Modèle de discrimination logistique

David Causeur
Laboratoire de Mathématiques Appliquées
Agrocampus Rennes
IRMAR CNRS UMR 6625
http://www.agrocampus-rennes.fr/math/causeur/

Plan du cours

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	$n_{22} = 20$

Plan du cours

Test global

Modèle complet contre modèle nul

$$\begin{cases} H_0 & : & x \text{ n'a pas d'influence sur } Y \\ H_1 & : & x \text{ a une influence sur } Y \end{cases}$$

Test global

Modèle complet contre modèle nul

$$\begin{cases} H_0 & : & x \text{ n'a pas d'influence sur } Y \\ H_1 & : & x \text{ a une influence sur } Y \end{cases}$$

Essai « Gras coloré »

$$\begin{cases} H_0 & : & \text{le mode d'alimentation n'a pas} \\ & & \text{d'influence sur la coloration du gras} \\ H_1 & : & \text{le mode d'alimentation a} \\ & & \text{une influence sur la coloration du gras} \end{cases}$$

Modèles de coloration du gras d'agneau

Modèle complet : $\text{logit}(\pi_1) = \mu,\ \text{logit}(\pi_2) = \mu + \alpha_2$

Modèles de coloration du gras d'agneau

Modèle complet : $\text{logit}(\pi_1) = \mu,\ \text{logit}(\pi_2) = \mu + \alpha_2$

Modèle nul : $\text{logit}(\pi_1) = \text{logit}(\pi_2) = \mu$

Modèles de coloration du gras d'agneau

Modèle complet : $\text{logit}(\pi_1) = \mu,\ \text{logit}(\pi_2) = \mu + \alpha_2$

Modèle nul : $\text{logit}(\pi_1) = \text{logit}(\pi_2) = \mu$

Modèle complet contre modèle nul

$$\begin{cases} H_0 & : & \alpha_2 = 0 \\ H_1 & : & \alpha_2 \neq 0 \end{cases}$$

Ajustement du modèle nul

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{Z_1+Z_2}{80} = 0.66$$

Ajustement du modèle nul

$\hat{\pi}_1 = \hat{\pi}_2 = \frac{Z_1+Z_2}{80} = 0.66$

Qualité de l'ajustement :

$$\begin{aligned} \mathcal{V}_0 &= \widehat{\mathbb{P}}(Z_1 = 22)\widehat{\mathbb{P}}(Z_2 = 31) \\ &= C_{40}^{22}\hat{\pi}_1^{22}(1-\hat{\pi}_1)^{40-22}C_{40}^{31}\hat{\pi}_2^{31}(1-\hat{\pi}_2)^{40-31} \\ &= 0.0019 \end{aligned}$$

Ajustement du modèle complet

$\hat{\pi}_1 = 0.55,\ \hat{\pi}_2 = 0.775$

Ajustement du modèle complet

$\hat{\pi}_1 = 0.55,\ \hat{\pi}_2 = 0.775$

Qualité de l'ajustement :

$$\begin{aligned} \mathcal{V} &= \widehat{\mathbb{P}}(Z_1 = 22)\widehat{\mathbb{P}}(Z_2 = 31) \\ &= C_{40}^{22}\hat{\pi}_1^{22}(1-\hat{\pi}_1)^{40-22}C_{40}^{31}\hat{\pi}_2^{31}(1-\hat{\pi}_2)^{40-31} \\ &= 0.019 \end{aligned}$$

Déviance du modèle

Rapport des vraisemblances

$\mathcal{V}$ vraisemblance du modèle complet :

$$\text{RV} = \frac{\mathcal{V}_0}{\mathcal{V}}.$$

Déviance du modèle

Rapport des vraisemblances

$\mathcal{V}$ vraisemblance du modèle complet :

$$\text{RV} \quad = \quad \frac{\mathcal{V}_0}{\mathcal{V}}.$$

Sous l'hypothèse d'absence d'effet de x

$$\underbrace{-2 \log \text{RV}}_{\text{déviance}} \quad \sim \quad \chi^2_{p-1},$$

où p est le nombre de paramètres du modèle complet.

Analyse de la déviance

$$\nu_0 = \frac{\nu_0}{\nu}\nu,$$

Analyse de la déviance

$$\begin{aligned} \mathcal{V}_0 &= \frac{\mathcal{V}_0}{\mathcal{V}}\mathcal{V}, \\ \underbrace{-2\log \mathcal{V}_0}_{\mathcal{D}_0} &= \underbrace{-2\log \mathrm{RV}}_{\mathcal{D}} \underbrace{-2\log \mathcal{V}}_{\mathcal{D}_r}, \end{aligned}$$

Analyse de la déviance

$$
\begin{aligned}
\mathcal{V}_0 &= \frac{\mathcal{V}_0}{\mathcal{V}}\mathcal{V}, \\
\underbrace{-2\log \mathcal{V}_0}_{\mathcal{D}_0} &= \underbrace{-2\log \mathrm{RV}}_{\mathcal{D}}\underbrace{-2\log \mathcal{V}}_{\mathcal{D}_r}, \\
\underbrace{\mathcal{D}_0}_{n-1 \text{ ddl}} &= \underbrace{\mathcal{D}}_{p-1 \text{ ddl}} + \underbrace{\mathcal{D}_r}_{n-p \text{ ddl}},
\end{aligned}
$$

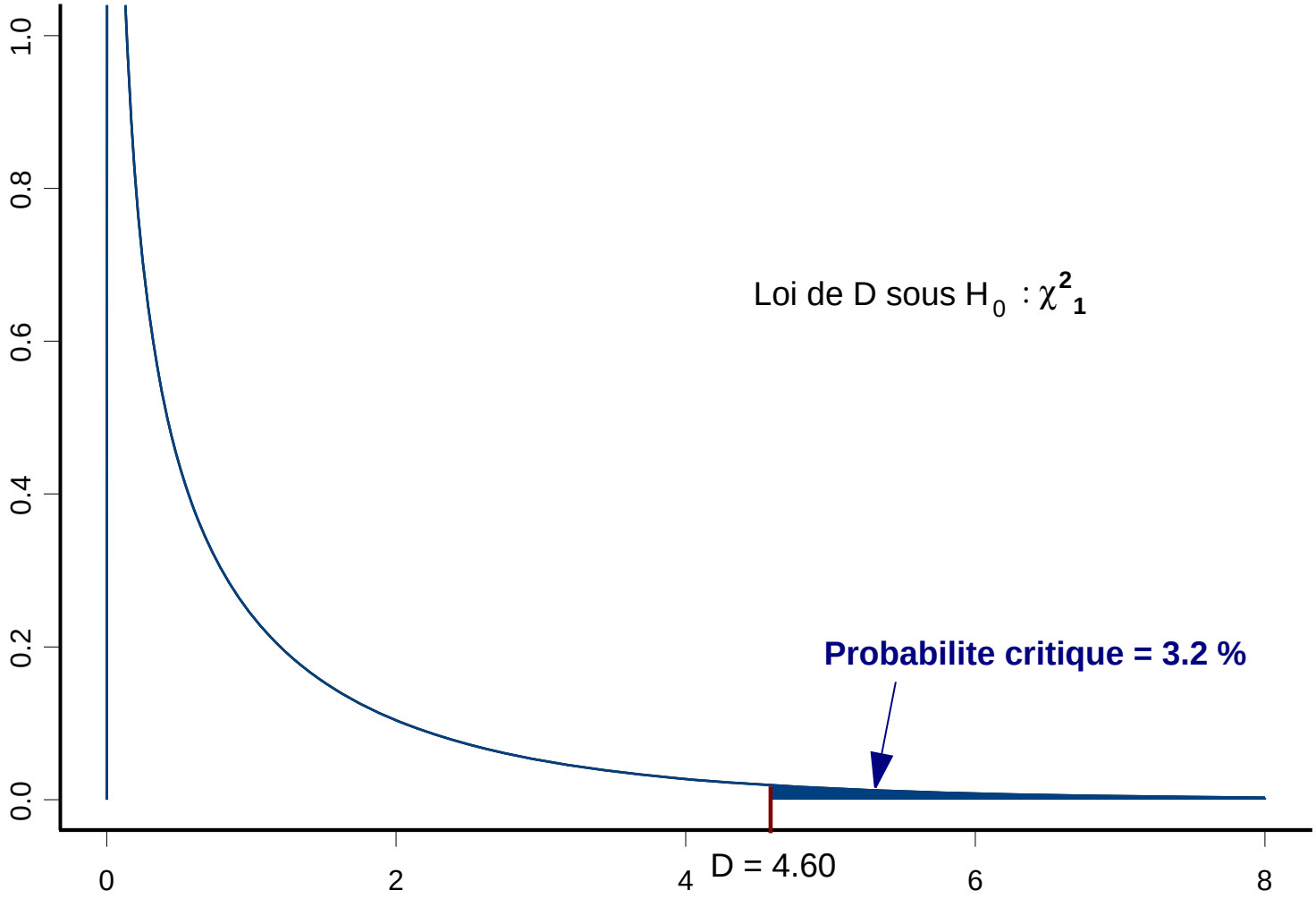
Analyse de la déviance

$$\begin{aligned} \mathcal{V}_0 &= \frac{\mathcal{V}_0}{\mathcal{V}}\mathcal{V}, \\ \underbrace{-2\log\mathcal{V}_0}_{\mathcal{D}_0} &= \underbrace{-2\log \mathrm{RV}}_{\mathcal{D}}\underbrace{-2\log\mathcal{V}}_{\mathcal{D}_r}, \\ \underbrace{\mathcal{D}_0}_{n-1 \text{ ddl}} &= \underbrace{\mathcal{D}}_{p-1 \text{ ddl}} + \underbrace{\mathcal{D}_r}_{n-p \text{ ddl}}, \end{aligned}$$

Essai « Gras coloré »

$$\underbrace{\mathcal{D}_0}_{12.54} = \underbrace{\mathcal{D}}_{4.60(1 \text{ ddl})} + \underbrace{\mathcal{D}_r}_{7.94}$$

Analyse de la déviance



Efficacité d'un fongicide

$\pi_i(x)$, probabilité qu'un plant de tournesol exposé au mildiou de race *i* et traité avec la log-dose *x* soit contaminé

Modèle complet

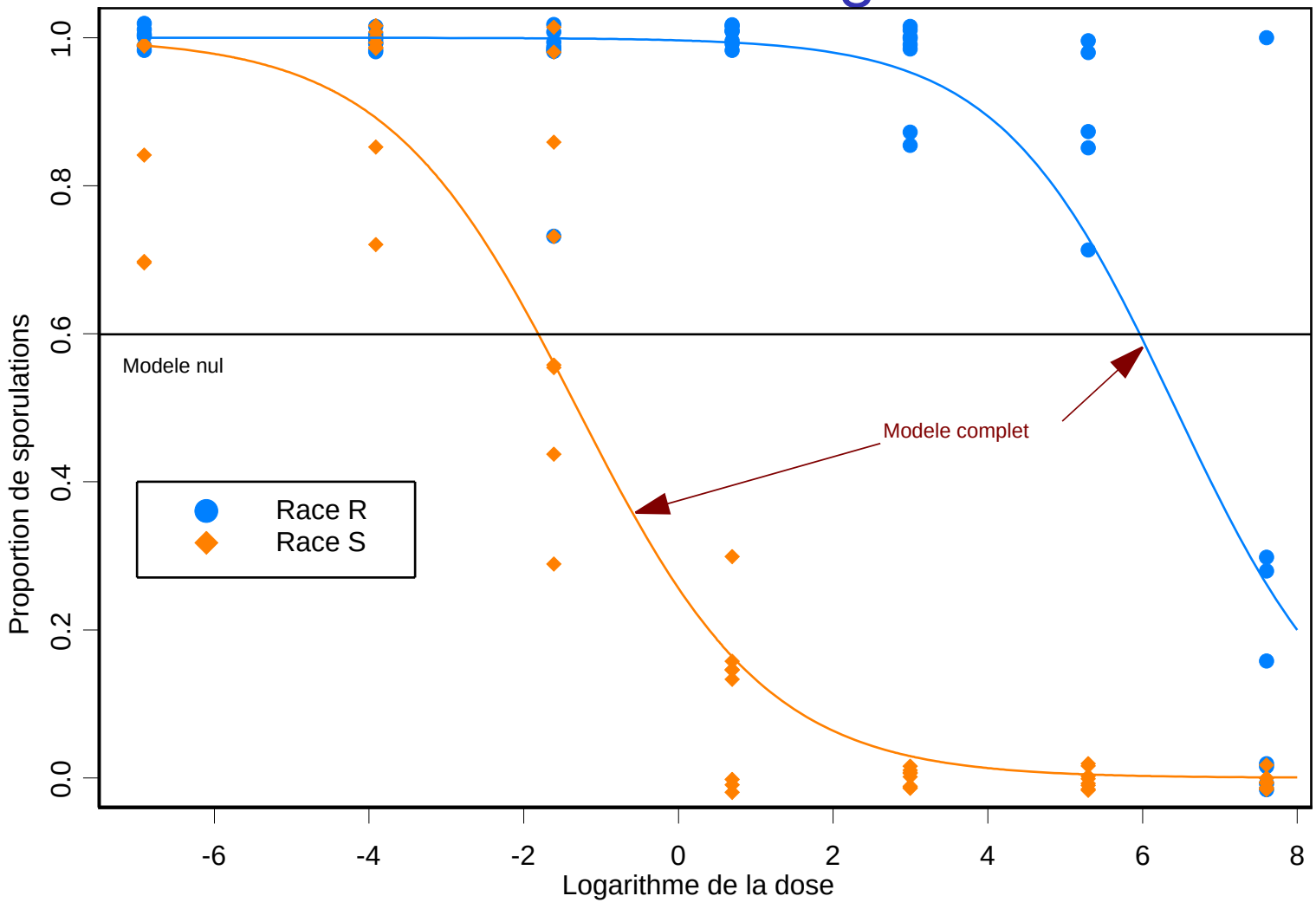
$$\text{logit}\left[\pi_i(x)\right] \quad = \quad \mu + \alpha_i + \left[\beta + \delta_i\right] x$$

où $\alpha_2 = \delta_2 = 0$.

Modèle nul

$$\text{logit}\left[\pi_i(x)\right] \quad = \quad \mu.$$

Efficacité d'un fongicide



Test global de l'efficacité d'un fongicide

Analyse de la déviance

	Déviance	Degré de liberté	Probabilité critique
Modèle nul	836.739	107	
Modèle complet	689.422	3	0
Résiduelle	147.316	104	

Comparaison à un sous-modèle

Modèle complet contre sous modèle

$$\begin{cases} H_0 & : & M_p \text{ est aussi intéressant que } M_q \ (q < p) \\ H_1 & : & M_p \text{ est plus intéressant que } M_q \end{cases}$$

Test de l'effet d'interaction Race $\times$ Dose

Modèle complet

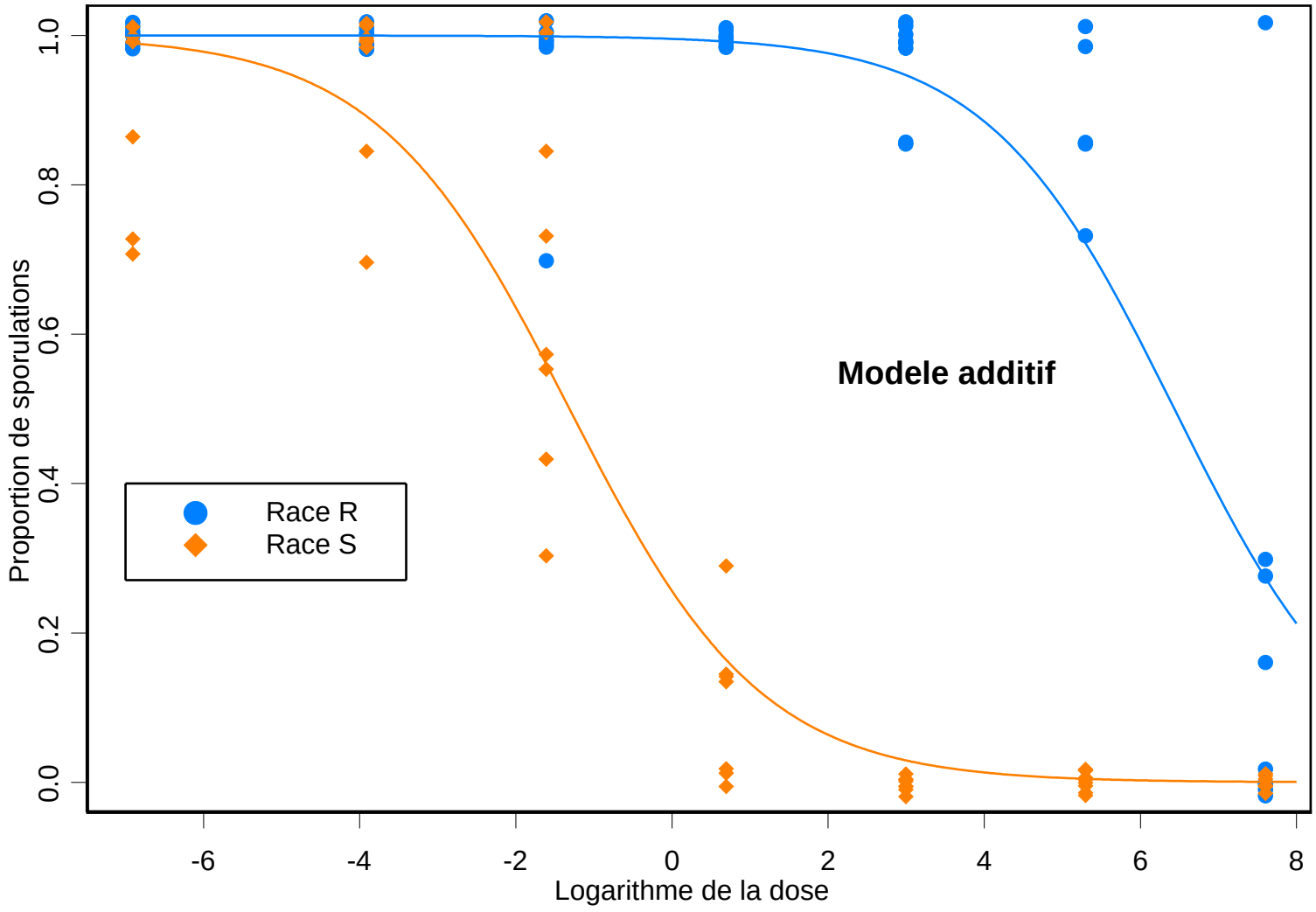
$$\text{logit}\left[\pi_i(x)\right] \quad = \quad \mu + \alpha_i + \left[\beta + \delta_i\right] x$$

où $\alpha_2 = \delta_2 = 0$.

Modèle sans interaction

$$\text{logit}\left[\pi_i(x)\right] \quad = \quad \mu + \alpha_i + \beta x.$$

Test de l'effet d'interaction Race × Dose



Rapport des vraisemblances

$\mathcal{V}_p$ vraisemblance du modèle complet, $\mathcal{V}_q$ pour le sous-modèle :

$$\mathrm{RV}_{p|q} = \frac{\mathcal{V}_q}{\mathcal{V}_p}.$$

Sous l'hypothèse d'absence d'effet de x

$$\underbrace{-2 \log \mathrm{RV}_{p|q}}_{\text{déviance}} \sim \chi^2_{p-q}.$$

Analyse de la déviance

$$\nu_0 = \frac{\nu_0}{\nu_q}\frac{\nu_q}{\nu_p}\nu_p,$$

Analyse de la déviance

$$\begin{aligned} \mathcal{V}_0 &= \frac{\mathcal{V}_0}{\mathcal{V}_q}\frac{\mathcal{V}_q}{\mathcal{V}_p}\mathcal{V}_p, \\ \underbrace{-2\log \mathcal{V}_0}_{\mathcal{D}_0} &= \underbrace{-2\log \mathrm{RV}_q}_{\mathcal{D}_q}\ \underbrace{-2\log \mathrm{RV}_{p|q}}_{\mathcal{D}_{p|q}}\ \underbrace{-2\log \mathcal{V}_p}_{\mathcal{D}_r}, \end{aligned}$$

Analyse de la déviance

$$
\begin{aligned}
\mathcal{V}_0 &= \frac{\mathcal{V}_0}{\mathcal{V}_q}\frac{\mathcal{V}_q}{\mathcal{V}_p}\mathcal{V}_p, \\
\underbrace{-2\log \mathcal{V}_0}_{\mathcal{D}_0} &= \underbrace{-2\log \mathrm{RV}_q}_{\mathcal{D}_q}\underbrace{-2\log \mathrm{RV}_{p|q}}_{\mathcal{D}_{p|q}}\underbrace{-2\log \mathcal{V}_p}_{\mathcal{D}_r}, \\
\underbrace{\mathcal{D}_0}_{n-1 \text{ ddl}} &= \underbrace{\mathcal{D}_q}_{q-1 \text{ ddl}} + \underbrace{\mathcal{D}_{p|q}}_{p-q \text{ ddl}} + \underbrace{\mathcal{D}_r}_{n-p \text{ ddl}}
\end{aligned}
$$

Test de l'efficacité d'un fongicide

Analyse de la déviance

	Déviance	Degré de liberté	Probabilité critique
Modèle nul	836.739	107	
Modèle sans interaction	689.21	2	0
Modèle avec interaction	0.21	1	0.64
Résiduelle	147.316	104	

Tests des effets «alimentation» et «logement» sur la coloration du gras d'agneau

$$\begin{cases} H_0 : \text{Seul le mode d'alimentation a une influence sur la coloration du gras} \\ H_1 : \text{les modes d'alimentation et de logement ont tous deux une influence sur la coloration du gras} \end{cases}$$

Comparaison de modèles

π_{ij} probabilité de coloration du gras avec l'alimentation *i* et le logement *j*

Modèle complet : $\text{logit}(\pi_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$
avec $\alpha_1 = \beta_1 = \gamma_{11} = \gamma_{12} = \gamma_{21} = 0$

Modèle sous H_0 : $\text{logit}(\pi_{ij}) = \mu + \alpha_i$

Modèle complet contre modèle nul

$$\begin{cases} H_0 & : & \beta_2 = 0 \\ H_1 & : & \beta_2 \neq 0 \end{cases}$$

Ajustement du modèle sous H_0

$\hat{\pi}_{11} = 0.55,\ \hat{\pi}_{12} = 0.55,\ \hat{\pi}_{21} = 0.775,\ \hat{\pi}_{22} = 0.775$

Qualité de l'ajustement :

$$\begin{aligned}
\mathcal{V} &= \widehat{\mathbb{P}}(Z_{11} = 10)\widehat{\mathbb{P}}(Z_{12} = 12)\widehat{\mathbb{P}}(Z_{21} = 15)\widehat{\mathbb{P}}(Z_{22} = 16) \\
&= C_{20}^{10}\hat{\pi}_1^{10}(1 - \hat{\pi}_{11})^{20-10}C_{20}^{1}\hat{\pi}_{12}^{12}(1 - \hat{\pi}_{12})^{20-12} \ldots \\
&= 0.0011
\end{aligned}$$

Ajustement du modèle complet

$\hat{\pi}_{11} = 0.5,\ \hat{\pi}_{12} = 0.6,\ \hat{\pi}_{21} = 0.75,\ \hat{\pi}_{22} = 0.8$

Qualité de l'ajustement :

$$\begin{aligned}
\mathcal{V} &= \widehat{\mathbb{P}}(Z_{11} = 10)\widehat{\mathbb{P}}(Z_{12} = 12)\widehat{\mathbb{P}}(Z_{21} = 15)\widehat{\mathbb{P}}(Z_{22} = 16) \\
&= C_{20}^{10}\hat{\pi}_{11}^{10}(1-\hat{\pi}_{11})^{20-10}C_{20}^{12}\hat{\pi}_{12}^{12}(1-\hat{\pi}_{12})^{20-12}\ldots \\
&= 0.0014
\end{aligned}$$

Rapport des vraisemblances

$\mathcal{V}_4$ vraisemblance du modèle complet, $\mathcal{V}_2$ pour le sous-modèle :

$$\mathrm{RV}_{4|2} = \frac{\mathcal{V}_2}{\mathcal{V}_4}.$$

Sous l'hypothèse d'absence H_0

$$\underbrace{-2 \log \mathrm{RV}_{4|2}}_{0.548} \sim \chi^2_2.$$

$$\text{Probabilité critique} = 0.76$$

Plan du cours

Descripteurs les plus impliqués dans l'intention d'achat

Variable	Définition
$x^{(1)}$	Impression visuelle
$x^{(2)}$	Appétance
$x^{(3)}$	Qualité du pain en bouche
$x^{(4)}$	Originalité du pain
$x^{(5)}$	Quantité de garniture
$x^{(6)}$	Qualité de la garniture en bouche
$x^{(7)}$	Originalité de la garniture
$x^{(8)}$	Pertinence de l'association des ingrédients de la garniture

Sélection pas à pas

Variable	1 x	2 x dont $x^{(6)}$	3 x dont $x^{(6)}$ et $x^{(3)}$	4 x dont $x^{(6)}$, $x^{(3)}$ et x
$x^{(1)}$	1.7e-2	*1.4e-1*	*4.2e-1*	*5.6e-1*
$x^{(2)}$	1.2e-4	4.0e-2	*2.8e-1*	*3.7e-1*
$x^{(3)}$	4.9e-6	**1.0e-3**		
$x^{(4)}$	*7.8e-1*	*9.6e-1*	*2.7e-1*	*2.0e-1*
$x^{(5)}$	*6.1e-1*	*8.8e-1*	*8.6e-1*	*8.2e-1*
$x^{(6)}$	**5.4e-10**			
$x^{(7)}$	7.7e-3	4.6e-1	*3.3e-1*	*6.9e-1*
$x^{(8)}$	1.8e-7	1.1e-2	**1.7e-2**	

Sélection selon le critère d'Akaike

Start: AIC= 147.76
achat ∼ attrait + appetance + painbouche + painorig + quantite + garnbouche + garnorig + association

	Df	Deviance	AIC
- appetance	1	129.77	145.77
- quantite	1	130.09	146.09
- attrait	1	130.89	146.89
<none>		129.76	147.76
- garnorig	1	132.24	148.24
- association	1	133.38	149.38
- painorig	1	134.40	150.40
- painbouche	1	137.40	153.40
- garnbouche	1	139.74	155.74

Sélection selon le critère d'Akaike

Step: AIC= 145.77
achat ∼ attrait + painbouche + painorig + quantite + garnbouche + garnorig + association

	Df	Deviance	AIC
- quantite	1	130.17	144.17
- attrait	1	131.44	145.44
<none>		129.77	145.77
- garnorig	1	132.37	146.37
- association	1	133.43	147.43
+ appetance	1	129.76	147.76
- painorig	1	134.72	148.72
- painbouche	1	138.00	152.00
- garnbouche	1	140.31	154.31

Sélection selon le critère d'Akaike

Step: AIC= 144.17
achat ∼ attrait + painbouche + painorig + garnbouche + garnorig + association

	Df	Deviance	AIC
- attrait	1	131.73	143.73
<none>		130.17	144.17
- garnorig	1	132.44	144.44
- association	1	133.63	145.63
+ quantite	1	129.77	145.77
+ appetance	1	130.09	146.09
- painorig	1	134.78	146.78
- painbouche	1	138.89	150.89
- garnbouche	1	140.51	152.51

Sélection selon le critère d'Akaike

Step: AIC= 143.73
achat ∼ painbouche + painorig + garnbouche + garnorig + association

	Df	Deviance	AIC
<none>		131.73	143.73
- garnorig	1	133.78	143.78
+ attrait	1	130.17	144.17
+ appetance	1	131.01	145.01
- painorig	1	135.31	145.31
+ quantite	1	131.44	145.44
- association	1	135.66	145.66
- painbouche	1	141.64	151.64
- garnbouche	1	142.97	152.97

Sélection selon le critère d'Akaike

Anova Table (Type III tests)

Response: achat

	LR Chisq	Df	Pr(>Chisq)	
painbouche	9.9057	1	0.0016476	**
painorig	3.5794	1	0.0585007	.
garnbouche	11.2399	1	0.0008006	***
garnorig	2.0503	1	0.1521783	
association	3.9306	1	0.0474142	*

Plan du cours

Discrimination

Exemple : mortalité du porcelet

On connaît le poids de naissance et le gain de poids à 24 h d'un porcelet.

Quel diagnostic ?

Discriminante logistique

- Estimation de la probabilité de survie $\hat{\pi}$ du porcelet
- Si $\hat{\pi}$ dépasse un seuil, on diagnostique la survie

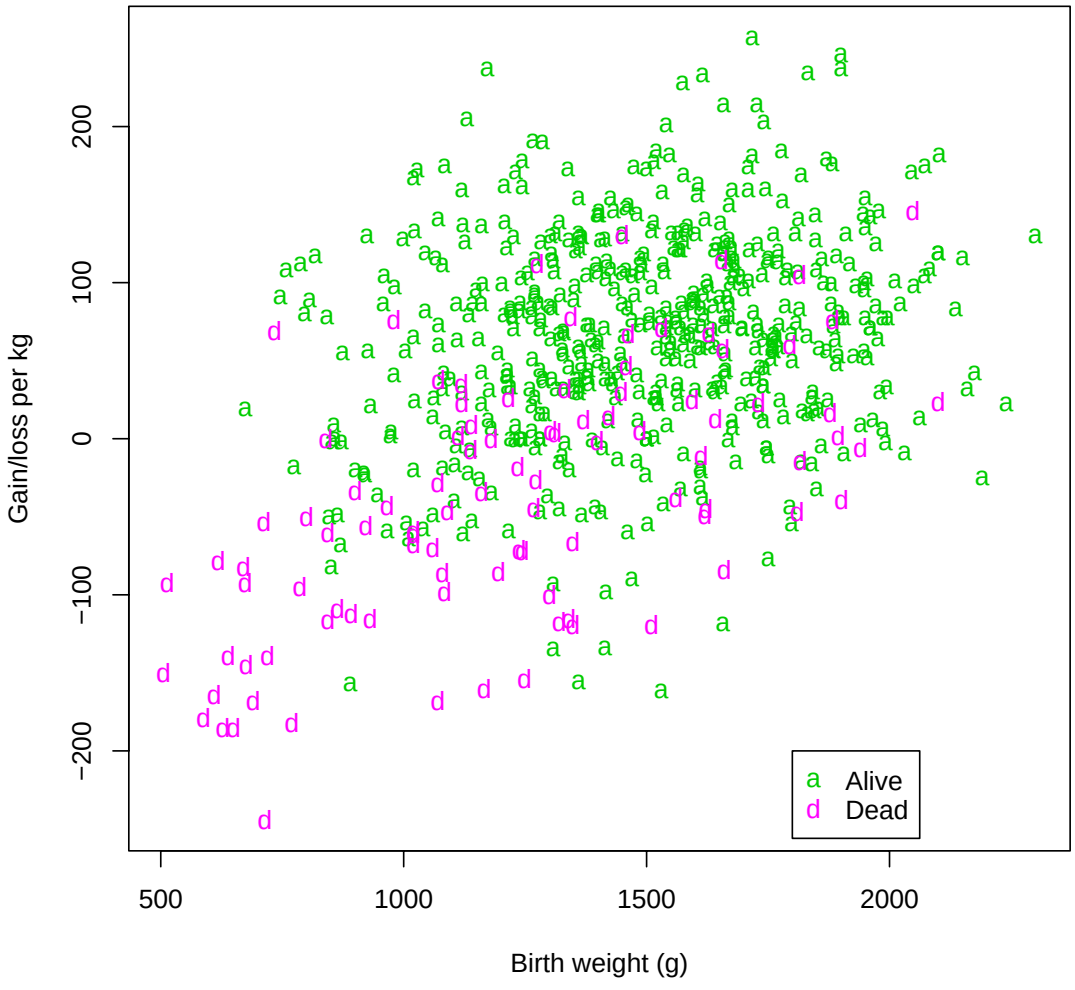
Validation sur la base du

- taux de diagnostics corrects de la survie (spécificité)
- taux de diagnostics corrects de la mort (sensibilité)

Mortalité du porcelet

Y : état du porcelet au sevrage (1=«Vivant», 0=«Mort»)
$x^{(1)}$: poids à la naissance
$x^{(2)}$: gain de poids (pour 1 kg) à 24h

Mortalité du porcelet



Mortalité du porcelet

Y : état du porcelet au sevrage (1=«Vivant», 0=«Mort»)
$x^{(1)}$: poids à la naissance
$x^{(2)}$: gain de poids (pour 1 kg) à 24h

Ajustement sur 526 porcelets vivants & 97 porcelets morts

Analyse de la déviance			
	Déviance	Degré de liberté	Probabilité critique
Modèle nul	538.85	622	
Poids de naissance	54.16	1	1.85e-13
Gain de poids à 24h	98.08	1	4.03e-23
Résiduelle	386.62	620	

Fonction discriminante logistique

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.94e-01	5.69e-01	0.34	0.73	
Poids de naissance	8.55e-04	4.16e-04	2.06	0.04	*
Gain de poids à 24h	17.41	2.07	8.41	$< 2e-16$	***

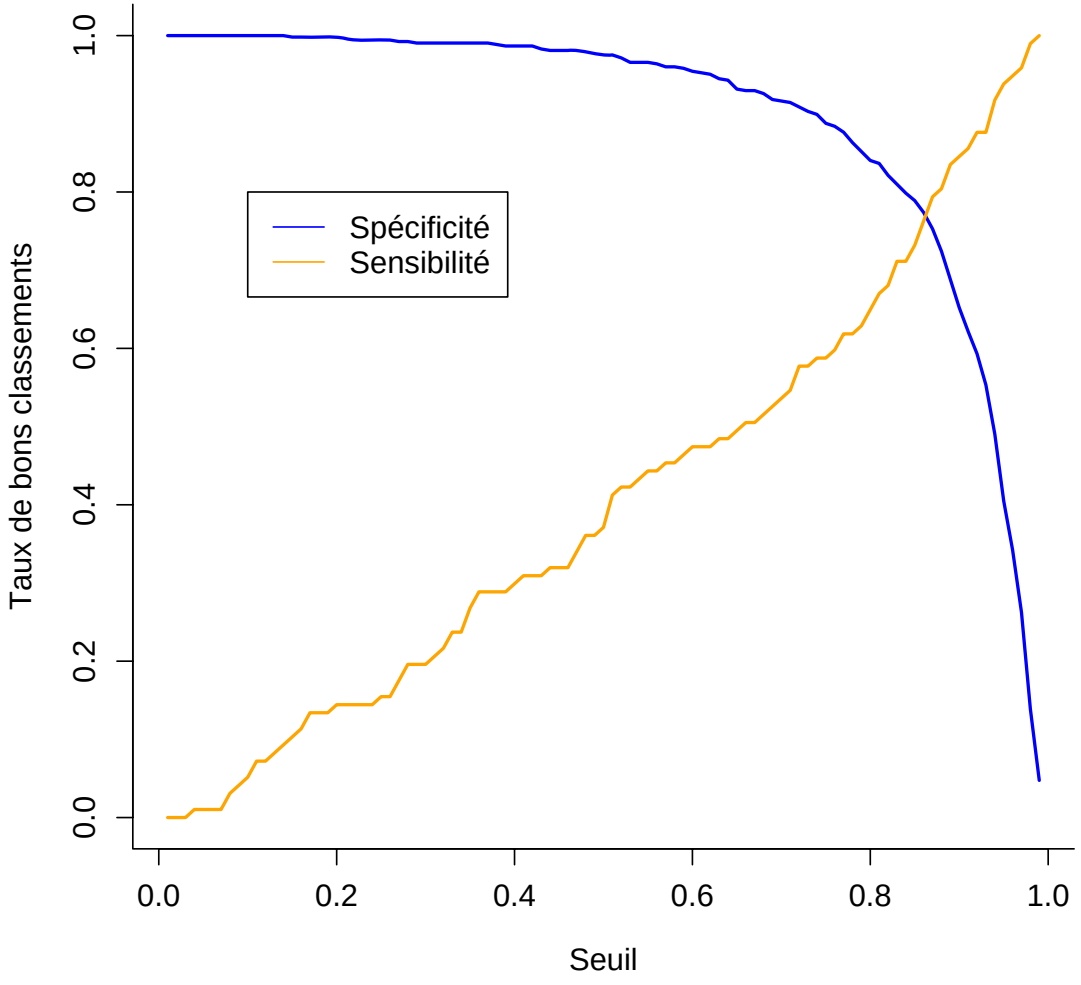
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring Iterations: 6

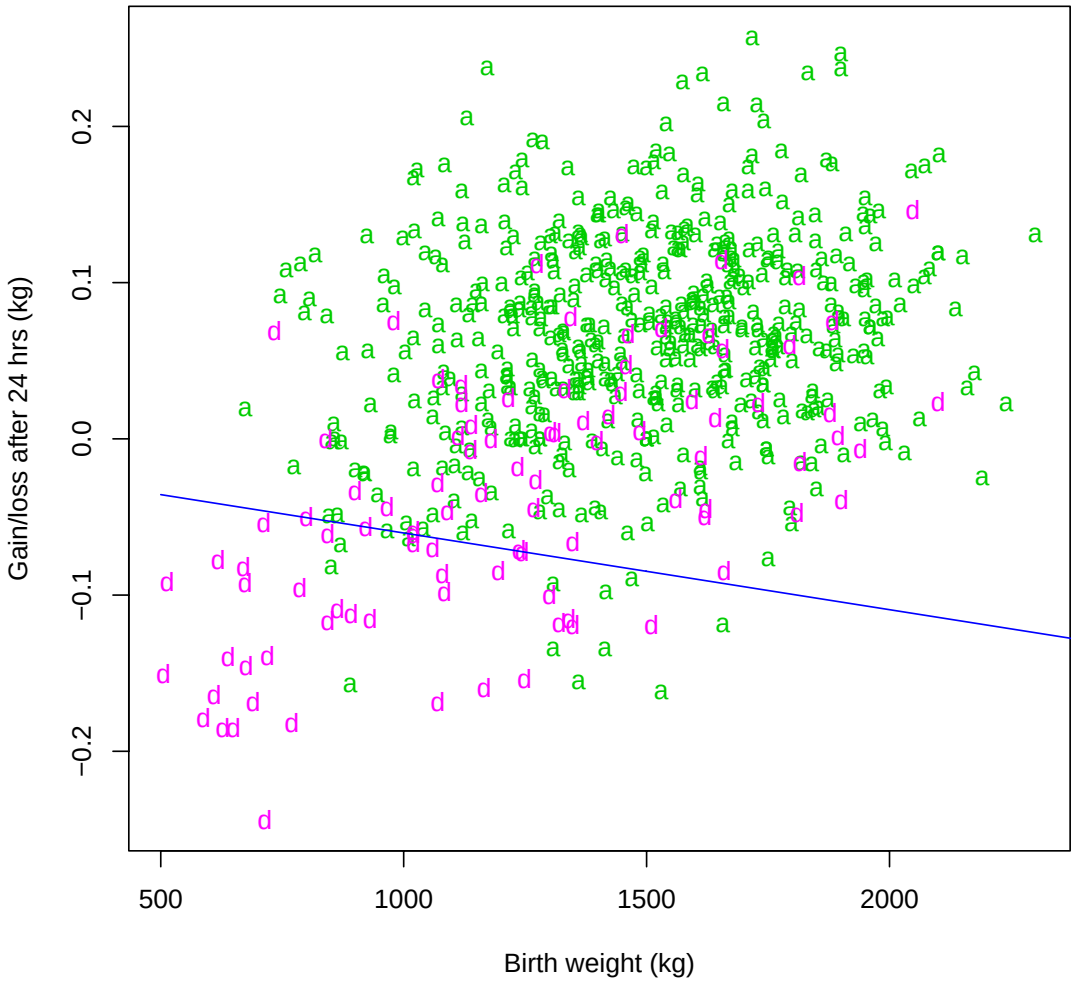
Règle de discrimination :

Si $\hat{\pi}(x) \geq k$, alors $\hat{Y} =$ « Vivant »

Capabilité de discrimination



Règle de discrimination



Validation d'une règle de discrimination

Validation croisée : confrontation de la règle de décision à des données externes

2 étapes:

- Échantillon d'apprentissage (de calibration)
  - Tirage au hasard uniforme dans l'échantillon disponible (2/3)
  - Sert à ajuster la règle de décision
- Échantillon test
  - Complémentaire de l'échantillon d'apprentissage
  - Sert à confronter la règle de décision à la réalité (calcul des taux d'erreurs)

Validation croisée robuste

Pour rendre la procédure moins sensible au découpage «apprentissage-test»:

- Répétition de la validation croisée
  - Taux de bons classements : moyennes des taux d'erreurs à chaque validation croisée
- Si échantillon de taille faible, «leave-one-out»

Mortalité du porcelet

Matrice de confusion - hors validation croisée:

	Observations	
Affectations	Mort	Vivant
Mort	0.76	0.23
Vivant	0.24	0.77

Matrice de confusion - Validation croisée (10×):

	Observations	
Affectations	Mort	Vivant
Mort	0.75	0.22
Vivant	0.25	0.78

Ouverture

- Lien avec la discriminante linéaire de Fisher

Ouverture

- Lien avec la discriminante linéaire de Fisher
- Extension au cas multi-classes

Ouverture

- Lien avec la discriminante linéaire de Fisher
- Extension au cas multi-classes
- Arbres de décision