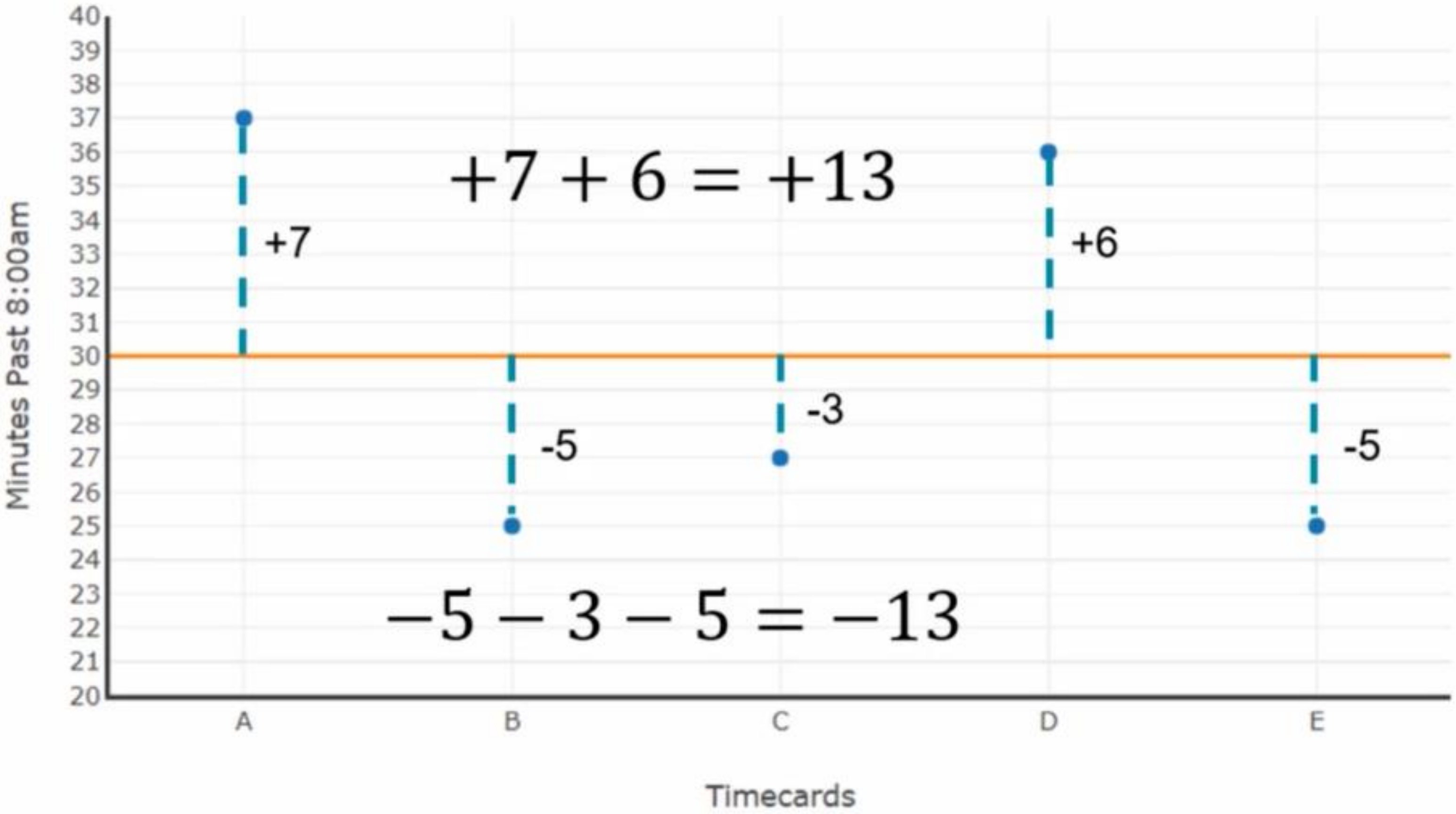
What makes $y = 30$ a best-fit line?

Consider the error



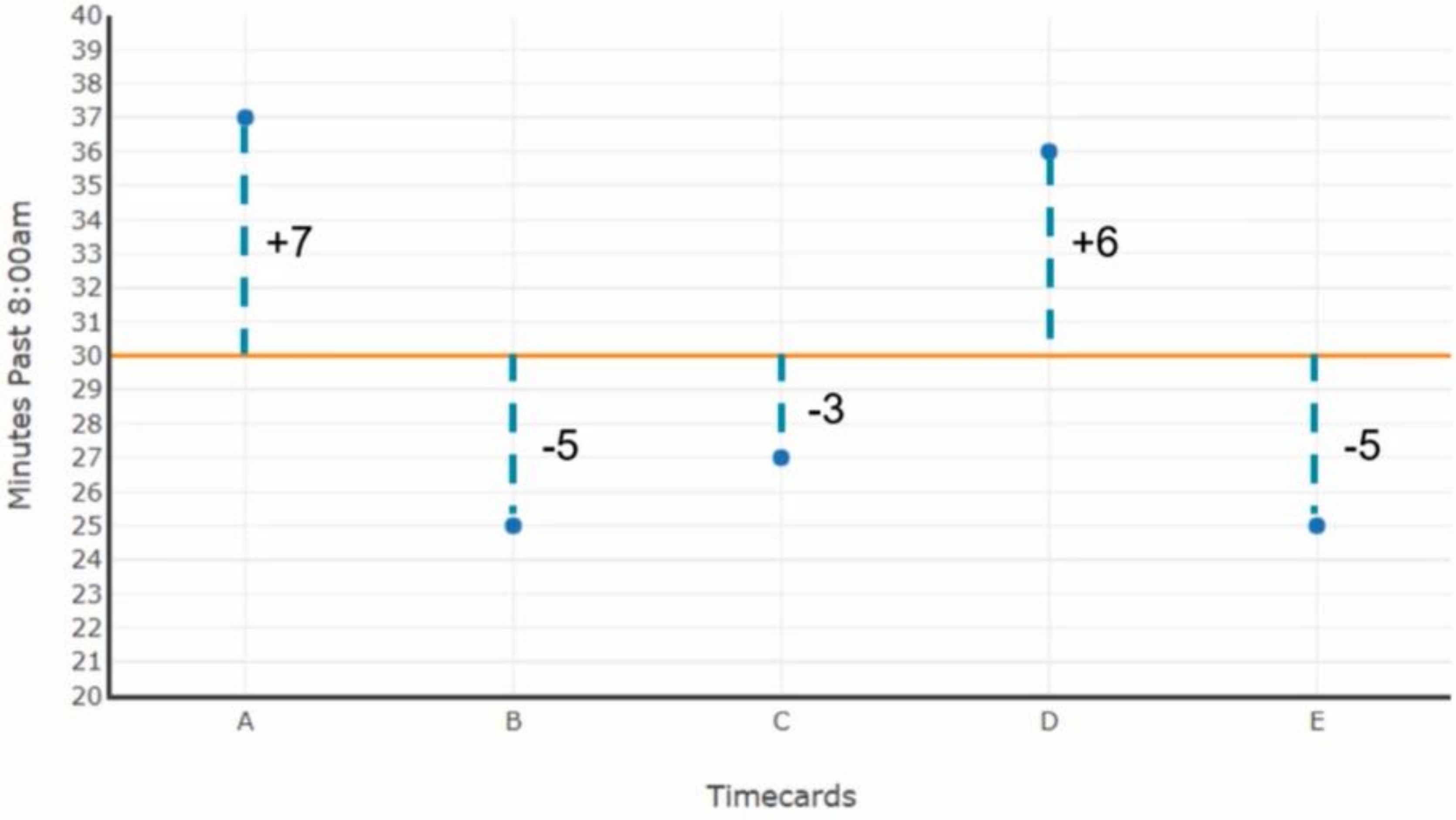
Linear Regression

See that the sum of the distances above the line balances the sum of those below the line



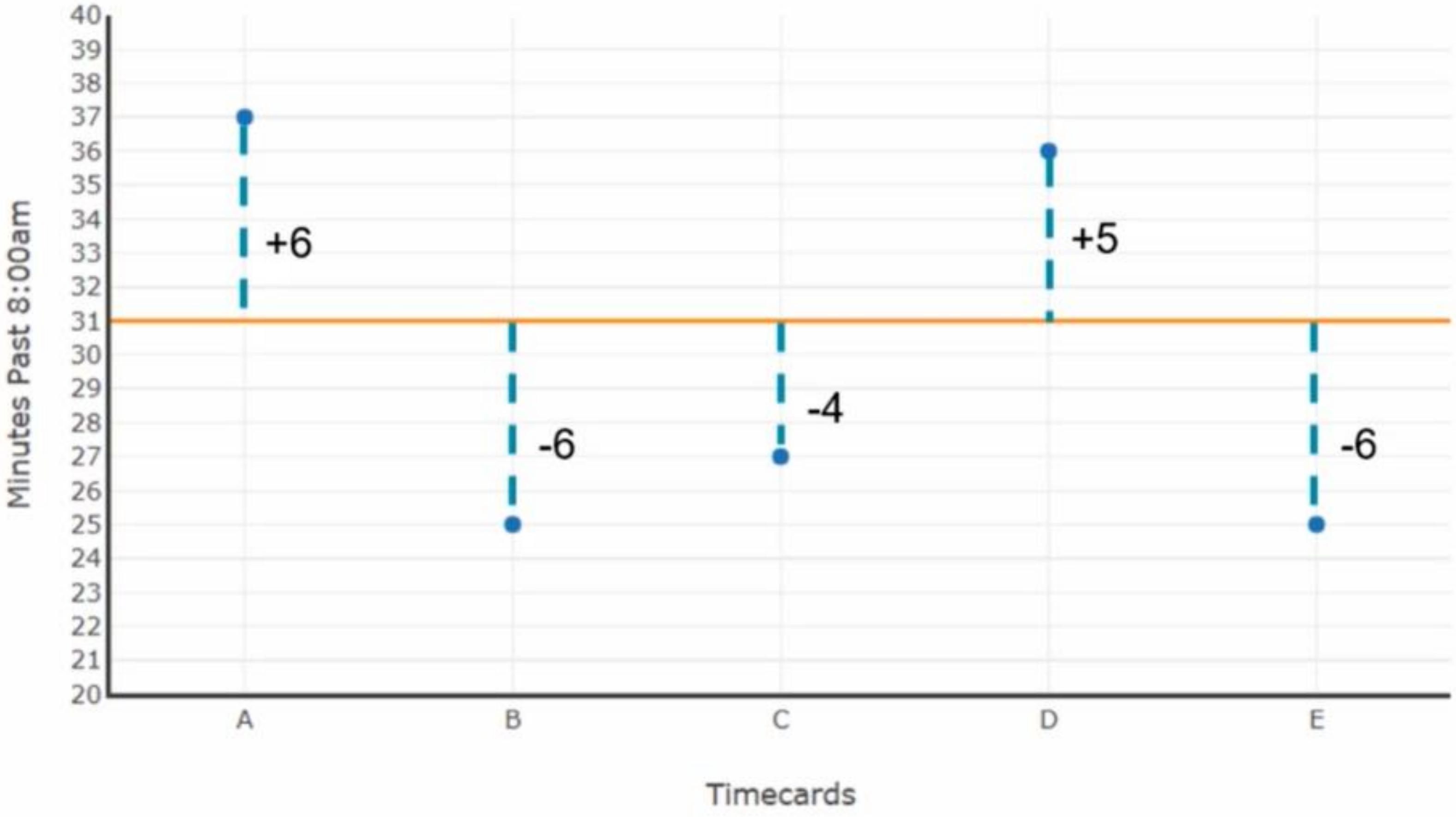
Linear Regression

Error (E)	Square Error (SE)
+7	49
-5	25
-3	9
+6	36
-5	25
Sum of Squares Error (SSE)	**144**



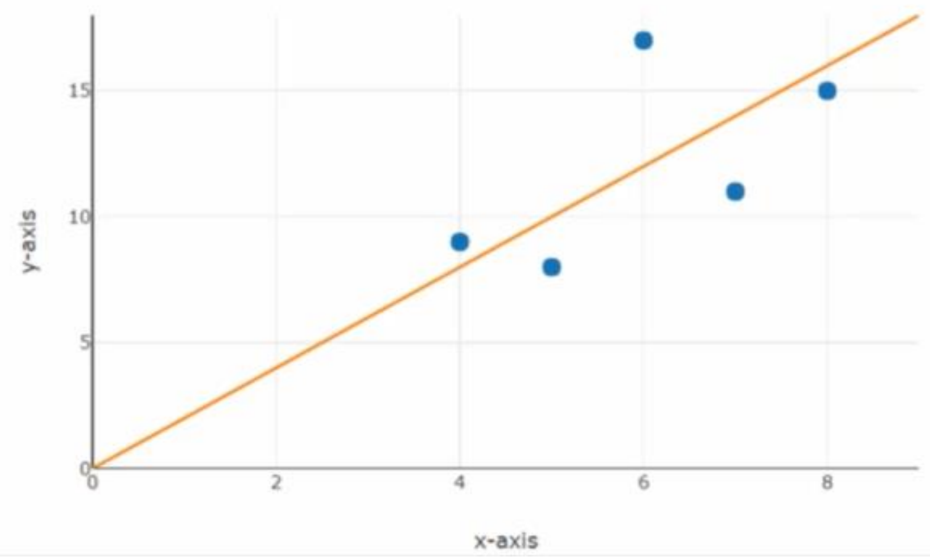
Linear Regression

Error (E)		Square Error (SE)	
+7	+6	49	36
-5	-6	25	36
-3	-4	9	16
+6	+5	36	25
-5	-6	25	36
Sum of Squares Error (SSE)		144	**149**



Linear Regression

That's it! The goal of regression is to find the line that best describes our data.

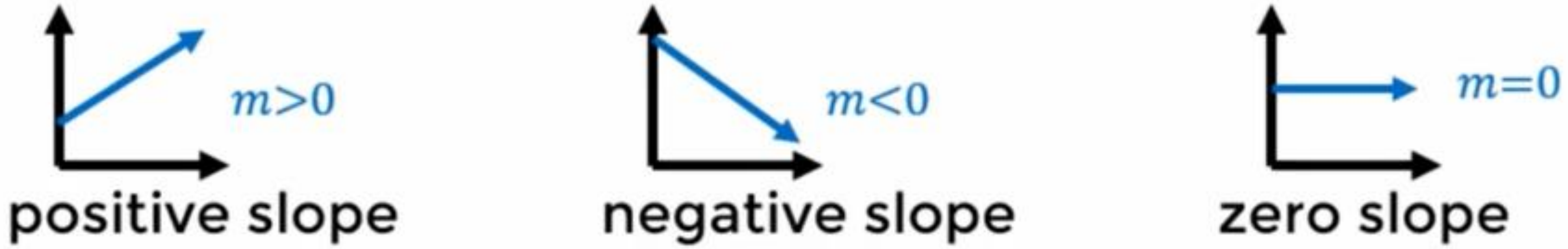


Linear Regression

Recall that the equation of a line follows the form $y = mx + b$ where

m is the **slope** of the line, and

b is where the line crosses the y-axis when x=0 (b is the **y-intercept**)



Linear Regression Example

A manager wants to find the relationship between the number of hours that a plant is operational in a week and weekly production.

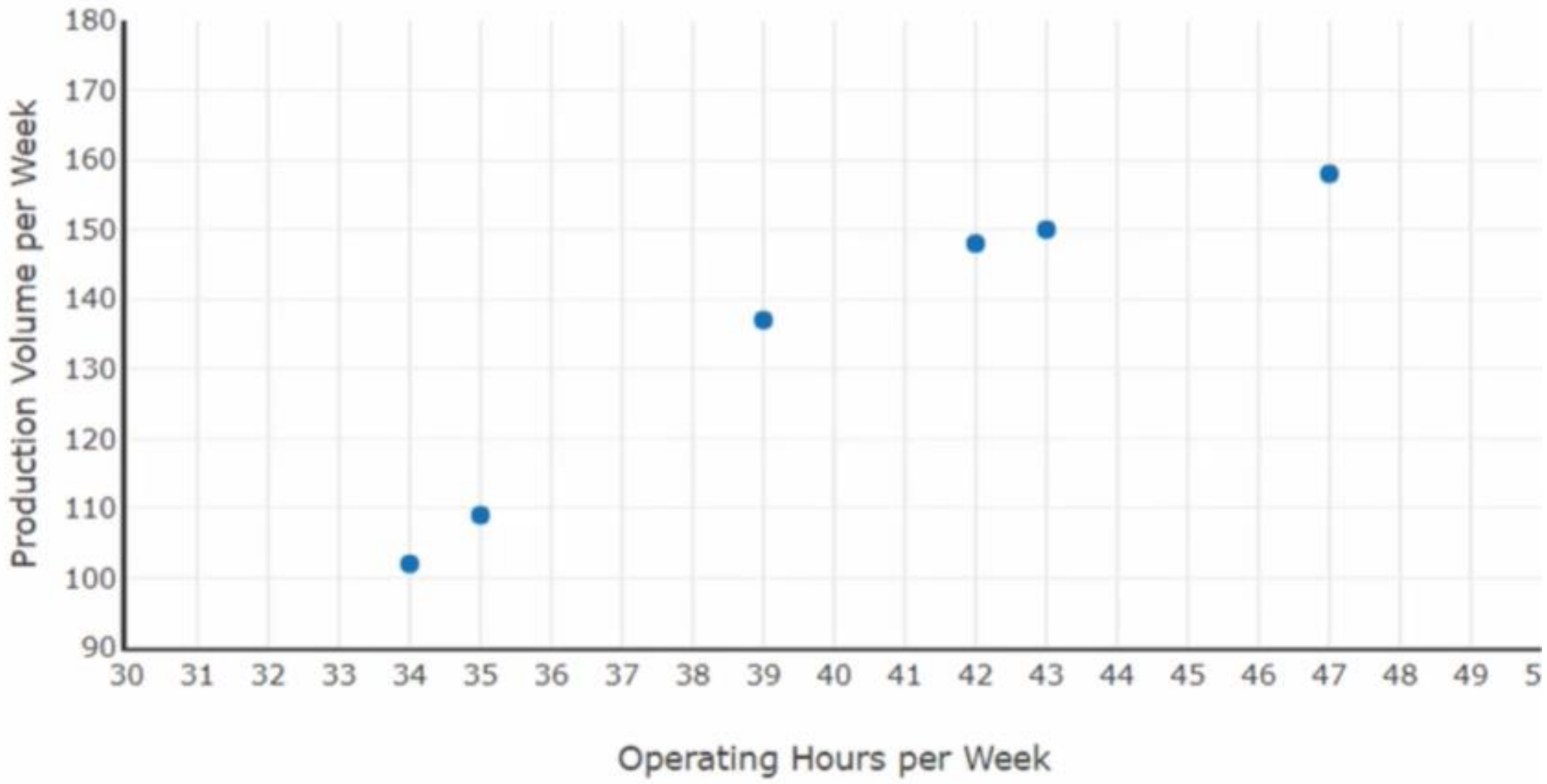
Linear Regression Example

Here the **independent variable** x is hours of operation, and the **dependent variable** y is production volume.

Linear Regression Example

First, plot the data

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158



Linear Regression Example

$$\hat{y} = b_0 + b_1 x$$
$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

	Production Hours (x)	Production Volume (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	34	102	-6	-32	192	36
	35	109	-5	-25	125	25
	39	137	-1	3	-3	1
	42	148	2	14	28	4
	43	150	3	16	48	9
	47	158	7	24	168	49
$\bar{x}, \bar{y}$	**40**	**134**		Sum:	**558**	**124**

$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

Linear Regression Example

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

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158
40	**134**

$\bar{x}, \bar{y}$: 40, 134

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{558}{124} = \mathbf{4.5}$$

$$b_0 = \bar{y} - b_1\bar{x} = 134 - (4.5 \times 40) = \mathbf{-46}$$

$$\boldsymbol{\hat{y} = -46 + 4.5x}$$

Sum:	558	124
	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

Linear Regression Example

Based on the formula, if the manager wants to produce 125 units per week, the plant should run for:

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158

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

$$125 = -46 + 4.5x$$

$$x = \frac{171}{4.5} = \mathbf{38}\ hours\ per\ week$$