Introduction to Logistic Regression for Classification

Imad EL BADISY

CM6RI-UM6SS (Rabat) | SESSTIM-AMU (Marseille)

Dernière modification 18 octobre 2023

Introduction

- Logistic regression is a statistical method used for **binary classification**.
- It models the probability of a binary outcome based on **one or more predictor variables**.
- It's widely used in various fields like healthcare, finance, and marketing for predicting outcomes like whether a customer will buy a product or not.
- For example in health, examples of binary outcomes include the **presence or absence** of certain behaviors or conditions, such smoking, having diabetes or being in depression.

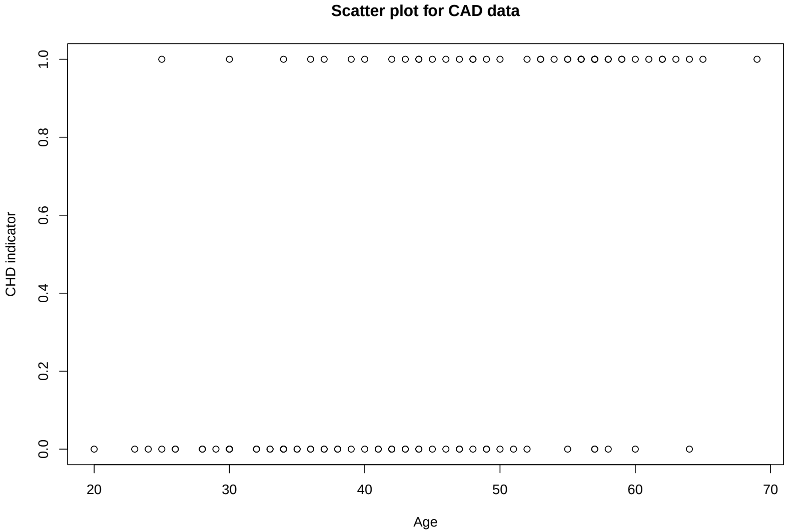
Binary outcome

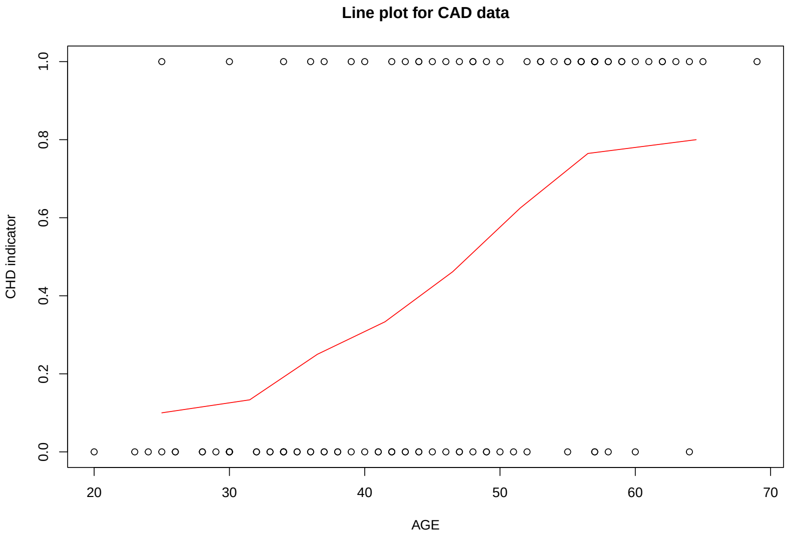
- Binary outcomes can reflect the **occurrence or nonoccurrence** of a specific event (cancer progression after primary treatment, cancer recurrence, spam, ...)
- Unlike continuous outcomes, categorical outcomes do not have a default numerical scale. Consequently, standard statistical summaries such as the mean, median, quantile or variance are not meanful.
- The prediction of a categorical variable can be likened to a classification task : one predicts the **probability of belonging to a given category**

Why not a linear regression ?

- The linear regression model assumes that the response variable is quantitative. However, in many situations the **variable is rather binary/categorical**.
- For discrete and bounded response-variables, the values predicted by the simple linear regression model may **exceed the limits of the interval [0, 1]**.
- In addition to the **violation of the validation assumptions of the linear regression model**

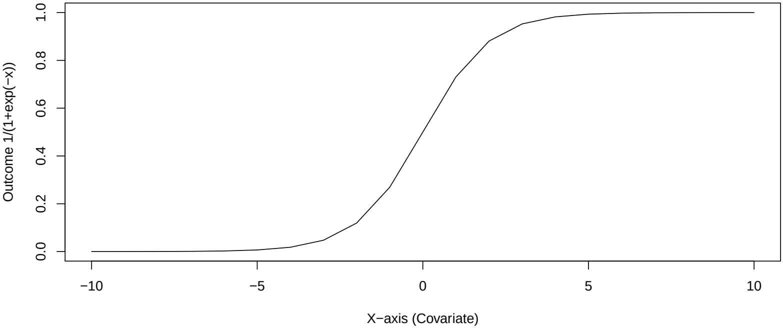
Example: Understanding the relationship between coronary artery disease and patient age




Line plot for CAD data
CHD indicator
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
AGE

Logistic function

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



Logistic regression model

- The estimator of the logistic regression model is the **maximum likelihood** (beyond the scope of this course).
- The probability of distribution (probability law) of a two-mode variable is the **binomial distribution**.
- It is known that the **probability of success is affected by multiple factors**. Hence the interest of a model that includes **explanatory variables** (i.e. features) in order to refine predictions.

Theoretical formulation of the logistic regression model

Let the binary variable $Y_i, i = 1, 2, ..., n$ and $X = (x_1, ..., x_k)$ be a vector of associated covariates of k elements, the probability of success is specified by $P(Y = 1) = \pi(x)$, the **logistic model** is written as fallows :

$$P(Y = 1|x_1, x_2, x_3, ..., x_k) = \pi(x) = \frac{e^{\ \beta\ X}}{1 + e^{\beta\ X}} = \frac{e^{\sum_{n=1}^{k}\ \beta_i x_i}}{1 + e^{\sum_{n=1}^{k}\ \beta_i x_i}}$$

with $\beta = (\beta_0, \beta_1, ..., \beta_p)$ a vector of regression coefficients.
From the previous expression, we can express the **probability** of failure by :

$$P(Y = 0|x_1, x_2, x_3, ..., x_k) = 1 - \pi(x) = 1 - \frac{e^{\ \beta\ X}}{1 + e^{\beta\ X}} = \frac{1}{1 + e^{\beta\ x}}$$

This nonlinear function is a sigmoidal function of the model terms and constrains the probability estimates to between 0 and 1.

Odds-ratios An important formula can be deduced from the expressions of the two probabilities, the **odds ratio** which is the **ratio of the probability of success and the probability of failure** :

$$OR = \frac{\pi(X)}{1-\pi(X)} = \frac{e^{\beta\ X}}{1+e^{\beta\ X}} \cdot \left[\frac{1}{1+e^{\beta\ x}}\right]^{-1} = e^{\beta X}$$

By adding the log to the equation, we find the **linear form of the logistic model**:

$$ln(OR) = ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = X\beta = \sum_{i=0}^{k} \beta_i x_i$$

- Odds ratio quantifies the relationship between a predictor variable and the outcome.

Variable importance

- Relative masure of the model predictors contribution in increasing a given performance metric.

$$\text{Variable Importance} = |\beta_i|$$

- β_i is the coefficient associated with predictor variable X_i.

Confusion Matrix

A confusion matrix provides a breakdown of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It is not a formula but a tabular representation of model performance.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Accuracy Formula

Accuracy measures the proportion of correctly classified instances and is calculated as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}}$$

AUC-ROC Formula: The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) quantifies a model's ability to discriminate between classes. The ROC curve is formed by plotting the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various thresholds.

- AUC measures the area under this curve. A perfect model has an AUC of 1, while a random model has an AUC of 0.5.

Cross-Validation

repeated K-fold cross-validation to repeatedly partition the full dataset into K folds. For a given partitioning, prediction is performed on each of the K-folds with models fit on all remaining folds (repeated split train/test on all the dataset).

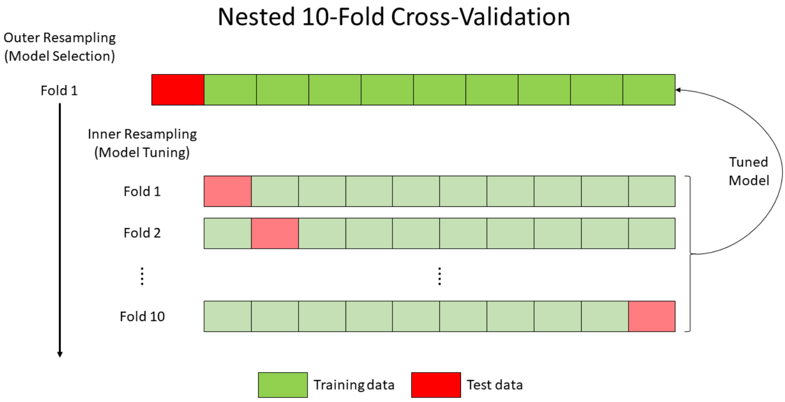


Figure 1: Cross-Validation

Recap

- Logistic regression is a powerful tool for binary classification.
- Train/test split and cross-validation are crucial for model evaluation.
- Variable importance, odds ratio, and p-values help interpret the model.
- Comparing to other ML models provides insights into model performance.
- Evaluation metrics like accuracy, AUC-ROC, and confusion matrix assess model quality.

Applications

Application : Predicting heart attacks for patients with arthritis