Intro to Inferential Statistics with R

Workshop 5

Course: VSK1004 Applied Researcher

Workshop structure (draft)

MONDAY
Intro to Statistic Inference

1. Descriptive vs Inferential statistics
2. Population, sample and sampling distribution
3. Null Hypothesis testing
4. Correlation and interpretation

WEDNESDAY
More about inferential stats

1. Choosing a statistical test
2. t-test family
3. chi-squared
4. correlation
5. Chi-squared distribution

TODAY
Linear & Logistic Regression

1. Linear Regression
2. Multiple Linear Regression
3. Model Assumption
4. Logistic Regression

Our goal in the next 40 min

In this session, we will cover some of the **basic statistical models and its properties such as:**

1. Simple Linear Regression
2. Multiple Linear Regression
3. Linear Model Assumptions
4. Logistic Regression

Simple Linear Regression

The mathematical equation

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

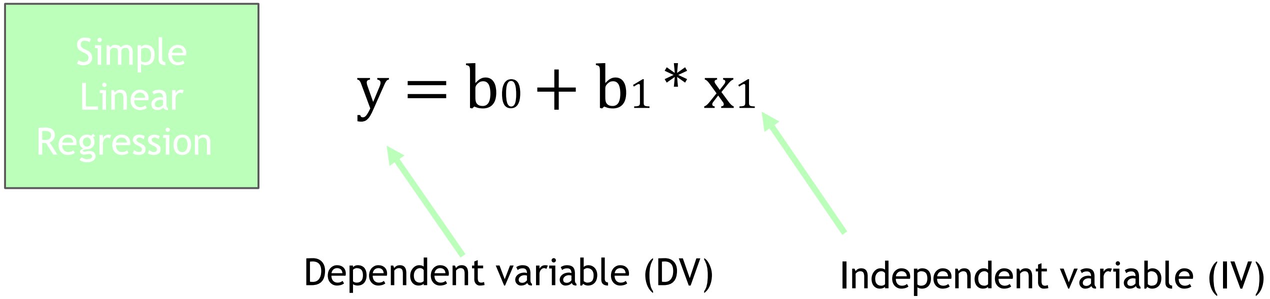
The mathematical equation

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

The mathematical equation



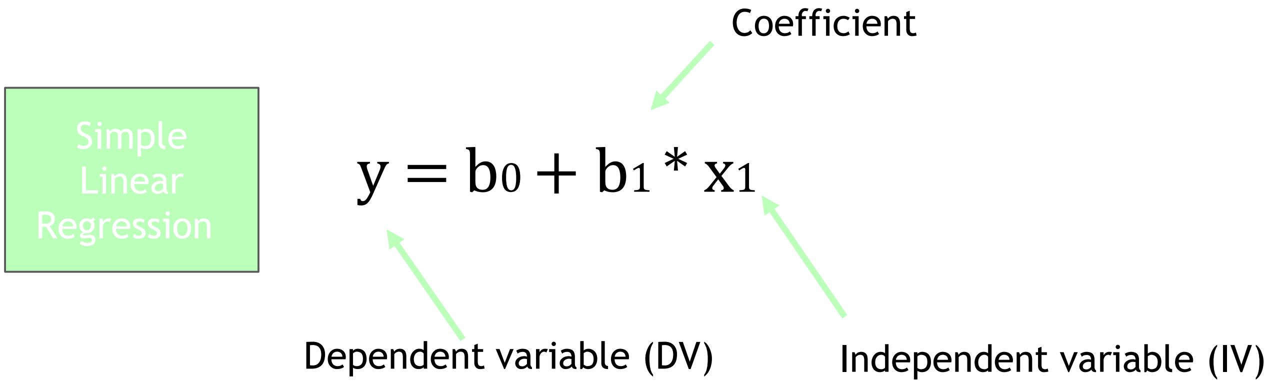
Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

Independent variable (IV)

The mathematical equation



Simple Linear Regression

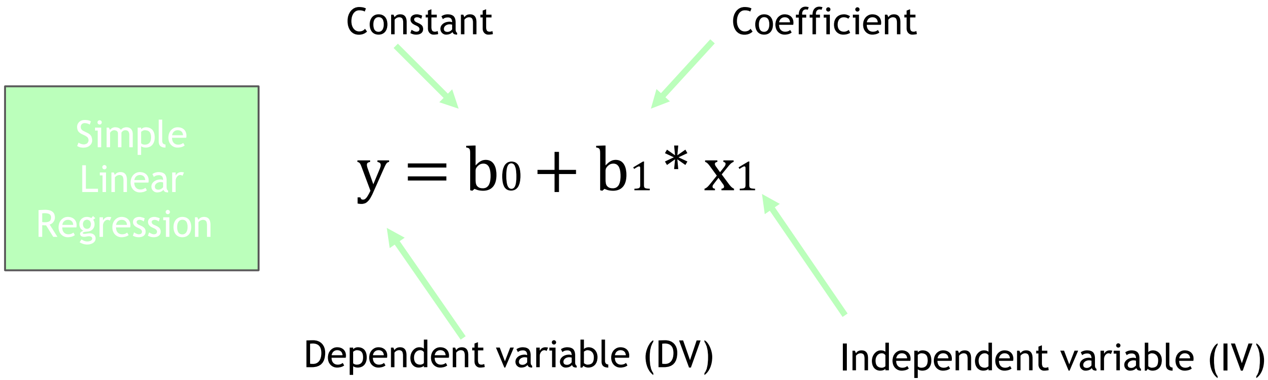
$$y = b_0 + b_1 * x_1$$

Coefficient

Dependent variable (DV)

Independent variable (IV)

The mathematical equation



Simple Linear Regression

Constant

Coefficient

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

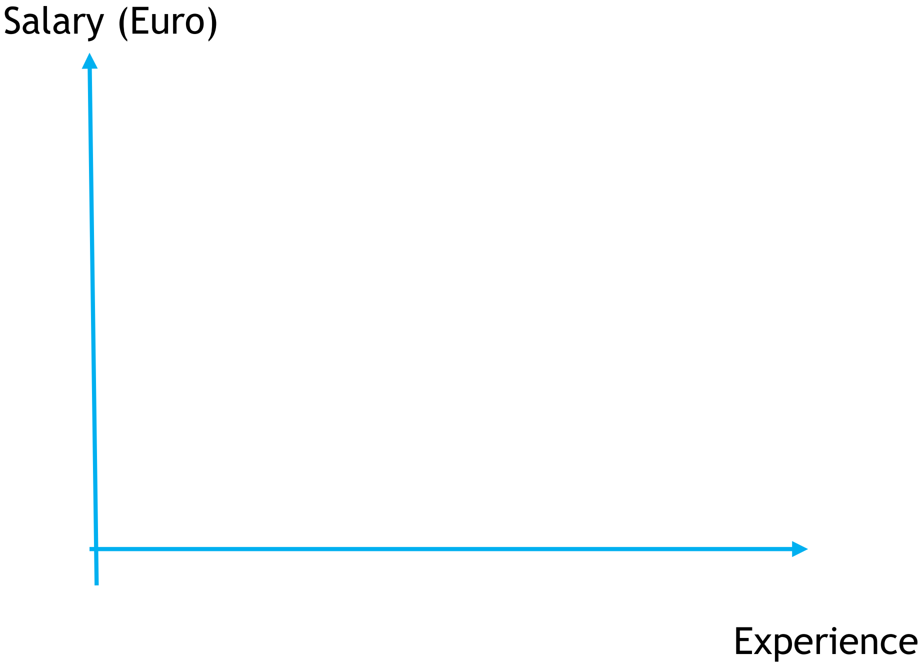
Independent variable (IV)

Look at the simple Linear regression

Simple Linear Regression:

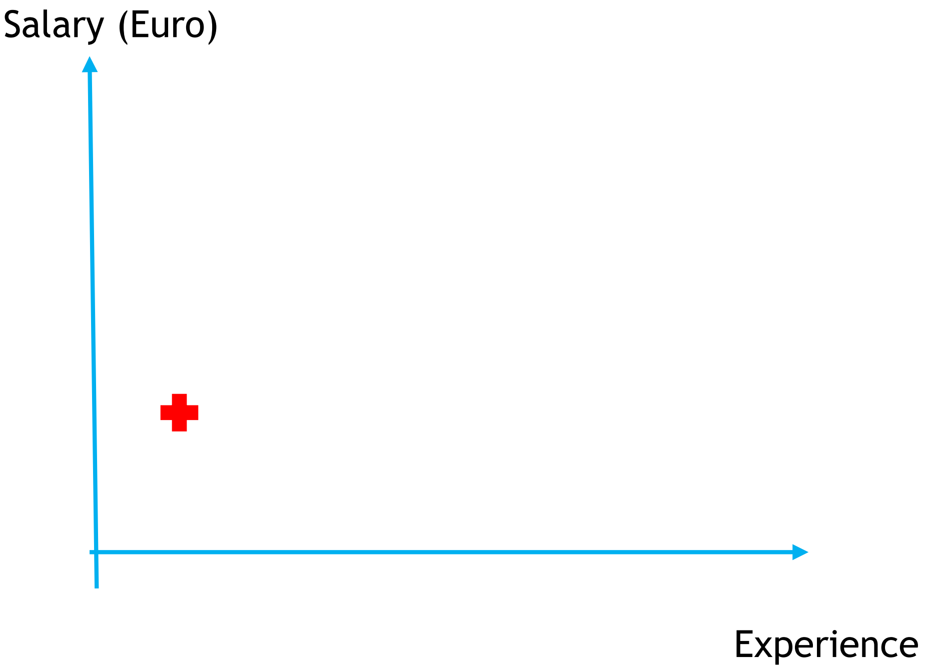
Representation in a graph: x and y axis

Simple Linear Regression:



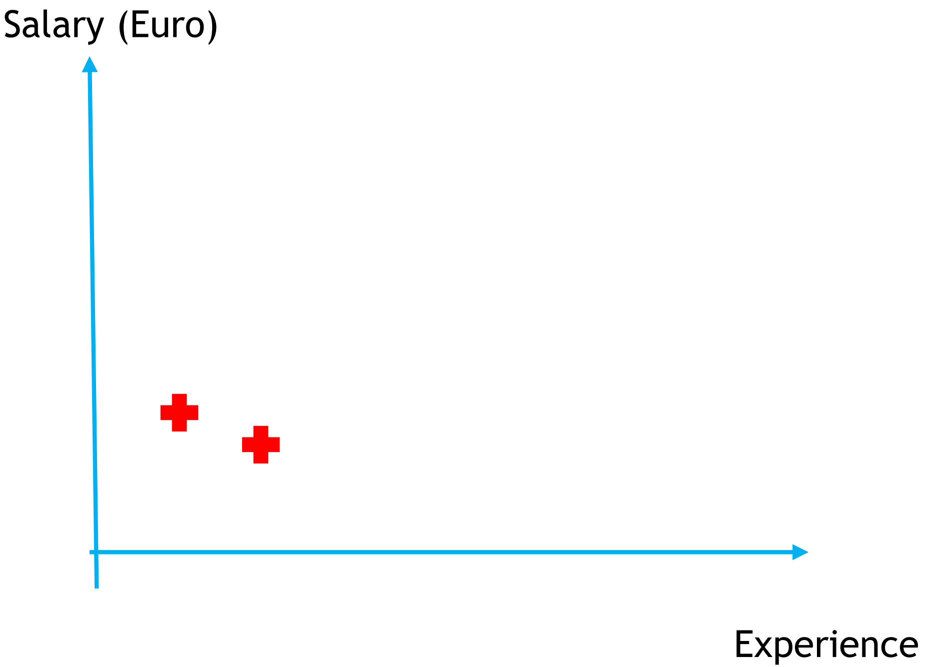
Repesentation in a graph: dots as observed data

Simple Linear Regression:



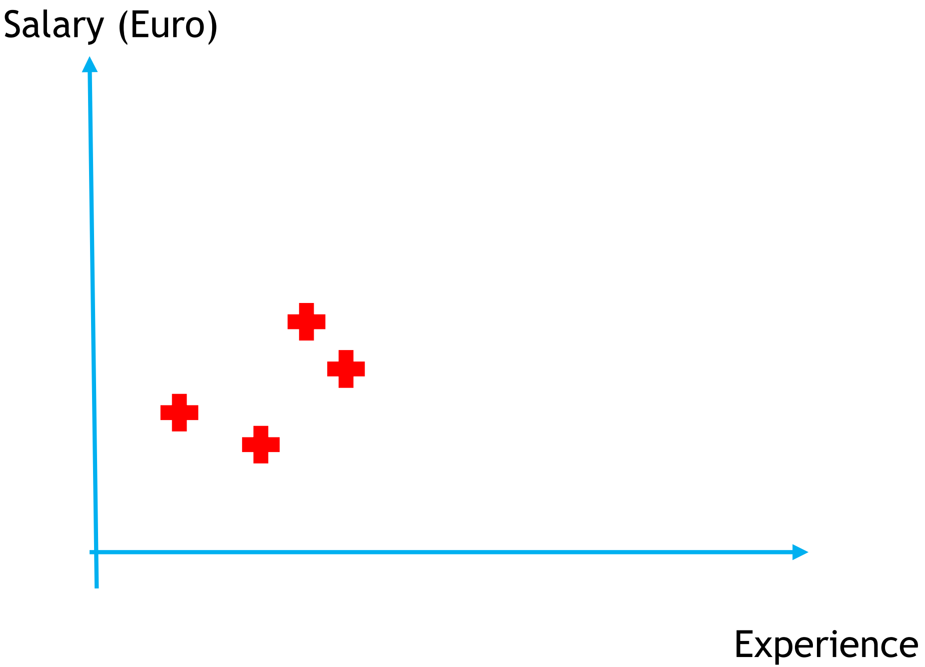
Repesentation in a graph: dots as observed data

Simple Linear Regression:



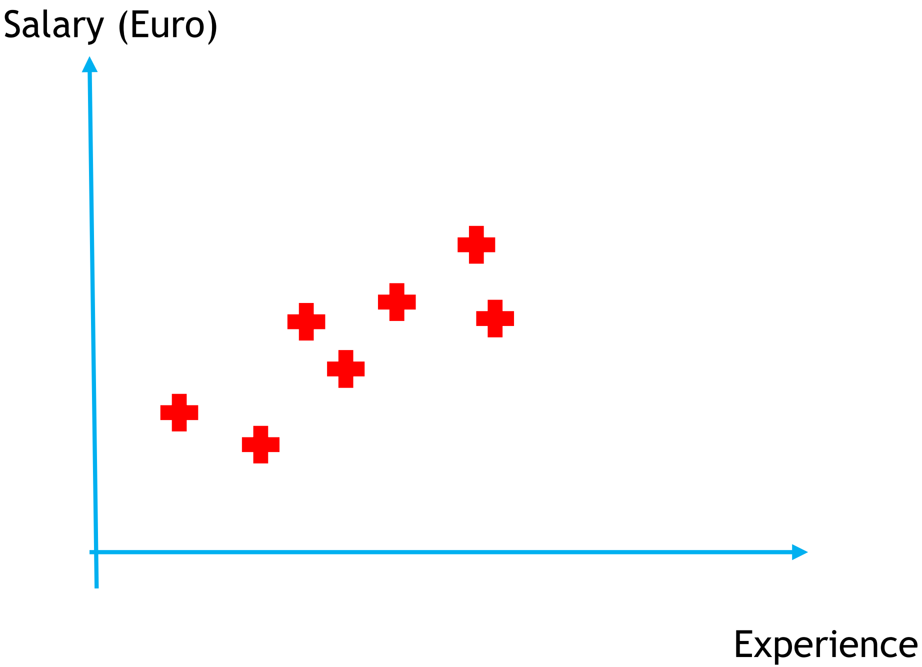
Repesentation in a graph: dots as observed data

Simple Linear Regression:



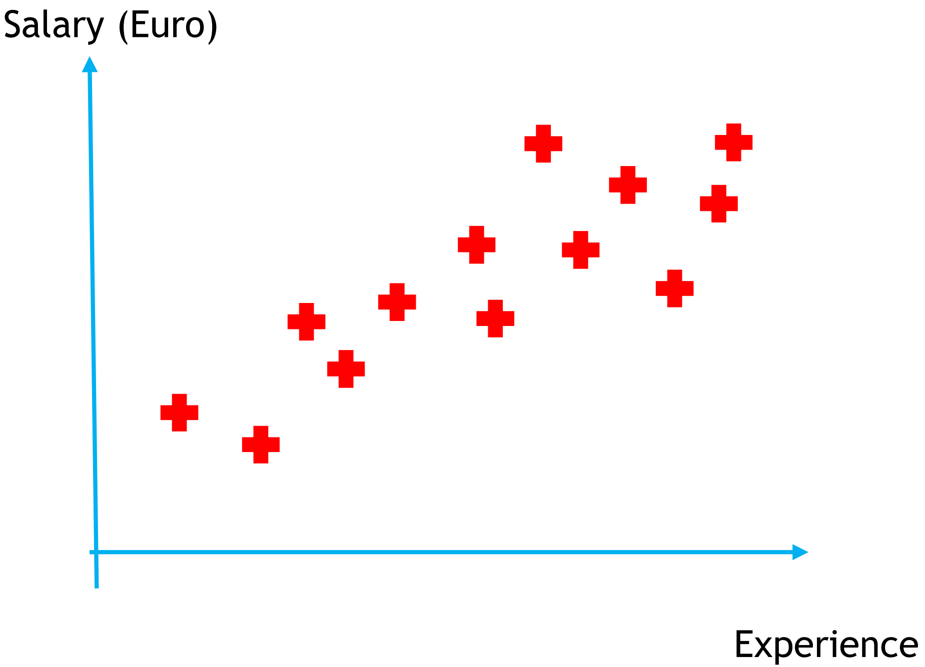
Repesentation in a graph: dots as observed data

Simple Linear Regression:



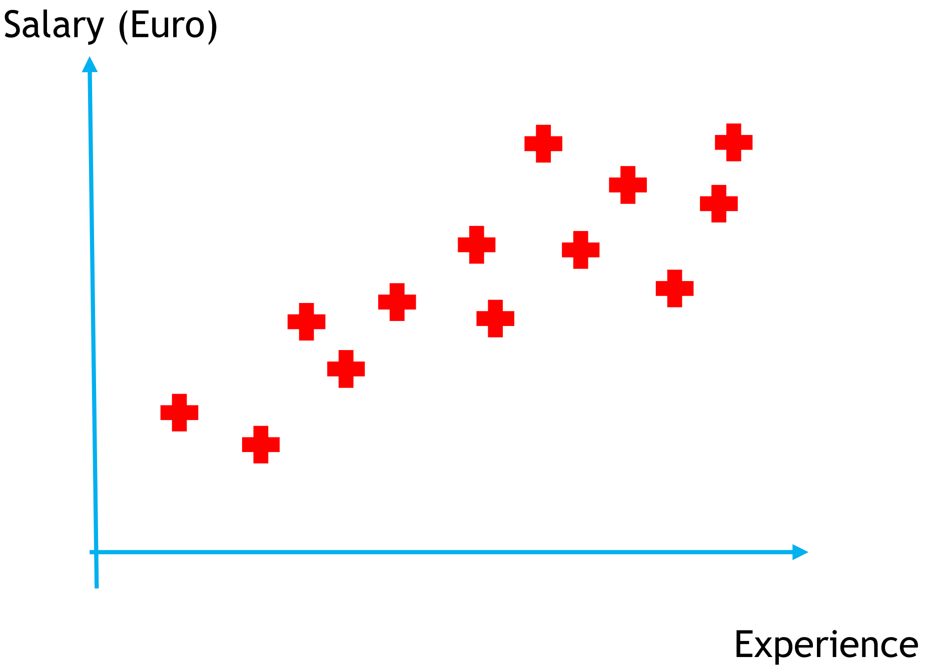
how salary is distributed among people

Simple Linear Regression:



Repesentation in a graph: observed data

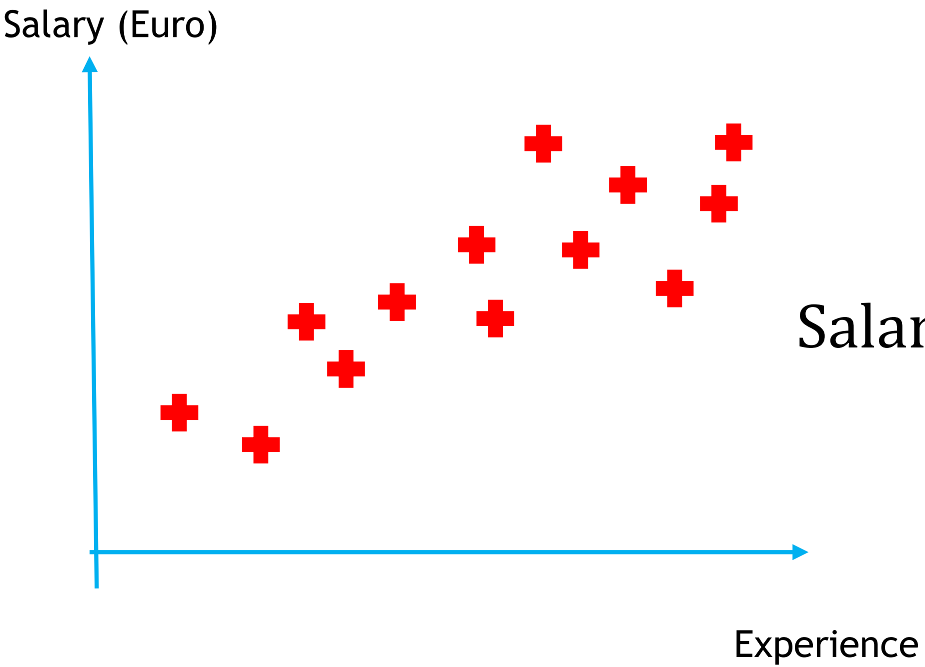
Simple Linear Regression:



$$y = b_0 + b_1 * x_1$$

Repesentation in a graph with the equation

Simple Linear Regression:

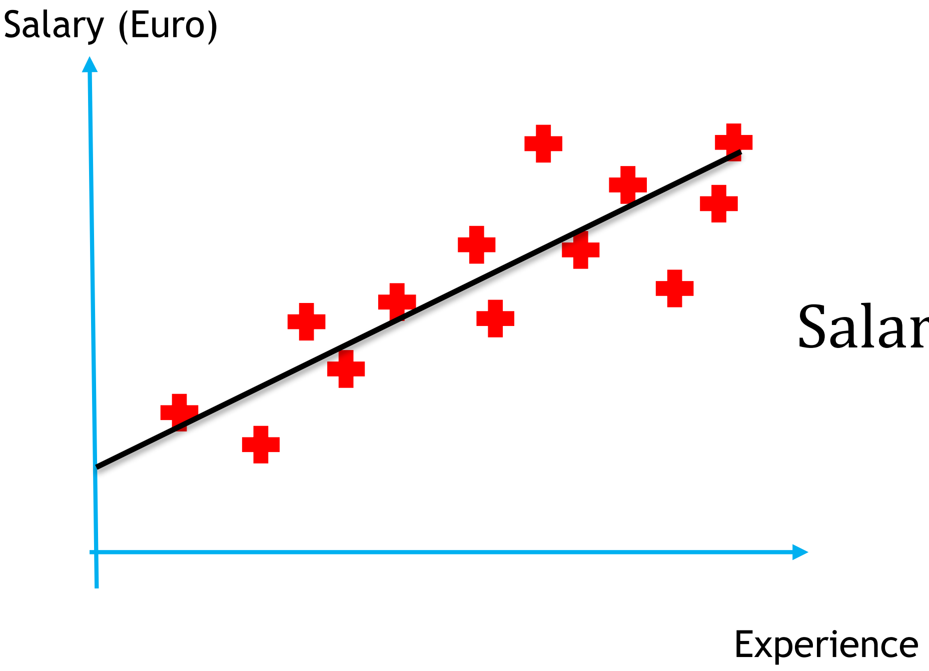


$$y = b_0 + b_1 * x_1$$

$$Salary = b_0 + b_1 * Experience$$

Add best fitting line for linear regression

Simple Linear Regression:



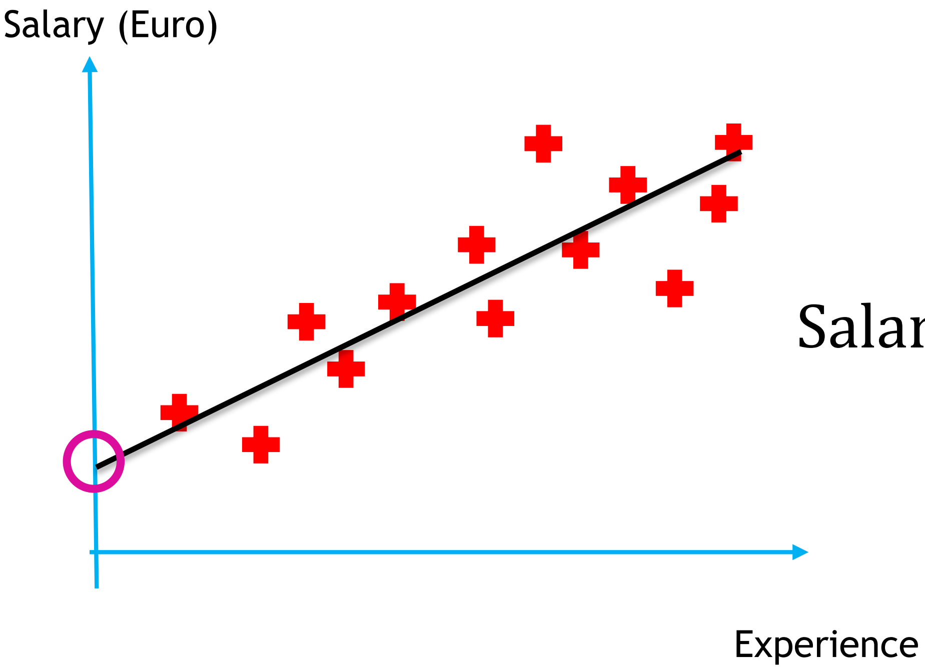
$$y = b_0 + b_1 * x_1$$

$$\Downarrow$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

Identify parameters in the graphs: Constant

Simple Linear Regression:

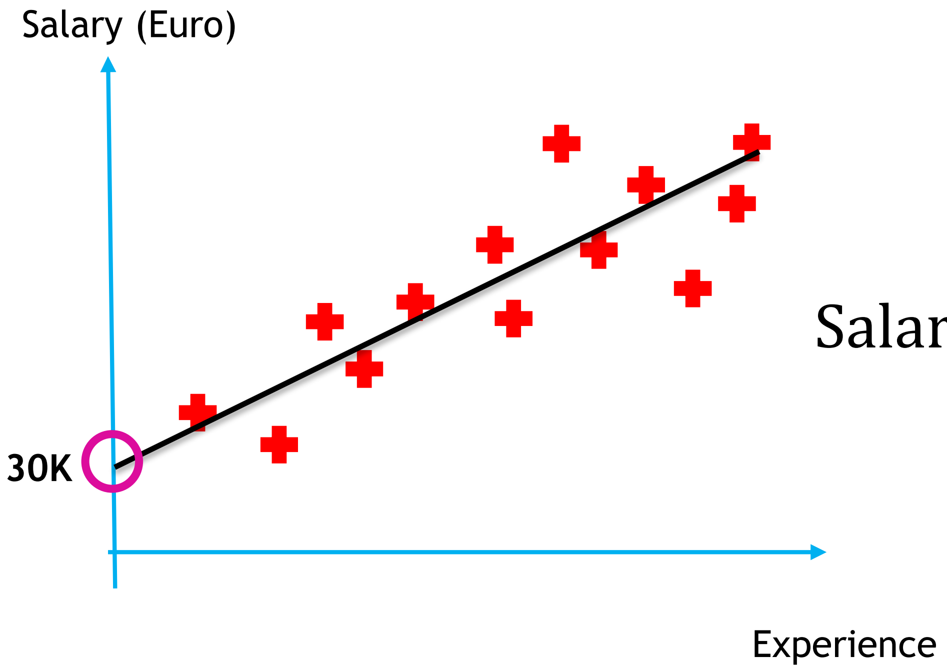


$$y = b_0 + b_1 * x_1$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

Identify parameters in the graphs: Constant

Simple Linear Regression:

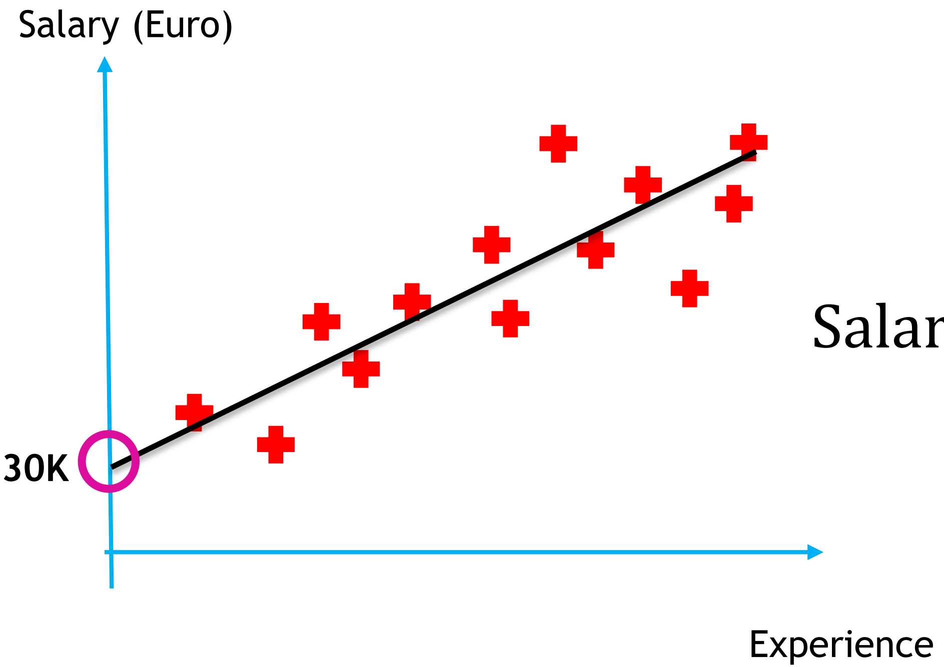


$$y = b_0 + b_1 * x_1$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

Identify parameters in the graphs: Slope

Simple Linear Regression:

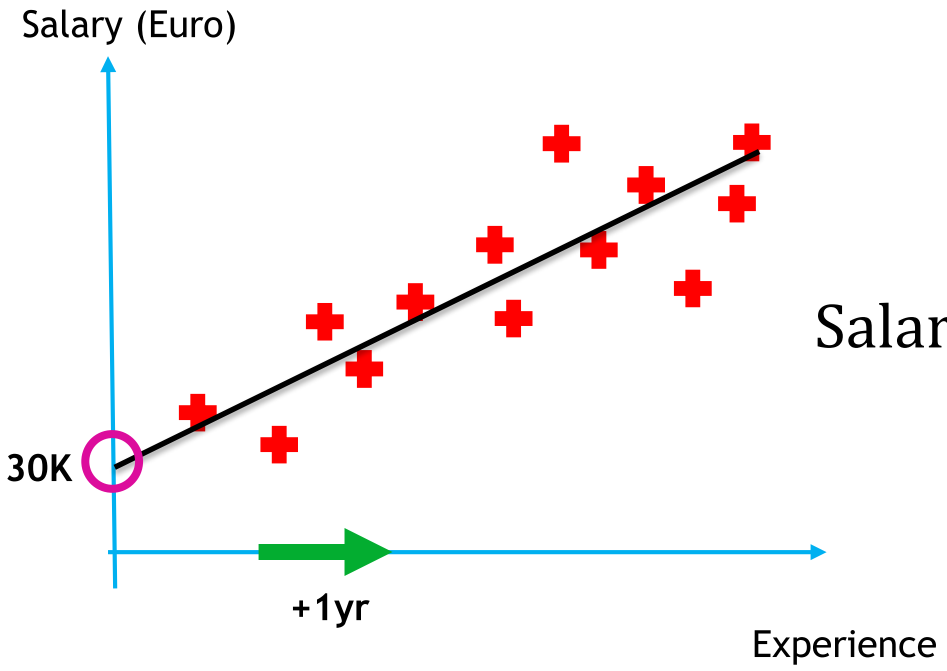


$$y = b_0 + b_1 * x_1$$

$$\text{Salary} = \boxed{b_0} + \boxed{b_1} * \text{Experience}$$

Identify parameters in the graphs: Slope

Simple Linear Regression:

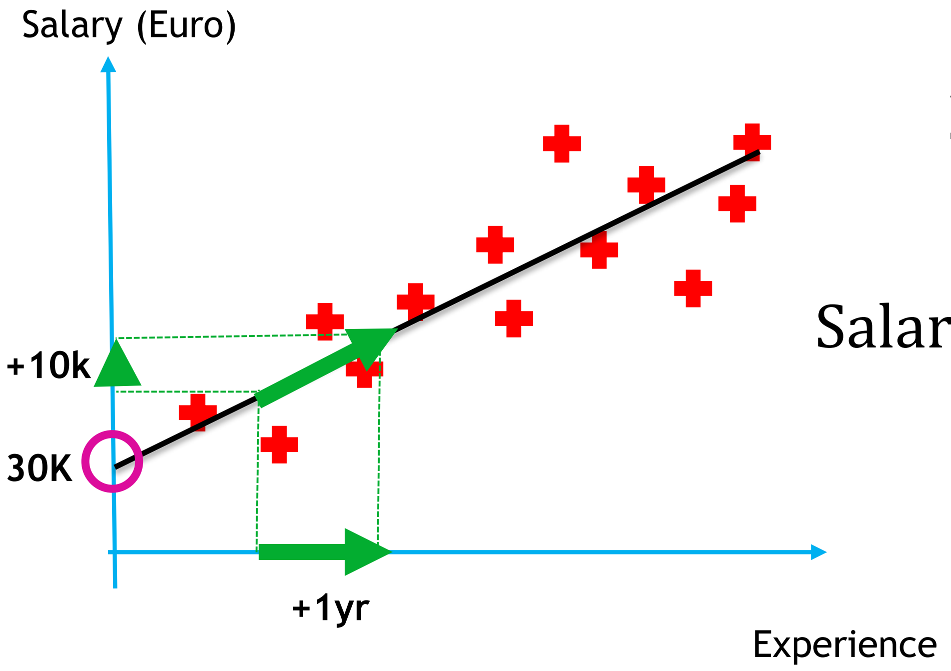


$$y = b_0 + b_1 * x_1$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

Identify parameters in the graphs: Slope

Simple Linear Regression:

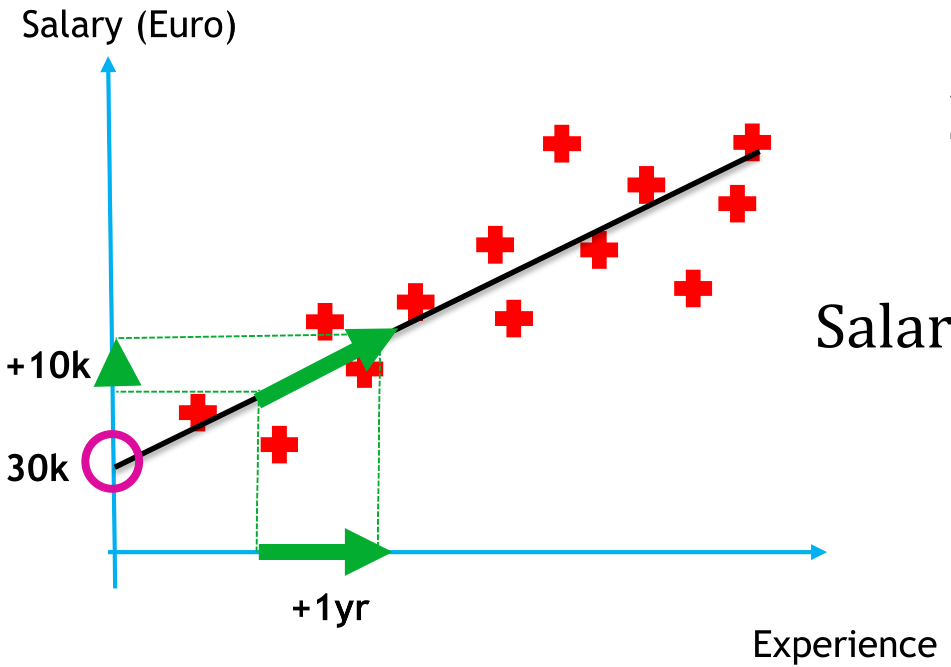


$$y = b_0 + b_1 * x_1$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

All linear parameters in the best fitted line

Simple Linear Regression:

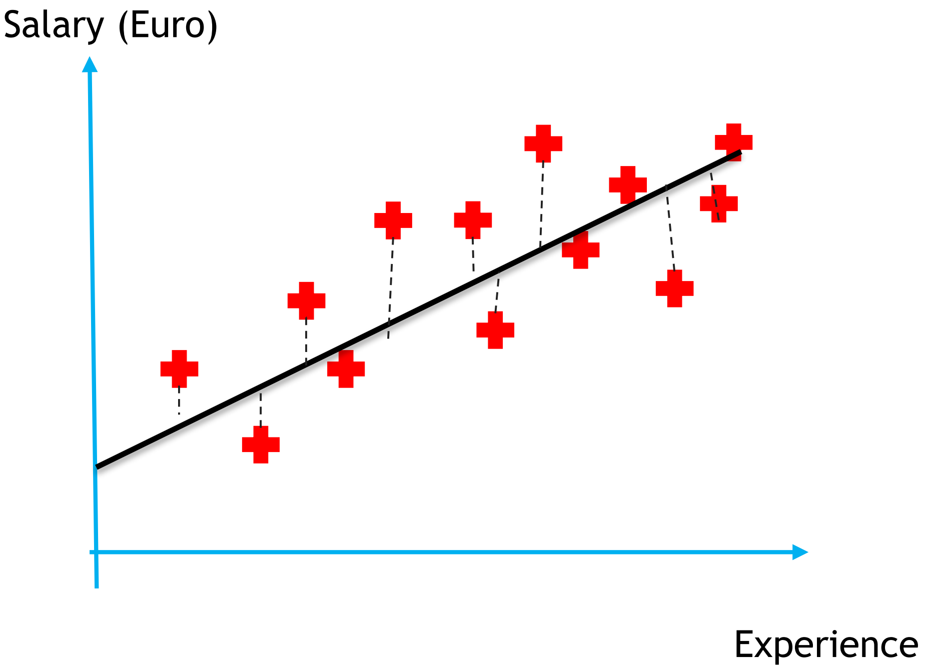


$$y = b_0 + b_1 * x_1$$

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

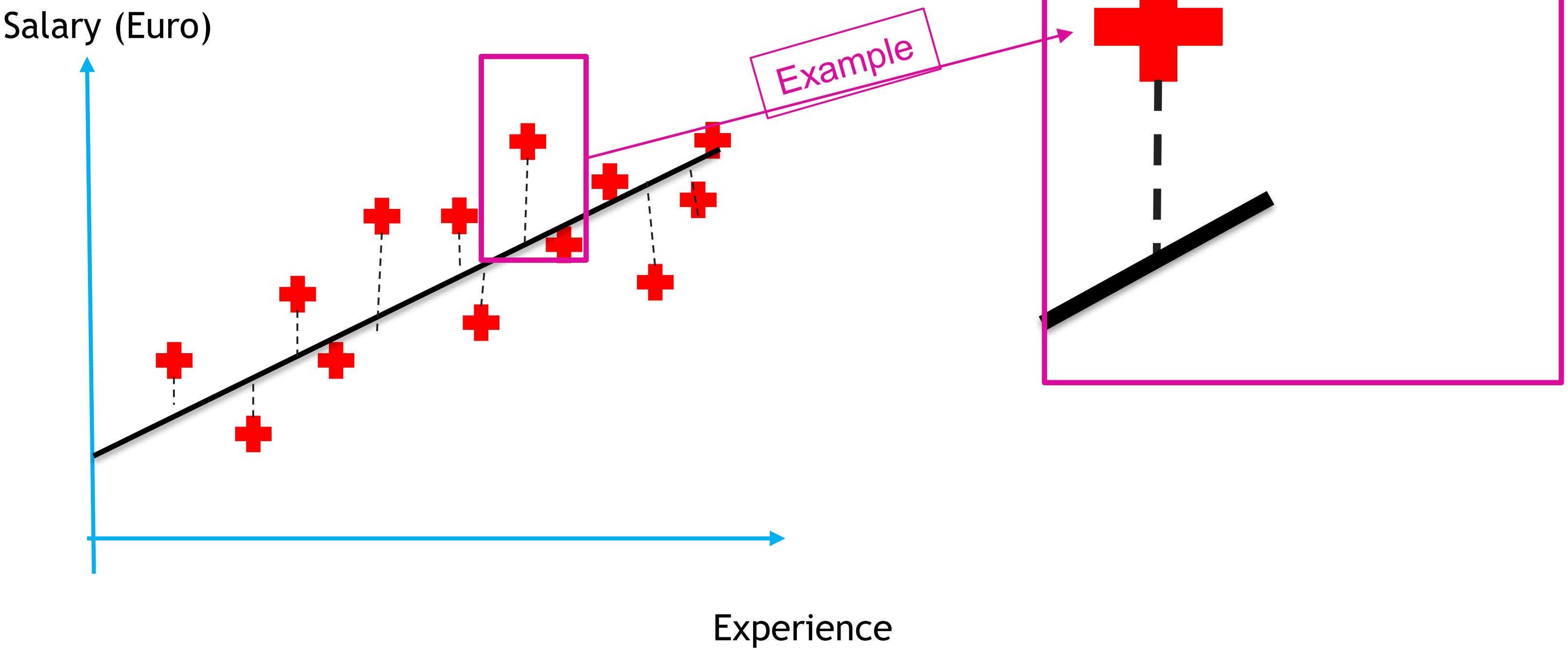
How linear regression find best fitting line?

Simple Linear Regression:



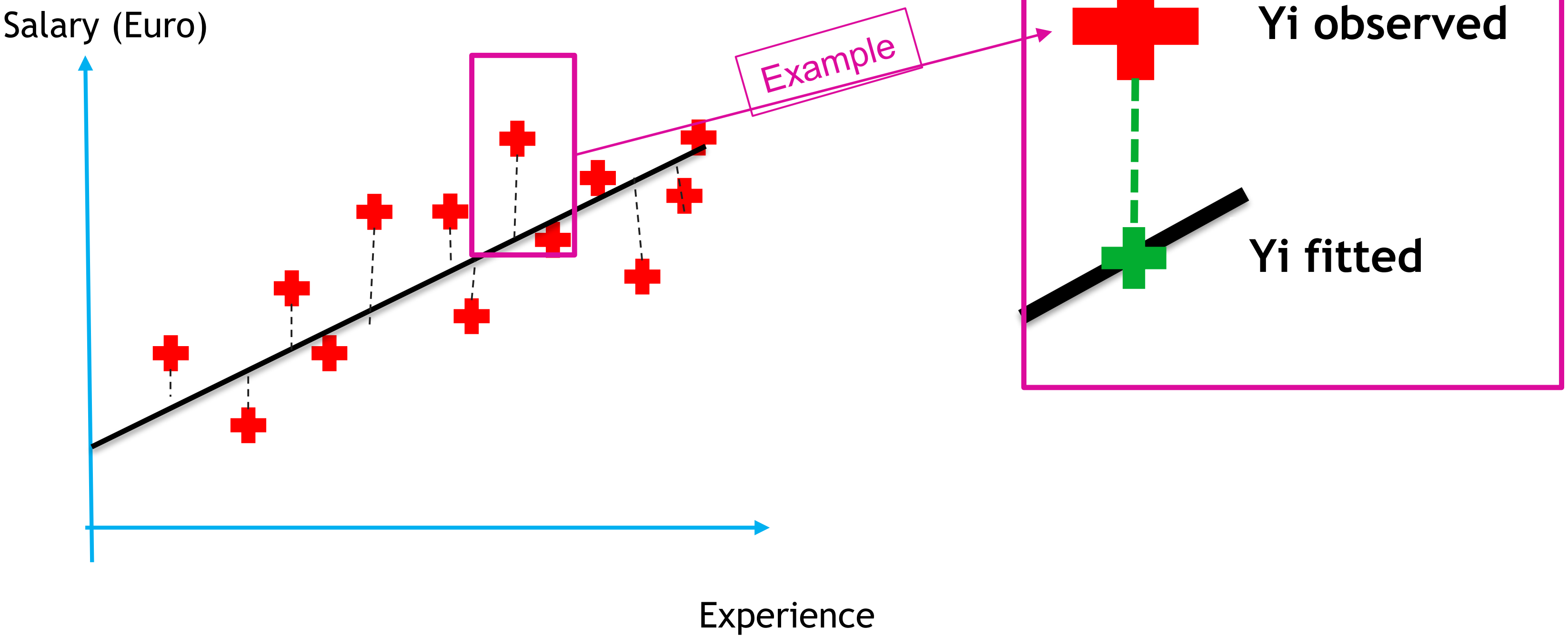
How linear regression find best fitting line

Simple Linear Regression:



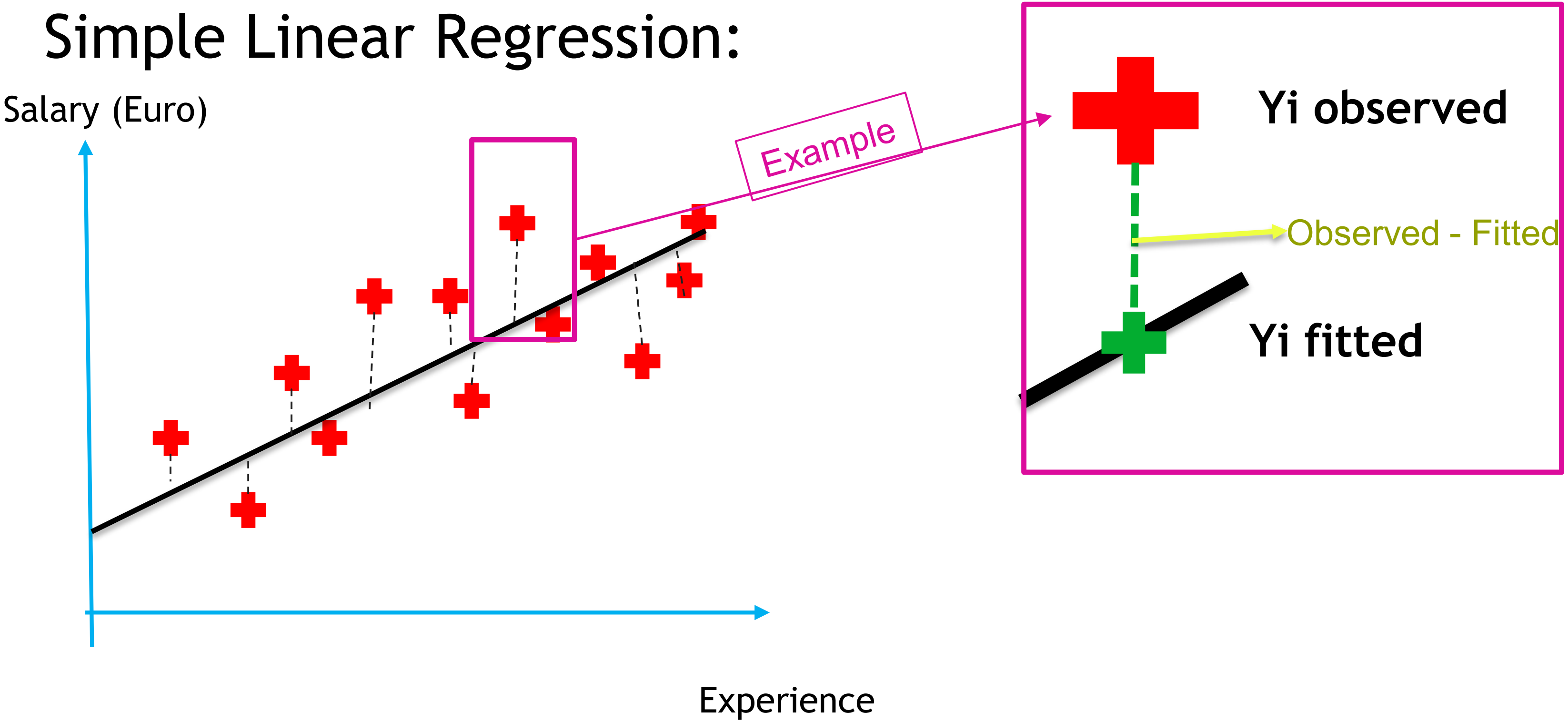
How linear regression find best fitting line

Simple Linear Regression:



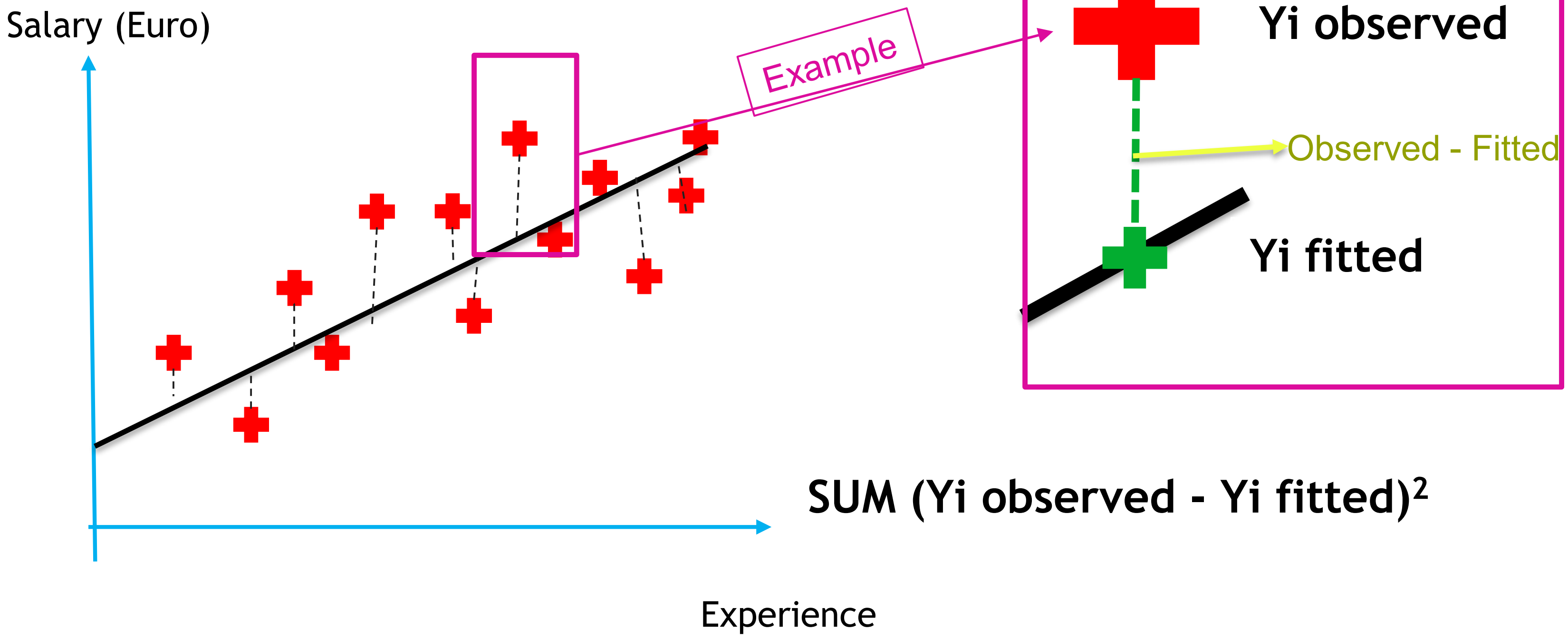
How linear regression find best fitting line

Simple Linear Regression:



How linear regression find best fitting line

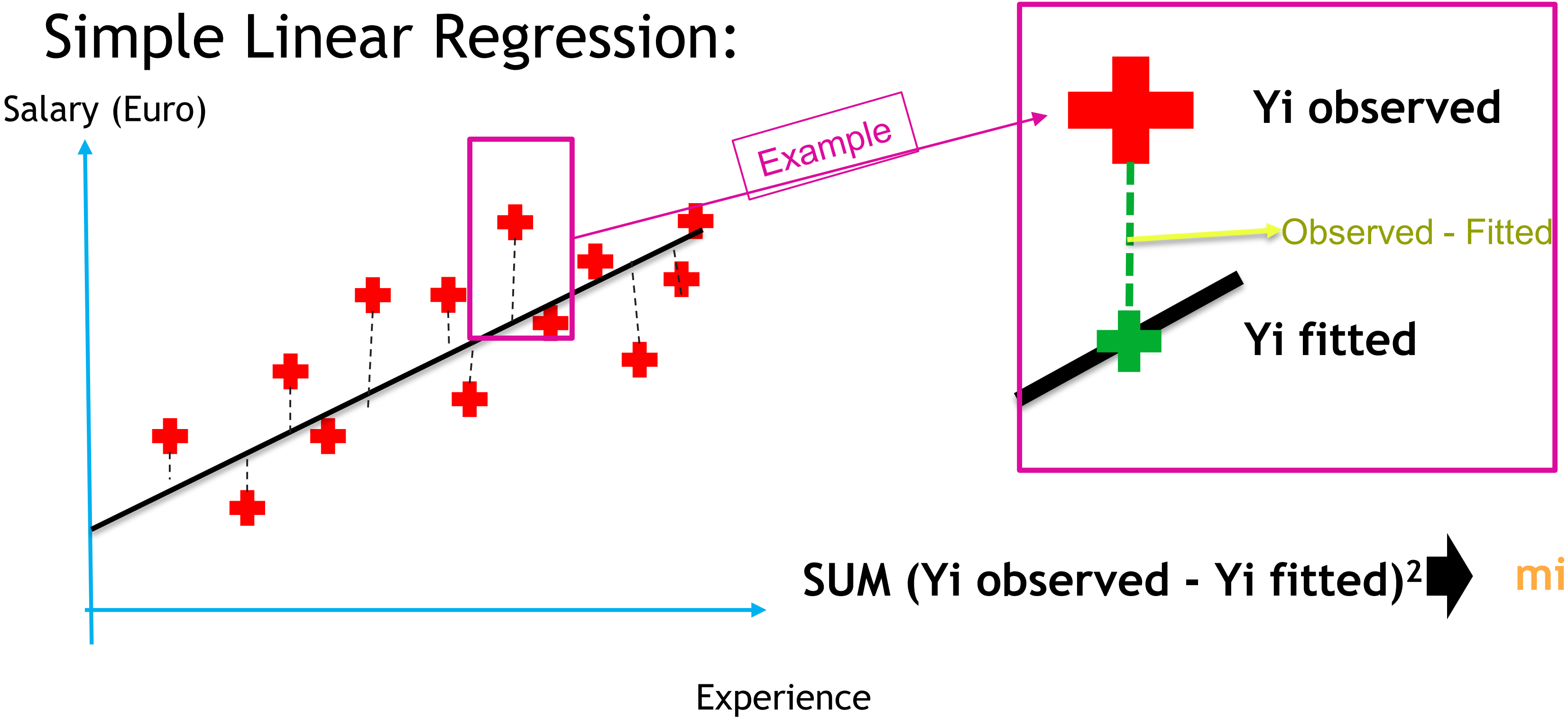
Simple Linear Regression:



SUM (Yi observed - Yi fitted)^2

How linear regression find best fitting line

Simple Linear Regression:



SUM (Yi observed - Yi fitted)^2 → min

Linear Regression looks for min sum of squares to find the line which has the smallest sum squares possible, and its called, the best fitting line

Multiple Linear Regression

Same thing but many variables into the model

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * X_n$$

Same thing but many variables into the model

Simple Linear Regression

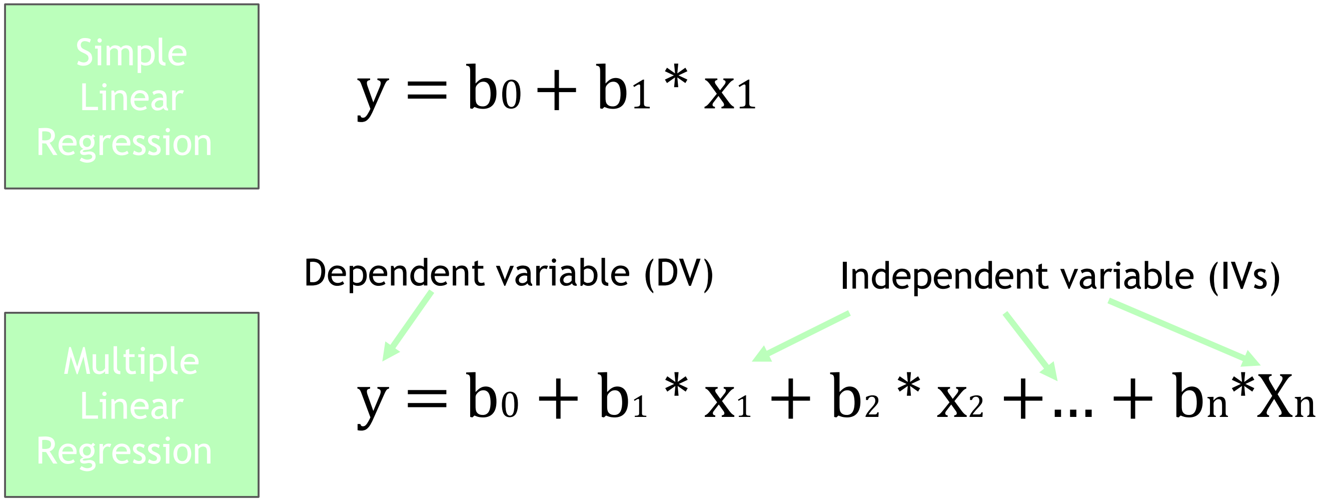
$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n{*}X_n$$

Same thing but many variables into the model



Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

Independent variable (IVs)

Multiple Linear Regression

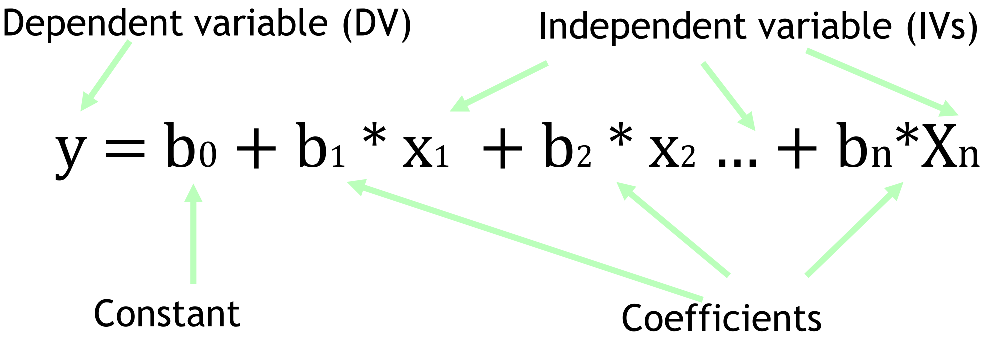
$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * X_n$$

Regressions

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression



Dependent variable (DV)

Independent variable (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 \ldots + b_n * X_n$$

Constant

Coefficients

Model Assumptions

A Caveat: Assumptions of Linear Regression

1. Linearity
2. Homoscedasticity
3. Multivariate Normality
4. Independence of errors
5. Lack of multicollinearity

Logistic Regression

What we know

Linear Regression:

What we know

Linear Regression:

- Simple

$y = b_0 + b_1 * x$

What we know

Linear Regression:

- Simple

$y = b_0 + b_1 * x$

- Multiple:

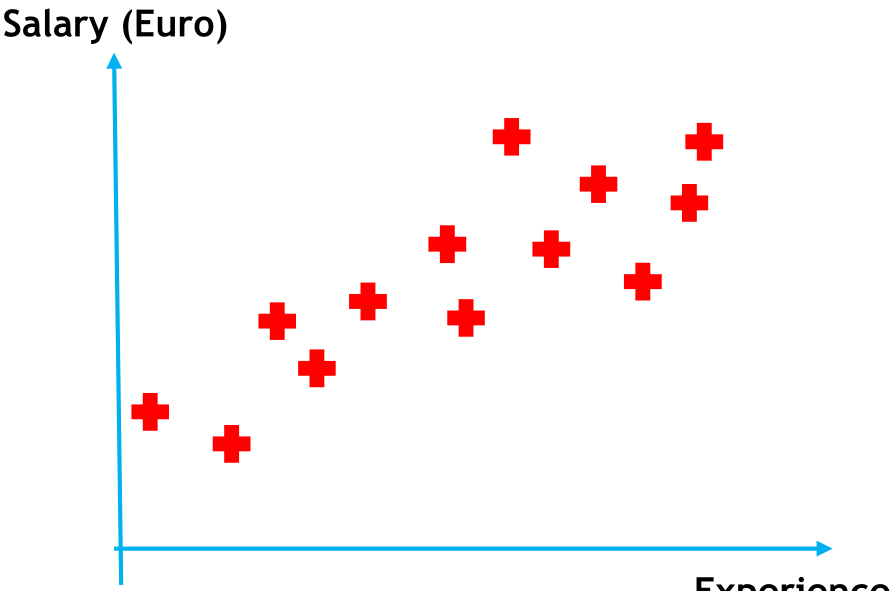
$y = b_0 + b_1 * x_1 + ... + b_n*X_n$

What we know

We know this:

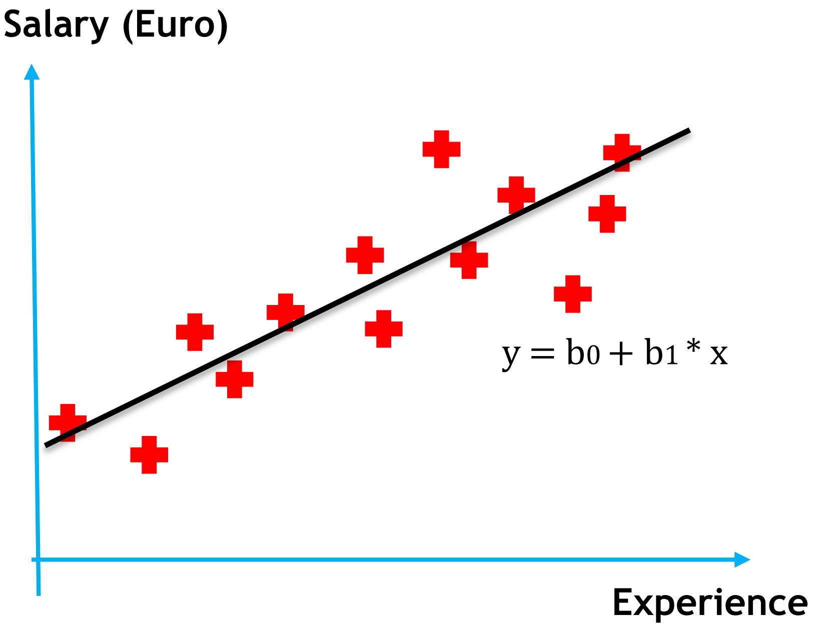
What we know

We know this:



What we know

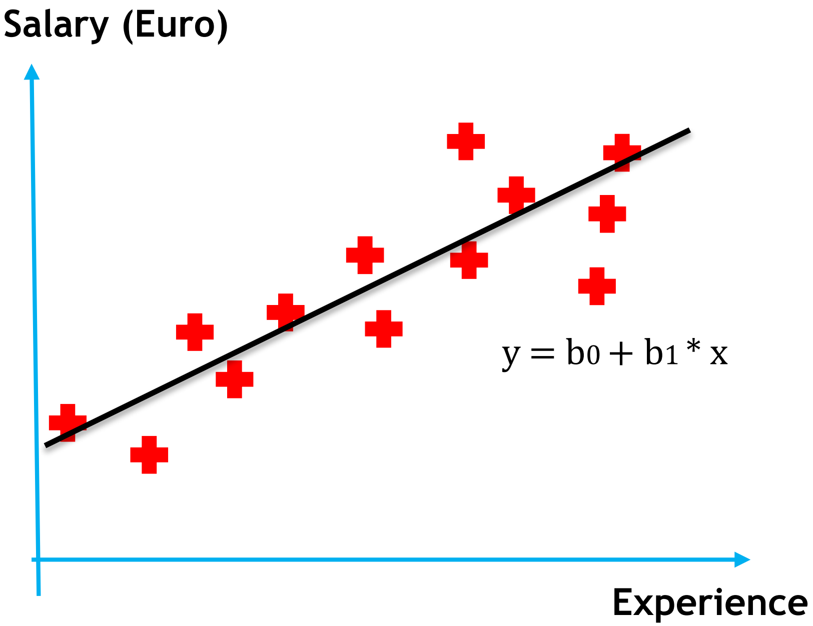
We know this:



What is new: Logistic regression

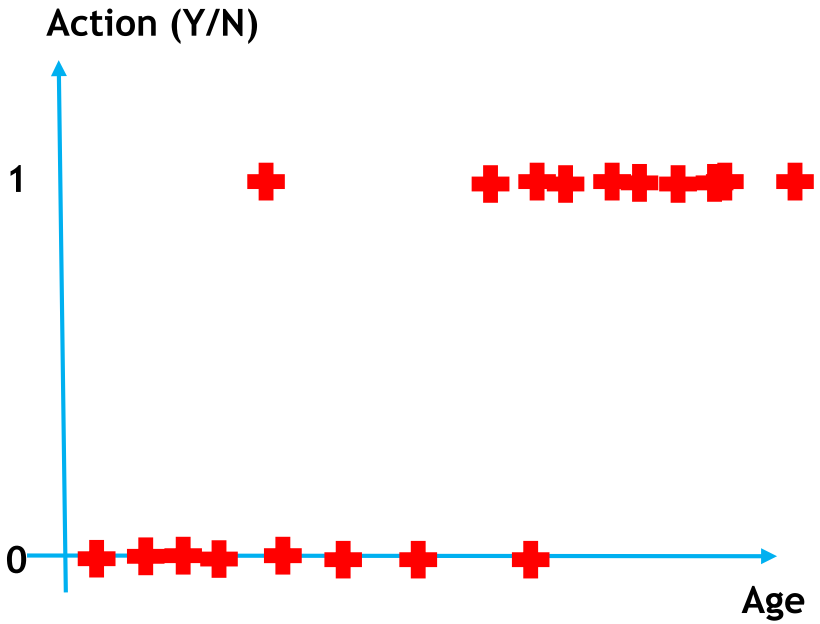
This is new:

We know this:

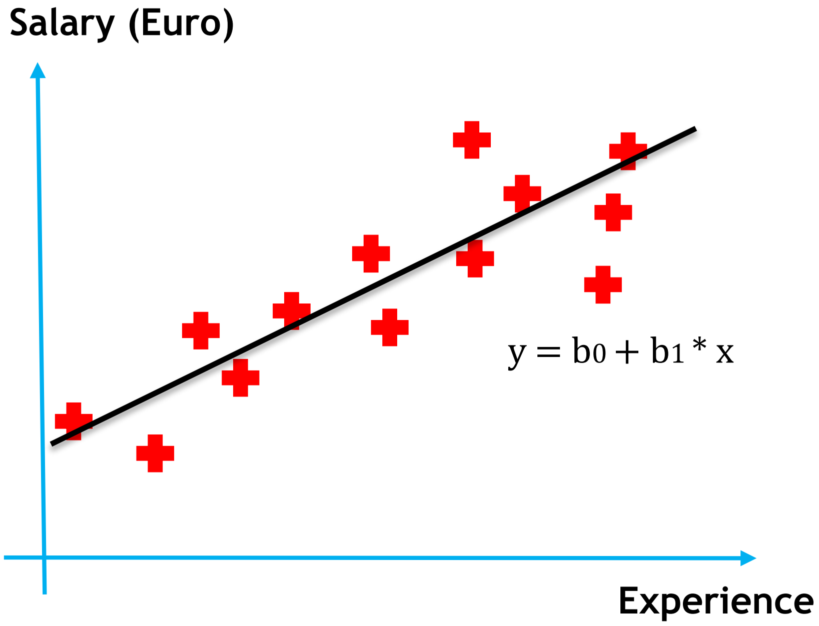


Logistic regression

This is new:

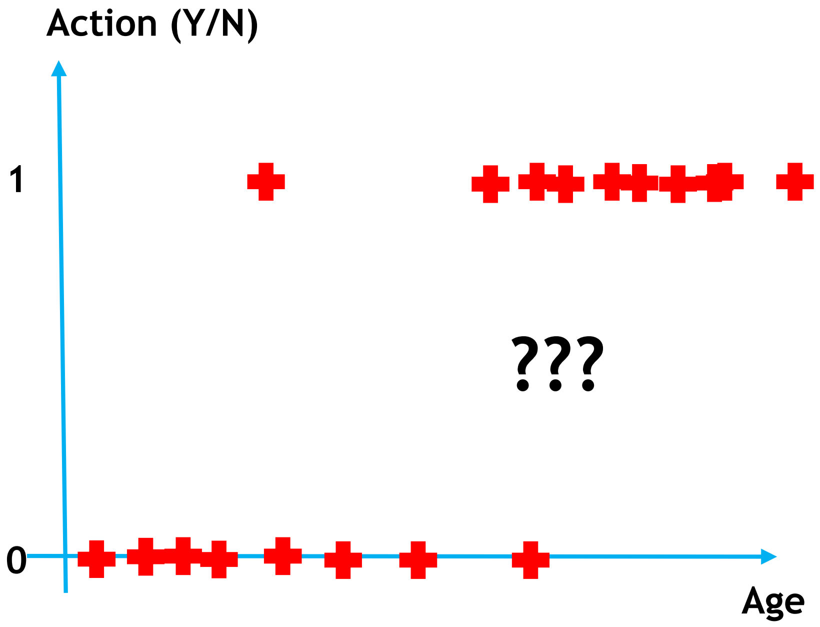


We know this:

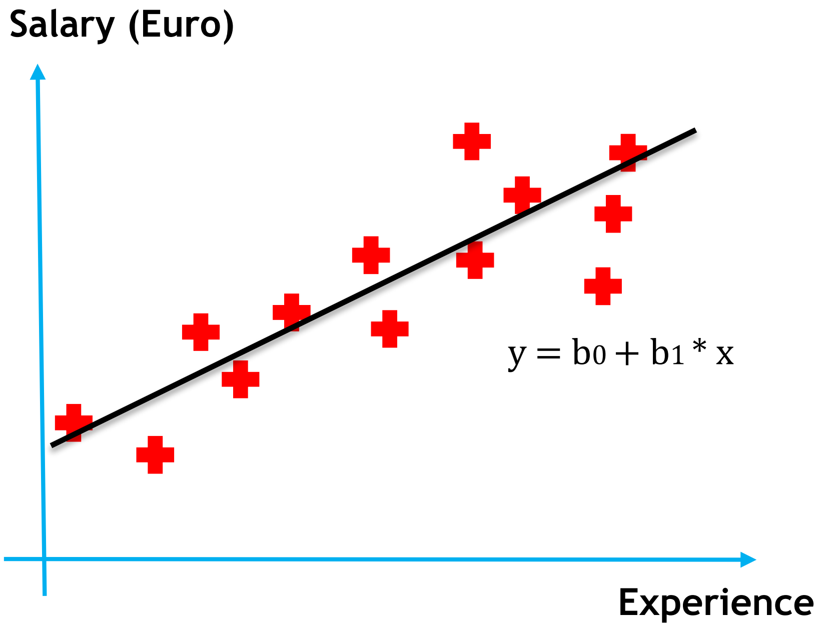


Logistic Regression

This is new:

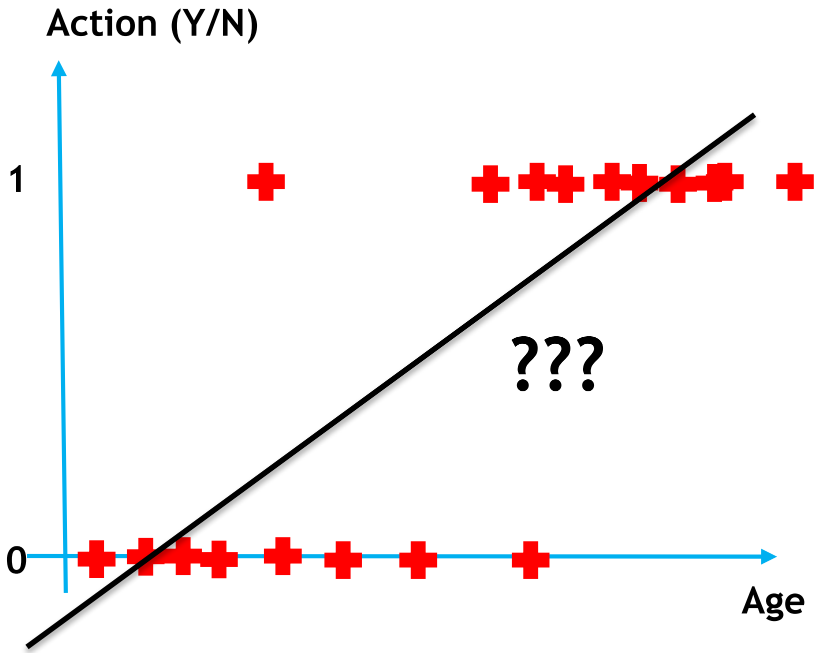


We know this:

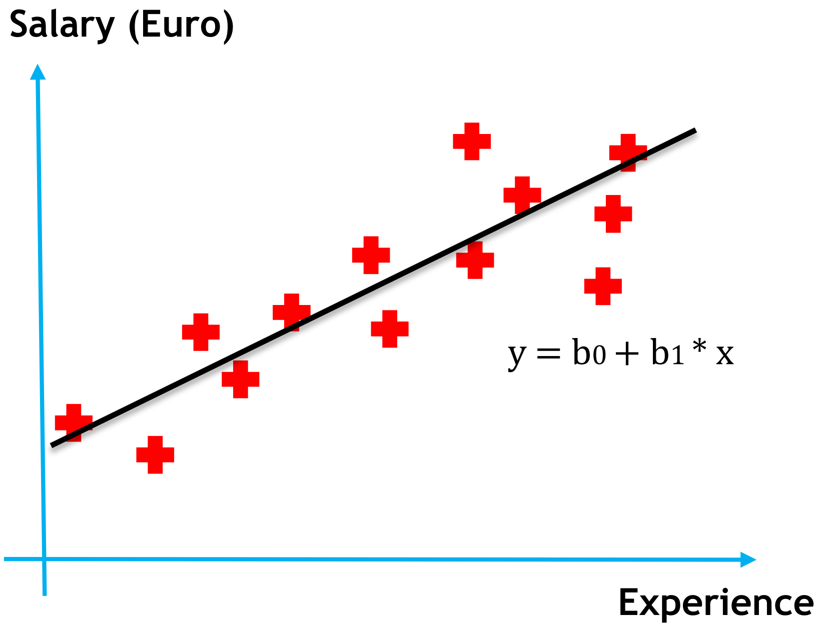


Logistic Regression

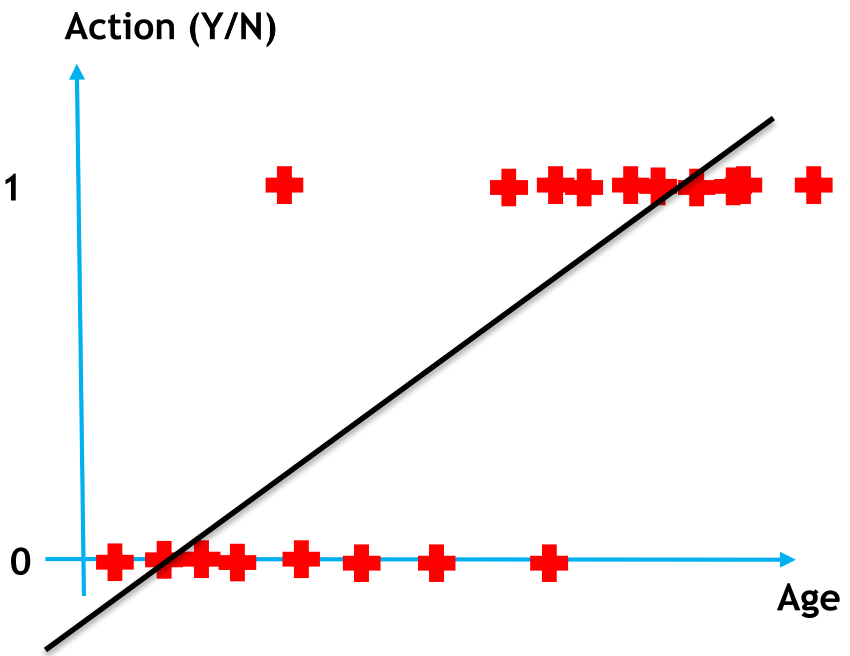
This is new:



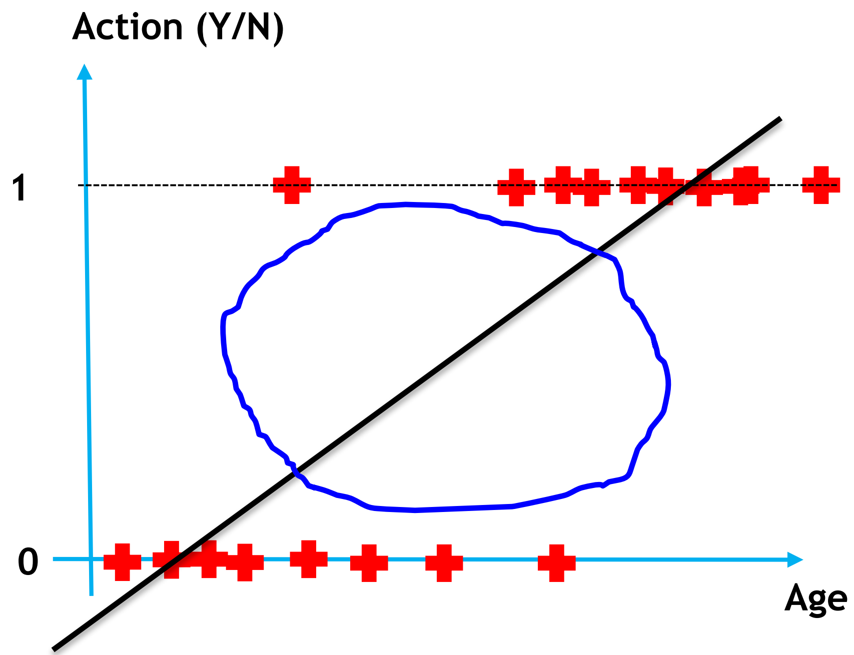
We know this:



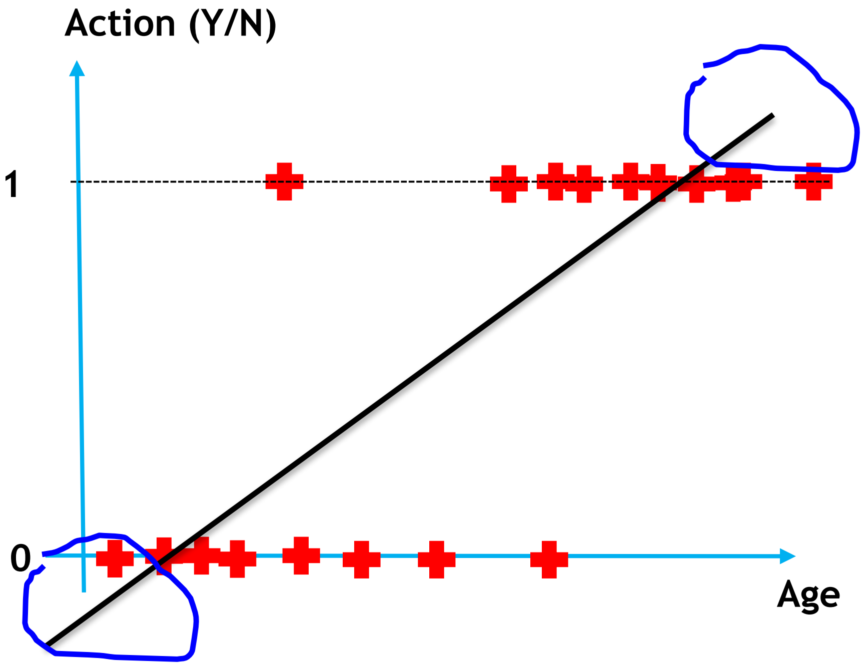
Logistic Regression: what if predict the probability or likelihood a person taking the offer?



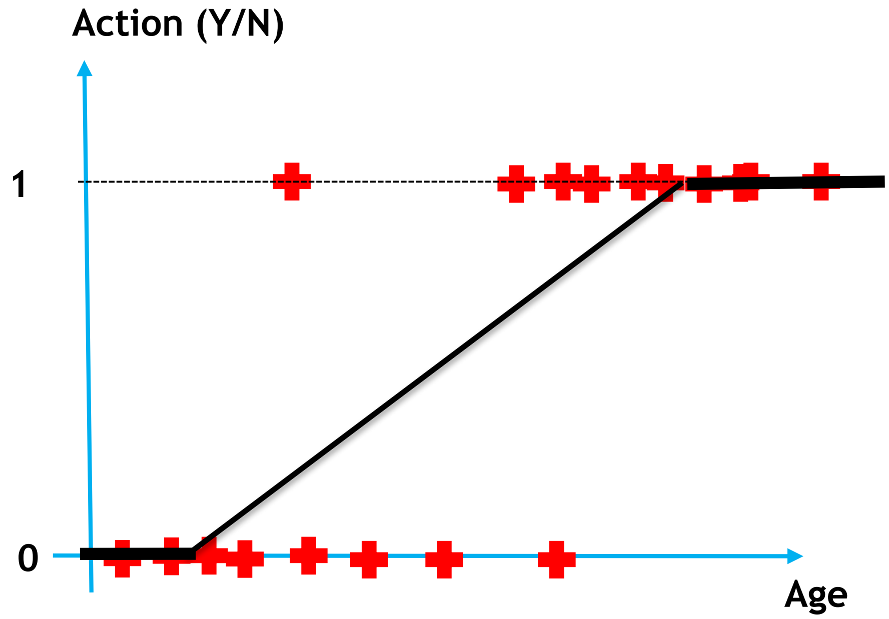
Logistic Regression: this part make sense to get probabilities



Logistic Regression: this part does not make sense to get probabilities (<0>1)

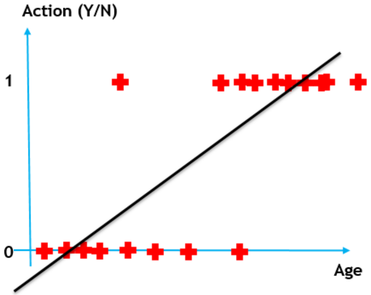


Logistic Regression



Logistic Regression

$$y = b_0 + b_1 * x$$

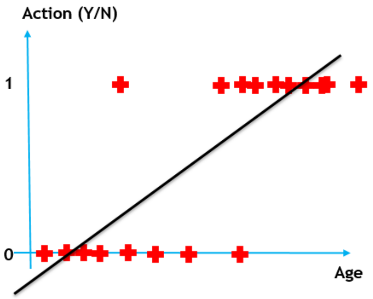


Logistic Regression

$$y = b_0 + b_1 * x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$



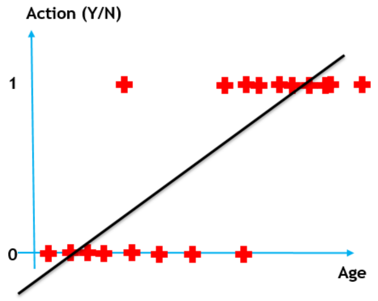
Logistic Regression

$$y = b_0 + b_1 * x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\boxed{ln\,(\frac{p}{1-p}) = b_0 + b_1 * x}$$



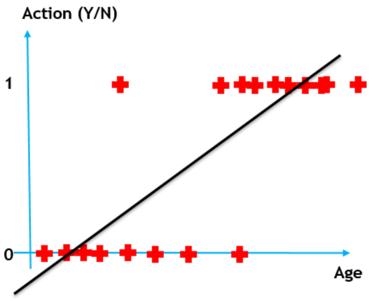
Logistic Regression

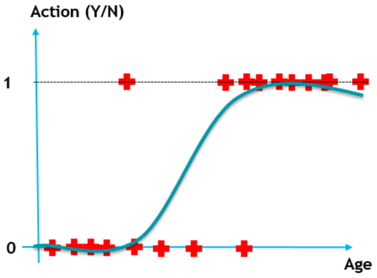
$$y = b_0 + b_1 * x$$

Sigmoid Function

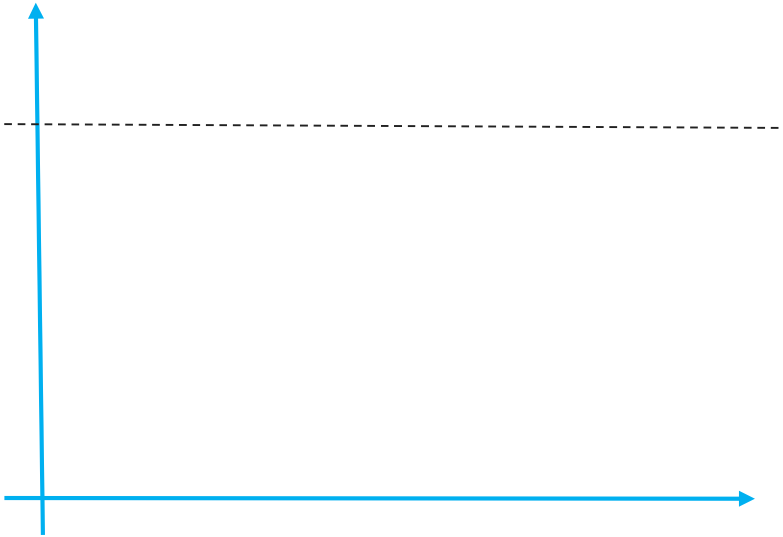
$$p = \frac{1}{1 + e^{-y}}$$

$$\boxed{ln\,(\frac{p}{1-p}) = b_0 + b_1 * x}$$

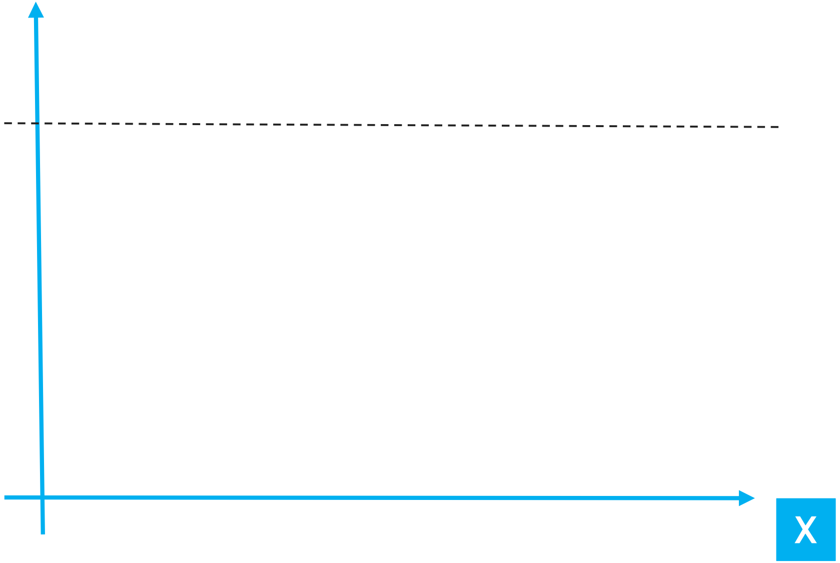




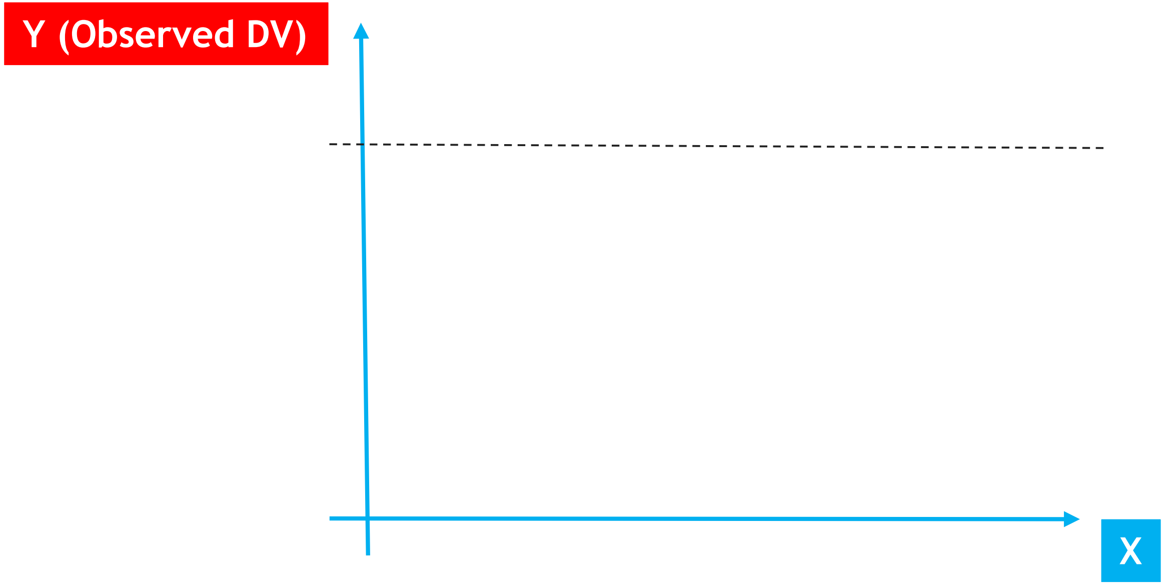
Logistic Regression



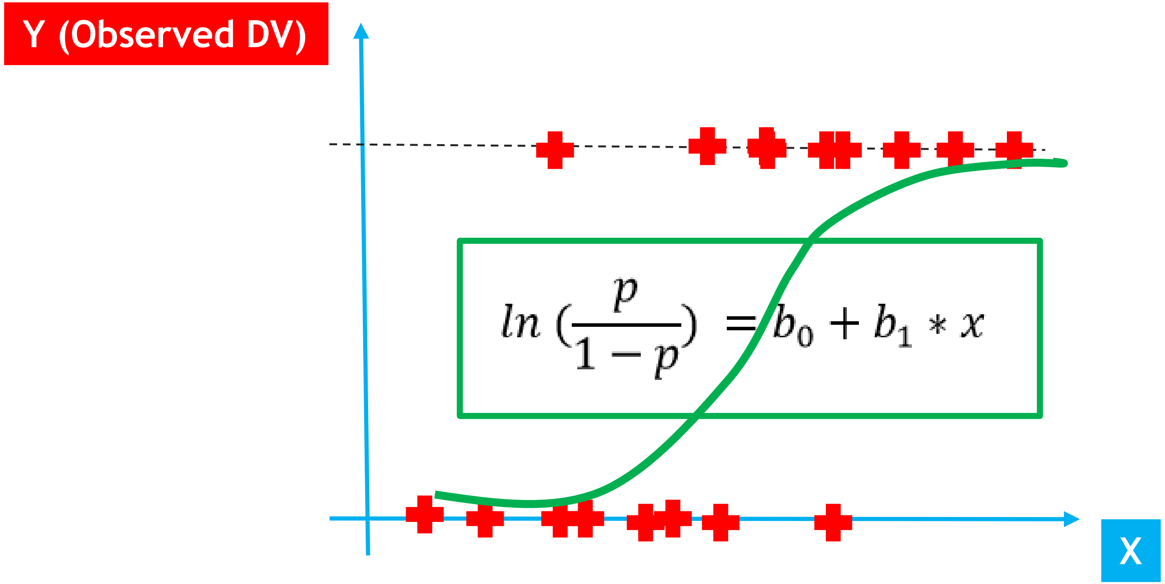
Logistic Regression



Logistic Regression

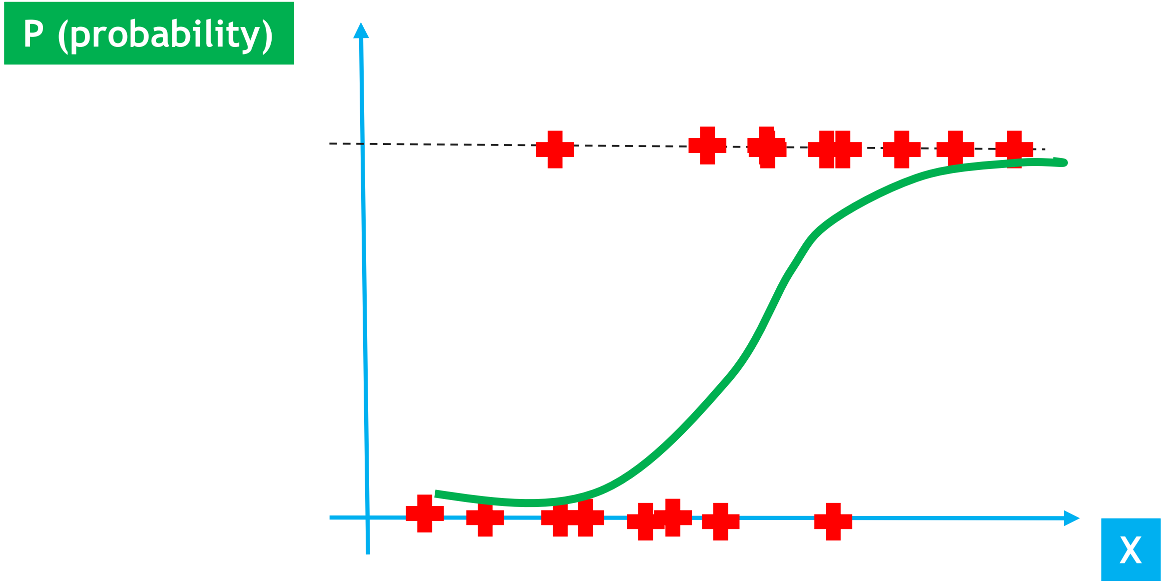


Logistic Regression: best fitting line that fits our data

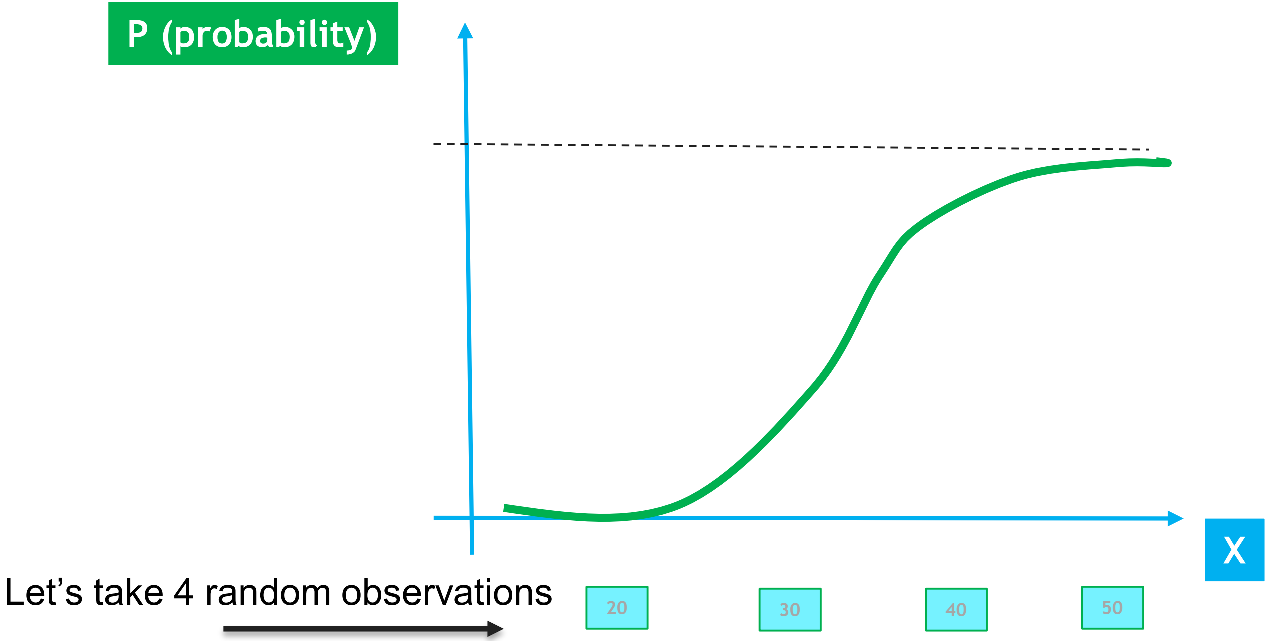


$$ln\,(\frac{p}{1-p}) = b_0 + b_1 * x$$

Logistic Regression: we created the model

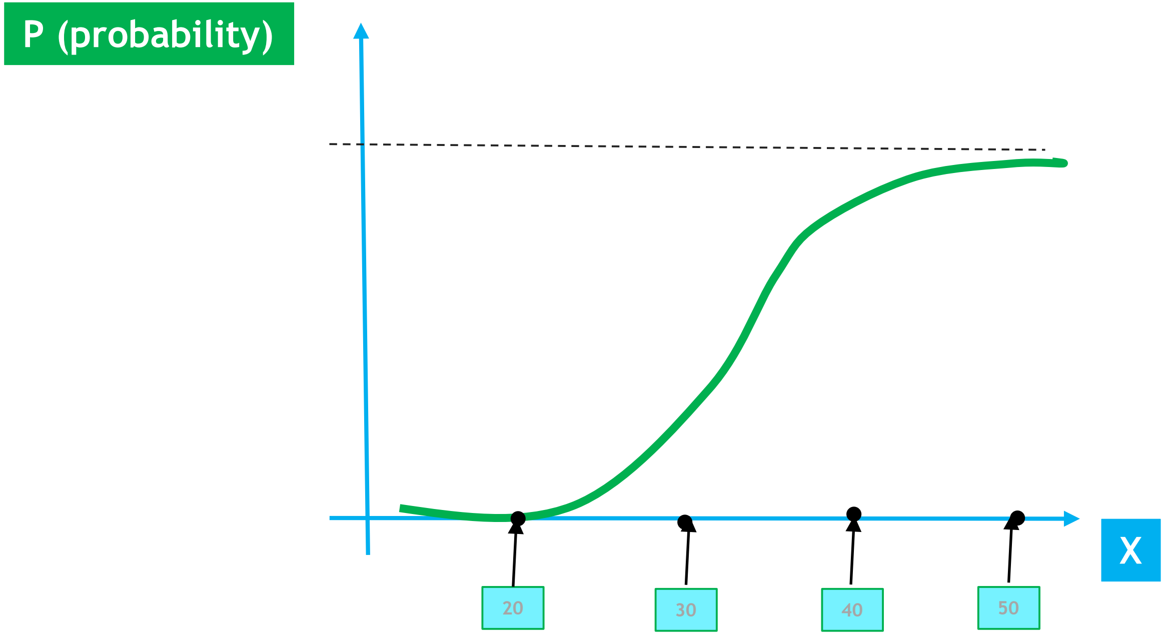


Logistic Regression: we predict this probability

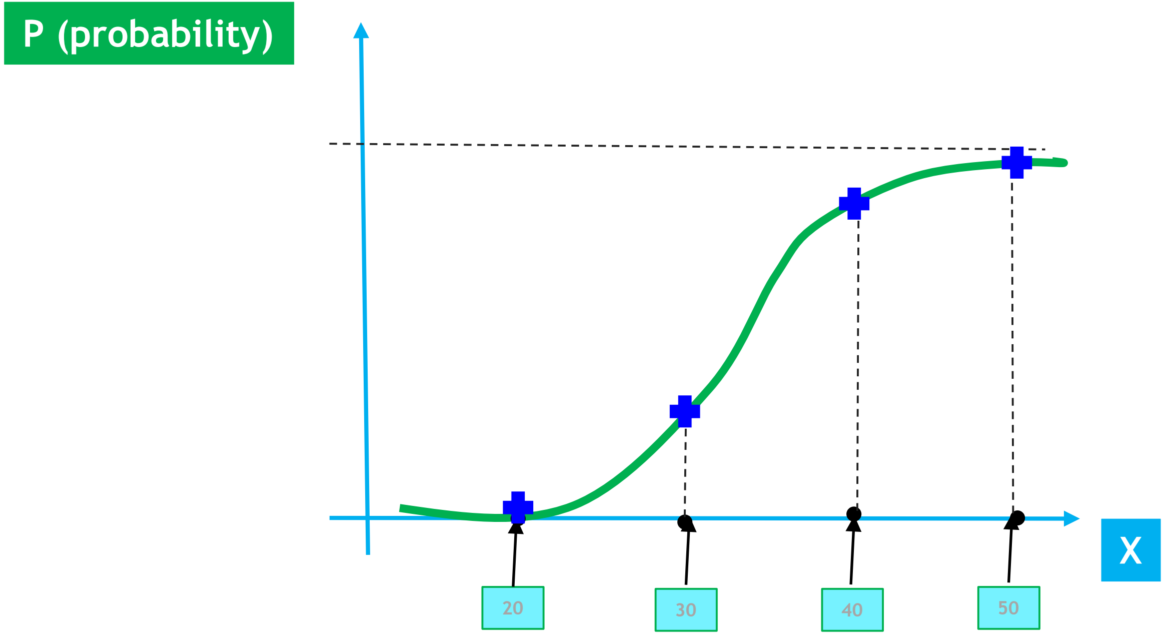


Let’s take 4 random observations

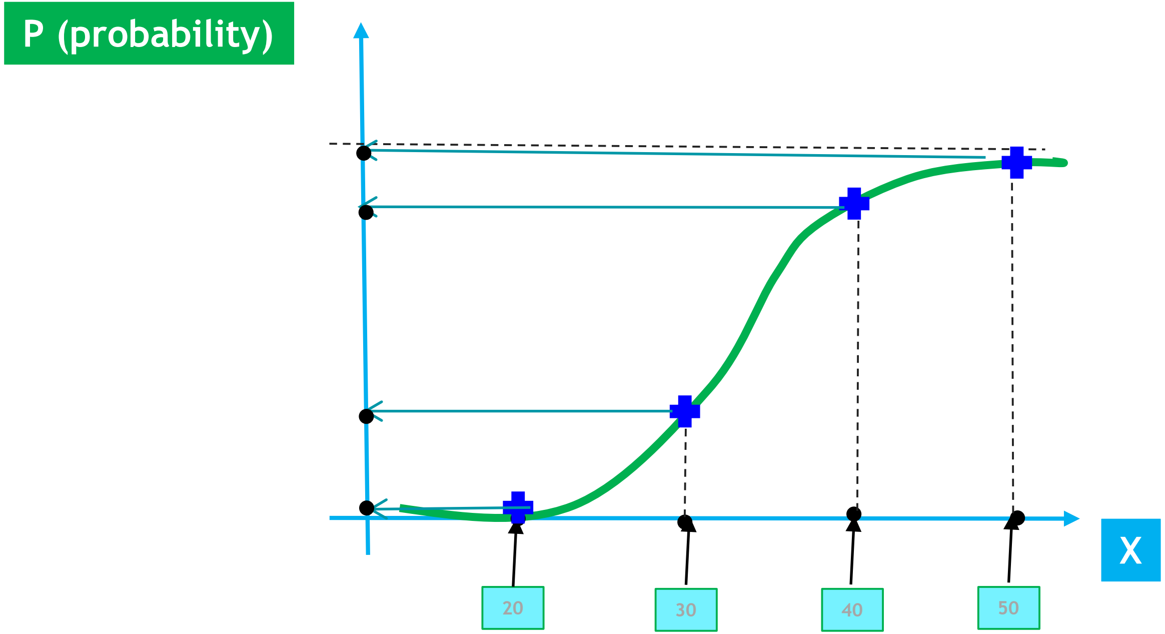
Logistic Regression: what we need to do to get the probability of this 4 random variables?



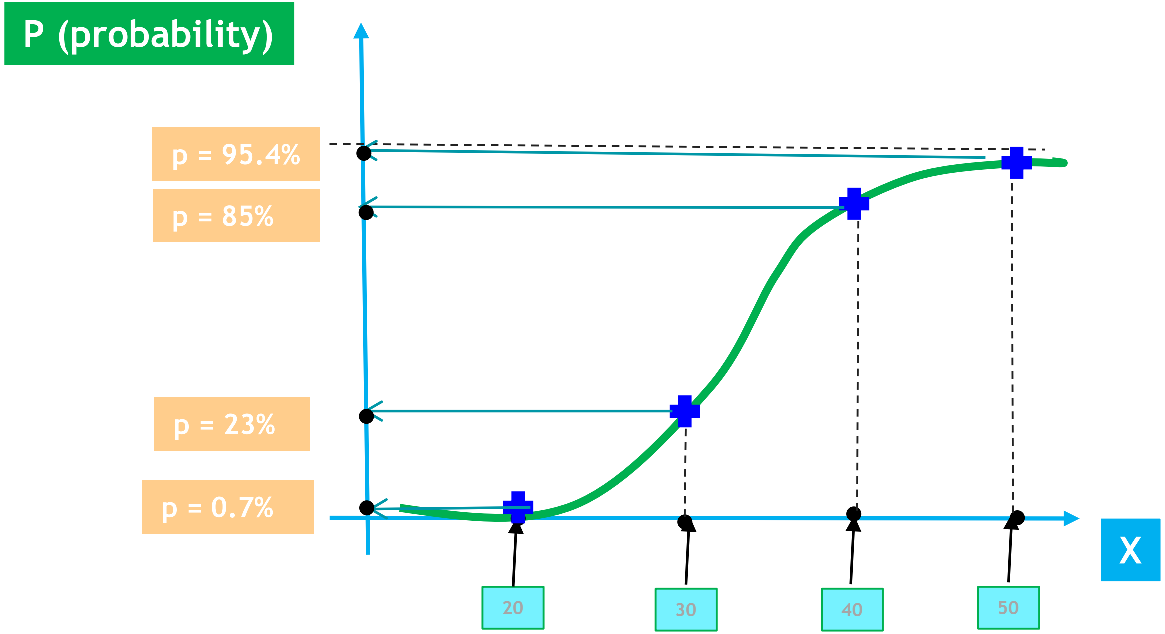
Logistic Regression: we need to project them in the curve



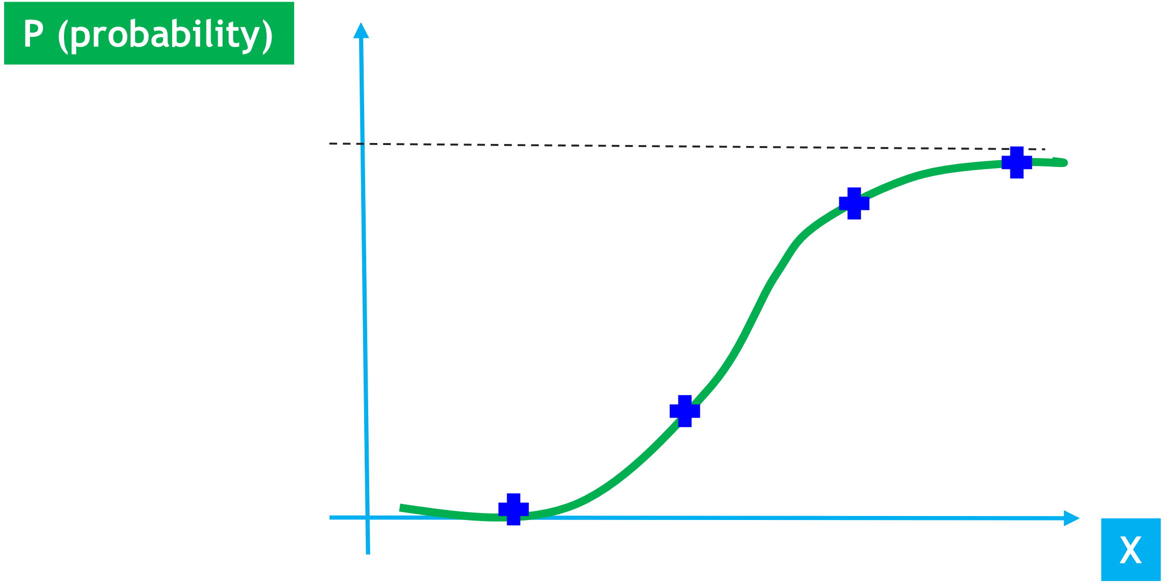
Logistic Regression: if you need the probabilities, then it has to be projected to the left



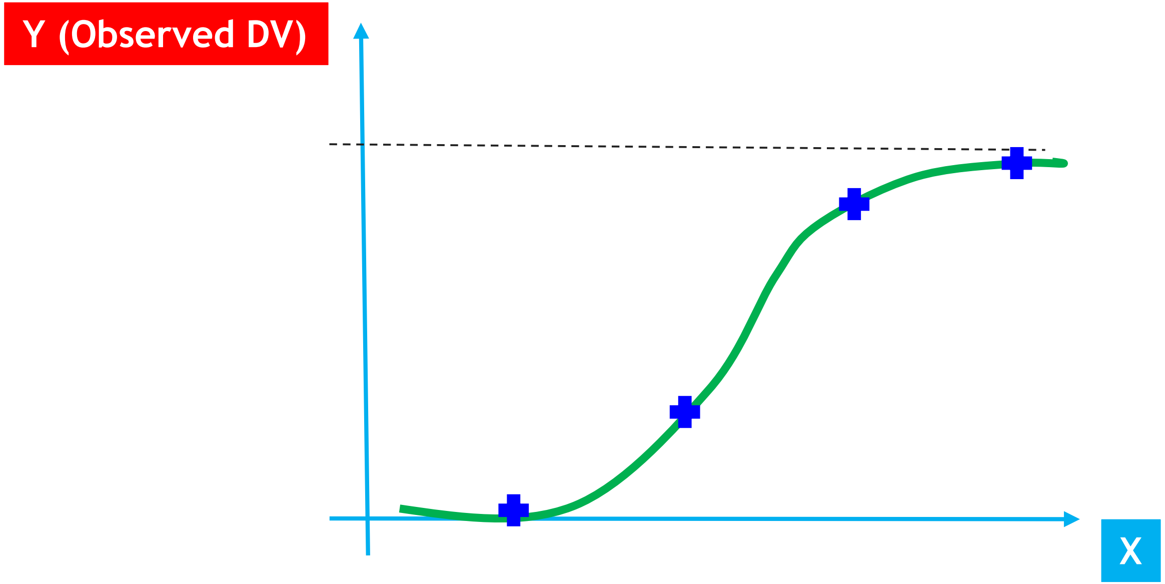
Logistic Regression: who is the least/most likely to take the offer?



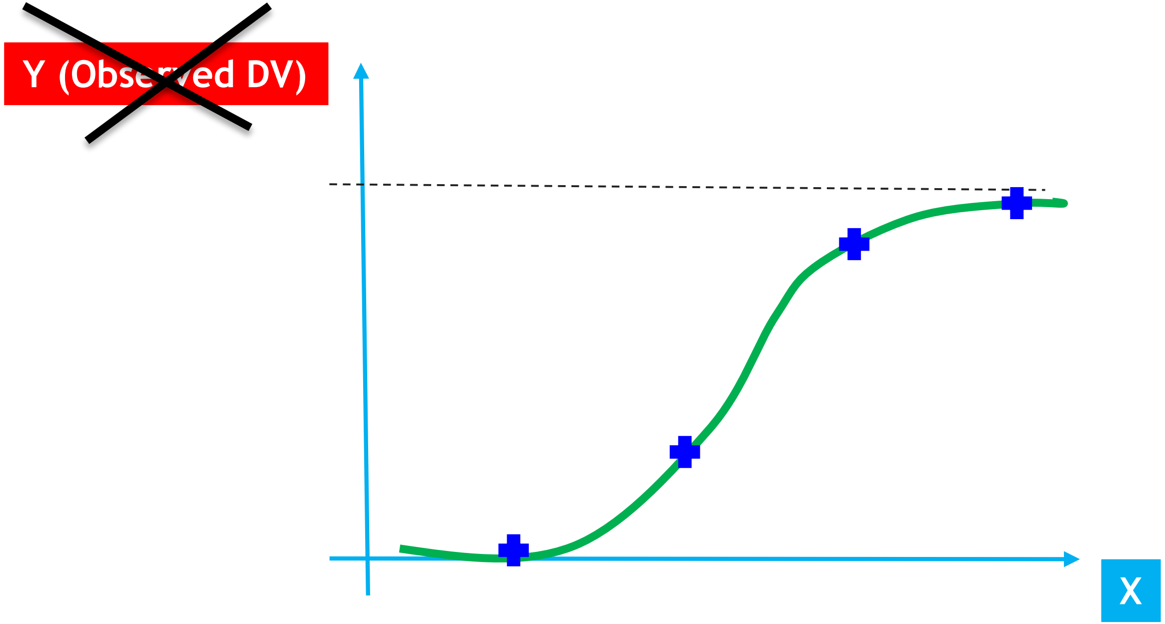
Logistic Regression: I want a prediction instead probabilities



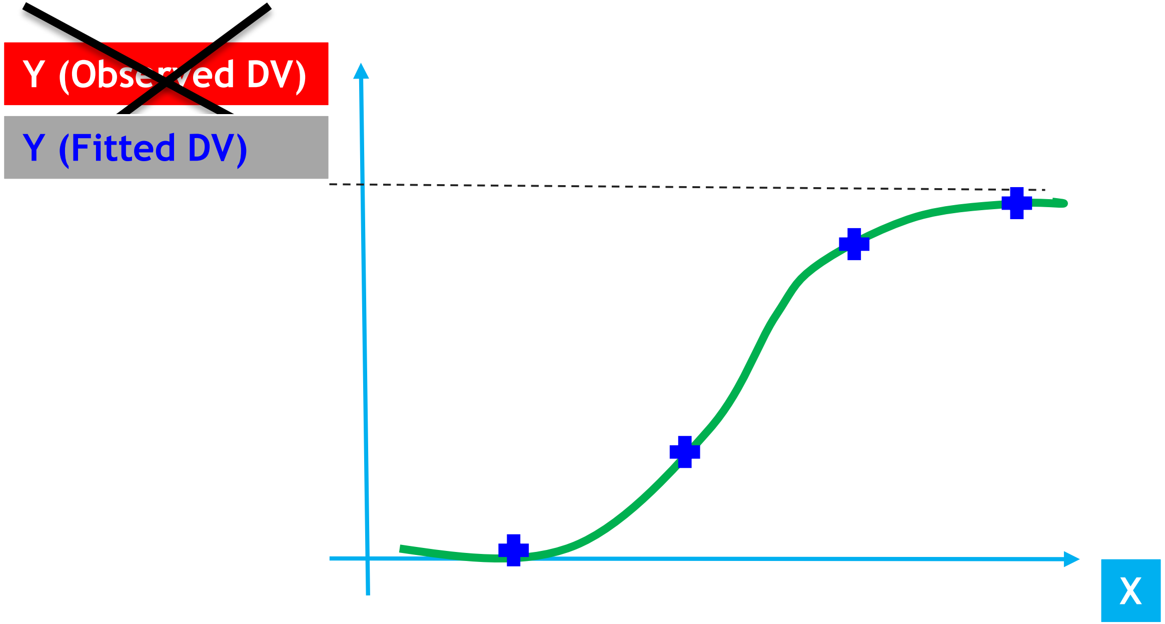
Logistic Regression: get predicted (fitted) values



Logistic Regression: get predicted (fitted) values



Logistic Regression: get predicted (fitted) values

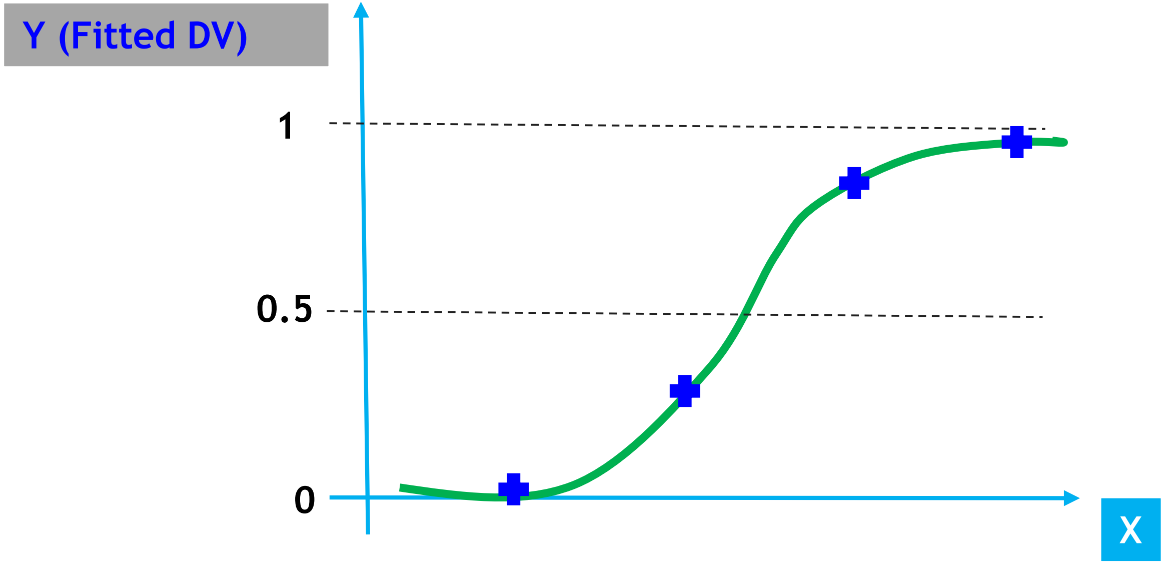


~~Y (Observed DV)~~

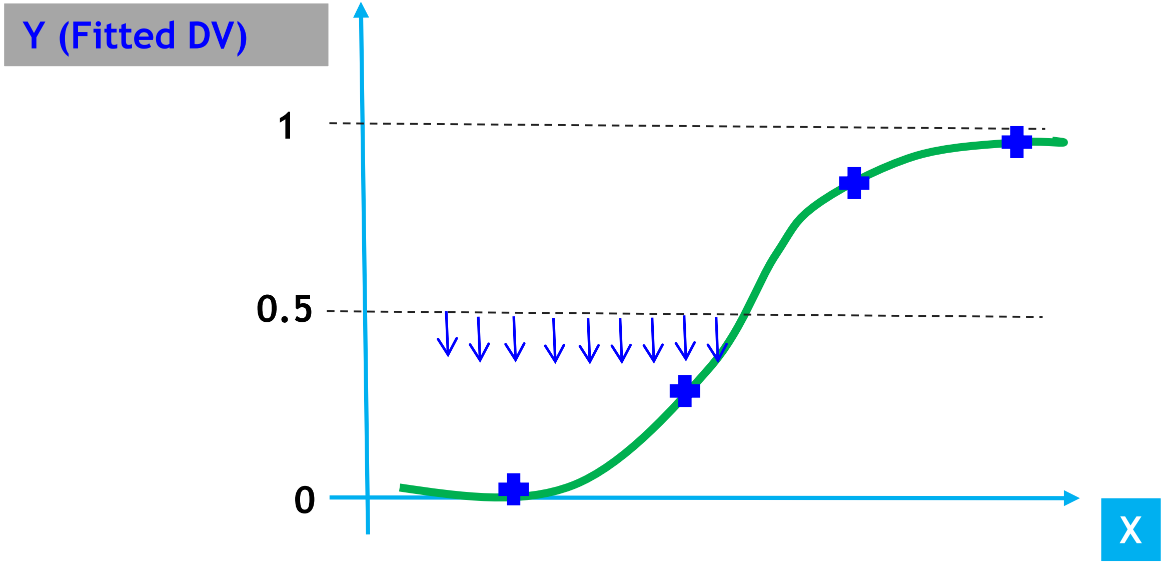
Y (Fitted DV)

X

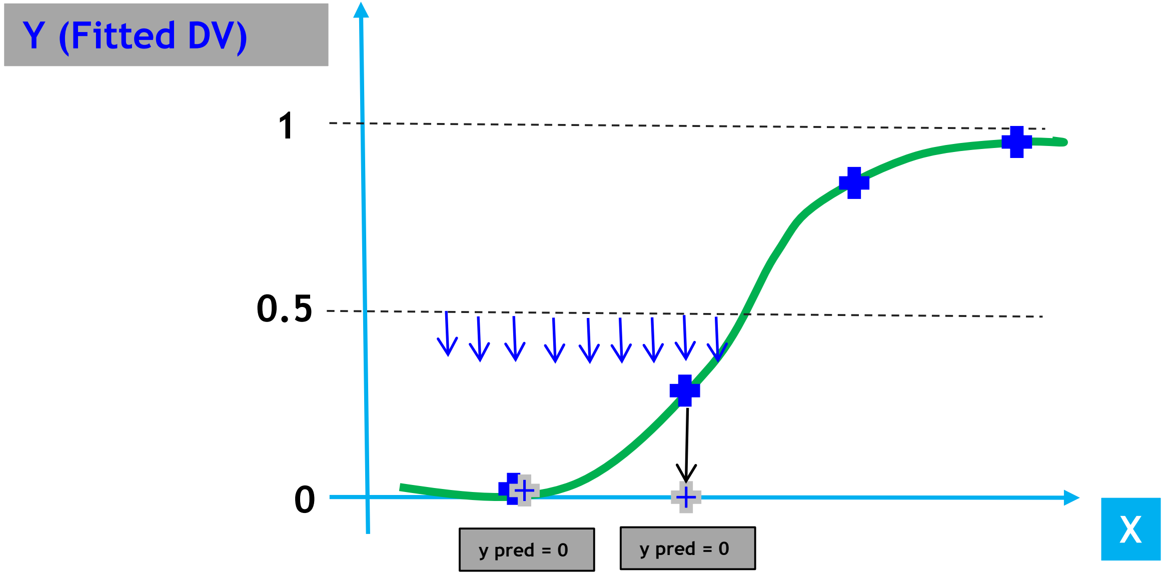
Logistic Regression: select a threshold line



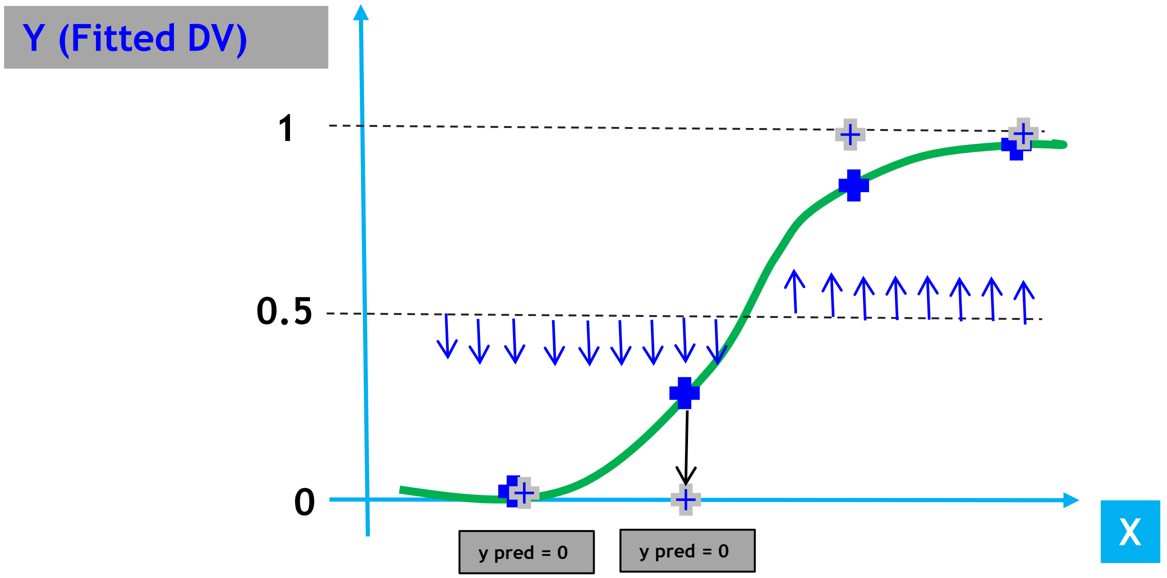
Logistic Regression: if predicted probability is less than 50% then we predict 0



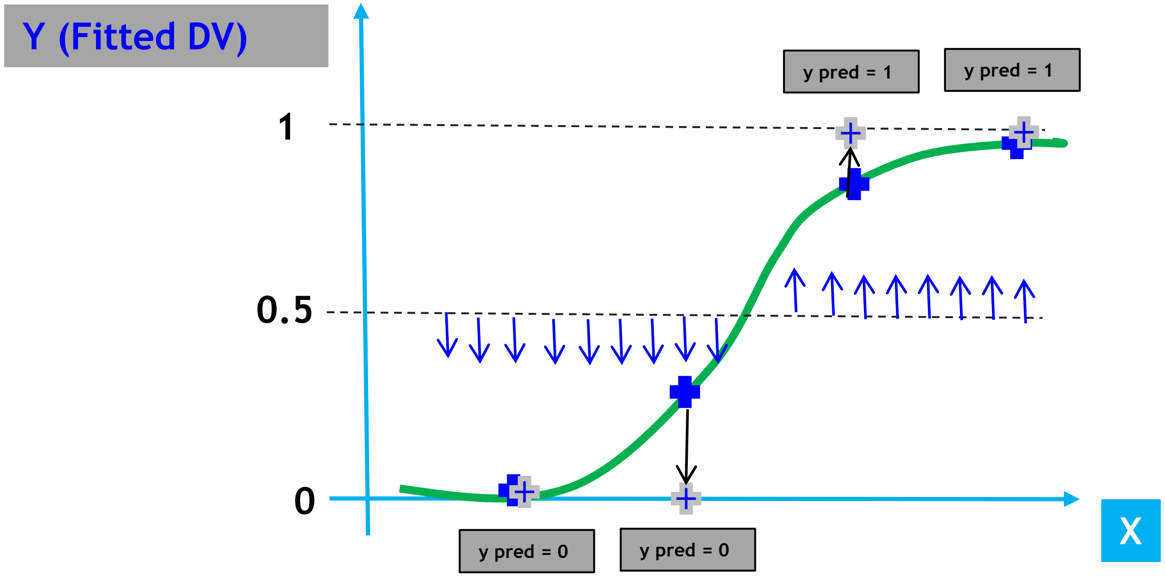
Logistic Regression: if predicted probability is less than 50% then we predict 0



Logistic Regression: anything above this threshold are predicted YES



Logistic Regression: anything above this threshold are predicted YES



Let’s do it in R!!