

# Régression Logistique

# Méthodes explicatives

Variable à  
expliquer

Variables explicatives  
 $X_1, X_2, \dots, X_k$

Y	Quantitatives	Qualitatives	Mélange
Quantitatif	Régression multiple	Analyse de la variance	Modèle linéaire général
Qualitatif	<ul style="list-style-type: none"> <li>- Régression Logistique</li> <li>- Segmentation</li> <li>- Analyse factorielle discriminante</li> <li>- Analyse discriminante bayésienne</li> </ul>	<ul style="list-style-type: none"> <li>- Régression Logistique</li> <li>- Segmentation</li> <li>- Analyse factorielle discriminante</li> </ul>	<ul style="list-style-type: none"> <li>- Régression Logistique</li> <li>- Segmentation</li> <li>- Analyse factorielle discriminante</li> </ul>

# Régression logistique binaire

- Les données:

$Y$  = variable à expliquer binaire

$X_1, \dots, X_k$  = variables explicatives numériques ou binaires  
(indicatrices de modalités)

- Régression logistique simple ( $k = 1$ )
- Régression logistique multiple ( $k > 1$ )

# Régression logistique simple

- Variable dépendante :  $Y = 0 / 1$
- Variable indépendante :  $X$
- Objectif : Modéliser

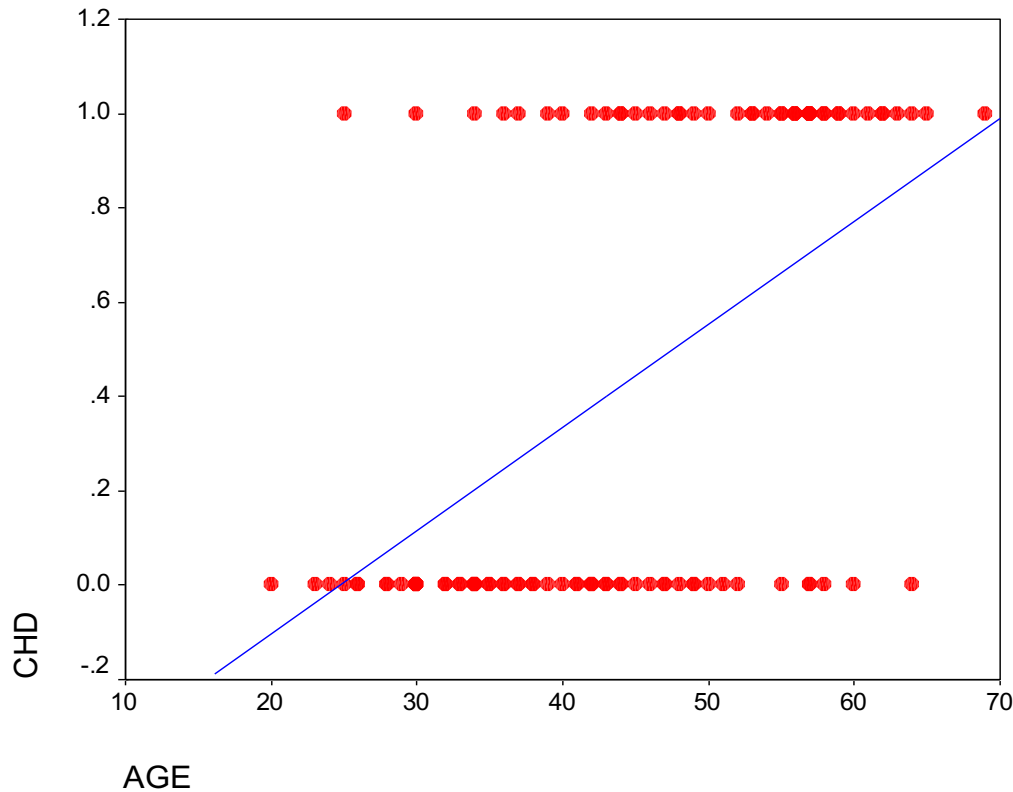
$$\pi(x) = \text{Prob}(Y = 1/X = x)$$

- Le modèle linéaire  $\pi(x) = \beta_0 + \beta_1 x$  convient mal lorsque  $X$  est continue.
- Le modèle logistique est plus naturel.

# Exemple

## Age and Coronary Heart Disease Status (CHD)

Plot of CHD by Age



Les données

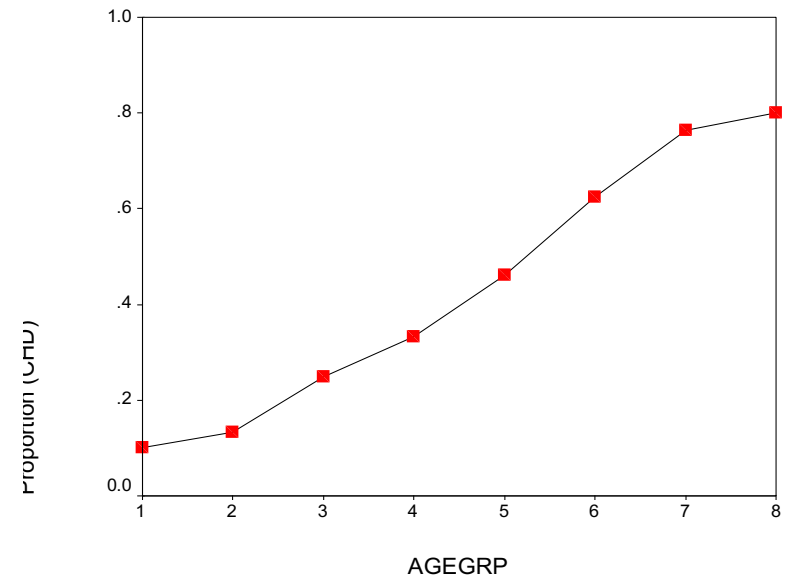
ID	AGRP	AGE	CHD
1	1	20	0
2	1	23	0
3	1	24	0
4	1	25	0
5	1	25	1
⋮	⋮	⋮	⋮
97	8	64	0
98	8	64	1
99	8	65	1
100	8	69	1

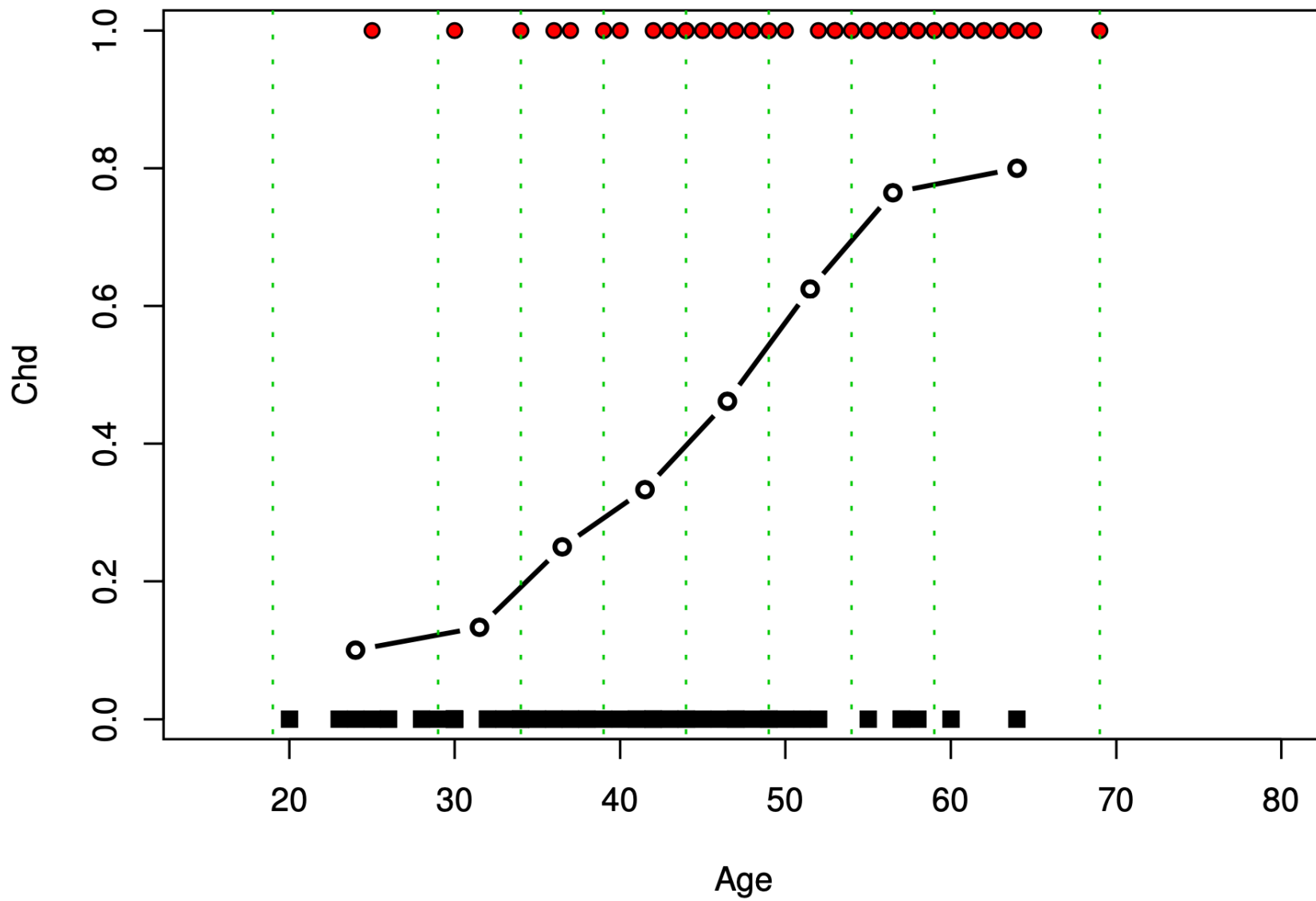
# Description des données regroupées par classe d'âge

Tableau des effectifs  
de CHD par classe d'âge

Age Group	n	CHD absent	CHD present	Mean (Proportion)
20 – 29	10	9	1	0.10
30 – 34	15	13	2	0.13
35 – 39	12	9	3	0.25
40 – 44	15	10	5	0.33
45 – 49	13	7	6	0.46
50 – 54	8	3	5	0.63
55 – 59	17	4	13	0.76
60 – 69	10	2	8	0.80
Total	100	57	43	0.43

Graphique des proportions  
de CHD par classe d'âge

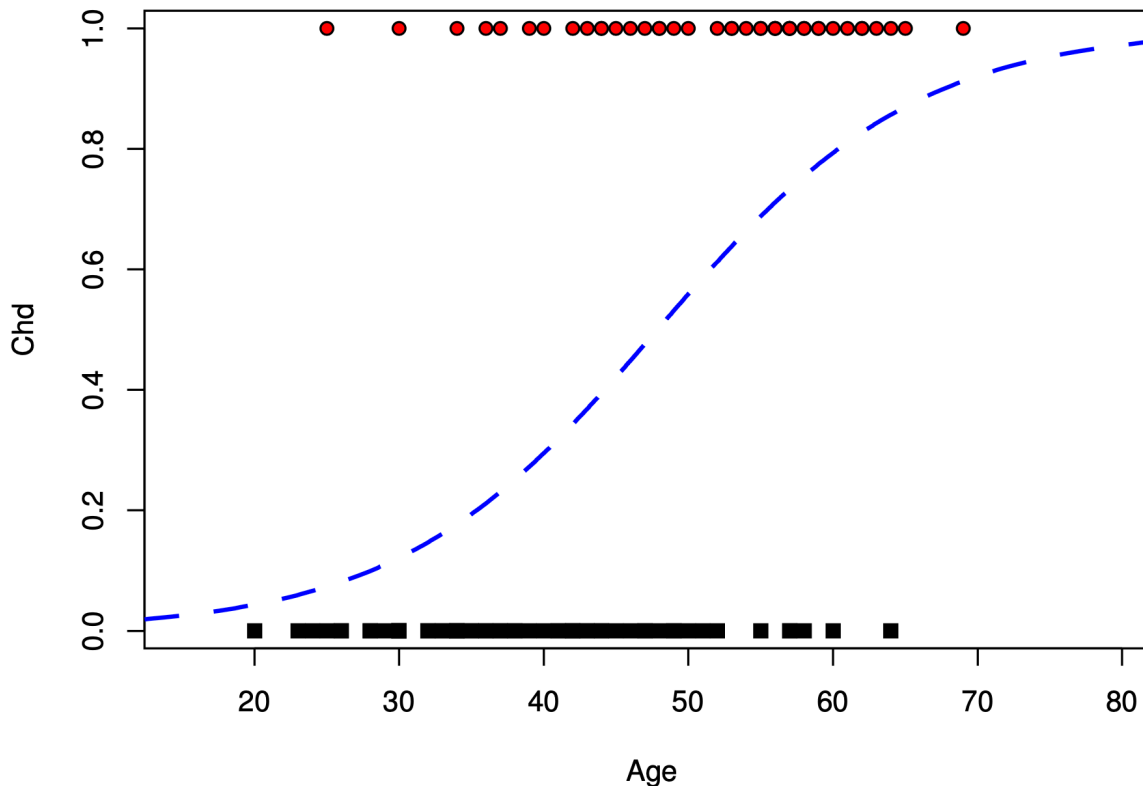




# Fonction souhaitée

On souhaiterait trouver une fonction:

- un peu plus régulière
  - qui utilise toutes les données (sinon faire des classes qui varient avec  $x$ )
- pour obtenir par exemple

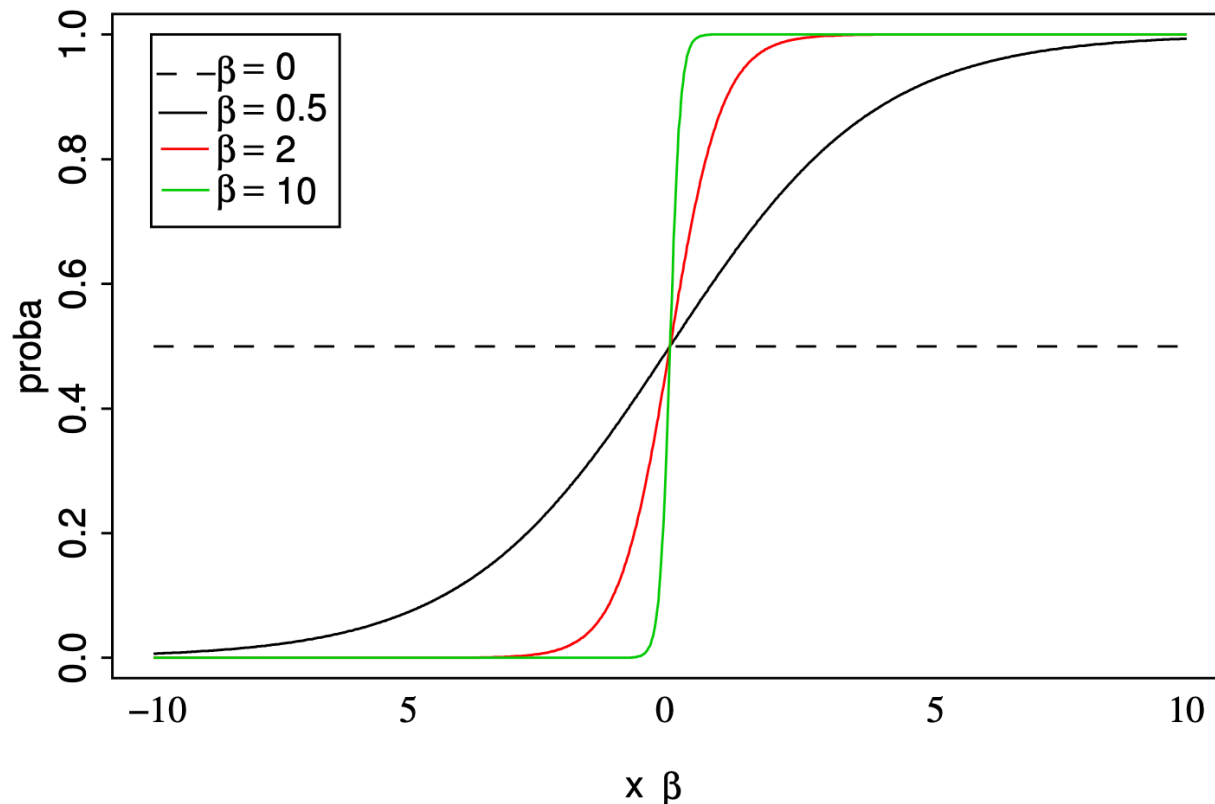




# Equation d'une courbe en S

Une première façon d'obtenir une courbe en S est de considérer

$$x \rightarrow \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$



## $Y$ variable binaire

Ici la variable  $Y$  prend 2 valeurs, modélisons

$$(Y|X = x) \sim \mathcal{B}(p(x))$$

$$\mathbb{P}(Y = 1|X = x) = p(x) \quad \text{et} \quad \mathbb{P}(Y = 0|X = x) = 1 - p(x)$$

Nous avons donc

$$\mathbb{E}_x(Y) = p(x)$$

$$\text{Var}_x(Y) = p(x)(1 - p(x)) \quad \text{hétéroscédasticité}$$

## Comparaison modèle linéaire

Dans le modèle linéaire

$$\mathbb{E}(Y|x) = x'\beta$$

Quand  $Y$  est binaire, on a

$$\mathbb{E}(Y|x) = p(x) \text{ à valeurs dans } [0, 1]$$

mais il existe des transformations  $g$  (appelées fonctions de lien) tq

$$g(p(x)) = x'\beta$$

## La fonction « logit »

$$\mathbb{E}(Y|X = x) = p(x) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

La fonction « logit » :

$$p \mapsto g(p) = \log\left(\frac{p}{1-p}\right)$$

est bijective (dérivable) et nous avons

$$g(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = x'\beta$$

# Modèle logistique

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

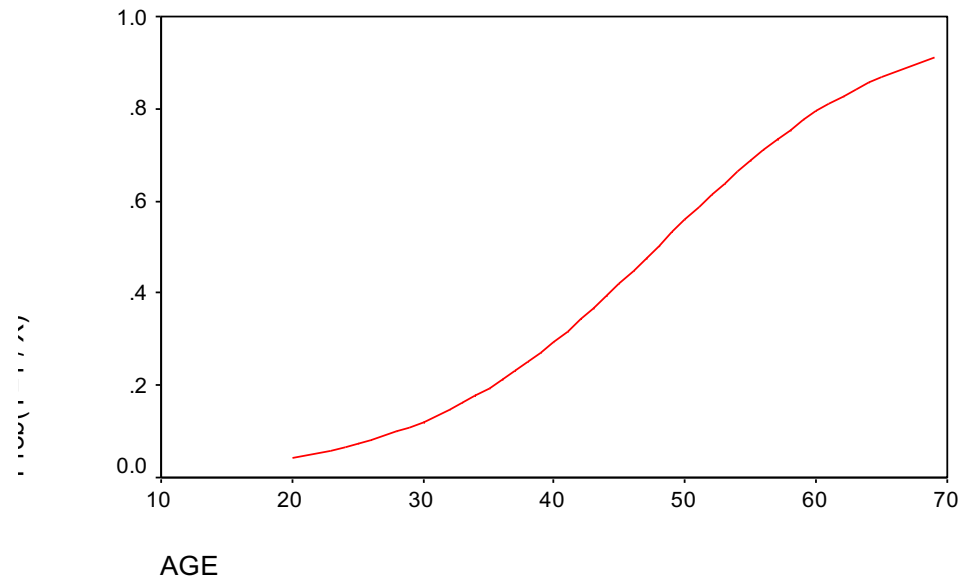
ou

$$\text{Log}\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x$$



*Fonction de lien : Logit*

Probabilité d'une maladie cardiaque  
en fonction de l'age



# Fonctions de lien

- Fonction **logit**

$$g(p) = \log(p / (1 - p))$$

- Fonction **normit** ou **probit**

$$g(p) = \Phi^{-1}(p)$$

où  $\Phi$  est la fonction de répartition de la loi normale réduite

- Fonction ‘**complementary log-log**’

$$g(p) = \log(-\log(1-p))$$

# Estimation des paramètres du modèle logistique

## Les données

<b>X</b>	<b>Y</b>
<b>x<sub>1</sub></b>	<b>y<sub>1</sub></b>
<b>⋮</b>	<b>⋮</b>
<b>x<sub>i</sub></b>	<b>y<sub>i</sub></b>
<b>⋮</b>	<b>⋮</b>
<b>x<sub>n</sub></b>	<b>y<sub>n</sub></b>

$y_i = 1$  si caractère présent,  
0 sinon

## Le modèle

$$\begin{aligned}\pi(x_i) &= P(Y = 1 / X = x_i) \\ &= \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\end{aligned}$$

## Définition

1. « Choix » d'une loi pour  $(Y|X = x)$  : Bernoulli
2. Choix d'une fonction  $g$  : fonction logit
3. Modéliser  $\mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$  grâce à

$$g \{ \mathbb{P}(Y = 1|X = x) \} = x' \beta$$

Les paramètres  $\beta$  sont inconnus !



## Estimation de $\beta$ par MV

### Definition

La vraisemblance du modèle est définie par :

$$L_n(y_1, \dots, y_n, \beta) = \prod_{i=1}^n \mathbf{P}(Y = y_i | X = x_i)$$

que nous noterons simplement  $L_n(\beta)$ .

## Ecriture de la vraisemblance

Exprimons la vraisemblance en fonction de  $\beta$  :

$$L_n(\beta) = \prod_{i=1}^n \mathbf{P}(Y = y_i | X = x_i) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}.$$

En passant au log, on obtient

$$\mathcal{L}_n(\beta) = \sum_{i=1}^n \{y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))\}$$

après quelques calculs à faire en exercice

$$= \sum_{i=1}^n \{y_i x_i' \beta - \log(1 + \exp(x_i' \beta))\}$$

On cherche le maximum

On calcule les dérivées partielles et on les annule pour obtenir les équations normales :

$$\sum_{i=1}^n [x_i(y_i - p(x_i))] = X'(Y - P_{\beta}) = 0$$

Rappels du modèle linéaire

$$X'(Y - X\beta) = 0$$

# Maximisation de la vraisemblance

## Malheureusement...

Il n'existe pas de solutions explicites pour maximiser la vraisemblance (on n'aura donc pas d'écriture explicite pour  $\hat{\beta}$ ).

## Mais

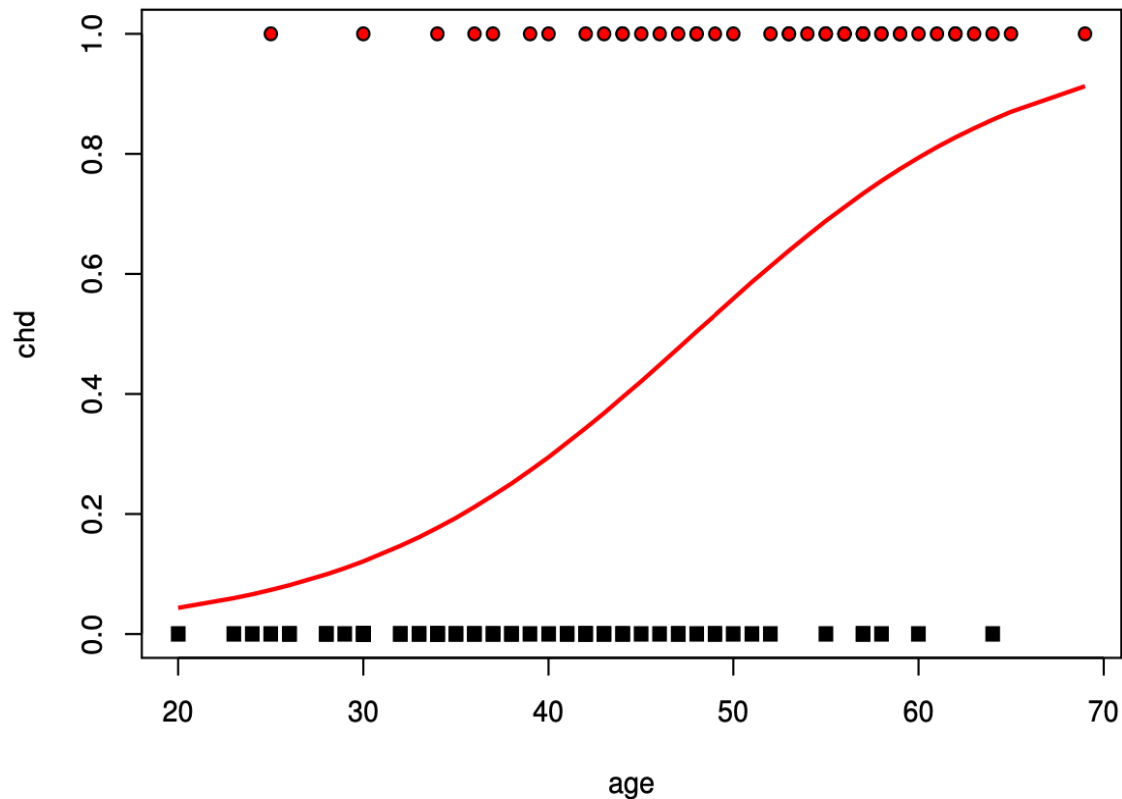
La vraisemblance possède (généralement) un unique maximum, et il existe des algorithmes numériques itératifs permettant d'obtenir ce maximum :

- ▶ algorithme de Newton ;
- ▶ algorithme du score de Fisher.

## Modèle ajusté

$$\hat{P}(Y = 1|age) = \frac{\exp(-5.30945 + 0.11092 \times age)}{1 + \exp(-5.30945 + 0.11092 \times age)}.$$

## Fonction estimée



## Interprétation directe

Quand le coefficient  $\beta_j$  associé à la variable  $X_j$  est

- ▶ positif :  $X_j$  augmente  $\rightarrow p$  augmente
- ▶ négatif :  $X_j$  augmente  $\rightarrow p$  diminue

Ici,  $\hat{\beta}_{age} = 0.11$ , donc la probabilité augmente avec l'âge !

# Régression logistique avec R

- Tutoriel:

<http://perso.ens-lyon.fr/lise.vaudor/realiser-une-regression-logistique-avec-r/>

# Exercice

- Dataset:  
<http://web1.sph.emory.edu/dkleinb/allDatasets/datasets/evans.dat>
- **Présentation des données:** Les données à analyser contiennent les informations sur une cohorte de 609 hommes ayant été suivis sur une période de 7 ans. Il s'agit d'étudier la variable d'intérêt apparition ou non d'une maladie cardiaque des coronaires.

```
evans <- read.table("http://web1.sph.emory.edu/dkleinb/allDatasets/datasets/evans.dat")
```

```
head(evans)
```

```
names(evans) <- c("id", "chd", "cat", "age", "chl", "smk", "ecg", "dbp", "sbp", "hpt", "ch", "cc")
```

```
head(evans)
```



Les variables du data frame sont définies ci-dessous.

- **id**: identifiant du sujet. Chaque observation a un identifiant unique soit une observation par individu.
- **chd**: une variable dichotomique prenant la valeur 1 si la maladie est présente, 0 sinon.
- **cat**: une variable dichotomique indiquant si le niveau de catecholamine est élevé (1) ou non (0).
- **age**: une variable continue exprimée en années.
- **chl**: une variable continue définissant le taux de cholestérol.
- **smk**: une variable dichotomique indiquant si le sujet est fumeur (1) ou s'il n'a jamais fumé (0).
- **ecg**: une variable dichotomique indiquant la présence d'un électrocardiogramme anormal (1) ou non (0).
- **dbp**: une variable continue indiquant la pression artérielle diastolique.
- **sbp**: une variable continue indiquant la pression artérielle systolique.
- **hpt**: une variable dichotomique indiquant la présence (1) ou non (0) d'une forte pression sanguine.
- **ch**: une variable construite à partir du produit  $\text{cat} \times \text{hpt}$ .
- **cc**: une variable construite à partir du produit  $\text{cat} \times \text{chl}$ .

- L'objectif est de discuter le modèle logistique. Soit chd la présence (1) ou l'absence (0) de la maladie coronarienne avec comme variable d'exposition le niveau de catecholamine cat.  
age, chl, smk, ecg et hpt sont les variables de contrôle.