# Algèbre Linéaire

Semaine 3

February 26, 2019

## La régression logistique cherche à:

- modéliser la probabilité qu'un évènement Y se produise en fonction de variables indépendantes X
- estimer la probabilité qu'un évènement se produise pour une observation aléatoire
- prédire l'effet d'une série de variables sur une variable binomiale (pouvant prendre deux valeurs, 0 et 1, *success/fail*, *malade/en bonne sante*)
- classifier les observations en estimant la probabilité qu'une observation soit dans une catégorie particulière

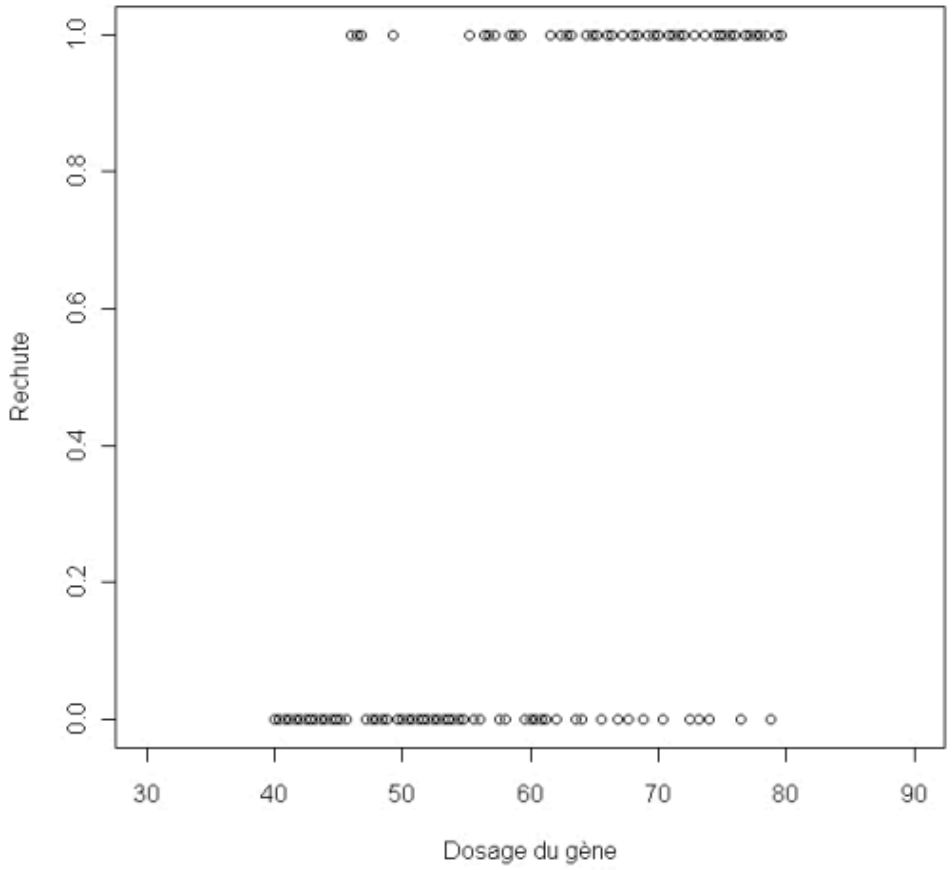
## Différence avec les autres méthodes de régression:

- la régression linéaire simple fonctionne avec une variable quantitative qui prédit une autre variable quantitative (1 & 0 ne sont pas quantitatifs)
- la multiple régression s'agit de régression linéaire simple mais avec un plus grand nombre de variables indépendantes.

Donc, effectuer une régression linéaire typique sur les données à l'œuvre provoque des problèmes majeurs:

- Les données binaires (0 ou 1) n'ont pas de distribution normale
- Les valeurs prévues des variables dépendantes peuvent aller au delà de 0 ou 1 et violent donc les propriétés des probabilités
- La représentation graphique de variables probabilistes est souvent non linéaire (comme les courbes en U) où la probabilité est très élevé selon qu'on se situe à l'extrême haut ou bas des valeurs x.

On parle ici de probabilités donc de valeurs contenues entre 0 et 1. Si elles sont représentées telles quelles sur un graphique, le résultat n'aurait aucun sens (voir ci-dessous).




Pour comprendre la régression logistique, il faut bien comprendre les probabilités. Ainsi, la probabilité qu'une chose se produise **p** correspond à:

**p =** ce que l'on veut qui se produise / tout ce qui peut se produire
ex: je veux tirer une carte du signe carreau. ¼ des cartes sont des carreaux. J'ai donc 1 chance sur 4 de tirer un carreau.

Partant de notre probabilité, nous pouvons déterminer la cote (*odds*) de tirer un carreau.
**Odds = p / 1 – p,** soit **0.25/0.75** noté **1:3.** La cote est donc de 1 pour 3.

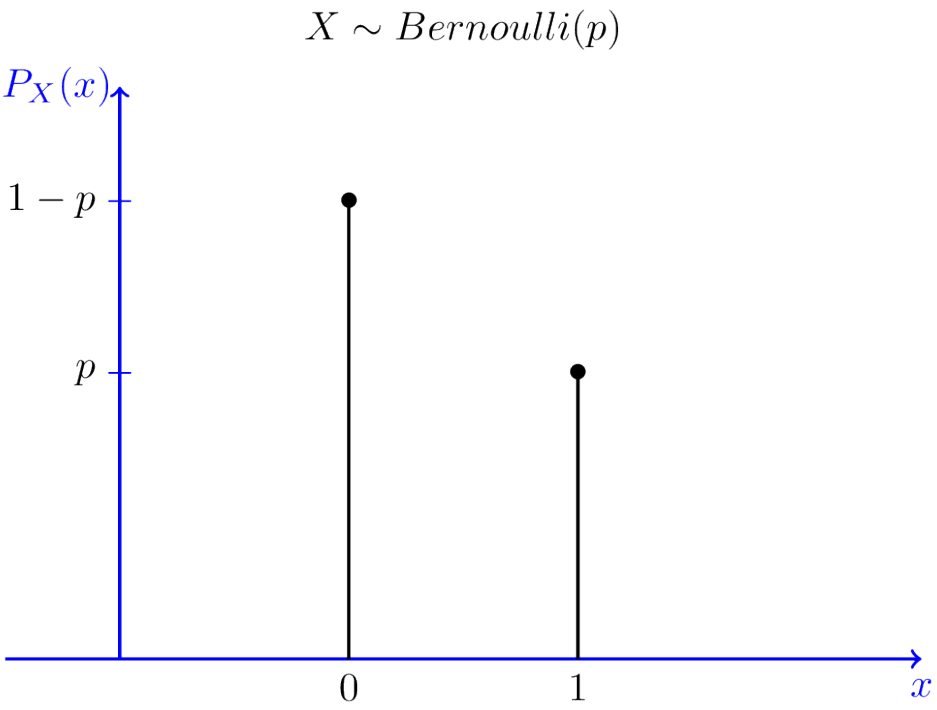
Une fois la cote de tirer une carte carreau déterminée, nous pouvons aller plus loin en calculant un ratio de cotes (*odds ratio*). L'odd ratio pour une variable en régression logistique représente comment les cotes changent quand on augmente cette variable de 1 en gardant toutes les autres variables constantes. Pour ce faire nous allons prendre l'exemple de pièces tirées avec deux potentiels résultats : pile ou face.

La première pièce a une probabilité p de faire pile égale à 0,5 et face à 0,5. La cote (odd) est donc de 1 : 1 (0,5/0,5).
La deuxième pièce est biaisée et a une probabilité de faire pile égale à 0,7 et une face à 0,3. La cote pour pile est donc de **0.7/0.3 = 2.33 :1**. Nous voulons maintenant déterminer le odd ratio avec la formule suivante :
Odd1 : odd de faire pile pour la pièce normale
Odd2 : odd de faire pile pour la pièce biaisée

Odd1/odd2 = **0.5/0.5 / 0.7/0.3** = 0.7*0.5 / 0.5 * 0.3 = 0.35/0.15 = 2.33.
Cela signifie que nous avons 2,33 fois plus de chance d'avoir pile en utilisant la pièce biaisée que la pièce normale.
Un autre exemple : Grace à un ensemble de données, on a pu déterminer que le poids du corps et l'apnée du sommeil (exprimée de façon binaire 1 ou 0) ont un odd ratio de 1.07. Ce qui signifie qu'une personne gagnant 1 kilo aura 1.07 (7%) de chance en plus de développer l'apnée du sommeil. Une augmentation de 10 kilos augmentera l'odd à 1.98 (double les odds) et 20 kilos d'augmentation quadruplera les odds (3.87).
**Attention** à ne pas confondre l'odd et la probabilité. Si le poids est faible à la base, la probabilité reste faible alors que la cote a elle bien augmenté.

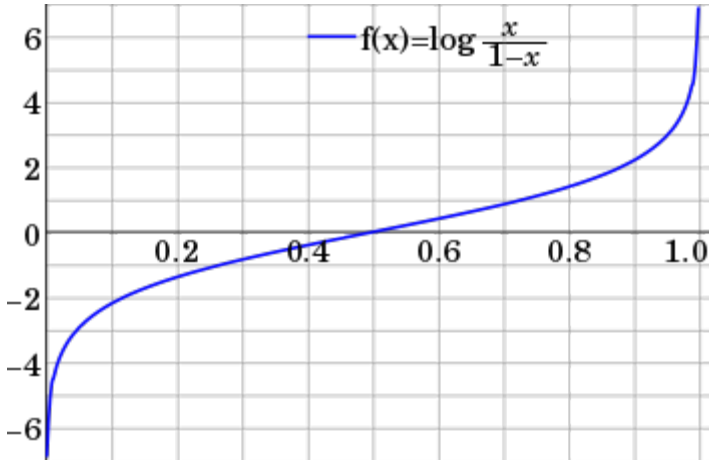
Le but de la régression logistique est d'estimer **p** selon une combinaison linéaire de variables indépendantes. On cherche donc à trouver p̂.

Ce que l'on souhaite faire, c'est donc utiliser une fonction qui va combiner ensemble notre distribution linéaire de variables et la distribution de Bernoulli, où les variables sont représentées par des 1 et des 0 (voir image ci-dessous).

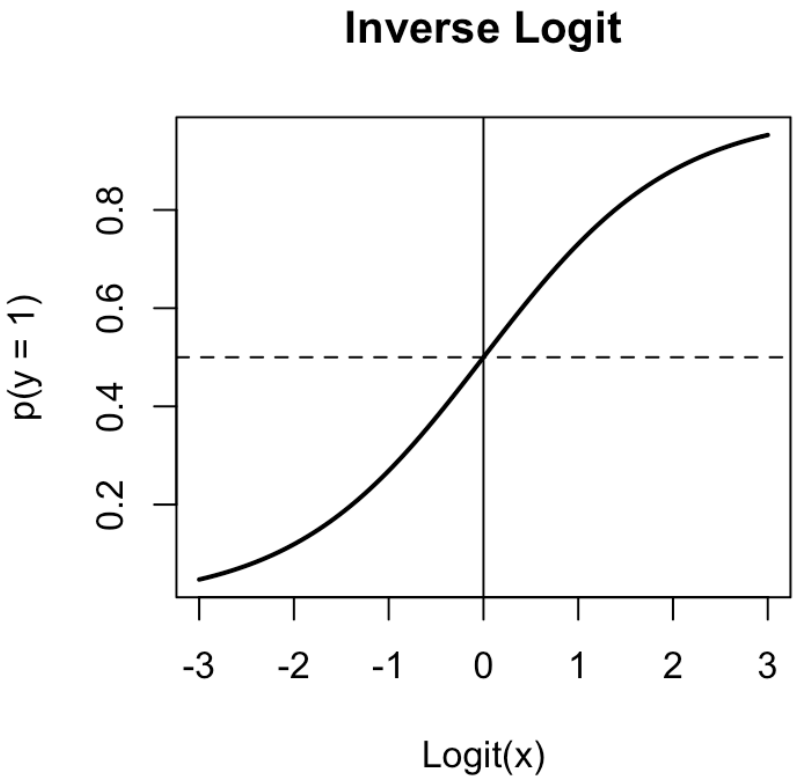



Autrement dit, on veut représenter la combinaison linéaire de variables qui pourraient avoir pour résultat n'importe quelle valeur entre 0 et 1.

La fonction que nous allons utiliser est le **logit.** Nous utilisons donc le **logit,** en appliquant un logarithme (de base 1) à l'odd.
Logit = ln(odds) = ln(p/1-p) ⇔ ln(p) – ln(1-p).




On obtient ainsi une distribution de variables entre 0 et 1. Mais nous voulons avoir les probabilités sur l'axe Y et non pas sur l'axe X. Dans ce cas, nous prenons l'inverse du logit pour avoir les probabilités sur l'axe Y, qui est la sigmoïde.



$Logit^{-1}(x) = 1 / 1 + e^{-x} = e^{x} / 1 + e^{x}$
La **sigmoïde**, ou l'inverse du logit, va nous retourner la probabilité qu'une variable soit un 1.

Nous avons donc notre formule pour calquer nos variables indépendantes sur une distribution de Bernoulli. Maintenant nous allons avoir besoin de nos coefficients de régression pour les intégrer dans notre formule de la sigmoid. Ces coefficients nous sont fournis par le Maximum Likelihood Estimation algorithme, qui n'a pas besoin d'être détaillé pour l'instant. Le logarithme naturel du odds ratio est équivalent à la fonction linéaire des variables indépendantes.
b0 et b1 sont nos coefficients de régression. x1 étant la variable dont on souhaite estimer la probabilité.

Logit (p) $= \ln (p / 1 - p) = b0 + b1x1.$

- ⇨ $p / 1 - p = e^{b0+b1x1}$
- ⇨ $p = e^{b0+b1x1}(1 - p)$
- ⇨ $p = e^{b0+b1x1} - e^{b0+b1x1} * p$
- ⇨ $p + e^{b0+b1x1} * p = e^{b0+b1x1}$
- ⇨ $p(1 + e^{b0+b1x1}) = e^{b0+b1x1}$
- ⇨ $e^{b0+b1x1}/1 + e^{b0+b1x1}$

Ainsi notre formule devient :
$\hat{p} = e^{b0+b1x1}/1 + e^{b0+b1x1}$
Cette formule porte pour nom la « Estimated Regression Equation » et nous permet donc d'obtenir la probabilité d'une variable selon une combinaison de variables linéaires.

Sources:
https://www.youtube.com/watch?v=zAULhNrnuL4 - la très bonne playlist de vidéos sur la regression logistique, jusqu'à la video 4.
http://udsmed.u-strasbg.fr/labiostat/IMG/pdf/RegLog_EAS_SMA1.pdf - un cours allant plus loin sur la regression logistique appliquees au monde de la pharmacie
https://fr.wikipedia.org/wiki/Logit
https://en.wikipedia.org/wiki/Logistic_regression
https://fr.wikipedia.org/wiki/R%C3%A9gression_logistique
https://en.wikipedia.org/wiki/Sigmoid_function