3.6 Predictive Analytics for Fraud Detection

3.6.1 Intro

The aim is to build an analytical model predicting a target measure of interest

Two types of predictive analytics can be distinguished depending on the measurement level of the target:

- *Regression*.
- *Classification*.

3.6.2 Regression #Regression

Target variable:

- *Continuous*.
- *Varies* along a predefined interval.
    - Limited (e.g., between 0 and 1).
    - Unlimited (e.g., between 0 and infinity).

3.6.3 Classification #Classification

Target variable:

- *Categorical*.
- It can only take on a *limited set of predefined values*:
    - **Binary classification**: only two classes are considered (e.g., fraud versus no-fraud).
    - **Multiclass classification**: the target can belong to more than two classes (e.g., severe fraud, medium fraud, no fraud).

3.6.4 Target Variable Definition #TargetVariable

The target fraud indicator is usually hard to (obtain) and determine.

> - *One can never be fully sure that a certain transaction is fraudulent.*
> - *The target labels are typically not noise-free.*

🕵️ => Complex analytical modeling exercise.

3.6.5 Linear Regression #LinearRegression

Technique to model a continuous target variable.

The *general formulation of the linear regression model*:
$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N$

- *Y* represents the *target variable*.
- $X_1, \ldots, X_N$ the *explanatory variables*.
- β *parameters measure the impact on the target variable Y of each of the individual explanatory variables*.

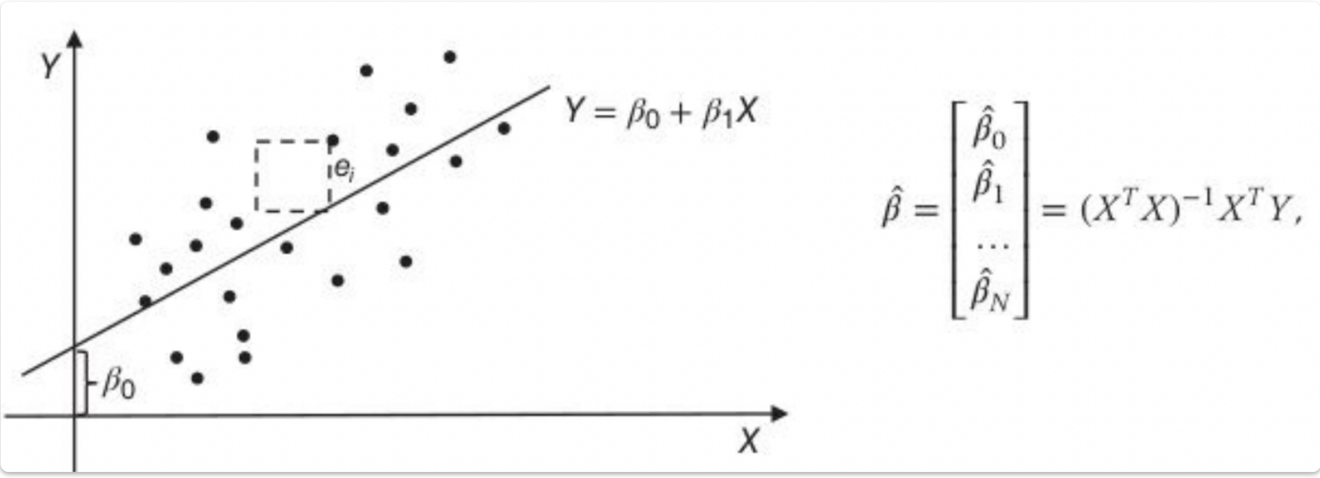
3.6.5.1 Parameter Estimation

The β parameters can then be estimated by *minimizing a squared error function*:
$\frac{1}{2}\sum_{i=1}^{n} e_i^2 = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2 = \frac{1}{2}\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_{1i} + \ldots + \beta_N X_{Ni}))^2$

Observation	X_1	X_2	…	X_N	Y
1	X_{11}	X_{21}	…	X_{N1}	Y_1
2	X_{12}	X_{22}	…	X_{N2}	Y_2
…					
n	X_{1n}	X_{2n}	…	X_{Nn}	Y_n

3.6.5.2 Ordinary least squares (OLS) regression #OLS

Minimizing the sum of all error squares.



Goal of Linear regression: find the *best fit line that can accurately predict the output* for the continuous dependent variable with the help of independent variables.

Example:

Company	Revenue	Employees	VATCompliant	…	Fraud	Y
ABC	3,000k	400	Y		No	0
BCD	200k	800	N		No	0
CDE	4,2000k	2,200	N		Yes	1
…						
XYZ	34k	50	N		Yes	1

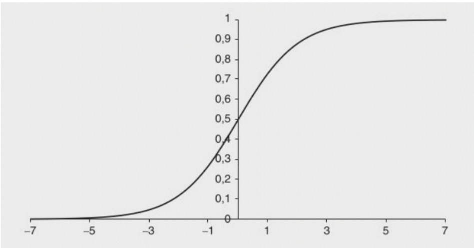
Linear regression: $Y = \beta_0 + \beta_1 Revenue + \beta_2 Employees + \beta_3 VATCompliant$

When estimating this using OLS, **two key problems arise**:

1. The *errors/target are not normally distributed* but follow a Bernoulli distribution with only two values.
2. There is *no guarantee that the target is between 0 and 1*, which would be handy since it can then be interpreted as a probability.

3.6.6 Logistic Regression

Consider now the following **bounding function**:



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

The *outcome is always between 0 and 1.*

#DEF **Logistic Regression Model**: **Combination of the linear regression with the bounding function**.

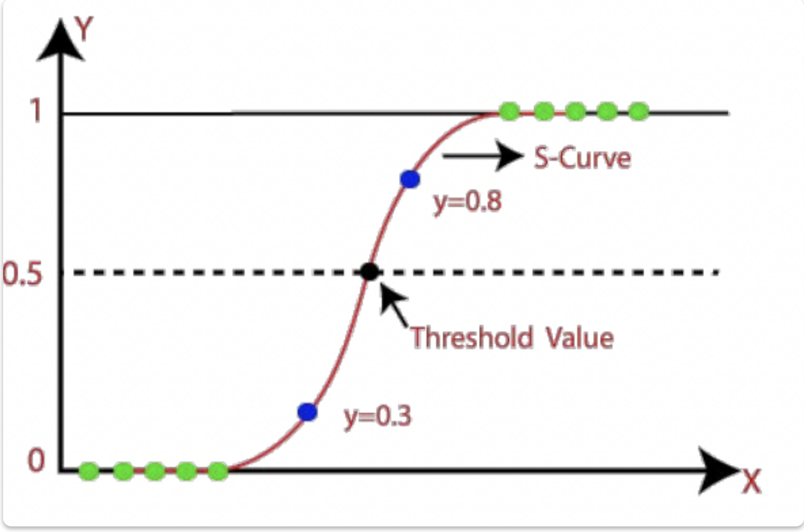
Given $Z = \beta_0 + \beta_1 Revenue + \beta_2 Employees + \beta_3 VATCompliant$
$f(Z) = \frac{1}{1+e^{-(\beta_0 + \beta_1 Revenue + \beta_2 Employees + \beta_3 VATCompliant)}}$

Outcome: *bounded between 0 and 1* == **probability**.

Then we have: $P(fraud = yes \mid Revenue, Employees, VATCompliant)$

3.6.6.1 ACTIVATION FUNCTION

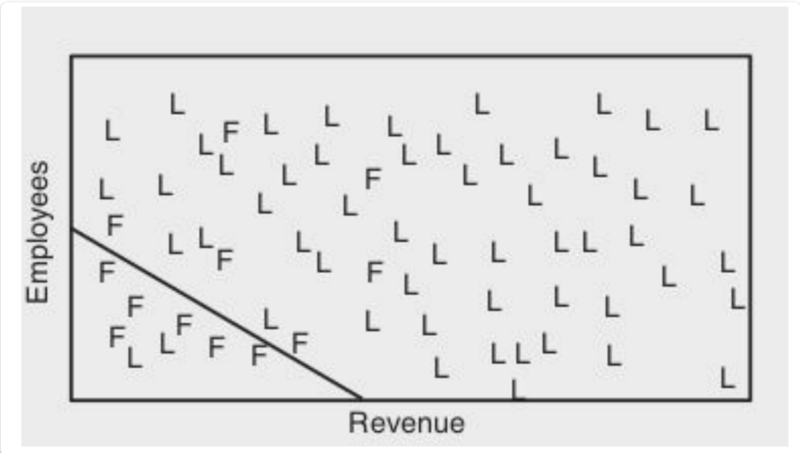
We pass the *weighted sum of inputs* through an **activation function** *that can map values in between 0 and 1*.
Such activation function is known as **sigmoid function** and the curve obtained is called as *sigmoid curve or S-curve*.



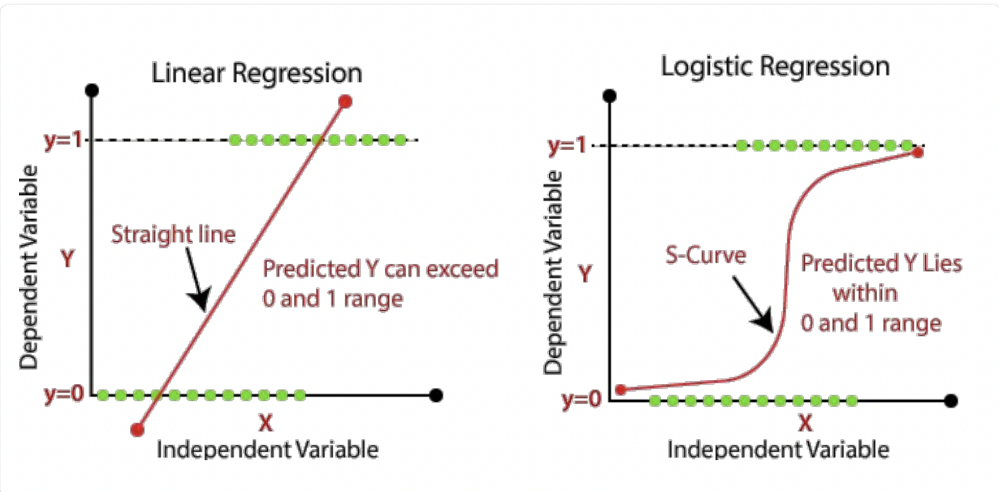
The β_i parameters of a *logistic regression model* are estimated using the **maximum likelihood optimization**.

3.6.6.2 LOGISTIC REGRESSION PROPERTY

It estimates a linear decision boundary to separate both classes.



3.6.7 Linear and Logistic Regression



Linear Regression	***Logistic Regression***

Linear Regression	***Logistic Regression***
Predicting the **continuous** dependent variable with independent variables.	Predict the **categorical** dependent variable with independent variables.
Predict the output for the continuous dependent variable.	It estimates a **linear decision boundary** to separate both classes.
Finds the **linear relationship** between dependent variable and independent variable.	Based on the concept of **Maximum Likelihood estimation**.
Based on **Ordinary Least Squares**	Used for: Classification/Regression/where the probabilities is required.
Output: continuous values	**Output**: between the 0 and 1.

Next chapter: Decision Trees