DATA MINING 1
Introduction Regression

Dino Pedreschi, Riccardo Guidotti

a.a. 2024/2025

Contains edited slides from StatQuest

UNIVERSITÀ DI PISA

Linear Regression

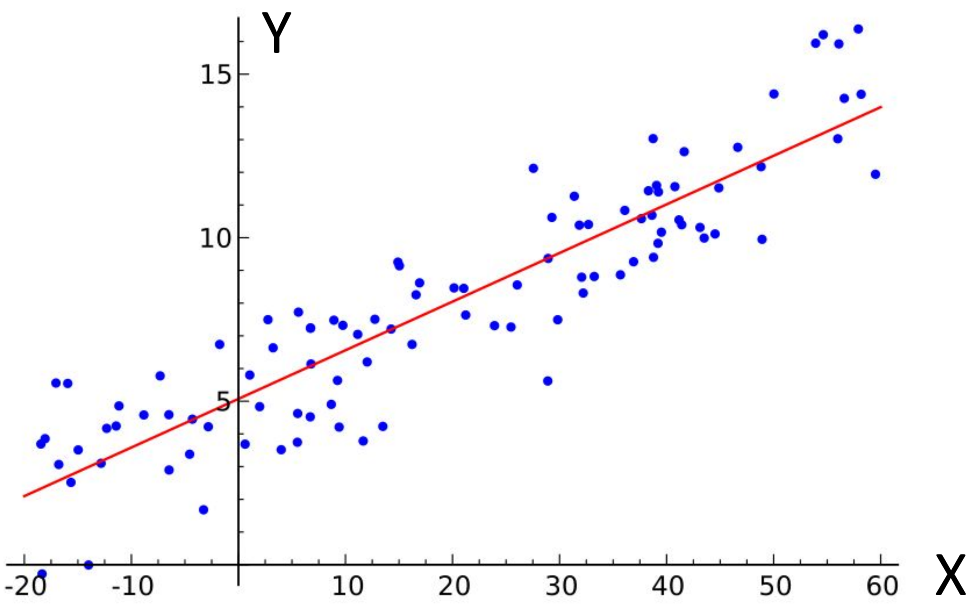
Regression

- Given a dataset containing N observations X_i, Y_i $i = 1, 2, \ldots, N$
- **Regression** is the task of learning a target function *f* that maps each input attribute set *X* into an output *Y*.
- The goal is to find the target function that can fit the input data with minimum error.
- The error function can be expressed as
  - Absolute Error = $\sum_i |y_i - f(x_i)|$
  - Squared Error = $\sum_i (y_i - f(x_i))^2$

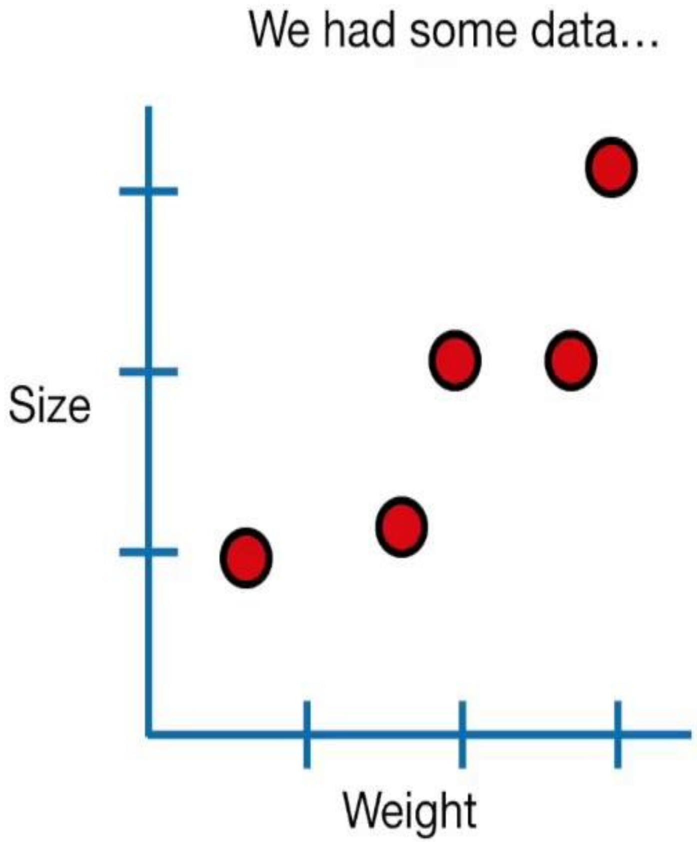
residuals

Linear Regression

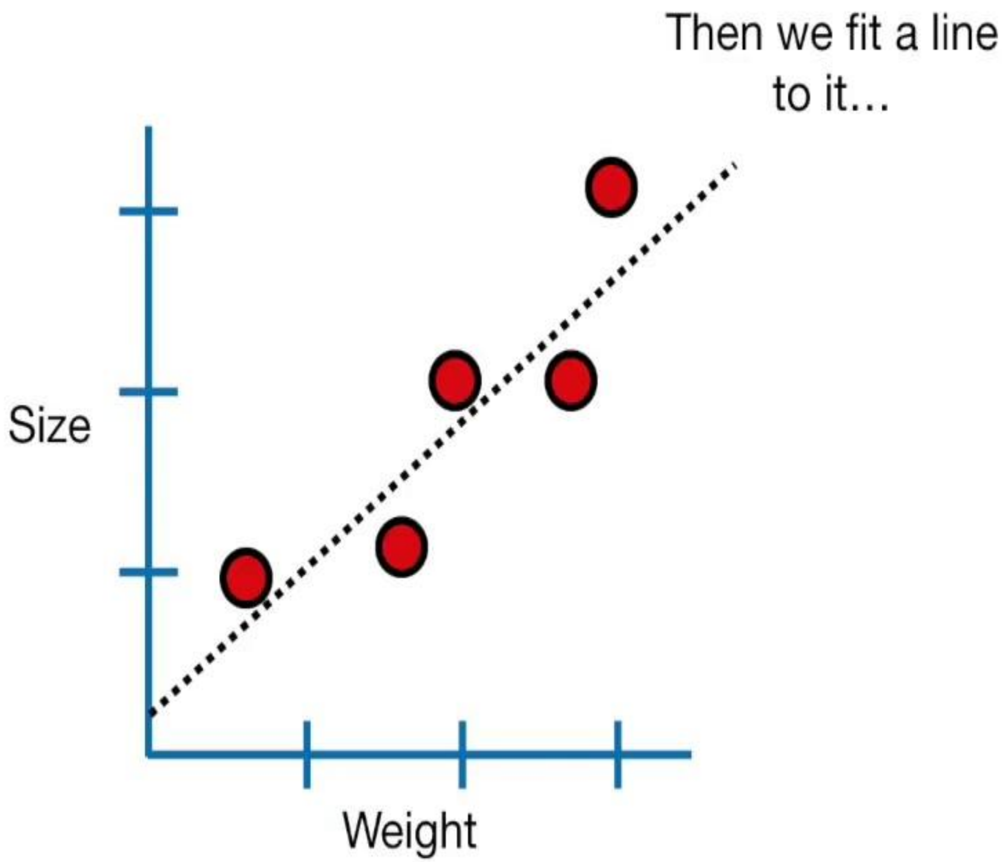
- **Linear regression** is a linear approach to modeling the relationship between a *dependent variable Y* and one or more *independent* (explanatory) variables *X*.
- The case of *one* explanatory variable is called **simple linear regression**.
- For *more than one explanatory variable*, the process is called **multiple linear regression**.
- For *multiple correlated dependent variables*, the process is called **multivariate linear regression**.



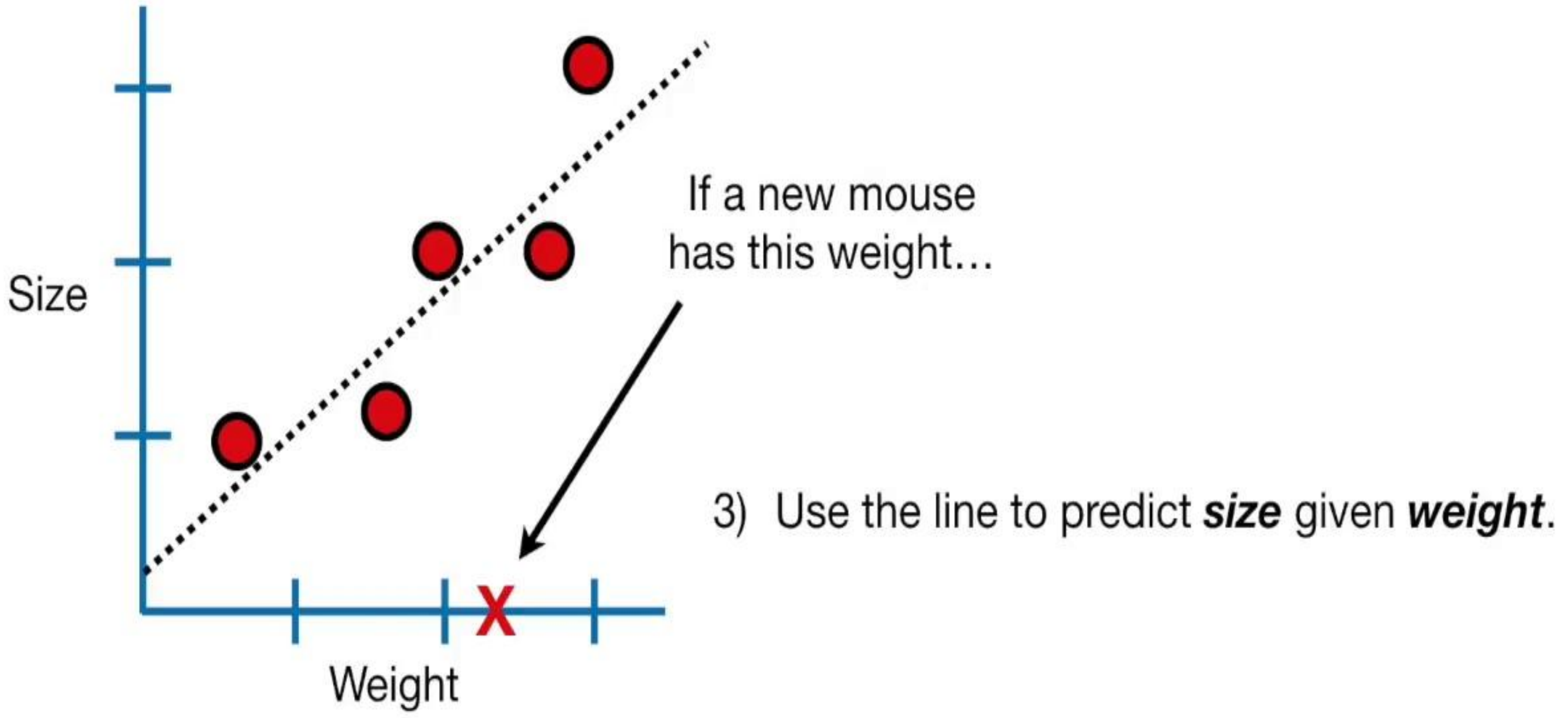
What does it mean to predict Y?



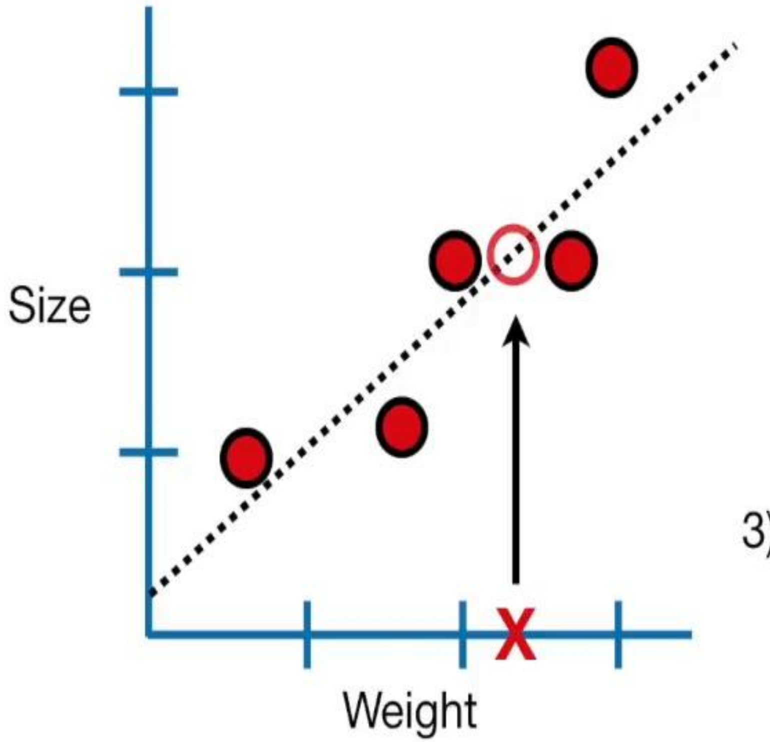
What does it mean to predict Y?



What does it mean to predict Y?

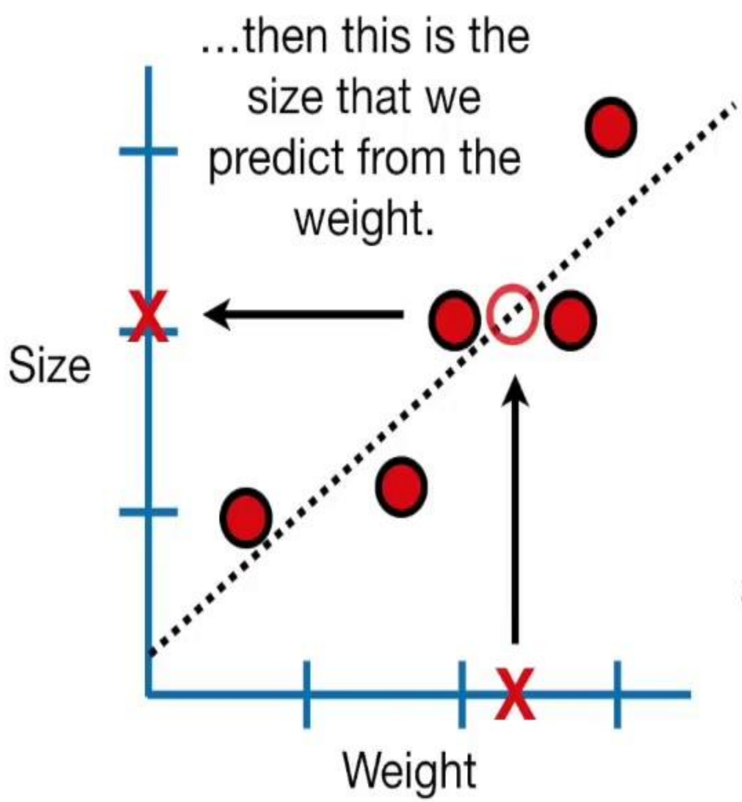


What does it mean to predict Y?



3) Use the line to predict ***size*** given ***weight***.

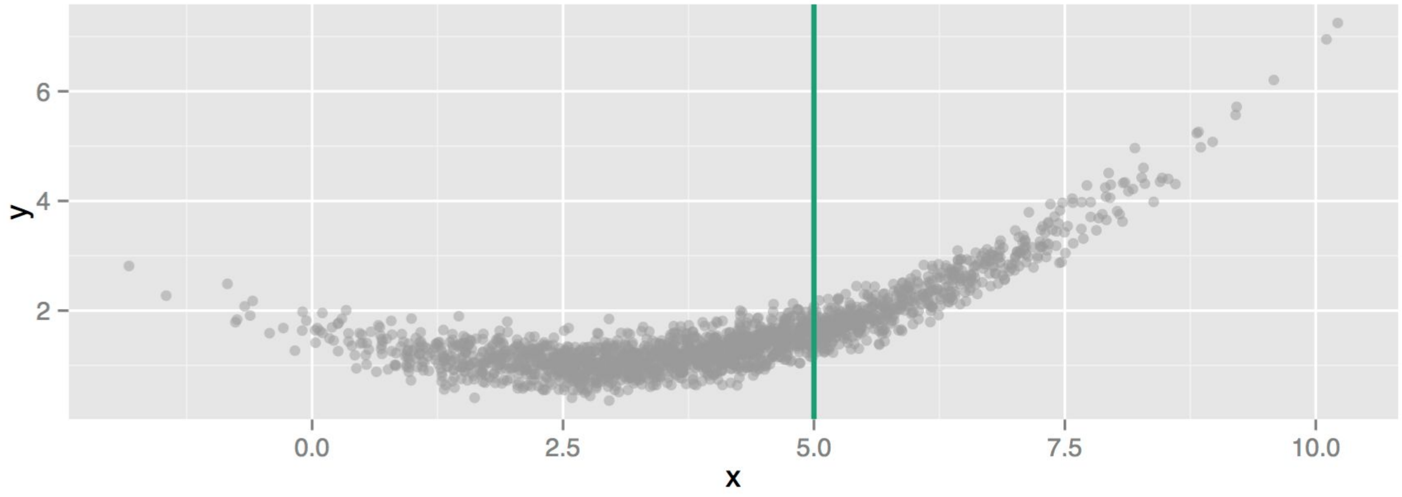
What does it mean to predict Y?



3) Use the line to predict ***size*** given ***weight***.

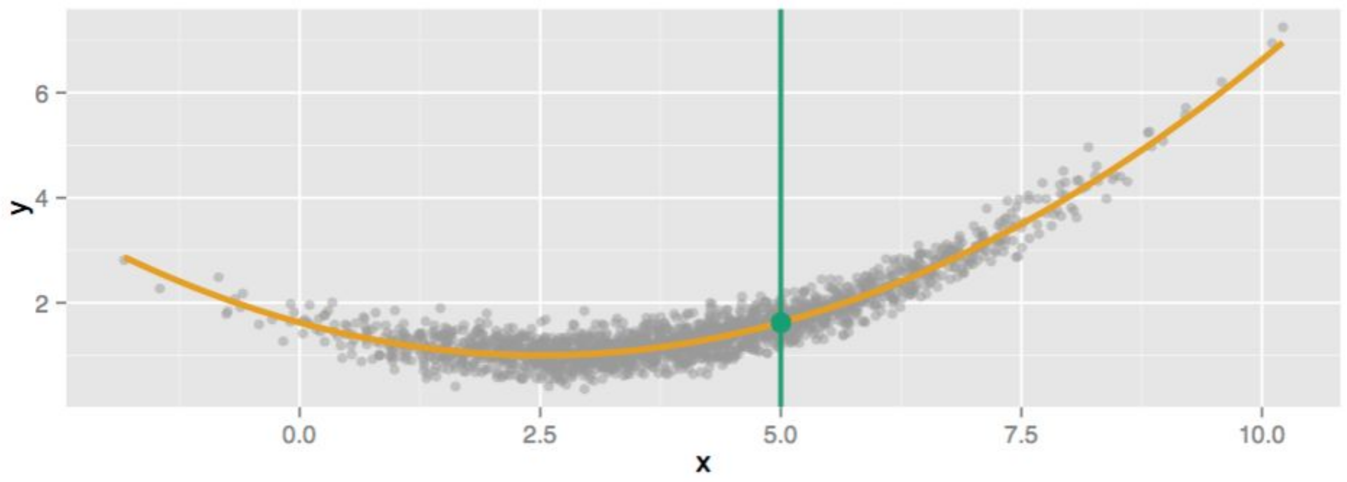
What does it mean to predict Y?

- Look at *X = 5*. There are many different *Y* values at *X=5*.
- When we say predict *Y* at *X =5*, we are really asking:
- What is the expected value (average) of *Y* at *X = 5*?

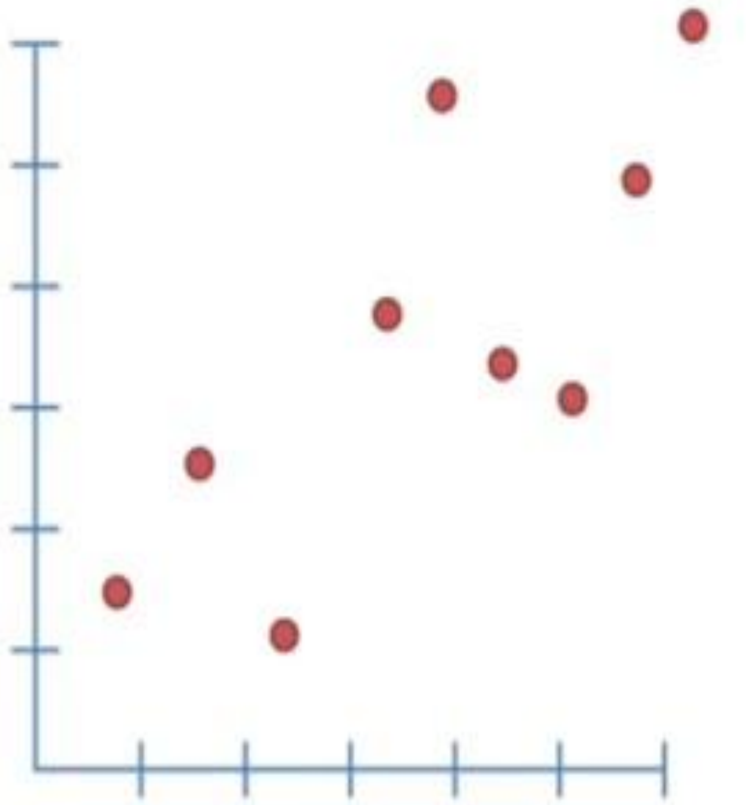


What does it mean to predict Y?

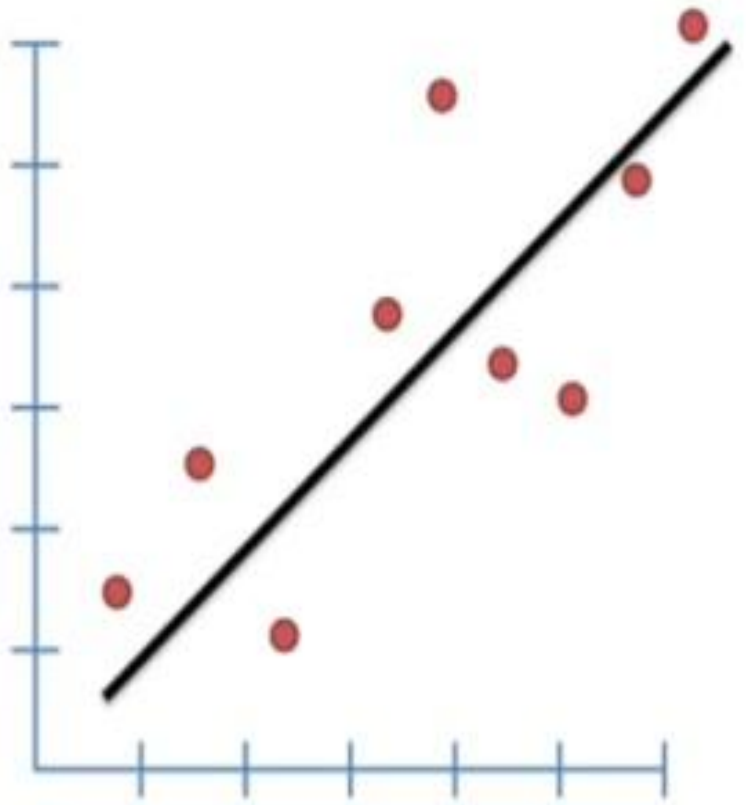
- Formally, the ***regression function*** is given by *E(Y|X=x)*. This is the expected value of Y at X=x.
- The ideal or optimal predictor of *Y* based on *X* is thus
  - f(X) = E(Y | X=x)



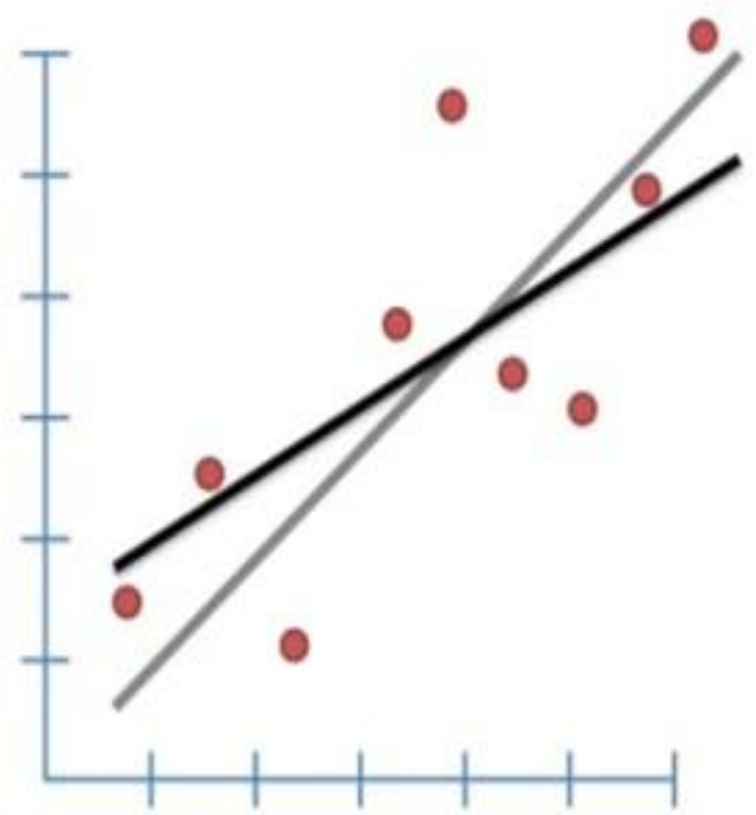
Fitting a Line Example



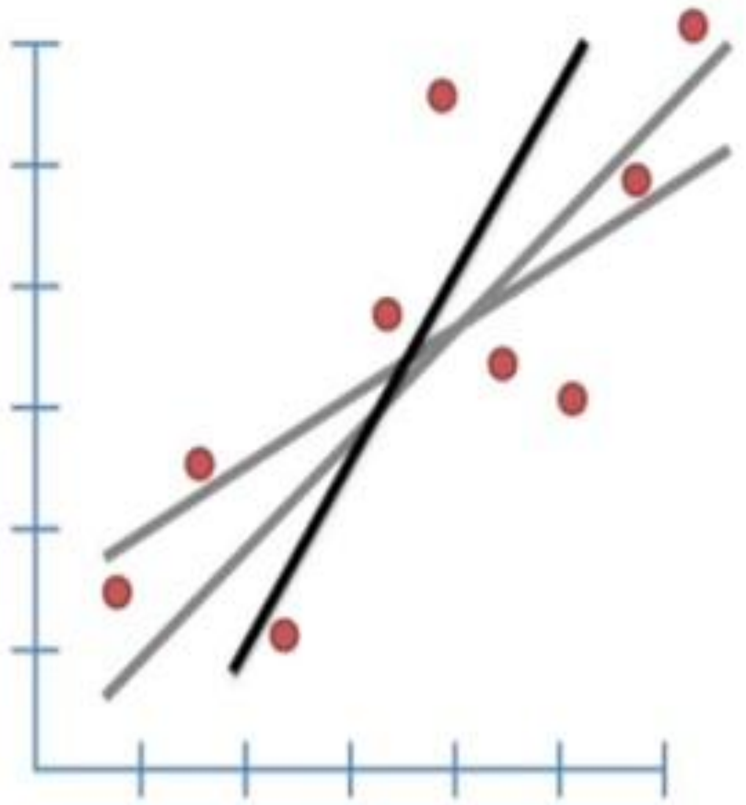
Fitting a Line Example



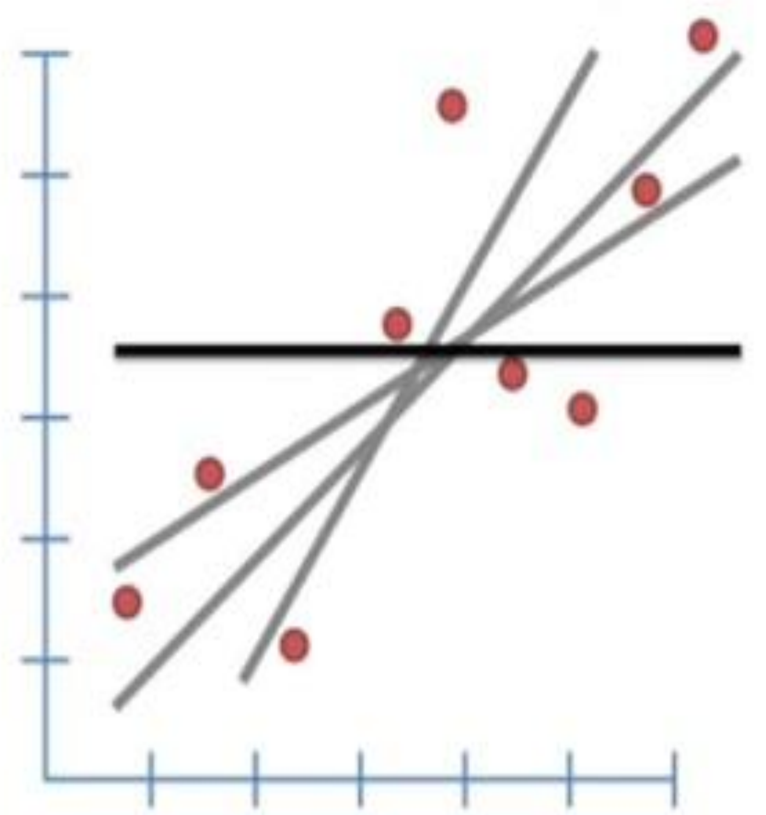
Fitting a Line Example



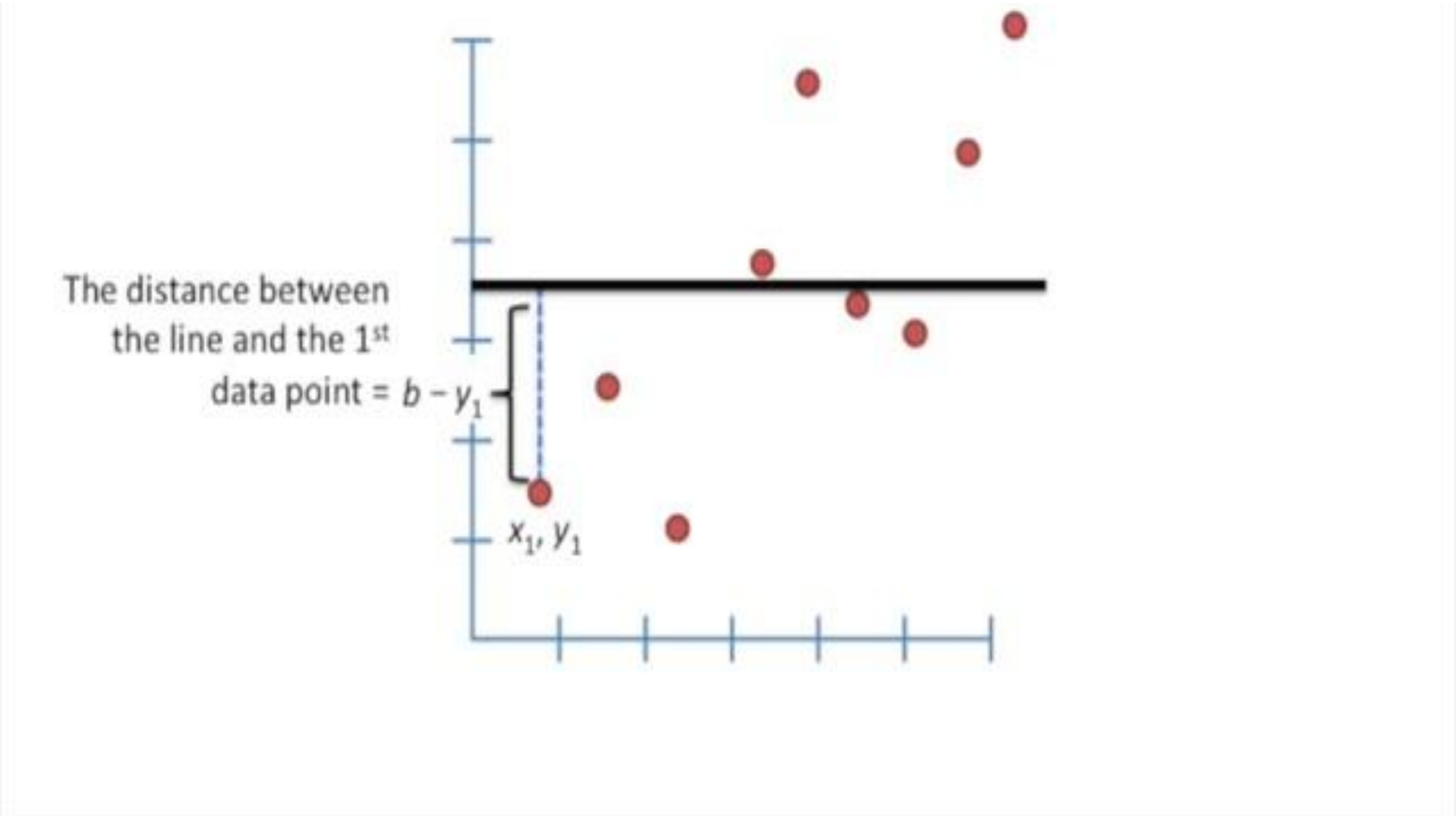
Fitting a Line Example



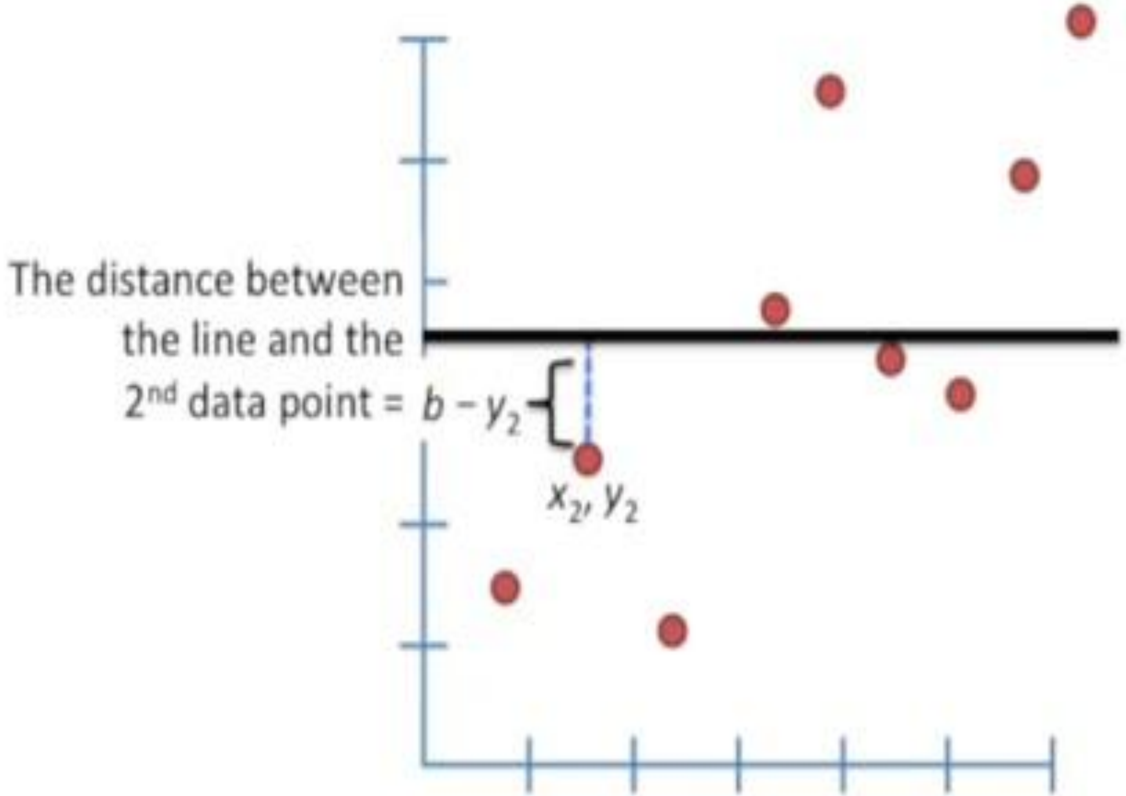
Fitting a Line Example

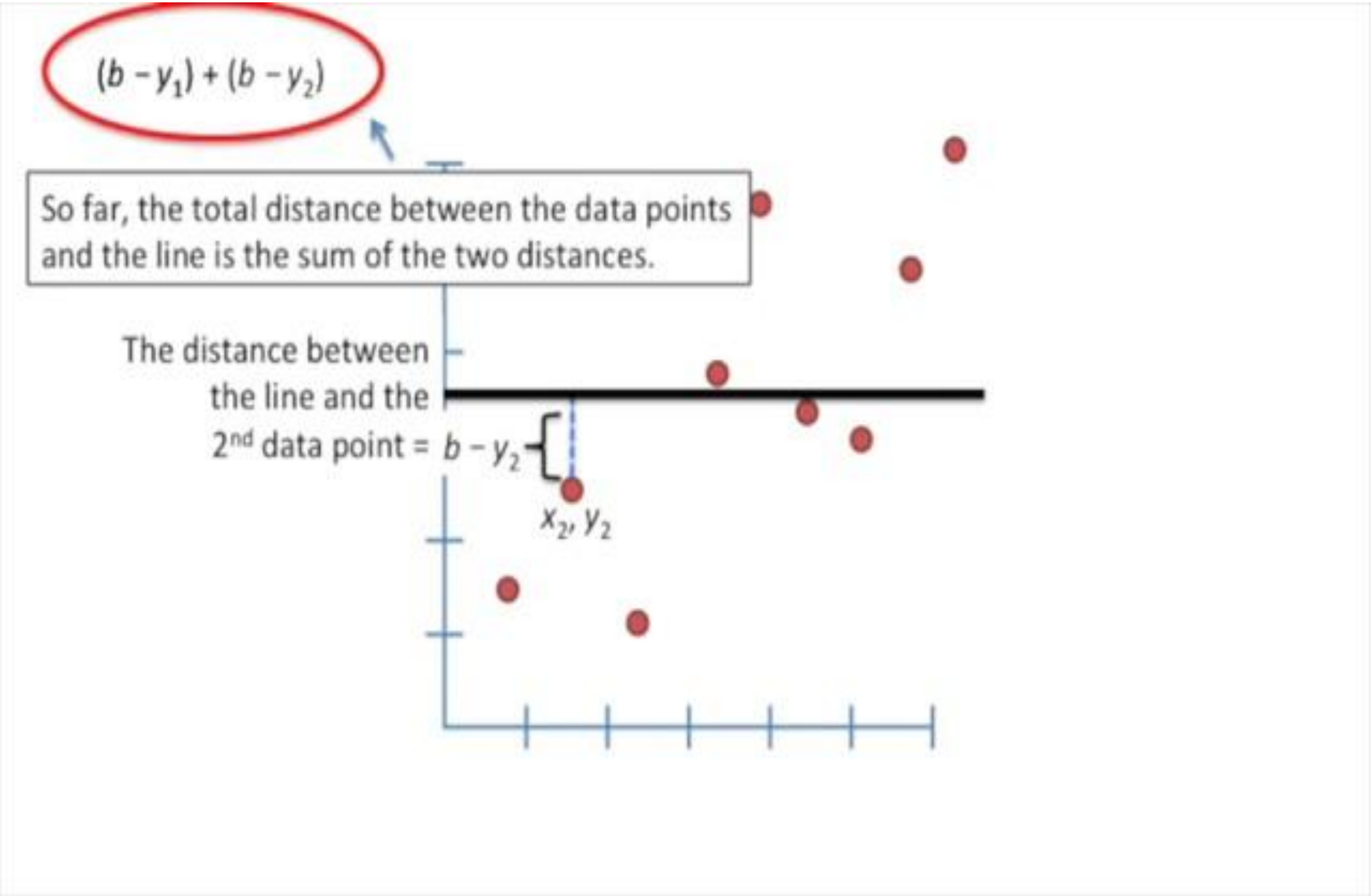


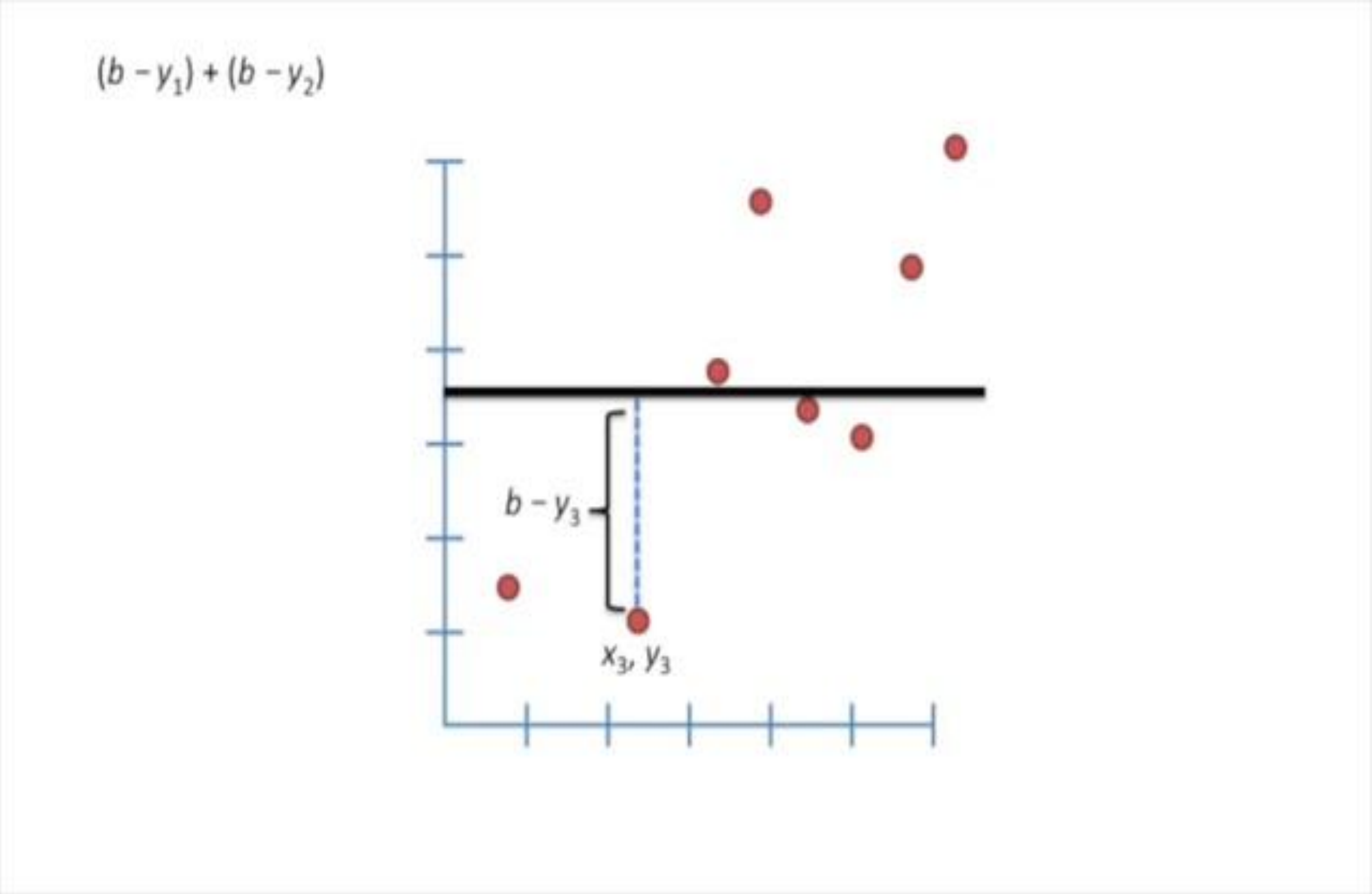
Measuring Line Error



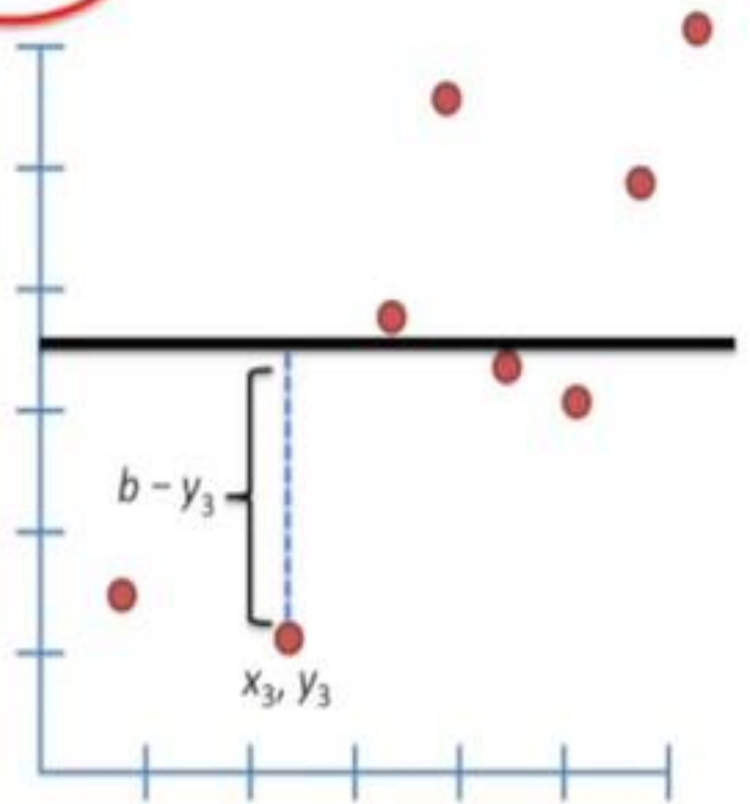
Measuring Line Error



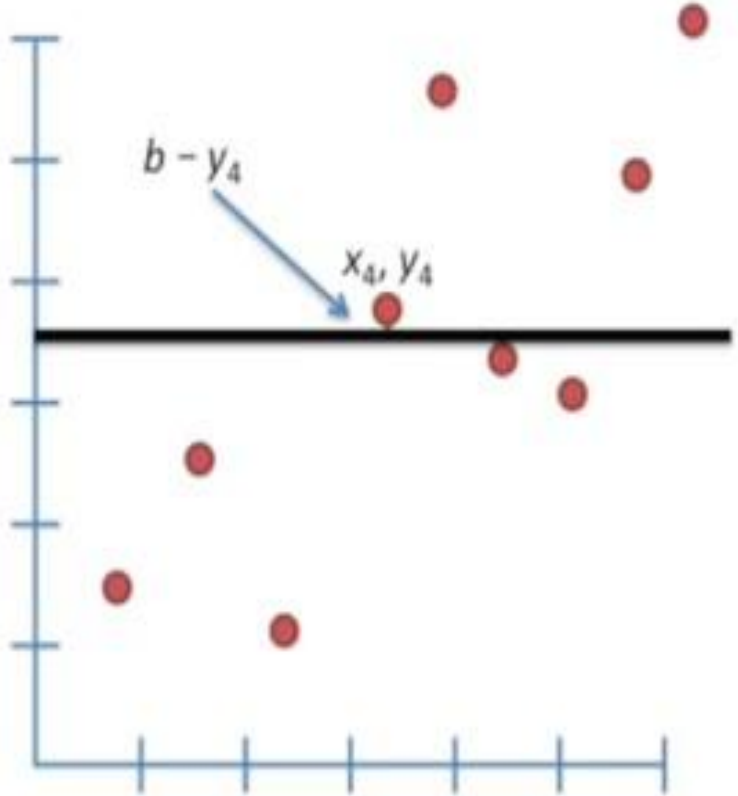

$(b - y_1) + (b - y_2)$
So far, the total distance between the data points and the line is the sum of the two distances.
The distance between the line and the 2nd data point = $b - y_2$
x_2, y_2


$(b - y_1) + (b - y_2)$
$b - y_3$
x_3, y_3

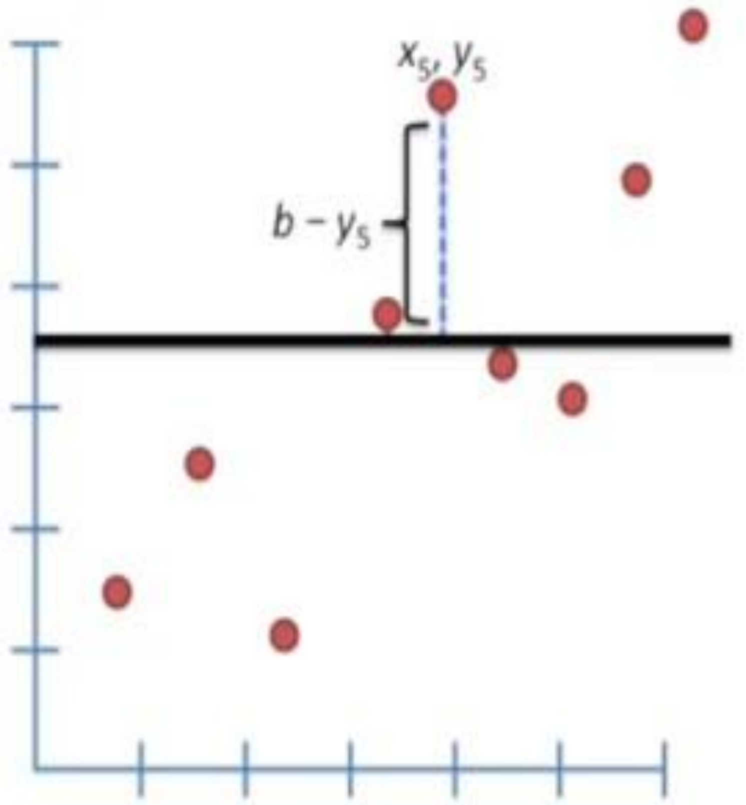
$(b - y_1) + (b - y_2) + (b - y_3)$ Now we've add the 3rd distance to our total sum.



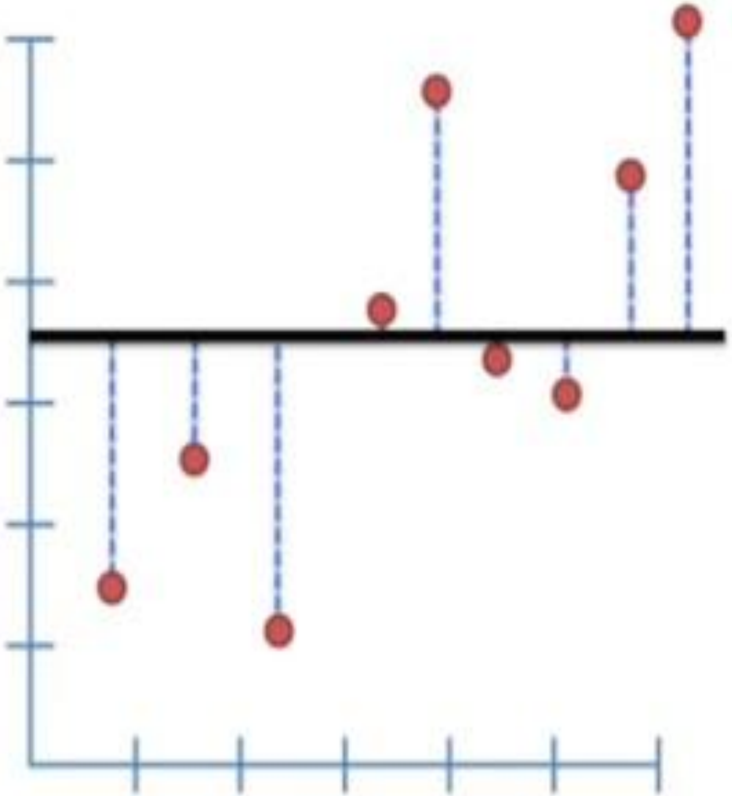
$$(b - y_1) + (b - y_2) + (b - y_3)$$



$(b - y_1) + (b - y_2) + (b - y_3) + (b - y_4)$

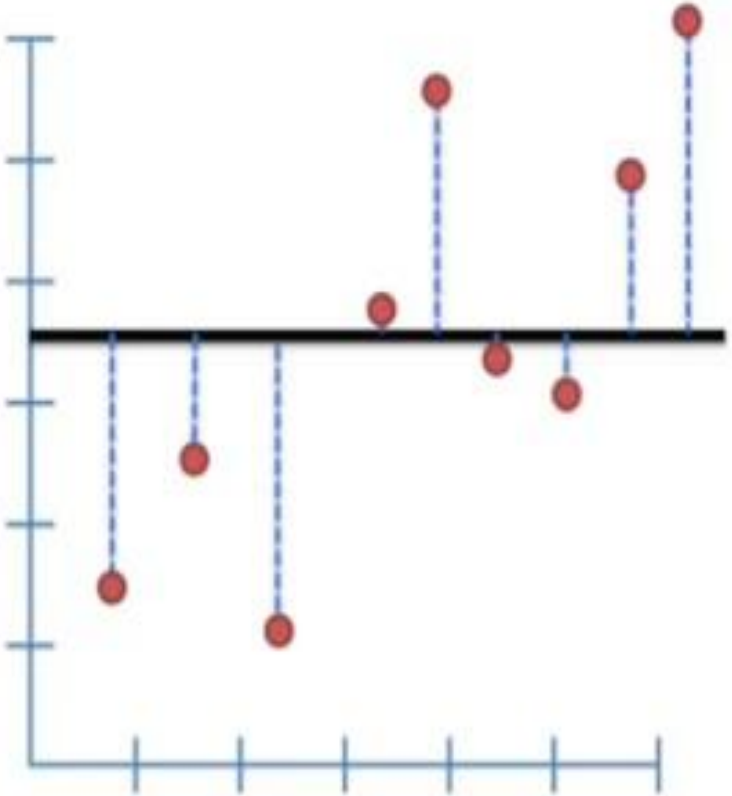


$$(b-y_1)^2+(b-y_2)^2+(b-y_3)^2+(b-y_4)^2+(b-y_5)^2+(b-y_6)^2+(b-y_7)^2+(b-y_8)^2+(b-y_9)^2$$



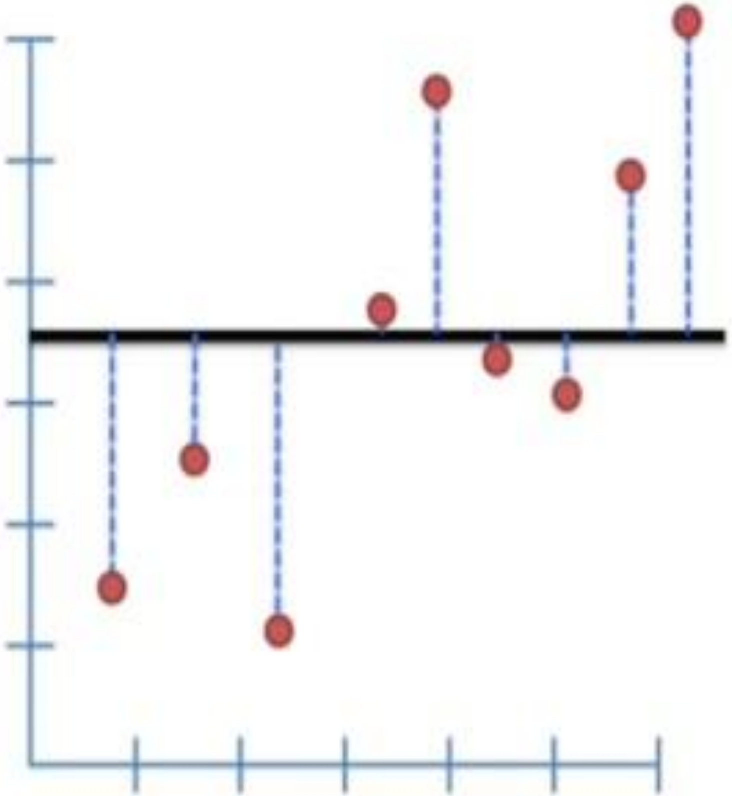
$$(b-y_1)^2+(b-y_2)^2+(b-y_3)^2+(b-y_4)^2+(b-y_5)^2+(b-y_6)^2+(b-y_7)^2+(b-y_8)^2+(b-y_9)^2$$

$$= 24.62$$



This is our measure of how well this line fits the data.

$$(b - y_1)^2 + (b - y_2)^2 + (b - y_3)^2 + (b - y_4)^2 + (b - y_5)^2 + (b - y_6)^2 + (b - y_7)^2 + (b - y_8)^2 + (b - y_9)^2$$

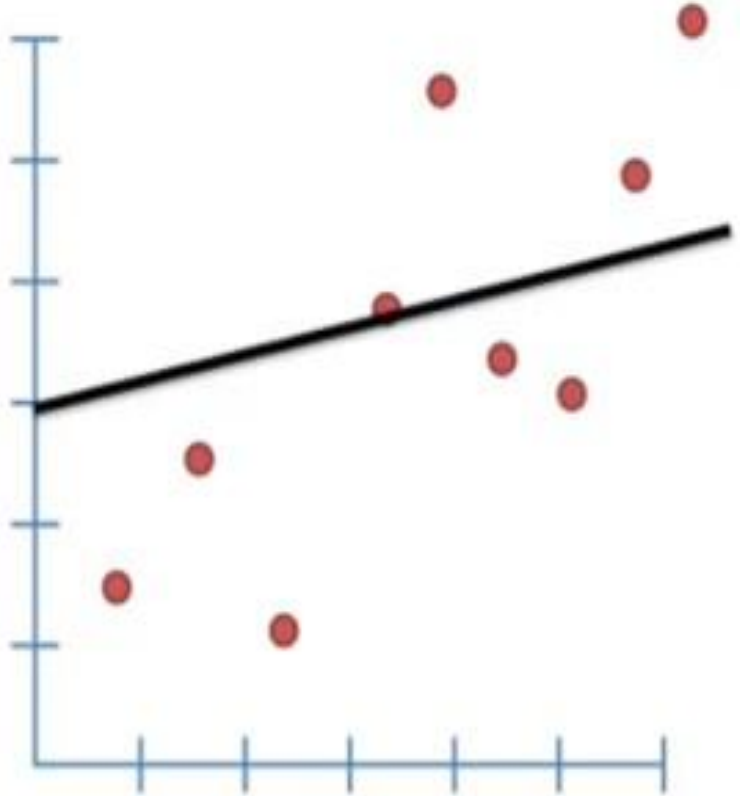


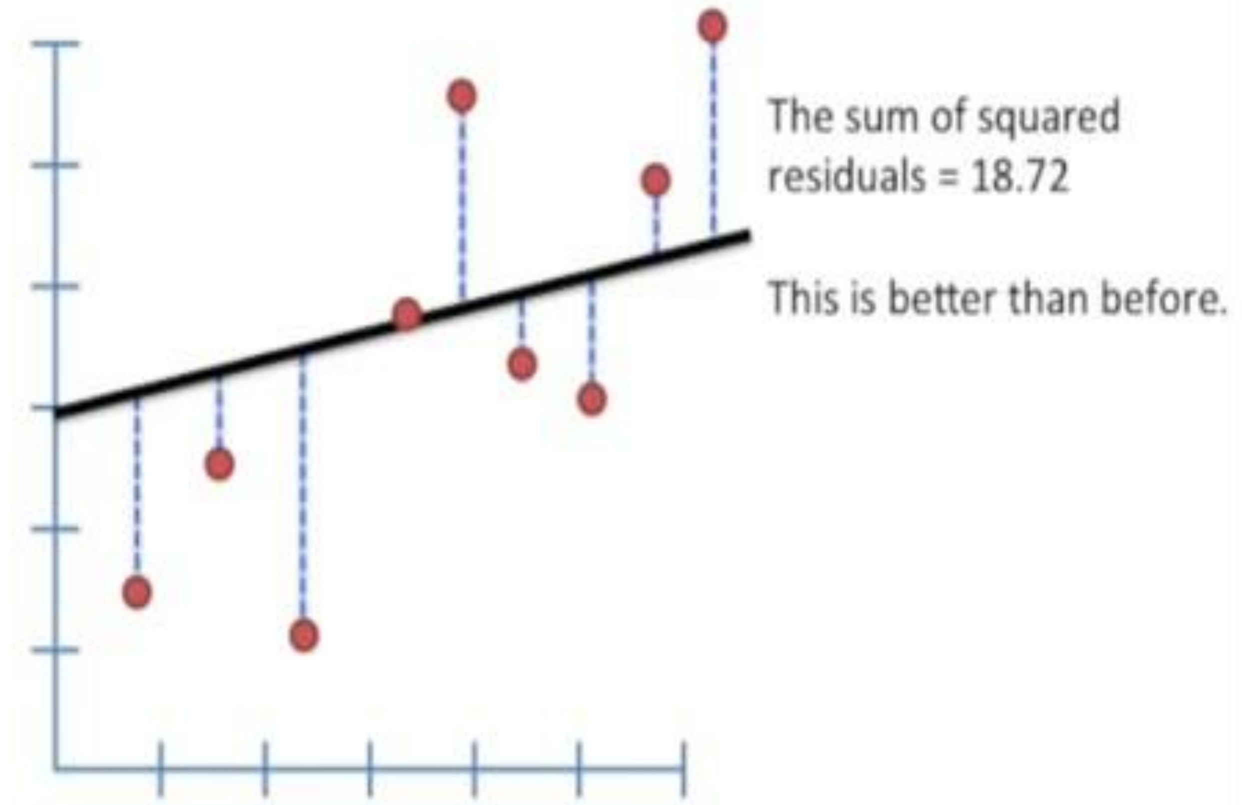
$= 24.62$

This is our measure of how well this line fits the data.

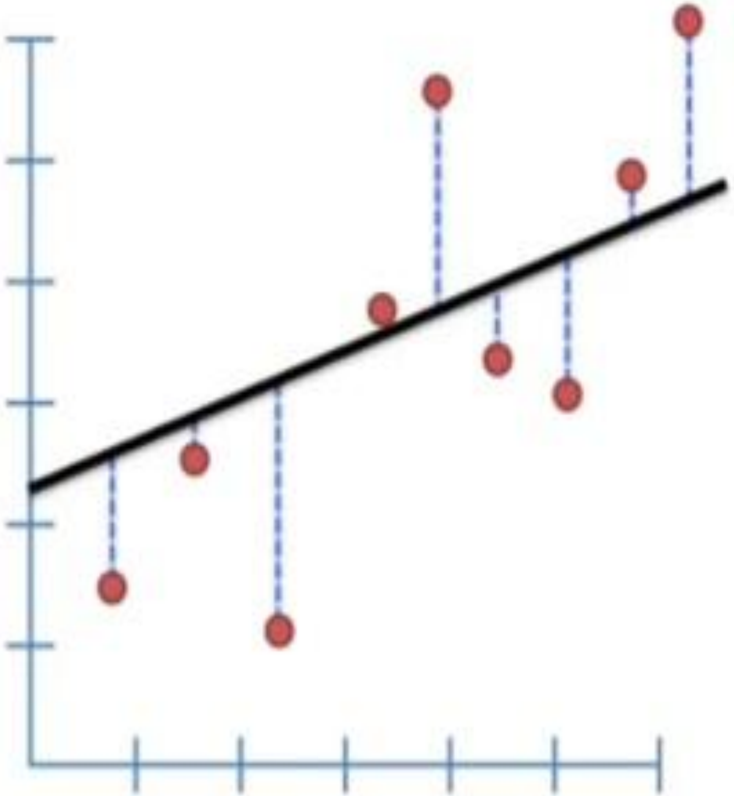
It's called the "sum of squared residuals, because the residuals are the differences between the real data and the line, and we are summing the square of these values.

Now let's see how good the fit is if we rotate the line a little bit.

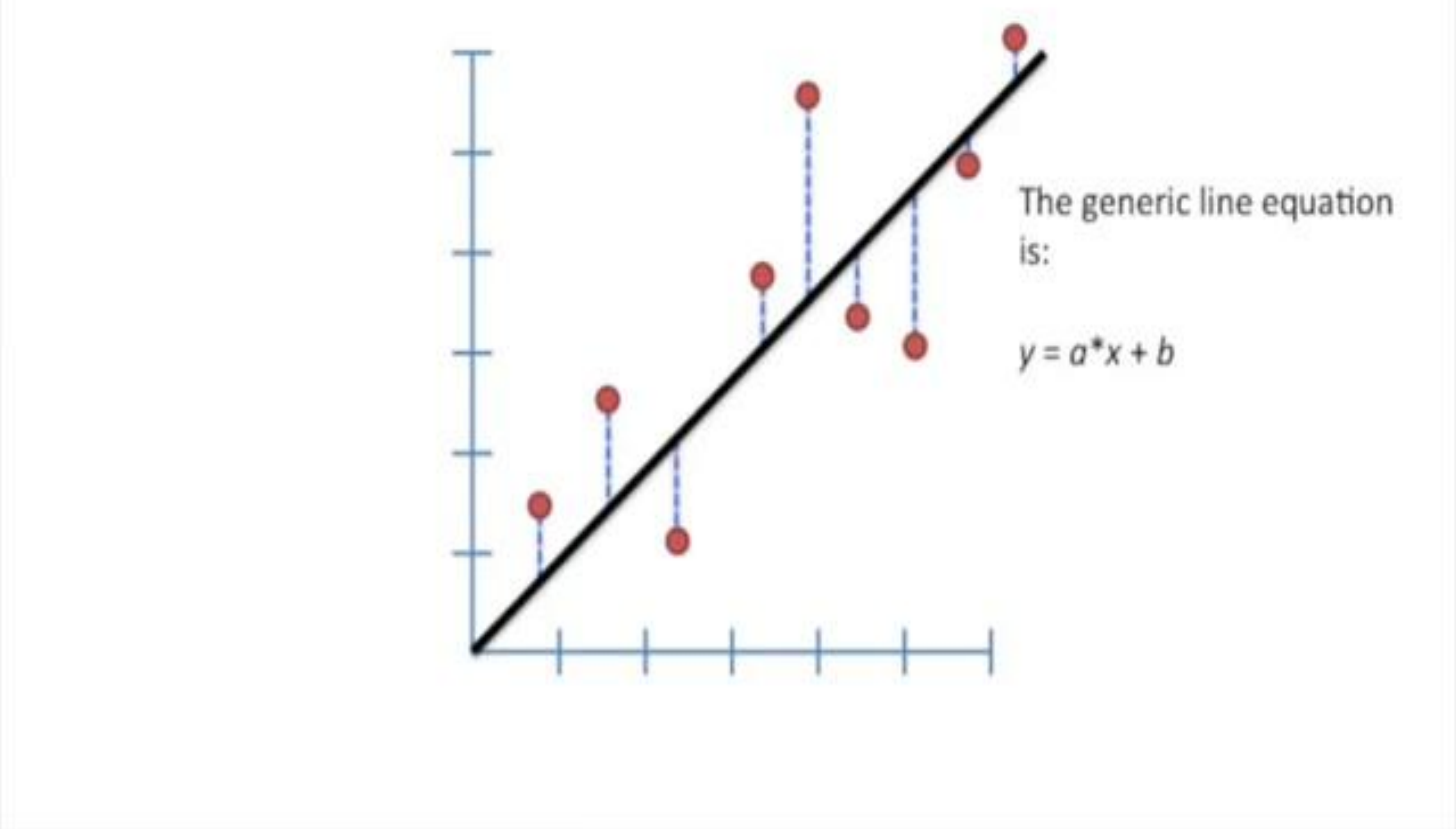



The sum of squared
residuals = 18.72
This is better than before.

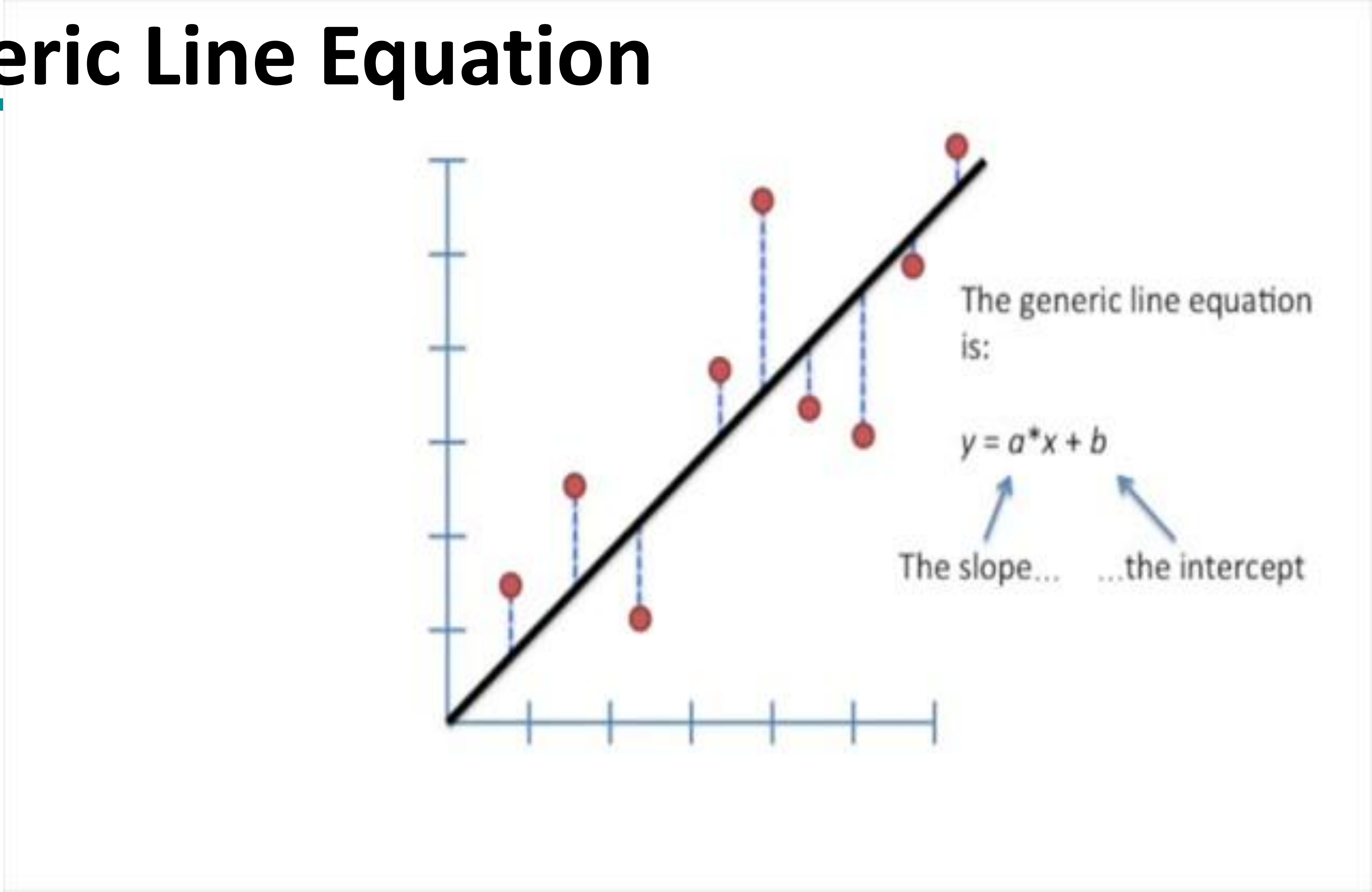
Does this fit improve if we rotate a little more?



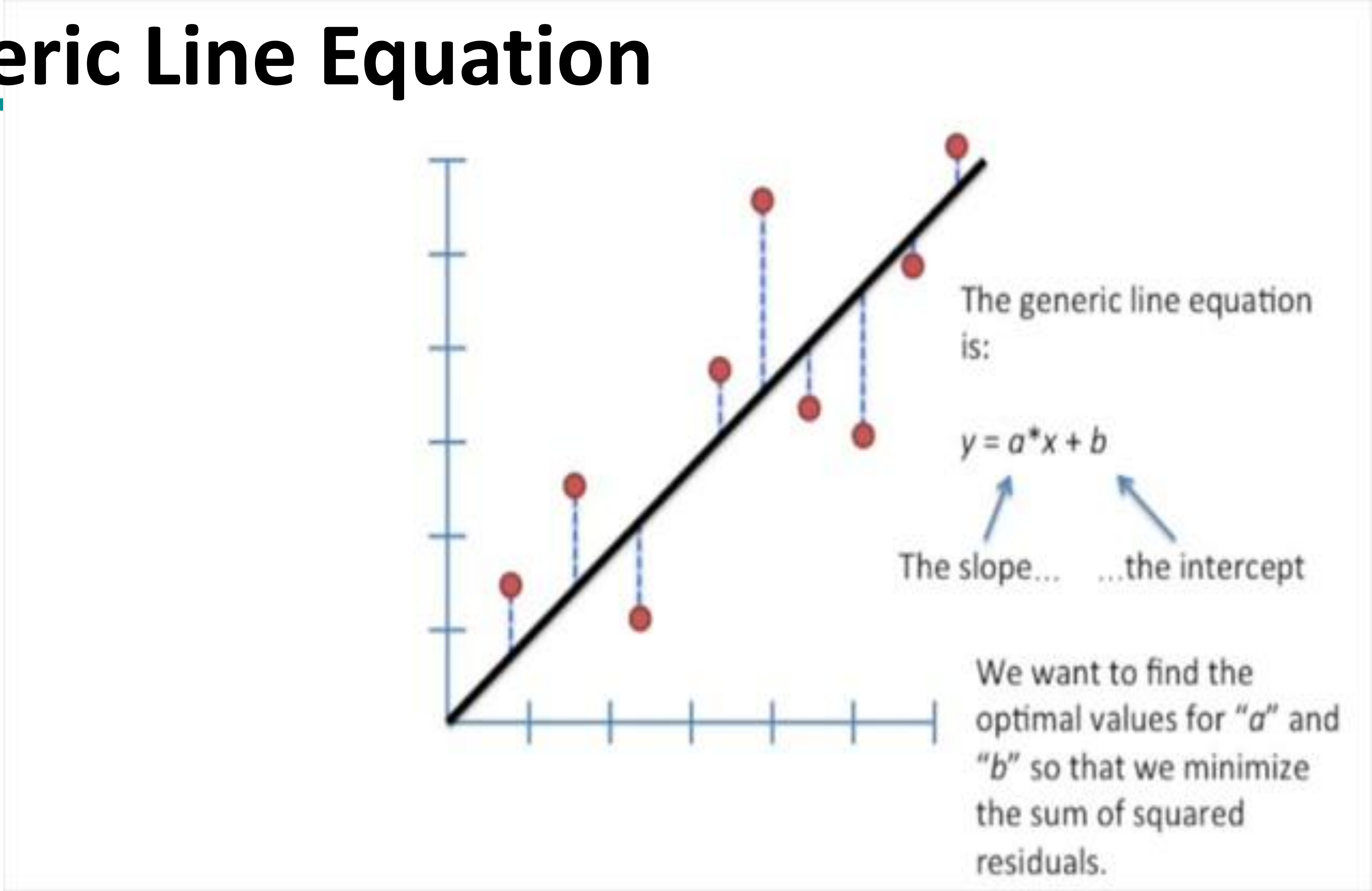
Generic Line Equation



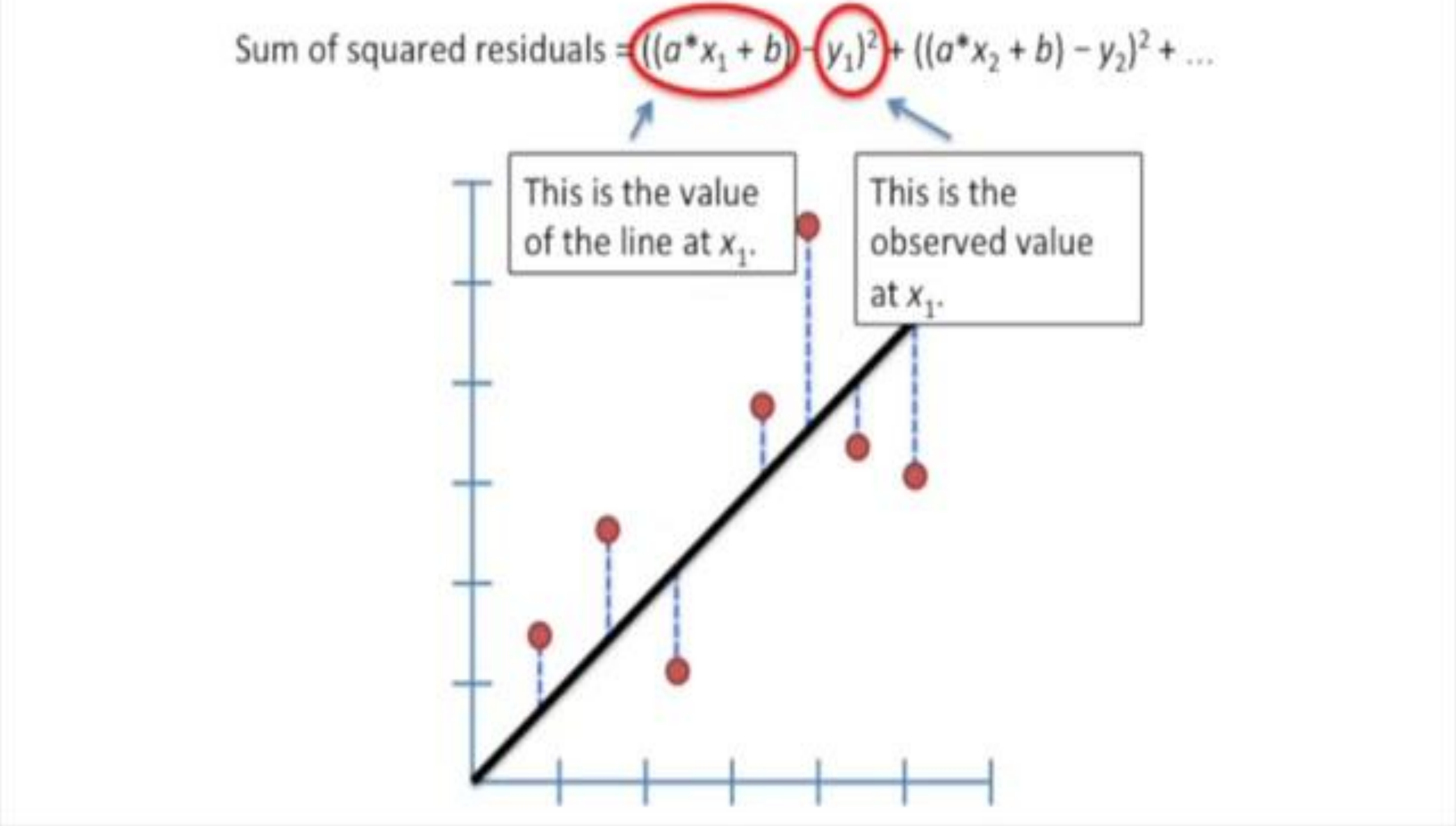
Generic Line Equation



Generic Line Equation



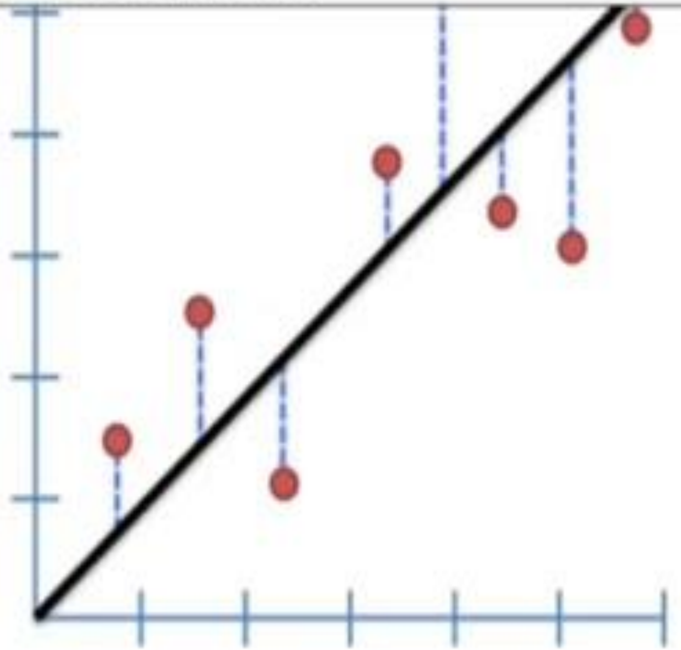
Least Square



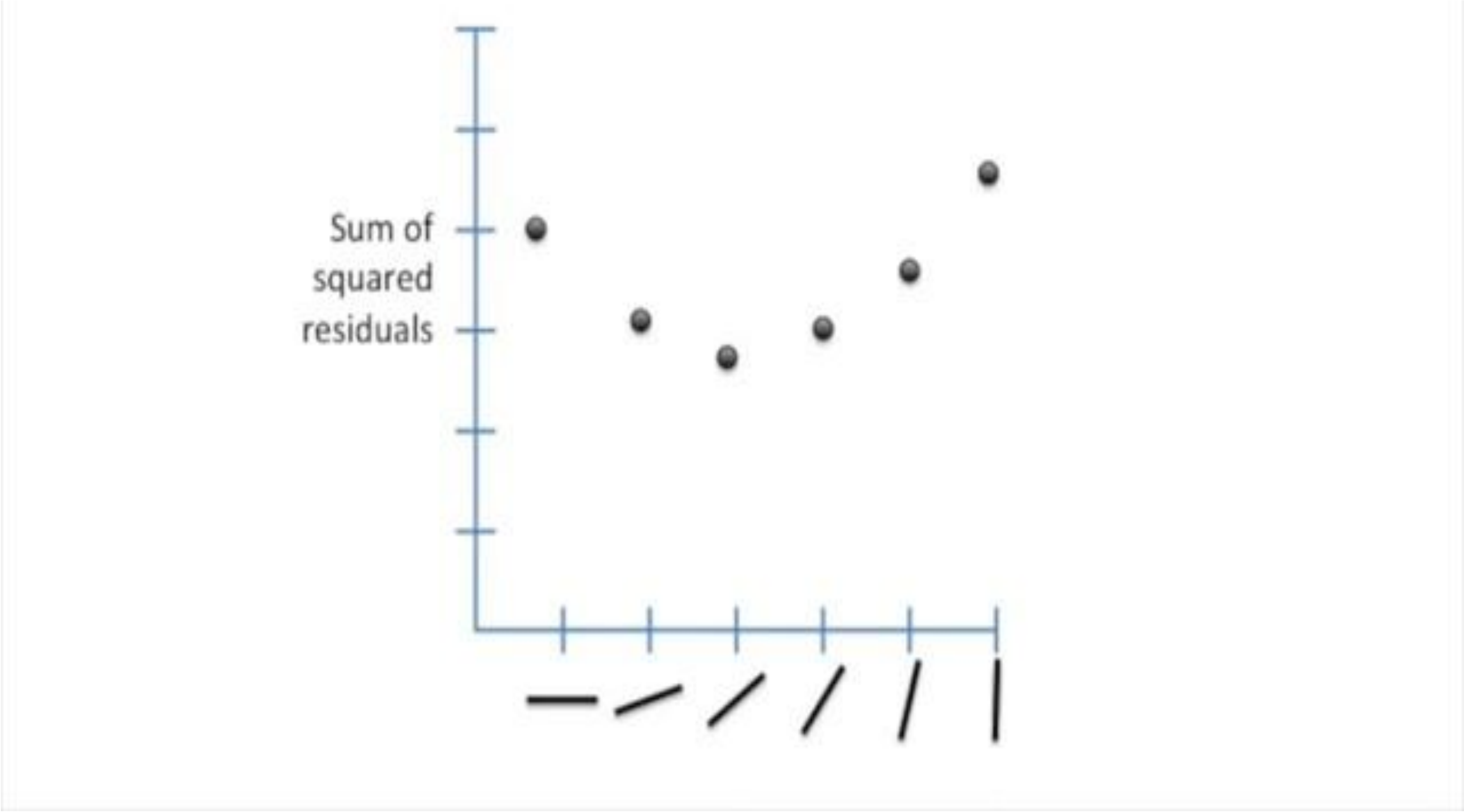
Least Square

Sum of squared residuals = $((a^*x_1 + b) - y_1)^2 + ((a^*x_2 + b) - y_2)^2 + \ldots$

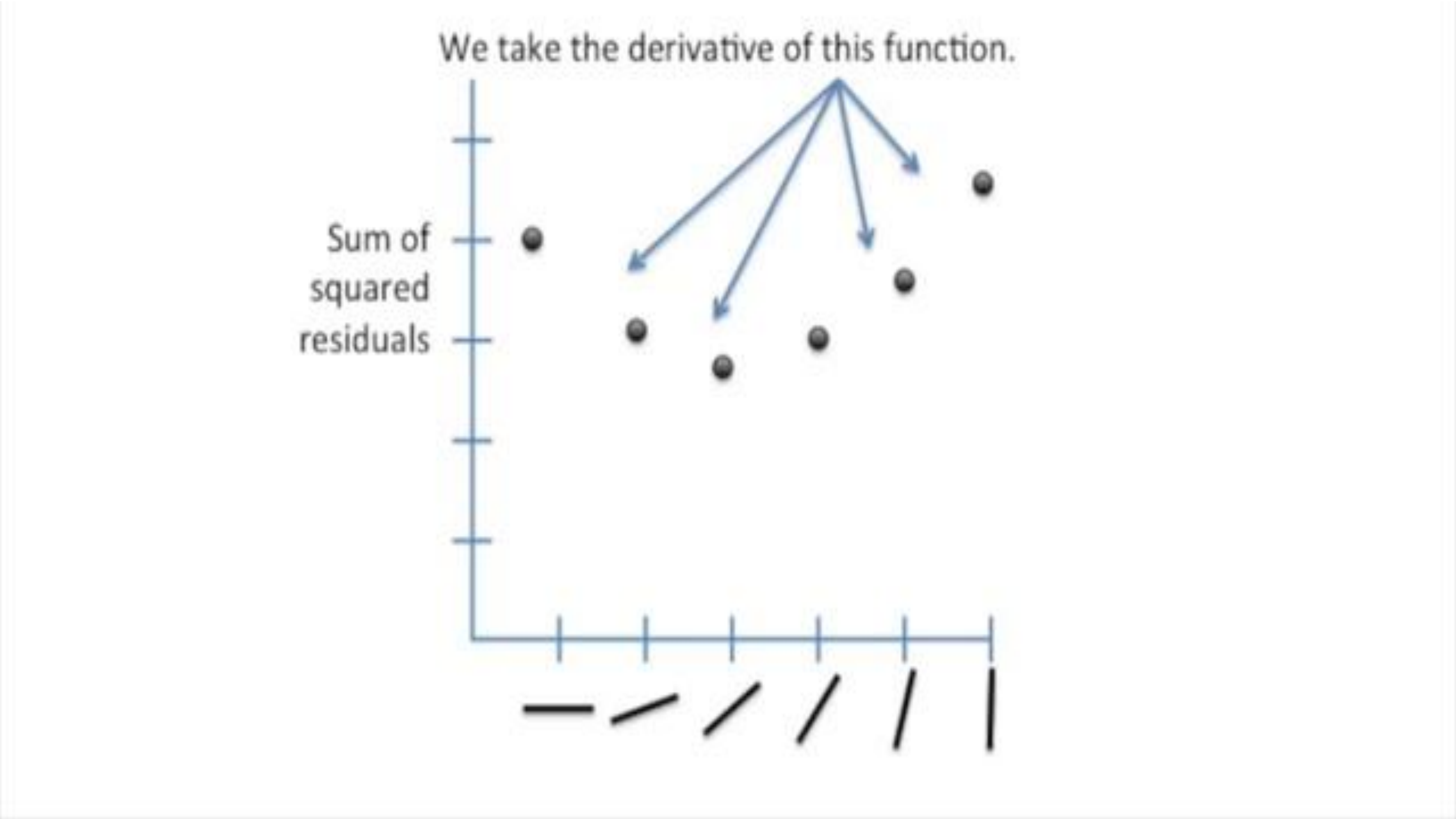
Since we want the line that will give us the smallest sum of squares, this method for finding the best values for "*a*" and "*b*" is called "Least Squares".



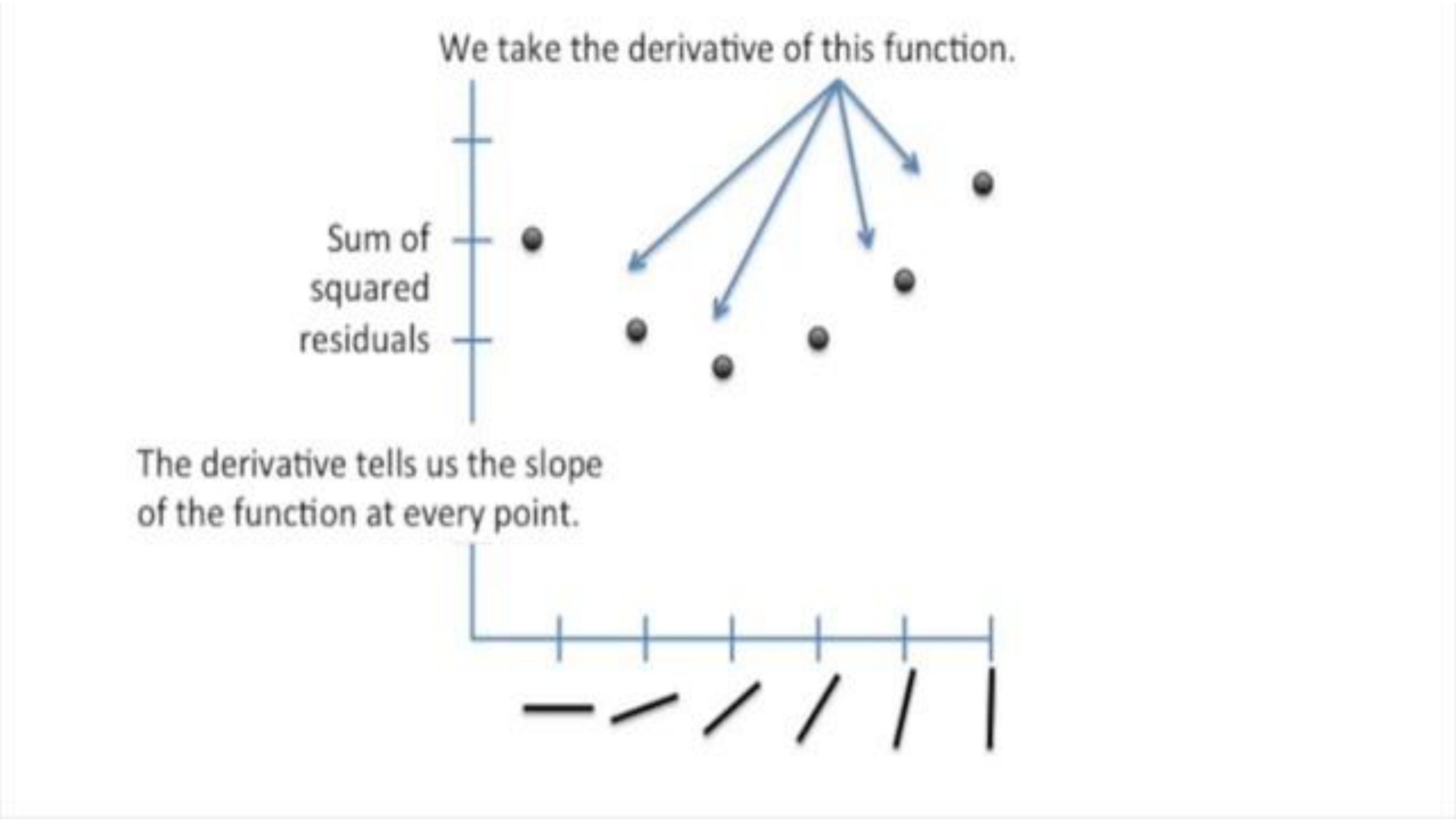
Sum of Squares Residuals



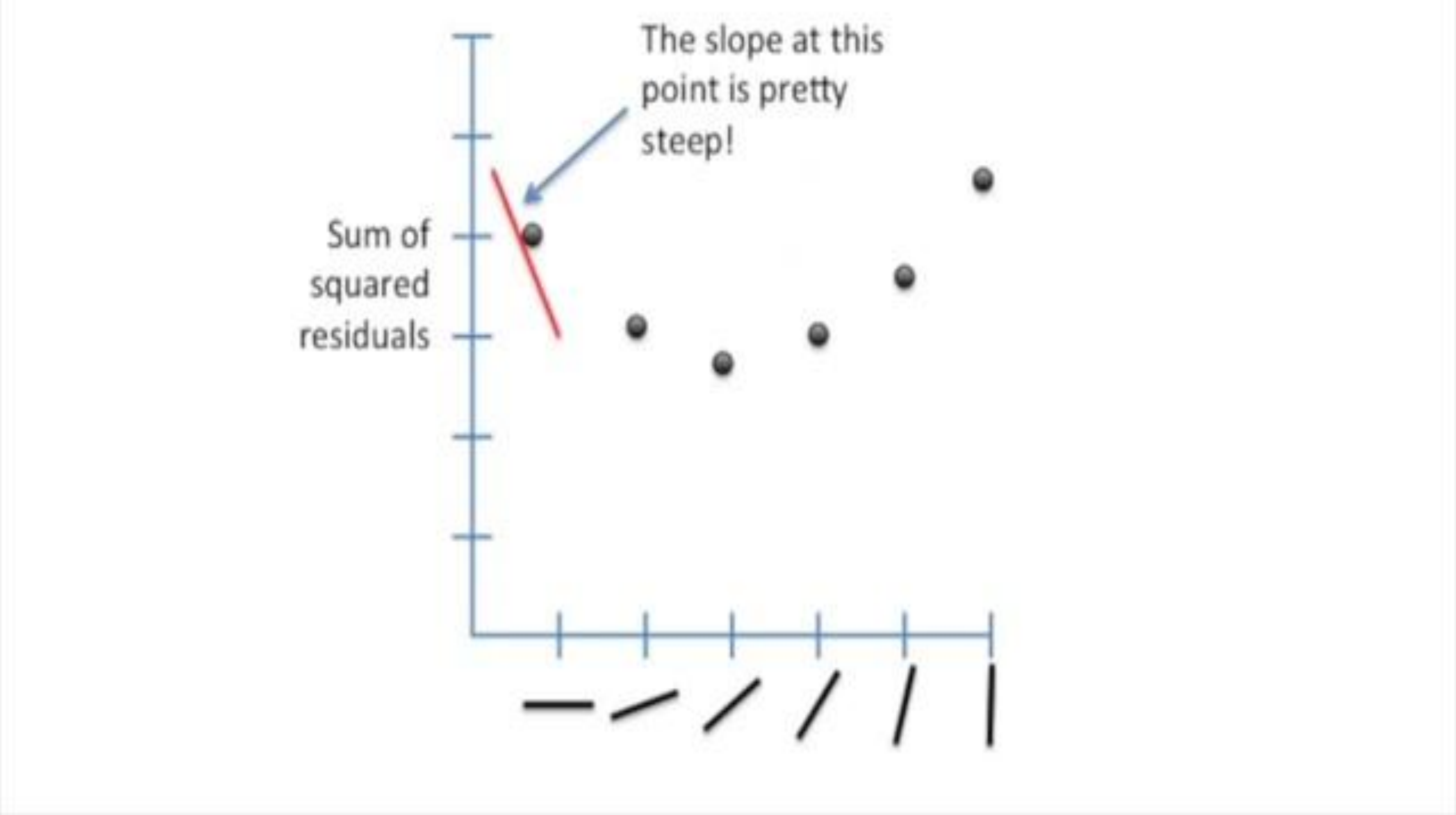
Finding Best Rotation



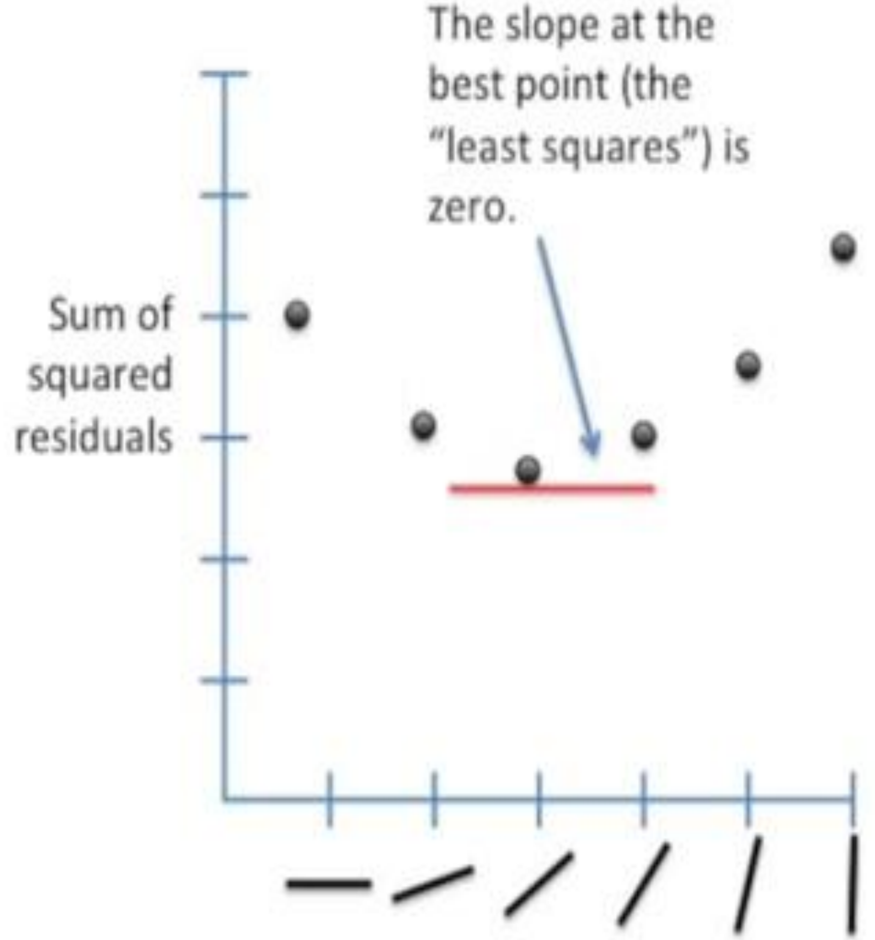
Finding Best Rotation



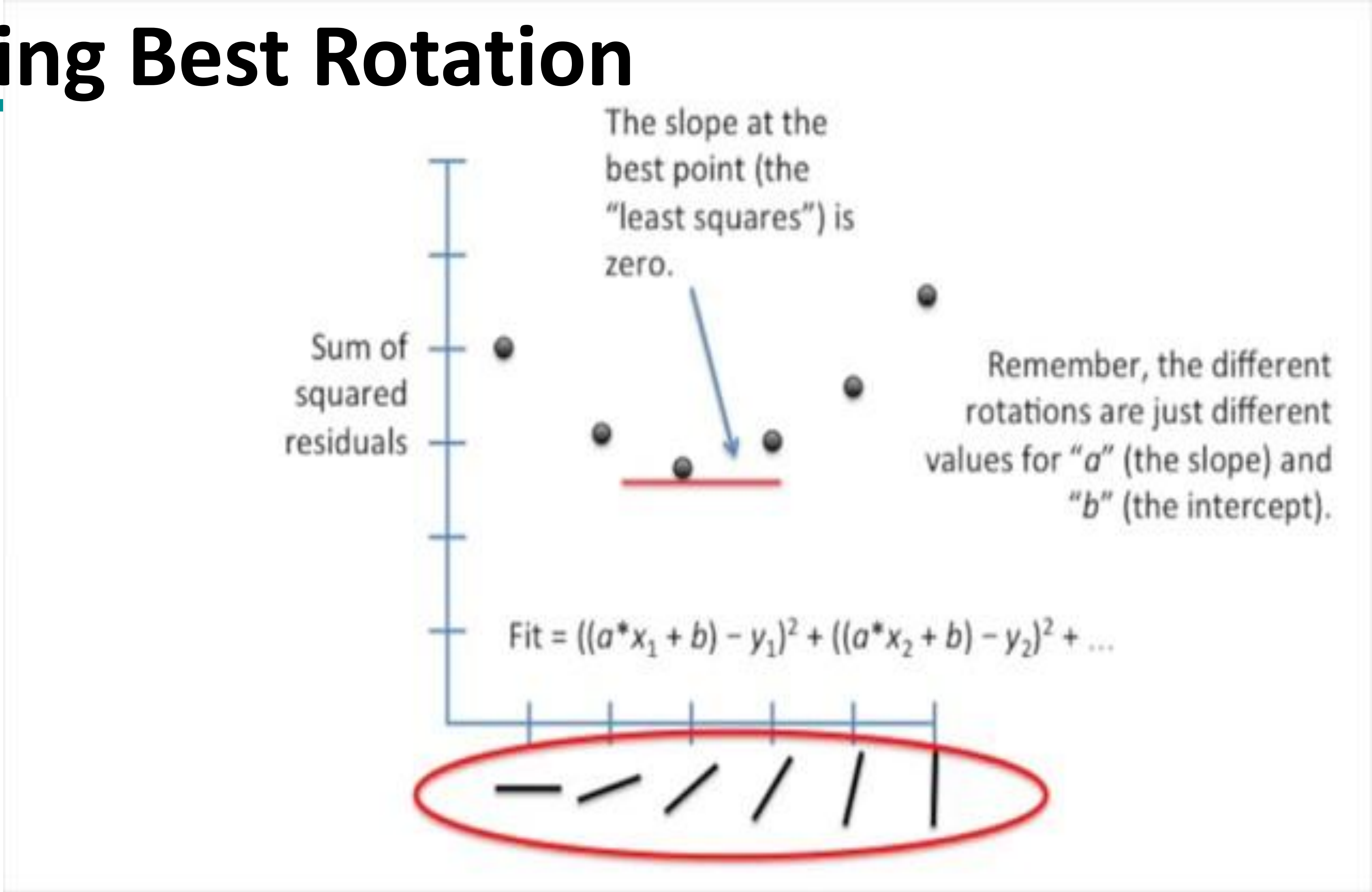
Finding Best Rotation



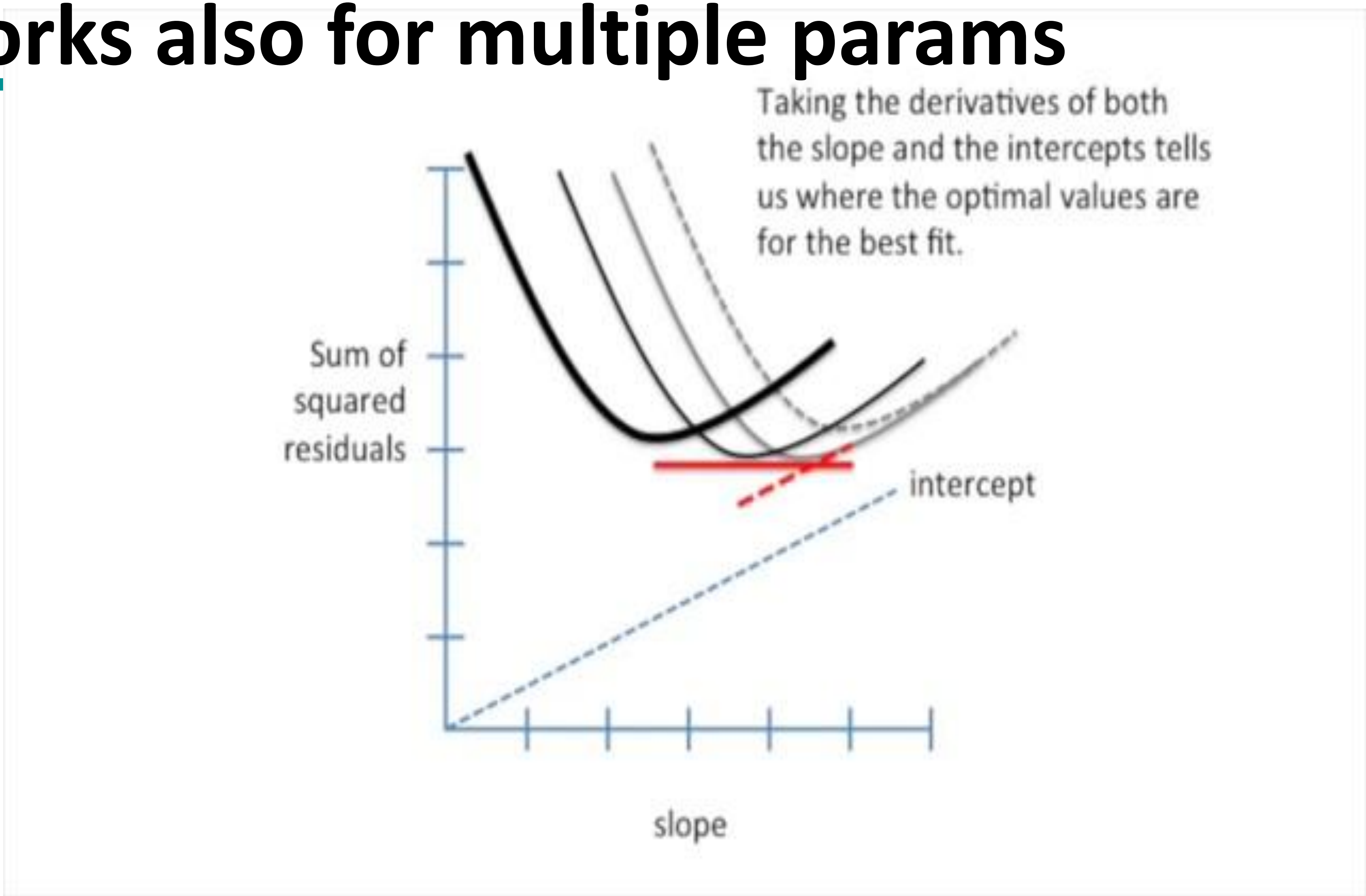
Finding Best Rotation



Finding Best Rotation



It works also for multiple params



Simple Linear Regression

Dependent Variable | Independent Variable

Linear Model: $$Y = mX + b \qquad Y = \beta_1 X + \beta_0$$

Slope | Intercept (bias)

- In general, such a relationship may not hold exactly for the largely unobserved population
- We call the unobserved deviations from *Y* the errors.
- The goal is to find estimated values *m'* and *b'* for the parameters *m* and *b* which would provide the "best" fit for the data points.

Least Square Method (LSM)

- A standard approach for doing this is to apply the **method of least squares** which attempts to find the parameters *m*, *b* that minimizes the sum of squared error.
- SSE = $\sum_i (y_i - f(x_i))^2$ = $\sum_i (y_i - mx_i \ - b)^2$
- also known as the **residual sum of squares**.
- The LSM finds *m, b* by setting to zero the first partial derivative of the above function w.r.t. *m* and *b* which are therefore calculated as follows:
- $m = (n \sum(xy) - \sum x \sum y) / (n \sum(x^2) - (\sum x)^2)$
- $b = (\sum y - m \sum x) / n$

- LSM can be extended to multiple linear regression.
- An alternative to find *m, b*, typically adopted in case of multivariate regression is the Gradient Descent method (see next lectures)

$m = (n \sum(xy) - \sum x \sum y) / (n \sum(x^2) - (\sum x)^2)$

$b = (\sum y - m \sum x) / n$

LSM - Example

"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

Let us find the best **m** (slope) and **b** (y-intercept) that suits that data

$y = mx + b$

$m = (n \sum(xy) - \sum x \sum y) / (n \sum(x^2) - (\sum x)^2)$

$b = (\sum y - m \sum x) / n$

LSM - Example

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 1: Calculate x^2 and xy

$m = (n \sum(xy) - \sum x \sum y) / (n \sum(x^2) - (\sum x)^2)$
$b = (\sum y - m \sum x) / n$

LSM - Example

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	Σy: 41	Σx^2: 168	Σxy: 263

Step 2: Sum all the columns

m = (n ∑(xy) - ∑x ∑y) / (n ∑(x²) – (∑x)²)
b = (∑y - m ∑x) / n

LSM - Example

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	**Σy: 41**	**Σx²: 168**	**Σxy: 263**

$$\mathbf{m} = \frac{N\ \Sigma(xy) - \Sigma x\ \Sigma y}{N\ \Sigma(x^2) - (\Sigma x)^2}$$

$$= \frac{5 \times 263 - 26 \times 41}{5 \times 168 - 26^2}$$

$$= \frac{1315 - 1066}{840 - 676}$$

$$= \frac{249}{164} = 1{,}5183...$$

$$\mathbf{b} = \frac{\Sigma y - m\ \Sigma x}{N}$$

$$= \frac{41 - 1{,}5183 \times 26}{5}$$

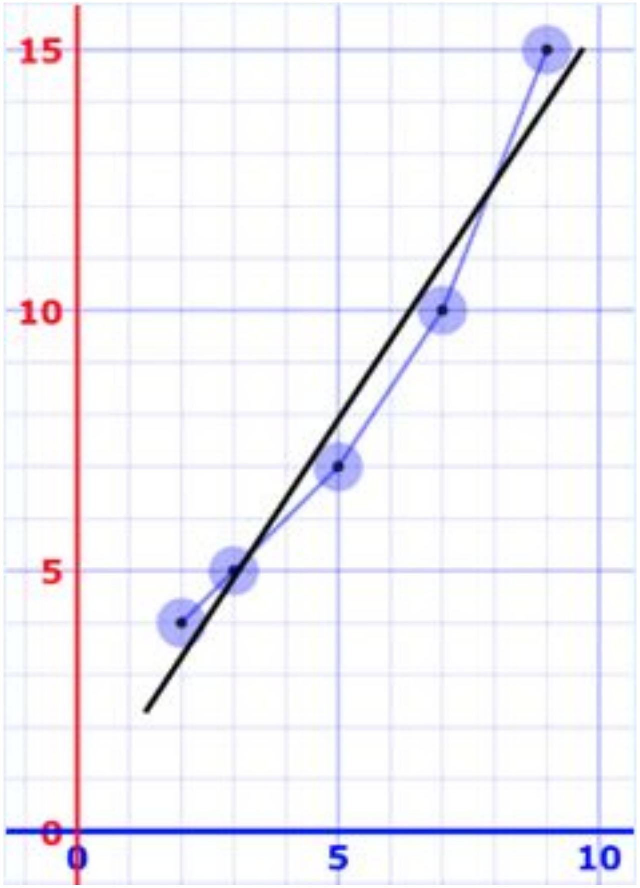
$$= 0{,}3049...$$

Step 3: Calculate the slope and the intercept with N = 5

$m = (n \sum(xy) - \sum x \sum y) / (n \sum(x^2) - (\sum x)^2)$
$b = (\sum y - m \sum x) / n$

LSM - Example

x	y	y = 1,518x + 0,305	error
2	4	3,34	−0,66
3	5	4,86	−0,14
5	7	7,89	0,89
7	10	10,93	0,93
9	15	13,97	−1,03

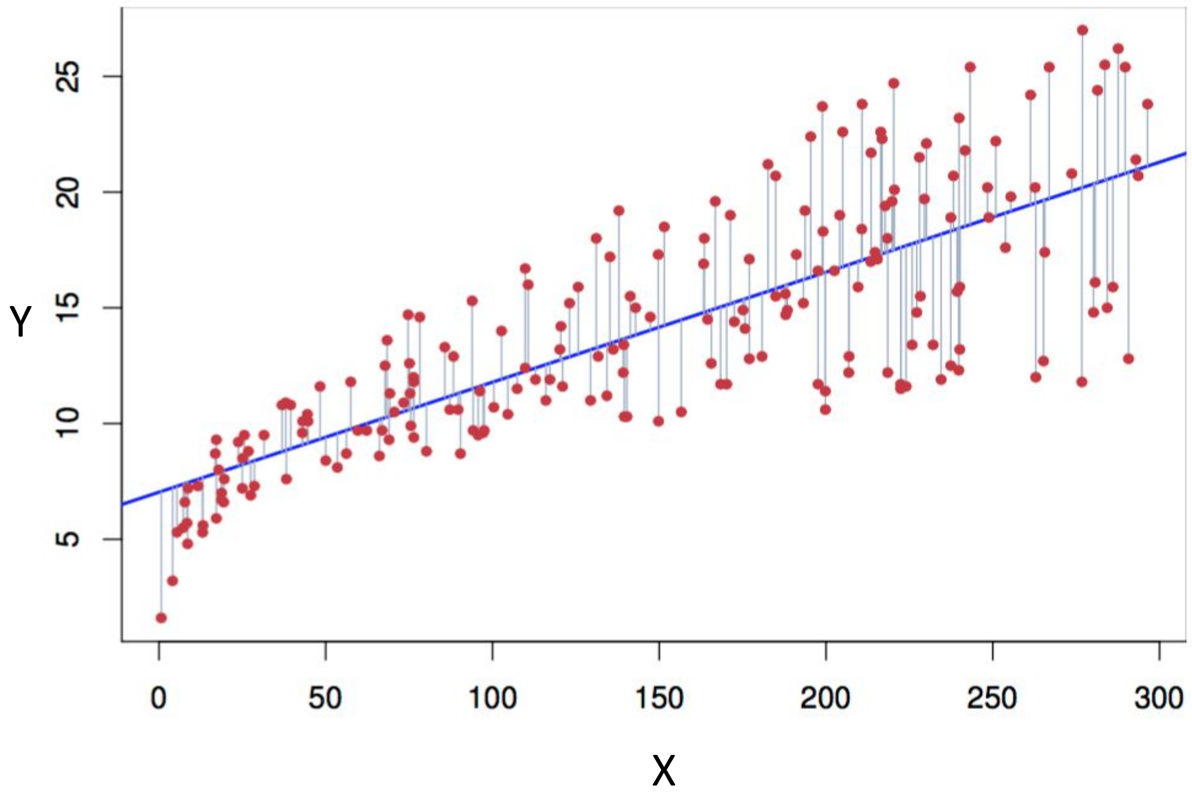


Step 4: test y = 1,518x + 0,305

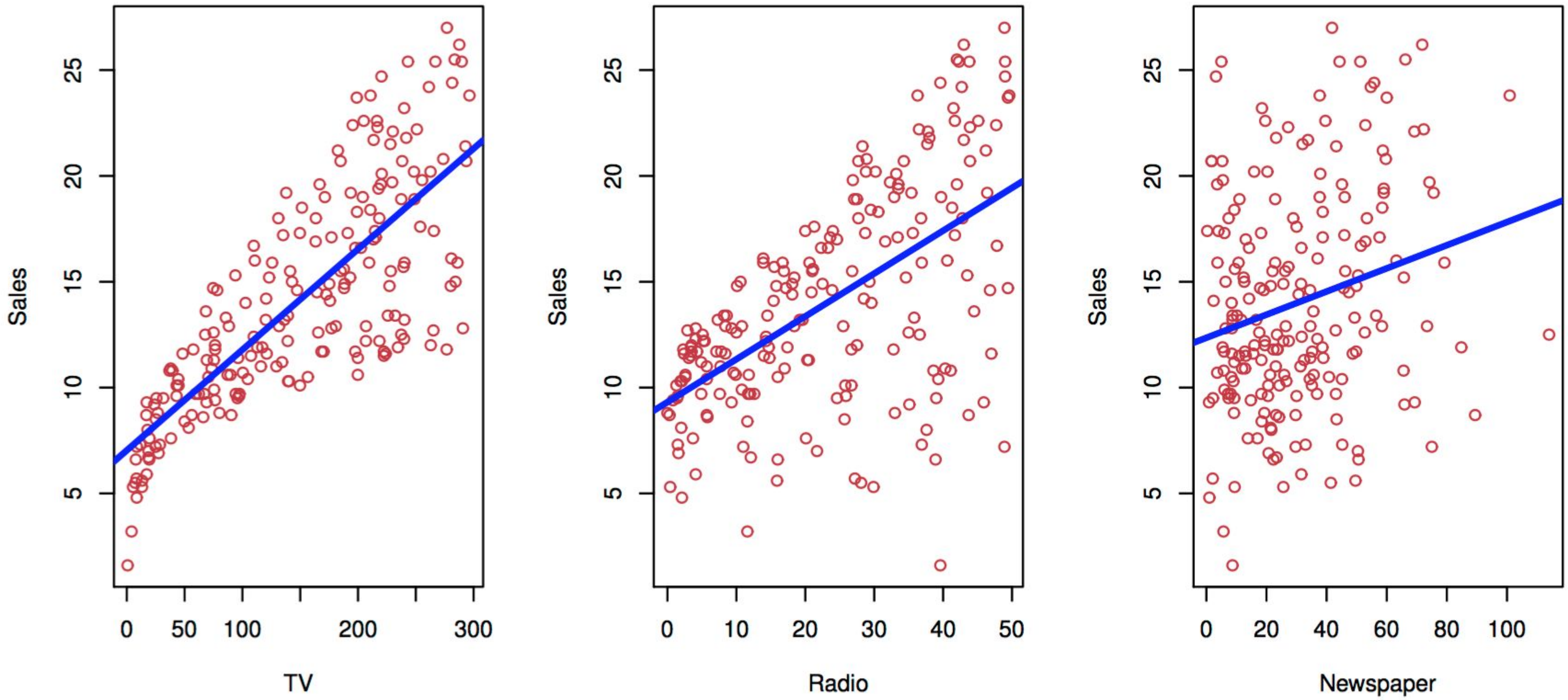
If x = 8 then we expect to sell 12,45 ice creams

Least Square Method

- Blue line shows the least square fit. Lines from red points to the regression line illustrate the residuals.
- For any other choice of slope *m* or intercept *b* the SSE between that line and the observed data would be larger than the SSE of the blue line.



Examples


Sales
25
20
15
10
5
0
50
100
200
300
TV
Sales
25
20
15
10
5
0
10
20
30
40
50
Radio
Sales
25
20
15
10
5
0
20
40
60
80
100
Newspaper

Alternative Fitting Methods

- Linear regressions fitted using gradient descent can benefit from some regularizations.
- However, they can be fitted in other ways, such as by minimizing a penalized version of the least squares cost function as in **ridge regression** (L2-norm penalty) and **lasso** (L1-norm penalty).
- **Tikhonov** regularization, also known as *ridge regression*, is a method of regularization of ill-posed problems particularly useful to mitigate the multicollinearity, which commonly occurs in models with large numbers of parameters.
- **Lasso** (least absolute shrinkage and selection operator) performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

Multicollinearity: is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

Linear Regression Models Objective Functions

- Simple $\beta_0 + \beta_1 x - y$
- Multiple $\beta_0 + \sum_i (y_i - \beta_i x_i)^2$
- Ridge $\beta_0 + \sum_i (y_i - \beta_i x_i)^2 + \lambda \sum_j {\beta_j}^2$
- Lasso $\beta_0 + \sum_i (y_i - \beta_i x_i)^2 + \lambda \sum_j |\beta_j|$

regularization

Ridge: mitigate the problem of multicollinearity

Lasso: variable selection, i.e., minimizes the number of coefficient different from zeros

Evaluating Regression

- **Coefficient of determination** R^2
  - is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

hat means predicted

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$

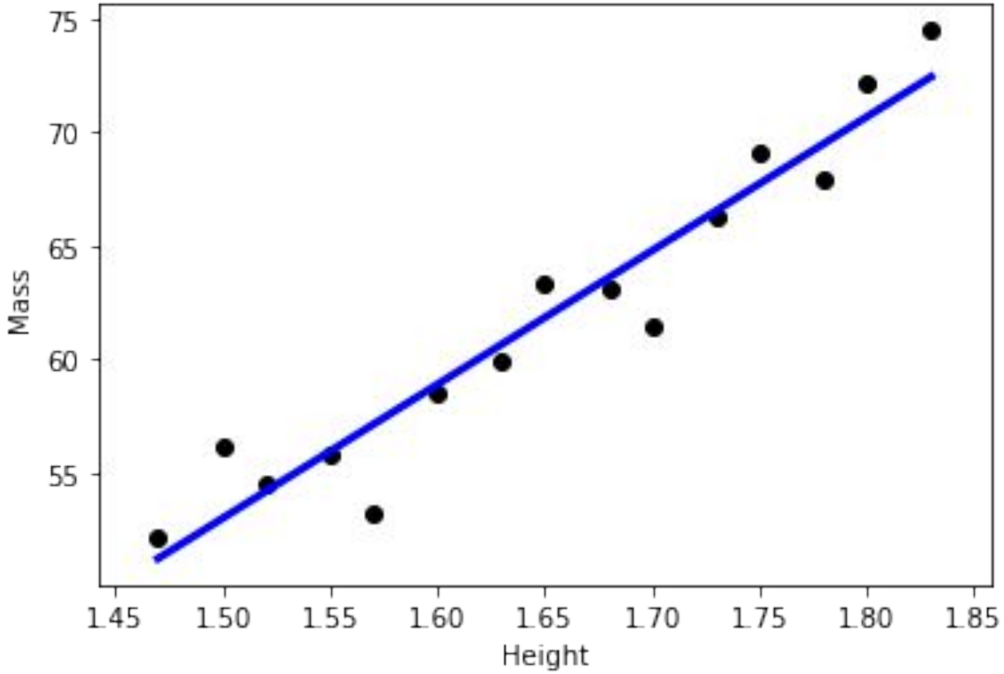
- **Mean Squared/Absolute Error** MSE/MAE
  - a risk metric corresponding to the expected value of the squared (quadratic)/absolute error or loss

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \quad \text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

Example

- Height (m): 1.47, 1.50, 1.52, 1.55, 1.57, 1.60, 1.63, 1.65, 1.68, 1.70, 1.73, 1.75, 1.78, 1.80, 1.83
- Mass (kg): 52.21, 56.12, 54.48, 55.84, 53.20, 58.57, 59.93, 63.29, 63.11, 61.47, 66.28, 69.10, 67.92, 72.19, 74.46

- Intercept: -35.30454824113264
- Coefficient: 58.87472632
- R^2: 0.93
- MSE: 3.40
- MAE: 1.43

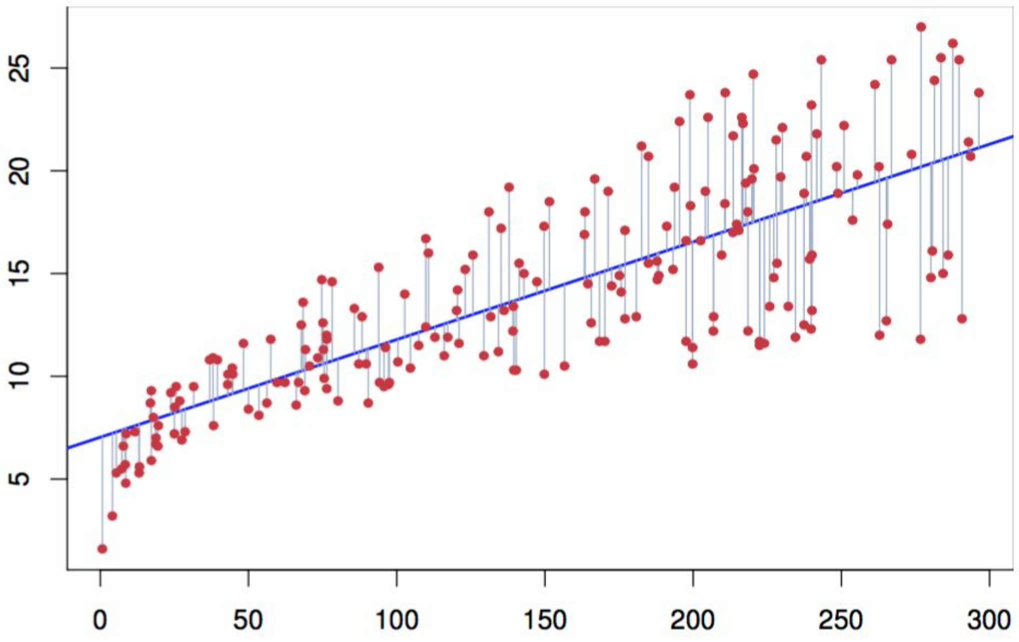


Nonlinear Regression

Linear Regression Recap

- Linear regression is used to fit a linear model to data where the dependent variable is continuous.
- Given a set of points *(X_i,Y_i)*, we wish to find a linear function (or line in 2 dimensions) that “goes through” these points.
- In general, the points are not exactly aligned.
- The objective is to find the line that best fits the points.

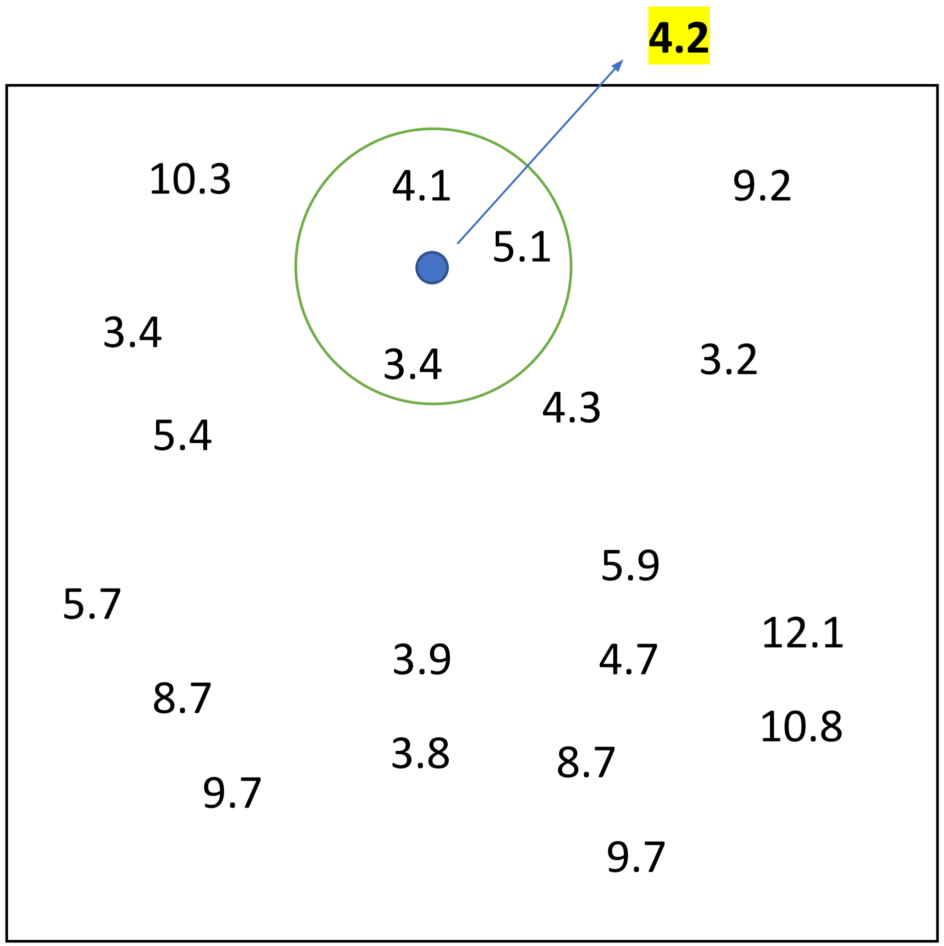
$$Y = \beta_1 X + \beta_0$$



k-NN for Regression

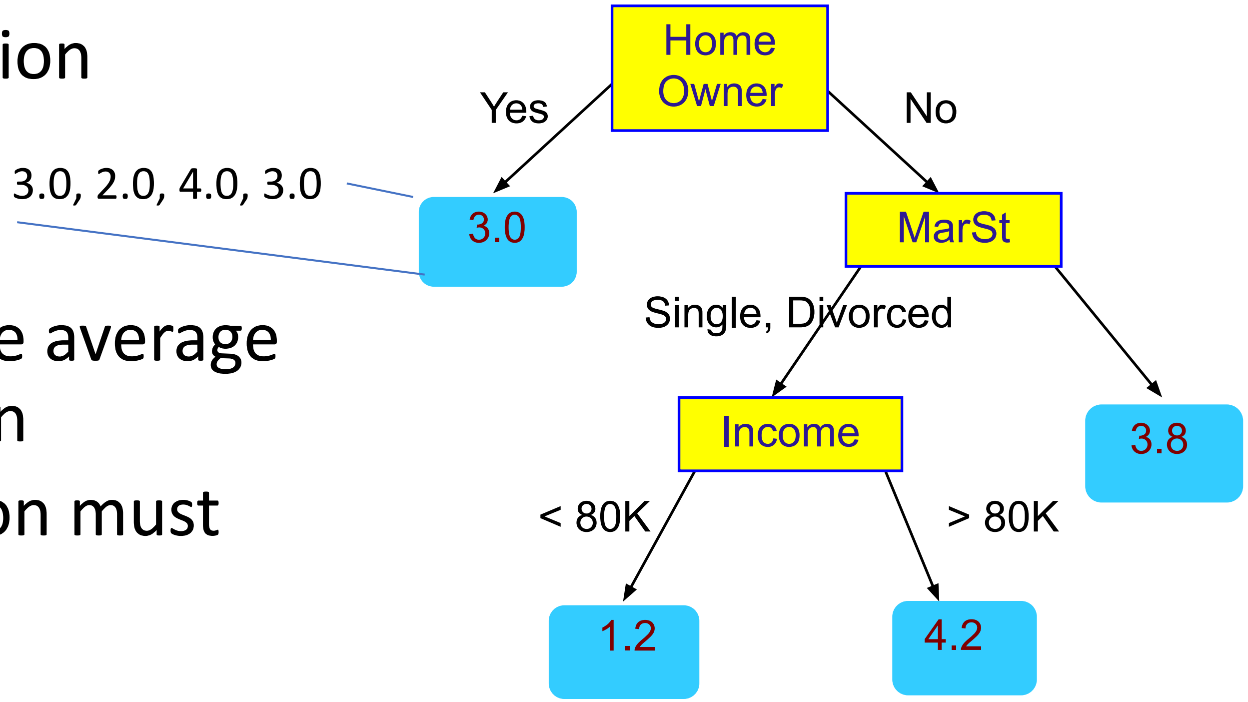
Given a set of training records (memory), and a test record:

1. ***Compute the distances*** from the records in the training to the test.
2. ***Identify the k "nearest" records***.
3. Use target value of nearest neighbors to ***determine the value*** of unknown record (e.g., by averaging the values).



Decision Trees for Regression

- The same induction and application procedures can be used.
- The only differences are:
  - When leaves are not pure, the average value is returned as prediction
  - Different optimization criterion must be used such as
    - MSE
    - MAE



$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \quad \text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

References

- Regression. Appendix D. Introduction to Data Mining.