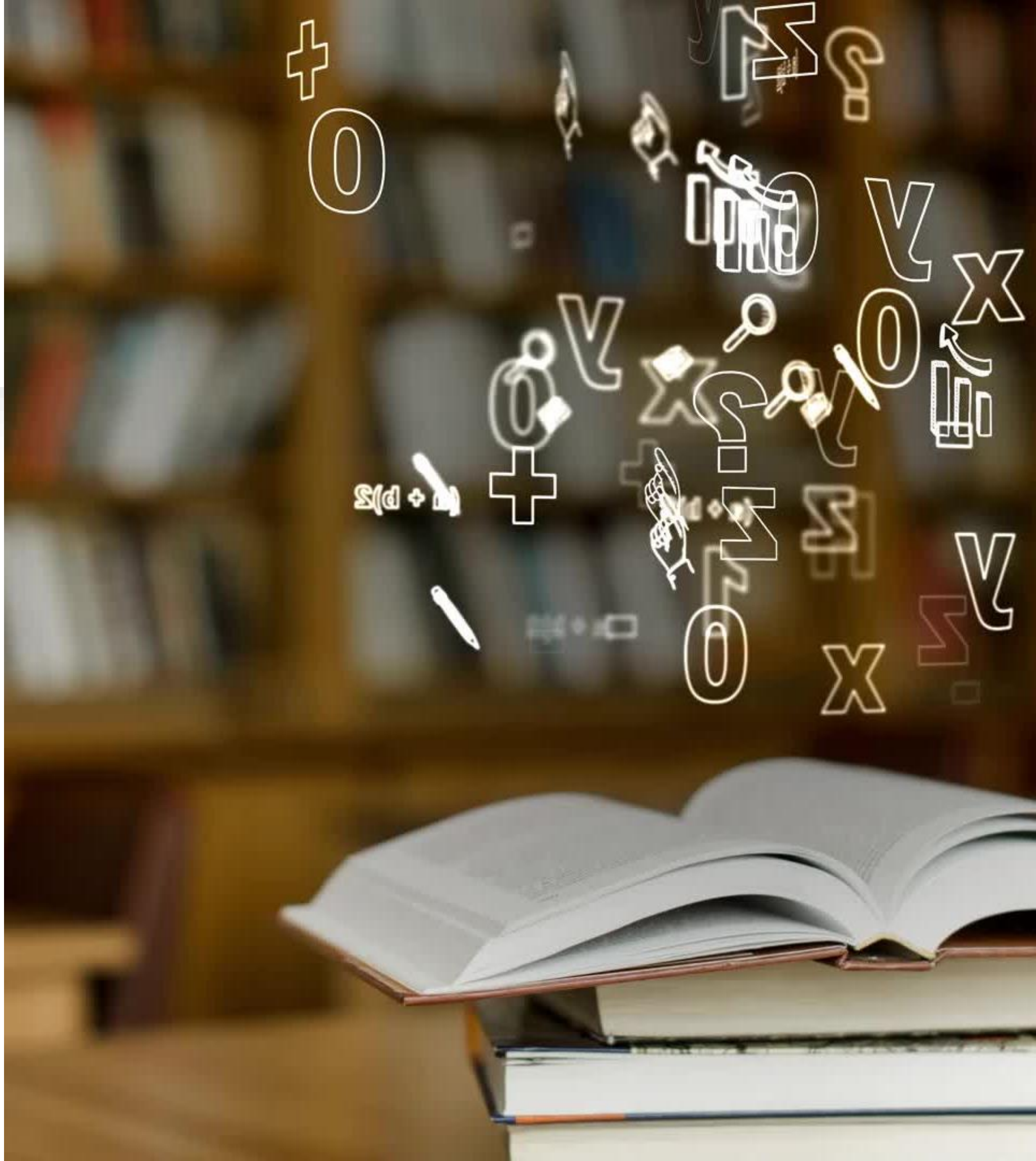
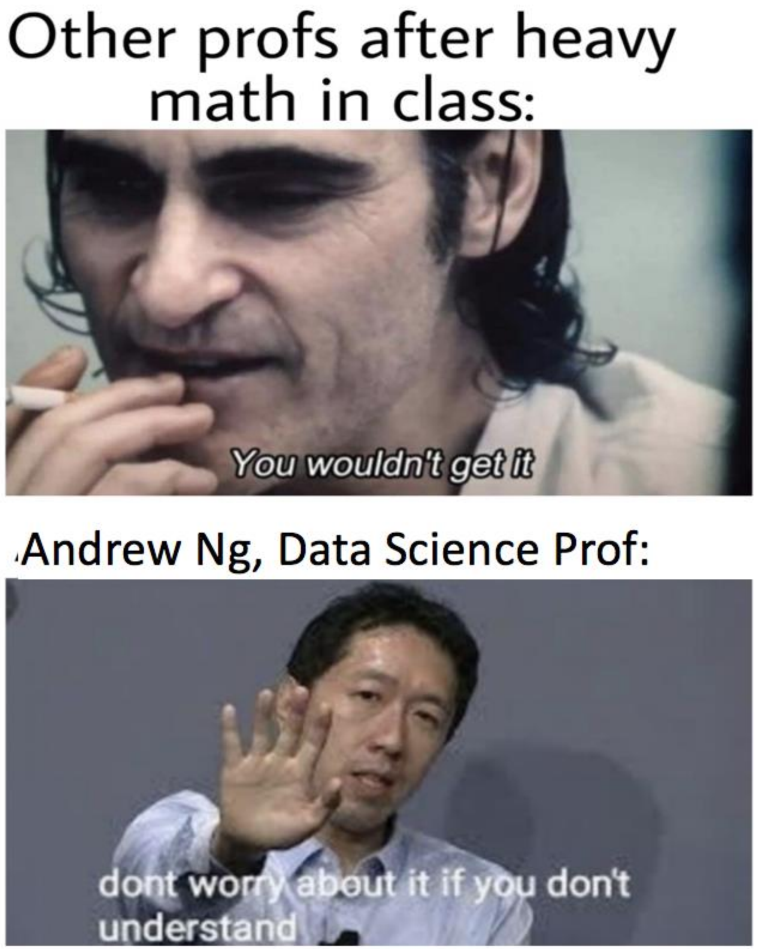
Regression

Week 3

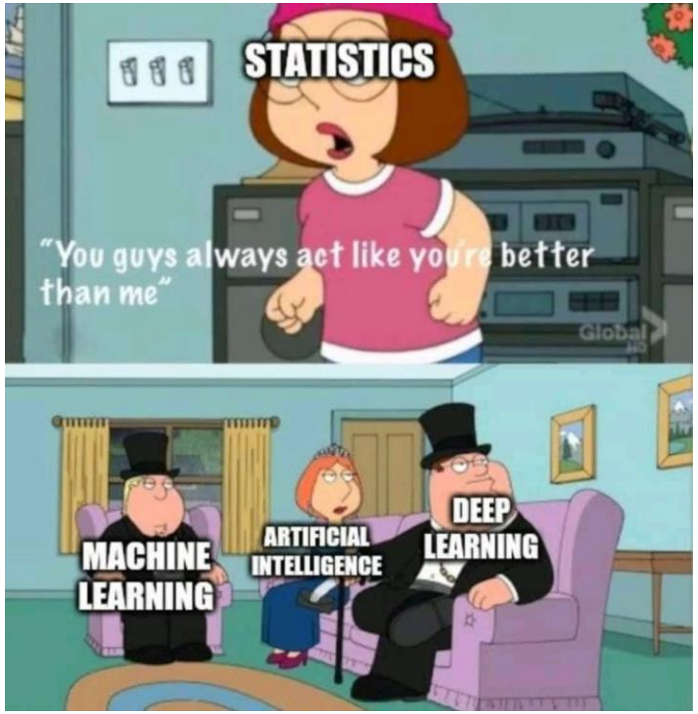
Disclaimer

- The topics may require a bit of knowledge about data analysis, statistics, linear algebra, and multivariable/variate calculus.
- We will not go in the entirety of each topics, as it can deplete too much of our time.
- Let’s focus more on machine learning on this topic. Let the statistics or data analysis class do the rest.
- Topics will be implemented using Python and appropriate libraries.
- Don’t worry, we aren't solving things on paper, we’re going to code things here. ;)
- This is a Machine Learning class! :D




Other profs after heavy math in class:
You wouldn't get it
Andrew Ng, Data Science Prof:
dont worry about it if you don't understand

Https://expert-programming-tutor.com
Machine Learning
Statistics
Calculus
Discrete Structure
Problem solving
Data Structure and Algorithm
python


STATISTICS
"You guys always act like you're better than me"
MACHINE LEARNING
ARTIFICIAL INTELLIGENCE
DEEP LEARNING

"If you don't understand, don't worry about it"
me
LINEAR ALGEBRA, CALCULUS, PROBABILITY & STATISTICS...

Andrew Ng, Ph.D

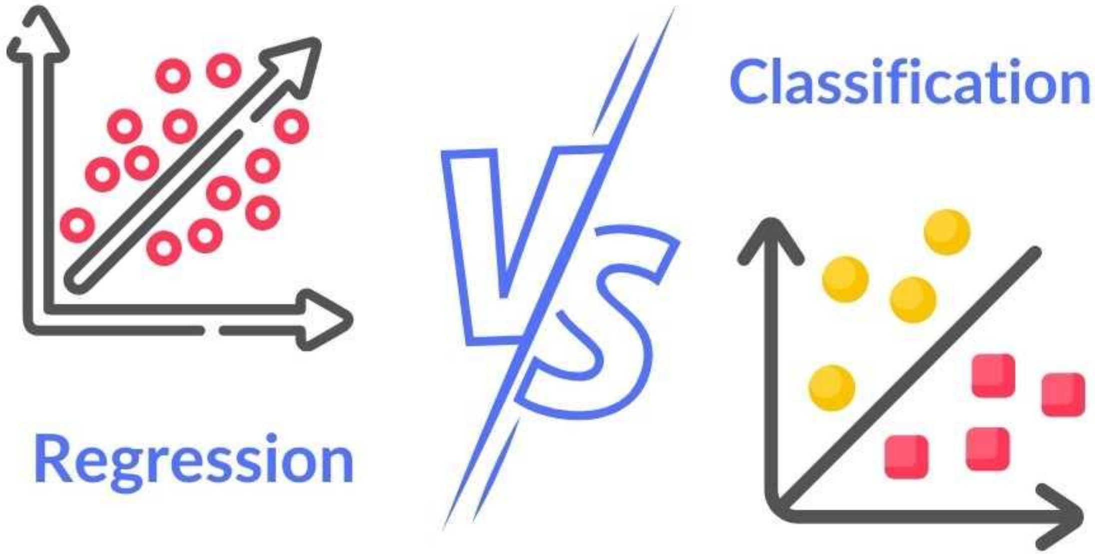
- Dr. Andrew Ng is a globally recognized leader in AI (Artificial Intelligence). He is Founder of DeepLearning.AI, Founder & CEO of Landing AI, General Partner at AI Fund, Chairman and Co-Founder of Coursera and an Adjunct Professor at Stanford University's Computer Science Department.

- As a pioneer in machine learning and online education, Dr. Ng has changed countless lives through his work in AI and has authored or co-authored over 200 research papers in machine learning, robotics and related fields. In 2013, he was named to the Time 100 list of the most influential persons in the world.

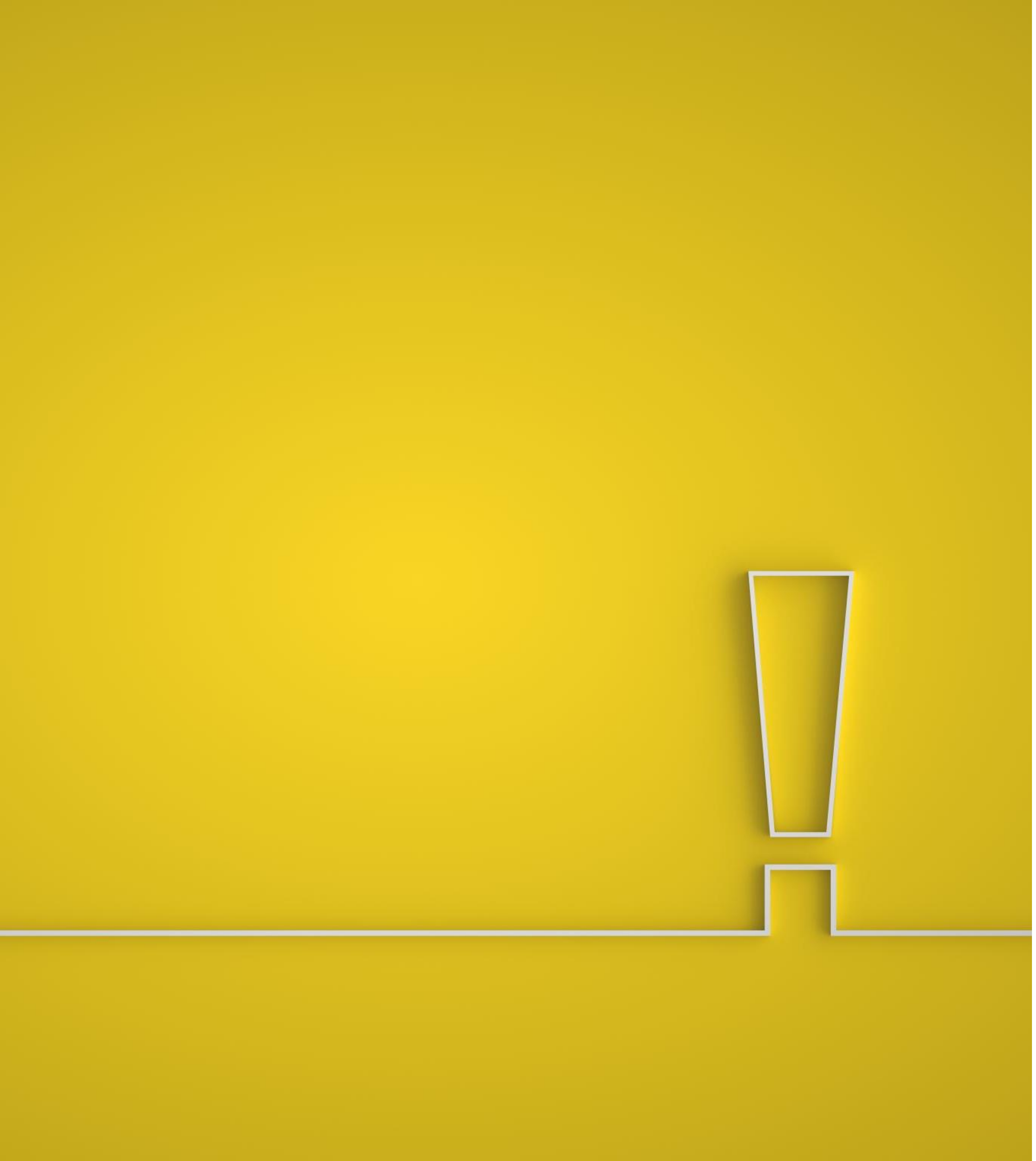
Core Topics

- Linear Regression
- Multiple Linear Regression
- Cost Function
- Gradient Descent

Classification vs Regression



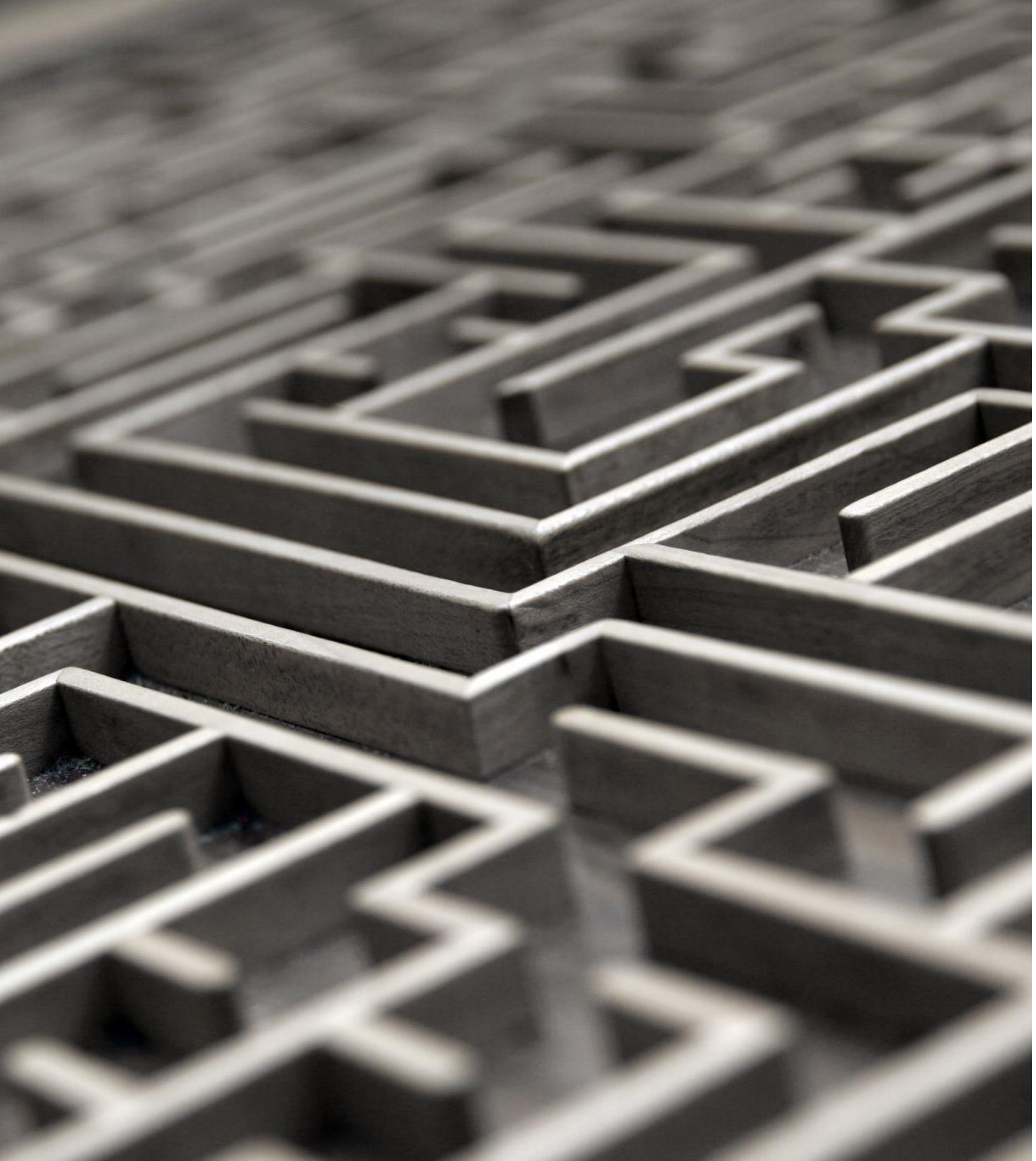
- Classification: Predicts a **categorical** or **discrete** class of the dataset based on the independent input variable.
- Regression: It predicts the **continuous** output variables based on the independent input variable.



Regression Models

- Describe **relationship between variables by fitting a line** to the observed data.
- Linear **regression** models **use a straight line**, while nonlinear **logistic use a curved line**.
- Regression allows us to **estimate** how a **dependent variable changes as the independent variable(s) change**.

Linear Regression



What is Linear Regression?

- Computes the **linear relationship** between a **dependent variable** and one or more **independent** features.
- Determines the **strength** of **relationships**.
- Learns from the **labeled** datasets and maps the data points to the most optimized **linear** functions.
- Can perform **predictions** on new datasets.

Types of Linear Regression

- **Simple** Linear Regression (**Univariate**)
- **Multiple** Linear Regression (**Multivariate**)

- When the number of the **independent feature is 1** then it is known as **Univariate** Linear regression, and in the case of **more than one feature**, it is known as **multivariate** linear regression.

Simple Linear Regression

A simple example of Linear Regression

- You are a social researcher interested in the **relationship** between **income** and **happiness**.
- You survey 500 people whose **incomes range from 15k to 75k** and ask them to rank their **happines**s on a **scale from 1 to 10**.
- Your **independent variable (income)** and **dependent variable (happiness)** are both **quantitative**, so you can do a regression analysis to see **if there is a linear relationship** between them.

Assumptions of a Simple Linear Regression

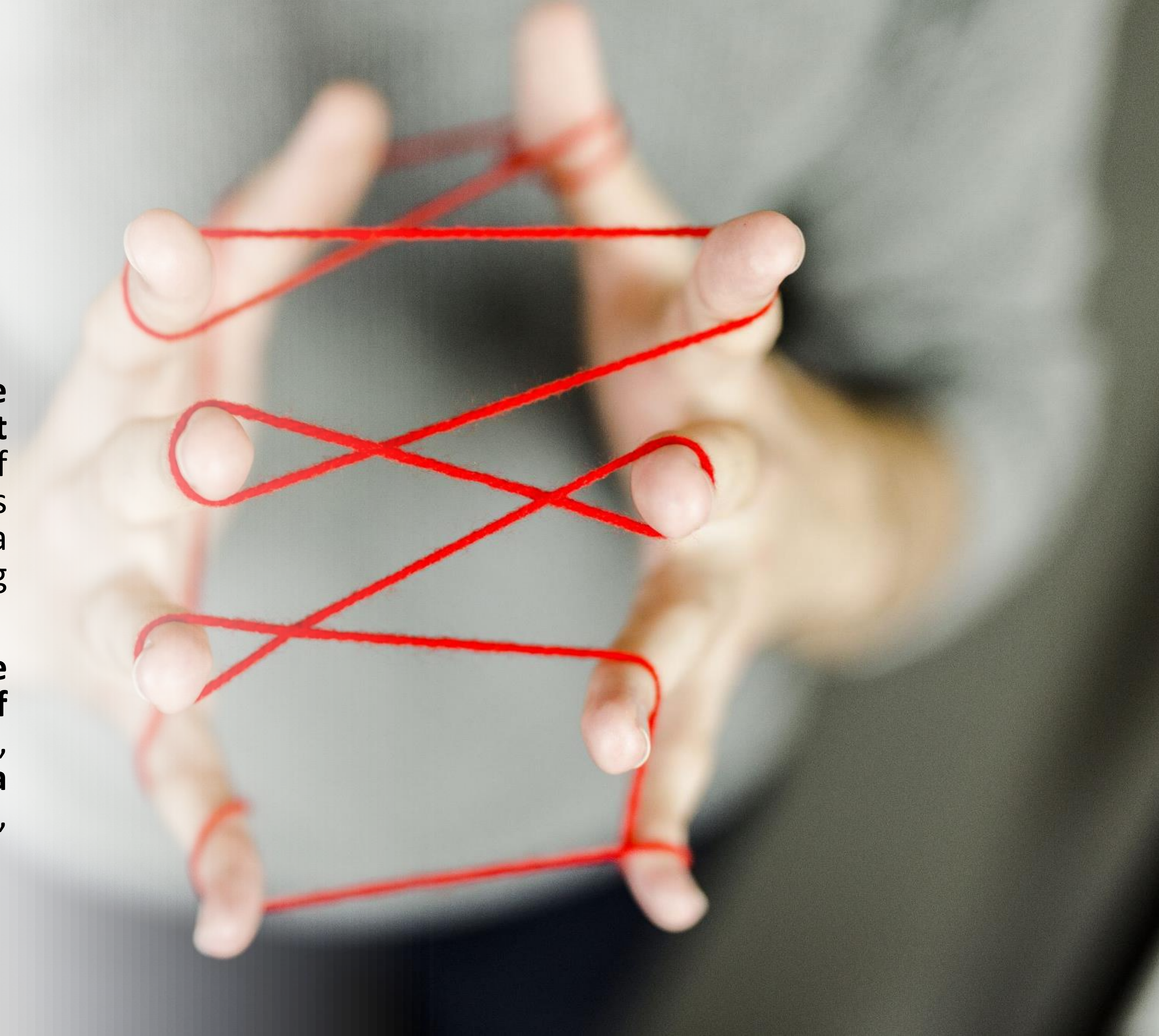
- Simple linear regression is a **parametric test**, meaning that it **makes certain assumptions about the data**.

These assumptions are:

- **Homogeneity of variance (homoscedasticity):** Size of the error in our prediction doesn’t change significantly across the values of the independent variable.
- **Independence of observations:** the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
- **Normality:** The data follows a normal distribution.

Additional Assumption about Simple Linear Regression

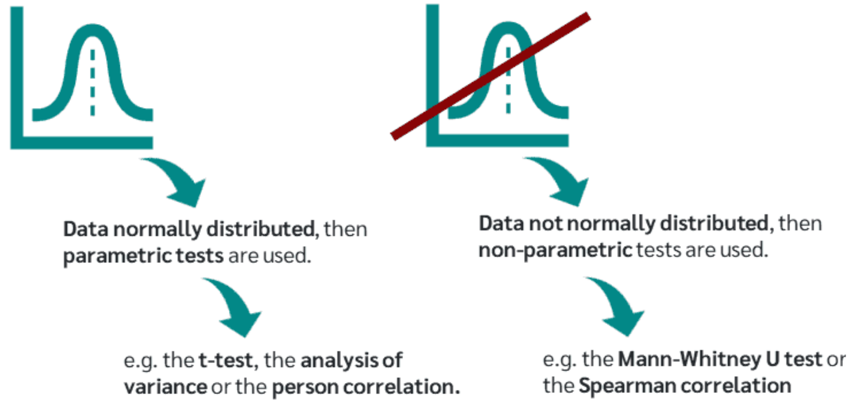
- The **relationship between the independent and dependent variable is linear**: the line of **best fit** through the data points is a **straight line** (rather than a curve or some sort of grouping factor).
- If your data **do not meet the assumptions of homoscedasticity or normality**, you may be able to **use a nonparametric test instead**, such as the Spearman rank test.



Example when Data does not meet the Assumptions

- **You think there is a linear relationship** between cured meat consumption and the incidence of colorectal cancer in the U.S.
- However, you find that **much more data has been collected at high rates of meat consumption than at low rates of meat consumption**, with the result that there is much more variation in the estimate of cancer rates at the low range than at the high range.
- Because the data **violate the assumption of homoscedasticity**, it **doesn't work for regression**, but you **perform a nonparametric Spearman rank test** instead.

Simplified:



Data normally distributed, then **parametric tests** are used.

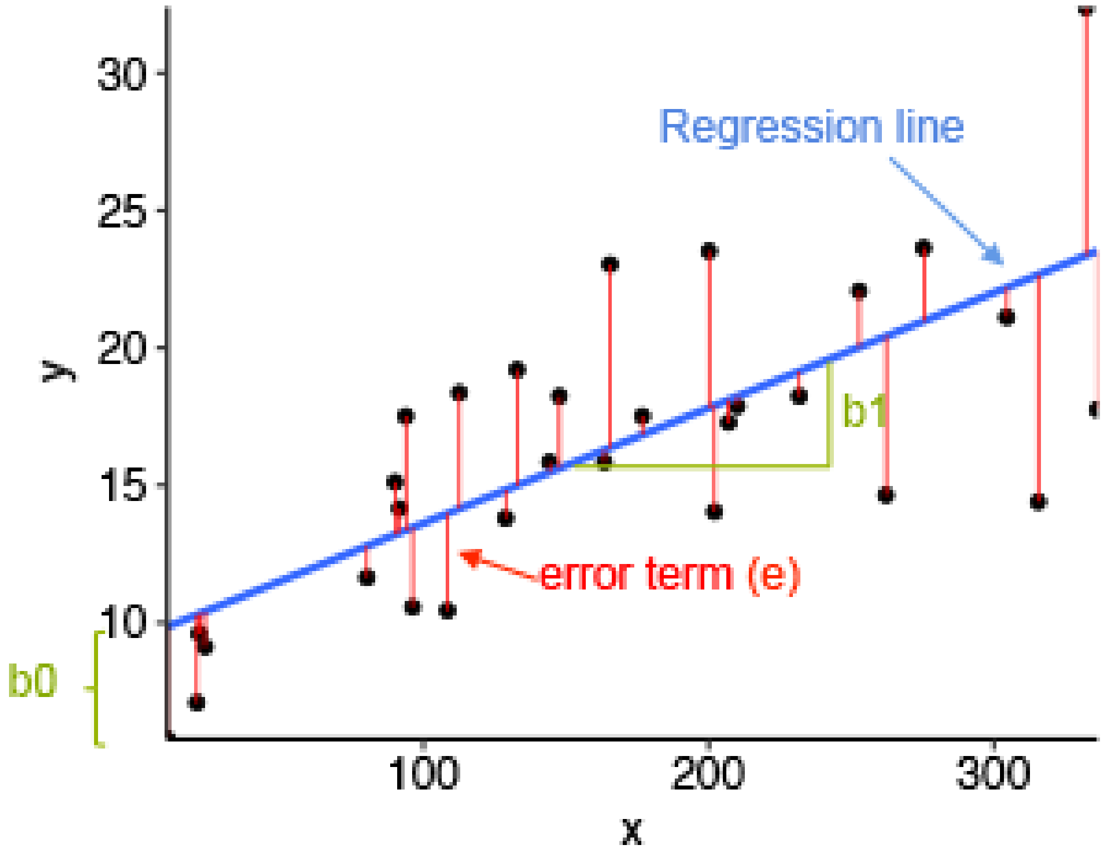
e.g. the **t-test**, the **analysis of variance** or the **person correlation.**

Data not normally distributed, then **non-parametric** tests are used.

e.g. the **Mann-Whitney U test** or the **Spearman correlation**

The Regression Line

- A straight line that describes how a response variable y changes as an explanatory variable x changes.
- A regression line can be used to predict the value of y for a given value of x.
- The line of best fit



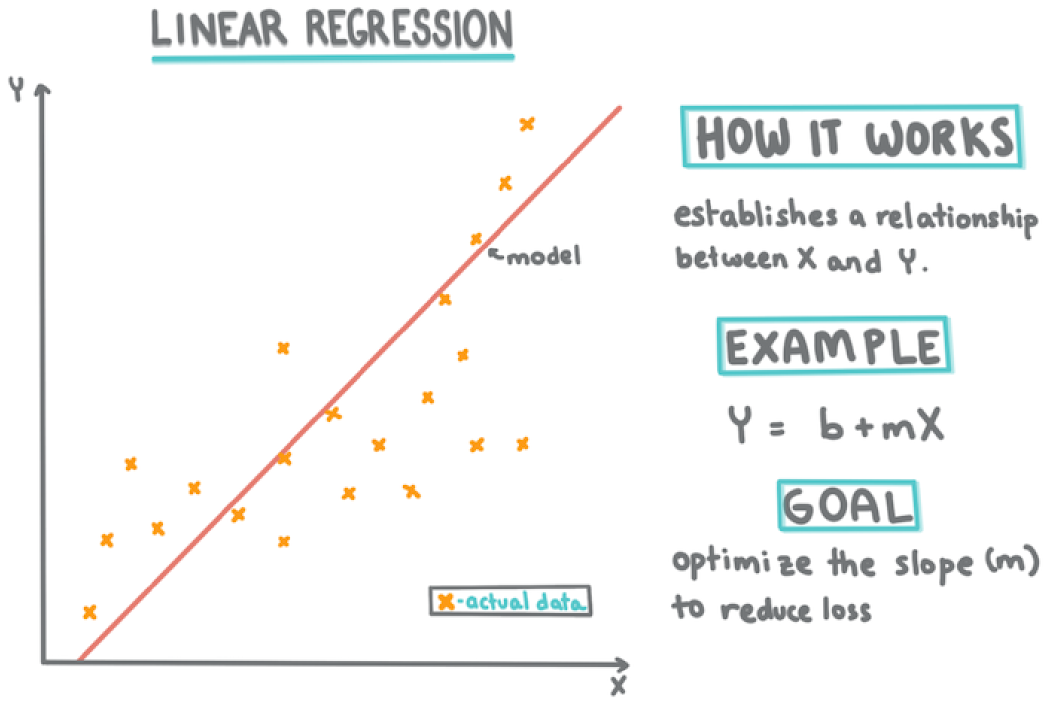
Simple Linear Regression

- In simplest form, linear regression involves only one independent variable (y) and one dependent variable (X).
- The equation for simple linear regression is:

$$y = b + mX$$

where:

- y is the dependent variable or response
- X is the independent variable or predictor
- m is the estimated slope
- b is the estimated y-intercept, which represents y (X = 0)



establishes a relationship between X and Y.

optimize the slope (m) to reduce loss

Solving a Simple Linear Regression

Finding m

- This is not identified randomly.
- Chosen to minimize the sum of the squared differences between the observed values of the dependent variable and the values predicted by the regression line.
- Involves calculating the residuals (the differences between observed and predicted values) and adjusting the slope until the sum of the squared residuals is minimized.

Finding m

- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- n is the number of data points
- y_i is the **observed value** of the **dependent variable** for i^{th} data point.
- $\hat{y}_i$ is the **predicted value** of the **dependent variable** for the i^{th} data point.

- To find m:
- $m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\bar{x}$ = mean of the independent variable.
- $\bar{y}$ = mean of the dependent variable.

Calculating the mean of x and y

- Example:
- $x = (x_1, x_2, x_3 \cdots x_n)$
- $\bar{x} = \frac{(x_1, x_2, x_3 \cdots x_n)}{n}$
- Same goes for $\bar{y}$
- The mean is just, the average…

Determining b like m

- Like m, its not identified randomly, but systematically.
- We can use the mean values of the dependent and independent variables.
- $\bar{x}$ = mean of the independent variable.
- $\bar{y}$ = mean of the dependent variable.
- $m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- x is the independent variable or predictor

- $b_0 = \bar{y} - m\bar{x}$

Finding the $\hat{y}$

- To find $\hat{y}$:
- $\hat{y}_i$=$mx_i + b$
- Where:
- x_i is the **value** of the **independent variable** for the i^{th} data point.

- These determines $m.$ Once identified, we can now identify the y-intercept b using the mean values of x and y.

Example

- Let's consider a hypothetical example involving the relationship between the number of years of experience (independent variable x) and salary (dependent variable y) for employees in a certain profession.
- We want to build a simple linear regression model to predict salary based on years of experience.
- Data:

x	y
2	45000
3	50000
5	60000
7	70000
8	75000

- Calculate the mean values $\bar{x}$ and $\bar{y}$ → solve for m → solve for b → solve for y → solve for $\hat{y}$ → use the regression line equation with an outside data x_n with $y = 65000$.

Solution

- Data

x	y
2	45000
3	50000
5	60000
7	70000
8	75000

$\boldsymbol{n = 5}$

$$\bar{x} = \frac{2+3+5+7+8}{5} = \frac{25}{5} = \mathbf{5}$$

$$\boldsymbol{\bar{x} = 5}$$

$$\bar{y} = \frac{45000+50000+60000+70000+75000}{5} = \frac{300000}{5} = 60000$$

$$\boldsymbol{\bar{y} = 60000}$$

$$\boldsymbol{m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$m = \frac{(2-5)(45000-60000)+(3-5)(50000-60000)+(5-5)(60000-60000)+(7-5)(70000-60000)+(8-5)(75000-60000)}{(2-5)^2+(3-5)^2+(5-5)^2+(7-5)^2+(8-5)^2}$$

$$= \frac{(-3)(-15000)+(-2)(-10000)+(0)(0)+(2)(10000)+(3)(15000)}{9+4+0+4+9}$$

$$= \frac{130000}{26}$$

$$\boldsymbol{m = 5000}$$

$$\boldsymbol{b = \bar{y} - m\bar{x}}$$

$$= 60000 - (5000)(5)$$

$$= 60000 - 25000 = 35000$$

$$\boldsymbol{b = 35000}$$

$$\boldsymbol{y = mx + b}$$

$$= (5000)(x) + 35000$$

$$\boldsymbol{y = 5000x + 35000}$$

x	$\boldsymbol{\hat{y}}$
6	65000

$$\hat{y} = 5000(6) + 35000$$

$$= 30000 + 35000$$

$$\boldsymbol{\hat{y} = 65000}$$

$$\boldsymbol{n = 5}$$

$$\boldsymbol{\bar{x} = 5}$$

$$\boldsymbol{\bar{y} = 60000}$$

$$\boldsymbol{m = 5000}$$

$$\boldsymbol{b = 35000}$$

$$\boldsymbol{y = 5000x + 35000}$$

$$\boldsymbol{\hat{y} = 65000}$$

Looking back at the equation

- $y = mx + b$ Linear Algebra style
- $y = b_0 + b_1x$ Statistics style
- $slope = m = b_1$
- $y - intercept = b = b_0$

Using the Least Squares Method: Ordinary Least Squares

Least Squares Method

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

- Where:
- y = dependent variable
- x = independent variable
- m = slope
- b = y-intercept

Example

- Write a linear equation that “best fits” the data in the table.

x	y
1	1.5
2	3.8
3	6.7
4	9.0
5	11.2
6	13.6
7	16

Taking the Sums

$\boldsymbol{x}$	y	xy	$\boldsymbol{x^2}$
1	1.5	1.5	1
2	3.8	7.6	4
3	6.7	20.1	9
4	9.0	36	16
5	11.2	56	25
6	13.6	81.6	36
7	16	112	49

$\sum x = 28$ $\sum y = 61.8$ $\sum xy = 314.8$ $\sum x^2 = 140$

Calculating the slope m

- $m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{7(314.8)-(28)(61.8)}{7(140)-(28)^2}$
- $m = \frac{473.2}{196} = 2.4142857$
- $m = 2.4142857$
- Do not round, it will affect the value of b

$\sum x = 28$ $\sum y = 61.8$ $\sum xy = 314.8$ $\sum x^2 = 140$

Calculating the y-intercept or b

- $b = \frac{\sum y - m \sum x}{n} = \frac{61.8 - (2.4142857)(28)}{7}$
- $b = -0.828571$

$m = 2.4142857$

$n = 7$

$\sum x = 28$

$\sum y = 61.8$

Solving for y

- $y = mx + b$
- We can now round the m to 2.41.
- We can now round the b to -0.83.
- Plug-in the values
- $y = 2.41x - 0.83$
- This is now the approximation for the data.

$m = 2.4142857 round\ to\ 2.41$

$b = -0.828571$

Checking the approximation with the data

- $x_2 = 2$
- $\hat{y} = 2.41(x_2) - 0.83$
- $\hat{y} = 2.41(2) - 0.83$
- $\hat{y} = 3.99$
- $y_2 = 3.8$
- $\hat{y} = 3.99$

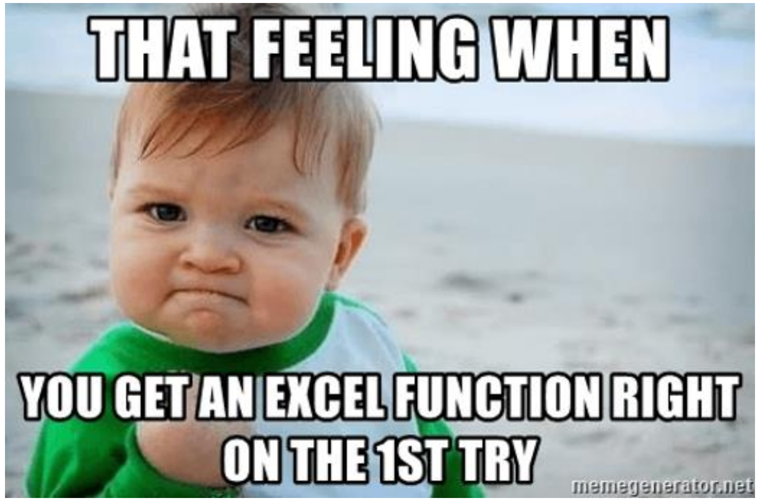
$$x_5 = 5$$
$$\hat{y} = 2.41(x_5) - 0.83$$
$$\hat{y} = 2.41(5) - 0.83$$
$$\hat{y} = 11.22$$
$$y_5 = 11.2$$
$$\hat{y} = 11.22$$

$$x_7 = 7$$
$$\hat{y} = 2.41(x_7) - 0.83$$
$$\hat{y} = 2.41(7) - 0.83$$
$$\hat{y} = 16.04$$
$$y_7 = 16$$
$$\hat{y} = 16.04$$

Doing it in excel

- SLOPE() function
- INTERCEPT() function

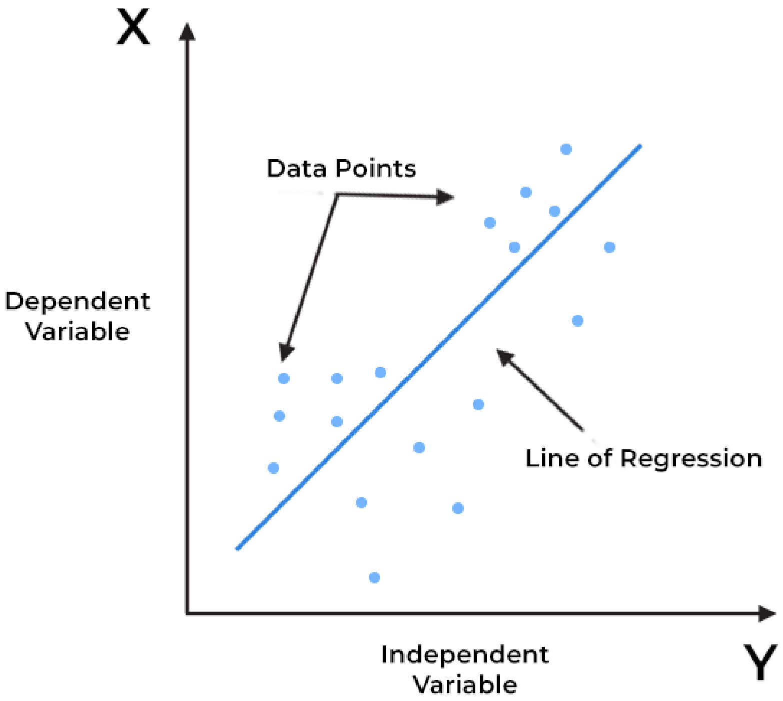
The one and only commandment
Thou shalt not use excel for regression analysis
@economist_memes

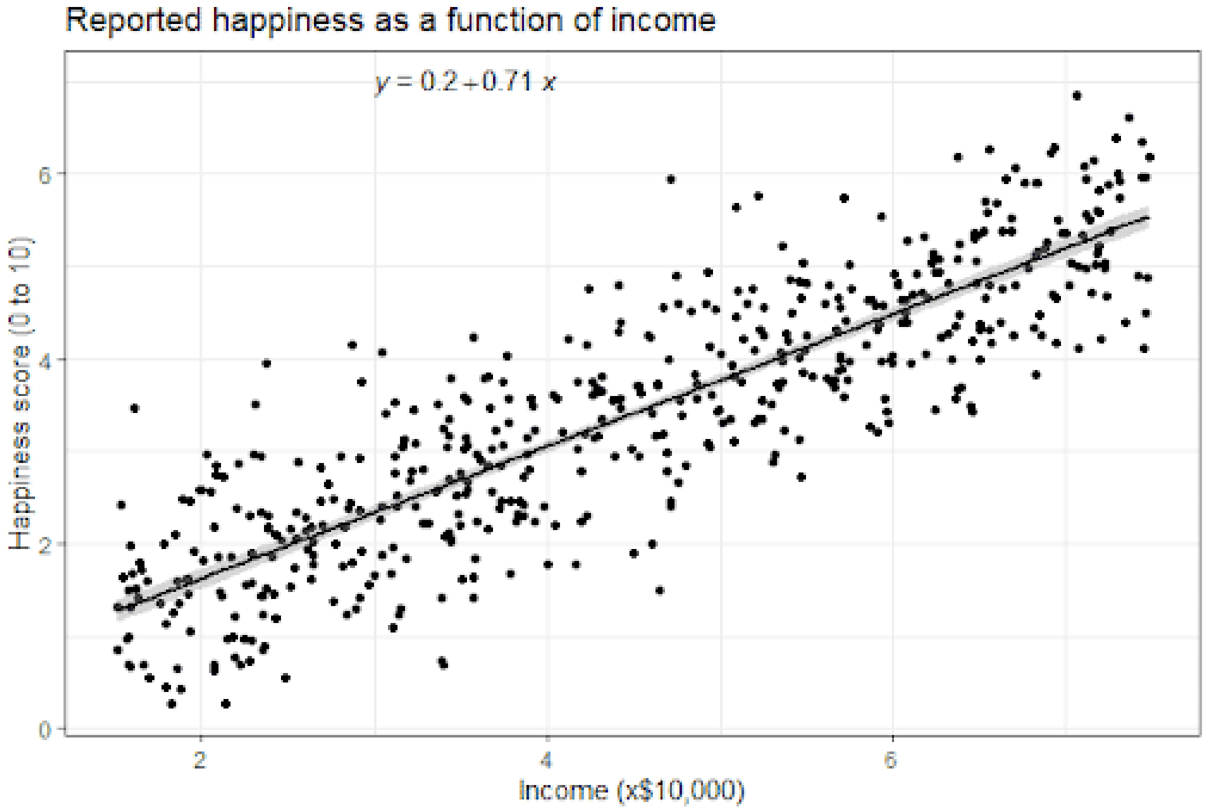


Key thing to remember about the example!

- Its just an example!
- This does not exactly apply in reality.
- Other factors can come at play that must be considered.
- These amount of data is not realistic.
- Calculating real-world data will require a statistical software or an automated approach.
- We will not do things by-hand >.<
- OLS is a common approach in training a regression model and a way to determine the possible best fit.
- Though we don’t do it by hand, its important to still know how it works and how things go!

The Linear Regression in Plot





Simple Linear Regression Example

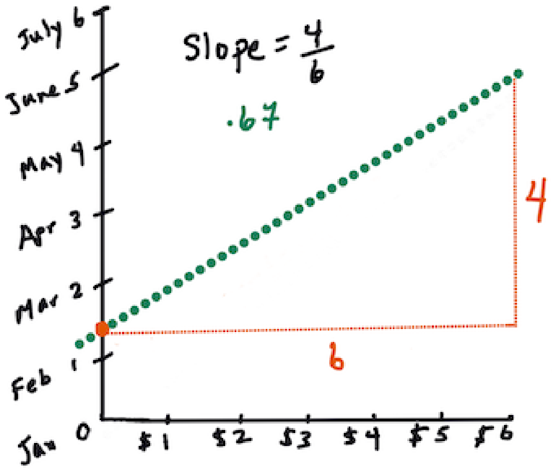
$y = a + bX$

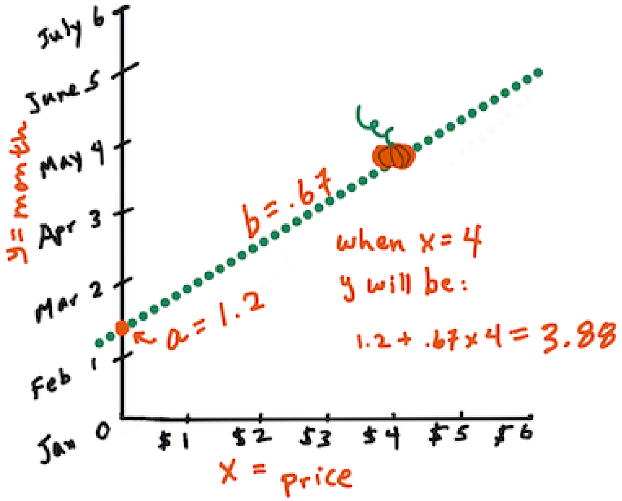
y is the dependent variable or response = ?

a is the estimated intercept = 1.2

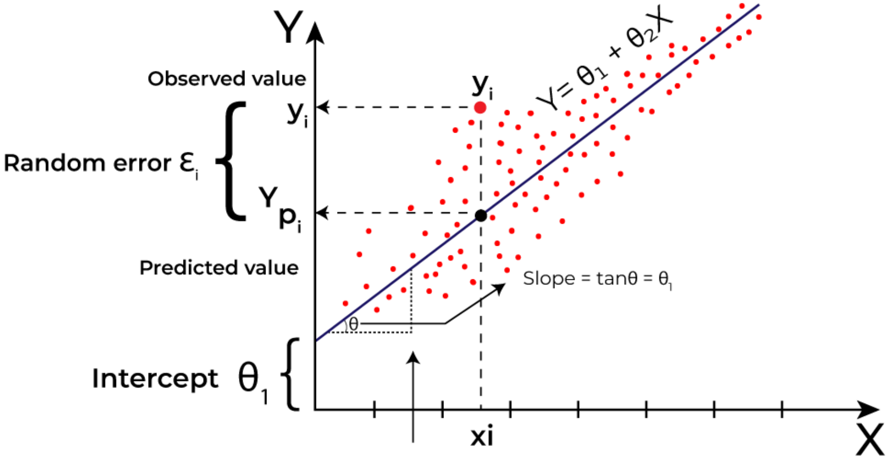
b is the estimated slope = 0.67

X is the independent variable or predictor = 4





Linear Regression Example



- Perform the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression.
- In the figure, X (input) is the **work experience** and Y (output) is the **salary of a person**. The regression line is the best-fit line for our model.
- We utilize the cost function to compute the best values to get the best fit line since different values for weights or the coefficient of lines result in different regression lines.

Before using Linear Regression

- Determine a relation between the variables in the dataset.
- A simple linear regression only uses a single variable, having more requires a multiple linear regression.

Example:

- Buying a house. The area of a house can be x (independent variable).
- Having only the area of the house x, a simple linear regression is applicable to predict the price of the house.
- The price of the house is y.
- But know, determining the price of a house may require more variables like the location, number of rooms, and date of purchase, making multiple linear regression a better option.

Implementing a Simple Linear Regression

- While you can perform a linear regression by hand, this is a tedious process, specifically when finding the best fit.
- Using statistical programs or tools are still the go to process.

Mini Quiz

- A researcher wants to perform a simple linear regression to find out if the socio-economic status of a teacher can predict whether they work at a primary or a secondary school. Why can’t this be done?

A. Because there are not enough variables for the analysis

B. Because socio-economic status can not be used as a predictor variable

C. Because the outcome variable is nominal not continuous

Mini Quiz

- A researcher wants to perform a simple linear regression to find out if the socio-economic status of a teacher can predict whether they work at a primary or a secondary school. Why can’t this be done?

A. Because there are not enough variables for the analysis (This is incorrect because simple linear regression only requires one predictor variable and one outcome variable.)

B. Because socio-economic status can not be used as a predictor variable (This is incorrect because socio-economic status can be used as a predictor variable, but it's not the issue in this case.)

C. Because the outcome variable is nominal not continuous

Mini Quiz

- The slope (b) in the prediction equation for Simple Linear Regression represents which of the following?

A. The value of $\hat{y}$ (y-hat) when x = 1.5.

B. The amount of change in the independent variable when the dependent variable increases by one unit.

C. The angle of the best fitting diagonal line when x = 2.

D. The amount of change in the dependent variable when the independent variable increases by one unit.

Mini Quiz

- The slope (b) in the prediction equation for Simple Linear Regression represents which of the following?

A. The value of $\hat{y}$ (y-hat) when x = 1.5. (This statement describes the predicted value of the dependent variable when the independent variable is 1.5, not the slope itself.)

B. The amount of change in the independent variable when the dependent variable increases by one unit. (This concept is reversed. It talks about the change in x for a change in y, not the other way around.)

C. The angle of the best fitting diagonal line when x = 2. (This statement is true in a way, but the slope describes the general "angle" of the line, not just at a specific x value.)

D. The amount of change in the dependent variable when the independent variable increases by one unit.

Mini Quiz

- What is a straight line that attempts to predict the relationship between two points, also known as a trend line or line of best fit?

- A. Slope
- B. Regression Line
- C. Intercept
- D. Const Function

Mini Quiz

- What is a straight line that attempts to predict the relationship between two points, also known as a trend line or line of best fit?

- A. Slope (Slope measures how steep a line is. While important, it doesn't fully describe a line's predictive ability.)
- B. Regression Line
- C. Intercept (Where the regression line crosses the y-axis. It's a part of the regression line equation but doesn't fully explain the line's predictive power.)
- D. Cost Function (This measures how well a model fits the data by calculating the difference between predicted and actual values. While important, it's not the same as the straight line we're looking for in the question.)

Mini Quiz

- What is the mathematical expression depicting the relationship between the independent variable x and the dependent variable y within the framework of simple linear regression, with the options:

- A. $y = mX + B$
- B. $y = m^2 + Bx$
- C. $y = B_x + m^2$
- D. $y = B_m + x$

- m: Represents the rate of change or slope of the line
- x: Represents the independent variable
- B: Represents the y-intercept, the value of y when x is zero

Mini Quiz

- What is the mathematical expression depicting the relationship between the independent variable x and the dependent variable y within the framework of simple linear regression, with the options:

A. $y = mX + B$

B. $y = m^2 + Bx$ It doesn't represent a linear relationship between y and x, and m is squared, which is not linear. B is also not multiplied by x, which does not align with the y-intercept.

C. $y = B_x + m^2$ Does not follow the standard. The m^2 does not represent a slope and B is multiplied by x that does not align with the y-intercept.

D. $y = B_m + x$ B is multiplied by m, there is no term for the slope m multiplied by x.

- m: Represents the rate of change or slope of the line
- x: Represents the independent variable
- B: Represents the y-intercept, the value of y when x is zero

Multiple Linear Regression

What is Multiple Linear Regression

- Used to estimate the relationship between **two or more independent variables** and **one dependent variable**. You can use multiple linear regression when you want to know:

1. 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).

2. 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).

Example of Multiple Linear Regression

- You are a public health researcher interested in social factors that influence heart disease. You survey 500 towns and gather data on the percentage of people in each town **who smoke (x)**, the percentage of people in each town **who bike to work (x)**, and the percentage of people in each town **who have heart disease (y)**.

- Because you have **two independent variables (who smoke and bike to work)** and one **dependent variable (with heart disease)**, and all your variables are **quantitative**, you can use multiple linear regression to analyze the relationship between them.

Assumptions for Multiple Linear Regression

- For Multiple Linear Regression, all four of the assumptions from Simple Linear Regression apply. In addition to this, below are few more:

- No multicollinearity
- Additivity
- Feature Selection
- Overfitting

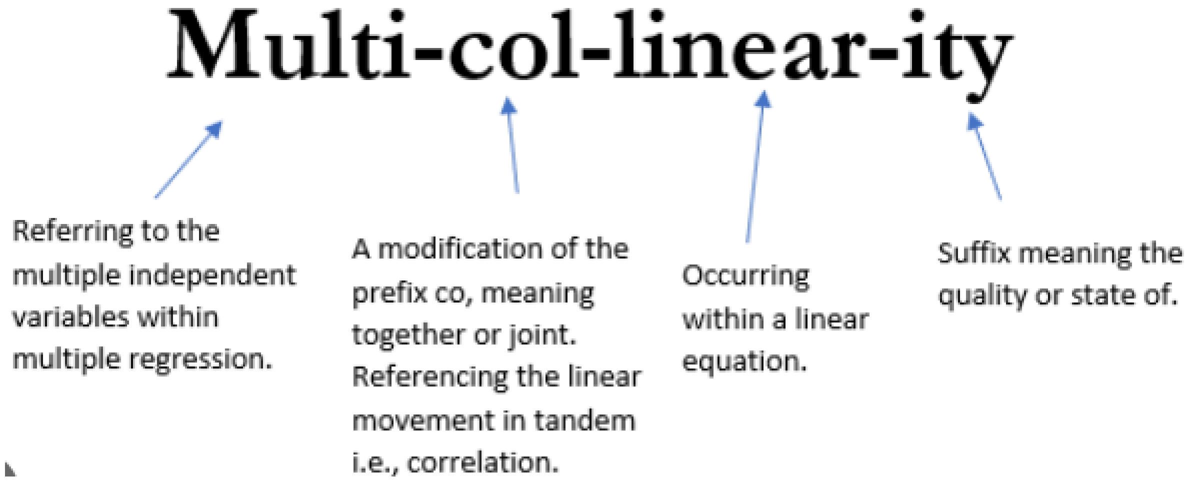
No multicollinearity

No high correlation between the independent variables.

Indicates that there is little or no correlation between the independent variables.

Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable.

If there is multicollinearity, then multiple linear regression will not be an accurate model.



Multi-col-linear-ity

Referring to the multiple independent variables within multiple regression.

A modification of the prefix co, meaning together or joint. Referencing the linear movement in tandem i.e., correlation.

Occurring within a linear equation.

Suffix meaning the quality or state of.

Additivity

- The model assumes that the effect of changes in a predictor variable on the response variable is consistent regardless of the values of the other variables.
- This assumption implies that there is no interaction between variables in their effects on the dependent variable.

Feature Selection

- In multiple linear regression, it is essential to carefully select the independent variables that will be included in the model. Including irrelevant or redundant variables may lead to overfitting and complicate the interpretation of the model.

Model Overfitting

- Occurs when the model fits the training data too closely, capturing noise or random fluctuations that do not represent the true underlying relationship between variables.
- Can lead to poor generalization performance on new, unseen data.

Solving Multiple Linear Regression

Multiple Linear Regression

- Involves **>1 independent variable** and **one dependent variable**.
- The equation for multiple linear regression is:
- $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \epsilon$

- where:
- y is the predicted value of the **dependent variable**
- β_0 is the **$\boldsymbol{y}$-intercept** (value of $\boldsymbol{y}$ when all other independent variables are 0)
- β_1, β_2,…, β_n are the **regression coefficients** of the independent variable or **the slopes**
- X_1, X_2,…, X_n are the **independent variables**
- ϵ the **error term** or **residual**, representing the unexplained variation in the dependent variable $\boldsymbol{y}$.

- The equation **captures the relationship between the dependent variable $\boldsymbol{y}$ and multiple independent variables**, with β_0 representing the intercept and β_1, β_2,…, β_n representing the slopes or coefficients associated with each independent variable.
- The interpretation of the regression coefficients depends on the units of the independent variables X_1, X_2,…, X_n, indicating the change in the dependent variable y for a one-unit change in the corresponding independent variable, holding all other variables constant.

Sample use of Multiple Linear Regression

- Predicting sales based on the money spent on TV, Radio, and Newspaper for marketing.
- In this case, there are three independent variables, i.e., **money spent on TV, Radio ($\boldsymbol{x_1}$, $\boldsymbol{x_2}$)**, and **Newspaper for marketing ($\boldsymbol{x_3}$)**, and one dependent variable, i.e., **sales ($\boldsymbol{y}$)**, that is the value to be predicted.

Solving a multiple linear regression problem

- We will only have to independent variable, as it can be VERY LENGTHY!!!
- $\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$
- $\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$
- $\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$

Important Equations to remember

- The equations will only differ based on the number of independent variables. Here we only have 2.

- $\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$
- $\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y)-(\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2)-(\sum x_1 x_2)^2}$
- $\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y)-(\sum x_1 x_2)(x_1 y)}{(\sum x_1^2)(\sum x_2^2)-(\sum x_1 x_2)^2}$

$$\sum x_1^2 = \sum x_1^2 - \frac{(\sum x_1)^2}{n}$$

$$\sum x_2^2 = \sum x_2^2 - \frac{(\sum x_2)^2}{n}$$

$$\sum x_1 y = \sum x_1 y - \frac{\sum x_1 \sum y}{n}$$

$$\sum x_2 y = \sum x_2 y - \frac{\sum x_2 \sum y}{n}$$

$$\sum x_1 x_2 = \sum x_1 x_2 - \frac{\sum x_1 x_2}{n}$$

The data

y	x_1	x_2	x_1^2	x_2^2	x_1y	x_2y	x_1x_2
64	57	8	3249	64	3648	512	456
71	59	10	3481	100	4189	710	590
53	49	6	2401	36	2597	318	294
67	62	11	3844	121	4154	737	682
55	51	8	2601	64	2805	440	408
58	50	7	2500	49	2900	406	350
77	55	10	3025	100	4235	770	550
57	48	9	2304	81	2736	513	432
502	**431**	**69**	**23405**	**615**	**27264**	**4406**	**3762**

$$\sum y = 502$$

$$\sum x_1 = 431$$

$$\sum x_2 = 69$$

$$\sum x_1^2 = 23405$$

$$\sum x_2^2 = 615$$

$$\sum x_1 y = 27264$$

$$\sum x_2 y = 4406$$

$$\sum x_1 x_2 = 3762$$

Solving

$$\sum \boldsymbol{x_1^2} = \sum \boldsymbol{x_1^2} - \frac{(\sum \boldsymbol{x_1})^2}{\boldsymbol{n}}$$

- $23405 - \frac{(431)^2}{8}$
- $23405 - 23220.125$
- $\sum \boldsymbol{x_1^2} = \mathbf{184.875}$

$$\sum \boldsymbol{x_2^2} = \sum \boldsymbol{x_2^2} - \frac{(\sum \boldsymbol{x_2})^2}{\boldsymbol{n}}$$

- $615 - \frac{(69)^2}{8}$
- $615 - 595.125$
- $\sum \boldsymbol{x_2^2} = \mathbf{19.875}$

$$\sum \boldsymbol{x_1 y} = \sum \boldsymbol{x_1 y} - \frac{\sum \boldsymbol{x_1} \sum \boldsymbol{y}}{\boldsymbol{n}}$$

- $27264 - \frac{(431 \times 502)}{8}$
- $27264 - 27045.25$
- $\sum \boldsymbol{x_1 y} = \mathbf{218.75}$

$$\sum \boldsymbol{x_2 y} = \sum \boldsymbol{x_2 y} - \frac{\sum \boldsymbol{x_2} \sum \boldsymbol{y}}{\boldsymbol{n}}$$

- $4406 - \frac{(69 \times 502)}{8}$
- $4406 - 4329.75$
- $\sum \boldsymbol{x_2 y} = \mathbf{76.25}$

$$\sum \boldsymbol{x_1 x_2} = \sum \boldsymbol{x_1 x_2} - \frac{\sum \boldsymbol{x_1 x_2}}{\boldsymbol{n}}$$

- $3762 - \frac{(431 \times 69)}{8}$
- $3762 - 3717.375$
- $\sum \boldsymbol{x_1 x_2} = \mathbf{44.625}$

$$\sum y = 502$$
$$\sum x_1 = 431$$
$$\sum x_2 = 69$$
$$\sum x_1^2 = 23405$$
$$\sum x_2^2 = 615$$
$$\sum x_1 y = 27264$$
$$\sum x_2 y = 4406$$
$$\sum x_1 x_2 = 3762$$

Summary of values

- $\sum x_1^2 = 184.875$
- $\sum x_2^2 = 19.875$
- $\sum x_1 y = 218.75$
- $\sum x_2 y = 76.25$
- $\sum x_1 x_2 = 44.625$

Solving for $\beta_1, \beta_2, \beta_0$

$$\boldsymbol{\beta_1 = \frac{\left(\sum x_2^2\right)\left(\sum x_1 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_2 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2}}$$

- $\beta_1 = \frac{(19.875)(218.75)-(44.625)(76.25)}{(184.875)(19.875)-(44.625)^2}$
- $\frac{4347.656-3402.656}{3674.390-199.391}$
- $\boldsymbol{\beta_1 = 0.5615}$

$$\boldsymbol{\beta_2 = \frac{\left(\sum x_1^2\right)\left(\sum x_2 y\right) - \left(\sum x_1 x_2\right)\left(x_1 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2}}$$

- $\beta_2 = \frac{(184.875)(76.25)-(44.625)(218.75)}{(184.875)(19.875)-(44.625)^2}$
- $\frac{14096.719-9761.719}{3674.390-199.391}$
- $\boldsymbol{\beta_2 = 2.5758}$

$$\boldsymbol{\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2}$$

- $\beta_0 = \frac{502}{8} - \left(0.5615 \times \frac{431}{8}\right) - \left(2.57 \times \frac{69}{8}\right)$
- $62.75 - 30.2508 - 22.2163$
- $\boldsymbol{\beta_0 = 10.2829}$

$$\boldsymbol{\sum x_1^2 = 184.875}$$

$$\boldsymbol{\sum x_2^2 = 19.875}$$

$$\boldsymbol{\sum x_1 y = 218.75}$$

$$\boldsymbol{\sum x_2 y = 76.25}$$

$$\boldsymbol{\sum x_1 x_2 = 44.625}$$

Plugging the values

- $\boldsymbol{\beta_0 = 10.2829}$
- $\boldsymbol{\beta_1 = 0.5615}$
- $\boldsymbol{\beta_2 = 2.5758}$

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

- $y = \mathbf{10.2829} + \mathbf{0.5615} x_1 + \mathbf{2.5758} x_2 + \epsilon$

Multiple Linear Regression with 3 independent variables

By Hand Multiple Linear Regression Example

Step 1: Calculate $X_1{}^2$, $X_2{}^2$, $X_3{}^2$ X_1y, X_2y, X_3y and $X_1X_2X_3$

y=Sales	x_1=TV	x_2=Radio	x_3=Newspaper
22	230	38	69
10	45	39	45
9	17	46	69

x_1^2	x_2^2	x_3^2	x_1y	x_2y	x_3y	$x_1x_2x_3$
52900	1444	4761	5060	836	1518	603060
2025	1521	2025	450	390	450	78975
289	2116	4761	153	414	621	53958
			SUM			
55214	5081	11547	5663	1640	2589	735993

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

By Hand Multiple Linear Regression Example

- Mean values

	y=Sales	**x_1=TV**	**x_2=Radio**	**x_3=Newspaper**
	22	230	38	69
	10	45	39	45
	9	17	46	69
Mean	**13.667**	**97.33**	**41**	**61**
Sum	41	292	123	183

By Hand Multiple Linear Regression Example

- Step 2: Calculate Regression Sums.
- $\sum x_1^2 = \sum x_1^2 - \frac{(\sum x_1)^2}{n} = 55214 - \frac{(292)^2}{3} = 26792.67$
- $\sum x_2^2 = \sum x_2^2 - \frac{(\sum x_2)^2}{n} = 5081 - \frac{(123)^2}{3} = 38$
- $\sum x_3^2 = \sum x_3^2 - \frac{(\sum x_3)^2}{n} = 11547 - \frac{(183)^2}{3} = 384$
- $\sum x_1 y = \sum x_1 y - \frac{(\sum x_1 \sum y)}{n} = 5663 - \frac{(292*41)}{3} = 1672.33$
- $\sum x_2 y = \sum x_2 y - \frac{(\sum x_2 \sum y)}{n} = 1640 - \frac{(123*41)}{3} = -41$
- $\sum x_3 y = \sum x_3 y - \frac{(\sum x_3 \sum y)}{n} = 2589 - \frac{(183*41)}{3} = 88$
- $\sum x_1 x_2 x_3 = \sum x_1 x_2 x_3 - \frac{(\sum x_1 \sum x_2 \sum x_3)}{n} = 735993 - \frac{(292*123*183)}{3} = -1454883$

By Hand Multiple Linear Regression Example

- Step 3: Calculate b0, b1, and b2.

- $\sum x_1^2$,$\sum x_2^2$,$\sum x_3^2$ the sum of squares of the respective predictor variables.
- $\sum x_1 x_2 \sum x_1 x_3 \sum x_2 x_3$ are the sums of products of pairs of predictor variables.
- $\sum x_1 x_2 x_3$ is the sum of the products of all three predictor variables.
- $\sum x_1 y, \sum x_2 y \sum x_3 y$ are the sums of products of each predictor variable with the response variable.

- $b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - b_3 \bar{x}_3$

By Hand Multiple Linear Regression Example

$$A = \sum x_1^2 \cdot \sum x_2^2 \cdot \sum x_3^2$$
$$B = \sum x_1 x_2 x_3$$
$$C = \sum x_1 x_2 \cdot \sum x_3^2$$
$$D = \sum x_1 x_3 \cdot \sum x_2^2$$
$$E = \sum x_2 x_3 \cdot \sum x_1^2$$

$$b_1 = \frac{A+B^2\cdot\sum x_3^2+C^2\cdot\sum x_2^2+D^2\cdot\sum x_1^2-B^2\cdot\sum x_3^2-D^2\cdot\sum x_2^2-E^2\cdot\sum x_1^2}{A+B^2\cdot\sum x_3^2+C^2\cdot\sum x_2^2+D^2\cdot\sum x_1^2-B^2\cdot\sum x_3^2-D^2\cdot\sum x_2^2-E^2\cdot\sum x_1^2}$$

$$A = \sum x_1^2 \cdot \sum x_2^2 \cdot \sum x_3^2$$
$$B = \sum x_1 x_2 x_3$$
$$C = \sum x_1 x_2 \cdot \sum x_3^2$$
$$D = \sum x_1 x_3 \cdot \sum x_2^2$$
$$E = \sum x_2 x_3 \cdot \sum x_1^2$$

$$b_2 = \frac{A+B^2\cdot\sum x_1^2+C^2\cdot\sum x_3^2+D^2\cdot\sum x_2^2-B^2\cdot\sum x_1^2-D^2\cdot\sum x_3^2-E^2\cdot\sum x_2^2}{A+B^2\cdot\sum x_1^2+C^2\cdot\sum x_3^2+D^2\cdot\sum x_2^2-B^2\cdot\sum x_1^2-D^2\cdot\sum x_3^2-E^2\cdot\sum x_2^2}$$

$$A = \sum x_1^2 \cdot \sum x_2^2 \cdot \sum x_3^2$$
$$B = \sum x_1 x_2 x_3$$
$$C = \sum x_1 x_2 \cdot \sum x_3^2$$
$$D = \sum x_1 x_3 \cdot \sum x_2^2$$
$$E = \sum x_2 x_3 \cdot \sum x_1^2$$

$$b_3 = \frac{A+B^2\cdot\sum x_1^2+C^2\cdot\sum x_2^2+D^2\cdot\sum x_3^2-B^2\cdot\sum x_1^2-C^2\cdot\sum x_2^2-E^2\cdot\sum x_3^2}{A+B^2\cdot\sum x_1^2+C^2\cdot\sum x_2^2+D^2\cdot\sum x_3^2-B^2\cdot\sum x_1^2-C^2\cdot\sum x_2^2-E^2\cdot\sum x_3^2}$$

By Hand Multiple Linear Regression Example

- Step 5: Place b0, b1, b2, and b3 in the estimated multiple linear regression equation.

- $\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

Finding the best fit-line

- $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \epsilon$

- Identifying the values for multiple regression is not the same as a simple linear regression.
- Finding the best-fit or regression lines in multiple linear regression are done via Gradient Descent or Normal Equation.

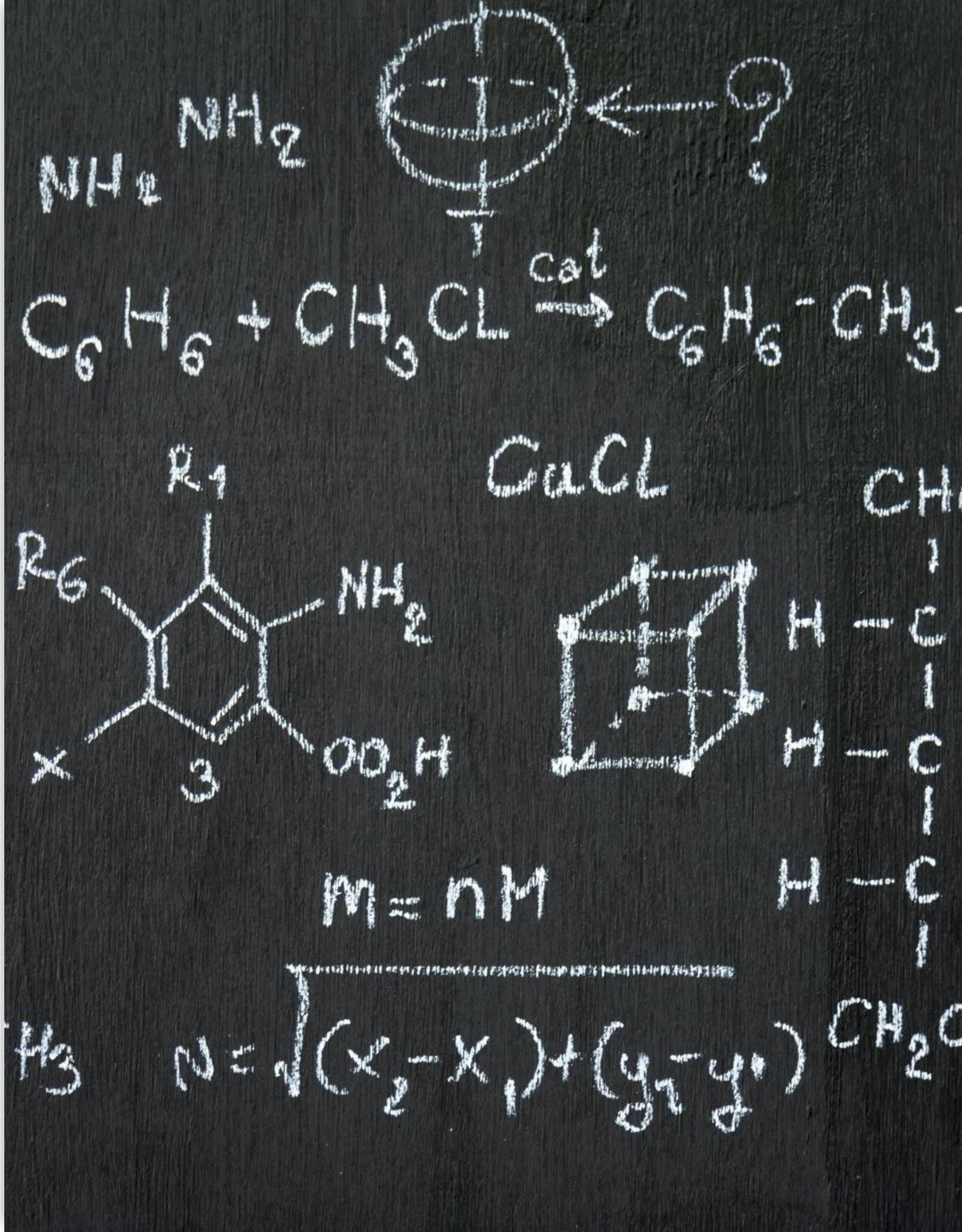
Improving a Regression Model

What is the goal?

- Find the best Fit Line equation that can predict the values based on the independent variables.

- In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used.

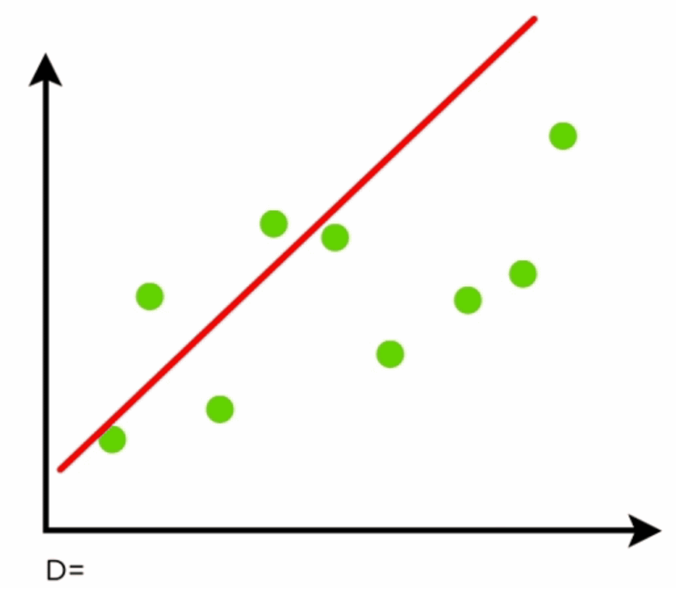
- In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.



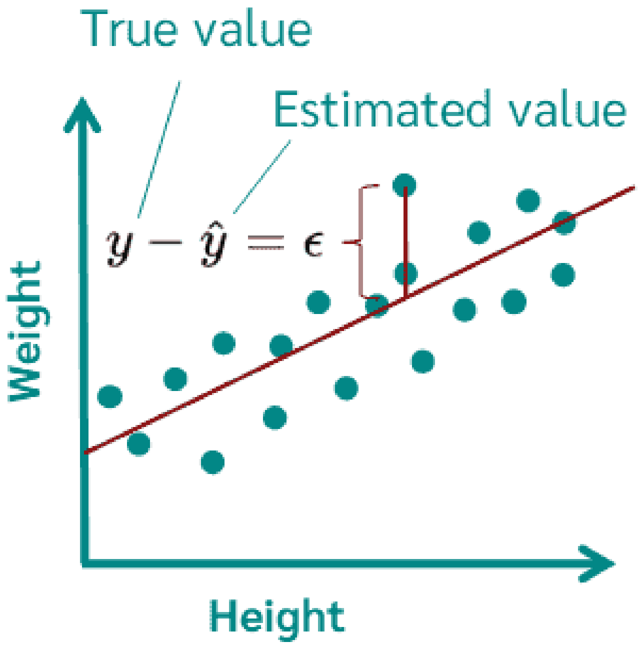
The best fit line

- The primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

- The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).



Cost Function in Linear Regression



Error epsilon

$$y = b \cdot x + a + \boxed{\epsilon}$$

- The goal is to minimize the difference between the estimated y value and the actual $\hat{y}$ value.
- This ca be achieved by updating the w0 and w1 values to reduce the difference.
- Cost function of the linear regression is the Root Mean Squared Error between predicted $\hat{y}$ and real y value.
- This is also referred to as the objective function.

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

Cost Function in Linear Regression

- $J = \frac{1}{2n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$ = Mean Squared Error
- Find the difference between the actual y and predicted $\hat{y}$ value $\hat{y} = mx + B$, which is $(\widehat{y_i} - y_i)^2$for a given x or n_i datapoint.
- n is the number of datapoints
- $\frac{1}{n}$ is the normalization and is used for averaging purposes, it can also be $\frac{1}{2n}$ sometimes as a technical choice to simplify optimization calculations in MSE.
- Square the difference
- Find the mean of the squares for every value in X.
- The y_i is the actual value and $\hat{y}$ is the predicted value.
- Substitute the value of $\hat{y}_i$, as it is equal to $(mx_i + B)$:
- $J = \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - (mx_i + B)\big)^2$
- So we square the error and find the mean. hence the name Mean Squared Error. Now that we have defined the loss function, lets get into the interesting part — minimizing it and finding m and B.
- Another could be the Mean Absolute Error: $J = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \widehat{y}_i|)$

Cost Function Example with Mean Squared Error

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

$$= \frac{1}{2(1)} \left((\mathbf{3.99}) - (\mathbf{3.8})\right)^2$$

$$= \frac{1}{1} (0.19)^2$$

$$= 1(0.0361)$$

$$= 0.0361$$

$$J = \mathbf{0.0361}$$

Where:

- $y_i = \mathbf{3.8}$
- $\hat{y}_i = \mathbf{3.99}$
- n=**1**

Cost Function Example with Mean Squared Error

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

Where:

- $y_1 = \mathbf{3.8}$
- $\widehat{y_1} = \mathbf{3.99}$
- $y_2 = \mathbf{16}$
- $\widehat{y_2} = \mathbf{16.04}$
- n=**2**

$$= \frac{1}{2(2)}\Big(\big((\mathbf{3.99}) + (\mathbf{16.04})\big) - \big((\mathbf{3.8}) + (\mathbf{16})\big)\Big)^2$$

$$= \frac{1}{4}\big((20.03) - (19.8)\big)^2$$

$$= \frac{1}{4}(0.23)^2$$

$$= 0.25(0.0529)$$

$$= 0.013225$$

$$J = \mathbf{0.013225}$$

Cost Function Example with Mean Squared Error

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

Where:

- $y_1 = \mathbf{3.8}$
- $\widehat{y_1} = \mathbf{3.99}$
- $y_2 = \mathbf{16}$
- $\widehat{y_2} = \mathbf{16.04}$
- $y_1 = \mathbf{11.77}$
- $\widehat{y_3} = \mathbf{15.22}$
- n=**3**

$$= \frac{1}{2(\mathbf{3})}\Big(\big((\mathbf{3.99}) + (\mathbf{16.04}) + (\mathbf{15.22})\big)$$
$$- \big((\mathbf{3.8}) + (\mathbf{16}) + (\mathbf{11.77})\big)\Big)^2$$
$$= \frac{1}{6}\big((35.25) - (31.57)\big)^2$$
$$= \frac{1}{6}(3.68)^2$$
$$= 0.1667(13.5424)$$
$$= 2.25751808$$
$$J = \mathbf{2.25751808}$$

Loss Functions for Regression

- $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$
- $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$
- $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$
- Root Mean Squared Error (RMSE)
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

MSE

- Calculated by taking the average of the squared differences between the actual values and the predicted values.
- It squares the errors, which gives more weight to large errors and less weight to small errors.
- Sensitive to outliers because of the squaring operation.
- Commonly used as a loss function for training regression models, as it is differentiable and convex.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

MAE

- Calculated by taking the average of the absolute differences between the actual values and the predicted values.
- Measures the average magnitude of errors without considering their direction.
- Less sensitive to outliers compared to MSE because it does not square the errors.
- More interpretable than MSE since it gives the average absolute deviation from the actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

RMSE

- Calculated by taking the square root of the average of the squared differences between the actual values and the predicted values.
- It is essentially the square root of the MSE and is measured in the same units as the dependent variable.
- Penalizes large errors more heavily than small errors, similar to MSE.
- Commonly used when you want to express the errors in the same units as the target variable, making it more interpretable.

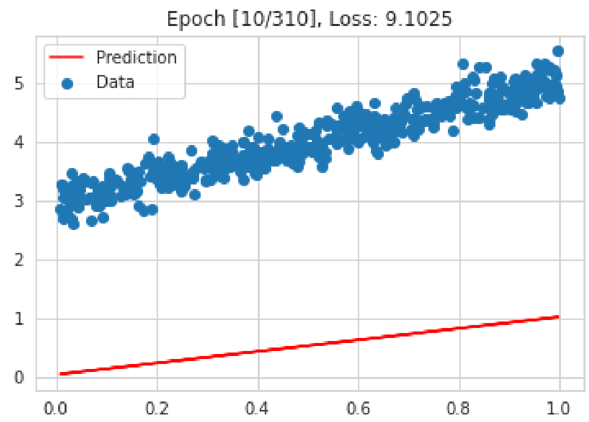
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

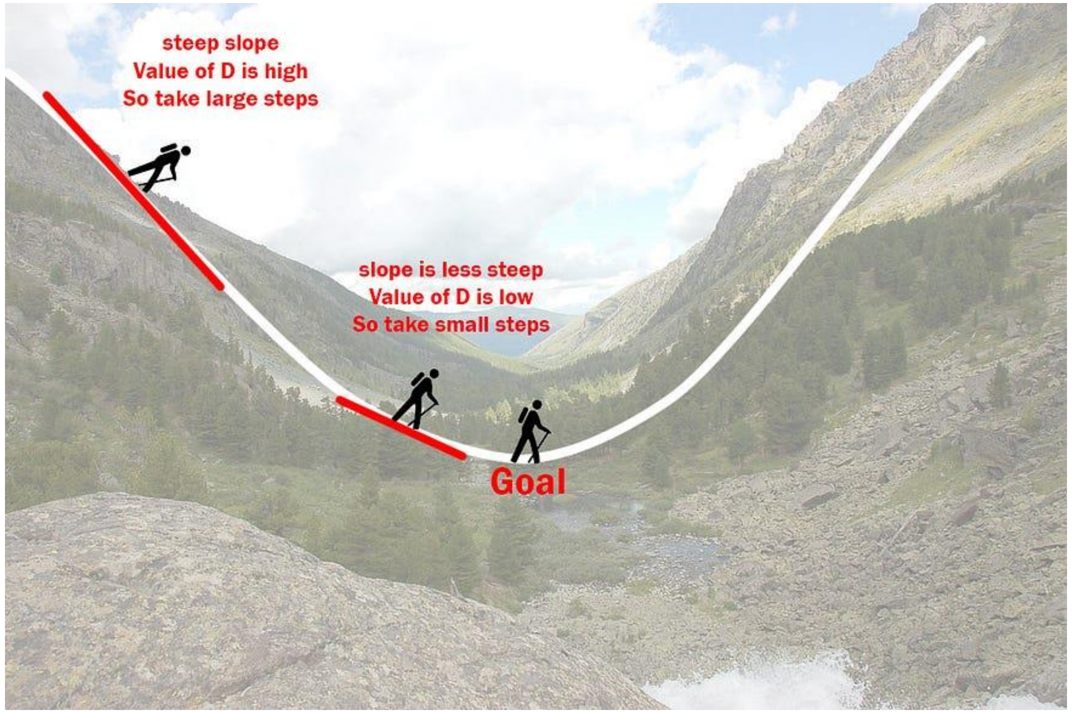
Summary of Loss Functions

- MSE and RMSE both emphasize larger errors more than smaller ones due to the squaring operation, while MAE treats all errors equally.
- MAE is generally more robust to outliers than MSE and RMSE.
- RMSE is a combination of MSE and MAE, providing a balance between interpretability and sensitivity to large errors.

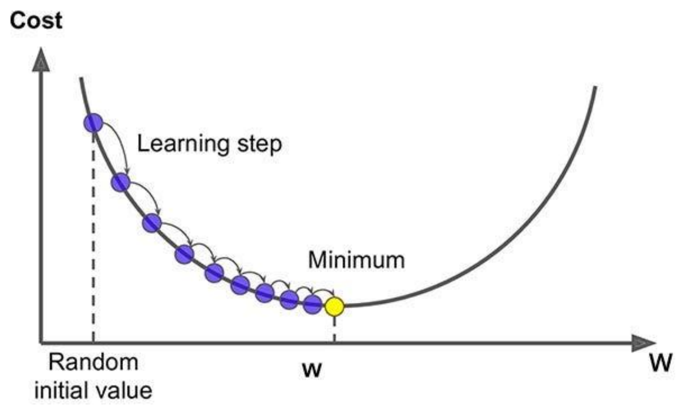
Gradient Descent

- By the definition of gradient descent, you have to find the direction in which the error decreases constantly. This can be done by finding the difference between errors.
- The small difference between errors can be obtained by differentiating the cost function and subtracting it from the previous gradient descent to move down the slope.



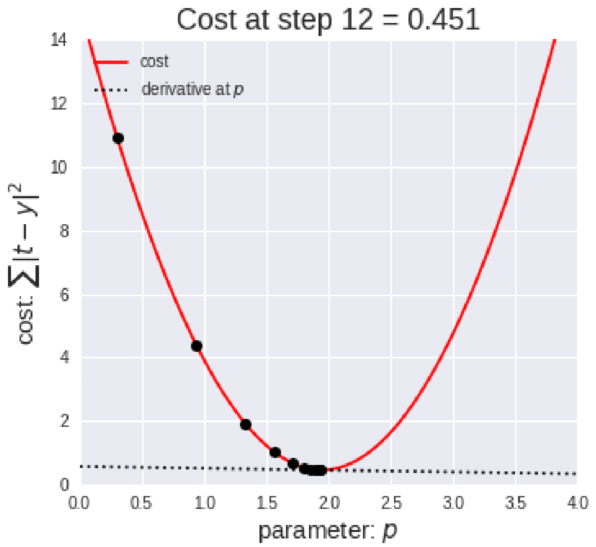


The update rule



- $\theta = \theta - \alpha \left(\frac{\partial cost}{\partial \theta}\right)$
- θ is the parameter to be optimized
- α is a positive constant or the learning rate, which determines the size of steps in each iteration.
- Smaller learning rate = slower convergence but more stable learning.
- Faster learning = faster convergence but can overshoot and oscillate heavily.
- $slope = \frac{\partial cost}{\partial \theta}$ this is typically negative, which is also referred to as the gradient ∇
- Multiplying a negative slope with a negative sign, yields a positive result. This makes it increase.
- Every time this equation repeats, θ will continue to get closer to the minima.
- If θ exceeds the local minimum, it will backtrack to the next iteration, adjusting θ again to get closer to the minima.
- Doing this process repetitively, will get θ to the local minima.

Calculating Gradient Descent



- Initialize the parameters used by the prediction $\hat{y} = \theta_0 + \theta x$
- Define the cost function. We can use MSE.
- $J(\theta)$=$\frac{1}{2n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$
- Where:
- $\widehat{y_i}$ is the predicted i-th sample.
- y_i is the actual or truth i-th sample.
- n is the number of all samples.
- Calculate the gradient of the Cost Function.
- Calculate first the partial derivative of the cost function $J(\theta)$ wrt each parameter θ_0 and θ_1. This will produce the gradients that guide the parameter updates.
- $\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{2n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)$
- $\frac{\partial J(\theta)}{\partial \theta_1} = \frac{1}{2n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)\, x_i$
- No squaring is done as we are taking partial derivatives, hence, only us getting the differences between the y_i and $\widehat{y_i}$.
- Apply the update equation.
- $\theta_0 = \theta_0 - \alpha\left(\frac{\partial J(\theta)}{\partial \theta_0}\right)$
- $\theta_1 = \theta_1 - \alpha\left(\frac{\partial J(\theta)}{\partial \theta_1}\right)$
- Plug the values into the equation $\hat{y} = \theta_0 + \theta x$ then calculate its loss using MSE. If the loss is far from the local minima or 0 error, repeat!

Another explanation Gradient Descent

- Initially let m = 0 and B = 0. Let L be our learning rate. This controls how much the value of m changes with each step. L could be a small value like 0.0001 for good accuracy.
- Calculate the partial derivative of the loss function with respect to m, and plug in the current values of x, y, m and B in it to obtain the derivative value D wrt m.
- $D_m = \frac{1}{n}\sum_{i=1}^{n} 2\big(y_i - (mx_i + B)\big)(-x_i)$
- $D_m = \frac{-2}{n}\sum_{i=1}^{n} x_i(y_i - \hat{y}_i)$
- $D_m$is the value of the partial derivative with respect to m. Similarly lets find the **partial derivative wrt** B, D_B:
- $D_B = \frac{-2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$
- Now, we just update the values of m and *B*
- $m = m - L \times D_m$
- $B = m - L \times D_B$
- Plug the values into the equation $\hat{y} = mx + B$ then calculate its loss using a specific Loss Function (e.g., MSE).
- We repeat this process until our loss function is a very small value or ideally 0 (which means 0 error or 100% accuracy). The value of m and c that we are left with now will be the optimum values.

Making sense of Gradient Descent

In analogy, m can be considered the current position of the person.

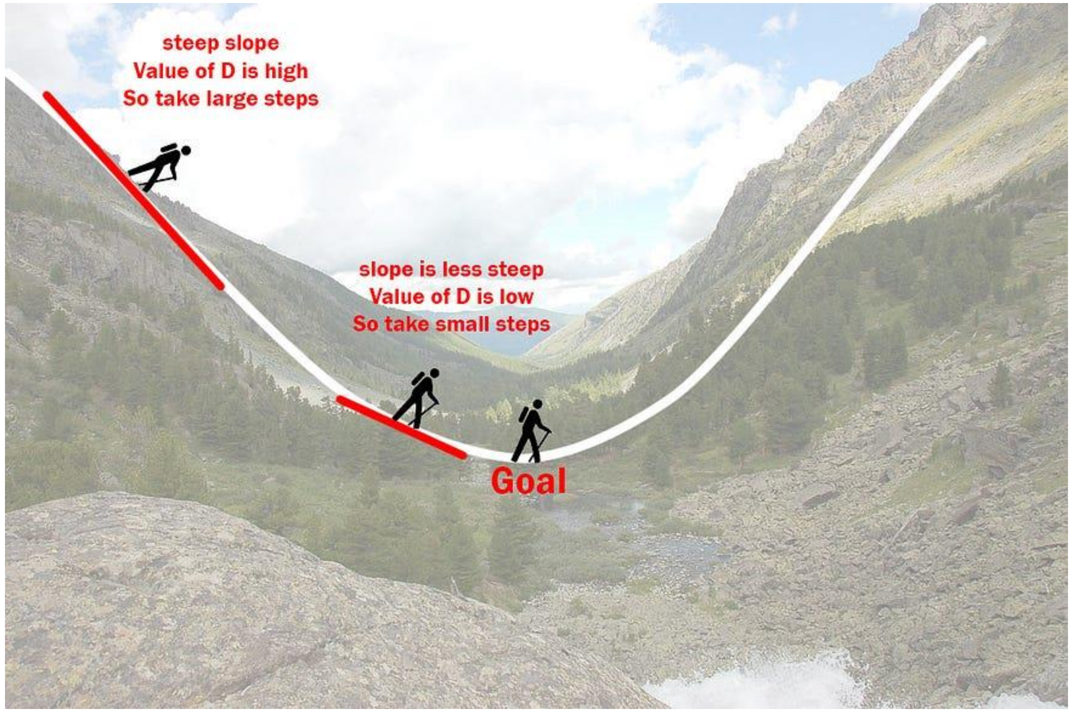
D is equivalent to the steepness of the slope and L can be the speed with which he moves.

Now the new value of m will be the next position, and $L \times D$ will be the size of the steps taken.

When the slope is more steep (D is more) it will take longer steps and when it is less steep (D is less), smaller steps are taken.

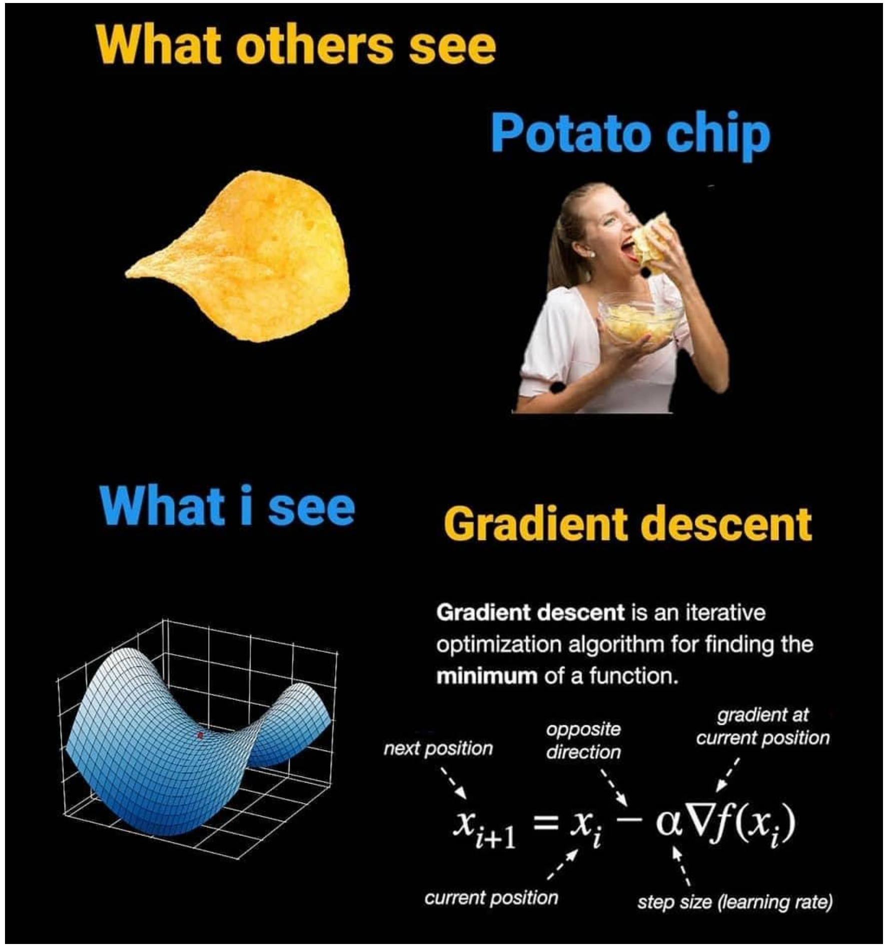
Finally, arriving at the bottom of the valley corresponds to loss = 0.

Now with the optimum value of m and c our model is ready to make predictions !



Summary of Equations in Gradient Descent

- $h_\theta(x) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots, \theta_n X_n$
- $J(\theta)$=$\frac{1}{2n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$
- $\theta_j \coloneqq \theta_j - \alpha\left(\frac{\partial J(\theta)}{\partial \theta_j}\right);\ \theta_j \coloneqq \theta - \alpha\nabla J(\theta)$



Gradient descent is an iterative optimization algorithm for finding the **minimum** of a function.

$$x_{i+1} = x_i - \alpha\nabla f(x_i)$$

Key things to remember!

- The cost function represents the error between the actual and predicted values.
- The lesser the cost function value or lower error, the better the model is.
- The cost function will be different for other models.
- The presented cost function is a general cost function for linear regression.
- Loss calculated by the cost function is reduced with the help of gradient descent.
- Values or parameters used to generate a prediction are optimized using a gradient descent and are identified optimal depending on the result of a loss function.

python + colab

NumPy

matplotlib

Pandas

scikit learn

Seaborn

Next meeting: Lab

Linear and Multiple Linear