Data Analysis

Model Development

Simple & Multiple Linear Regression

Model Development

- A model can be thought of as a mathematical equation used to predict a value given one or more other values
- Relating one or more independent variables to dependent variables.

independent variables or features

dependent variables

'highway-mpg'

'predicted price'

$55\ mpg$

Model

$\$5000$

Model Development

- Usually the more **relevant data** you have the more accurate your model is

'highway-mpg'
'curb-weight'
'engine-size'
'highway-mpg'

Model

'price'

$5400

Model Development

To understand why more data is important consider the following situation:

1. you have two almost identical cars
2. Pink cars sell for significantly less

'highway-mpg'
'curb-weight'
'engine-size'
'highway-mpg'

→ Model → $Y = \$5400$

'highway-mpg'
'curb-weight'
'engine-size'
'highway-mpg'

→ Model → $Y = \$5400$

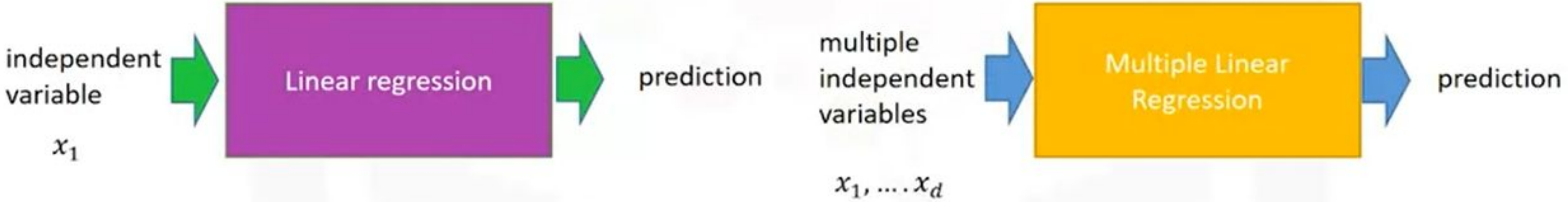
Linear Regression and Multiple Linear Regression

What is regression

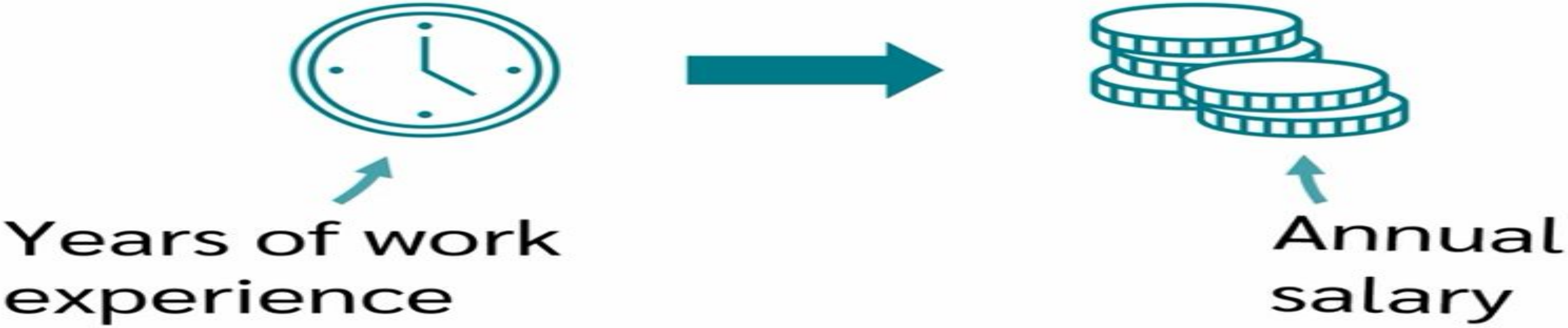
- A regression is **a statistical technique that relates a dependent variable to one or more independent (explanatory) variables**. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.
- **Regression models** are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

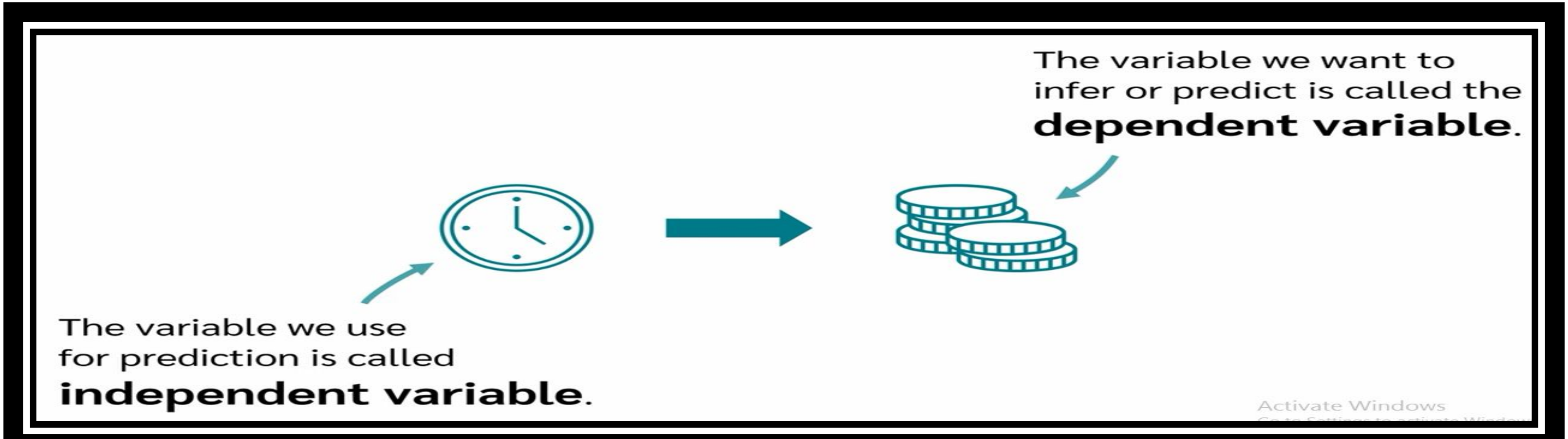
Introduction

- Linear regression will refer to one independent variable to make a prediction
- Multiple Linear Regression will refer to multiple independent variables to make a prediction

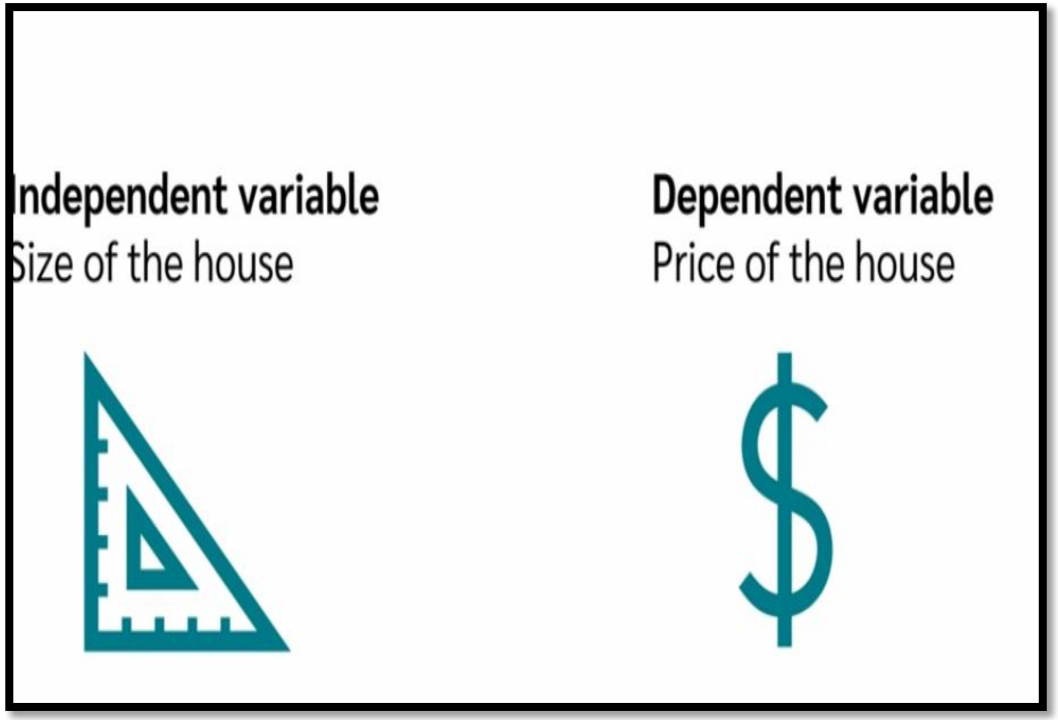


Simple Linear Regression
can help us understand how...





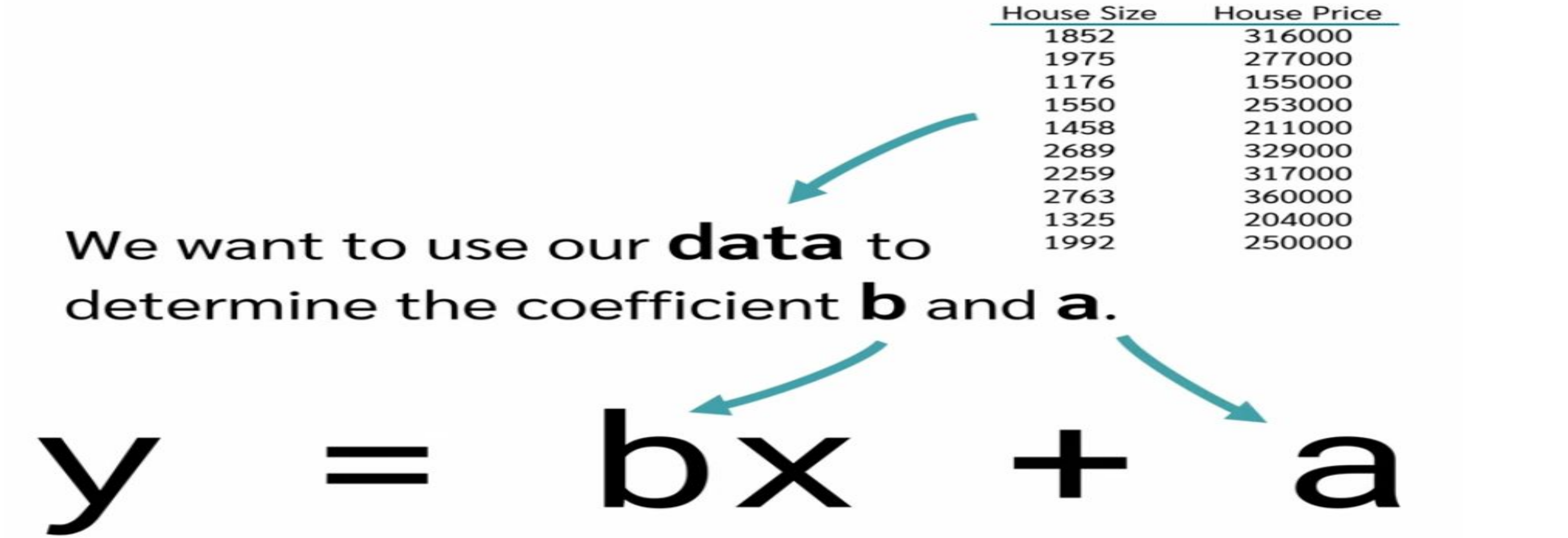
Predict house prices



- As mentioned above, linear regression is a predictive modeling technique. It is used whenever there is a linear relation between the dependent and the independent variables.

  - $Y = b0 + b1 * x$

- It is used in estimating exactly how much of y will change, when x changes a certain amount.



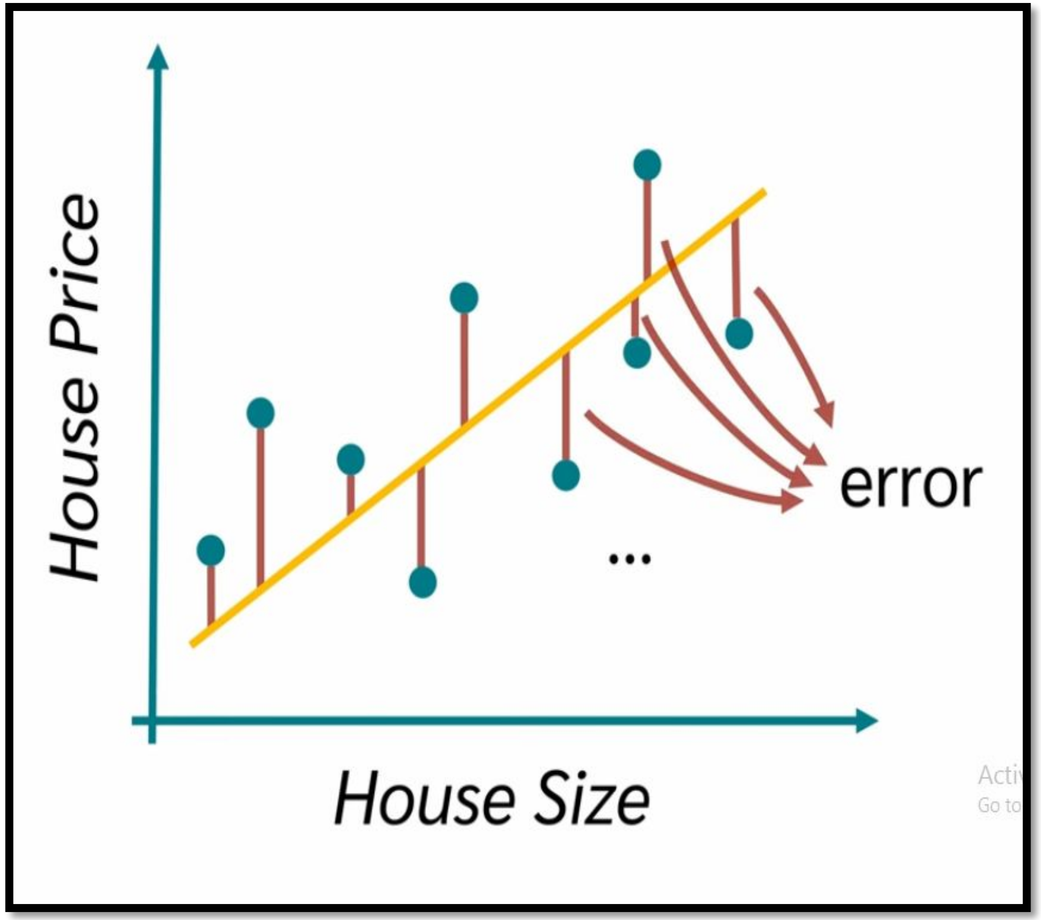
Simple Linear Regression

1. The predictor (independent) variable - x
2. The target (dependent) variable - y

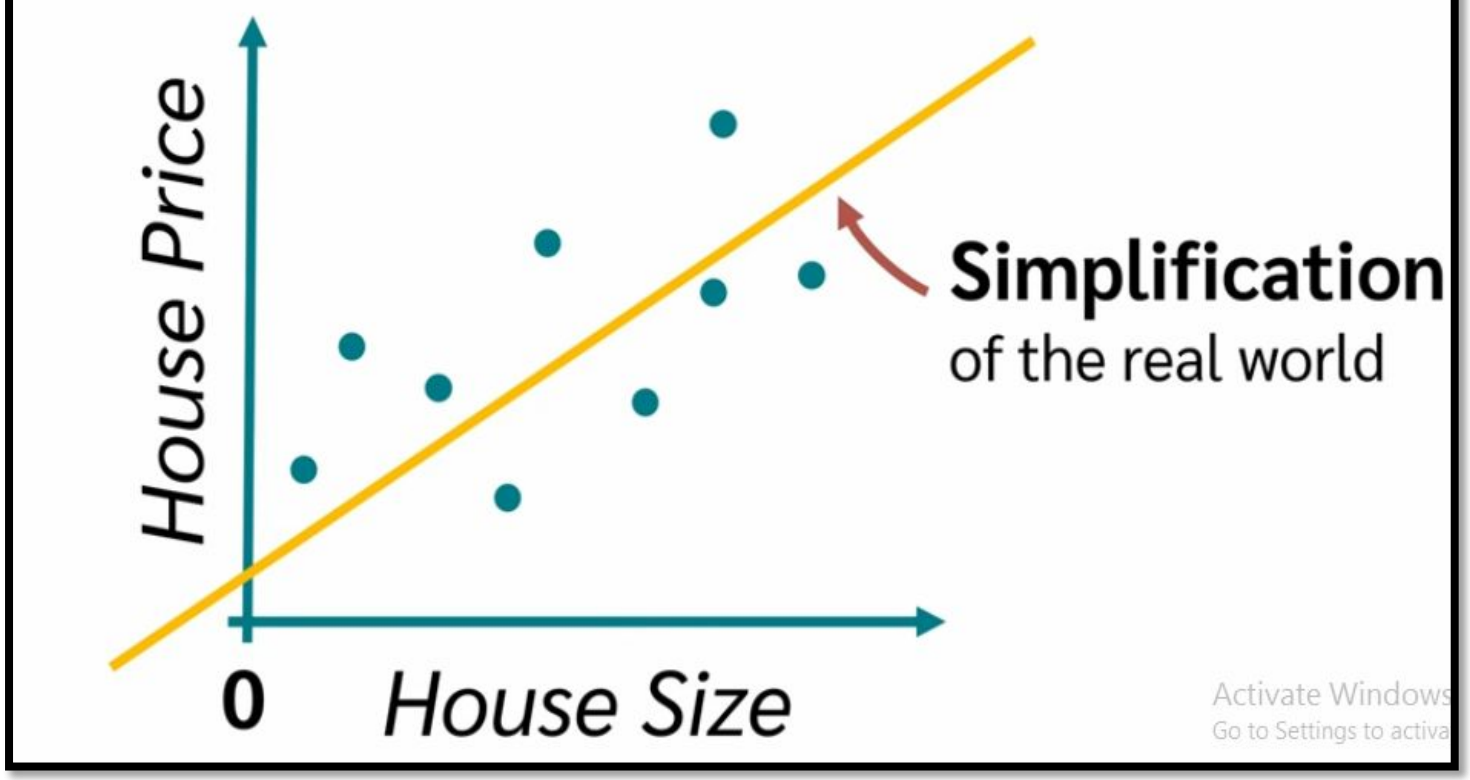
$$y = b_0 + b_1 x$$

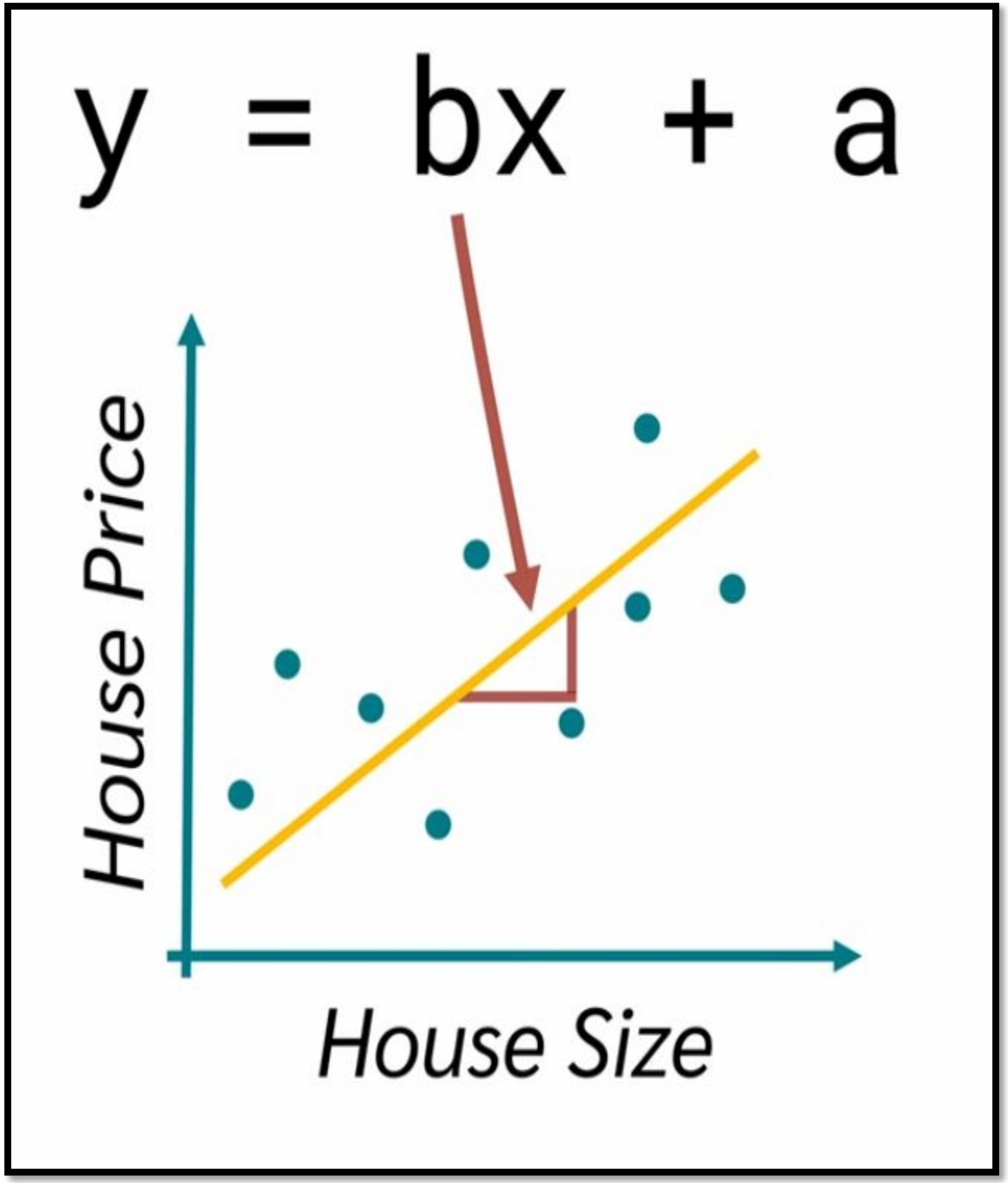
- b_0: **the intercept**
- b_1 : **the slope**

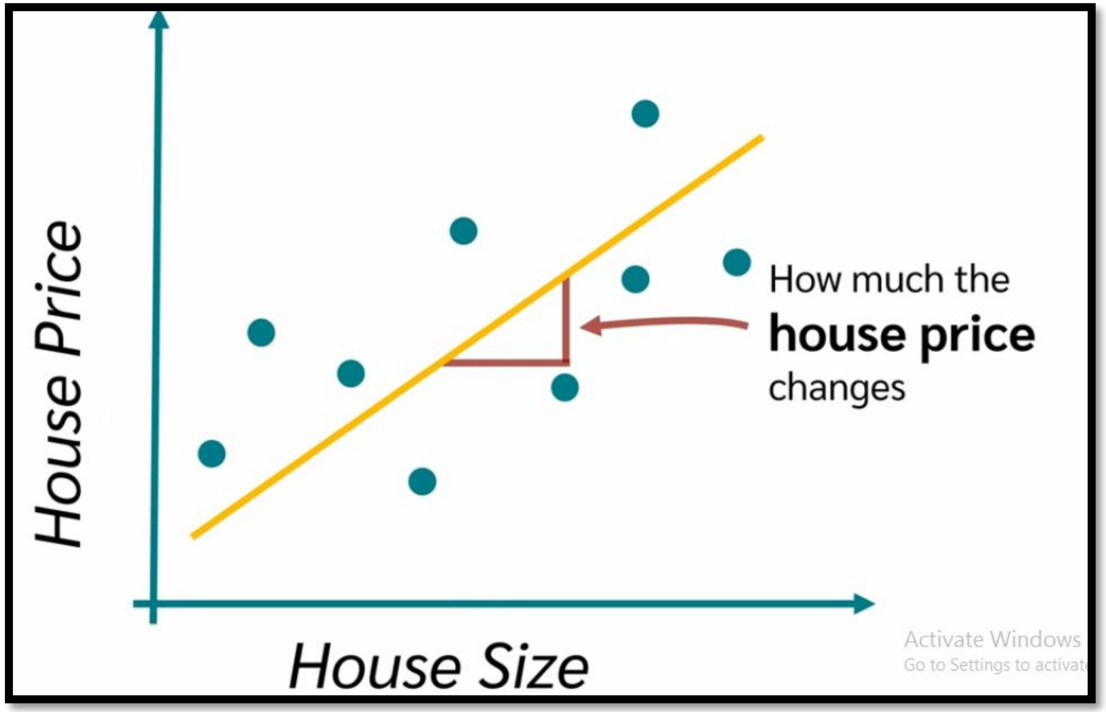
Reseduals

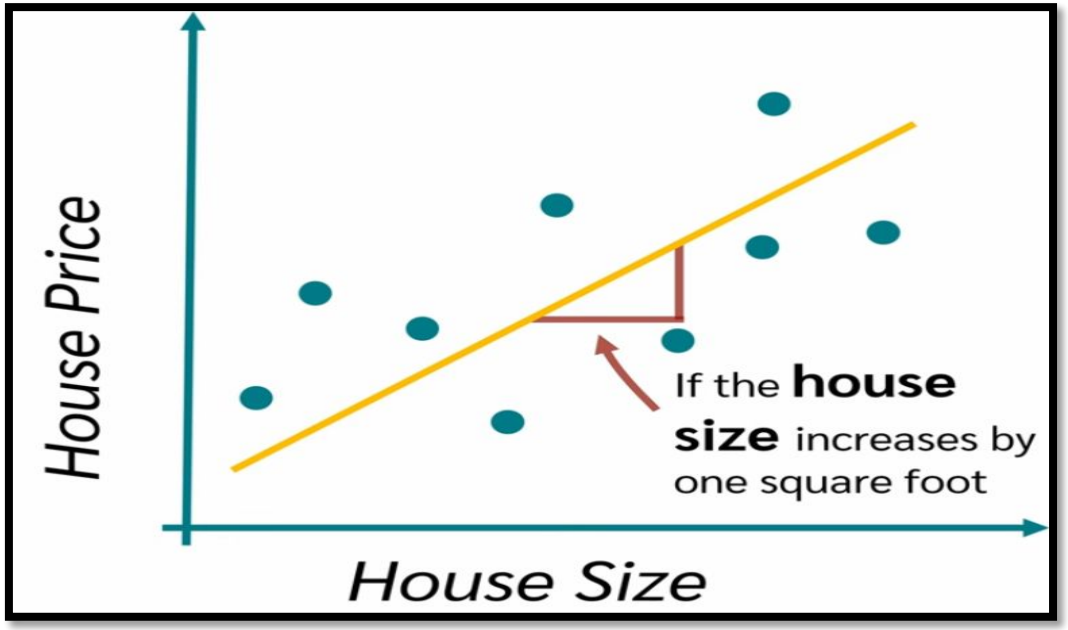


$$y = bx + a$$



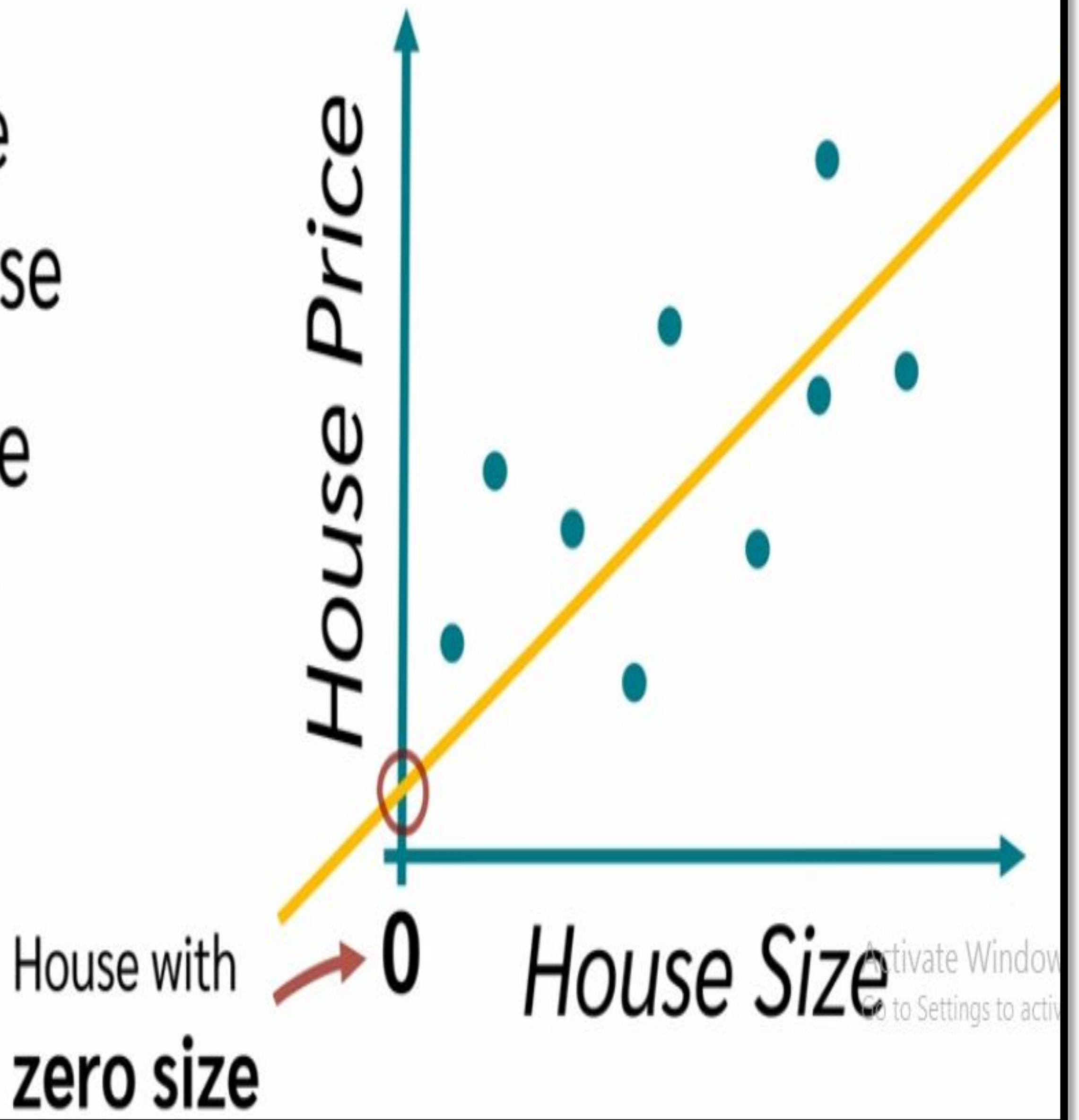

y = bx + a
House Price
House Size

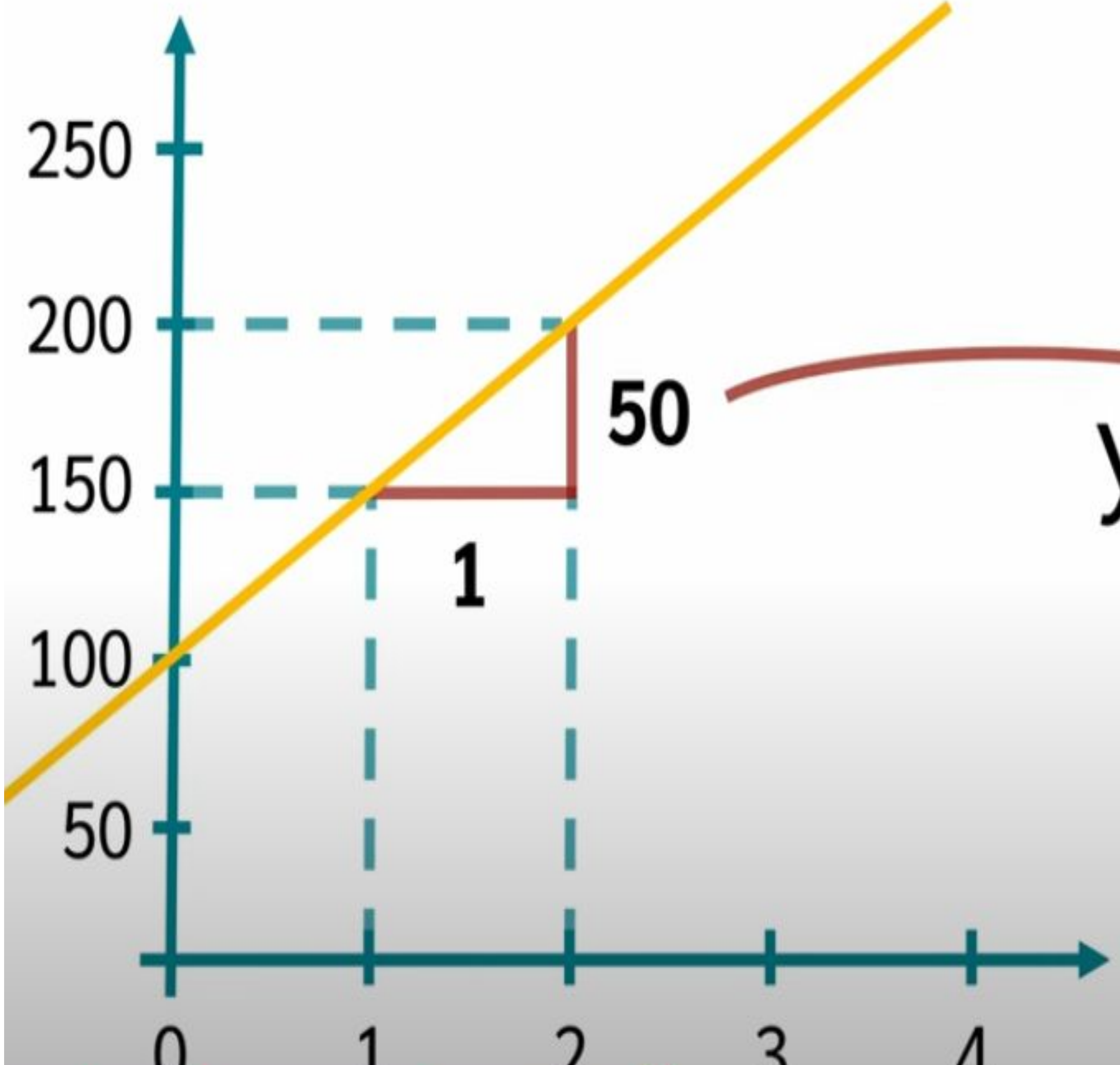

House Price
How much the **house price** changes
House Size


House Price
If the **house size** increases by one square foot
House Size

$$y = bx + a$$

Predicting the price of a house with **zero** size doesn't make sense.





$$y = bx + 100$$

$$y = bx + a$$

$$b = r\frac{s_y}{s_x}$$

Correlation coefficient

House Size	House Price
1852	316000
1975	277000
1176	155000
1550	253000
1458	211000
2689	329000
2259	317000
2763	360000
1325	204000
1992	250000

$$y = bx + a$$

$$b = r\frac{s_y}{s_x}$$

61341.34

0.92

Standard deviations

House Size	House Price
1852	316000
1975	277000
1176	155000
1550	253000
1458	211000
2689	329000
2259	317000
2763	360000
1325	204000
1992	250000

$$y = bx + a$$

$$b = r\frac{s_y}{s_x}$$

108.35

0.92

61341.34

518.17

$$y = bx + a$$

House Size	House Price
1852	316000
1975	277000
1176	155000
1550	253000
1458	211000
2689	329000
2259	317000
2763	360000
1325	204000

mean

$$a = \bar{y} - b \cdot \bar{x}$$

267200

$$y = bx + a$$

House Size
1852
1975
1176
1550
1458
2689
2259
2763
1325
1992

$$a = \bar{y} - b \cdot \bar{x}$$

267200 108.35

$$b = r\frac{s_y}{s_x}$$

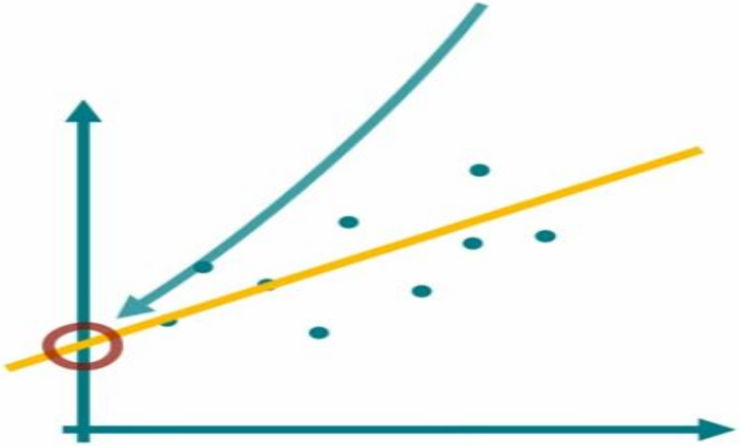
$$\sigma = \sqrt{\boxed{\frac{1}{n}} \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad s = \sqrt{\boxed{\frac{1}{n-1}} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

60919.44

$$a = \bar{y} - b \cdot \bar{x}$$

267200 108.35 1903.9

$$y = 108.35 \cdot 0 + 60919.44 = 60919.44$$



$$y = \boxed{108.35} \cdot 0 + 60919.44 = 60919.44$$

$$y = 108.35 \cdot 1 + 60919.44 = 61027.79$$

$$y = 108.35 \cdot 2 + 60919.44 = 61136.14$$

$$y = 108.35 \cdot 3 + 60919.44 = 61244.49$$

$$y = 108.35 \cdot 4 + 60919.44 = 61352.84$$

How to Find Linear Regression Slope:

Find the following data from the information given: Σx, Σy, Σxy, Σx^2, Σy^2. If you don't remember how to get those variables from data, see this Pearson's correlation coefficient.
In the linear regression formula, the slope is the a in the equation **y' = b + ax**.

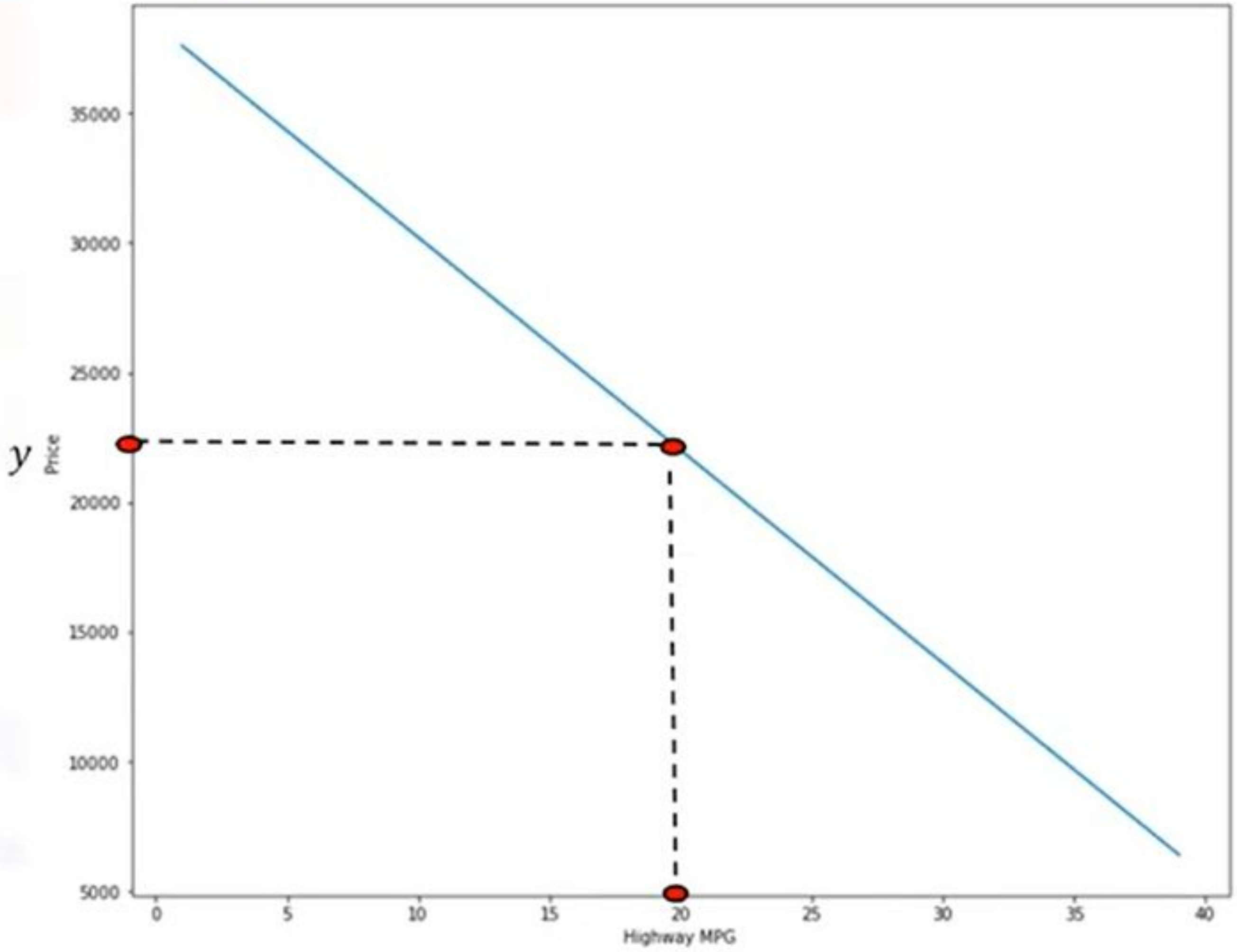
$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

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

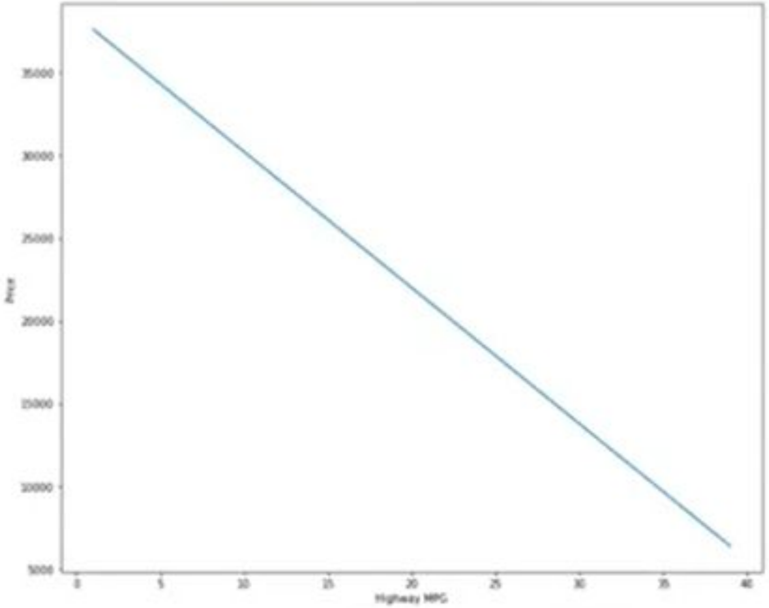
Simple Linear Regression: Prediction

$$y = 38423 - 821x$$

$$= 38423 - 821\,(20)$$
$$= 22\,003$$



Simple Linear Regression: Fit

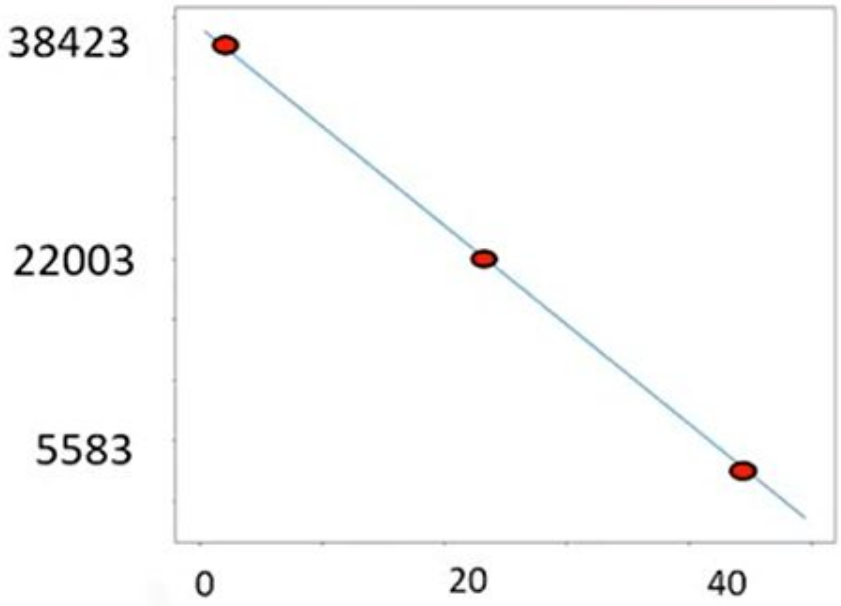


Fit

(b_0, b_1)

x

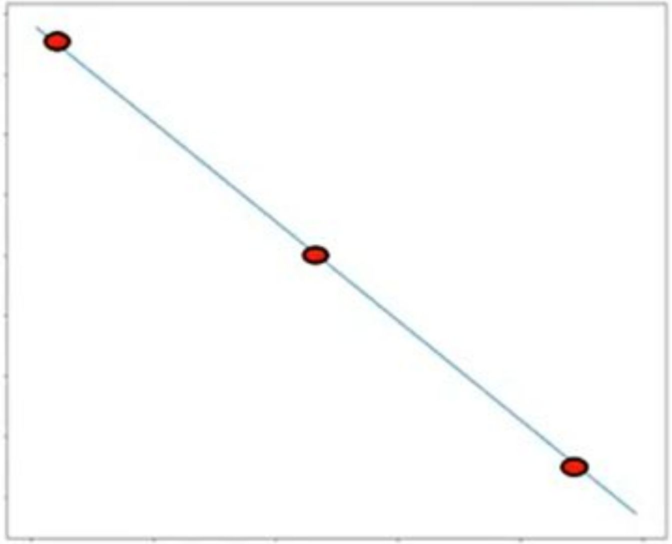
Simple Linear Regression: Fit



$$X = \begin{bmatrix} \\ \\ \end{bmatrix} \quad Y = \begin{bmatrix} \\ \end{bmatrix}$$

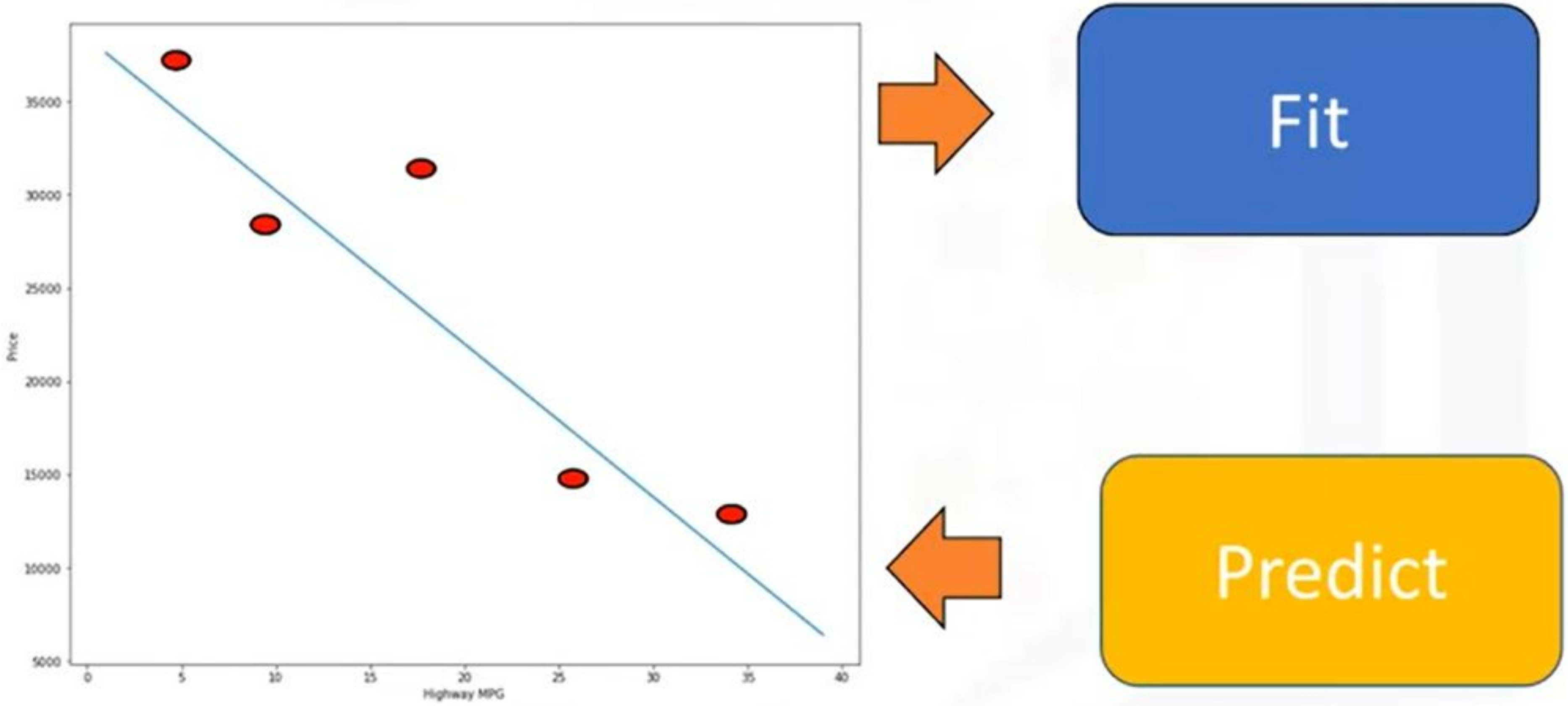
x

Simple Linear Regression: Fit



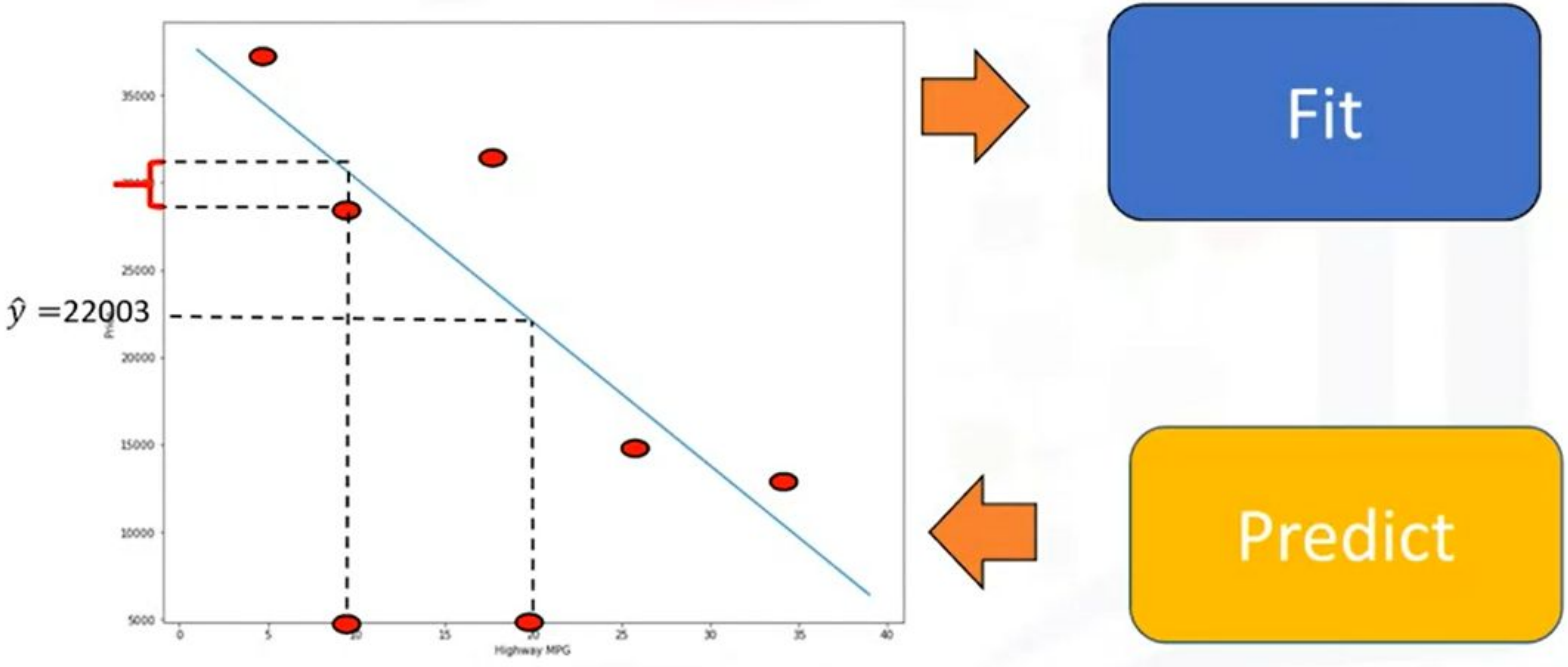
$$X = \begin{bmatrix} 0 \\ 20 \\ 40 \end{bmatrix} \quad Y = \begin{bmatrix} 38423 \\ 22003 \\ 5583 \end{bmatrix}$$

Simple Linear Regression



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

Simple Linear Regression



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

Fitting a Simple Linear Model Estimator

- X :Predictor variable
- Y: Target variable

1. Import linear_model from scikit-learn

```
from sklearn.linear_model import LinearRegression
```

2. Create a Linear Regression Object using the constructor :

```
lm=LinearRegression()
```

Fitting a Simple Linear Model

- We define the predictor variable and target variable

```
X = df[['highway-mpg']]
Y = df['price']
```

- Then use lm.fit (X, Y) to fit the model , i.e fine the parameters b_0 and b_1

```
lm.fit(X, Y)
```

- We can obtain a prediction

```
Yhat=lm.predict(X)
```

Yhat	X
2	5
:	
3	4

SLR – Estimated Linear Model

- We can view the intercept (b_0):

```
lm.intercept_
38423.305858
```

- We can also view the slope (b_1):

```
lm.coef_
-821.73337832
```

- The Relationship between Price and Highway MPG is given by:
- **Price = 38423.31 - 821.73 * highway-mpg**

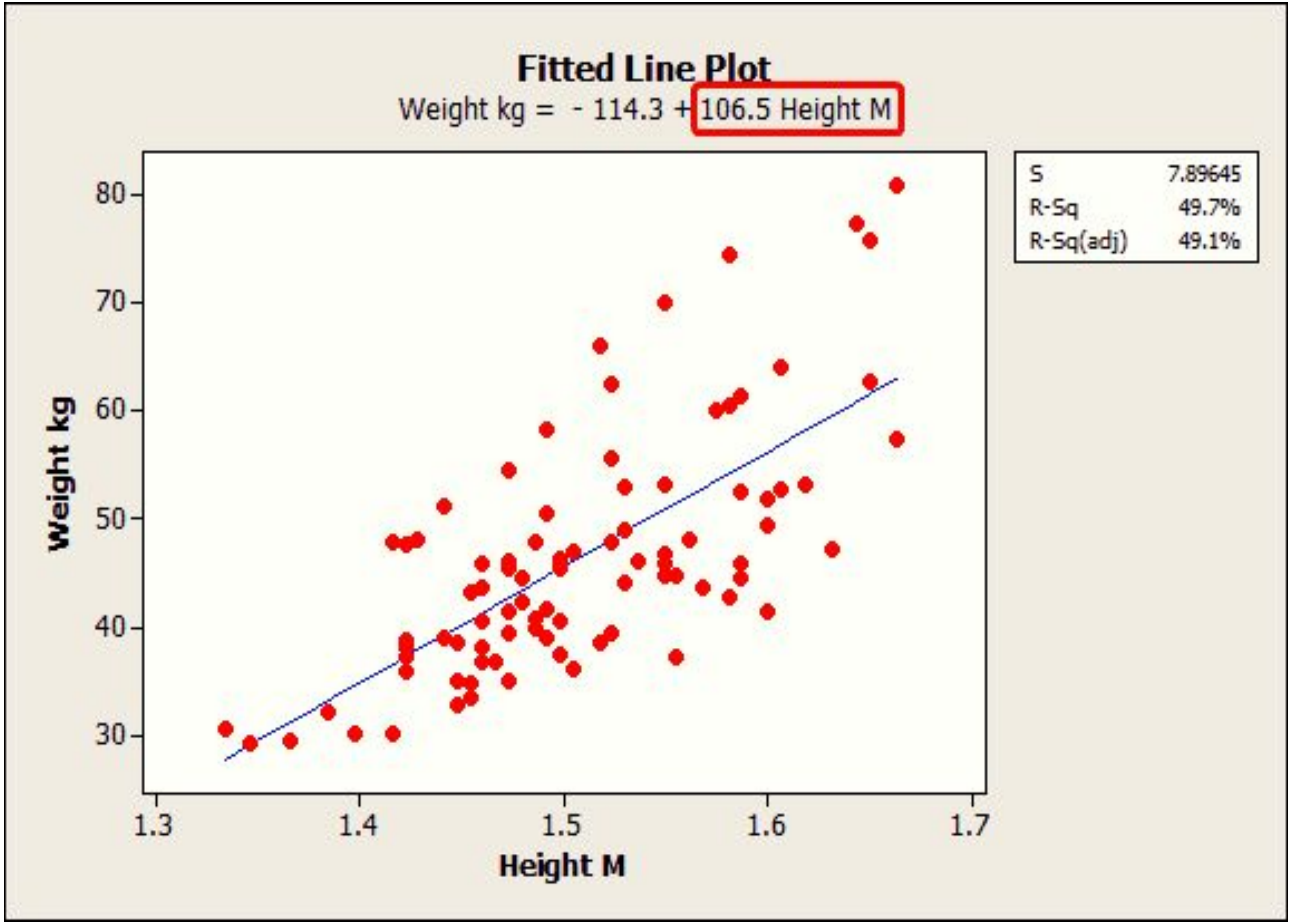
$$\hat{Y} = b_0 + b_1 x$$

Cont…

- To find out the best fit line, we have something called **residual sum of squares (RSS)**. In RSS, we take the square of residuals and sum them up.
- The line with the **lowest value of RSS** is the best fit line.
- In simple linear regression, if the coefficient of x is positive, then we can conclude that the relationship between the independent and the dependent variables is positive.
- Here, if the value of x increases, the value of y also increases

example

The height coefficient in the regression equation is 106.5. This coefficient represents the mean increase of weight in kilograms for every additional one meter in height. If your height increases by 1 meter, the average weight increases by 106.5 kilograms.



1. *Linear Relationship*
2. *Independence of Errors*
3. *Homoscedasticity*
4. *Normally Distributed Errors*

Coefficients

Copy AI Interpretation

	Unstandardized Coefficients	Standardized Coefficients			
Model	B	Beta	Standard error	t	p
(Constant)	60919.44		33271.8	1.83	.104
House Size	108.35	0.92	16.86	6.43	<.001

$$y = 108.35\,x + 60919.44$$

p-value small

< 0.05

Reject
null hypothesis

Suggesting a **significant relationship** between the variables.

p-value large

≥ 0.05

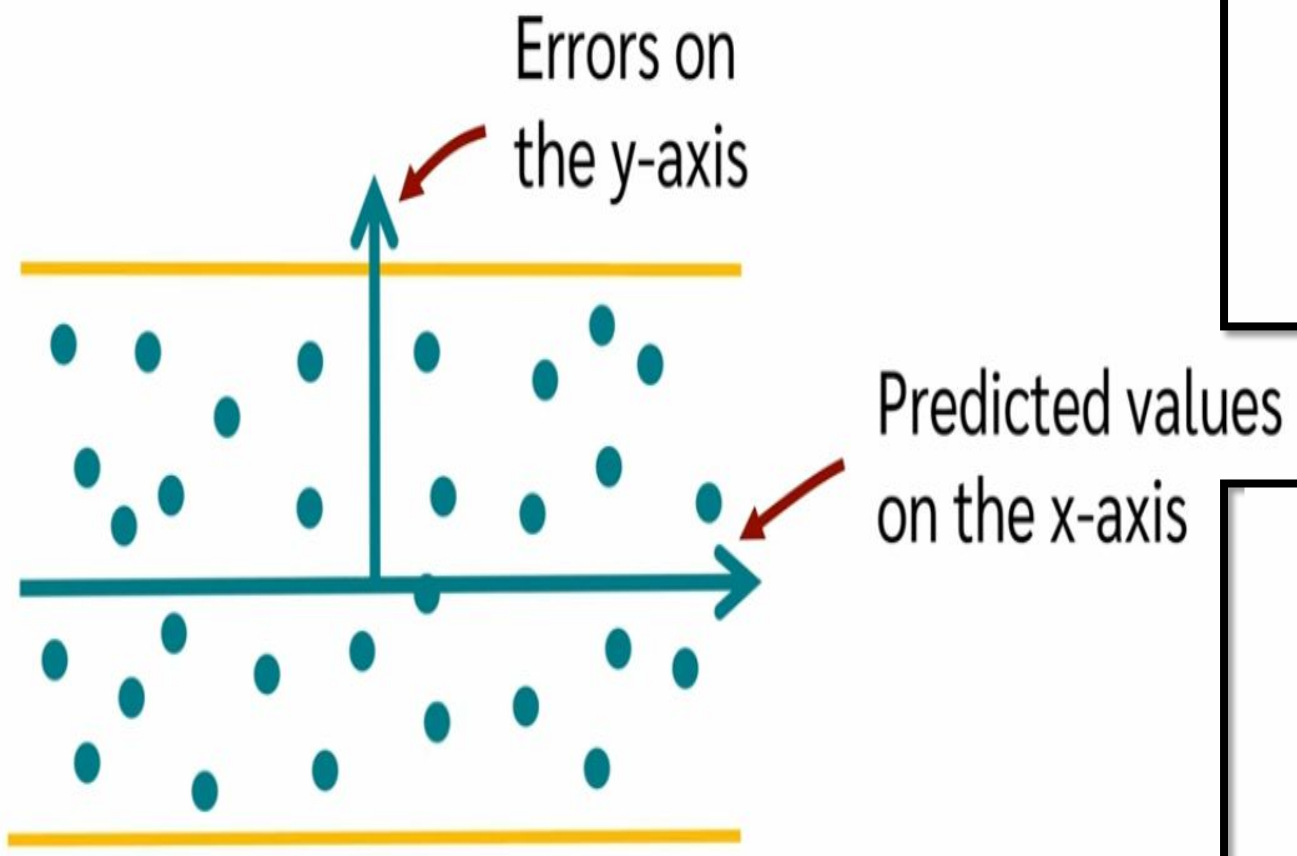
Fail to reject the
null hypothesis

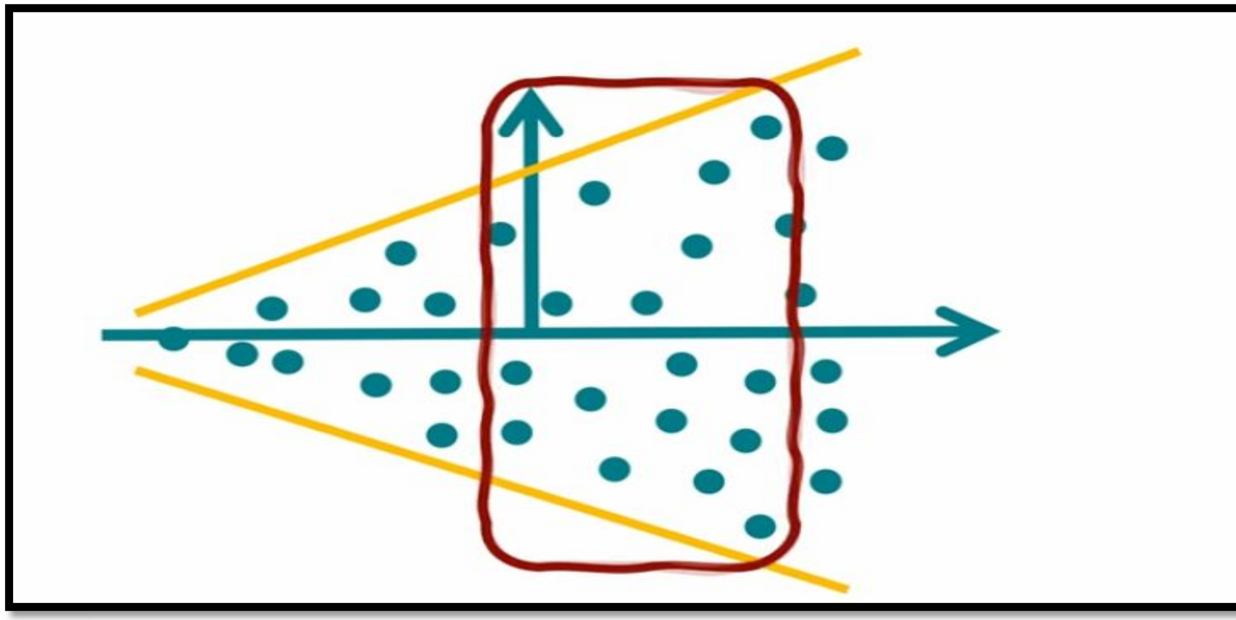
The observed data may have occurred by chance with no strong evidence of a relationship.

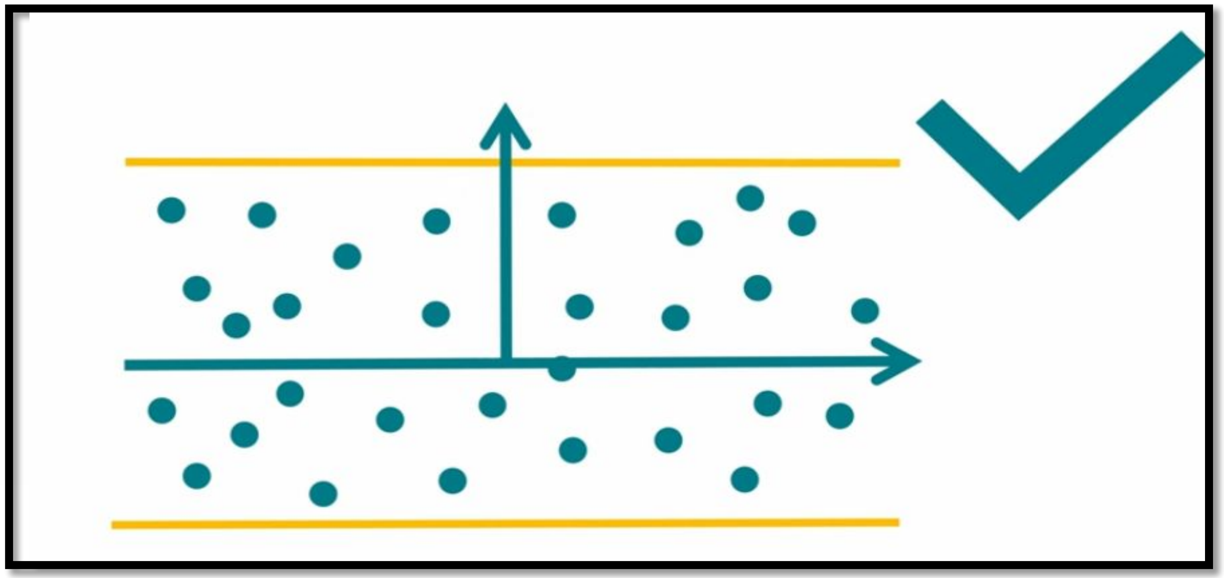
relationship
between house price and size

	Unstandardized Coefficients	Standardized Coefficients			
Model	B	Beta	Standard error	t	p
(Constant)	60919.44		33271.8	1.83	.104
House Size	108.35	0.92	16.86	6.43	$\boxed{<.001}$

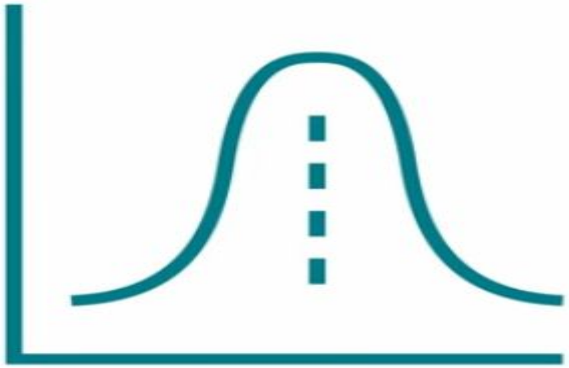
3. *Homoscedasticity*



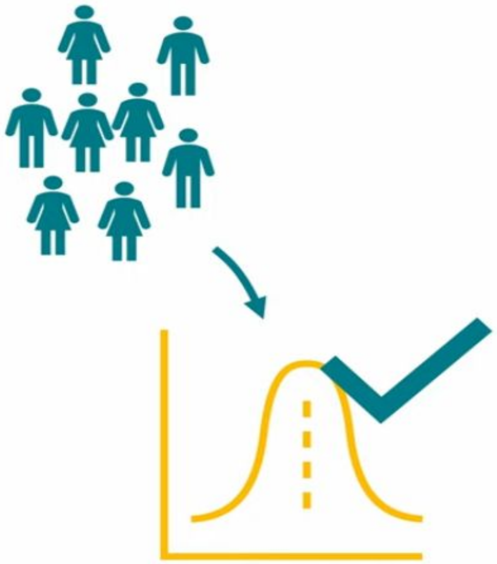


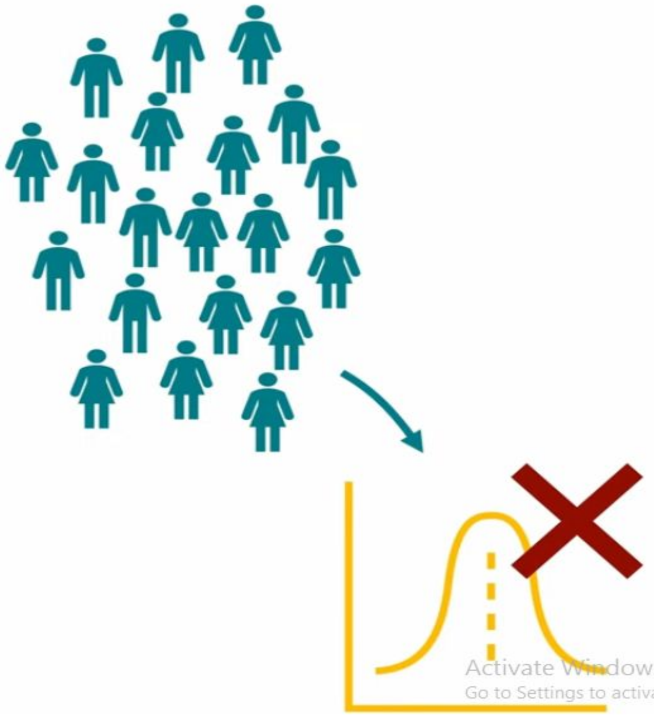


4. Normally Distributed Errors



The errors should be **normally distributed**.



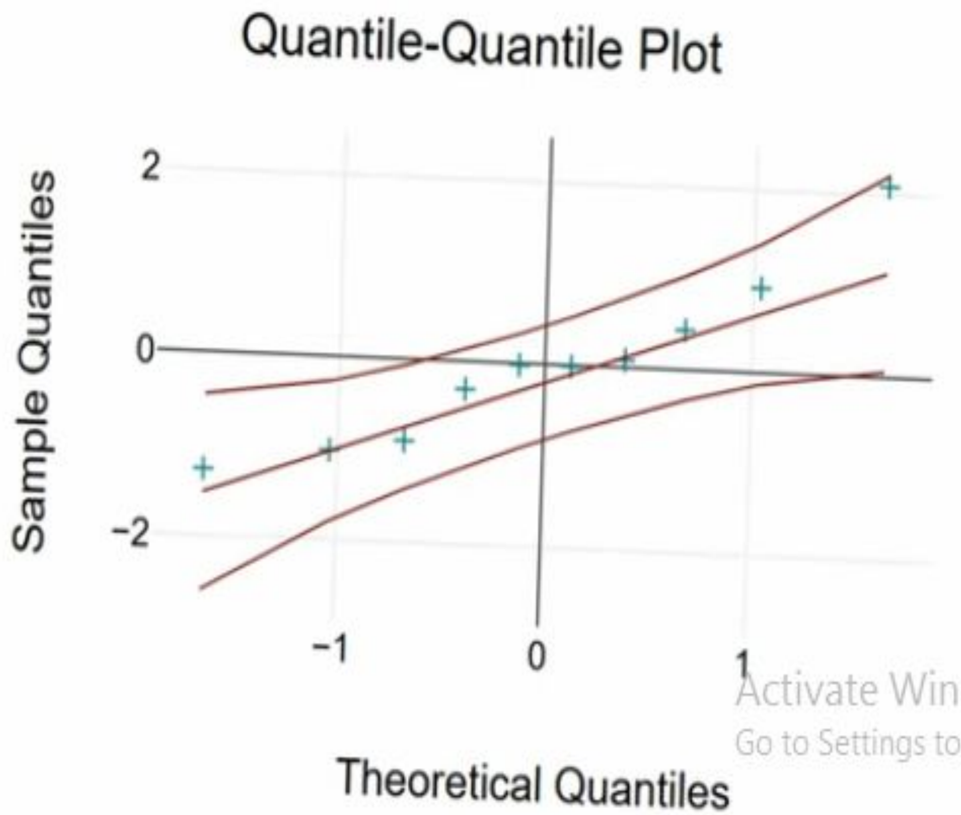


Normally Distributed Errors

Analytical

	Statistics	p
Kolmogorov-Smirnov	0.17	.898
Kolmogorov-Smirnov (Lilliefors Corr.)	0.17	.687
Shapiro-Wilk	0.94	.544
Anderson-Darling	0.32	.535

Graphic



Multiple Linear Regression (MLR)

This method is used to explain the relationship between:

- One continuous target (Y) variable
- Two or more predictor (X) variables

Predicting House Price using multiple features

What is multiple regression and why is it used?

- Multiple regression is **a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables**.
- The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.
- Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

You can use multiple linear regression :

- How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
- The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Multiple Linear Regression (MLR)

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

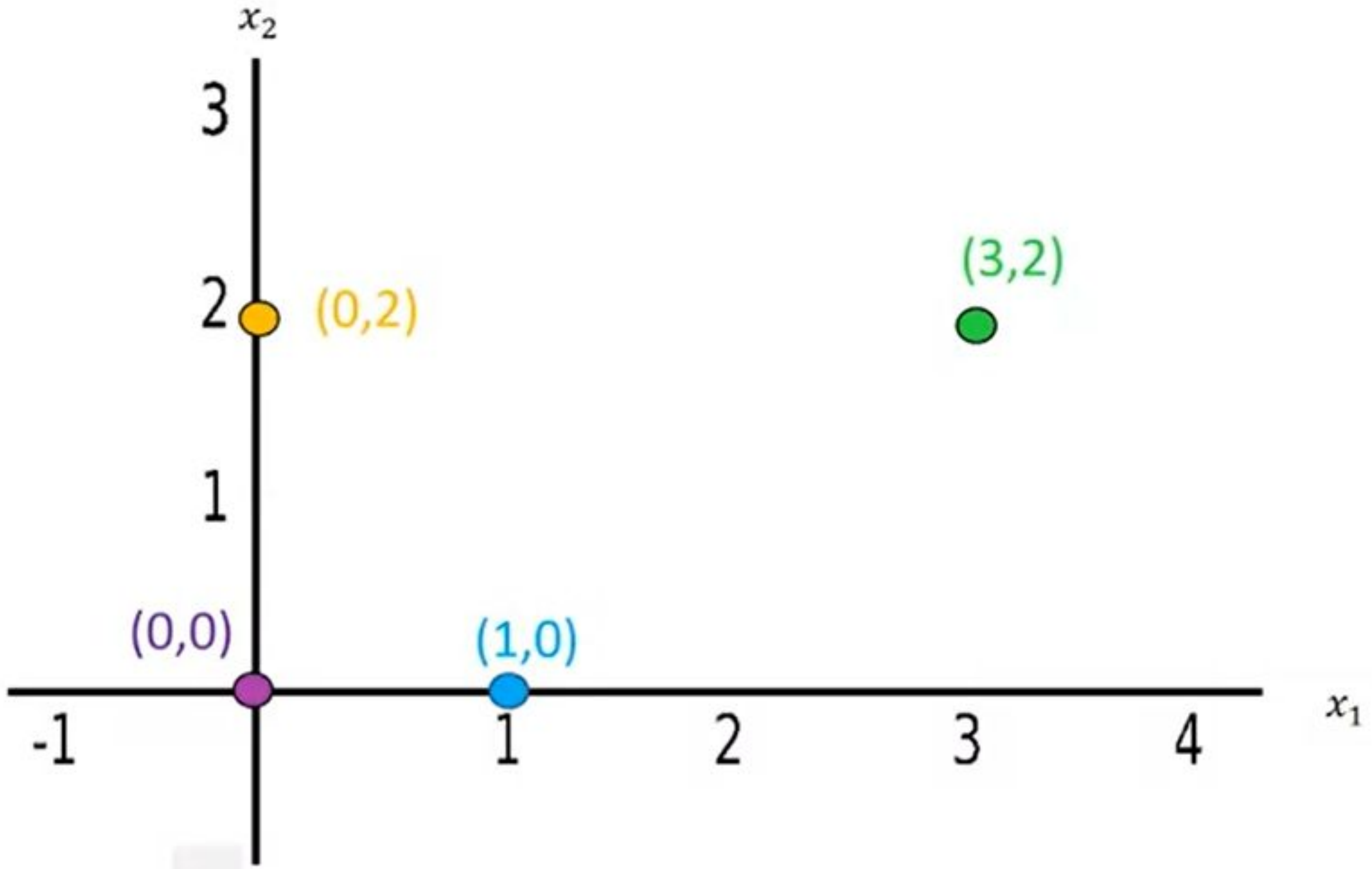
- b_0 **: intercept (X=0)**
- b_1**:** the **coefficient** or **parameter** of x_1
- b_2**:** the **coefficient** of **parameter** x_2 and so on..

Multiple Linear Regression (MLR)

$$\hat{Y} = 1 + 2x_1 + 3x_2$$

- The variables x_1 and x_2 can be visualized on a 2D plane, lets do an example on the next slide

n	x_1	x_2
1	0	0
2	0	2
3	1	0
4	3	2

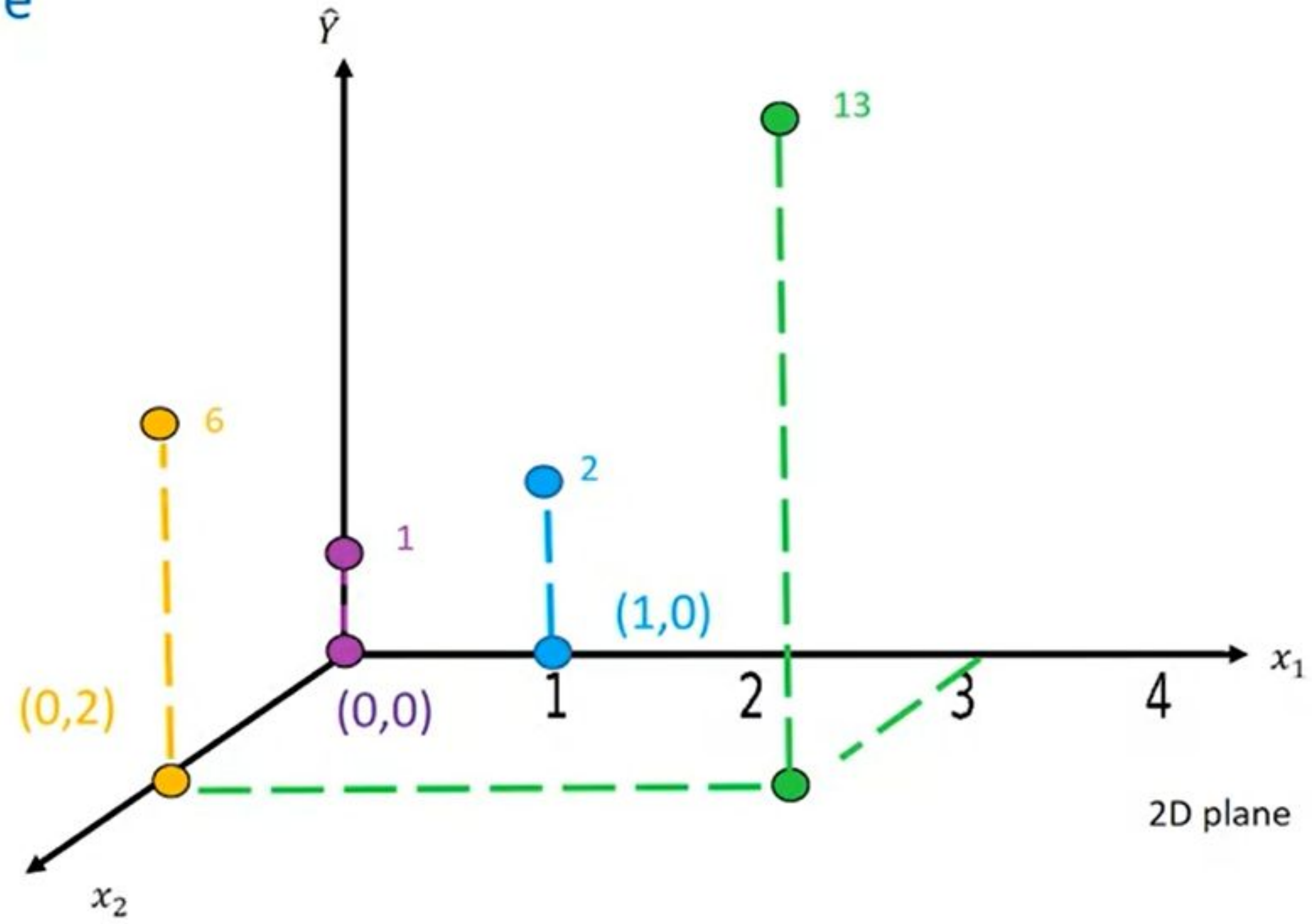


- This is shown below where

$$\hat{Y} = 1 + 2x_1 + 3x_2$$

n	x_1	x_2	$\hat{Y}$
1	0	0	1
2	0	2	6
3	1	0	2
4	3	2	13

$\boldsymbol{x}$



Fitting a Multiple Linear Model Estimator

1. We can extract the for 4 predictor variables and store them in the variable Z

```
Z = df[['horsepower', 'curb-weight', 'engine-size', 'highway-mpg']]
```

2. Then train the model as before:

```
lm.fit(Z, df['price'])
```

3. We can also obtain a prediction

```
Yhat=lm.predict(X)
```

x_1	x_2	x_3	x_4
3	5	-4	3
:	:	:	:
2	4	2	-4

Yhat
2
:
3

MLR – Estimated Linear Model

1. Find the intercept (b_0)

```
lm.intercept_
-15678.742628061467
```

2. Find the coefficients (b_1, b_2, b_3, b_4)

```
lm.coef_
array([52.65851272 ,4.69878948,81.95906216 , 33.58258185])
```

The Estimated Linear Model:

- **Price** = -15678.74 + (52.66) * **horsepower** + (4.70) * **curb-weight** + (81.96) * **engine-size** + (33.58) * **highway-mpg**

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$