Machine Learning: Single Linear Regression

Notes by Mannan Ul Haq (BDS-3C)

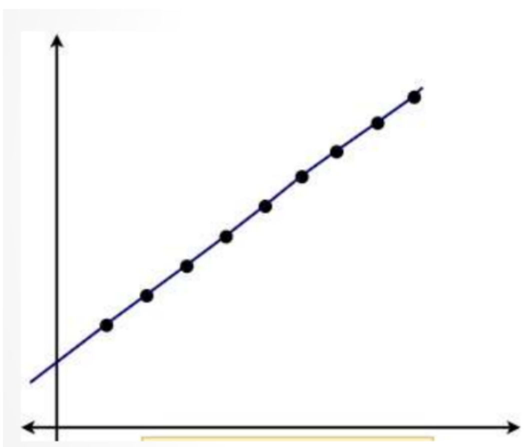
Linear regression has a wide range of applications beyond machine learning, including:

- **Mathematics**: It's a fundamental concept in mathematical modeling.
- **Statistical Modeling**: Used for analyzing relationships between variables.
- **Economics**: For modeling economic relationships.
- **Forecasting**: Predicting trends in various fields.

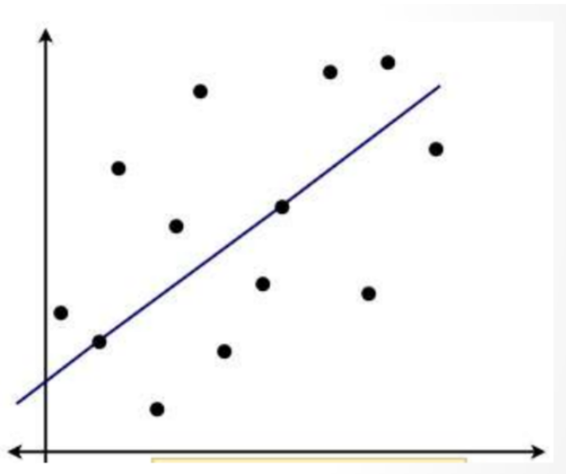
Single linear regression is a simple and commonly used model that explores the relationship between two continuous quantitative variables. It is used to predict or explain the behavior of a dependent variable (Y) based on a single independent variable (X).

Types of Relationships:

1. **Deterministic Relationship**: This implies that there is a perfect, predictable relationship between the independent and dependent variables. Changes in the independent variable(s) directly and completely determine changes in the dependent variable.



1. **Random Relationship**: In some cases, there may be no relationship at all between the variables. The dependent variable may change randomly, regardless of the values of the independent variable.
2. **Statistical Relationship**: Most real-world relationships fall into this category. In a statistical relationship, there's a pattern or correlation between the variables, but it's not perfectly deterministic. The relationship is influenced by other factors, represented by the error term in the linear regression equation.



Example:

Linear regression helps us quantify and understand these relationships, allowing us to make predictions, draw inferences, and model real-world phenomena in a wide range of fields.

Scenario: You have a dataset containing information about houses, including their areas (in square feet) and their selling prices.

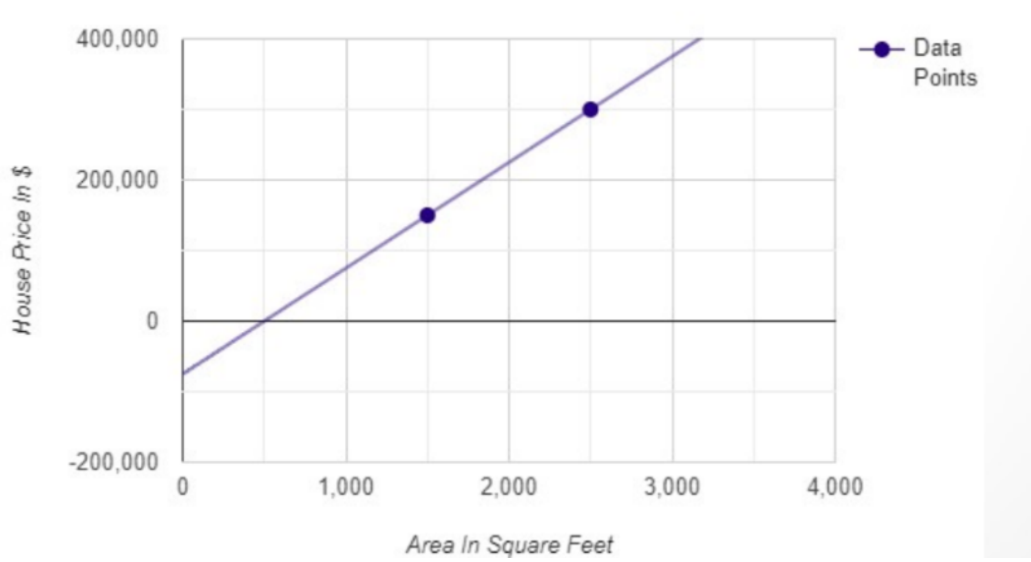
Objective: You want to visualize the relationship between the area of houses and their prices using a scatter plot to understand whether there's a connection between these two variables.

Scatter Plot:

In the scatter plot below, the x-axis represents the area of houses, and the y-axis represents the selling prices. Each point on the plot corresponds to a specific house in your dataset.

```
  Area                 |    Price
-----------------------------------
  1500                 |    150,000
  1800                 |    ???????
  2500                 |    300,000
```

Analysis:



Looking at the scatter plot, you can make a few observations:

1. There's a general trend where, as the area of a house increases, its price tends to increase as well.
2. Based on this scatter plot, it appears that there's a positive relationship between house area and price, suggesting that these two variables are related.
3. By using Linear Regression, we can predict the value of 1800. For this we have to understand this concept.

Let’s start with some basics:

Basics of Coordinate Geometry:

1. The convention in coordinate geometry is to read and name points from left to right.

2. Points on a coordinate plane are typically named (x1, y1), (x2, y2), and so on based on their order from left to right.
3. Horizontal lines have a slope of 0, meaning they don't change in the vertical direction.
4. Vertical lines have an "infinite" slope because they are perfectly vertical.
5. The slope of a line can be determined by comparing the Y-coordinates of two points; a positive slope means the line goes up as you move from left to right, and a negative slope means it goes down.
6. Points at the same vertical distance from the X-axis have the same Y-coordinate.
7. Points at the same vertical distance from the Y-axis have the same X-coordinate.

Single Linear Regression:

- The equation of a line in a coordinate plane is often represented as **Y = mX + b**, where **Y** is the dependent variable, **X** is the independent variable, m is the slope, and b is the intercept.
- You can determine the slope **(m)** of a line given two points on it.

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

$Where,$

$m \quad = slope\ of\ the\ line$

$(X_1, Y_1) = \ X\ and\ Y\ coordinates\ of\ the\ first\ data\ point$

$(X_2, Y_2) = \ X\ and\ Y\ coordinates\ of\ the\ second\ data\ point$

From the rules mentioned above, we can infer that in our graph:

(X1 , Y1) = (1500 , 150000)
(X2 , Y2) = (2500 , 300000)

Next, we can easily find the slope of the two points.

$$m = \frac{300000 - 150000}{2500 - 1500}$$

$$m = \frac{150000}{1000}$$

$$m = 150$$

Taking our example into consideration, in our equation, Y represents the house's price, and X represents the area of the house.

- The intercept **(b)** can be calculated once you have the slope and the coordinates of one point.

Now since we have all the other values, we can calculate the value of intercept b.

$$Y = mX + b$$
$$b = Y - mX$$
$$(X, Y) = (1500{,}150000)$$
$$b = 150000 - 150 * 1500$$
$$b = -75000$$

$$Y = mX + b$$
$$b = Y - mX$$
$$(X, Y) = (2500{,}300000)$$
$$b = 300000 - 150 * 2500$$
$$b = -75000$$

10/16/202

- The equation of a line can be used to make predictions. In our example, it's used to estimate house prices based on their areas.

Next, since we have all our parameters, we can write the equation of line as:

$$Y = mX + b$$
$$Y = 150X - 75000$$

To find the price of Squidward's house, we need to plug-in X=1800 in the above equation.

$$Y = mX + b$$
$$Y = 150X - 75000$$
$$Y = 150 * 1800 - 75000$$
$$Y = 195000$$

- In the real world, linear regression with only two data points may not provide accurate predictions. To generalize and improve prediction performance, large datasets and more advanced regression techniques are typically used.

The concept of a **"best fit"** or regression line is essential when dealing with real-world datasets that contain numerous data points. It aims to find a line that provides the best overall fit to the data, considering all the variability and noise present in the real world. This best-fit line is used to make predictions and understand the underlying relationships between variables.

Calculating the Linear Best Fit

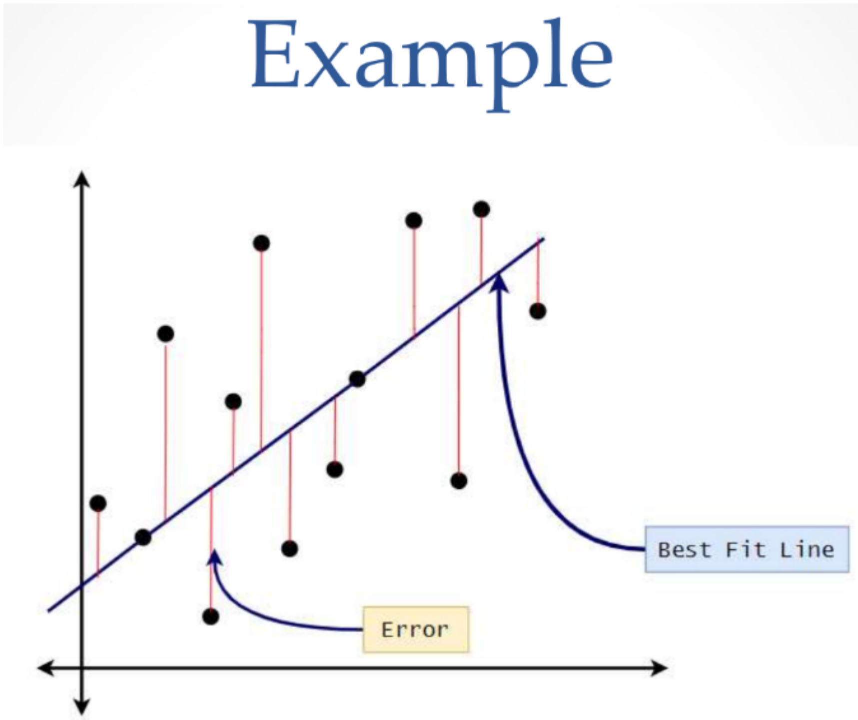
The Problem:

- We want to find a straight line that best fits our data points.
- But, it's impossible to have one line pass through all the points perfectly.

Minimizing Error:

- So, we aim to minimize the prediction errors.
- We look at each data point and see how much our line's prediction is off from the actual value.

- Our goal is to find the line with the smallest errors - that's called the **"linear best fit"**.



Example:

- Let's say the actual price of a house with an area of 1800 sq. ft. is $220,000.
- We predict the price using our line: Y = 150X - 75,000, which gives us $195,000.
- This prediction has an error.
- To measure this error, we use a technique called the **Sum of Squared Error**.
- We randomly choose line parameters, calculate the error, and adjust the parameters. We repeat this until we find the line with the least error.
- This process is part of the gradient descent algorithm, a way to find the best-fitting line.

Positive and Negative Errors:

- Errors can be positive or negative.
- For example, if the predicted price is lower than the actual price, it's a positive error.
- If the predicted price is higher than the actual price, it's a negative error.

- To work with negative errors, we square them (make them positive).

$$Errors = \sum_{i=1}^{i=N} \left(Y_{i\,actual} - Y_{i\,predicted}\right)^2$$

The squaring is necessary to remove any negative signs.

Mean Squared Error:

- MSE measures the average of the squared differences between the predicted values (from a model) and the actual values (from the data). It quantifies how "wrong" the model's predictions are on average. A lower MSE indicates a better fit of the model to the data.

MSE formula = (1/n) * Σ(actual – prediction)^2

Example Problem: Find the MSE for the following set of values: (43,41), (44,45), (45,49), (46,47), (47,44).

when, regression line **y = 9.2 + 0.8x**

Step 1: Find the new **Y’** values:

- 9.2 + 0.8(43) = 43.6
- 9.2 + 0.8(44) = 44.4
- 9.2 + 0.8(45) = 45.2
- 9.2 + 0.8(46) = 46
- 9.2 + 0.8(47) = 46.8

Step 2: Find the error **(Y – Y’)**:

- 41 – 43.6 = -2.6
- 45 – 44.4 = 0.6
- 49 – 45.2 = 3.8
- 47 – 46 = 1
- 44 – 46.8 = -2.8

Step 3: Square the Errors:

- 2.6^2 = 6.76
- 0.6^2 = 0.36
- 3.8^2 = 14.44
- 1^2 = 1
- 2.8^2 = 7.84

Height (X)	Weight (Y)	Estimated (Y')	Error (Y-Y')	Error Squared
43	41	43.6	-2.6	6.76
44	45	44.4	0.6	0.36
45	49	45.2	3.8	14.44
46	47	46	1	1
47	44	46.8	-2.8	7.84
Regression line = y=9.2+0.8x				

Step 4: Add all of the squared errors up: 6.76 + 0.36 + 14.44 + 1 + 7.84 = 30.4.

Step 5: Find the mean squared error:

30.4 / 5 = 6.08.

Derivation of the Simple Linear Regression Formula

Issue in the Used Approach:

- Finding the best-fit line for a specific dataset can be cumbersome, especially for larger datasets.

Proposed Solution:

- Use a formula to calculate the required parameter values (intercept and slope) for the best-fit line.

Derivation:

1. You have a total of **n** data points (X, Y), ranging from **i=1** to **i=n**.

$$(X_i, Y_i)$$

2. Define the linear best fit as:

$$\hat{Y}_i = a + BX_i$$

3. Write the error function, often represented as S, which measures the difference between actual and predicted values.

$$S = \sum_{i=1}^{i=n} \left(Y_i - \hat{Y}_i\right)^2$$

4. Substitute the equation for the best fit (from step 2) into the error function:

$$S = \sum_{i=1}^{i=n} (Y_i - a - BX_i)^2$$

Minimizing Error (S):

To minimize S (the error function), you find where the first derivative of S with respect to a and b is equal to 0. The closer a and b are to 0, the less total error for each point is.

Finding Intercept (a):

1. Find the partial derivative of S with respect to a:

$$\frac{\partial S}{\partial a} = \frac{\partial}{\partial a}\left[\sum_{i=1}^{i=n} (Y_i - a - BX_i)^2\right]$$

2. Use the chain rule of partial derivatives:

$$Y_i - a - BX_i = u$$

$$S = \sum_{i=1}^{i=n} u^2$$

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial u} * \frac{\partial u}{\partial a}$$

3. Find partial derivatives:

$$\frac{\partial S}{\partial u} = \frac{\partial}{\partial u}\left[\sum_{i=1}^{i=n} u^2\right] = 2 * \sum_{i=1}^{i=n} u$$

$$\frac{\partial u}{\partial a} = \frac{\partial}{\partial u}(Y_i - a - BX_i) = 0 - 1 - 0 = \overset{23}{-}1$$

$$\frac{\partial S}{\partial a} = -2 * \sum_{i=1}^{i=n} u = -2 * \sum_{i=1}^{i=n}(Y_i - a - BX_i)$$

4. To find the extreme values, we take the derivative = 0:

$$-2 * \sum_{i=1}^{i=n}(Y_i - a - BX_i) = 0$$

5. Simplify the equation and find the value of a:

$$\sum_{i=1}^{i=n}(Y_i - a - BX_i) = 0$$

$$\sum_{i=1}^{i=n} a = (a + a + a \ldots + a)_{n\,times} = n * a$$

$$\sum_{i=1}^{i=n} Y_i - na - B \sum_{i=1}^{i=n} X_i = 0$$

6. Further simplifying:

$$a = \frac{\sum_{i=1}^{i=n} Y_i - B \sum_{i=1}^{i=n} X_i}{n}$$

$$\frac{\sum_{i=0}^{i=n} Y_i}{n} = \bar{Y}$$

$$\frac{\sum_{i=0}^{i=n} X_i}{n} = \bar{X}$$

$$a = \bar{Y} - B\bar{X}$$

Finding Slope (B):

1. Find the partial derivative of S with respect to B:

$$\frac{\partial S}{\partial B} = \frac{\partial}{\partial B}\left[\sum_{i=1}^{i=n} (Y_i - a - BX_i)^2\right]$$

2. Use the chain rule of partial derivatives:

$$Y_i - a - BX_i = u$$

$$S = \sum_{i=1}^{i=n} u^2$$

$$\frac{\partial S}{\partial B} = \frac{\partial S}{\partial u} * \frac{\partial u}{\partial B}$$

3. Find partial derivatives, simplify, and use the derivative = 0 to find extreme values:

$$\frac{\partial S}{\partial u} = \frac{\partial}{\partial u}\left[\sum_{i=1}^{i=n} u^2\right] = 2 * \sum_{i=1}^{i=n} u$$

$$\frac{\partial u}{\partial B} = \frac{\partial}{\partial B}(Y_i - a - BX_i) = 0 - 0 - X_i = -X_i$$

$$\frac{\partial S}{\partial B} = -2\sum_{i=1}^{i=n} X_i * u = -2\sum_{i=1}^{i=n} X_i * (Y_i - a - BX_i)$$

$$\frac{\partial S}{\partial B} = -2\sum_{i=1}^{i=n}(X_iY_i - X_ia - BX_iX_i)$$

$$\sum_{i=1}^{i=n}(X_iY_i - X_ia - BX_i^2) = 0$$

4. Substitute the value of a into the equation:

$$a = \bar{Y} - B\bar{X}$$

$$\sum_{i=1}^{i=n}(X_iY_i - X_i(\bar{Y} - B\bar{X}) - BX_i^2) = 0$$

5. Simplify the equation and find the value of B:

$$\sum_{i=1}^{i=n}\left(X_iY_i - \bar{Y}X_i + BX_i\bar{X} - BX_i^2\right) = 0$$

$$\sum_{i=1}^{i=n}(X_iY_i - \bar{Y}X_i) + \sum_{i=1}^{i=n}(BX_i\bar{X} - BX_i^2) = 0$$

$$\sum_{i=1}^{i=n}(X_iY_i - \bar{Y}X_i) - B\sum_{i=1}^{i=n}(X_i^2 - X_i\bar{X}) = 0$$

$$B = \frac{\sum_{i=1}^{i=n}(X_iY_i - \bar{Y}X_i)}{\sum_{i=1}^{i=n}(X_i^2 - X_i\bar{X})}$$

Generalized Form for B:

$$B = \frac{\sum(XY - \frac{\sum Y}{n} X)}{\sum(X^2 - \frac{\sum X}{n} X)}$$

$$B = \frac{n \sum XY - \sum Y \sum X}{n \sum X^2 - \sum X \sum X}$$

$$B = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

10/16/2023

Generalized Form for a:

$$a = \bar{Y} - B\bar{X}$$

$$a = \bar{Y} - \left(\frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}\right)\bar{X}$$

$$a = \frac{\sum Y}{n} - \left(\frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}\right)\frac{\sum X}{n}$$

$$a = \frac{1}{n} * \frac{[\sum Y * (n\sum X^2 - (\sum X)^2) - (n\sum XY - \sum X \sum Y)\sum X]}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{1}{n} * \frac{n\sum X^2 \sum Y - (\sum X)^2 \sum Y - n\sum XY \sum X + \sum X \sum Y \sum X}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{1}{n} * \frac{n\sum X^2 \sum Y - (\sum X)^2 \sum Y - n\sum XY \sum X + (\sum X)^2 \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{1}{n} * \frac{n(\sum X^2 \sum Y - \sum XY \sum X)}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum X^2 * \sum Y - \sum X * \sum XY}{n\sum X^2 - (\sum X)^2}$$

10/16/2023 32

This process results in a formula that can be used to find the best-fit line for simple linear regression, given a dataset.

Example:

Question: Find linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10

Solution:

Construct the following table:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
$\sum x$ = 20	$\sum y$ = 25	$\sum x^2$ = 120	$\sum xy$ = 144

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

$b = \frac{4 \times 144 - 20 \times 25}{4 \times 120 - 400}$

b = 0.95

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

$a = \frac{25 \times 120 - 20 \times 144}{4(120) - 400}$

a = 1.5

Linear regression is given by:

y = a + bx

y = 1.5 + 0.95 x

Coefficient of Determination (r^2):

The coefficient of determination, often denoted as r^2, is a crucial statistic in regression analysis. It provides insight into how well a regression model fits the data and quantifies the proportion of the total variability in the dependent variable (Y) that can be explained by the independent variable(s) (X) in the model.

Formula for r^2:

r^2 is calculated by dividing the Sum of Squares Regression (SSR) by the Sum of Squares Total (SST):

r^2 = SSR / SST

Here's a more detailed explanation:

Sum of Squares Regression (SSR):

SSR quantifies the variation in Y that can be explained by the regression model. It measures how far the estimated regression line (ŷ) deviates from the horizontal "no relationship line", represented by the sample mean (ȳ). It is also called explained variation.

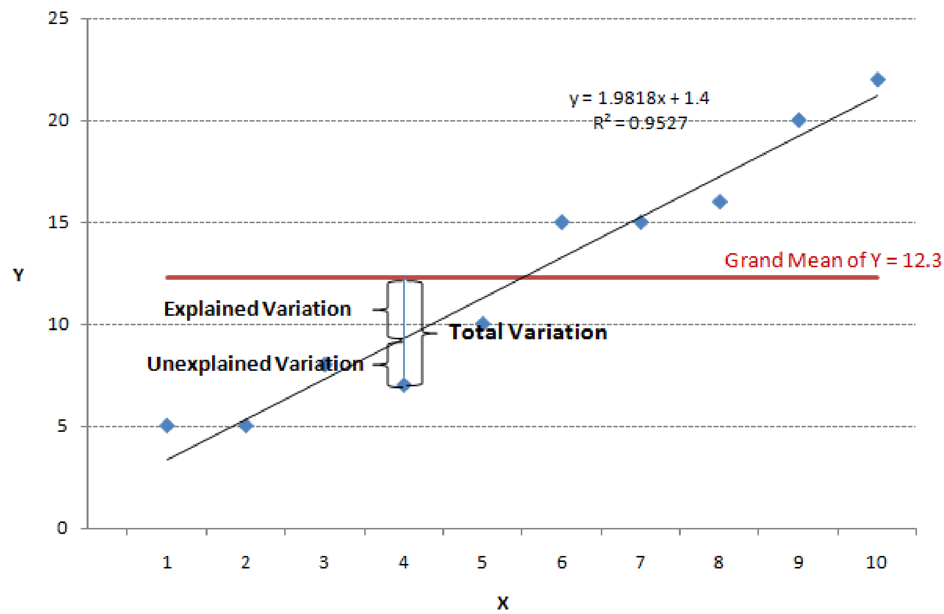
SSR = $\Sigma(\hat{y} - \bar{y})^2$

Sum of Squares Total (SST):

SST measures the total variation in the dependent variable (Y) without considering any regression model. It represents the variability in Y that is not explained by the model.

SST = SSR + SSE

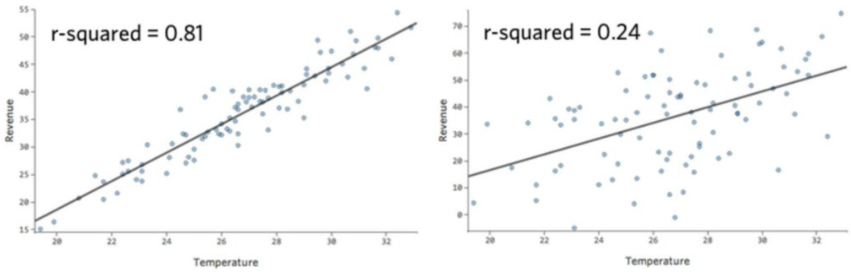
Here, SSE = Sum of Squares Error (Unexplained Variation).



Interpretation of r²:

- **0 ≤ r² ≤ 1**
- The higher the value of r², the better the fit of the regression to the data set
- Values of r² near one denote an extremely good fit of the regression to the data
- While values near zero denote an extremely poor fit.

Coefficient of Determination Example



- How “good” is regression model? Roughly:

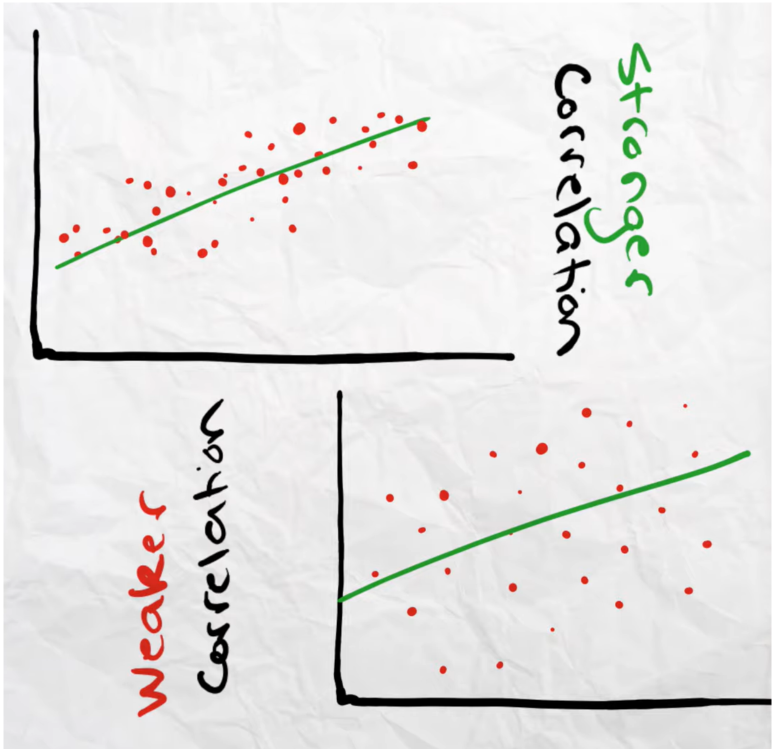
$0.8 <= R^2 <= 1$	strong
$0.5 <= R^2 < 0.8$	medium
$0 <= R^2 < 0.5$	weak

Coefficient of Correlation (r):

- The coefficient of correlation, denoted as r, is a measure of the strength and direction of the linear relationship between two variables, typically X and Y. It represents the degree to which X and Y move together linearly.
- Formula: r is calculated as the square root of r^2.

  $r = \sqrt{(r^2)}$

- r ranges from -1 to 1. A positive value of r indicates a positive linear relationship, while a negative value of r indicates a negative linear relationship. The magnitude of r (how close it is to 1 or -1) tells you how strong the relationship is, with 1 or -1 indicating a perfect linear relationship.



Standard Error of the Estimate (SEE):

- SEE provides an estimate of the typical deviation or "error" in the predicted values of Y by the regression model. In other words, it quantifies how much individual data points are expected to deviate from the regression line.
- It is essential for assessing the quality and precision of a regression model, as it helps in understanding the dispersion of the observed data points around the regression line.

Calculation:

- SEE is calculated as the square root of the Mean Squared Error (MSE), which is the average of the squared differences between the observed values of Y and the predicted values from the regression model. The formula for SEE is as follows:

  SEE = √(MSE)

  Where MSE is calculated as:

  MSE = $\Sigma(Y - \hat{Y})^2 / (n - m - 1)$

  - **Y** represents the observed values of the dependent variable.

- **Ŷ** represents the predicted values of the dependent variable from the regression model.
- **n** is the number of data points.
- **m** is the number of independent variables.

Interpretation:

- SEE is expressed in the same units as the dependent variable Y. It measures the typical "spread" of data points around the regression line.
- A smaller SEE indicates that the predicted values are close to the actual data points, which implies a better fit of the model to the data.
- A larger SEE suggests that there is more variability in the predictions, indicating a less precise model.

			Predicted Score	Error in Prediction	(Error in Prediction)^2
Subject	X = Time	Y = Distance	$\hat{y} = 6 + 2x$	$(y - \hat{y})$	$(y - \hat{y})^2$
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	−1	1
5	4	14	14	0	0
6	5	15	16	−1	1
7	6	20	18	2	4
8	7	18	20	−2	4
9	8	22	22	0	0
10	9	25	24	1	1
					$SSE = \sum(y - \hat{y})^2 = 12$

$$s = \sqrt{MSE} = \sqrt{\frac{12}{(10 - 1 - 1)}} = 1.2$$