

# Modèles log-linéaires et régression de Poisson

- I. Introduction
- II. Régression linéaire avec fonction logarithme
- III. Modèle log-linéaire sans offset
- IV. Modèle log-linéaire avec offset
- V. Modèle de Poisson avec ou sans offset, et autres modèles
- VI. Conclusion



# Introduction

- I. Objectifs
- II. Rappels sur les modèles de régression
- III. Fonctions exponentielle et logarithme



# *Introduction*

## Prérequis, objectifs



- Objectifs :
  - Comprendre le modèle log-linéaire et les bases du modèle de Poisson (détaillé plus tard)
- Prérequis :
  - Comprendre et savoir interpréter les régressions linéaires multiples
  - Connaître risque relatif et odds ratio
  - Comprendre et savoir interpréter les régressions logistiques
  - Facultatif : comprendre les analyses de survie



# *Introduction*

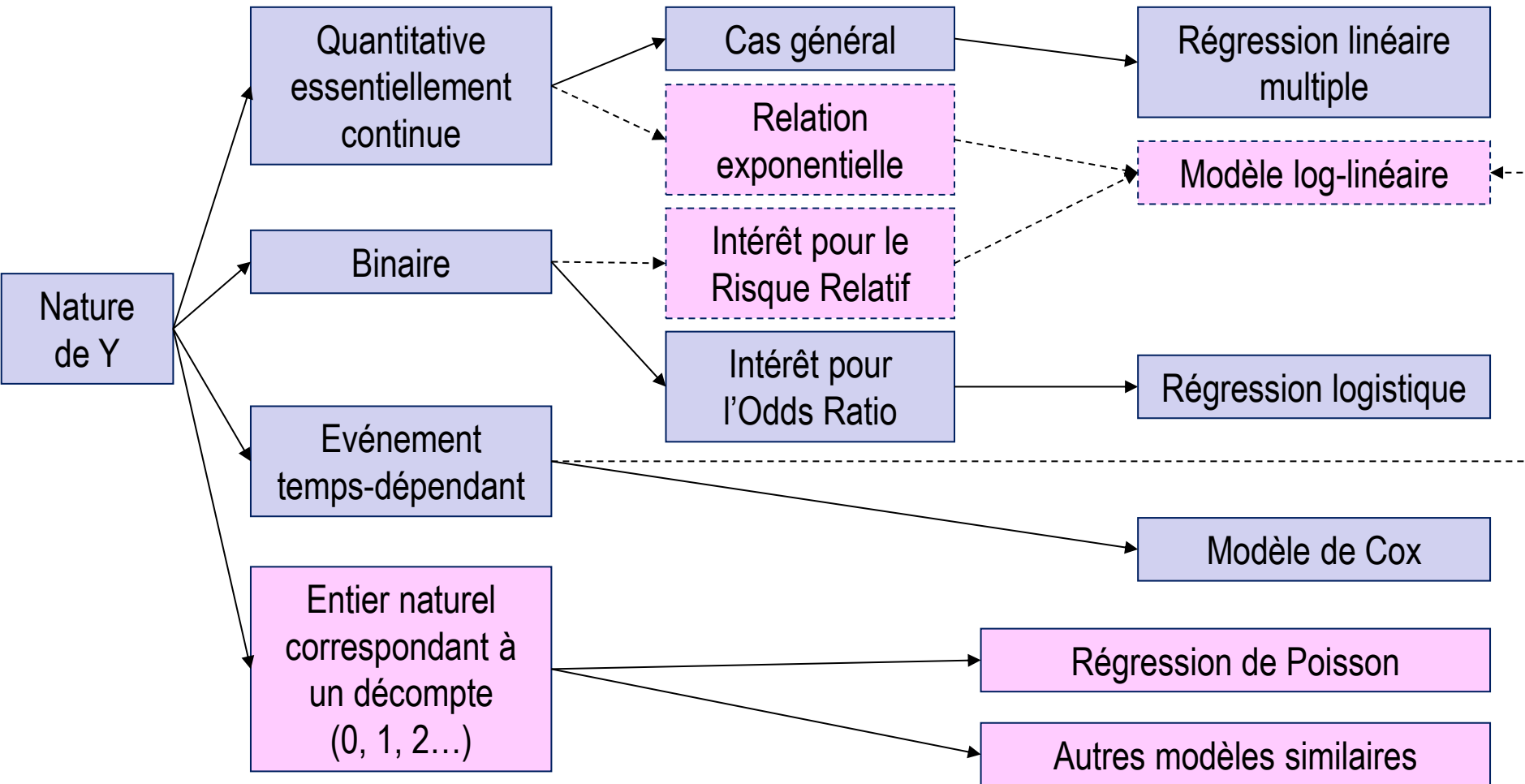
## Modèles de régression

- Familles de méthodes multivariées « phare » en recherche en santé :
  - **Explication** : dans un jeu de données, expliquer une variable  $Y$  par des variables  $X_i$  quantitatives/binaires/qualitatives (effet ajusté)
  - **Prédiction** : ensuite seulement, prédire la valeur  $\hat{y}_j$ , avec intervalle de confiance, d'un nouvel individu  $j$  dont les valeurs  $x_{i,j}$  sont connues
- Procédé :
  - Une variable à expliquer  $Y$
  - Des variables explicatives  $X_i$
  - Exemple avec 3 variables explicatives  $X_1, X_2$  et  $X_3$  :  
Ajustement :  $Y$  en fonction de  $(b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3)$
- Choix de la méthode : dépend principalement de la nature de  $Y$



# Introduction

## Modèles de régression



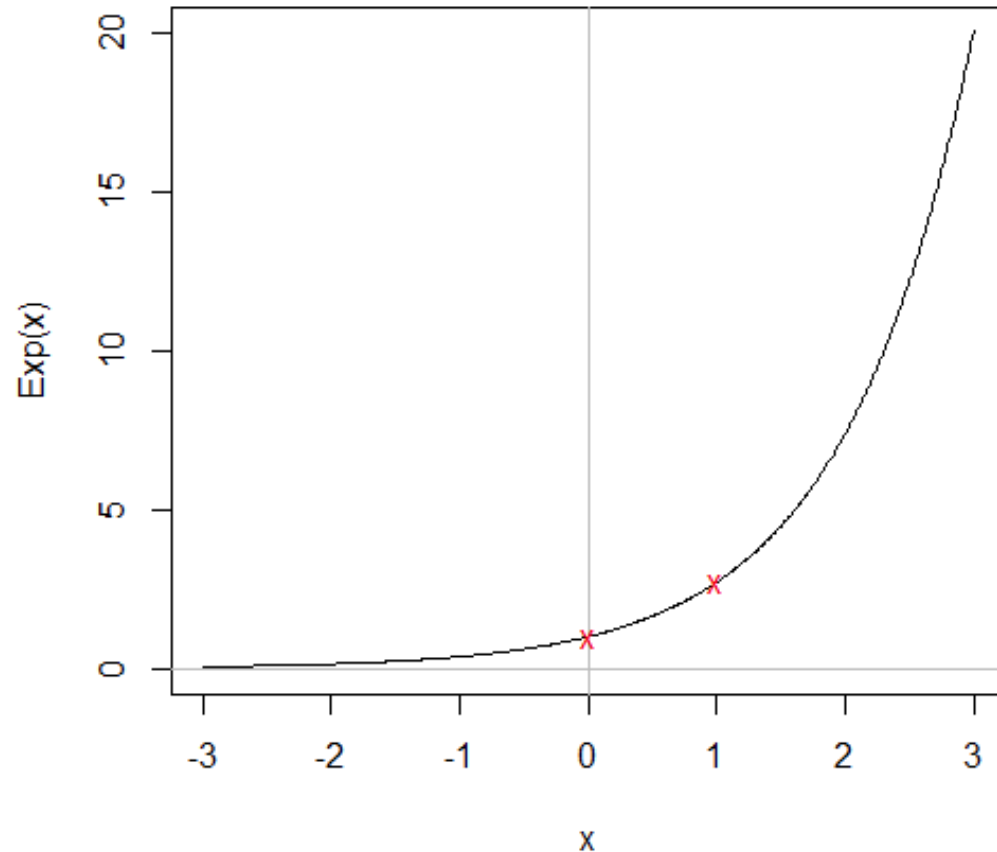
*Ce schéma sera complété à la fin*

# Introduction

## Fonction exponentielle



- On note :
$$y = \exp(x) = e^x$$
$$e^0 = 1$$
$$e^1 = e$$
- Nombre d'Euler, ou constante de Neper :
$$e \approx 2,71828$$
- Généralise les puissances, avec  $x$  éventuellement non-entier
- Quelques particularités :
  - Définie sur  $\mathbb{R}$
  - « sans effet mémoire » : une translation donnée sur  $X$  a toujours le même effet multiplicatif sur  $Y$ , quelle que soit la position sur  $X$
  - $e^{a+b} = e^a \times e^b$

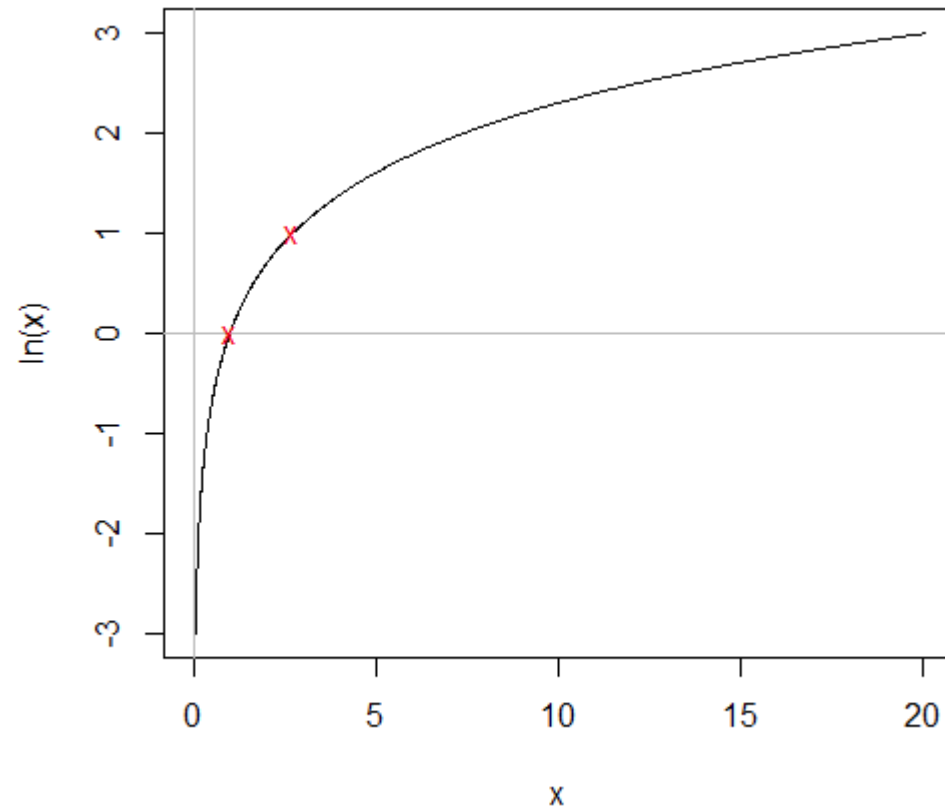


# Introduction

## Fonction Logarithme



- Fonction  $\ln()$ 
  - Logarithme népérien, ou naturel
  - Réciproque de l'exponentielle
  - Ne pas noter « log »\*
- On note :
  - $y = \ln(x) \Leftrightarrow x = e^y$
  - $\ln(1) = 0$
  - $\ln(e) = 1$
- Particularités :
  - Définie sur  $\mathbb{R}^{+*}$
  - Est une primitive de  $1/x$
  - $\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$



\*Normes AFNOR de 1961 et ISO 80000-2

# Régression linéaire avec fonction logarithme





# Expliquer $\ln(Y)$ par une combinaison linéaire de $X_i$



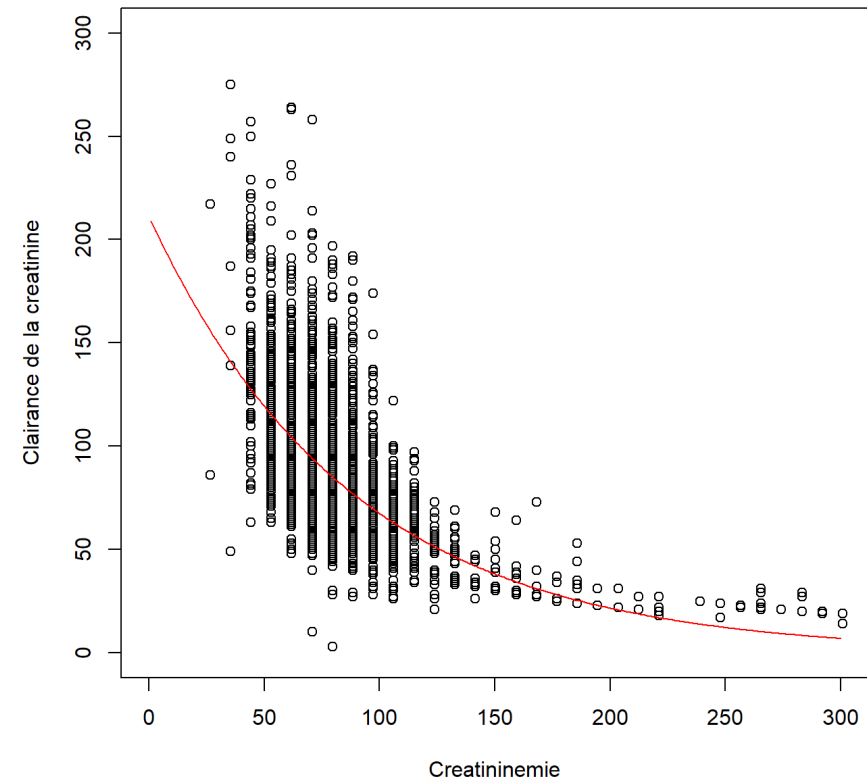
- Lorsque cela semble approprié :

$$\ln(Y) = b_0 + b_1 \cdot x_1 + \dots + b_i \cdot x_i$$

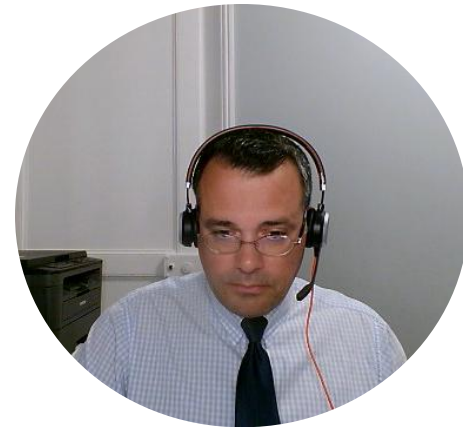
$$Y = e^{(b_0 + b_1 \cdot x_1 + \dots + b_i \cdot x_i)}$$

- Exemple :

- Clairance de la créatinine ( $Y$ ) en fonction du taux de créatinine ( $X$ )
- On obtient
$$b_0 = 5,35$$
$$b_1 = -0,11$$
- On trace la courbe
$$Y = e^{5,35 - 0,11 \cdot x}$$
- $e^{-0,11} = 0,90$  donc chaque augmentation d'un point de créatininémie est associée à une multiplication de la clearance de la créatinine par 0,90



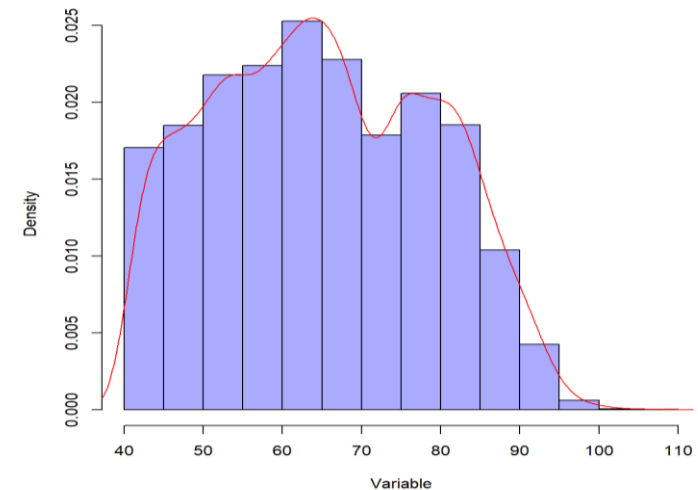
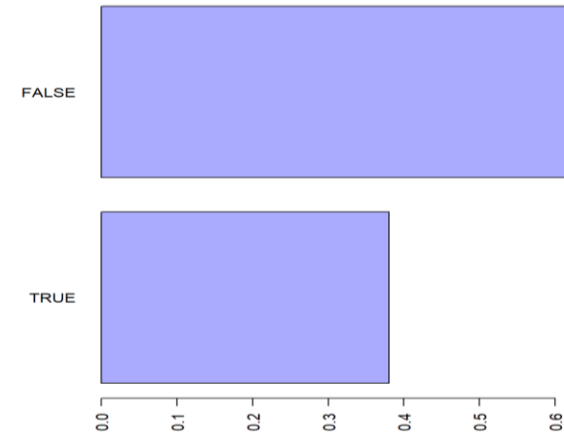
# Modèle log-linéaire (sans offset)



# $Y$ est une variable binaire



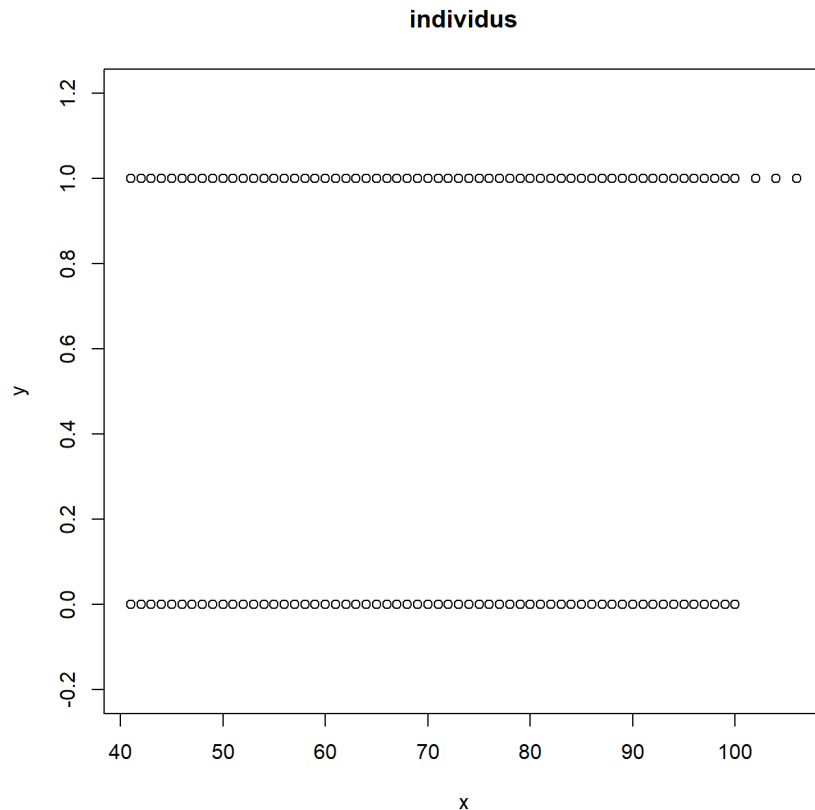
- Cas d'étude :  
10 000 séjours hospitaliers
- $Y$  : séjour lourd oui/non
  - 0=FALSE=non
  - 1=TRUE=oui  
(durée  $\geq 4$  ou décès ou transfert ou réanimation)
- $X$  : âge



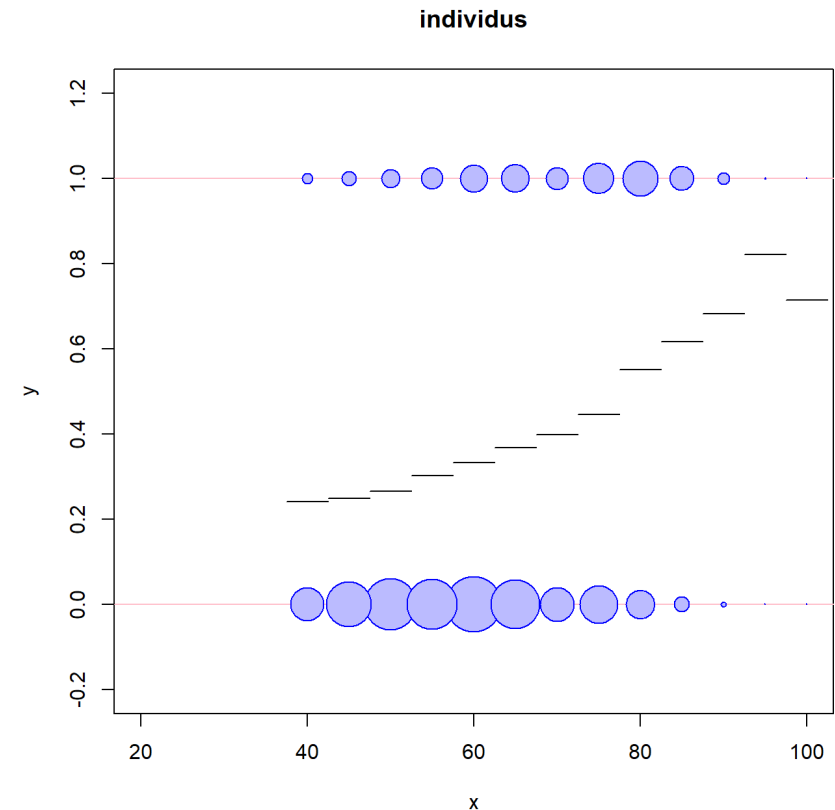
# Séjour lourd (Y) en fonction de l'âge (X)



- Nuage de points, avec âge entier :



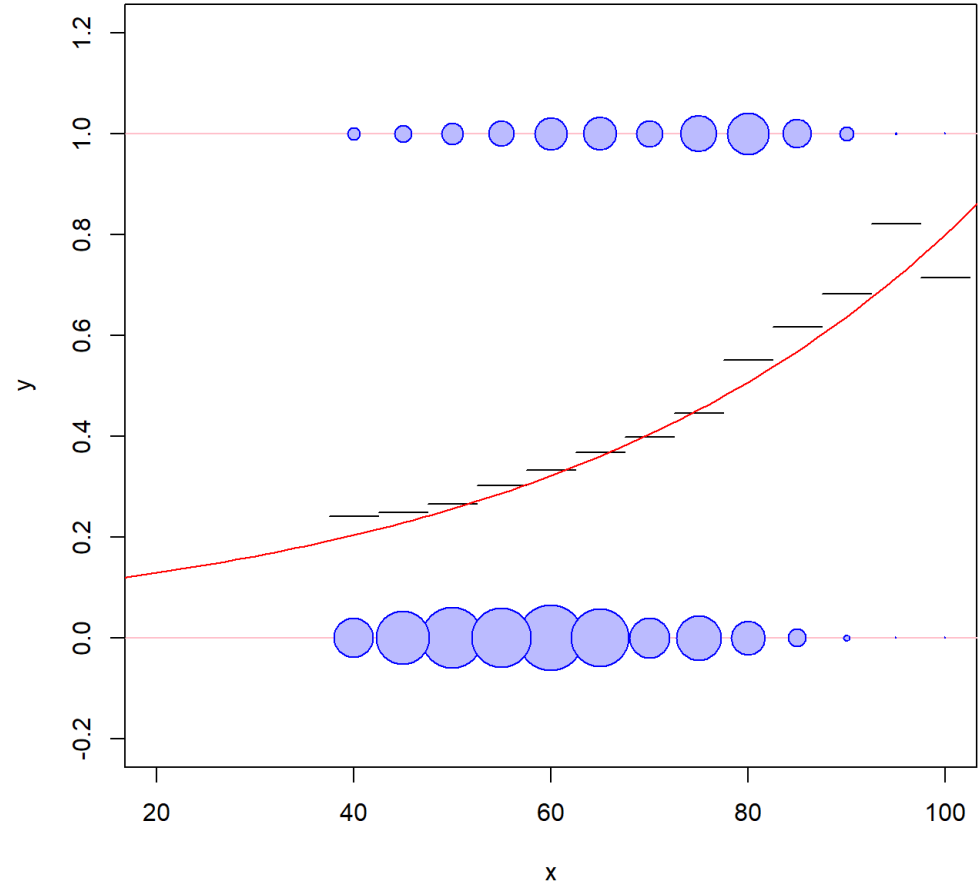
- Avec âge en classes de 5 ans, et moyenne de Y par classe (donc proportion de séjours lourds) :



# Modèle log-linéaire sur notre cas pratique



- $Y$  = séjour lourd
- $X$  = âge en années
- Procédé (automatisé)
  - Lancer le modèle  
 $Y = e^{(b_0 + b_1 \cdot X_1)}$
  - Ajustement par le « maximum de vraisemblance », calcule les valeurs de  $b_0$  et  $b_1$
  - $\hat{Y} = e^{(b_0 + b_1 \cdot X_1)}$   
est la probabilité prédite que  $Y$  vaille 1, sachant  $X$
- Ici :
  - $b_0 = -2,50$  et  $b_1 = 0,0228$
  - $e^{b_1} = 1,02$  pour chaque année d'âge supplémentaire, le risque est multiplié par 1,02

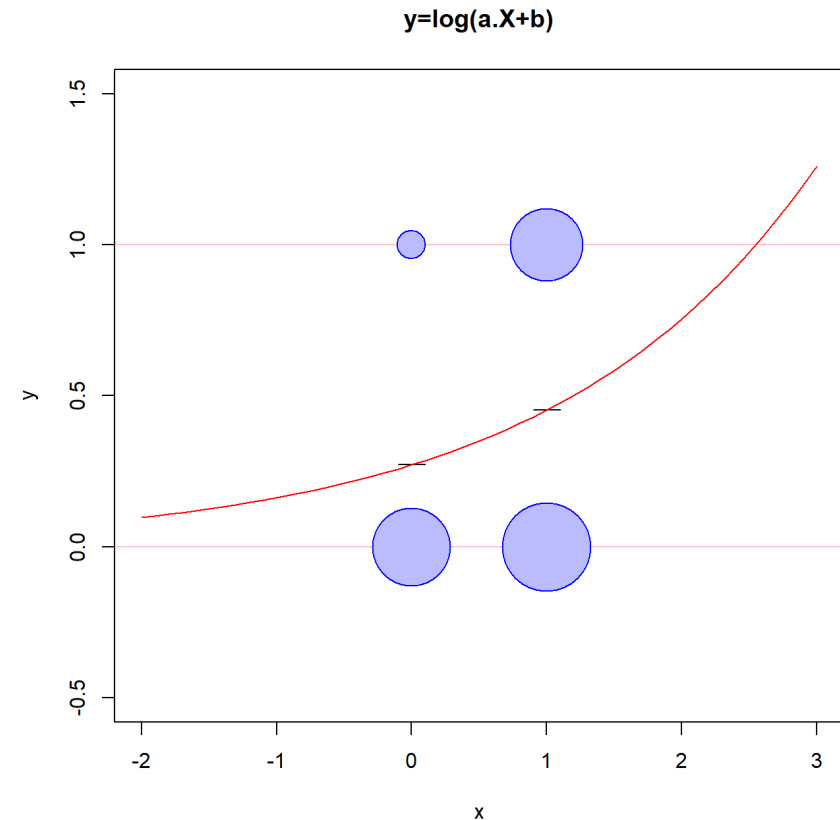


$$b_1 = 0,0228 \quad b_0 = -2,50$$

# Et quand X devient binaire ?



- $Y$  = « séjour lourd »
- $X$  = « âge > 60 ans », ne prend que deux valeurs :
  - 0 (pour faux, les moins de 60 ans)
  - 1 (pour vrai, les plus de 60 ans)
- L'ajustement  $Y = e^{(b_0 + b_1 \cdot X_1)}$  fonctionne aussi :
  - $b_1 = 0,512$
  - $b_0 = -1,31$
  - $e^{b_1} = 1,37$
- Calcul manuel :
  - risque relatif =  $1,37 = e^{b_1}$
- Il est démontré que pour les variables binaires,  $\exp(\text{coeff}) = \text{RR}$
- Ceci est généralisé pour tous les types de variables : l'exponentielle du coefficient sera toujours appelée « risque relatif » et interprétée de même : FDR si  $\text{RR} > 1$ , facteur protecteur si  $\text{RR} < 1$



# Comparaison avec la régression logistique



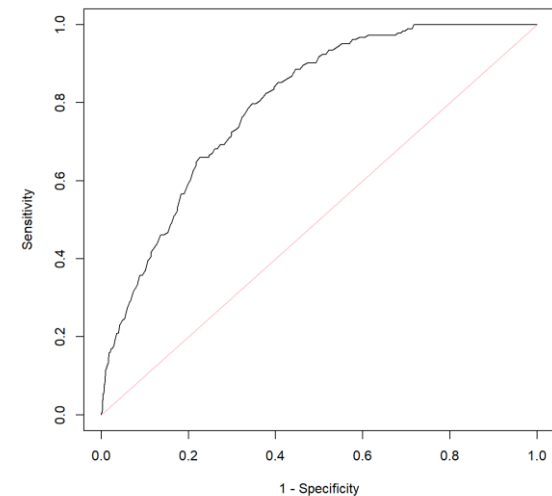
|                                                  | Modèle log-linéaire                            | Régression logistique                            |
|--------------------------------------------------|------------------------------------------------|--------------------------------------------------|
| Variable à expliquer                             | Binaire                                        |                                                  |
| Variables explicatives                           | Binaires, quantitatives, qualitatives          |                                                  |
| Fonction de lien                                 | Exponentielle<br>(réciproque : logarithme)     | Sigmoïde<br>(réciproque : logit)                 |
| Méthode d'ajustement                             | Maximum de vraisemblance                       |                                                  |
| Exp(coefficients)                                | Risque relatif ajusté                          | Odds ratio ajusté                                |
| Evaluation de la performance globale             | Aire sous la courbe ROC                        |                                                  |
| Test de la significativité de chaque coefficient | Test de Wald                                   |                                                  |
| Avantage différenciant                           | Prise en compte d'un effet volume : « offset » | Prise en compte d'une saturation des proportions |

# Exemple de modèle log-linéaire multivarié



- Variable à expliquer : décès (1,82%)
- Variables explicatives, simultanément :
  - Âge en dizaines (quantitatif continu)
  - Sexe masculin (binaire)
  - Cancer (qualitatif)
- Exponentielle des coefficients = risques relatifs ajustés

| Variable     | N     | Estimate |                   | p      |
|--------------|-------|----------|-------------------|--------|
| cancer       | 10000 |          | 2.90 (1.88, 4.29) | <0.001 |
| homme        | 10000 |          | 1.48 (1.10, 1.99) | 0.009  |
| age10continu | 10000 |          | 1.87 (1.69, 2.08) | <0.001 |

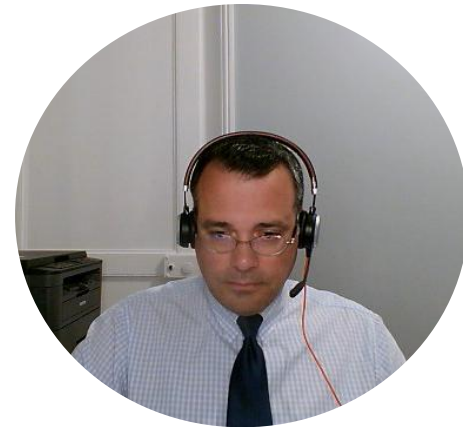


AUC : 0,797 [0,771 ; 0,824]

*Besoin d'explications plus détaillées ? Voir cours sur la régression logistique*



# Modèle log-linéaire AVEC offset



# Parfois besoin de prendre en compte un « effet volume »



- Exemple 1 :
  - Individu statistique : individus
  - Événement : subir un infarctus du myocarde
  - Effet volume : durée du suivi (variable), en jours
  
- Exemple 2 :
  - Individu statistique : familles
  - Événement : subir une infestation de poux (peu importe le nombre d'individus)
  - Effet volume : nombre d'enfants scolarisés
  
- Exemple 3 :
  - Individu statistique : hôpitaux
  - Événement : subir une fermeture administrative
  - Effet volume : nombre de services

# Pris en compte d'un « effet volume » par l'OFFSET



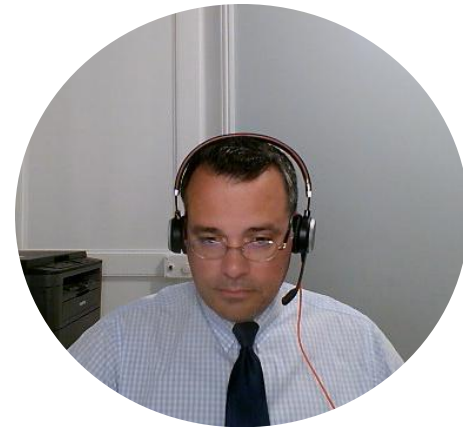
## ■ Modèle :

- $\ln\left(\frac{Y}{offset}\right) = b_0 + b_1.X_1 + \dots + b_i.X_i$
- $\Leftrightarrow \ln(Y) = b_0 + b_1.X_1 + \dots + b_i.X_i + \ln(offset)$
- $\Leftrightarrow Y = e^{(b_0 + b_1.X_1 + \dots + b_i.X_i + \ln(offset))}$

## ■ Commentaires :

- $\ln(offset)$  devient une nouvelle variable pour laquelle on ne calculera pas de coefficient
- Interprétation des autres coefficients inchangée (mais pertinente car tient compte de l'offset)
- Intéressant car pas d'effet mémoire dans l'exponentielle  
=> peu importe l'unité de l'offset !

# Introduction au modèle de Poisson

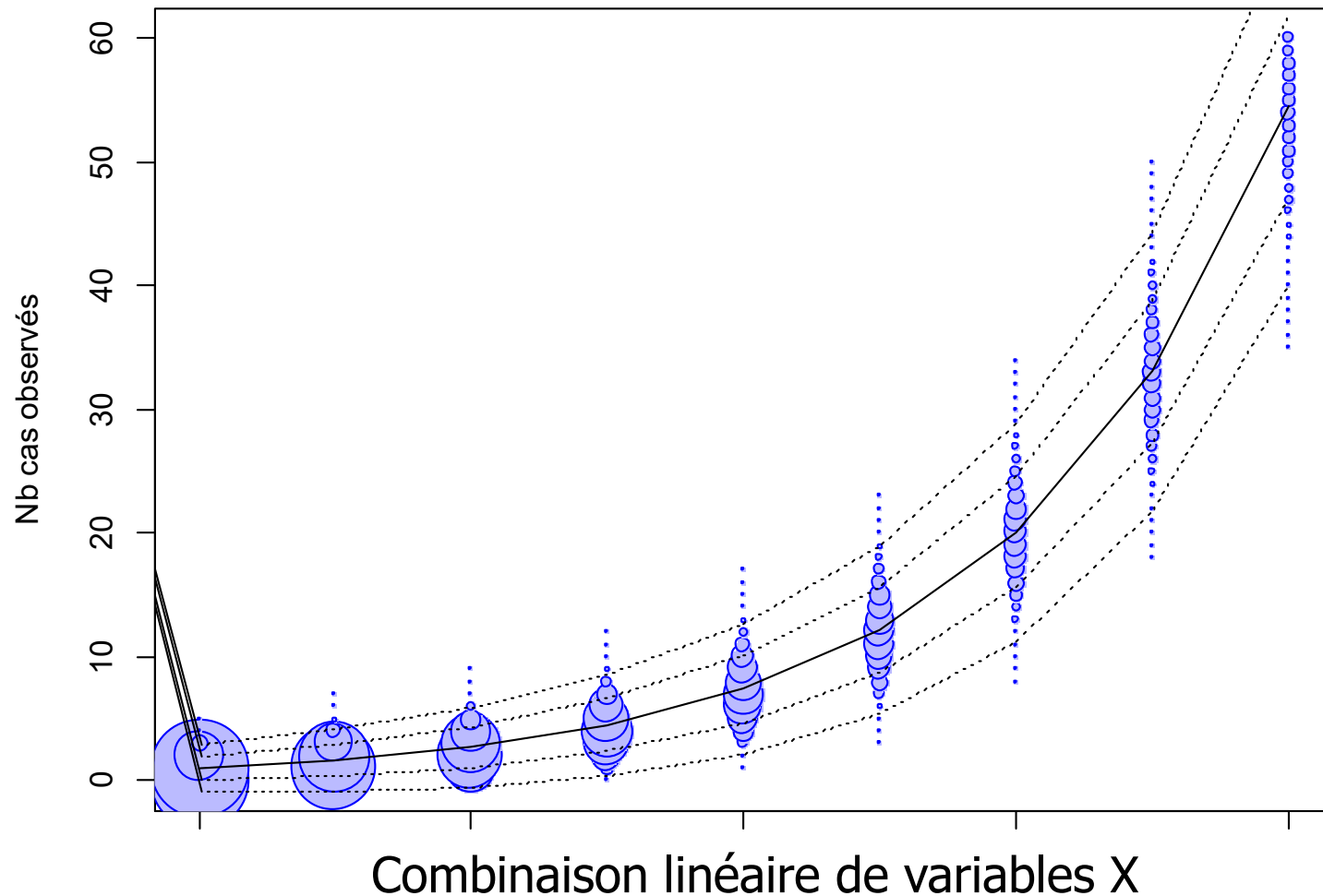


# Contexte



- Variable à expliquer :
  - Y observé :
    - Nombre d'événements, entier naturel  $\{0, 1, 2, 3 \dots\}$
    - Doit suivre une loi de Poisson : conditions de validité assez précises (cf. cours dédié)
  - Y prédit :
    - Nombre *moyen* d'événements attendu
    - = paramètre lambda d'une loi de Poisson
- Variables explicatives :
  - Binaires, quantitatives, qualitatives (idem précédemment)
- Modèle :
  - $\ln(Y) = b_0 + b_1 \cdot x_1 + \dots + b_i \cdot x_i$
  - $Y = e^{(b_0 + b_1 \cdot x_1 + \dots + b_i \cdot x_i)}$

# Principe de la modélisation d'une loi de Poisson



# Parfois besoin de prendre en compte un « effet volume »



- Avec un offset, exactement comme précédemment
- Exemple 1 :
  - Individu statistique : départements
  - Événement : nombre de cas de cancers
  - Offset : nombre d'habitants de chaque département
- Exemple 2 :
  - Individu statistique : patients
  - Événement : nombre de consultations durant le suivi
  - Offset : durée du suivi en jours
- Commentaires :
  - Aucune contrainte sur l'unité ou l'échelle de l'offset, seul le ratio importera (exponentielle : sans effet mémoire)
  - Pas capable de prendre en compte une saturation : idéal pour les événements rares

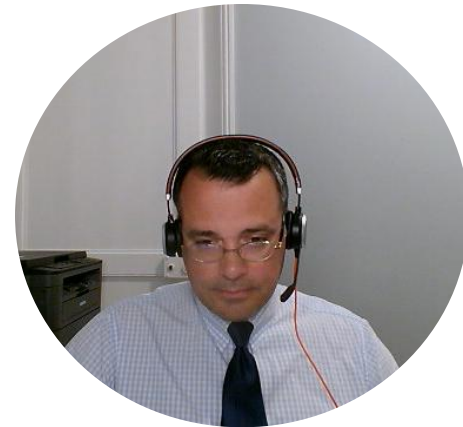
# Autres modèles



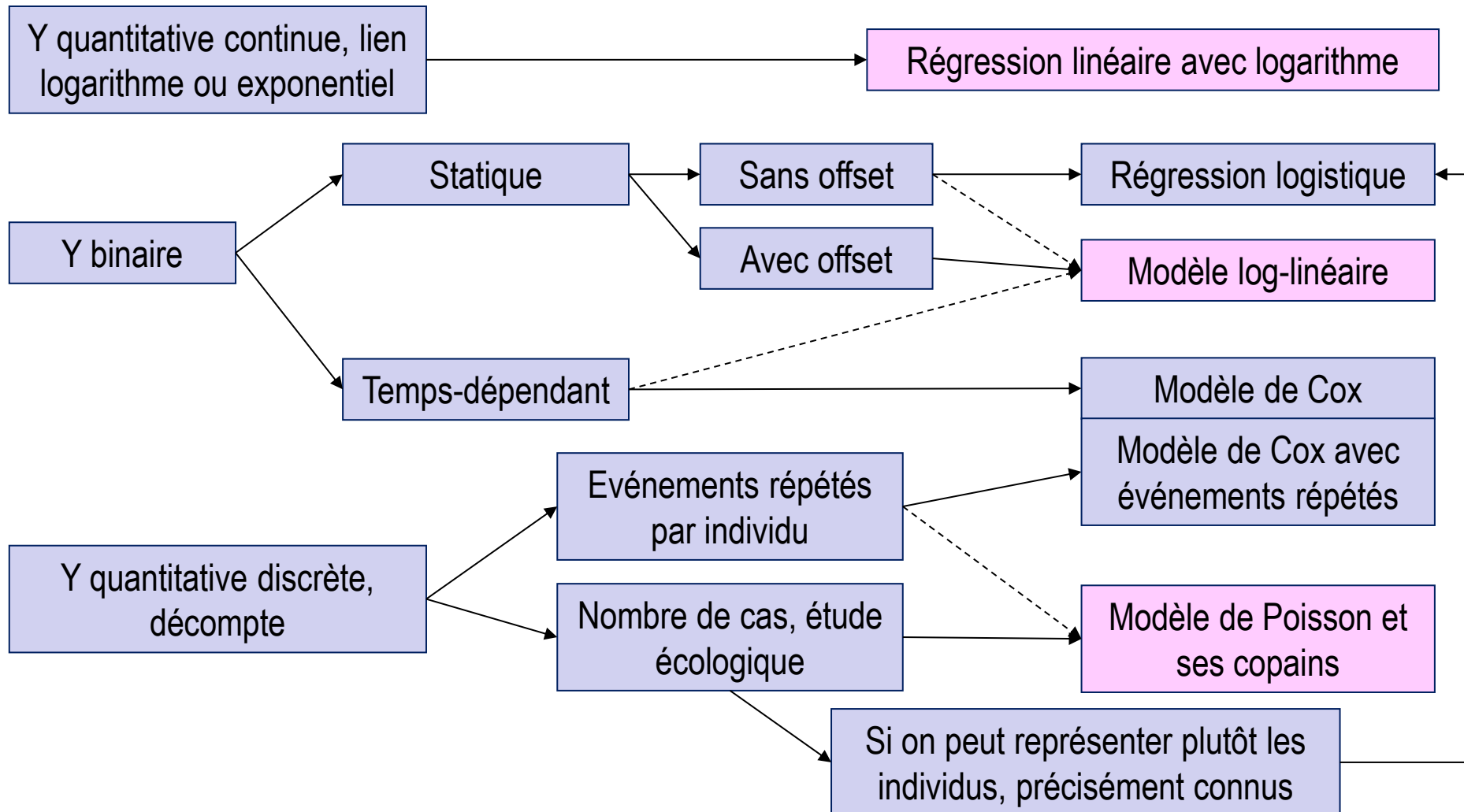
- Problème principal du modèle de Poisson :
  - Hypothèse très forte : le décompte suit une loi de Poisson
  - Conditionnellement aux  $X$ , *moyenne=variance=lambda*
  - Hypothèse souvent invalide !
    - « sur-dispersion » si variance > moyenne
    - « sous-dispersion » si variance < moyenne
  
- Autres modèles si hypothèse invalide :
  - Quasi-Poisson
  - Loi binomiale négative
  - Modèle de Poisson « zero-inflated »
  - Etc.



# Conclusion



# Utilisation pratique en recherche en santé



**Merci de votre attention !**  
**RDV sur**  
**<http://objectifthese.org>**

