Régression Logistique

Introduction

- La régression logistique est une approche statistique.
- Employer pour évaluer les relations entre une variable réponse de type binaire (variable à expliquer), et une, ou plusieurs, variables explicatives

- Variable à expliquer : *par exemple vrai/faux, succès/échec, malade/non malade…)*
- Variables explicatives : *de type catégoriel (par exemple le sexe H/F) ou numérique continu (par exemple l'âge)*.

Régression logistique binaire

- <u>Les données:</u>

Y = variable à expliquer binaire

$X_1, \ldots, X_k$ = 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
- <u>Objectif :</u> Modéliser

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

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

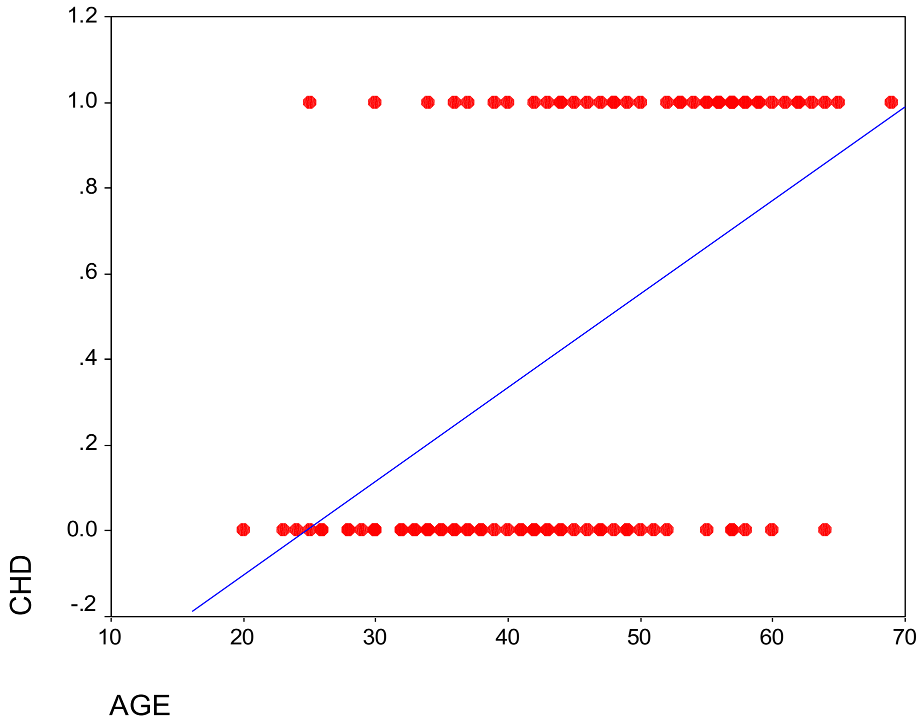
- Le modèle logistique est plus naturel.

Dans la régression logistique, ce n'est pas la réponse binaire qui est directement modélisée, mais la probabilité de réalisation d'une des deux modalités

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

La probabilité de réalisation ne peut pas être modélisée par une droite (car celle-ci conduirait à des valeurs < 0 ou > 1) ➔ impossible (une probabilité est forcément bornée par 0 et 1).

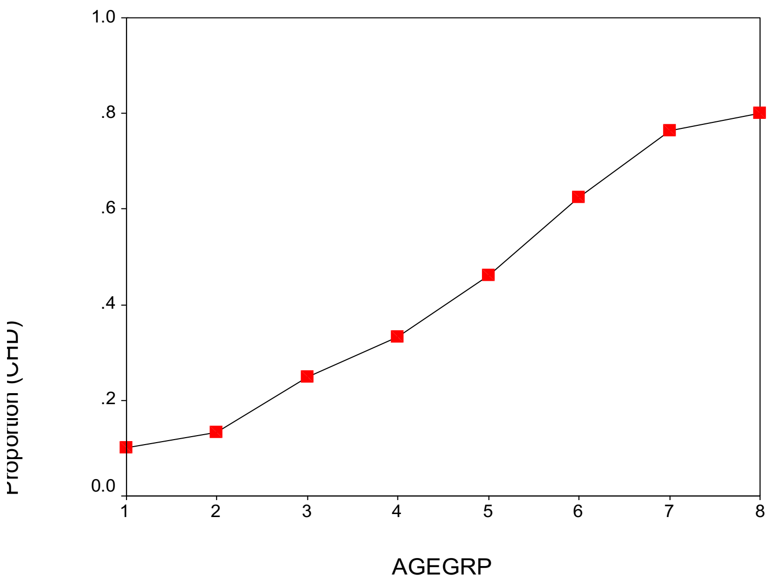
Cette probabilité, est alors modélisée par une courbe sigmoïde, bornée par 0, et 1

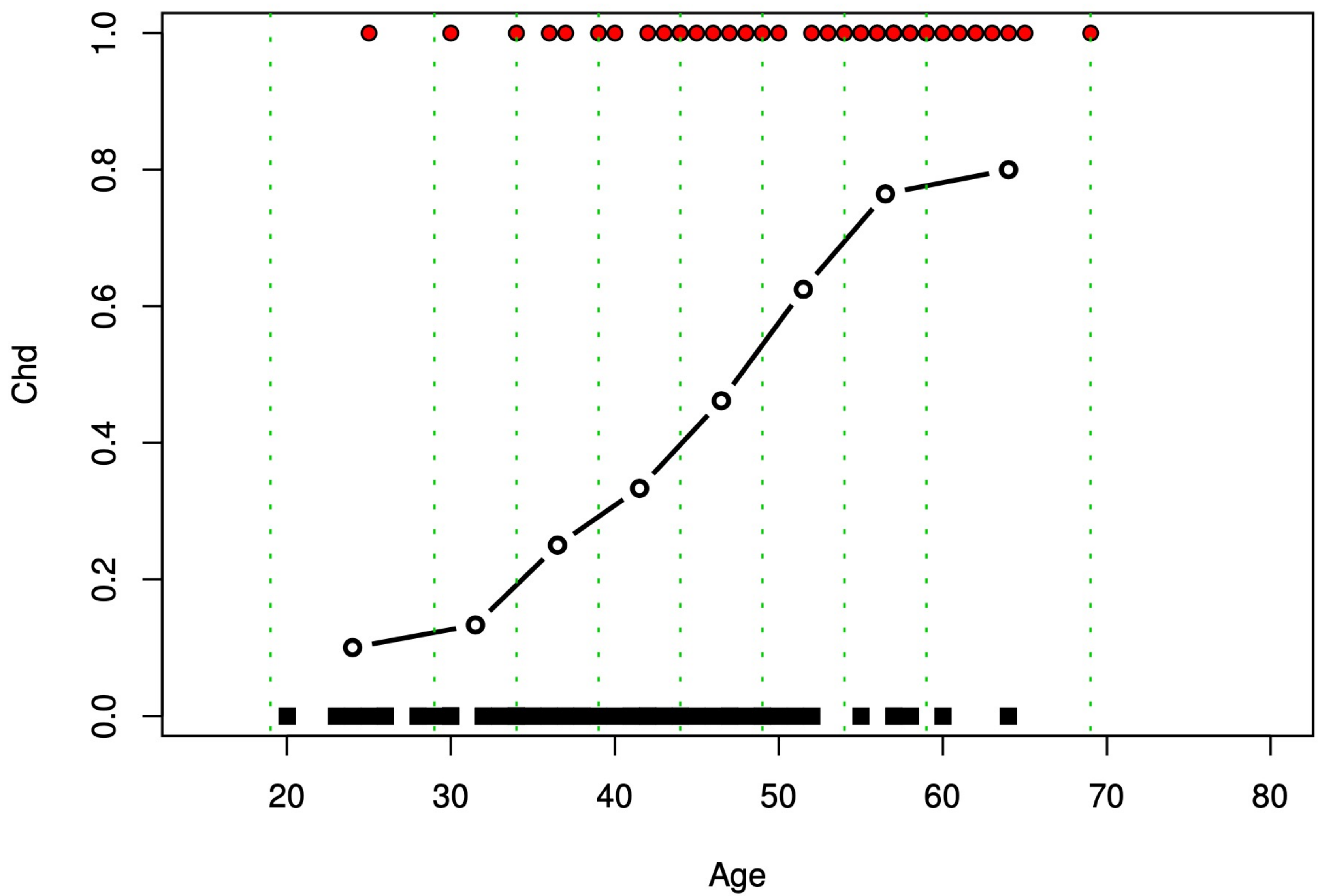
Description des données regroupées par classe d'âge

Tableau des effectifs de CHD par classe d'age

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'age



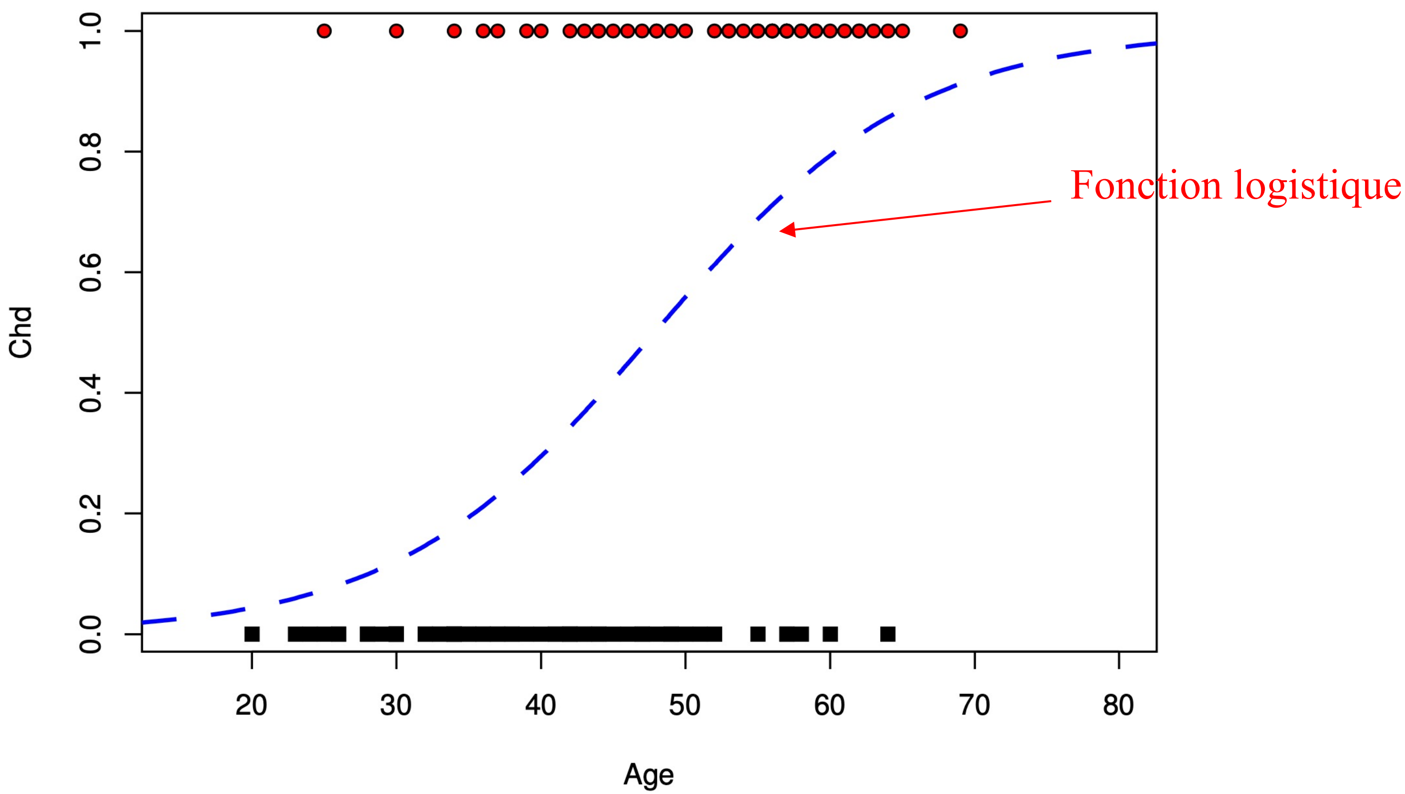

1.0
0.8
0.6
0.4
0.2
0.0
Chd
20
30
40
50
60
70
80
Age

Fonction souhaitée

On souhaiterait trouver une fonction:

- un peu plus régulière
- qui utilise toutes les données (sinon faire des classes qui varient avec x)

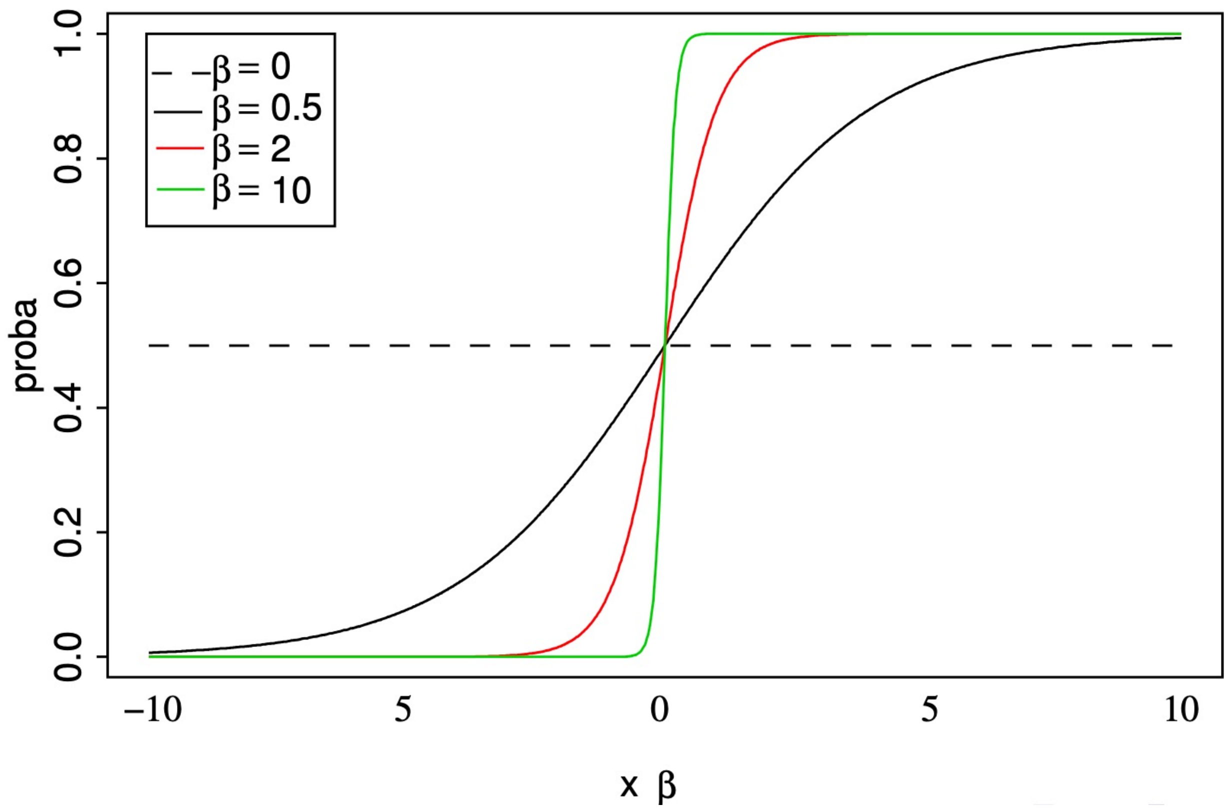
pour obteni



Equation d'une courbe en S

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

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



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

$$\begin{aligned} \mathbb{E}_x(Y) &= p(x) \\ Var_x(Y) &= p(x)(1 - p(x)) \quad \text{hétéroscédasticité} \end{aligned}$$

Comparaison modèle linéaire

Dans le modèle linéaire

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

Quand Y est binaire, on a

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

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

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

La fonction « logit »

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

La fonction « logit » :

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

est bijective (dérivable) et nous avons

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

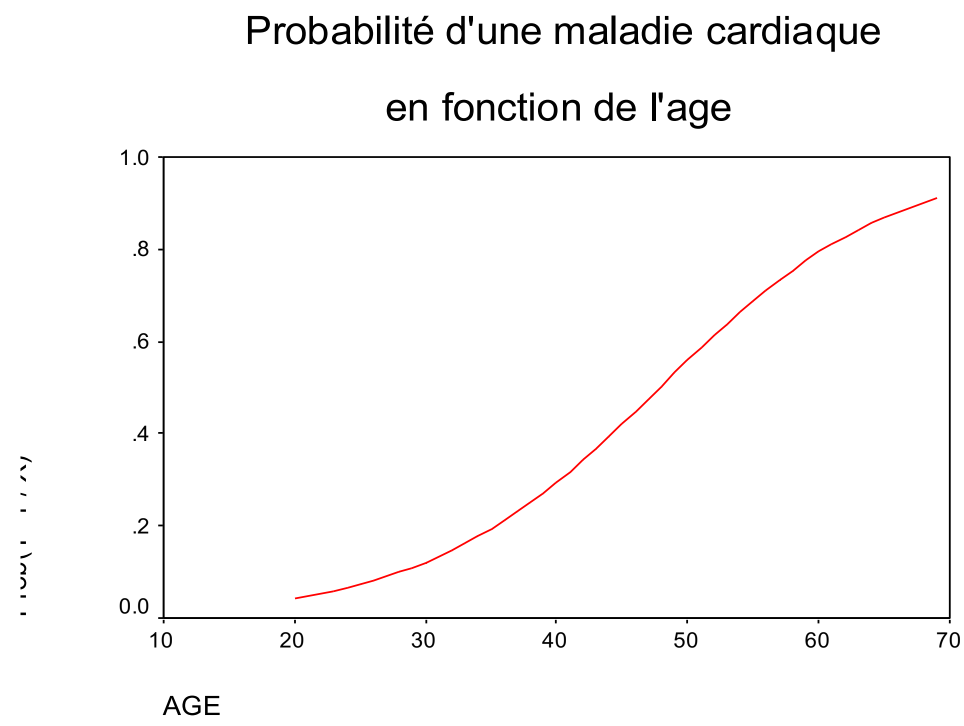
Modèle logistique

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

ou

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

Fonction de lien : Logit



Fonctions de lien

- Fonction logit

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

- Fonction normit ou probit

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

où Φ est la fonction de répartition de la loi normale réduite

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}t^2}\, \mathrm{d}t, \text{ pour tout } x \in \mathbb{R}.$$

- Fonction ‘complementary log-log’

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

Estimation des paramètres du modèle logistique

Les données

X	Y
$\mathbf{x_1}$	$\mathbf{y_1}$
$\vdots$	$\vdots$
$\mathbf{x_i}$	$\mathbf{y_i}$
$\vdots$	$\vdots$
$\mathbf{x_n}$	$\mathbf{y_n}$

$y_i = 1$ si caractère présent,
0 sinon

Le modèle

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

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

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

Les paramètres β sont inconnus !

Estimation de β par MV

Definition

La vraisemblance du modèle est définie par :

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

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

Ecriture de la vraisemblance

Exprimons la vraisemblance en fonction de β :

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

En passant au log, on obtient

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \{y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))\}$$
$$\text{après quelques calculs à faire en exercice}$$
$$= \sum_{i=1}^{n} \{y_i x_i' \beta - \log(1 + \exp(x_i' \beta))\}$$

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

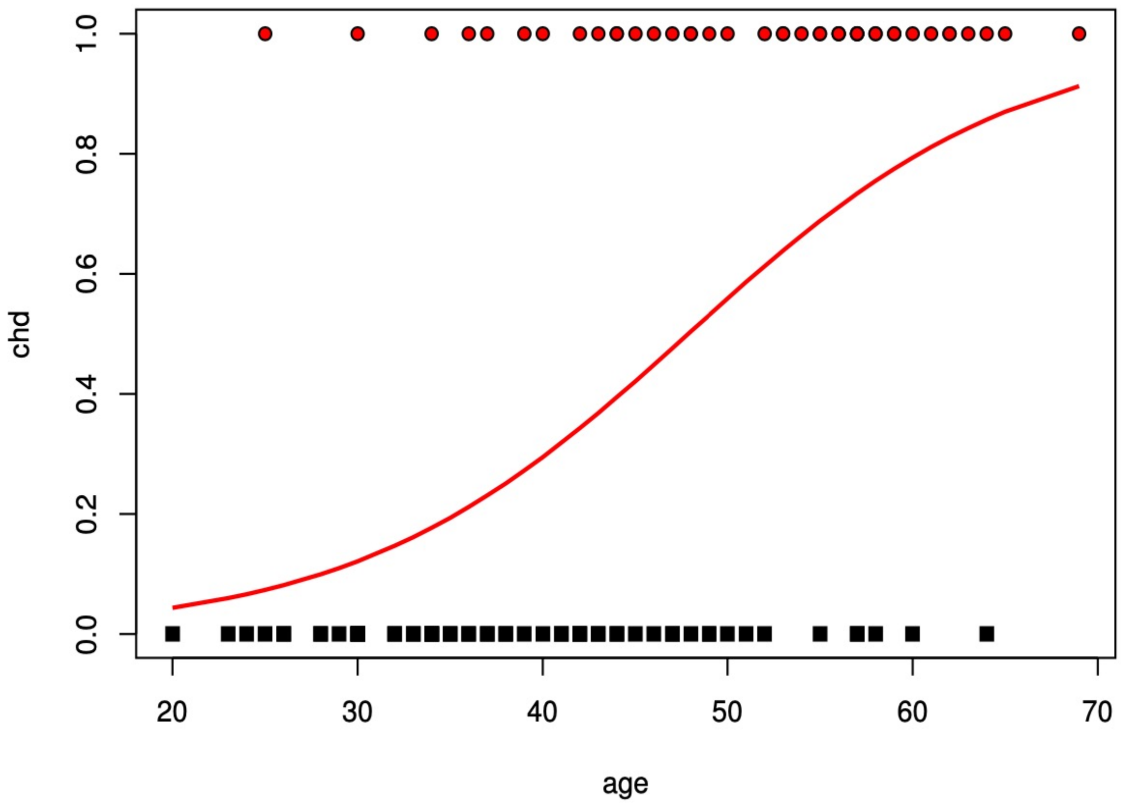
La vraisemblance possède (généralement) un unique maximum, et il existe des algorithmes numériques itératifs permettant d'obtenir ce maximum :

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

Modèle ajusté

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

Fonction estimée



Interprétation directe

Quand le coefficient β_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 !