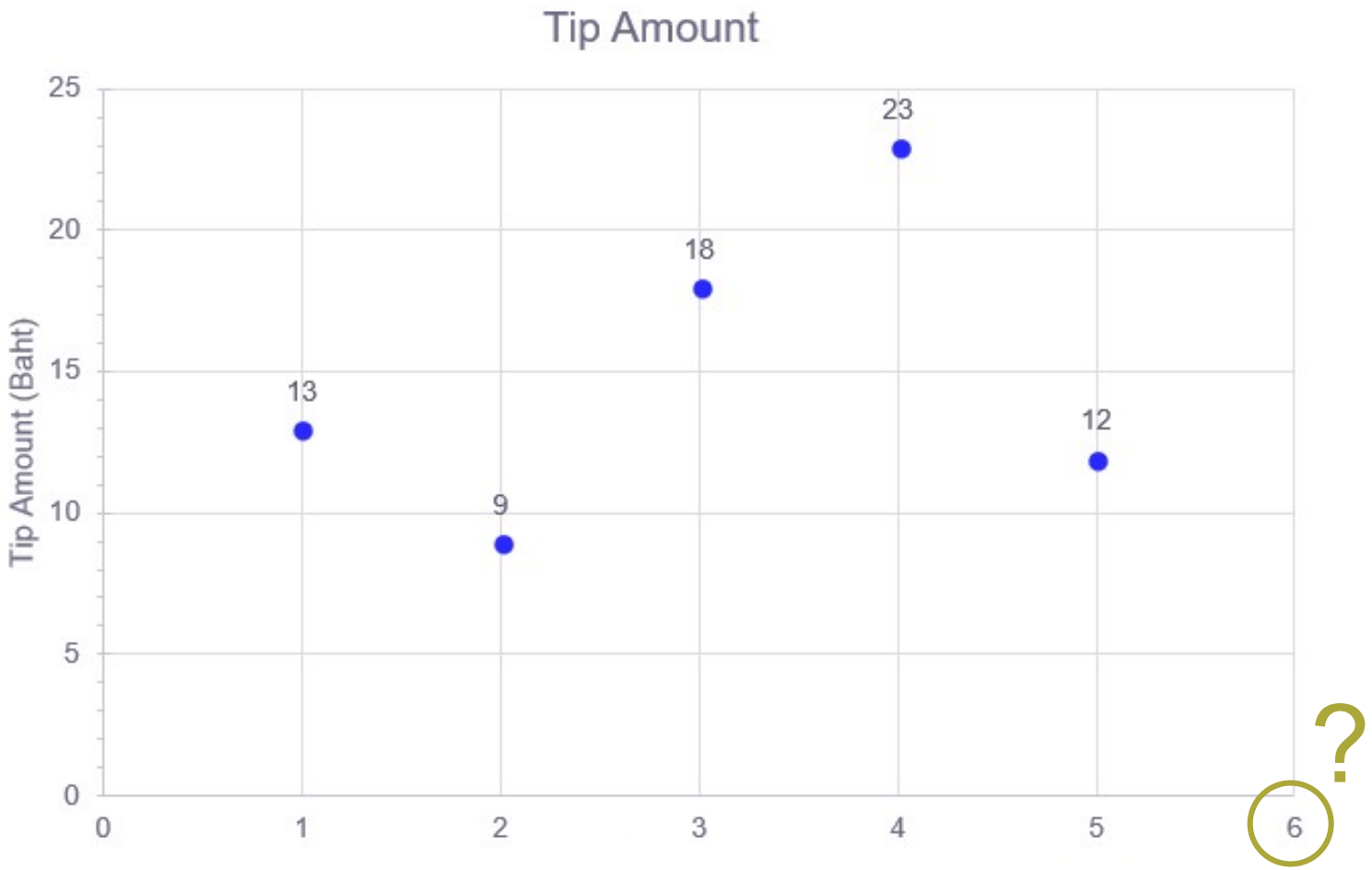
REGRESSION

Concept

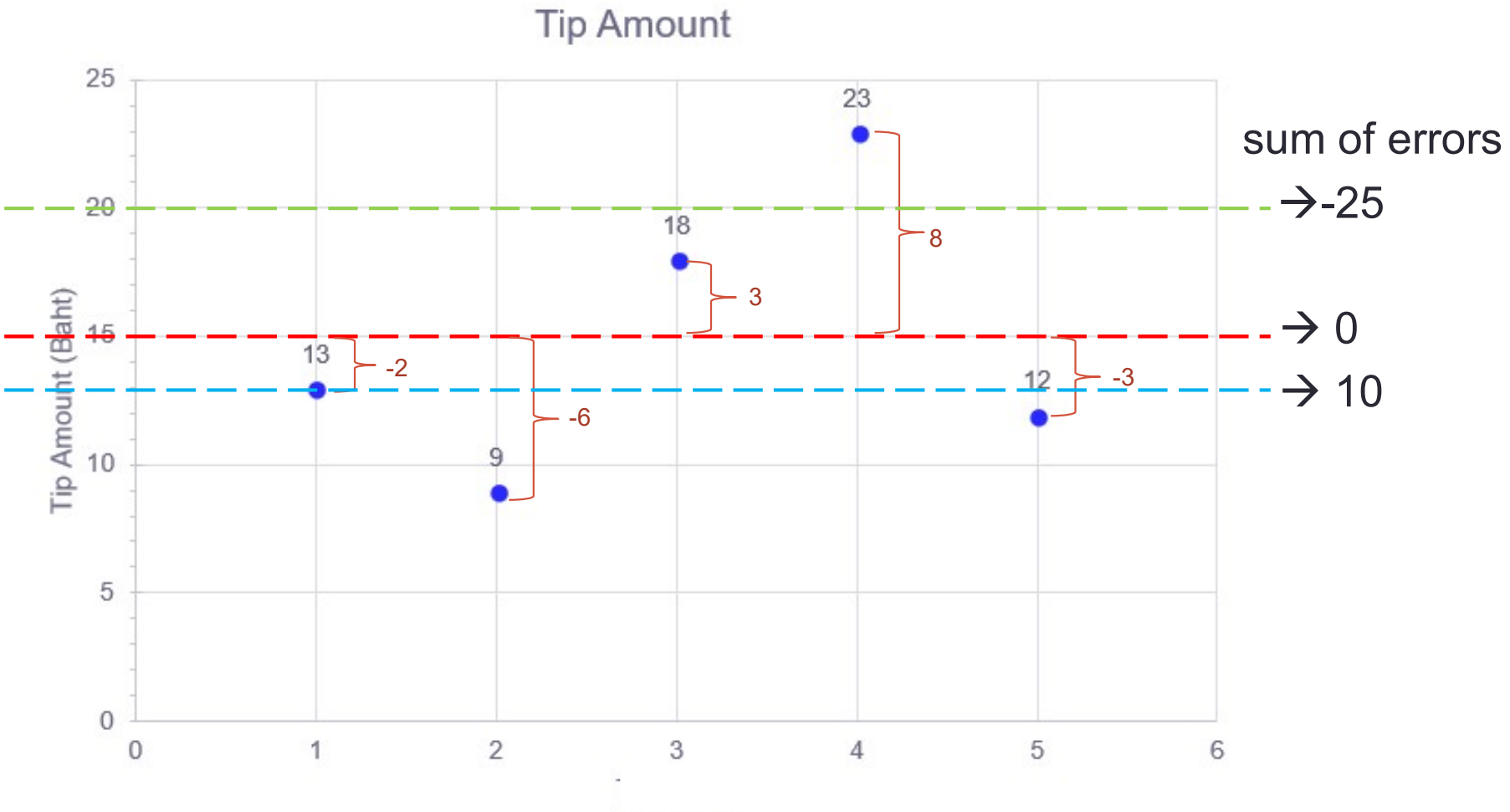
Simple Linear Regression

- One dependent variable – Tip Amount



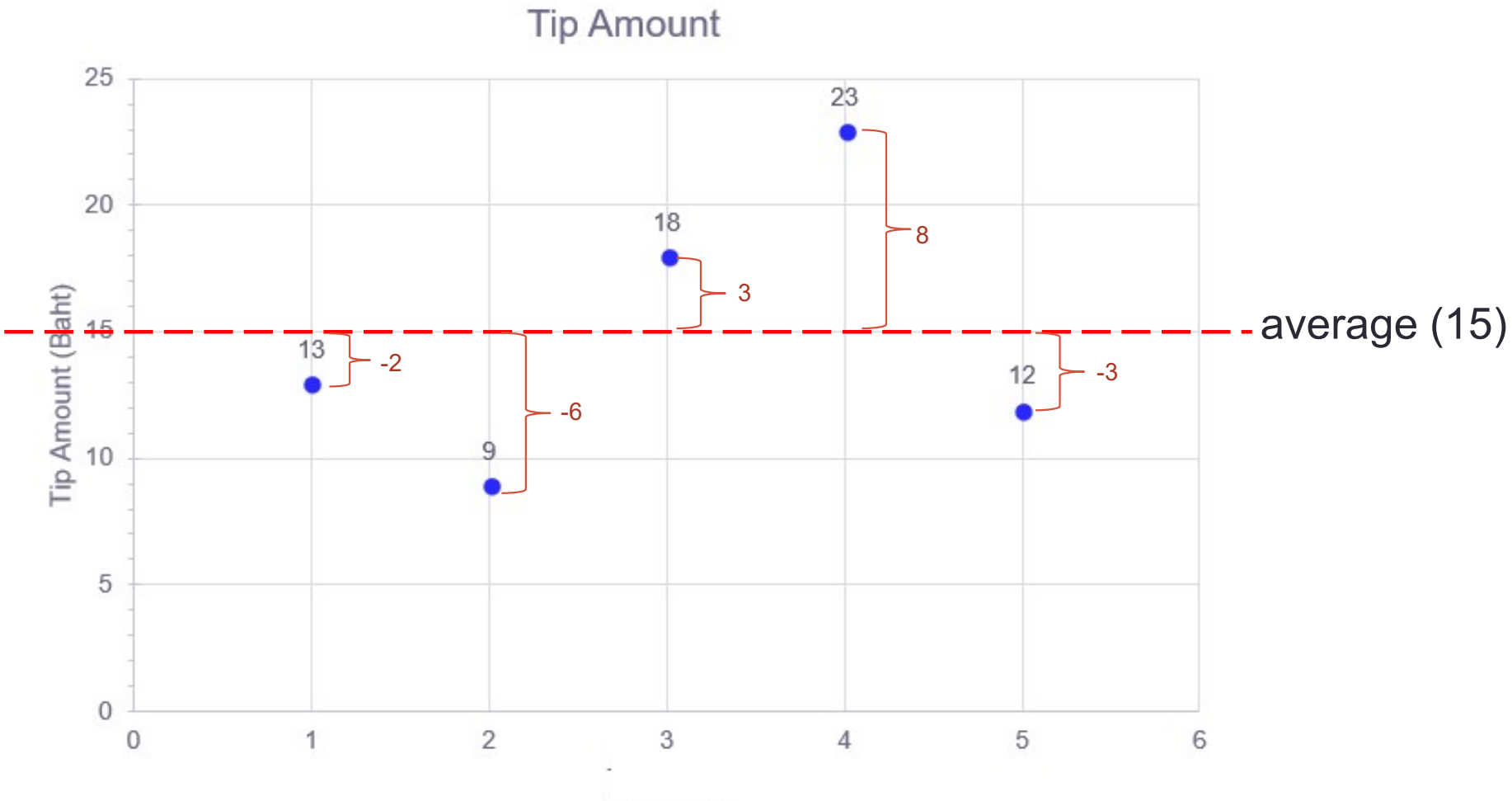
The Best Fit Line

error = distance from the estimated value



Estimated Tip Amount

Sum of residuals (errors) = -2 + -6 + 3 + 8 + -3 = 0



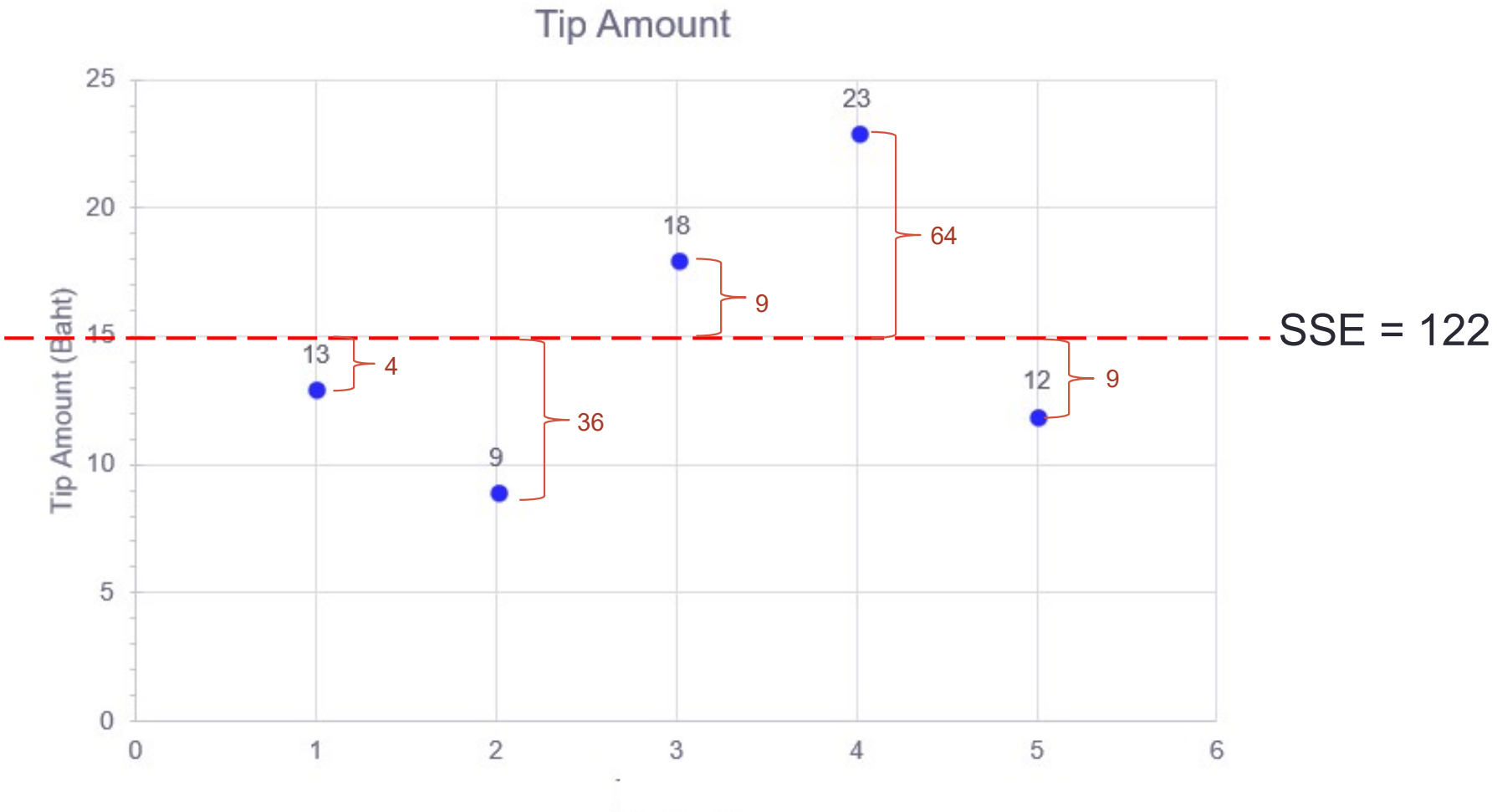
Squared Residuals / Errors

- Make numbers positive
- Emphasizes large deviation

Sum of Squared Errors (SSE)

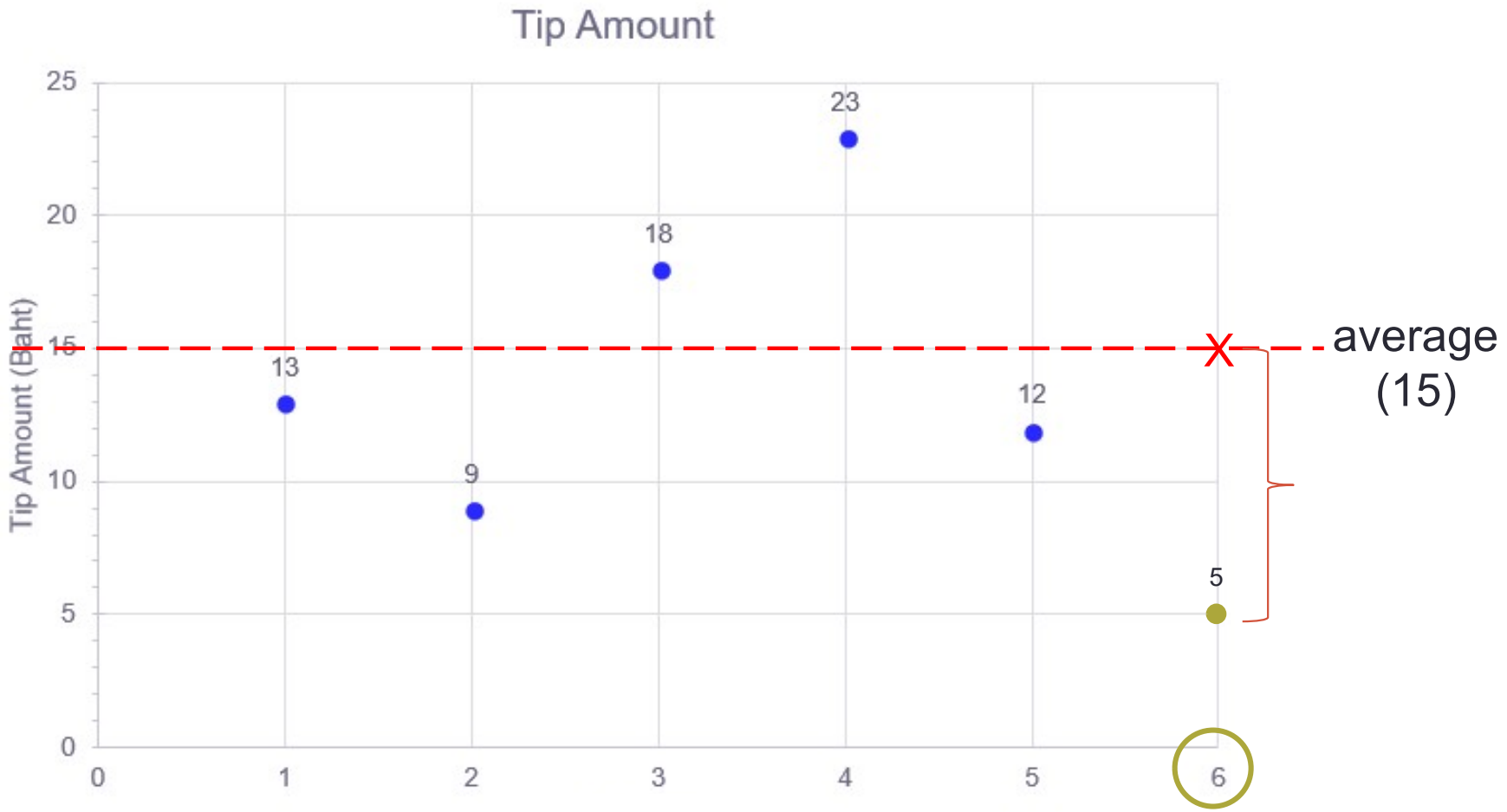
SSE

Sum of squared error = 4 + 36 + 9 + 64 + 9 = 122



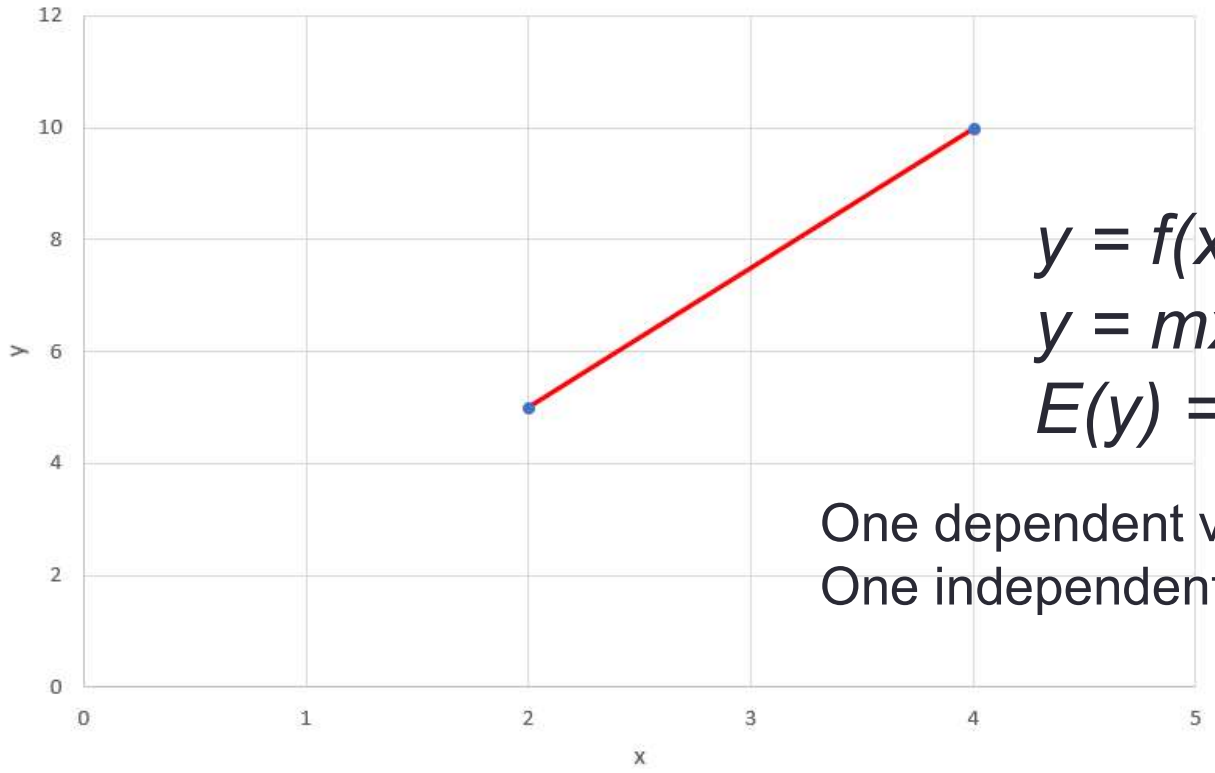
So, How much is the next tip ?

Sum of errors = -10
Sum of squared errors = 222



Linear Regression

- Regression line : a line that is as close to every dot as possible
- Sum of squared residuals/errors (SSE) is at the minimum

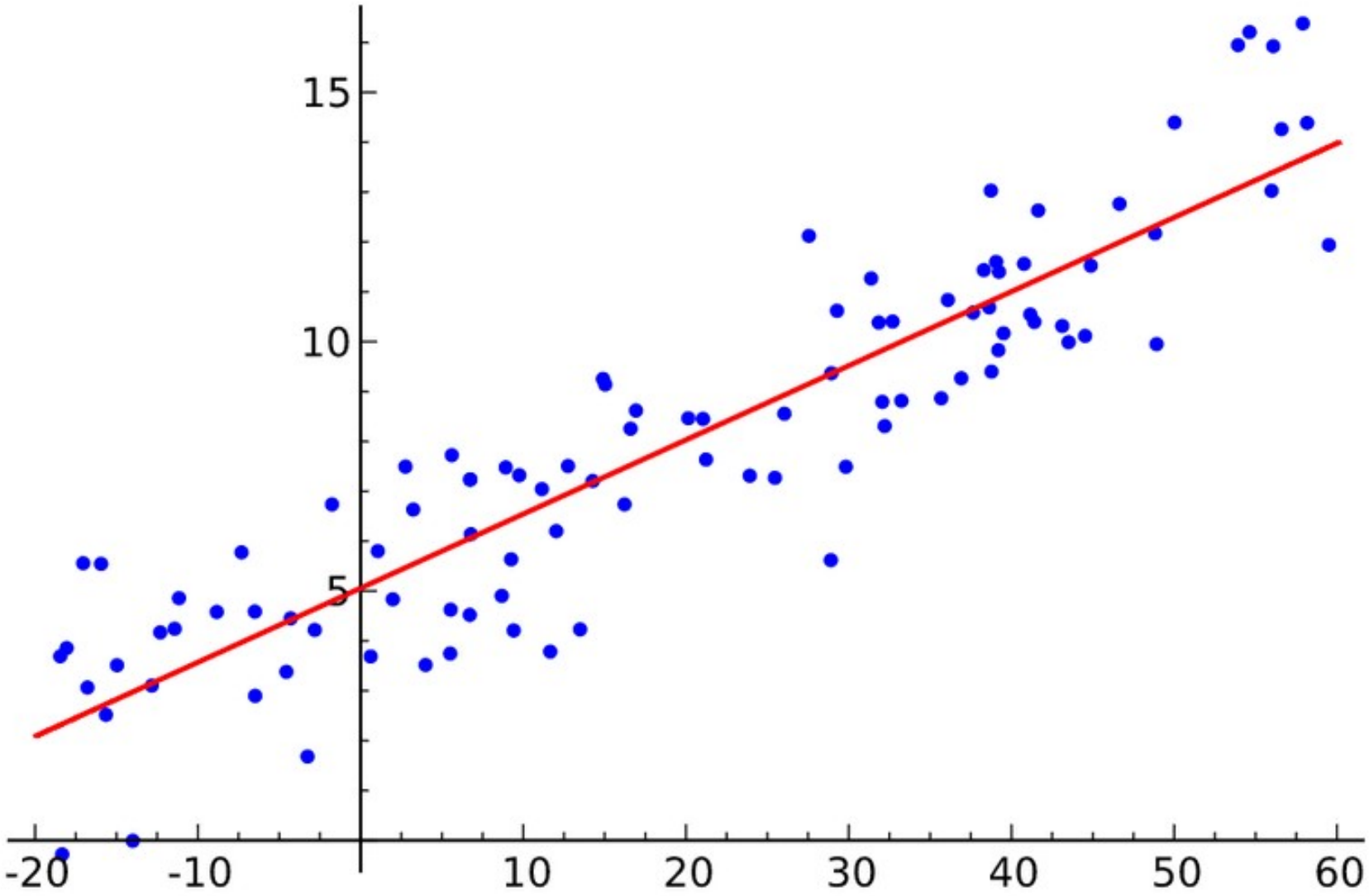


$y = f(x)$

$y = mx + b$

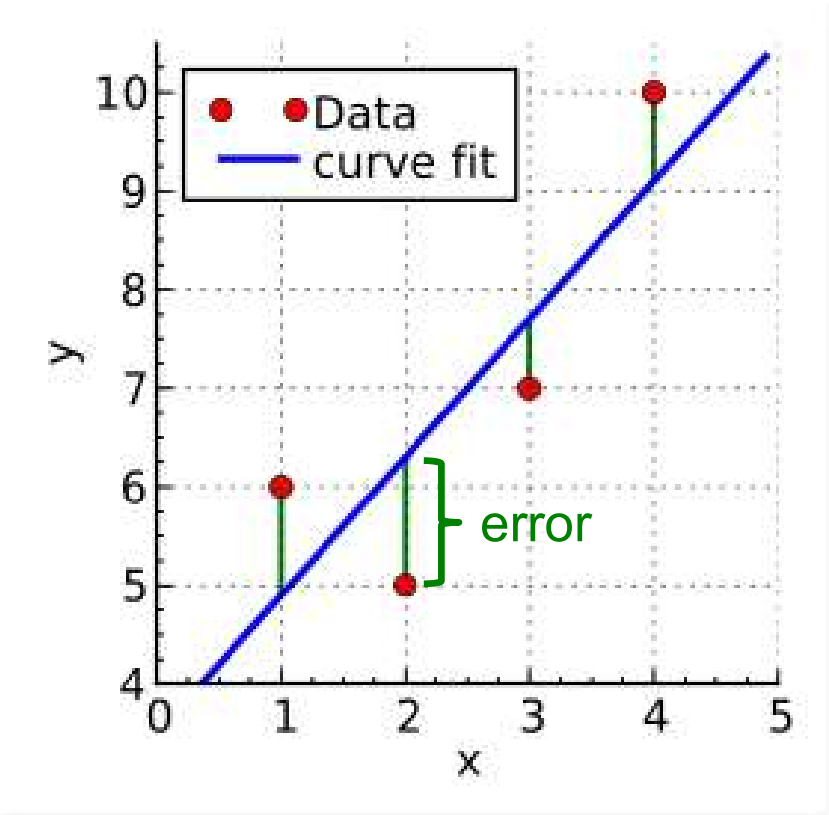
$E(y) = \beta_0 + \beta_1 x$

One dependent variable, y
One independent variable, x


15
10
5
-20
-10
10
20
30
40
50
60

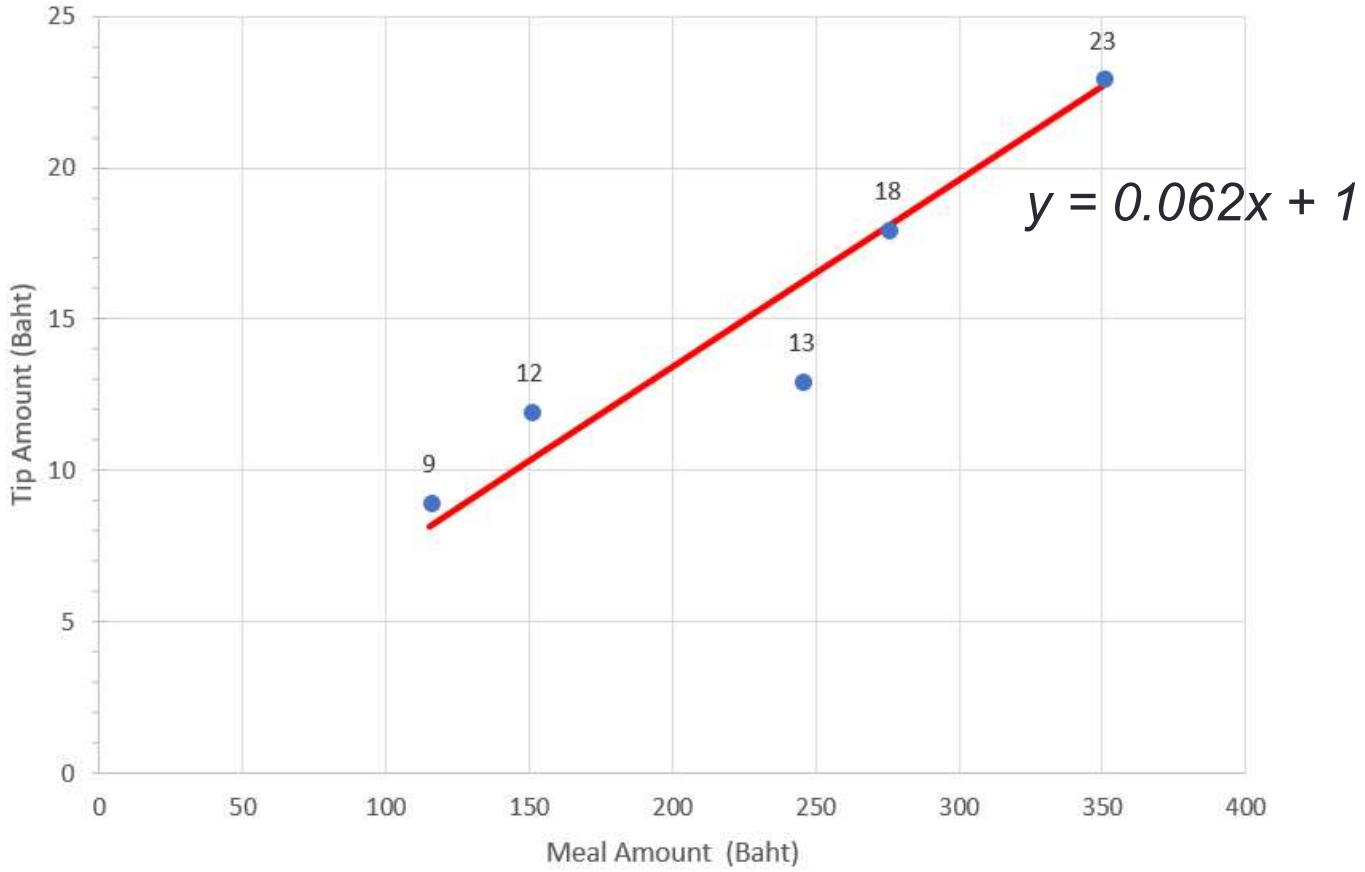
Error

- Vertical distance from regression line



Tip Amount vs Meal Amount

Sum of errors = -0.37
Sum of squared errors = 13.92



Least Squares Method

- Least squares criterion

$$min\ \Sigma\ (y_i - \hat{y}_i)^2$$

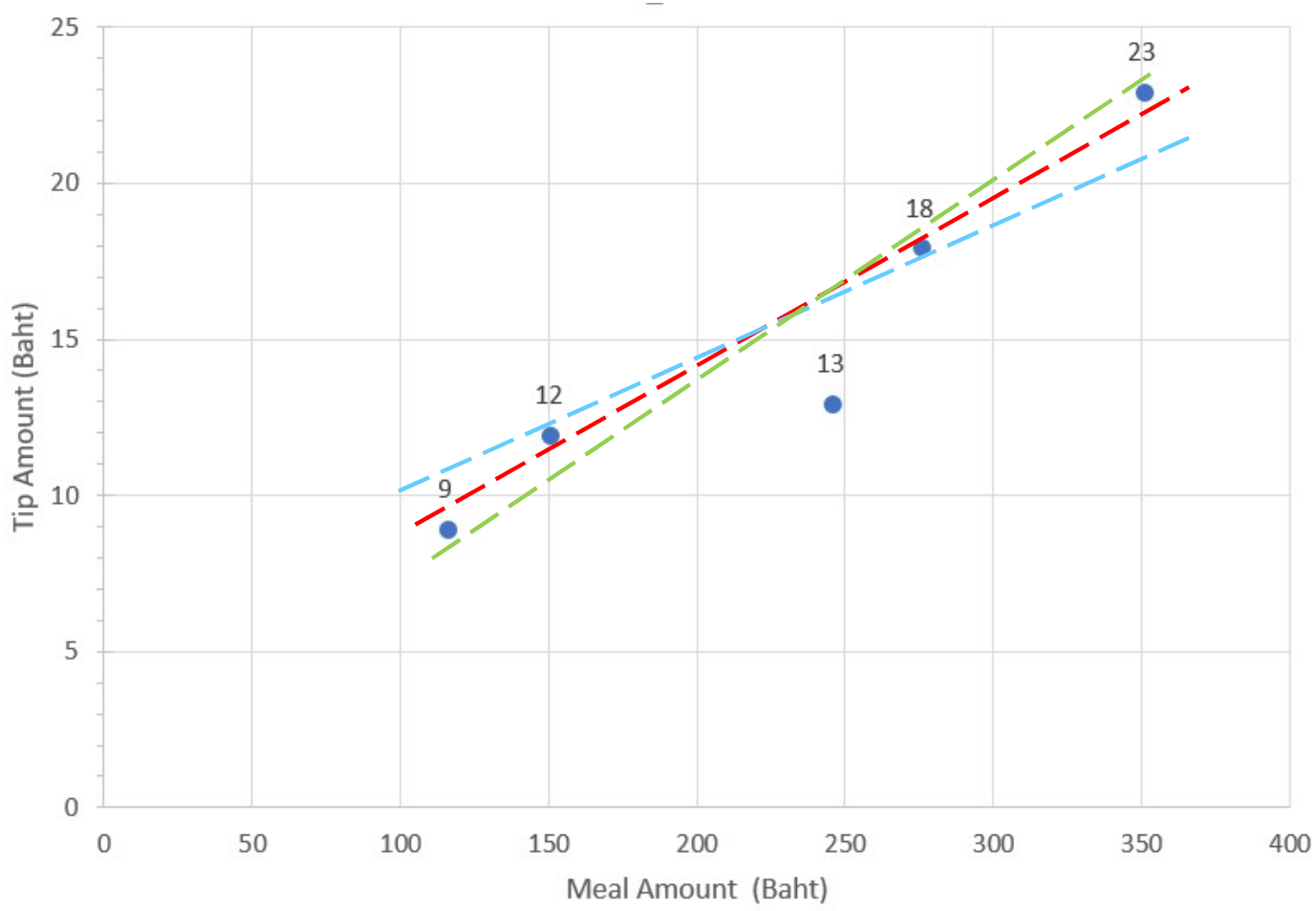
Tip Amount

y_i = *observed value of the dependent variable*
$\hat{y}_i$ = *estimated value of the dependent variable*

Note: sum of squared residuals should be much smaller than when only one dependent variable is considered

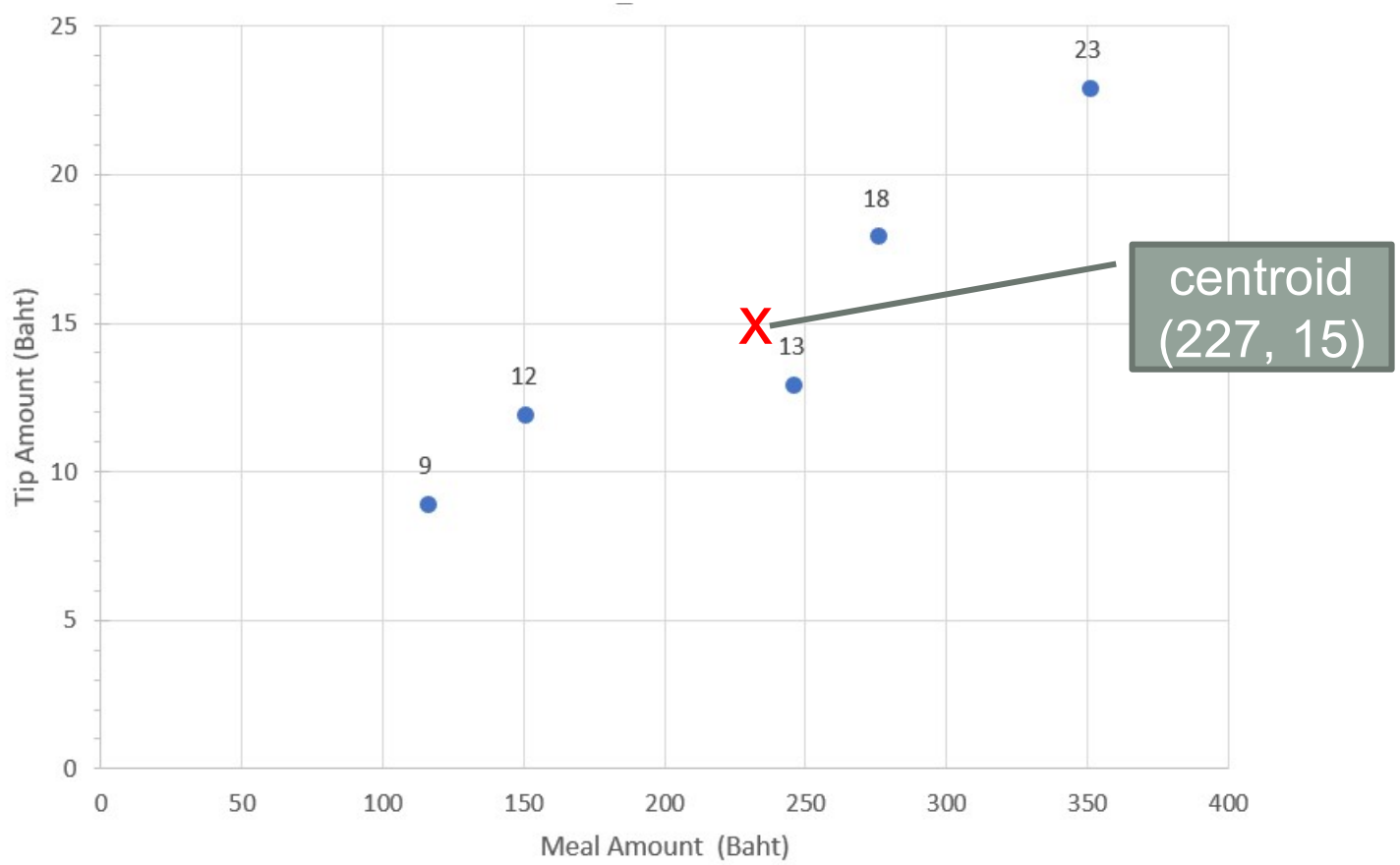
122 in our example

Which Line ?

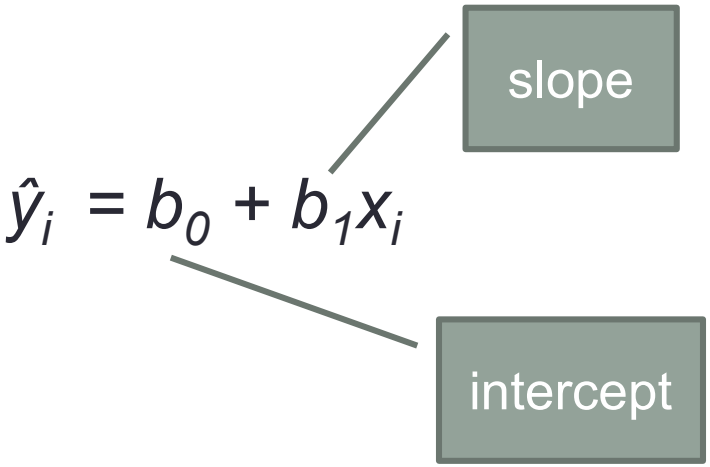

25
20
15
10
5
0
Tip Amount (Baht)
0
50
100
150
200
250
300
350
400
Meal Amount (Baht)
23
18
13
12
9

Centroid

- Average values of x and y → 227 and 15
- The best-fit regression line must pass the centroid



The Best Fit Line

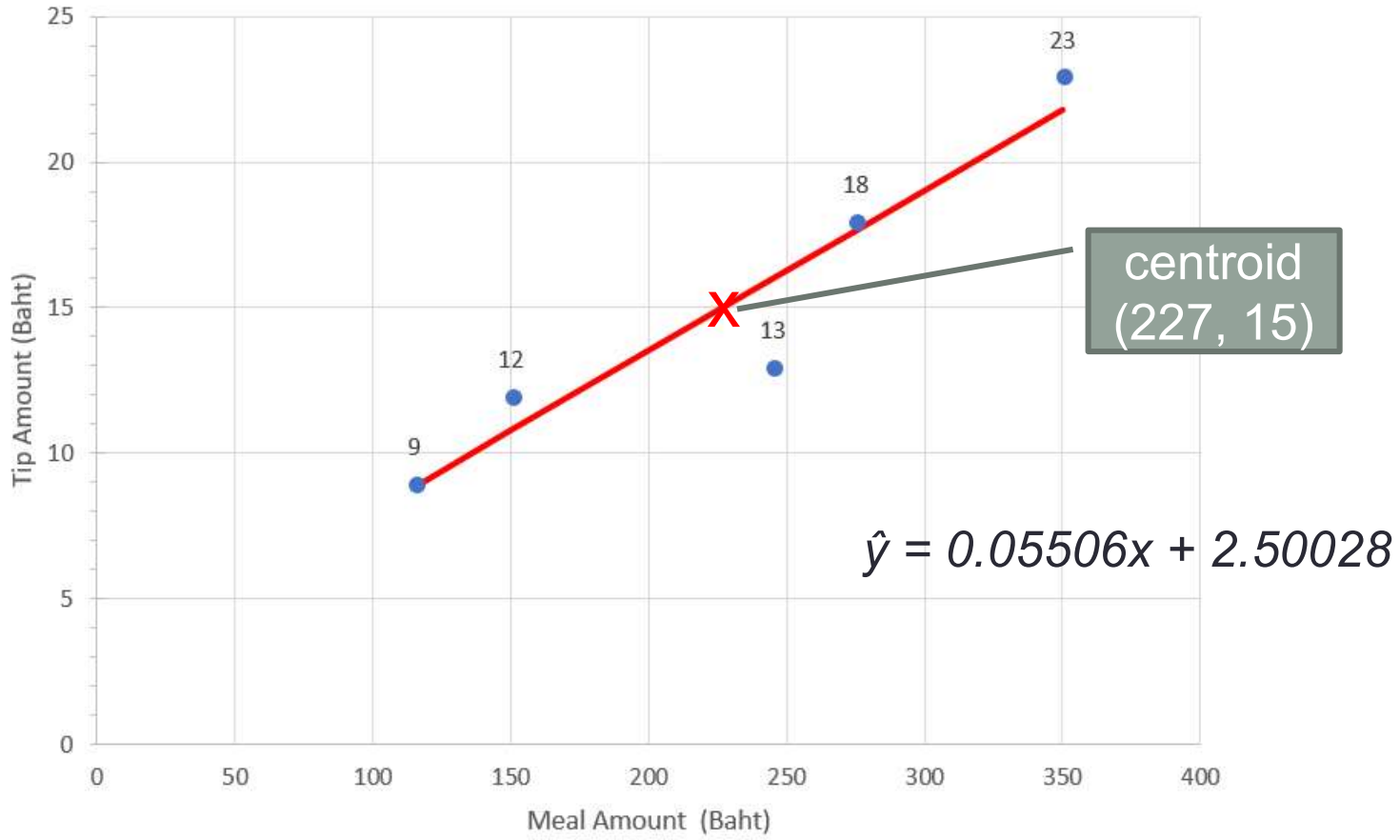


$$b_1 = \frac{\Sigma (x_i - \overline{x})(y_i - \overline{y})}{\Sigma (x_i - \overline{x})^2}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

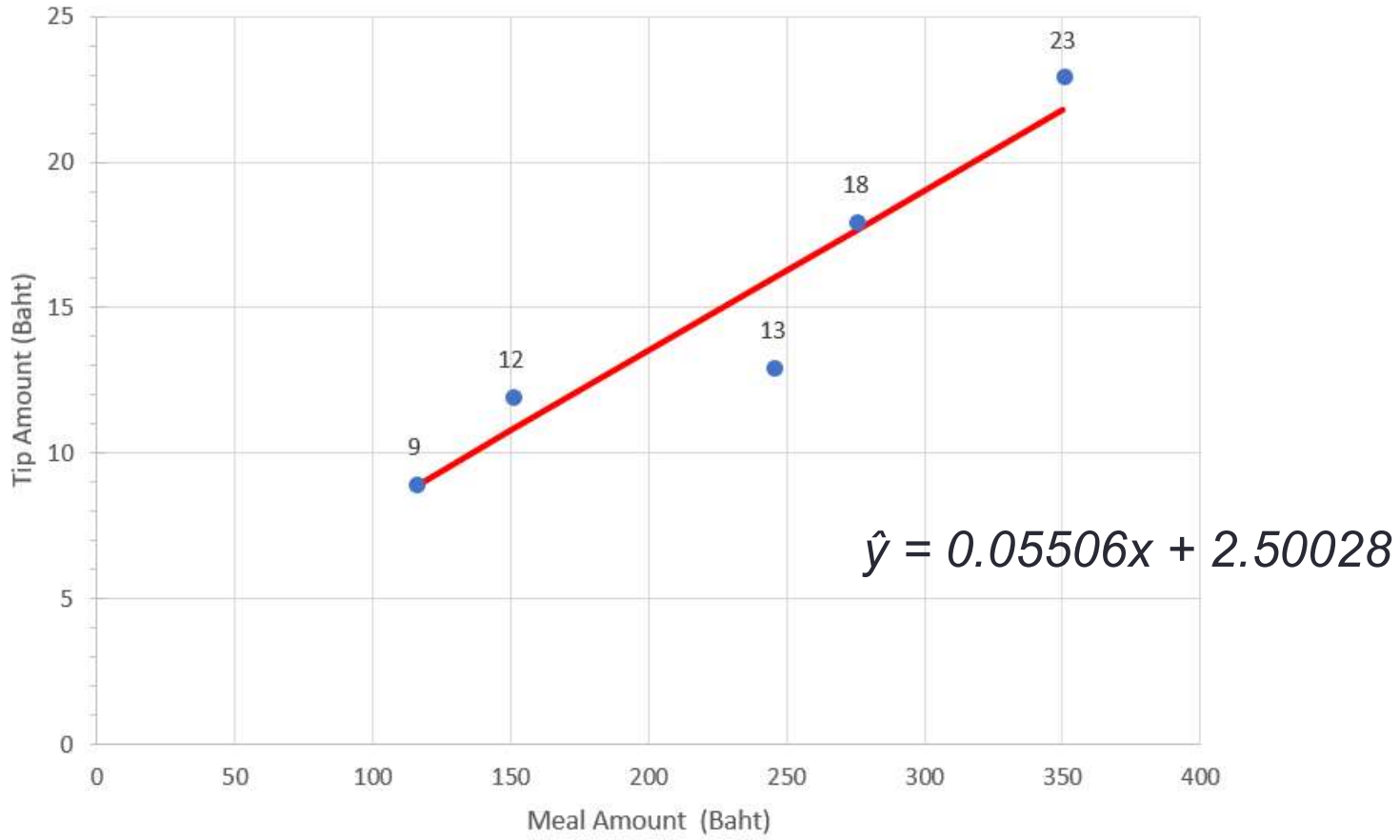
Our Example

$b_0 = 2.50028$
$b_1 = 0.05506$



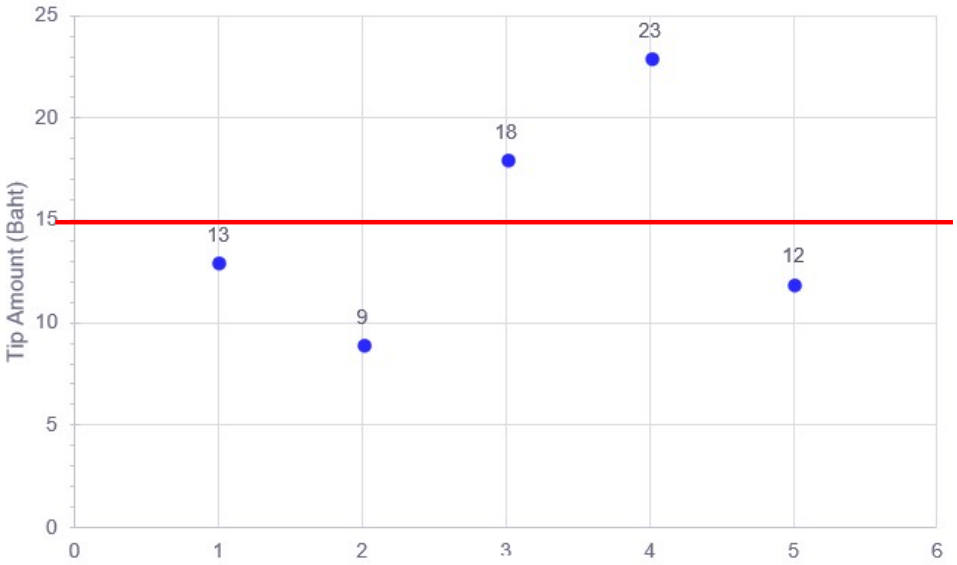
SSE

SSE = 12.1456



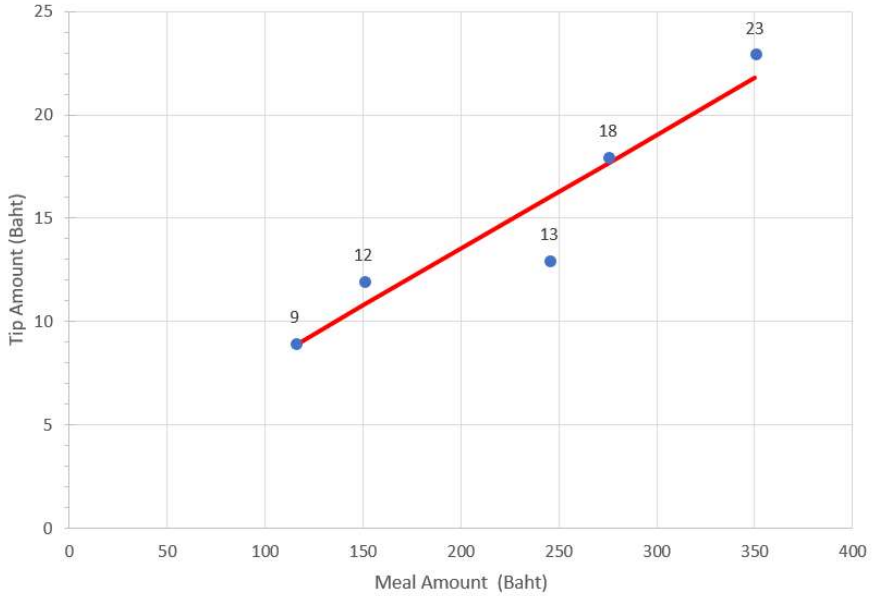
Two Models

SSE = 122 → SSE = 12.1456



One dependent variable

Average value



One dependent variable
One independent variable

Least squares method

Intercept and Coefficients

- Y-intercept is the expected mean value of Y when all X=0

- Coefficient, B_i, is the difference in the predicted value of Y for each one-unit difference in X_i, if all other X's remain constant

Common Regression Evaluation Metrics

- Mean Absolute Error (MAE) :

  the mean of the absolute value of the errors

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Squared Error (MSE) :

  mean of the squared errors

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

- Root Mean Squared Error (RMSE) :

  square root of the mean of the squared errors

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

Multiple Regression

- More than one independent variable
- One dependent variable

- Independent variables may relate to each other
- Ideal is that all independent variables to be correlated with dependent variable, not with each other

- Note
  - predictor variable is often called independent variable
  - response variable is often called dependent variable

Multiple Regression Model / Equation

Multiple regression model

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}_{\text{linear parameters}} + \overbrace{\varepsilon}^{\text{error}}$$

Estimated multiple regression equation

$$\underbrace{\hat{y}}_{\substack{\text{predicted value} \\ \text{(dependent variable)}}} = \overbrace{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}^{\substack{\text{estimated value of} \\ \beta_0, \beta_1, \beta_2, \ldots, \beta_n}}$$