Linear Regression

Problem: Suppose you are given following historical data:

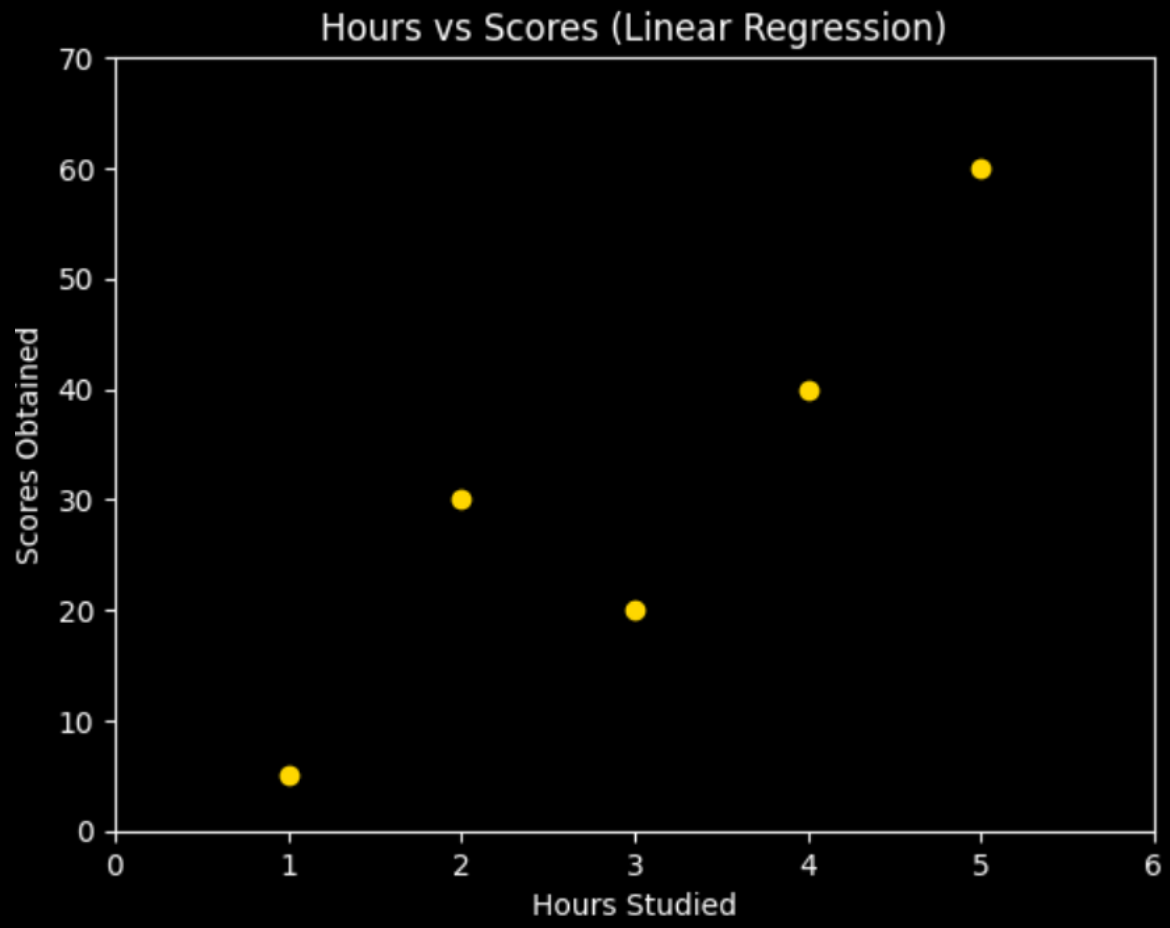
Hours Studied (x)	Score (y)
1	5
2	30
3	20
4	40
5	60

Based on above, predict the score of the student who studied for 5.5 hours.

Linear Regression:

So, you have 5 data points: number of hours student's study and their scores. Let's plot them using scatter plot.

Hours Studied (x)	Score (y)
1	5
2	30
3	20
4	40
5	60



Linear Regression:

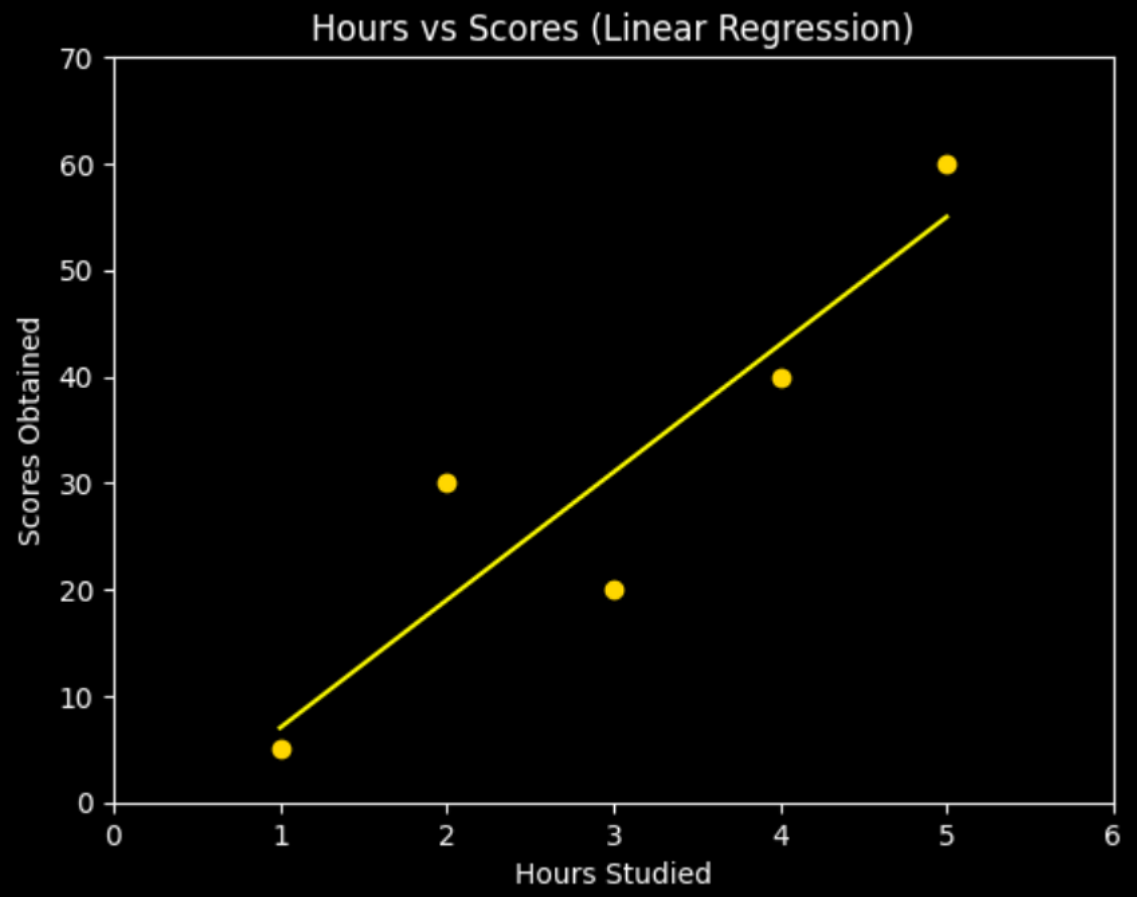
Now based on the distribution of the points, it seems like there is a straight-line model that can best fit this data. This line is called the **regression line.**

This line is defined by 2 parameters: a bias term, a_0 (also called y-intercept) and a weight, a_1 (also called slope).

In linear regression problems, we have to find this line: i.e., find a_0 and a_1:

$y = a_0 + a_1 x$

Hours Studied (x)	Score (y)
1	5
2	30
3	20
4	40
5	60



Linear Regression

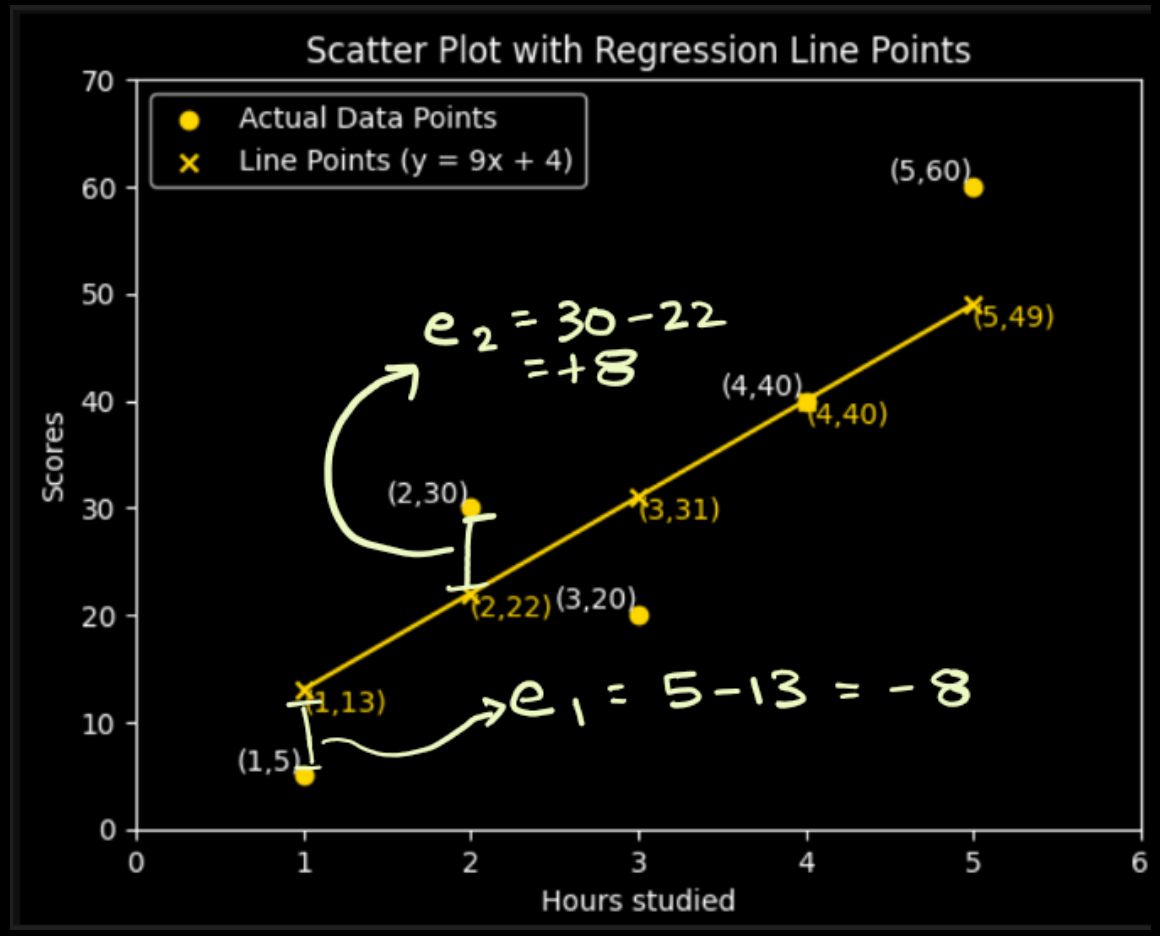
Notice that this line (say given by y = 4 + 9x) is not going through all the points: there are errors.

For each data point we can calculate the errors:

error = actual_y – predicted_y

$\Rightarrow e_i = y_i - \hat{y}_i$

Hours Studied (x)	Score (y)	Predicted ($\hat{y} = 4 + 9x$)	$e_i = y_i - \hat{y}_i$
1	5	13	e_1 = 5 – 13 = - 8
2	30	22	e_2 = 30 – 22 = + 8
3	20	31	e_3 = 20 – 31 = - 11
4	40	40	e_4 = 40 – 40 = 0
5	60	49	e_5 = 60 – 49 = + 11



These errors are also called **residuals.**

There are error above the regression line and errors below the regression line. Some of the errors would be positive and others negative.

Linear Regression

Look at an example shown in fig:
If we calculate the error taking into account the sign, we get
Total Error <u>with sign</u> = $e_1 + e_2 + e_3 + e_4 + e_5$

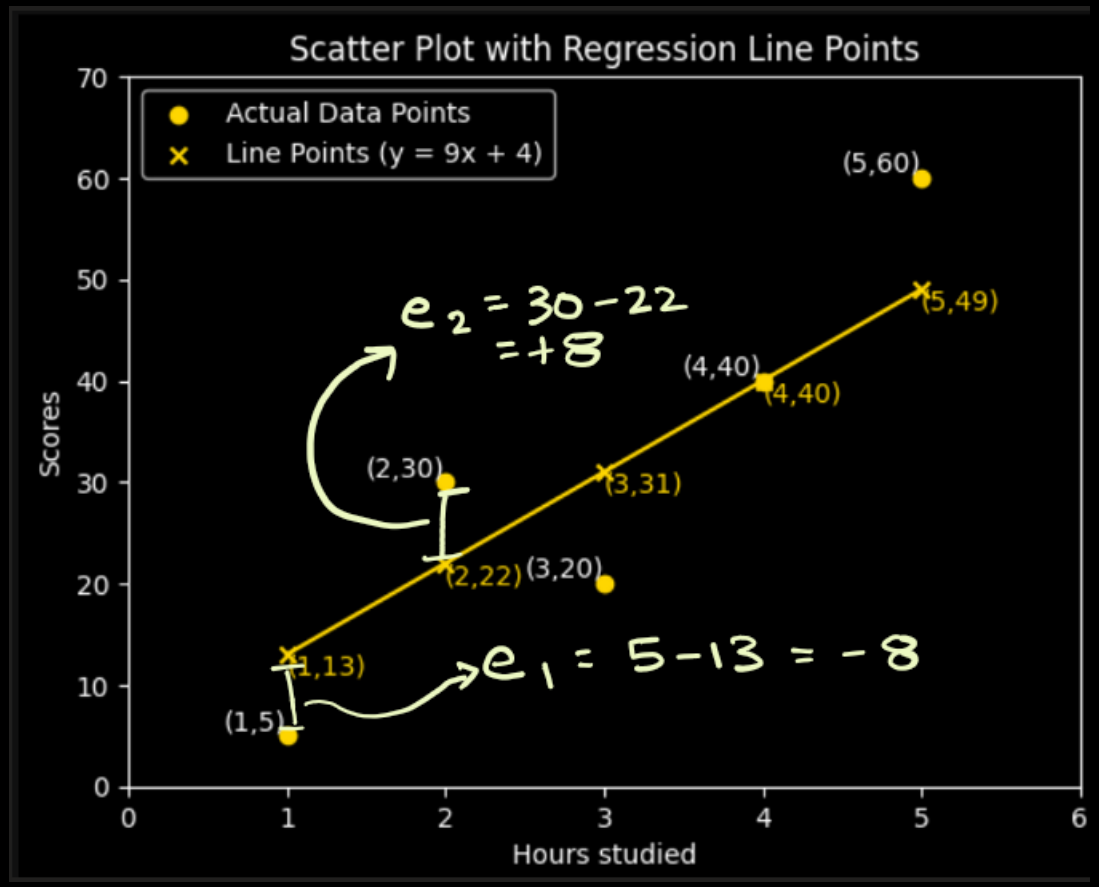
$= (5 - 13) + (30 - 22) + (20 - 31) + (40 - 40) + (60 - 49) = 0$

Now this is not true: We see that error is not 0.
The reason is coming out to 0 because of signs. So, we get rid of sign by taking square of the errors and adding them. This would give us Sum of Squared Errors (SSE).

SSE = $e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$

Another name of SSE is Residual Sum Of Squares (RSS).

For the line shown,
SSE = $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$
$= (-8)^2 + (8)^2(-11)^2 + (0)^2 + (11)^2$
$= 370$



Linear Regression

Different lines would give you different error (SSE).
For example, the line y = 5 + 10x would give us following:

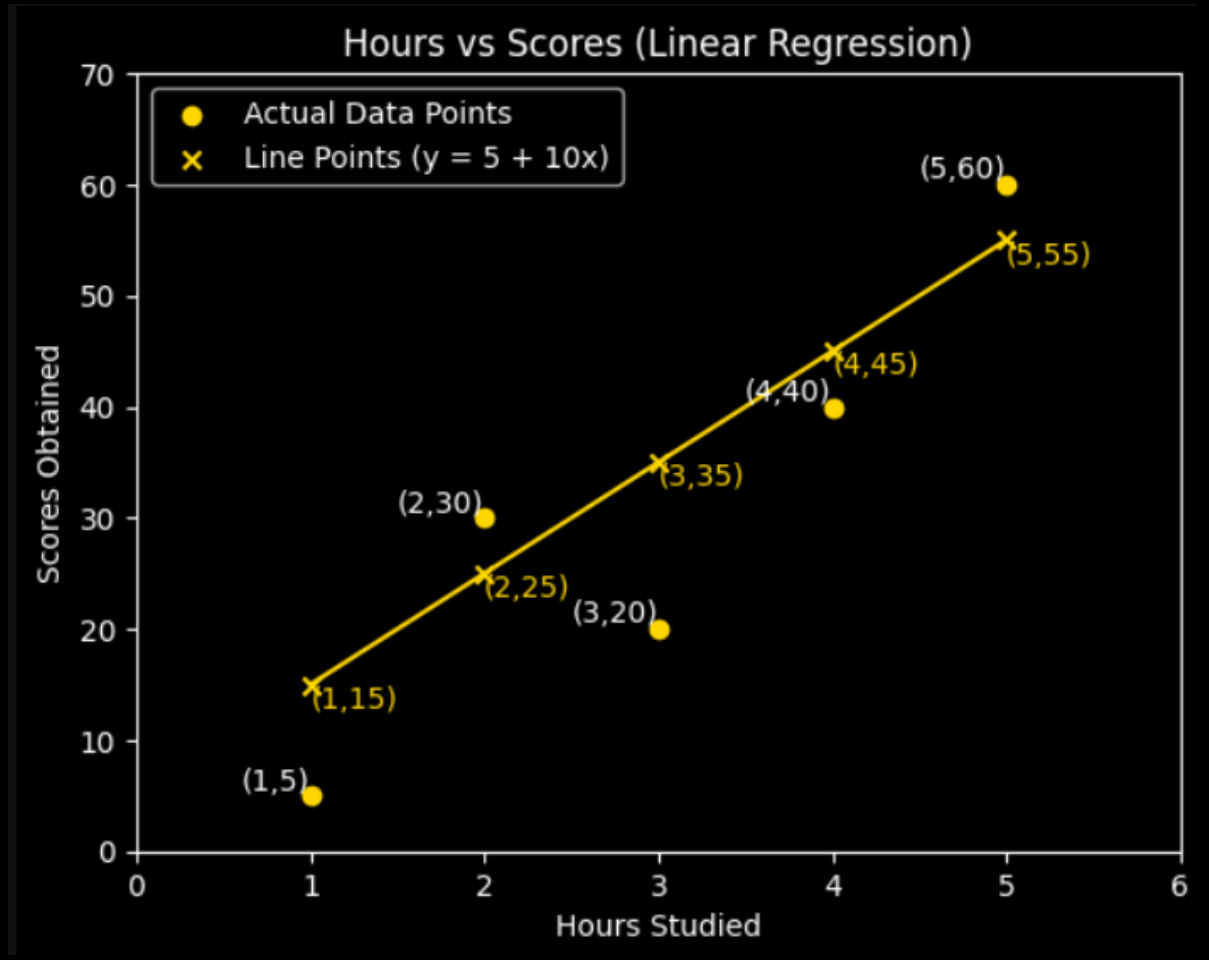
Hours Studied (x)	Score (y)	Predicted ($\hat{y} = 5 + 10x$)	$e_i = y_i - \hat{y}_i$
1	5	15	e_1 = 5 – 15 = - 10
2	30	25	e_2 = 30 – 25 = + 5
3	20	35	e_3 = 20 – 35 = - 15
4	40	45	e_4 = 40 – 45 = - 5
5	60	55	e_5 = 60 – 55 = + 5

For above line,

SSE = $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$

= $(-10)^2 + (5)^2(-15)^2 + (-5)^2 + (5)^2$

= **400**



Linear Regression

Different lines would give you different error (SSE):
For example, the line y = 1 + 10x would give us following:

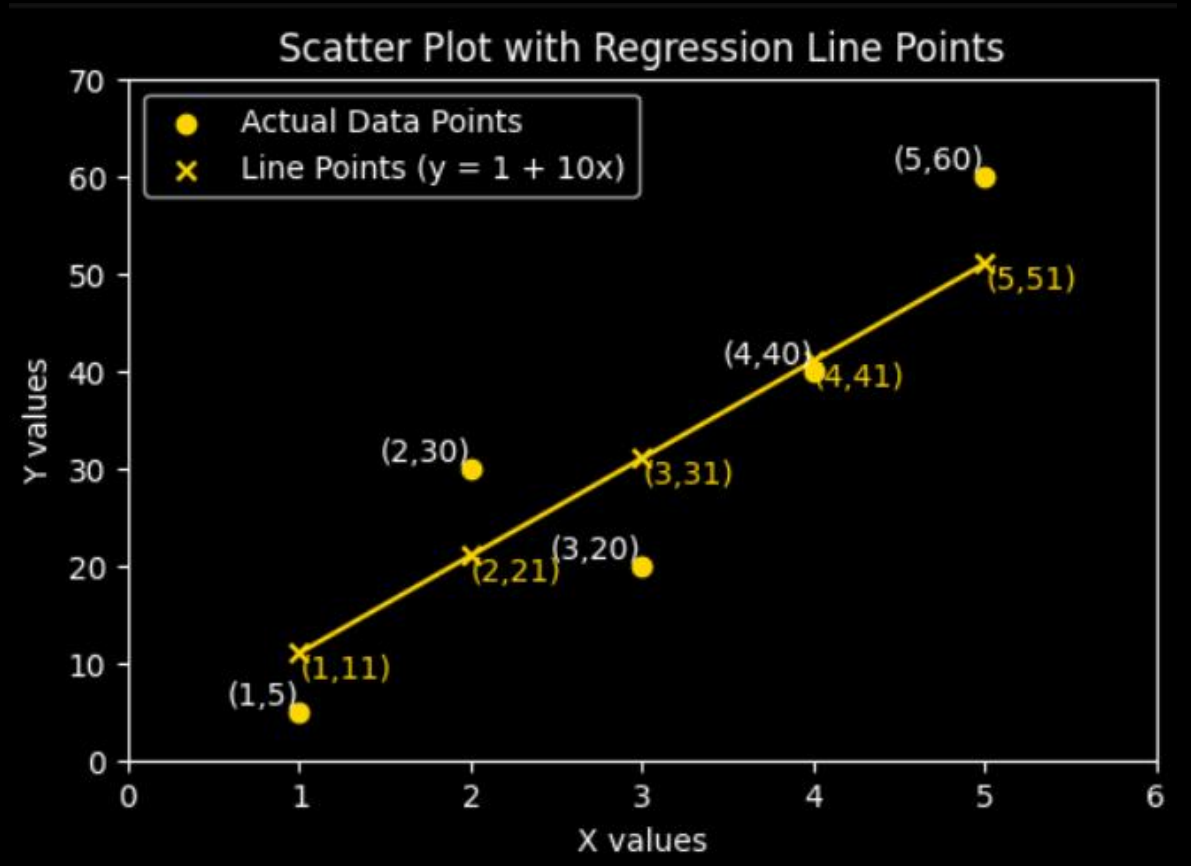
Hours Studied (x)	Score (y)	Predicted ($\hat{y} = 1 + 10x$)	$e_i = y_i - \hat{y}_i$
1	5	11	e_1 = 5 – 11 = - 6
2	30	21	e_2 = 30 – 21 = + 9
3	20	31	e_3 = 20 – 31 = - 11
4	40	41	e_4 = 40 – 41 = - 1
5	60	51	e_5 = 60 – 51 = + 9

For above line,

SSE = $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$

= $(-6)^2 + (9)^2(-11)^2 + (-1)^2 + (9)^2$

= **320**



Linear Regression

The goal in linear regression problem is to find the line ($y = a_0 + a_1 x$) such that the error, SSE, is minimum.
And in order to find line, we need to find a_0 and a_1

Using calculus, we can determine a_0 and a_1 that minimizes value of SSE:

$$SSE = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$= [y_1 - (a_0 + a_1 x_1)]^2 + [y_2 - (a_0 + a_1 x_2)]^2 + \cdots + [y_n - (a_0 + a_1 x_n)]^2$$

Solve,

$$\frac{d\,SSE}{da_0} = 0 \quad \text{and} \quad \frac{d\,SSE}{da_1} = 0$$

$$\Rightarrow \quad a_0 = \bar{y} - a_1 \bar{x}, \qquad a_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

where $\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$, $\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$

Linear Regression

Let's use this on our simple example:

Hours Studied (x)	Score (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	5	1-3 = -2	5 - 31 = -26	-2 x -26 = 52	4
2	30	2-3 = -1	30 - 31 = -1	-1 x 1 = 1	1
3	20	3-3 = 0	20 - 31 = -11	0 x -11 = 0	0
4	40	4-3 = 1	40 - 31 = 9	1 x 9 = 9	1
5	60	5-3 = 2	60 - 31 = 29	2 x 29 = 58	4
$\bar{x} = 3$	$\bar{y} = 31$			**$\Sigma(x_i - \bar{x})(y_i - \bar{y})$ = 120**	**$\Sigma(x_i - \bar{x})^2$ = 10**

Substituting above in the formula:

$$a_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{120}{10} = 12$$

$$a_0 = \bar{y} - a_1\bar{x} = 31 - 12 * 3 = -5$$

The regression line that minimizes the Sum of Squared Error is given by

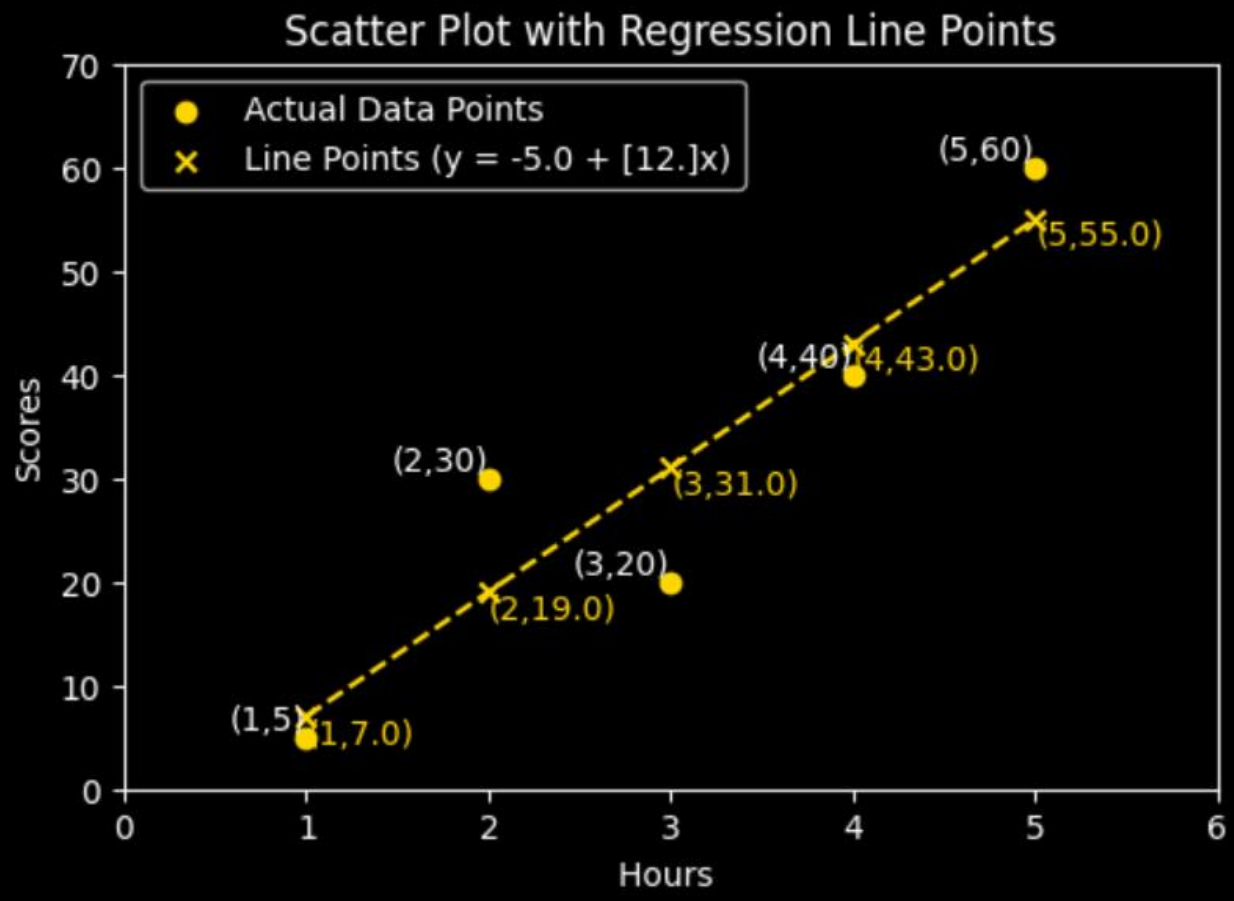
$$y = a_0 + a_1 x$$
$$y = -5 + 12x$$

Linear Regression

Now let us compute minimized SSE of above regression line:

$$y = -5 + 12x$$

Hours Studied (x)	Score (y)	Predicted ($\hat{y} = -5 + 12x$)	$e_i = y_i - \hat{y}_i$
1	5	7	e_1 = 5 – 7 = - 2
2	30	19	e_2 = 30 – 19 = + 11
3	20	31	e_3 = 20 – 31 = - 11
4	40	43	e_4 = 40 – 43= - 3
5	60	55	e_5 = 60 – 55 = + 5



For above line,

SSE = $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$

= $(-2)^2 + (11)^2(-11)^2 + (-3)^2 + (5)^2$

= **280**

This is lowest SSE one can achieve from all possible lines.

Linear Regression

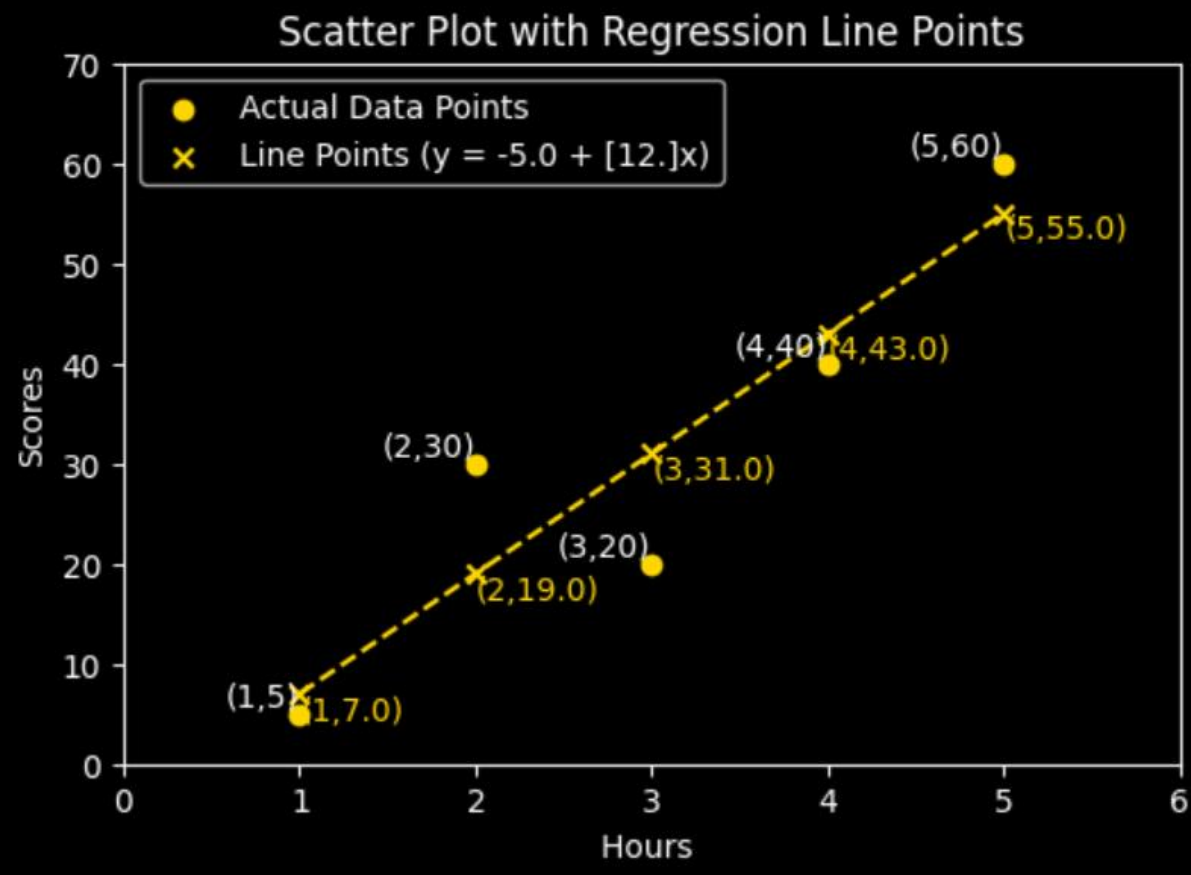
Problem: Suppose you are given following historical data:

Hours Studied (x)	Score (y)
1	5
2	30
3	20
4	40
5	60

Predict the score of the student who studied for 5.5 hours.

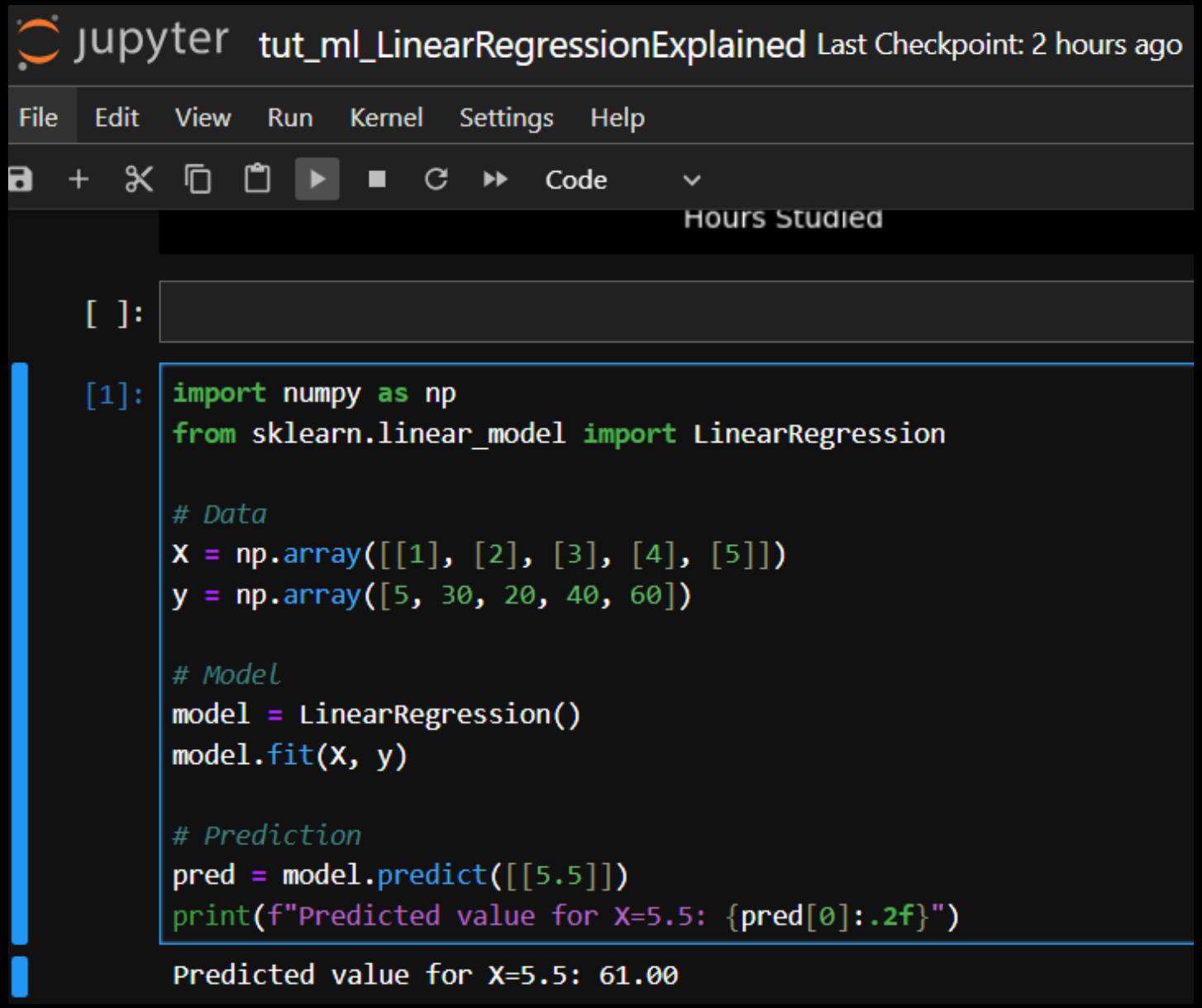
Ans:

$$y = -5 + 12x$$
$$= -5 + 12 * 5.5$$
$$= 61$$



Linear Regression: Solving using numpy and sklearn

Hours Studied (x)	Score (y)
1	5
2	30
3	20
4	40
5	60

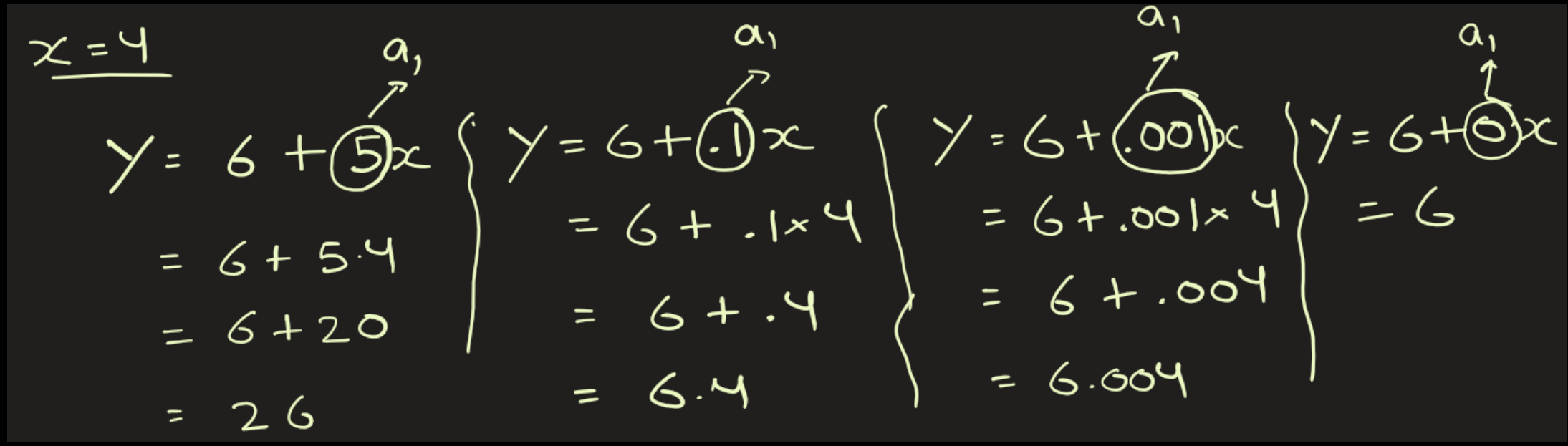


Linear Regression

Side note on the term bias, a_0 and weight a_1:

$$y = a_0 + a_1 x$$

The coefficient a_1 controls how much of x should influence the value of y. That is, how much weight should be given to x
This is shown below with an example: The effect of x diminishes as a_1 decreases from 5 -> 0.1 -> 0.001 -> 0.



Different ways of measuring errors in regression

Mean Absolute Error:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Mean Squared Error:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Root Mean Squared Error:

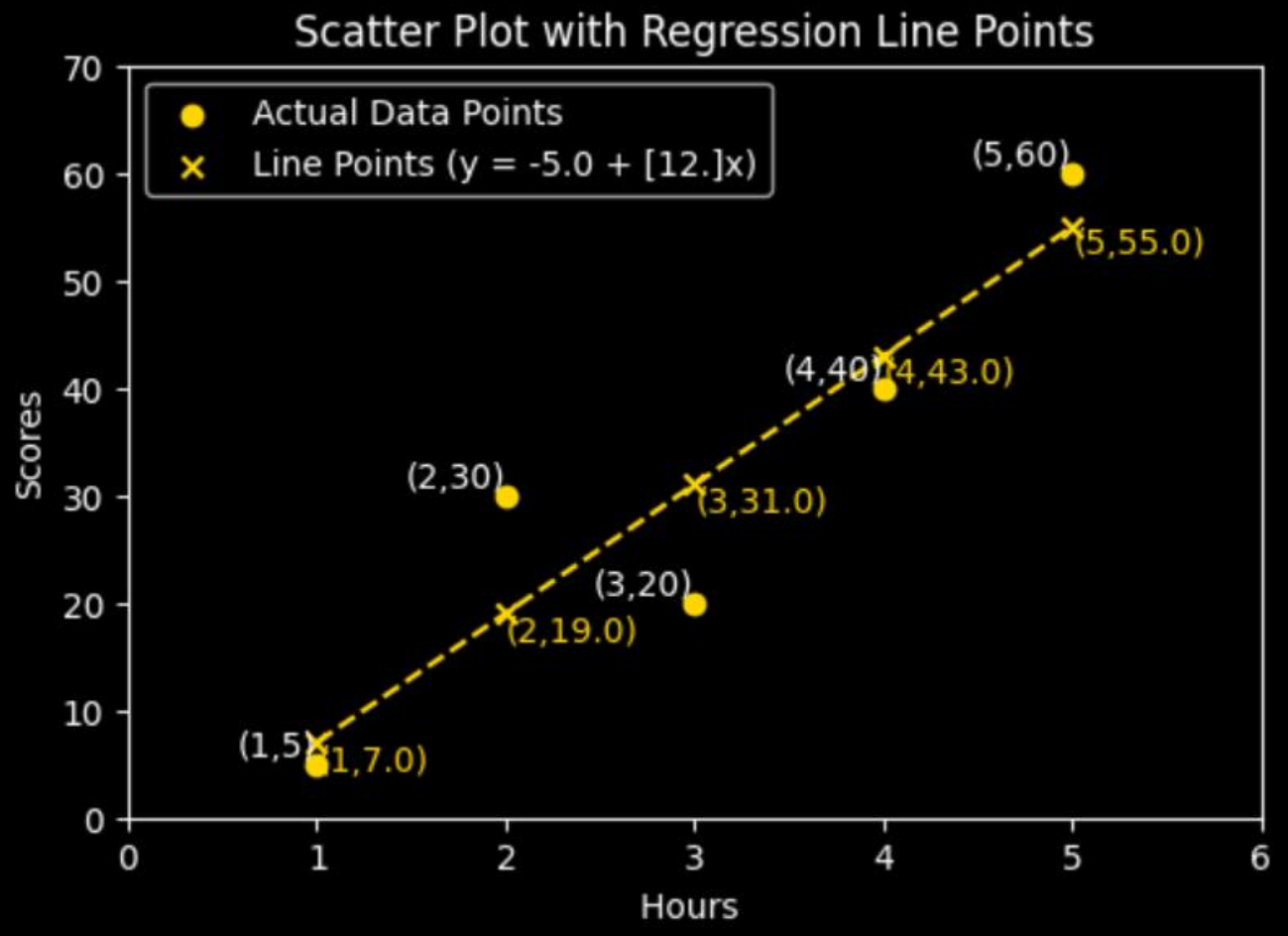
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Example:

MAE = $\frac{|5-7| + |30-19| + |20-31| + |40-43| + |60-55|}{5} = 6.2$

MSE = $\frac{[5-7]^2 + [30-19]^2 + \ldots + [60-55]^2)}{5} = 56$

RMSE = $\sqrt{MSE} = \sqrt{56}$

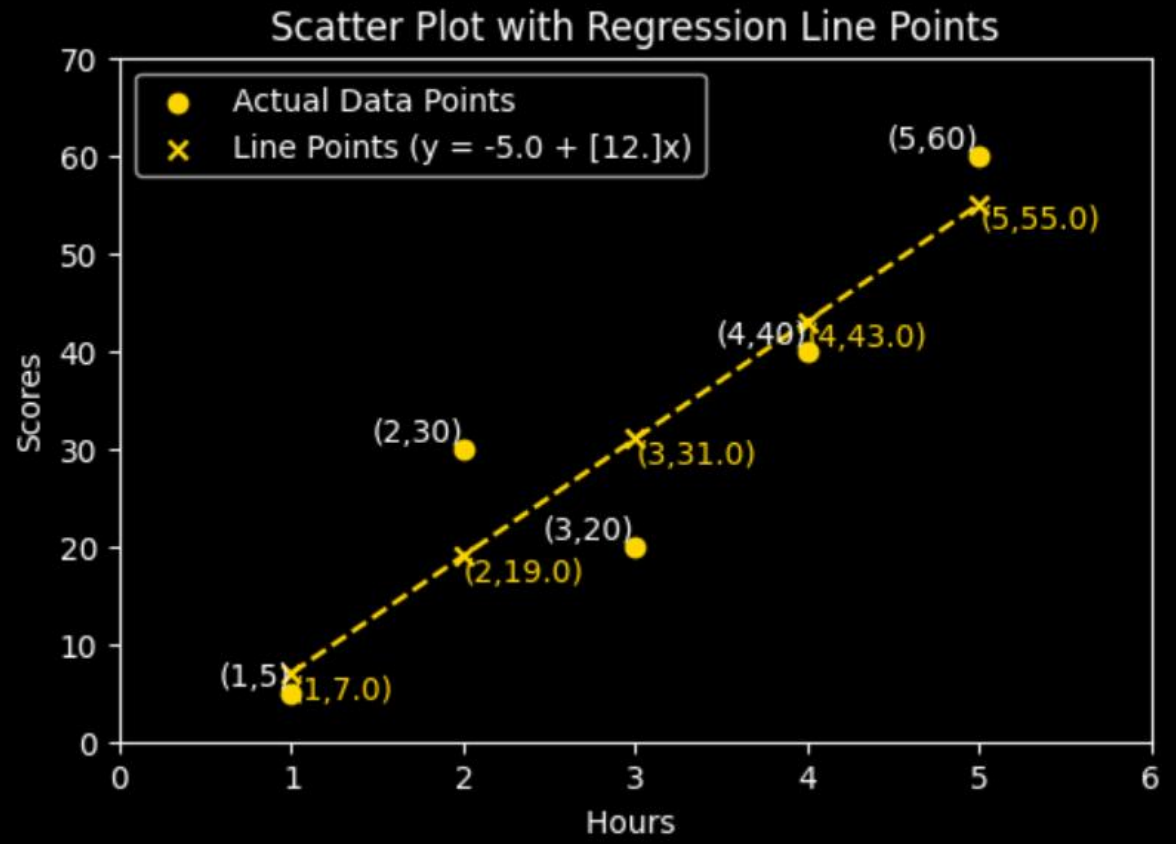


Different ways of measuring errors in regression

Coefficient of Determination (R^2 Score): R-square measures how much of the variance in Y is explained by the model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$SS_{res} = \sum(Y_i - \hat{Y}_i)^2$ → **Residual Sum of Squares**
$SS_{tot} = \sum(Y_i - \bar{Y})^2$ → **Total Sum of Squares**



Example: Using previous example,

$$\bar{Y} = \frac{5 + 30 + 20 + 40 + 60}{5} = \frac{155}{5} = 31.0$$

- Residual sum of squares (SSR or SS_{res}):

$$SS_{res} = \sum e_i^2 = 4 + 121 + 121 + 9 + 25 = 280$$

- Total sum of squares (SS_{tot}):

$$SS_{tot} = \sum(Y_i - \bar{Y})^2 = 676 + 1 + 121 + 81 + 841 = 1720$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{280}{1720} = 0.8372093$$

The model explains about **83.72%** of the variance in scores.

EXTRA


$R^2 = 1$: perfect fit

$R^2 = 0$: model predicts no better than the mean

$R^2 < 0$: model worse than just predicting the mean

Linear Regression: Derivation of the Coefficients

Goal of Simple Linear Regression

We want the best-fitting line:

$$\hat{Y}_i = a_0 + a_1 X_i$$

that **minimizes the sum of squared errors**:

$$S(a_0, a_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a_0 - a_1 X_i)^2.$$

Linear Regression: Derivation of the Coefficients

Step 1 — Objective function

$$S(a_0, a_1) = \sum (Y_i - a_0 - a_1 X_i)^2$$

Minimize S by taking partial derivatives w.r.t. a_0 and a_1 and setting them to zero.

Step 2 — Partial derivatives

With respect to a_0:

$$\frac{\partial S}{\partial a_0} = -2 \sum (Y_i - a_0 - a_1 X_i)$$

Set to zero:

$$\sum (Y_i - a_0 - a_1 X_i) = 0$$

Rearrange:

$$\sum Y_i - n a_0 - a_1 \sum X_i = 0$$

So

$$n a_0 + a_1 \sum X_i = \sum Y_i \quad ...(1)$$

Linear Regression: Derivation of the Coefficients

Step 2 — Partial derivatives

With respect to a_1:

$$\frac{\partial S}{\partial a_1} = -2\sum X_i(Y_i - a_0 - a_1X_i)$$

Set to zero:

$$\sum X_i(Y_i - a_0 - a_1X_i) = 0$$

Expand:

$$\sum X_iY_i - a_0\sum X_i - a_1\sum X_i^2 = 0$$

So

$$a_0\sum X_i + a_1\sum X_i^2 = \sum X_iY_i \quad ...(2)$$

Linear Regression: Derivation of the Coefficients

Step 3 — Solve the normal equations

We have the two linear equations:

1. $na_0 + a_1 \sum X_i = \sum Y_i$
2. $a_0 \sum X_i + a_1 \sum X_i^2 = \sum X_i Y_i$

From (1):

$$a_0 = \frac{\sum Y_i - a_1 \sum X_i}{n}$$

Substitute into (2):

$$\left(\frac{\sum Y_i - a_1 \sum X_i}{n}\right) \sum X_i + a_1 \sum X_i^2 = \sum X_i Y_i$$

Multiply by n:

$$(\sum Y_i)(\sum X_i) - a_1(\sum X_i)^2 + na_1 \sum X_i^2 = n \sum X_i Y_i$$

Collect a_1 terms:

$$a_1 \left[n \sum X_i^2 - (\sum X_i)^2\right] = n \sum X_i Y_i - (\sum X_i)(\sum Y_i)$$

Therefore the slope a_1 is:

$$\boxed{a_1 = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2}}$$

Linear Regression: Derivation of the Coefficients

Step 4 — Intercept a_0

Plug a_1 back into the expression from (1):

$$a_0 = \frac{\sum Y_i - a_1 \sum X_i}{n}$$

or equivalently, using means $\bar{X}, \bar{Y}$,

$$a_0 = \bar{Y} - a_1 \bar{X}$$

Alternative (mean-deviation) form — equivalent and often clearer

$$a_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$a_0 = \bar{Y} - a_1 \bar{X}$$

Linear Regression assumption:

- **Linear relationship**: The first important assumption of linear regression is that the dependent and independent variables **should be linearly related**. The relationship can be determined with the help of scatter plots that help in visualization. Also, one needs to check for outliers as linear regression is sensitive to them.
- **Normal distribution of residuals**: The second assumption relates to the normal distribution of residuals or error terms, i.e., if residuals are non-normally distributed, the model-based estimation may become too wide or narrow. The non-normal distribution also underscores that you need to closely observe some unusual data points to make a good model.
- **Multicollinearity**: The third assumption relates to multicollinearity, where several independent variables in a model are highly correlated. More correlated variables make it difficult to determine which variable contributes to predicting the target variable. Also, standard errors inevitably increase due to correlated variables. Moreover, with such a robust variable correlation, the predicted regression coefficient of a correlated variable further depends on the other variables available in the model, leading to wrong conclusions and poor performance. The goal, therefore, is to have minimal or lesser multicollinearity.
- **Autocorrelation**: One fundamental assumption of linear regression specifies that the given dataset should not be autocorrelated. This mostly happens when residuals or error terms are not independent of each other. In other words, the situation arises when the value of f(a+1) is not independent of the value of f(a). For example, in the case of stock prices, the price of one stock depends on the cost of the previous one.

Matrix math of Linear regression(>1 Dimension)

In linear regression:

$$\hat{y}_i = b_0 + b_1 x_i$$

We minimize **Mean Squared Error (MSE)**:

$$J(b_0, b_1) = \sum_i (y_i - \hat{y}_i)^2$$

If we take partial derivatives and set them to zero:

$$\frac{\partial J}{\partial b_0} = 0, \quad \frac{\partial J}{\partial b_1} = 0$$

we get **simple linear equations** in b_0, b_1.

Solving them gives an exact formula:

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (X^T X)^{-1} X^T y$$

This is a **closed-form solution** — meaning we can solve it directly with algebra.