Regression Analysis

Nisal Mihiranga

Facilitator

Nisal Mihiranga

Areas of Interest & Expertise:
AI, Technology, Science, Teaching, Consulting, Mentoring

Experience:
Head of AI and Data Science, Architect at
Zone24x7 pvt Ltd
Corporate Trainer

12 Years of Industry exposure to Data Engineering, Data Science and Business Intelligence

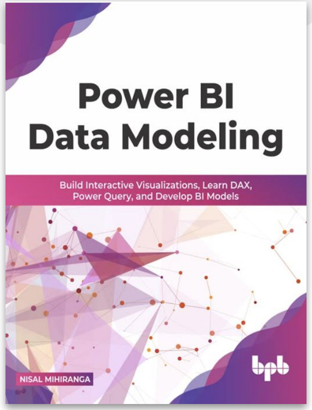
Credentials:
M.Sc in Data Science

B.Sc in Information Technology

Microsoft Certified Trainer

Microsoft MVP Most Valuable Professional

Microsoft MCT



Curriculum

PROFESSIONAL DEVELOPMENT UNIT

Week	Module
Week 1	Python for Machine Learning
Week 2	Introduction to Machine Learning
Week 3	Data Transformation and Analysis
Week 4	Classification
Week 5	**Regression**
Week 6	Clustering Algorithms
Week 7	Neural Networks
Week 8	MLOPS, Machine Learning in Cloud

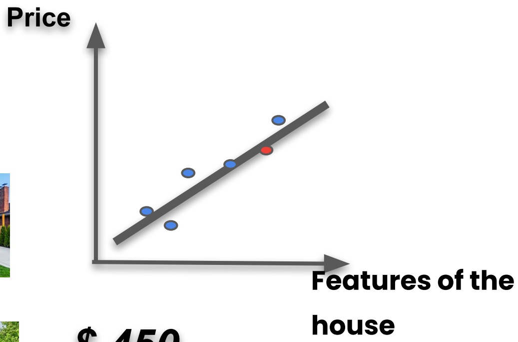
Learning Outcomes

1. **Understand the Purpose and Types of Regression Analysis**: Learn the fundamental purpose of regression, which is to model and analyze relationships between variables, and differentiate between types (e.g., linear, multiple, logistic) based on use cases and data types.
2. **Evaluate Model Performance and Fit**: Learn to use performance metrics (e.g., Mean Squared Error, R-squared, Adjusted R-squared) to assess the quality of a regression model, determining how well it explains the variability in the data.

Regression Analysis

Regression

How much or How many ?



$ 540

$ 450

$ 390

$ 500

Supervised learning

How much will this cost ?

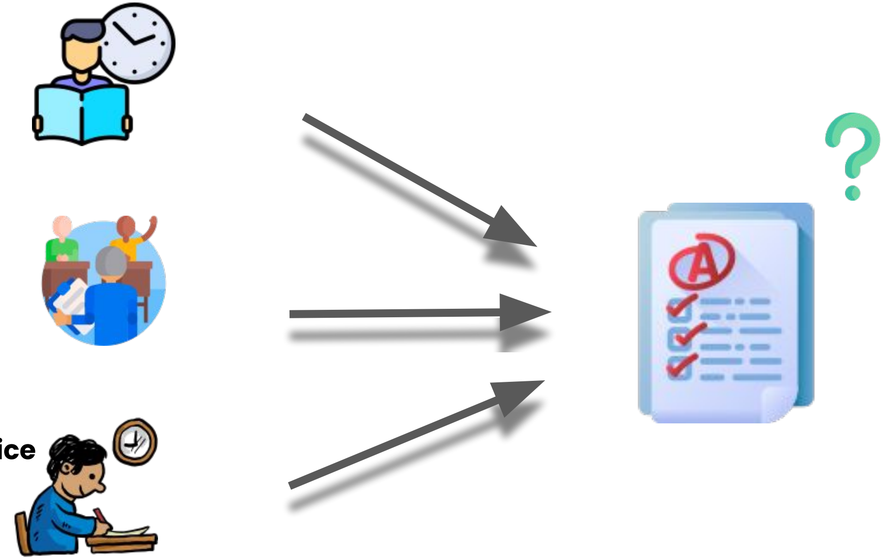
What Is Regression Analysis?

Regression analysis is statistical method that can allow to infer or predict one variable basis on one or more other variables.

Hours Studied

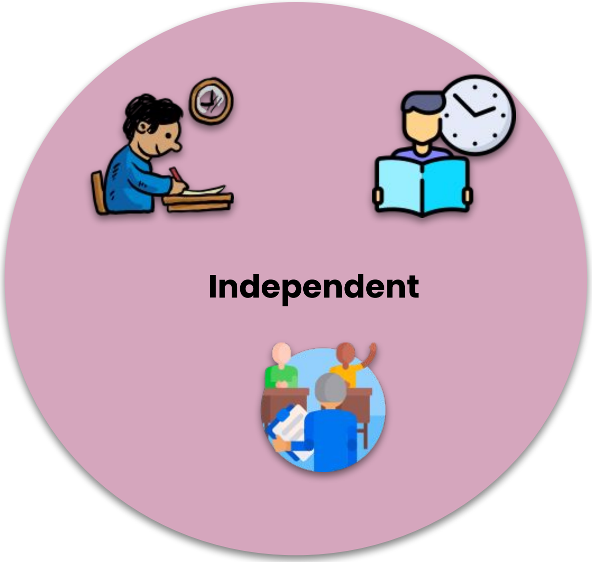
Attendance Percentage

Number of Practice tests



Dependent and Independent variable

- Variables used for prediction are called Independent variables.



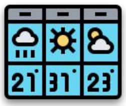
- Variable to be inferred is called Dependent variable.

When do we use Regression Analysis?

Regression analysis can be used to achieve two goals,

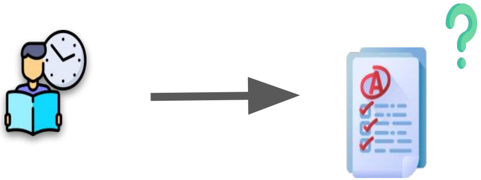
- Measurement of the influence of one or more variables on other variables.
    - **Economics Analysis**
      (Impact factors: Inflation, Interest Rate, etc...)
    - **Medical Research**
      (Blood pressure, cholesterol level, etc...)
    - **Marketing**
      (Advertising spread, Pricing Strategies, etc...)

- Regression analysis is also valuable for making predictions based on known data.
    - **Stock Market**
      (Predict stock price)
    - **Weather Forecasting**
      (Predict weather pattern)
    - **Customer Behaviors**
      (Predict customers behaviours)

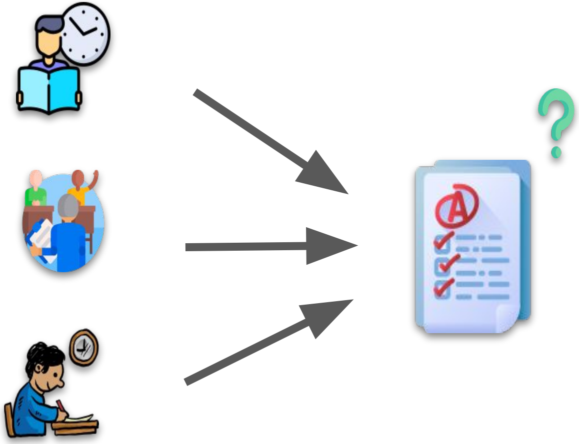


Types of Regression Analysis

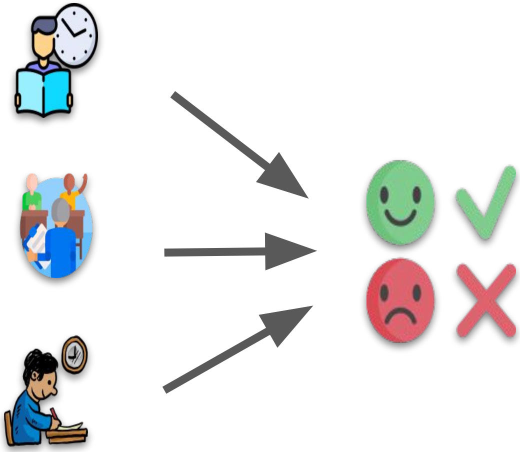
1. Simple Linear Regression



2. Multiple Linear Regression



3. Logistic Regression

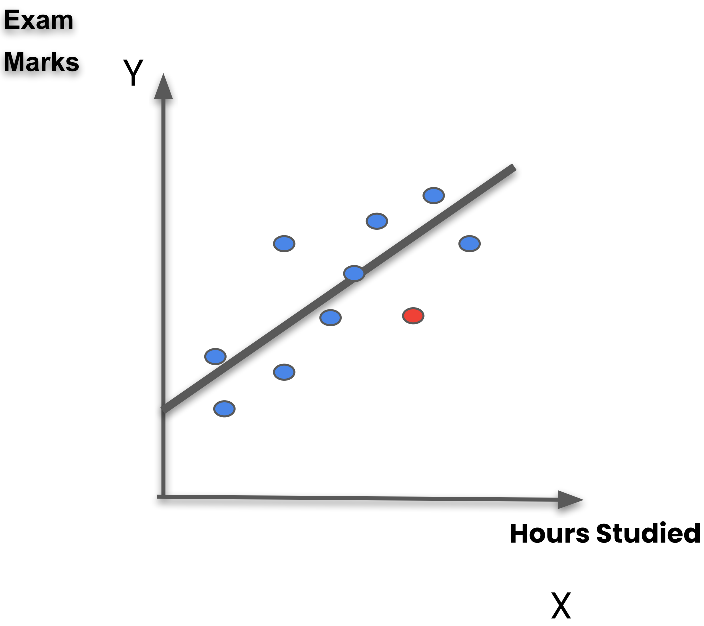


Simple Linear Regression

Linear Regression is used to find a relationship between a *dependent variable* (**Y**) and one or more *independent variables* (**X**). In simple linear regression, you have one independent variable, while in multiple linear regression, you have more than one.

Suppose you want to predict a student's test score (Y) based on the number of hours they study (X). The relationship can be represented as: **Y = aX + b**, where **'a'** is the ***slope*** and **'b'** is the ***y-intercept***.

Simple Linear Regression



Exam marks

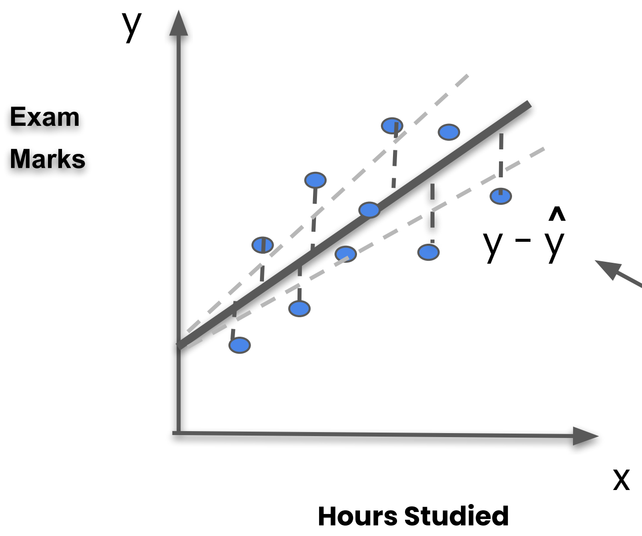
Hours Studied

$$y = a \,.\, x + b$$

Intercept

Slope

Fitting a line to data



Residuals

Regression Error

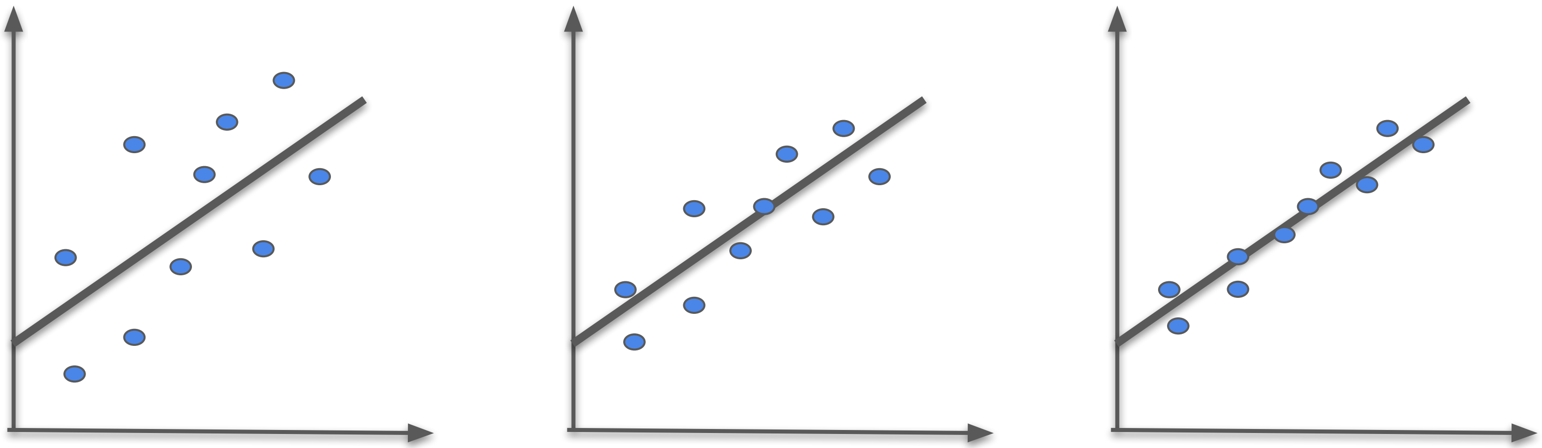
$$y = b \cdot x + a + e$$

Simple Linear Regression

PROFESSIONAL DEVELOPMENT UNIT

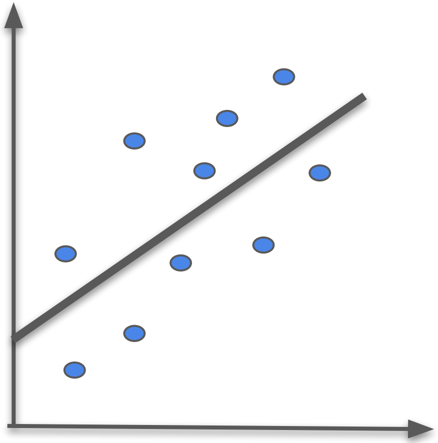
Linear Relationship

Low → **High**



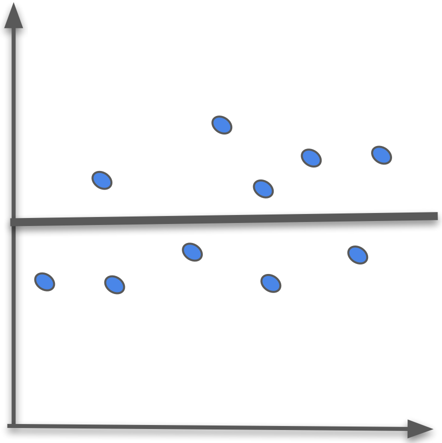
Regression Coefficient

There is positive correlation between x and y



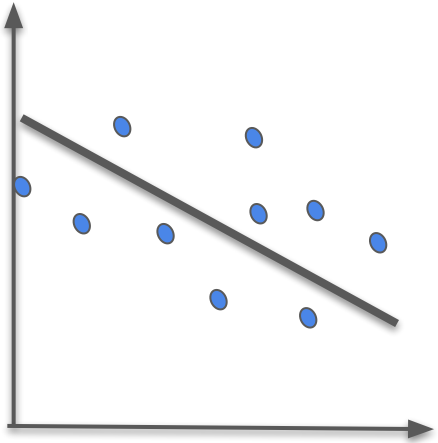
$b > 0$

There is no correlation between x and y



$b = 0$

There is negative correlation between x and y



$b < 0$

Activity

PROFESSIONAL DEVELOPMENT UNIT

https://mlu-explain.github.io/linear-regression

Implement Regression Model

There are several ways to implement linear regression, depending on the programming language and libraries you prefer

Method 1: Using sklearn library

Method 2: Using Statsmodels

Get familiar with **numpy**

Simple Linear Regression - code

PDU
PROFESSIONAL DEVELOPMENT UNIT

```
import numpy as np
from sklearn.linear_model import LinearRegression

# Sample data
hours_studied = np.array([1, 2, 3, 4, 5])
exam_scores = np.array([50, 60, 70, 80, 90])

# Reshape the data
hours_studied = hours_studied.reshape(-1, 1)

# Create a Linear Regression model
model = LinearRegression()

# Fit the model to the data
model.fit(hours_studied, exam_scores)

# Predict the exam score for a student who studies 6 hours
predicted_score = model.predict([[6]])
print("Predicted exam score:", predicted_score[0])
```

What is Mean Absolute Error (MAE), Mean Squared Error(MSE) and R Squared?

Mean Absolute Error(MAE): measures the average absolute error between actual and predicted values.

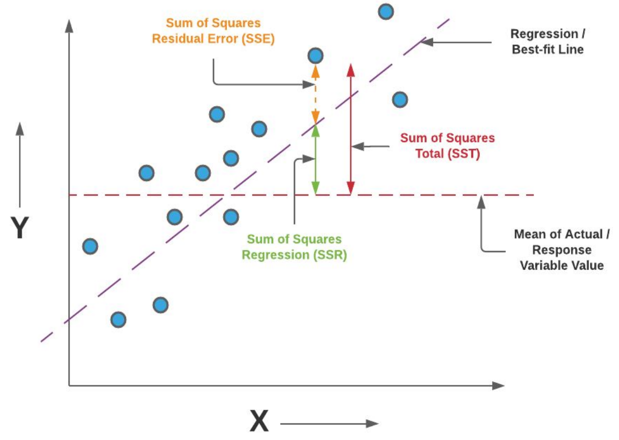
MAE = Actual values – Predicted values

The Mean Squared Error(MSE): measures the average squared error between actual and predicted values.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n}(\text{actual values} - \text{predicted values})^2$$

R- Squared score represents the goodness of fit, with values closer to 1 indicating a better fit. Determine the X and Y are correlated

$$R^2 = \frac{SSR}{SST}$$



What is Mean Absolute Error (MAE), Mean Squared Error(MSE) and R Squared?

PROFESSIONAL DEVELOPMENT UNIT

Mean Absolute Error(MAE) is the mean size of the mistakes in collected predictions. We know that an error basically is the absolute difference between the actual or true values and the values that are predicted. The absolute difference means that if the result has a negative sign, it is ignored.

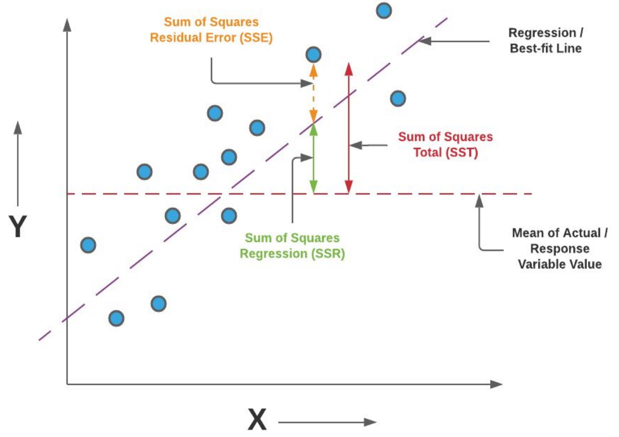
MAE = Actual values – Predicted values

The Mean Squared Error is the squared mean of the difference between the actual values and predictable values.

$$\textbf{MSE} = \frac{1}{N}\sum_{i=1}^{n}(\text{actual values} - \text{predicted values})^2$$

R- Squared is the proportion of the total variation in the dependent variable that can be explained by the independent variables in the model.

$$R^2 = \frac{SSR}{SST}$$



Simple Linear Regression - code

PDU PROFESSIONAL DEVELOPMENT UNIT

```
# Calculate the metrics
mae = mean_absolute_error(exam_scores, predicted_scores)
mse = mean_squared_error(exam_scores, predicted_scores)
r2 = r2_score(exam_scores, predicted_scores)

print("Mean Absolute Error:", mae)
print("Mean Squared Error:", mse)
print("R-squared Score:", r2)

# Plot the data and the regression line
plt.scatter(hours_studied, exam_scores, label="Actual Scores")
plt.plot(hours_studied, predicted_scores, color='red', label="Regression
Line")
plt.xlabel("Hours Studied")
plt.ylabel("Exam Scores")
plt.legend()
plt.show()
```

Multiple Linear Regression

Multiple Linear Regression is an extension of Linear Regression that allows you to predict a continuous target variable based on multiple input features. Instead of a line, it fits a hyperplane to the data.

Ex:
Suppose you want to predict a house's price based on its size (in square feet), the number of bedrooms, and the number of bathrooms.

Multiple Linear Regression

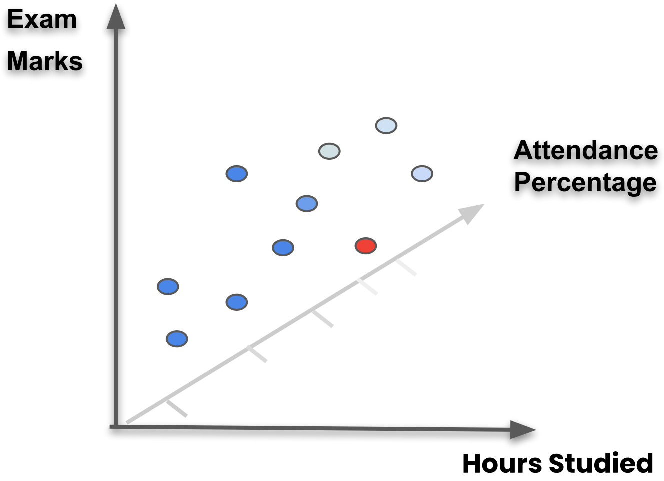
PROFESSIONAL DEVELOPMENT UNIT

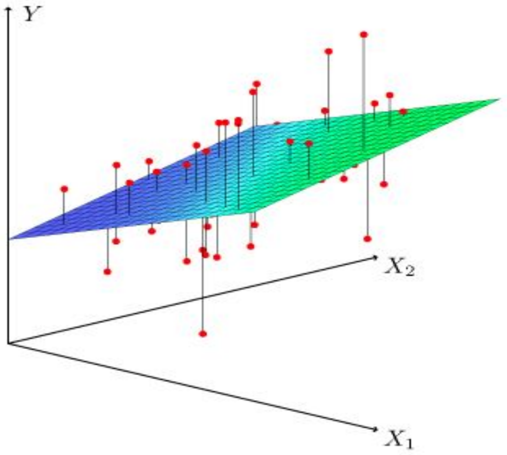
Simple

$$y = b \cdot x + a$$

Multiple

$$y = b1 \cdot x1 + b2 \cdot x2 + \ldots\ldots bk \cdot xk + a$$





Multiple Linear Regression

PDU PROFESSIONAL DEVELOPMENT UNIT

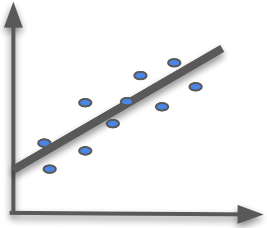
	Study Hours	Prep Exams	Final Exam Score
Student 1	3	2	76
Student 2	7	6	88
Student 3	16	5	96
Student 4	14	2	90
Student 5	12	7	98
Student 6	7	4	80
Student 7	4	4	86
Student 8	19	2	89
Student 9	4	8	68
Student 10	8	4	75
Student 11	8	1	72
Student 12	3	3	76

Multiple Linear Regression

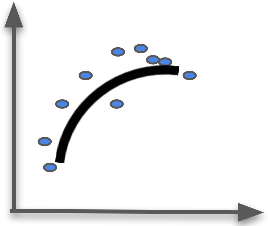
```
import numpy as np
from sklearn.linear_model import LinearRegression

# Sample data
sizes = np.array([1200, 1500, 1800, 2000, 2200, 2400, 2600, 2800, 3000,
3500])
bedrooms = np.array([2, 3, 3, 4, 4, 3, 4, 3, 4, 5])
bathrooms = np.array([1, 2, 2, 2, 3, 2, 3, 2, 3, 4])
prices = np.array([140000, 210000, 220000, 270000, 330000, 250000, 310000,
280000, 350000, 400000])


# Create a matrix of input features
X = np.column_stack((sizes, bedrooms, bathrooms))

# Create a Multiple Linear Regression model
model = LinearRegression()

# Fit the model to the data
model.fit(X, prices)

# Predict the price of a house with 2500 sq. ft., 3 bedrooms, and 2
bathrooms
predicted_price = model.predict([[2500, 3, 2]])
print("Predicted house price:", predicted_price[0])
```

Assumptions of Linear Regression

- **Linearity** - In linear regression, a straight line is drawn through the data. The straight line should represented all data points as good as possible, If it is can not represented its called non linear.

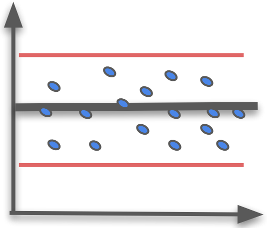


Linear

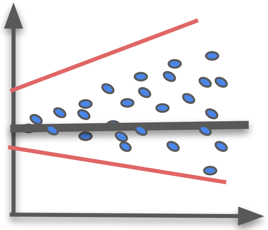


Non linear

- **Homoscedasticity** - The residuals must have a constant variance.



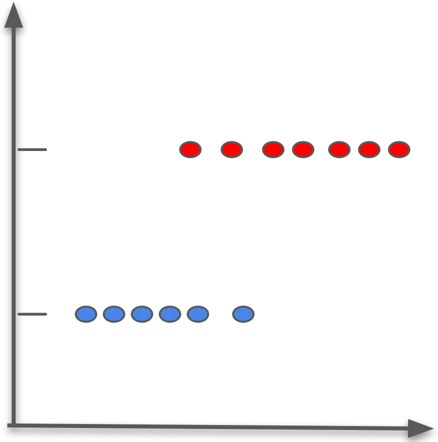
Homoscedasticity



Heteroscedasticity

Logistic Regression

Logistic regression predicts whether something is
Yes/No, **Win/Loss**, **Negative/Positive**, **True/False** and so on.

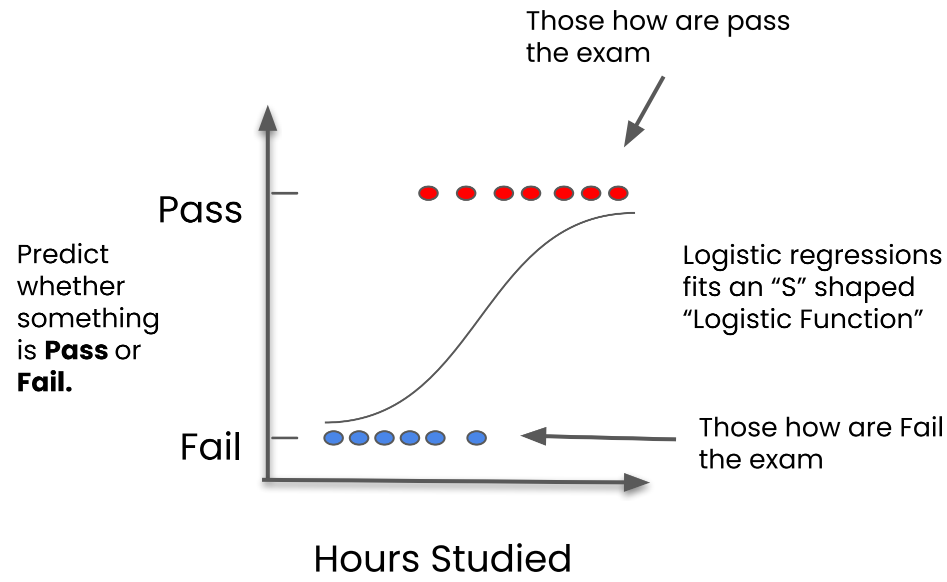


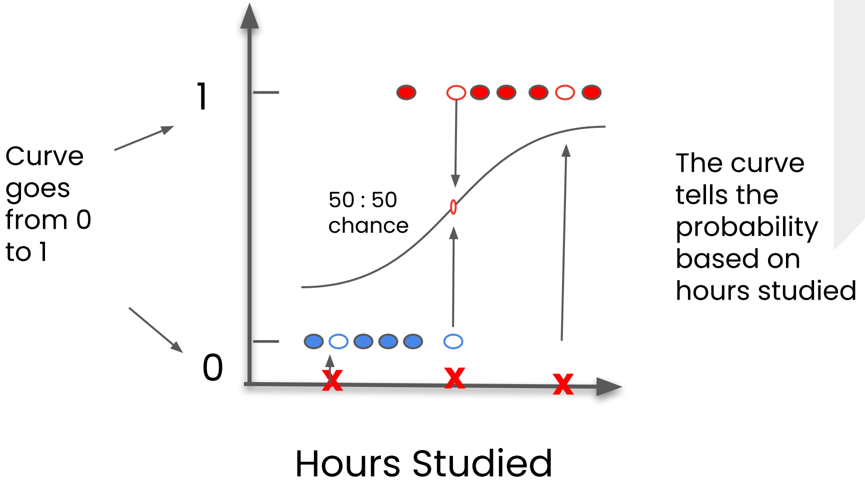
- Risk of a patient developing diabetes or not,
  (*based on **age**, **weight** and other factors..*)

- Likelihood of a borrower defaulting on a loan,
  (*Based on their **income**, **monthly payments**, and other financial factors…*)

- Detect fraudulent transactions,
  (*Based on **amount of transactions**, the t**ime of transactions** and other factors…*)

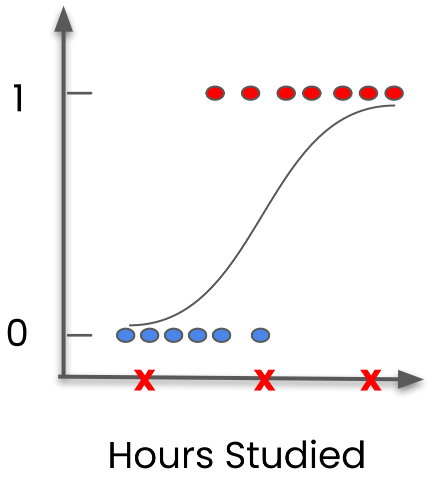
Logistic Regression

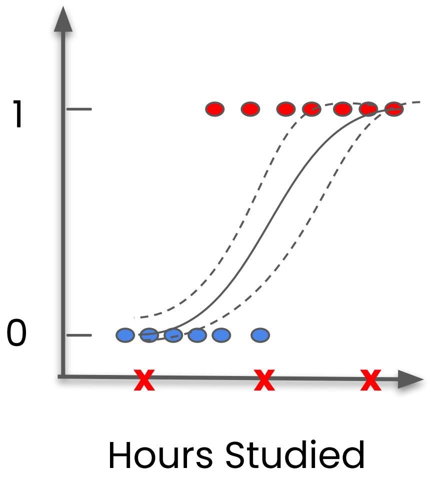




How the line is fit to the data?

Logistic regression doesn't have the same concept "***Residual***".
Instead it uses "**maximum likelihood**".

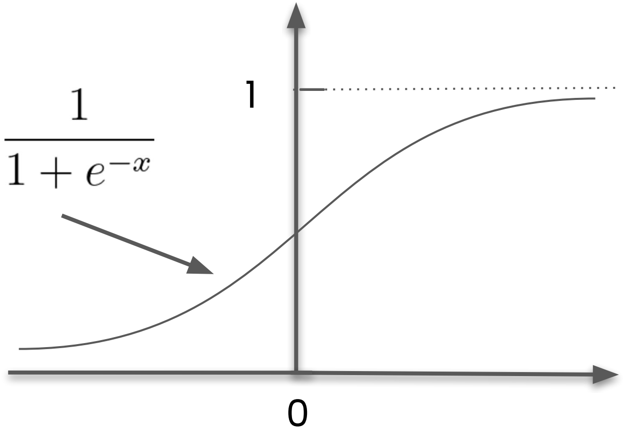




How the line is fit to the data?

The logistic model is based on the logistic function.
The important thing about the logistic function is only values between 0 and 1.

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



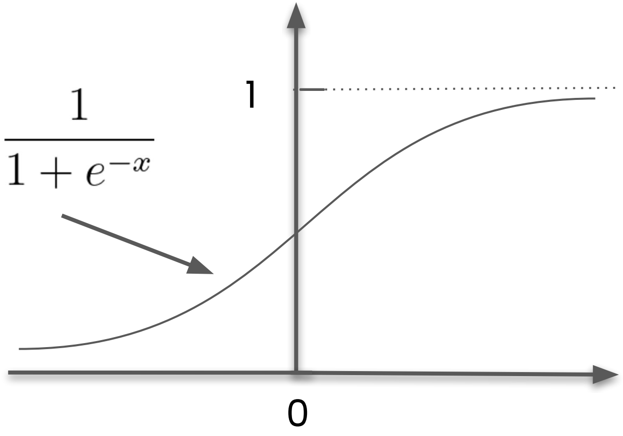
$$\hat{y} = \boxed{b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a}$$

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\boxed{(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}}$$

How to Train a Logistic Regression Model

The logistic model is based on the logistic function.
The important thing about the logistic function is only values between 0 and 1.

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



$$\hat{y} = \boxed{b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a}$$

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\boxed{(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}}$$

Activity

https://mlu-explain.github.io/logistic-regression/

Q&A

Thank You

Nisal Mihiranga
Head of AI & DS at **Zone24x7**
Consultant, Trainer
M: +94 71 726 3044