When we discuss solving classification problems, **Logistic Regression** should be the first supervised learning type algorithm that comes to our mind and is commonly used by many data scientists and statisticians. It is fundamental, powerful and easy to implement. More importantly, its basic theoretical concepts are integral to understanding deep learning. So, I believe everyone who is passionate about machine learning should have acquired a strong foundation of **Logistic Regression** and theories behind the code on Scikit Learn.

Alright...Let's start uncovering this mystery of Regression (the transformation from Simple Linear Regression to Logistic Regression)!

**Linear Regression**

I believe that everyone should have heard or even have learnt **Linear model** in Mathethmics class at high school. **Linear regression** is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line. There are two types of linear regression- **Simple** and **Multiple**.

**Quick reminder**: 4 Assumptions of Simple Linear Regression

1. *Linearity: The relationship between X and the mean of Y is linear.*

2. *Homoscedasticity: The variance of residual is the same for any value of X (Constant variance of errors).*

3. *Independence: Observations are independent of each other.*

4. *Normality: The error(residuals) follow a normal distribution.*



$$y = b_0 + b_1 * x$$
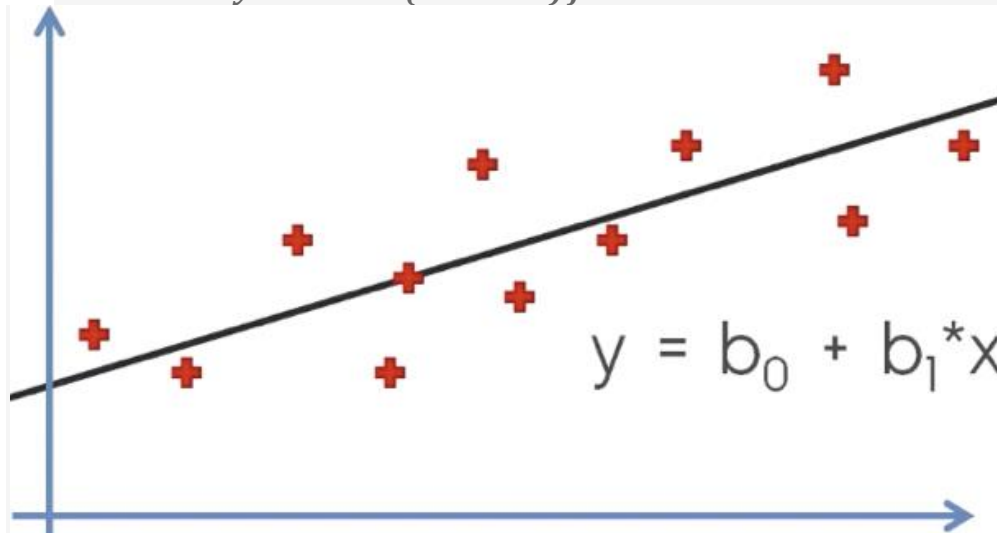
Simple Linear Regression with one explanatory variable (x): *The red points are actual samples, we are able to find the black curve (y), all points can be connected using a (single) straight line with linear regression.*

$$y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots + \beta n Xn$$

The equation of Multiple Linear Regression: X1, X2 … and Xn are explanatory variables

You might have a question "**How to draw the straight line that fits as closely to these (sample) points as possible?**" The most common method for fitting a regression line is the method of **Ordinary Least Squares** used to minimize the **sum of squared errors (SSE)** or **mean squared error (MSE)** between our observed value(yi) and our predicted value (ŷi).
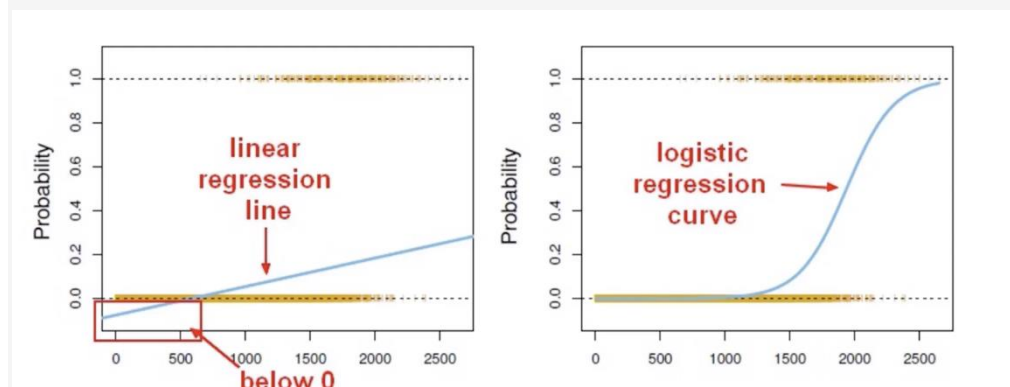


Simple Linear Regression:

$$\text{SUM } (y_i - \hat{y_i})^2 \rightarrow \min$$
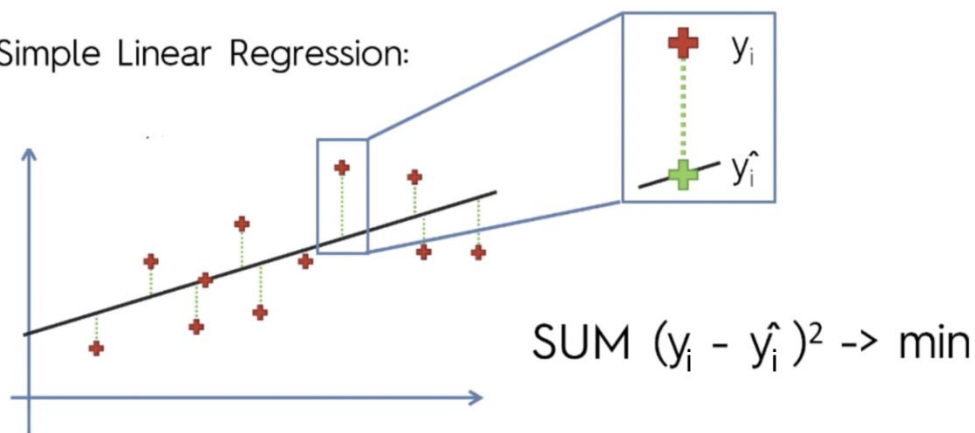
Residual: **e** = y — ŷ (Observed value — Predicted value)

# From Numerical to Binary

Now we have a classification problem, we want to predict the binary output variable Y (2 values: either 1 or 0). For example, the case of flipping a coin (Head/Tail). The response yi is binary: 1 if the coin is Head, 0 if the coin is Tail. This is represented by a Bernoulli variable where the probabilities are bounded on both ends (they must be between 0 and 1).



The linear regression line is below 0

Linear regression is only dealing with continuous variables instead of Bernoulli variables. The problem of Linear Regression is that these predictions are not sensible for classification since the true probability must fall between 0 and 1 but it can be larger than 1 or smaller than 0. Noted that classification is not normally distributed which is violated assumption 4: Normality. Moreover, both mean and variance depend on the underlying probability. Any factor that affects the probability will change not just the mean but also the variance of the observations which means the variance is no longer constantly violating the assumption 2: Homoscedasticity. As a result, we cannot directly apply linear regression because it won't be a good fit.

So...how can we predict a classificiation problem?



# Instead we can transform our liner regression to a logistic regression curve!

As we are now looking for a model for probabilities, we should ensure the model predicts values on the scale from 0 to 1. A powerful model Generalised linear

model (GLM) caters to these situations by allowing for response variables that have arbitrary distributions (other than only normal distributions), and by using a link function to vary linearly with the predicted values rather than assuming that the response itself must vary linearly with the predictor. As a result, GLM offers extra flexibility in modelling. In this case, we need to apply the logistic function (also called the 'inverse logit' or 'sigmoid function').

## Logistic Regression is a type of Generalized Linear Models.

Before we dig deep into logistic regression, we need to clear up some of the fundamentals of statistical terms — **Probablility** and **Odds.**
The probability that an event will occur is the fraction of times you expect to see that event in many trials. If the probability of an event occurring is Y, then the probability of the event not occurring is 1-Y. Probabilities always range between 0 and 1.

The odds are defined as the probability that the event will occur divided by the probability that the event will not occur. Unlike probability, the odds are not constrained to lie between 0 and 1, but can take any value from zero to infinity.

If the probability of Success is P, then the odds of that event is:

$$odds = \frac{P}{1 - P}$$

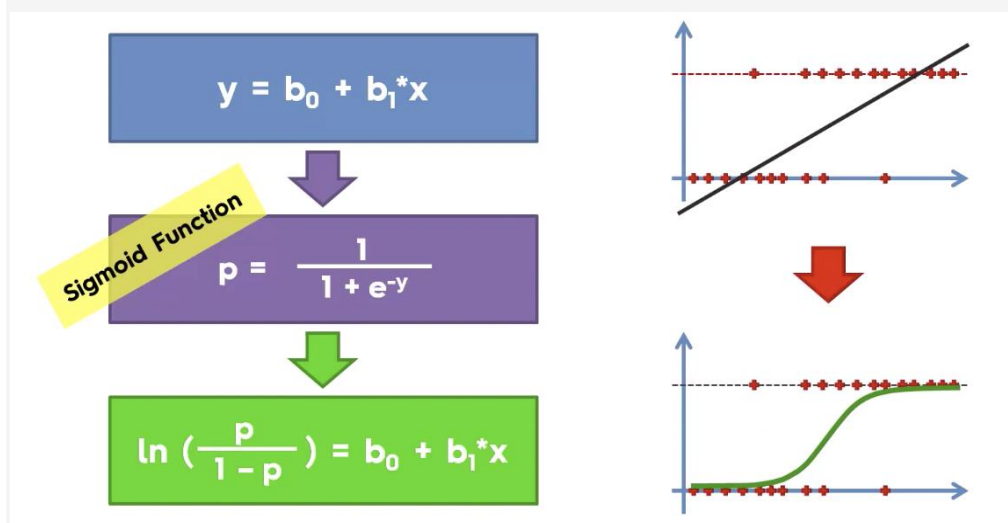| Y | 1 | 0 |
|---|---|---|
| Pr(Y=1) | P | 1- P |

*P= Success, 1-P = Failure

Odds: Success/ Failure

Example: If the probability of success (P) is 0.60 (60%), then the probability of failure(1-P) is 1–0.60 = 0.40(40%). Then the odds are 0.60 / (1–0.60) = 0.60/0.40 = 1.5.

It's time.... to transform the model from linear regression to logistic regression using the logistic function.

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

$$\frac{P}{1-P} = e^{b_0 + b_1 x}$$

$$P = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

In (odd)=bo+b1x



**logistic function** (also called the '**inverse logit**').

We can see from the below figure that the output of the linear regression is passed through a sigmoid function (logit function) that can map any real value between 0 and 1.

# Logistic Regression is all about predicting binary variables, not predicting continuous variables.

Don't get confused with the term 'Regression' presented in Logistic Regression. *I know it's pretty confusing, for the previous 'me' as well* 😀
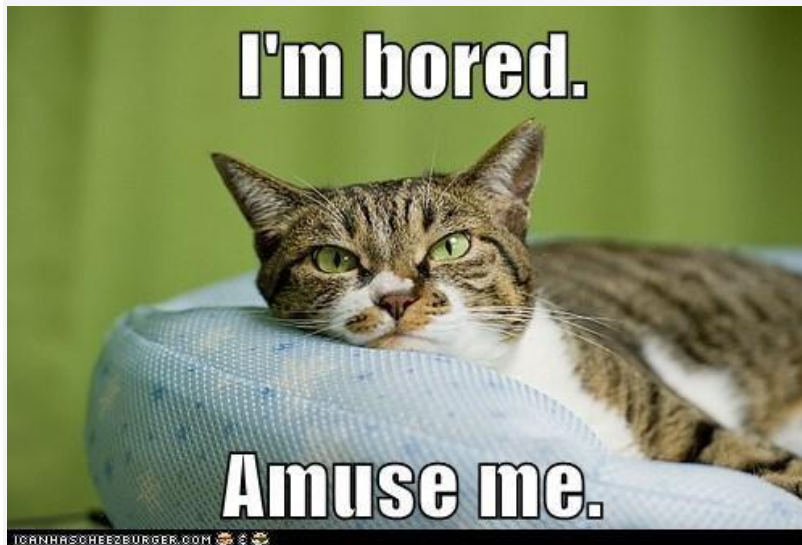


Congrats~you have gone through all the theoretical concepts of the regression model. Feel bored?! Here's a real case to get your hands dirty!

Imagine that you are a store manager at the APPLE store, increasing 10% of the sale revenue is your goal this month. Therefore, you need to know who the potential customers are in order to maximise the sale amount.

**The client information you have is including Estimated Salary, Gender, Age and Customer ID.**

| | User ID | 15714658 | 15727311 | 15624510 |
|---|---|---|---|---|
| | Age | 48 | 35 | 19 |
| | Salary | $41,000 | $65,000 | $19,000 |
| | Purchased? | ✅ | ❌ | ❌ |

3 samples of client information

If now we have a new potential client who is 37 years old and earns $67,000,can we predict whether he will purchase an iPhone or not (Purchase?/ Not purchase?)



Purchase?
Not purchased?

| | Age | 37 |
|---|---|---|
| | Salary | $67,000 |

# Topic: Calculating Probability in Logistic Regression

The dataset of pass/fail in an exam for 5 students is given in the table below.
If we use **Logistic Regression** as the classifier and assume the model suggested by
the optimizer will become the following for Odds of passing a course:
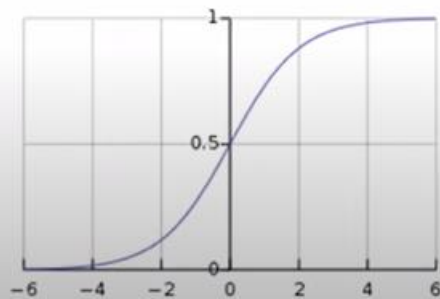
$$\log(Odds) = -64 + 2 \times hours$$

| Hours Studies | Result (1=Pass, 0=Fail) |
|:---:|:---:|
| 29 | 0 |
| 15 | 0 |
| 33 | 1 |
| 28 | 1 |
| 39 | 1 |

(1) How to calculate the **probability of Pass** for the student who studied 33 hours?

(2) **At least how many hours** the student should study that makes sure will pass the
course with the probability of more than 95%?

## Sigmoid Function

$$S(x) = \frac{1}{1 + e^{-x}}$$



$$P = \frac{1}{1 + e^{-z}}$$

$$Z = -64 + 2 \cdot Hours \qquad \# Hours = 33$$

$$Z = -64 + 2 \cdot 33$$

$$-64 + 66$$

$$Z = 2$$

$$P = \frac{1}{1 + e^{-2}}$$

$P = 0.88$      $\therefore$ A student who studies for 33 Hours
                    has a 88% Chance of passing
                    the course.

Q2.  $\underline{P} = \boxed{0.95}$          #  $P = \dfrac{1}{1 + e^{-z}}$

$\underset{\times \, 1 + e^{-z}}{0.95} = \underset{1 + e^{-z}}{\dfrac{1}{}} \quad \times \underset{}{1 + e^{-z}}$     # Solve for z

$0.95 \cdot (1 + e^{-z}) = 1$

$0.95 + 0.95 e^{-z} = 1$

$0.95 e^{-z} = 1 - 0.95$

$0.95 e^{-z} = 0.05$

$\dfrac{0.95 e^{-z}}{0.95} = \dfrac{0.05}{0.95}$     # Divide both sides
                          by 0.95

$e^{-z} = 0.0526$

$\ln(e^{-z}) = \ln(0.0526)$

$$\ln(e^{-z}) = \ln(0.0526) \quad \text{\# } \ln(e^{-z}) \to -z$$

$$-z = -2.94$$

$$\frac{-z}{-1} = \frac{-2.94}{-1} \quad \text{\# Divide } -1$$

$$\boxed{z = 2.94}$$

Refer back to the log (odds) equuation

$$\log(\text{odds}) = -64 + 2 \cdot \text{Hours}$$
$$z = -64 + 2 \cdot \text{Hours} \quad \to \quad \text{\# Solve for Hours}$$

Refer back to the log (odds) equuation

$$\log(\text{odds}) = -64 + 2 \cdot \text{Hours}$$
$$z = -64 + 2 \cdot \text{Hours} \quad \to \quad \text{\# Solve for Hours}$$
$$z = 2.94$$
$$2.94 = -64 + 2 \cdot \text{Hours}$$

$$2.94 + 64 = 2 \cdot \text{Hours}$$

$$66.94 = 2 \cdot \text{Hours}$$

$$\frac{66.94}{2} = \frac{2 \cdot \text{Hours}}{2}$$

\# Divided by 2 to solve for Hours

$$33.47 = \text{Hours}$$

$$\text{Hours} = 33.5 \text{ Hours}$$

Verify $\Rightarrow$

$$z = -64 + 2 \cdot 33.5$$
$$= -64 + 67$$
$$z = 3$$

$\therefore$ Student needs to study for approx. 33.5 hours

$$P = \frac{1}{1 + e^{-3}} \approx 0.952$$