



# Modèles pour variables qualitatives à deux modalités

David Causeur

*Laboratoire de Mathématiques Appliquées*

*Agrocampus Rennes*

*IRMAR CNRS UMR 6625*

*<http://www.agrocampus-rennes.fr/math/causeur/>*



# Plan du cours

- 1 Quelques situations concrètes
  - Modélisation de l'intention d'achat
  - Modélisation de l'efficacité d'un fongicide
- 2 Modèle pour variables qualitatives à deux modalités
  - Modèle logistique
  - Prédicteur linéaire et fonction de lien
- 3 Estimation des paramètres
  - Maximum de vraisemblance
  - Déviance résiduelle
  - Propriétés
- 4 Perspectives



## Intention d'achat

La promesse d'achat depend-t'elle de l'évaluation sensorielle ?

	Promesse d'achat	Note globale
1	0	6
2	N	7
3	N	5
4	0	9
5	0	8
6	0	5
⋮	⋮	⋮



## Intention d'achat

La promesse d'achat depend-t'elle de l'évaluation sensorielle ?

	Promesse d'achat	Note globale
1	0	6
2	N	7
3	N	5
4	0	9
5	0	8
6	0	5
⋮	⋮	⋮

Promesse d'achat : variable qualitative



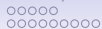
## Intention d'achat

La promesse d'achat depend-t'elle de l'évaluation sensorielle ?

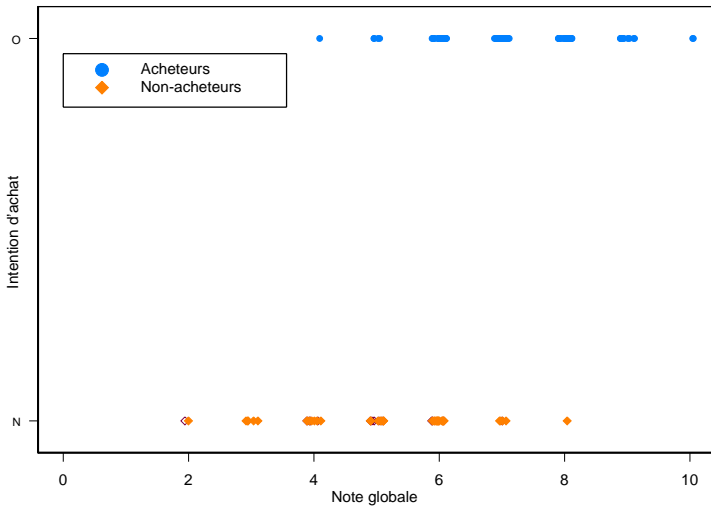
	Promesse d'achat	Note globale
1	0	6
2	N	7
3	N	5
4	0	9
5	0	8
6	0	5
⋮	⋮	⋮

Promesse d'achat : variable qualitative

Note globale : variable quantitative



# Intention d'achat





## Première approche

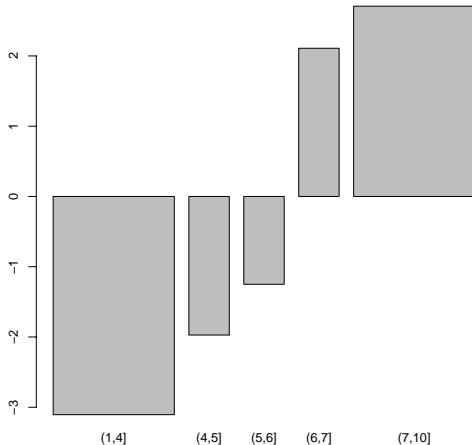
### Table de contingence

Note globale	Promesse d'achat		Promesse d'achat		Somme
	Non	Oui	Non	Oui	
$\leq 4$	18	1	0.95	0.05	1
5	13	4	0.76	0.24	1
6	19	15	0.56	0.44	1
7	5	34	0.13	0.87	1
$\geq 8$	1	33	0.03	0.97	1
Distance du $\chi^2 = 68.63$ , Probabilité critique = $4.41e-14$					



# Première approche

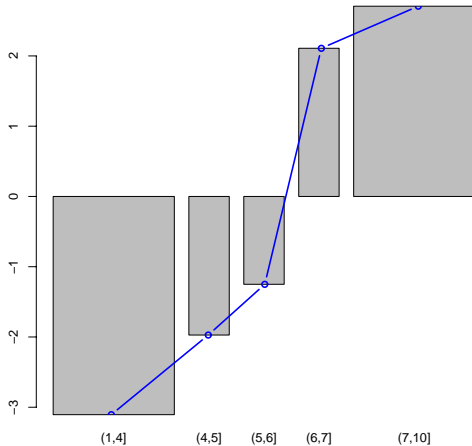
Contributions à la distance du  $\chi^2$





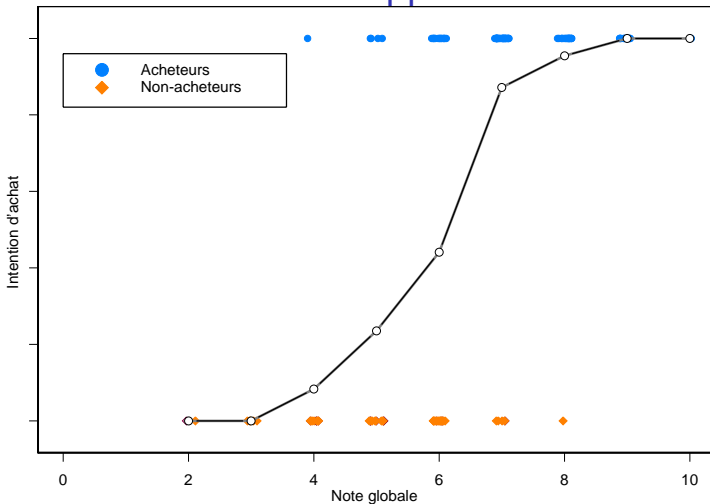
# Première approche

Contributions à la distance du  $\chi^2$





# Première approche

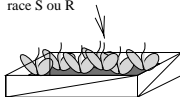




# Efficacité d'un traitement fongicide

mildiou

race S ou R



7 plantules

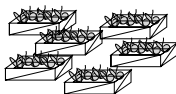
pour chaque plantule :

sporulation



RACE sensible

RACE résistante

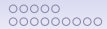


différentes doses

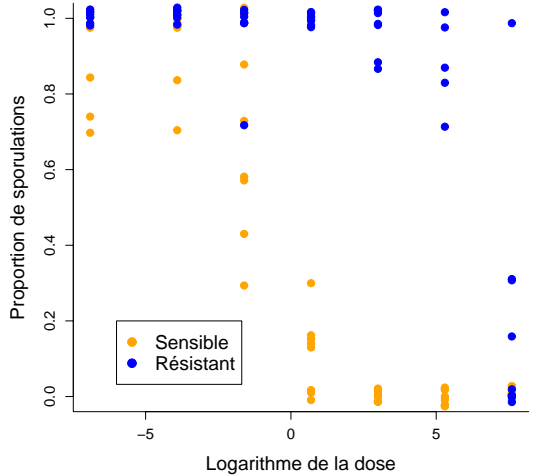


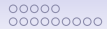
## Efficacité d'un traitement fongicide

Groupe	Race	Dose	Sporu- -lations	Prop. de sporulations	Nombre de plantules
1	S	0,001	5	0.71	7
2	S	0,001	6	0.86	7
3	S	0,001	7	1.00	7
4	S	0,001	7	1.00	7
5	S	0,001	7	1.00	7
6	S	0,001	5	0.71	7
7	S	0,02	7	1.00	7
8	S	0,02	7	1.00	7
⋮	⋮	⋮	⋮	⋮	

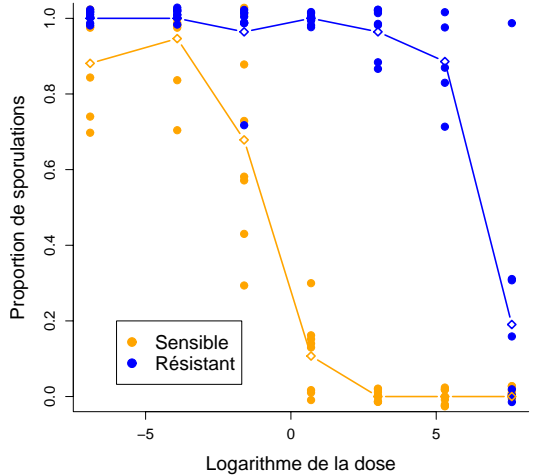


# Efficacité d'un traitement fongicide





# Efficacité d'un traitement fongicide





# Plan du cours

- 1 Quelques situations concrètes
  - Modélisation de l'intention d'achat
  - Modélisation de l'efficacité d'un fongicide
- 2 **Modèle pour variables qualitatives à deux modalités**
  - Modèle logistique
  - Prédicteur linéaire et fonction de lien
- 3 Estimation des paramètres
  - Maximum de vraisemblance
  - Déviance résiduelle
  - Propriétés
- 4 Perspectives



## Y qualitative binaire

**Y qualitative à 2 modalités (0 et 1) :**

$$\mathbb{P}(Y = 1) = \pi_1,$$

$$\mathbb{P}(Y = 0) = \pi_0.$$

**Pour  $x$  fixé :**  $Y \sim \text{Bernoulli}[\pi(x)]$

$$\mathbb{P}_x(Y = 1) = \pi(x),$$

$$\mathbb{E}_x(Y) = \pi(x), \quad \text{Var}_x(Y) = \pi(x)[1 - \pi(x)].$$



## Y qualitative binaire

**Y qualitative à 2 modalités (0 et 1) :**

$\mathbb{P}(Y = 1) = \pi_1$ , Exemple : risque de sporulation

$\mathbb{P}(Y = 0) = \pi_0$ . Exemple : proba. de non – sporulation

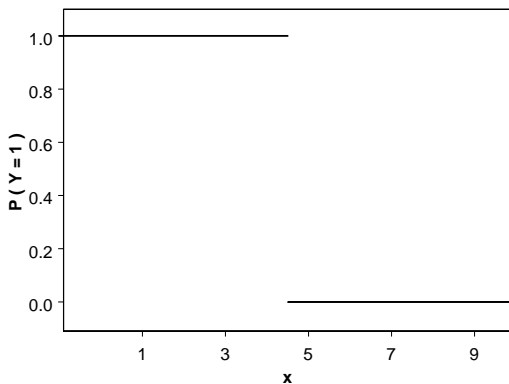
**Pour  $x$  fixé :**  $Y \sim \text{Bernoulli}[\pi(x)]$

$\mathbb{P}_x(Y = 1) = \pi(x)$ , Exemple : risque de sporulation en  
fonction du log – dose de fongicide

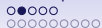
$$\mathbb{E}_x(Y) = \pi(x), \quad \text{Var}_x(Y) = \pi(x)[1 - \pi(x)].$$



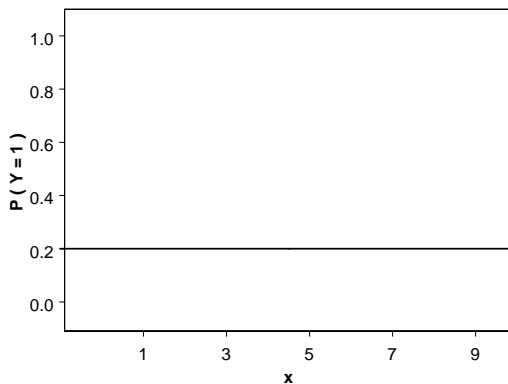
## Prédicteur en forme de «S»



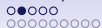
$$\mathbb{P}_x(Y = 1) = \begin{cases} 1 & \text{si } x < x_0 \\ 0 & \text{si } x \geq x_0 \end{cases}$$



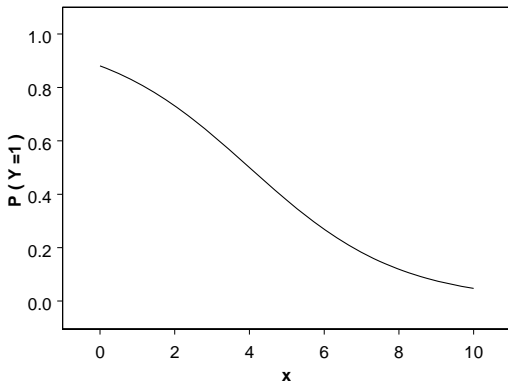
## Prédicteur en forme de «S»



$$\mathbb{P}_x(Y = 1) = \pi$$



## Prédicteur en forme de «S»



$$\mathbb{P}_x(Y = 1) = \pi(x)$$



## Notion de risque relatif

### Risque relatif

$$\begin{aligned}\frac{\mathbb{P}_x(Y = 1)}{\mathbb{P}_x(Y = 0)} &= \frac{\pi(x)}{1 - \pi(x)}, \\ &= \text{odds}(x),\end{aligned}$$

Si  $f(x; \mu_1, \sigma_1)$  est la densité de  $X$  si  $Y = 1$ .

$$\text{odds}(x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x)},$$

où  $f_1(x)$  est la densité de  $X$  si  $Y = 1$ .



## Modélisation du risque relatif

Si  $f_1(x) = f(x; \mu_1, \sigma)$  est la densité de la loi  $\mathcal{N}(\mu_1, \sigma)$ .

$$\text{odds}(x) = \frac{\pi_1}{\pi_0} \exp \left[ \left( \frac{\mu_1 - \mu_0}{\sigma^2} \right) x - \left( \frac{\mu_1^2 - \mu_0^2}{2\sigma^2} \right) \right].$$

$$\log [\text{odds}(x)] = \log \left[ \frac{\pi_1}{\pi_0} \right] + \left( \frac{\mu_1 - \mu_0}{\sigma^2} \right) x - \left( \frac{\mu_1^2 - \mu_0^2}{2\sigma^2} \right),$$

$$\text{logodds}(x) = \beta_0 + \beta_1 x$$

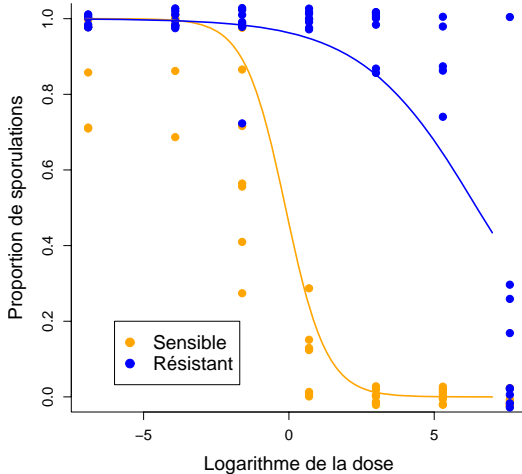


## Modélisation de l'efficacité d'un traitement fongicide

Paramètres	Mildiou sensible		Mildiou résistant	
	Sporulation	Non-sporulation	Sporulation	Non-sporulation
$\pi$	0.35	0.65	0.84	0.16
$\mu$	-3.88	3.48	-0.74	6.98
$\sigma$		2.99		3.92
$\beta_0$		-0.16		3.26
$\beta_1$		-1.42		-0.50



# Modélisation de l'efficacité d'un traitement fongicide





## Prédicteur du modèle

### Cas du modèle de régression linéaire multiple

$$\begin{aligned}\mathbb{E}_x(Y) &= \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)}, \\ &= L(x; \beta)\end{aligned}$$



## Prédicteur du modèle

### Cas du modèle de régression linéaire multiple

$$\begin{aligned}\mathbb{E}_x(Y) &= \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)}, \\ &= L(x; \beta)\end{aligned}$$

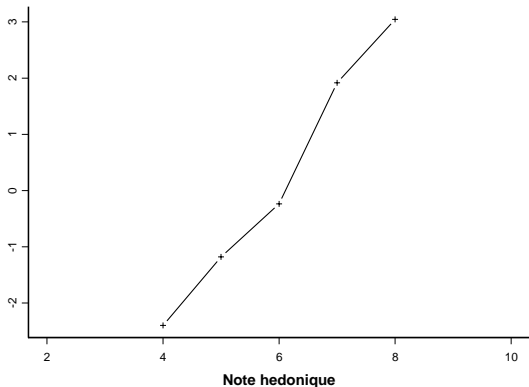
### Cas où $Y$ est binaire, $\mathbb{E}_x(Y) = \pi(x)$

$$\begin{aligned}\log \left[ \frac{\pi(x)}{1 - \pi(x)} \right] &= \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)}, \\ \text{logit} [\pi(x)] &= L(x; \beta).\end{aligned}$$



# Estimation empirique dans le modèle de l'intention d'achat

Logodds de la promesse d'achat



$$\beta_0 \approx -8, \beta_1 \approx 1.5$$

# Fonction de lien et prédicteur linéaire

## Prédicteur du modèle

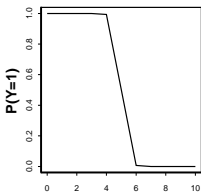
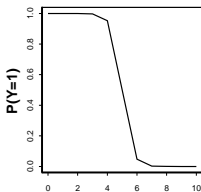
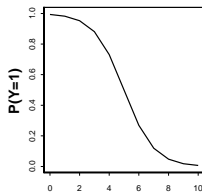
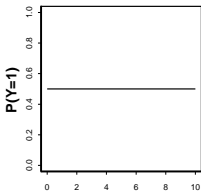
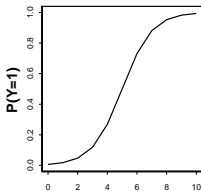
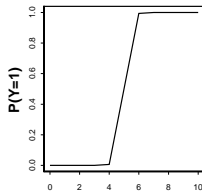
$$g[\pi(x)] = L(x; \beta)$$

où

- $g$  est la **fonction de lien**
  - **Modèle logistique** :  $g = \text{logit}$
  - **Modèle probit** :  $g = F^{-1}$ , où  $F$  est la fonction de répartition de la loi  $\mathcal{N}(0; 1)$ .
- $L$  est le prédicteur linéaire



# Interprétation des paramètres

 $\beta_1 = -5$ 

 $\beta_1 = -3$ 

 $\beta_1 = -1$ 

 $\beta_1 = 0$ 

 $\beta_1 = 1$ 

 $\beta_1 = 5$ 


# Odds ratio

## Evolution du risque entre $x$ et $x'$

$$\text{OR}(x'; x) = \frac{\text{odds}(x')}{\text{odds}(x)}.$$

## Odds ratio

### Evolution du risque entre $x$ et $x'$

$$\text{OR}(x'; x) = \frac{\text{odds}(x')}{\text{odds}(x)}.$$

Si  $x^{(j)'} = x^{(j)} + 1$

$$\text{OR}_j = \exp(\beta_j).$$

**Exemple** : OR pour la log-dose

**0.24** (pour mildiou sensible) et **0.60** (pour mildiou résistant)

## Essai « gras coloré »

**Problématique** : étudier l'effet de l'alimentation sur la coloration du gras d'agneaux

**Dispositif expérimental** : 80 agneaux

- 20 agneaux, non-nourris en continu, logement collectif
- 20 agneaux, non-nourris en continu, logement individuel
- 20 agneaux, nourris en continu, logement collectif
- 20 agneaux, nourris en continu, logement individuel



## Données « gras coloré »

Alimentation	Logement	Nombre d'agneaux à gras coloré	Effectif total
$A_1$	$L_1$ (collectif)	10	$n_{11} = 20$
$A_1$	$L_2$	12	$n_{12} = 20$
$A_2$ (continu)	$L_1$ (collectif)	15	$n_{21} = 20$
$A_2$ (continu)	$L_2$	16	$n_{22} = 20$



## Modèle « gras coloré »

Pour une combinaison  $(A_i, L_j)$  donnée,  $Y_{ijk}$  désigne la coloration, ou non, du gras du  $k$ ème agneau

$$Y_{ijk} = \begin{cases} O & \text{recodé } 1 \\ N & \text{recodé } 0 \end{cases}$$

$Z_{ij}$ , nombre d'agneaux à gras coloré :

$$\begin{aligned} Z_{ij} &= Y_{ij,1} + Y_{ij,2} + \dots + Y_{ij,20} \\ Z_{ij} &\sim \mathcal{B}(20; \pi_{ij}) \end{aligned}$$

## Risque de coloration du gras d'agneau

Pour un mode d'alimentation  $A_i$  fixé,  $Z_i =$  nombre d'agneaux à gras coloré

$$Z_i \sim \mathcal{B}(40; \pi_i);$$
$$\text{logit}(\pi_i) = \mu + \alpha_i,$$

avec  $\alpha_1 = 0$ .



## Risque de coloration du gras d'agneau

Pour un mode d'alimentation  $A_i$  fixé,  $Z_i =$  nombre d'agneaux à gras coloré

$$Z_i \sim \mathcal{B}(40; \pi_i);$$
$$\text{logit}(\pi_j) = \mu + \alpha_j,$$

avec  $\alpha_1 = 0$ .

### Risque relatif de coloration du gras

$$\text{OR}(A_2; A_1) = \frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}} = \exp(\alpha_2).$$



# Plan du cours

- 1 Quelques situations concrètes
  - Modélisation de l'intention d'achat
  - Modélisation de l'efficacité d'un fongicide
- 2 Modèle pour variables qualitatives à deux modalités
  - Modèle logistique
  - Prédicteur linéaire et fonction de lien
- 3 Estimation des paramètres
  - Maximum de vraisemblance
  - Déviance résiduelle
  - Propriétés
- 4 Perspectives



## Vraisemblance d'un modèle

### Exemple : modèle logit, $x$ quantitative

$(x_1 = 6, y_1 = 1)$ ,  $(x_2 = 7, y_2 = 0)$ ,  $(x_3 = 5, y_3 = 0)$ ,  $(x_4 = 9, y_4 = 1)$ , ...



## Vraisemblance d'un modèle

### Exemple : modèle logit, $x$ quantitative

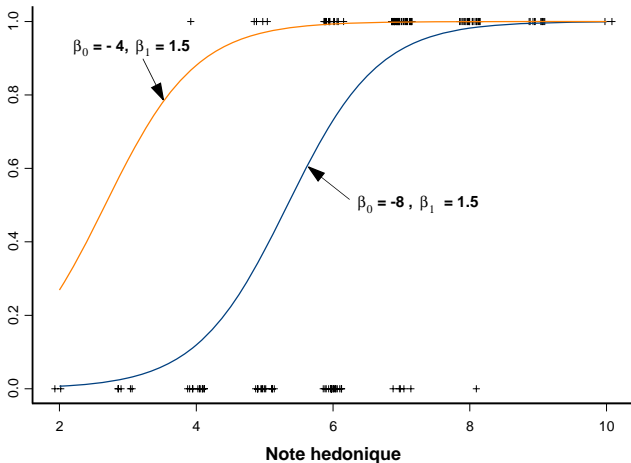
$(x_1 = 6, y_1 = 1)$ ,  $(x_2 = 7, y_2 = 0)$ ,  $(x_3 = 5, y_3 = 0)$ ,  $(x_4 = 9, y_4 = 1)$ , ...

$$\begin{aligned}\mathcal{V}(\beta) &= \mathbb{P}_x(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) \\ &= \mathbb{P}_{x_1}(Y_1 = y_1)\mathbb{P}_{x_2}(Y_2 = y_2) \dots \mathbb{P}_{x_n}(Y_n = y_n).\end{aligned}$$



# Vraisemblance d'un modèle

Promesse d'achat





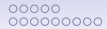
## Vraisemblance d'un modèle

### Exemple : modèle logit, $x$ quantitative

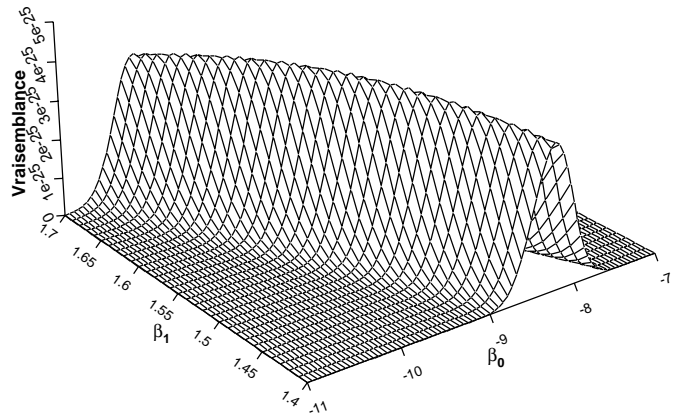
$(x_1 = 6, y_1 = 1), (x_2 = 7, y_2 = 0), (x_3 = 5, y_3 = 0), (x_4 = 9, y_4 = 1), \dots$

$$\begin{aligned}\mathcal{V}(\beta) &= \mathbb{P}_x(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) \\ &= \mathbb{P}_{x_1}(Y_1 = y_1)\mathbb{P}_{x_2}(Y_2 = y_2) \dots \mathbb{P}_{x_n}(Y_n = y_n).\end{aligned}$$

- $\mathcal{V}(-8, 1.5) = 2.88 \cdot 10^{-28}$ ,  $\log \mathcal{V}(-8, 1.5) = -63.41$
- $\log \mathcal{V}(-4, 1.5) = -210.70$



# Maximum de vraisemblance



## Qualité d'ajustement

**Un indicateur naturel** : la vraisemblance

$$\mathcal{V}(\hat{\beta}) = \hat{\mathbb{P}}(Y_1 = y_1)\hat{\mathbb{P}}(Y_2 = y_2)\dots\hat{\mathbb{P}}(Y_n = y_n),$$

Si la  $i$ ème donnée est mal ajustée,

- $\hat{\mathbb{P}}(Y_i = y_i)$  est faible
- elle contribue à diminuer la vraisemblance



## Contributions individuelles

**Un autre indicateur** : la déviance résiduelle

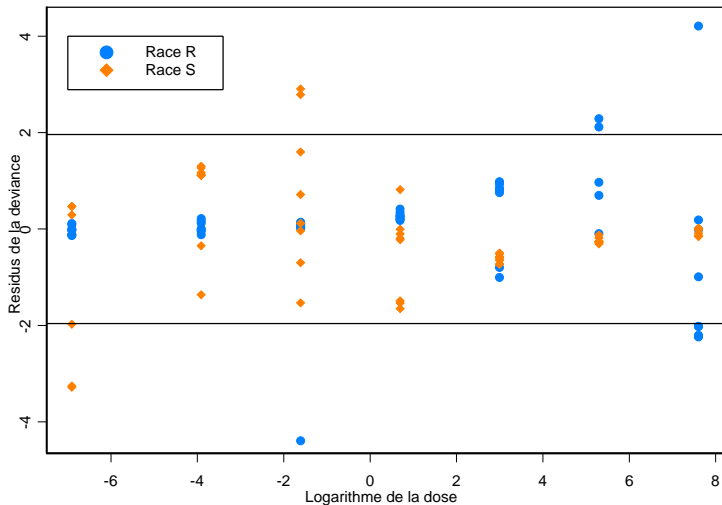
$$\begin{aligned}
 \mathcal{V}(\hat{\beta}) &= \hat{\mathbb{P}}(Y_1 = y_1) \hat{\mathbb{P}}(Y_2 = y_2) \dots \hat{\mathbb{P}}(Y_n = y_n), \\
 \underbrace{-2 \log \mathcal{V}(\hat{\beta})}_{\text{Déviance résiduelle}} &= \underbrace{-2 \log \hat{\mathbb{P}}(Y_1 = y_1)}_{d_1^2} - \underbrace{-2 \log \hat{\mathbb{P}}(Y_2 = y_2)}_{d_2^2} - \\
 &\quad \dots - \underbrace{-2 \log \hat{\mathbb{P}}(Y_n = y_n)}_{d_n^2}
 \end{aligned}$$

Si la  $i$ ème donnée est mal ajustée,  $d_i^2$  est grand

$$\pm d_i \sim \mathcal{N}(0; 1), \text{ approximativement}$$



## Résidus de la déviance



## Modèle de l'intention d'achat

---



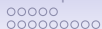
---

Coefficients:

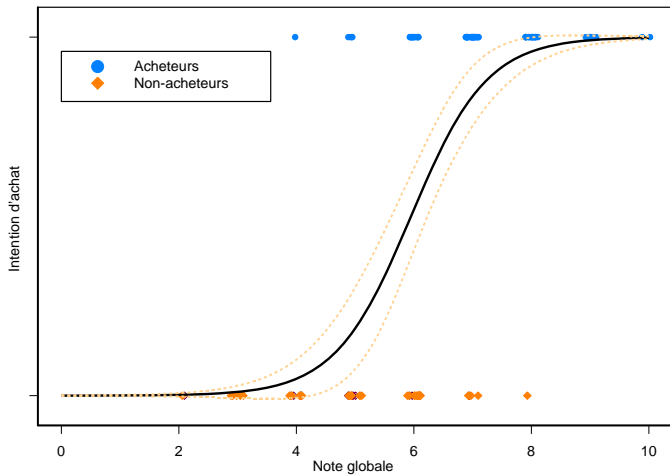
	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-9.188	1.659	-5.537	3.08e-08	***
globale	1.544	0.265	5.825	5.72e-09	***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring Iterations: 6



## Modèle de l'intention d'achat





# Plan du cours

- 1 Quelques situations concrètes
  - Modélisation de l'intention d'achat
  - Modélisation de l'efficacité d'un fongicide
- 2 Modèle pour variables qualitatives à deux modalités
  - Modèle logistique
  - Prédicteur linéaire et fonction de lien
- 3 Estimation des paramètres
  - Maximum de vraisemblance
  - Déviance résiduelle
  - Propriétés
- 4 Perspectives



# Perspectives

- Test des effets



## Perspectives

- Test des effets
- Construction de règles de discrimination



## Perspectives

- Test des effets
- Construction de règles de discrimination
- Validation d'un modèle de diagnostic