Introduction à la régression logistique

La régression logistique est une technique d'apprentissage supervisée de base pour résoudre les problèmes de classification.

Le nom de cet algorithme peut être un peu déroutant dans le sens où l'algorithme d'apprentissage automatique de régression logistique est destiné aux tâches de classification et non aux problèmes de régression. Le nom «Régression» implique ici qu'un modèle linéaire est ajusté dans l'espace des fonctionnalités. Cet algorithme applique une fonction logistique à une combinaison linéaire de caractéristiques pour prédire le résultat d'une variable dépendante catégorielle basée sur des variables prédictives. Les algorithmes de régression logistique aident à estimer la probabilité de tomber à un niveau spécifique de la variable dépendante catégorielle en fonction des variables prédictives données.

Supposons que vous vouliez prédire s'il y aura de la pluie demain à Toronto. Ici, le résultat de la prédiction n'est pas un nombre continu car il y aura soit de la pluie, soit pas de pluie et donc la régression linéaire ne peut pas être appliquée. Ici, la variable de résultat est l'une des nombreuses catégories et l'utilisation de la régression logistique aide.

Applications de la régression logistique

- L'algorithme de régression logistique est appliqué dans le domaine de l'épidémiologie pour identifier les facteurs de risque de maladies et planifier en conséquence des mesures préventives.
- Utilisé pour prédire si un candidat gagnera ou perdra une élection politique ou pour prédire si un électeur votera pour un candidat particulier.
- Utilisé dans les prévisions météorologiques pour prédire la probabilité de pluie.
- Utilisé dans les systèmes de notation de crédit pour la gestion des risques afin de prédire la défaillance d'un compte.

Ressources

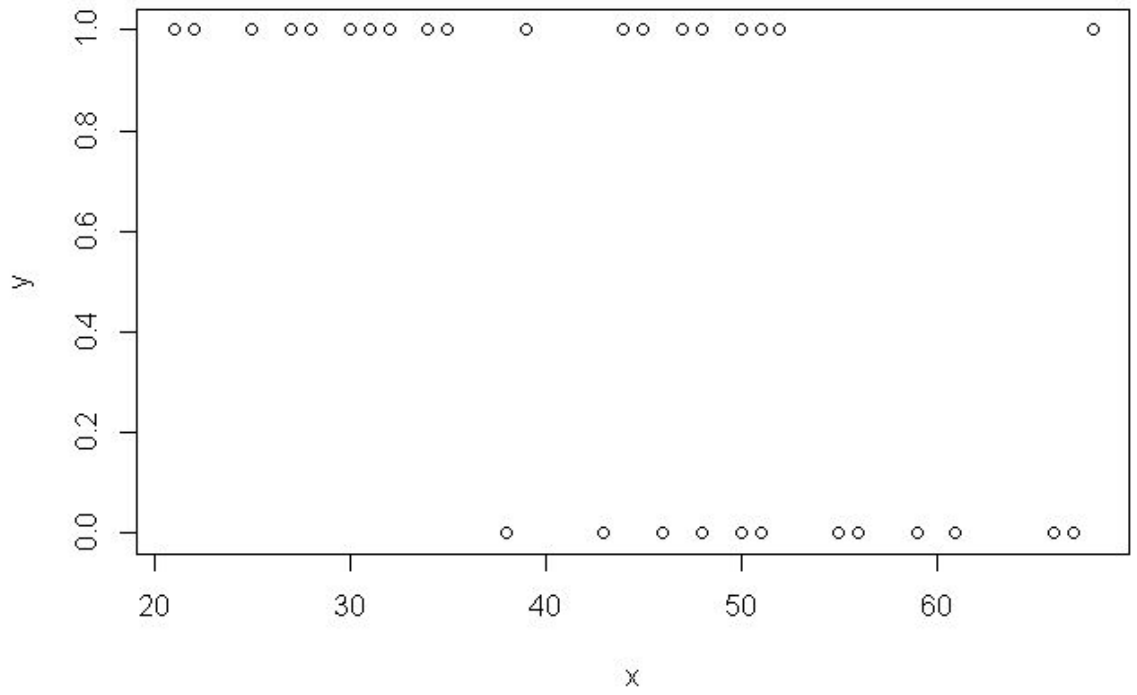
https://www.em-consulte.com/showarticlefile/143634/pdf_54287.pdf
https://fr.wikipedia.org/wiki/Loi_logistique
https://fr.wikipedia.org/wiki/R%C3%A9gression_logistique

1. **Présentation générale d'une régression logistique**

Pour analyser une variable binaire (dont les valeurs seraient VRAI/FAUX, 0/1, ou encore OUI/NON) en fonction d'une variable explicative quantitative, on peut utiliser une régression logistique.

Considérons par exemple les données suivantes (data_regression_logistique.csv), où x est l'âge de 40 personnes, et y la variable indiquant s'ils ont acheté un album de death metal au cours des 5 dernières années (1 si "oui", 0 si "non")

Graphiquement, on constate que vraisemblablement, plus les personnes sont âgées, moins elles achètent de death metal.



Vérifions cela à l'aide d'un modèle...
La régression logistique est un cas particulier de Modèle Linéaire Généralisé (GLM).
Avec un modèle de régression linéaire classique, on considère le modèle suivant:

$$Y = \alpha X + \beta$$

On prédit donc l'espérance de Y de la manière suivante :
$$E(Y) = \alpha X + \beta$$

Ici, du fait de la distribution binaire de Y, les relations ci-dessus ne peuvent pas s'appliquer.

Pour "généraliser" le modèle linéaire, on considère donc que :
$$g(E(Y)) = \alpha X + \beta$$
où g est une fonction de lien.

En l'occurrence, pour une régression logistique, la fonction de lien correspond à la fonction logit:

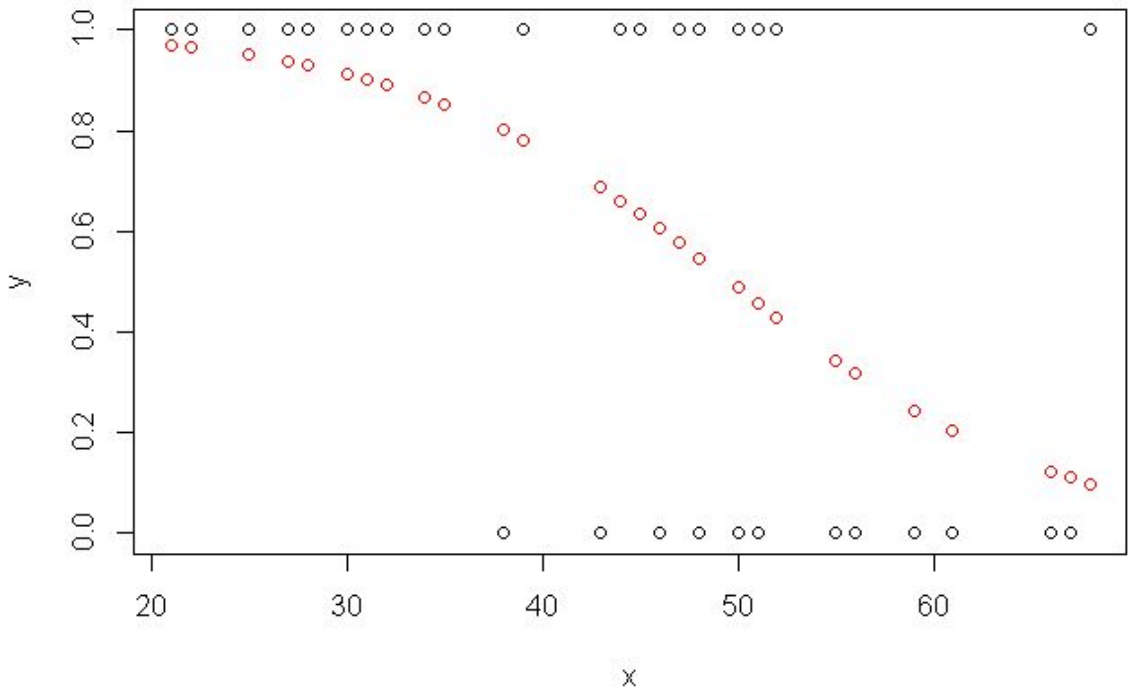
$$logit(p) = log(\frac{P}{1-P})$$

Notez que cette fonction logit transforme une valeur (p) comprise entre 0 et 1 (comme une probabilité par exemple) en une valeur comprise entre $-\infty$ et $+\infty$.

En réalisant la régression logistique avec un logiciel, on pourrait avoir :
$$logit(E(Y)) = -0.12X + 5.95$$

et l'on constate que l'influence (négative) de l'âge sur l'achat d'albums de death metal est bien significative au seuil de 5%.

Si l'on représente cette relation entre logit(E(Y) et X, on retrouve bien une relation linéaire. En revanche, l'échelle des ordonnées n'est pas aisée à interpréter... On procède donc à une transformation inverse de la relation:



2. Pour aller plus loin

La régression logistique est une approche statistique qui peut être employée pour évaluer et caractériser les relations entre une variable réponse de type binaire (par exemple : Vivant / Mort, Malade / Non malade, succés / échec), et une, ou plusieurs, variables explicatives, qui peuvent être de type catégoriel (le sexe par exemple), ou numérique continu (l'âge par exemple).

Tout comme la régression de Poisson, la régression logistique appartient aux modèles linéaires généralisés. Pour rappel, il s'agit de modèles de régression qui sont des extensions du modèle linéaire, et qui reposent sur trois éléments :

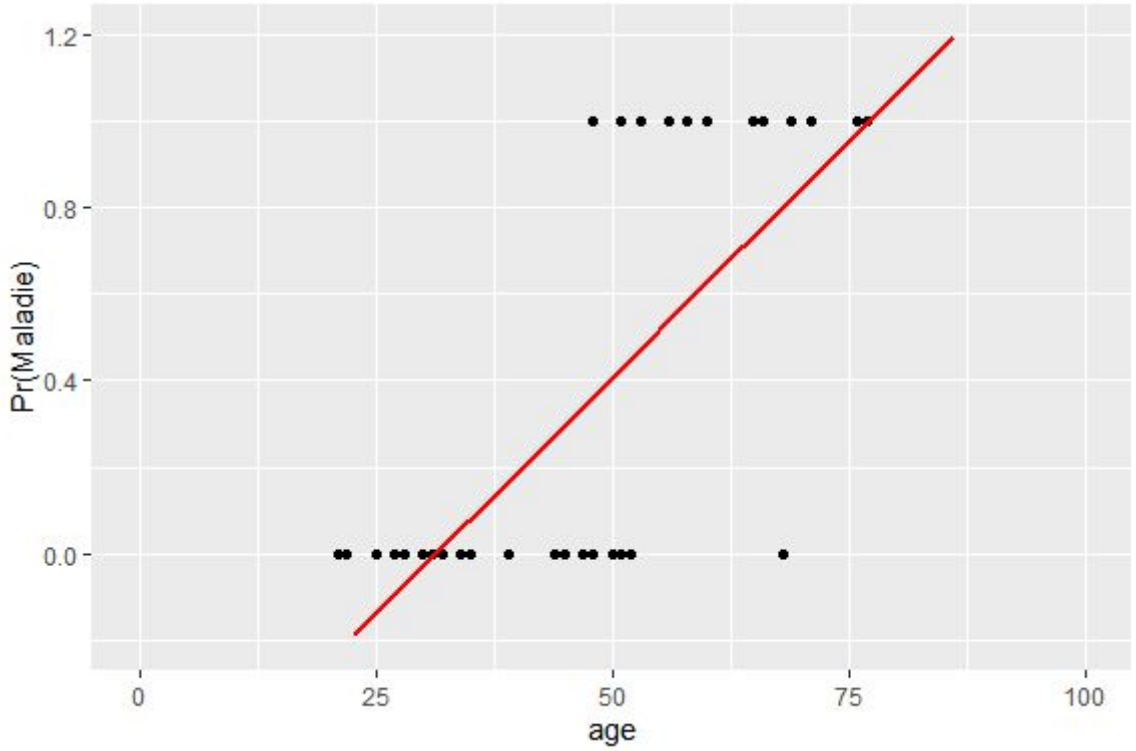
1. un prédicteur linéaire
2. une fonction de lien
3. une structure des erreurs

Les aspects théoriques et mathématiques de la régression logistique sont relativement complexes, c'est pourquoi nous ne les aborderons pas ici. Néanmoins, certains éléments caractérisant la régression logistique peuvent être retenus.
Remarque : La régression logistique peut également être utilisée comme un algorithme de classification supervisée, mais nous ne l'aborderons pas ici.

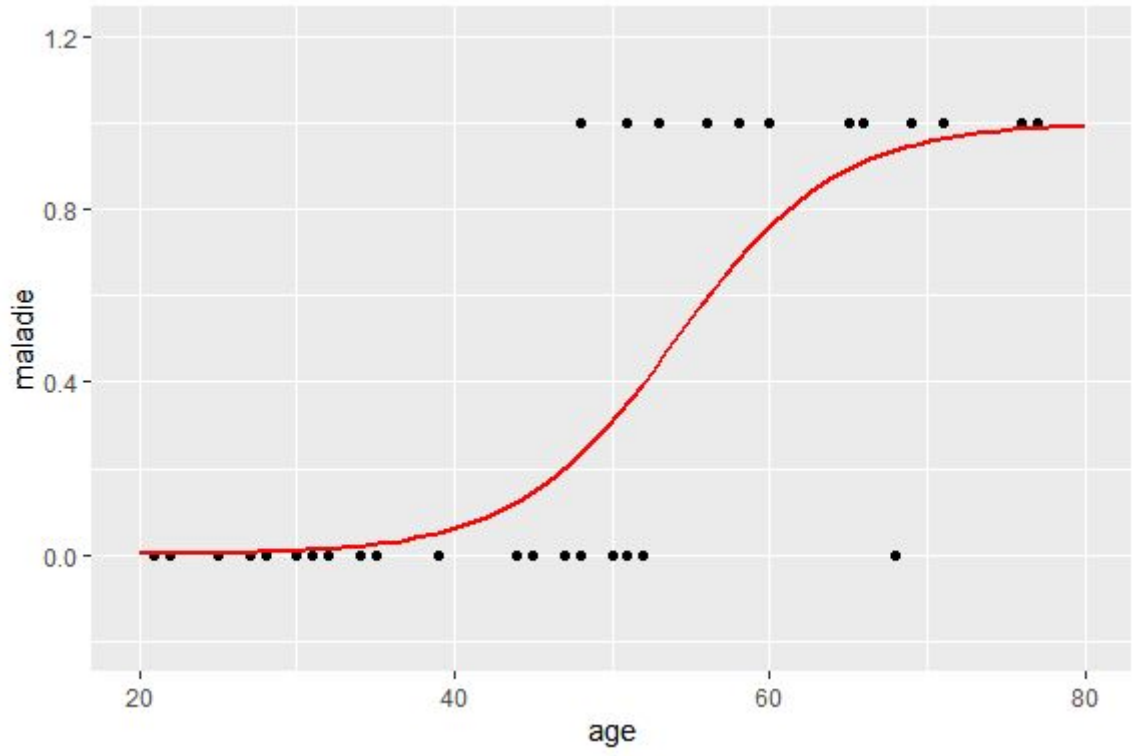
Les principaux éléments de la régression logistique

Modélisation de la probabilité

Dans la régression logistique, ce n'est pas la réponse binaire (malade/pas malade) qui est directement modélisée, mais la probabilité de réalisation d'une des deux modalités (être malade par exemple)
Cette probabilité de réalisation ne peut pas être modélisée par une droite car celle-ci conduirait à des valeurs <0 ou >1. Ce qui est impossible puisqu'une probabilité est forcément bornée par 0 et 1.



Cette probabilité, est alors modélisée par une courbe sigmoïde, bornée par 0, et 1 :



Cette courbe sigmoïde est définie par la fonction logistique (fonction sigmoïde), d'équation :

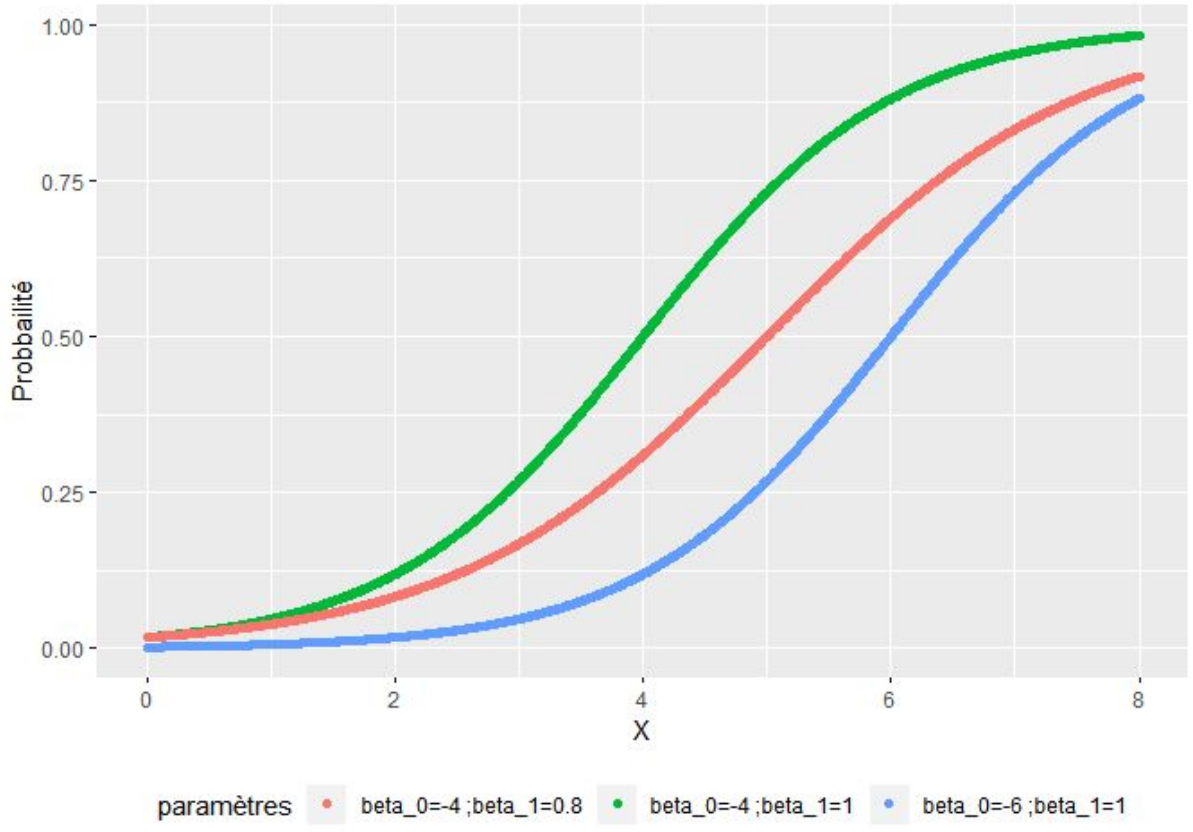
$f(x) = \frac{exp(x)}{1+exp(x)=p}$ =p

La fonction logistique

Lorsque la fonction logistique est ajustée à des données observées, la forme de la courbe sigmoïde s'adapte à ces données, par l'estimation de paramètres. Dans le cas d'une seule variable explicative (X), l'équation de la courbe logistique est alors :

$$P(x) = \frac{exp(\beta 0+\beta 1X)}{1+exp(\beta 0+\beta 1X)}$$

Voici 3 exemples de courbes logistiques, obtenues avec des paramètres Beta0 et Beta1 différents :



Dans une situation de variables explicatives multiples l'équation se généralise en :

$$P(x) = \frac{exp(\beta 0+\beta 1X1+\ldots+\beta nXn)}{1+exp(\beta 0+\beta 1X1+\ldots+\beta nXn)} = \frac{exp(\sum \beta X)}{1+exp(\sum \beta X)}$$

La fonction de lien logit

Le modèle précédent n'est pas linéaire dans l'expression des paramètres beta_X puisque la probabilité de réalisation ne s'exprime pas comme une addition des effets des différentes variables explicatives.
Pour obtenir un tel modèle (linéaire dans ses paramètres), il est nécessaire de passer par une transformation logit :

$$logit(p) = log(\frac{p}{1-P}) = \sum_{j=1}^{n} \beta_j X_{ij}$$

Cette transformation logit est la fonction de lien qui permet de mettre en relation la probabilité de réalisation (bornée entre 0 et 1), et la combinaison linéaire de variable explicatives.

La structure d'erreur

Les données de base employées dans une régression logistique dont des données binaires (oui/non). Celles-ci sont distribuées selon une loi binomiale B(1,p). Il en est alors de même pour les erreurs : elles sont distribuées selon une loi binomiale B(1,p).

Les coefficients estimés sont des log odds ratio

Le terme p/(1-p) est un rapport de cote (RC) ou Odds Ratio (OR), en anglais. Ce paramètre permet de mesurer la relation entre la variable explicative (X) et la réponse Y (vivant par exemple).
Les coefficients beta_j issus de la régression logistique sont donc des log odds ratio.
Un odds ou côte est le rapport de deux probabilités complémentaires : la probabilité P de survenue d'un événement (risque), divisé par la probabilité (1-P) que cet événement ne survienne pas (non risque, c'est-à-dire sans l'événement).
Par exemple, si on s'intéresse au risque de récidive d'une pathologie chez les hommes et les femmes, et que le risque de récidive est de 80% chez les hommes et de 40% chez les femmes, alors :

- la côte de récidives chez les hommes est 0.8/0.2 = 4 (il y a 4 fois plus de récidives que de non récidives chez les hommes)
- la côte de récidives chez les femmes est 0.4/0.6 = 0.67 (il y a 0.67 fois plus de récidives que de non récidives chez les femmes)
- l'OR correspond aux rapport de ces deux cotes :

$OR = \frac{\frac{0.8}{0.2}}{\frac{0.4}{0.6}}=6$
Ici l'odds des hommes est 6 fois plus élevé que celui des femmes. On dira, par la suite (voir plus loin) que le risque de récidive est plus important chez les hommes.

Exemple avec une variable explicative catégorielle

Il s'agit des résultats d'une régression logistique visant à étudier le lien entre la présence d'une maladie cardiaque et le sexe des patients :

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-1.05779	0.2321396	-4.556699	5.2e-06
gendermale	1.27220	0.2711647	4.691614	2.7e-06

Le coefficient (Estimate) de la ligne “gendermale” correspond au log OR. Pour obtenir l’OR, il est donc nécessaire d’employer une transformation
exponentielle :

```
OR_gender = exp(1.27)
OR_gender
## [1] 3.560853
```

Exemple avec une variable explicative numérique

Lorsque la variable explicative est de type numérique, le coefficient obtenu est également un log(OR). Sa transformation, par la fonction exponentielle, permettra d’obtenir un OR qui caractérisera la force de la relation entre la probabilité de réalisation et la variable explicative.
Ici, il s’agit des résultat d’une régression logistique visant à étudier le lien entre l’apparition d’une maladie et l’âge des patients :

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-10.496783	3.4901907	-3.007510	0.002634
age	0.194039	0.0665538	2.915519	0.003551

```
OR_age = exp(0.19)

OR_age
## [1] 1.20925
```

Interprétation de l'OR

Règles générales

- Si l'OR est significativement < 1 alors la variable explicative est un facteur protecteur.
- Si l'OR n’est pas significativement différent de 1, alors il n’y a pas de lien entre la réalisation (par exemple la maladie) et la variable explicative.
- Si l’OR est significativement > 1 alors la variable explicative est un facteur de risque

Lorsque la variable explicative est catégorielle

Dans cette situation, il existe deux cas de figure :

- La fréquence de la réalisation (maladie, récidive, etc..) est rare (<10%). Dans cette situation, on interprète l'OR comme un risque relatif (RR). Par exemple si on étudie la relation entre la récidive d'une maladie et le sexe (Feminin en référence), et que l'odds ratio = 4 alors on pourra dire,"être un homme multiplie le risque de récidive par 4″. Et on interprétera ensuite la significativité de cet odds ratio avec la p-value correspondante.
- La fréquence de réalisation n'est pas rare. Dans cette situation l'OR ne peut pas être interprété comme un risque relatif. De ce fait, on n'interprète pas la quantité de l'OR. On se contentera de dire, si la pvalue du log OR est <0.05, "être un homme est associé à un risque plus élevé de récidive".

Lorsque la variable explicative est numérique continue

Dans cette situation, on n'interprète pas non plus la valeur de l'OR. Dans l'exemple précédent l'OR relatif à l'âge = 1.23. On ne peut pas dire "une augmentation d'un an d'âge augmente le risque de maladie d'un facteur 1.23". Dans cette situation on se contente de regarder le signe de l'OR, et s'il est significativement différent de 1 (pvalue du log OR <0.05), on pourra dire "il existe une association significative entre l'âge et le risque de maladie, au risque de 5%, le risque de maladie augmente lorsque l'âge augmente".

A noter qu'il existe trois principaux types de régression logistique:

- **binomial:** la variable cible ne peut avoir que 2 types possibles: «0» ou «1» qui peut représenter «gagnant» vs «perte», «réussite» vs «échec», «mort» vs «vivant», etc.
- **multinomial:** la variable cible peut avoir 3 types possibles ou plus qui ne sont pas ordonnés (c'est-à-dire que les types n'ont pas de signification quantitative) comme «maladie A» vs «maladie B» vs «maladie C».
- **ordinal:** il traite des variables cibles avec des catégories ordonnées. Par exemple, un résultat de test peut être catégorisé comme suit: «très mauvais», «mauvais», «bon», «très bon». Ici, chaque catégorie peut recevoir un score tel que 0, 1, 2, 3.

La régression logistique est une technique d'apprentissage automatique supervisée largement utilisée. C'est l'un des meilleurs outils pour les statisticiens, les chercheurs et les data scientists en analyse prédictive. Elle offre plusieurs avantages comme il s'agit d'un algorithme robuste car les variables indépendantes n'ont pas besoin d'avoir une variance égale ou une distribution normale, ne supposent pas une relation linéaire entre les variables dépendantes et indépendantes et peuvent donc également gérer les effets non linéaires et elles sont également plus faciles à inspecter et moins complexe.