# Introduction à la régression logistique


Cédric Stéphane BATIONO

cedric-stephane.bationo@univ-amu.fr

Méthodologiste Biostatisticien, Msc MPH PhDc, Aix Marseille Univ

Juste Aristide GOUNGOUNGA

juste.goungounga@univ-amu.fr

Méthodologiste Biostatisticien, MD PhD, Aix Marseille Univ, Univ de Bourgogne


université virtuelle
Burkina Faso

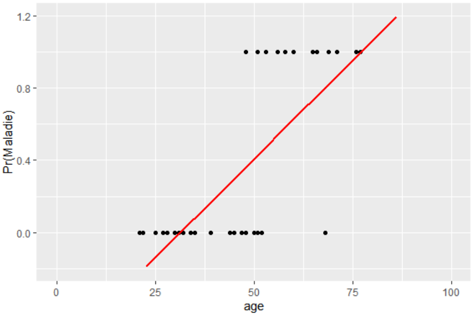
# Sommaire

- Approche statistique pour évaluer et caractériser les relations entre une variable réponse de type binaire ( par ex : Vivant / Mort, Malade / Non malade, succés / échec), et une, ou plusieurs, variables explicatives, qui peuvent être de type catégoriel (le sexe par ex), ou numérique continu (l'âge par ex)
- 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 :

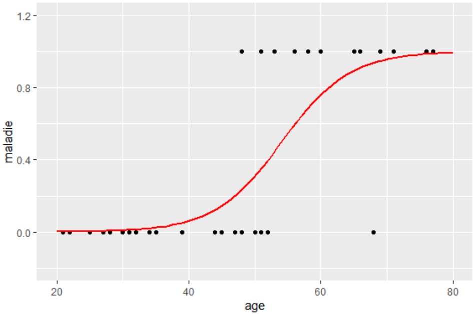
1. un prédicteur linéaire
2. une fonction de lien
3. une structure des erreurs

# Modélisation de la probabilité (1)

- 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 ex)
- Cette probabilité de réalisation ne peut pas être modélisée par une droite car conduirait à des valeurs $< 0$ ou $> 1$. Ce qui est impossible puisqu'une probabilité est forcément bornée par 0 et 1.



- Probabilité, est alors modélisée par une courbe sigmoide, bornée par 0, et 1 :



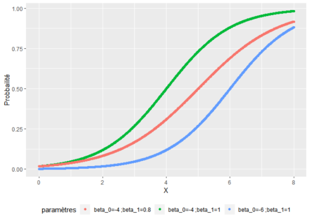

- Cette courbe sigmoide est définie par la fonction logistique, d'équation : $f(x) = \frac{\exp(x)}{1+\exp(x)} = p$

# La fonction logistique

Lorsque la fonction logistique est ajustée à des données observées, la forme de la courbe sigmoide 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:

$$P(X) = \frac{\exp\left(\beta_0 + \beta_1 X\right)}{1 + \exp\left(\beta_0 + \beta_1 X\right)}$$

- Ex avec $\beta_0$ et $\beta_1$ différents



# La fonction logistique (cont.)

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

$$P(X) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n\right)} = \frac{\exp\left(\sum \beta X\right)}{1 + \exp\left(\sum \beta_X\right)}$$

# La fonction de lien logit

- Le modèle précédent n'est pas linéaire dans l'expression des paramètres 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 :

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

- 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 ex).
- Les coefficients $\beta_j$ issus de la régression logistique sont donc des log odds ratio.
- Un odds 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).

# Les coefficients estimés sont des log odds ratio ( 2)

Ex, 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:

1. La cote 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)
2. La cote 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)
3. 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
```

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
```

- 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 est rare ($< 10\%$):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 OR$=4$ alors on pourra dire,"être un homme multiplie le risque de récidive par 4". Et on interpretera 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'interpretera 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".

- 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 (p-value 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".