



# Modèle pour variable multi-classes

## Arbres de décision

David Causeur

*Laboratoire de Mathématiques Appliquées*

*Agrocampus Rennes*

*IRMAR CNRS UMR 6625*

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



# Plan du cours

- 1 Préférence d'une marque
- 2 Modèles pour variables qualitatives
  - Modèles pour variables qualitatives nominales
  - Modèles pour variables qualitatives ordinales
  - Discrimination logistique
- 3 Arbres de classification
  - Règle binaire de classification
  - Arbre récursif binaire
- 4 Bilan



## Choix d'une marque par un consommateur

La marque préférée est-elle la même selon l'âge et le sexe ?

	Gender	Age	Brand
1	M	24	M1
2	M	26	M3
3	M	26	M1
4	F	27	M2
5	F	27	M3
6	M	27	M2
⋮	⋮	⋮	



## Choix d'une marque par un consommateur

La marque préférée est-elle la même selon l'âge et le sexe ?

	Gender	Age	Brand
1	M	24	M1
2	M	26	M3
3	M	26	M1
4	F	27	M2
5	F	27	M3
6	M	27	M2
⋮	⋮	⋮	

Age : variable quantitative



## Choix d'une marque par un consommateur

La marque préférée est-elle la même selon l'âge et le sexe ?

	Gender	Age	Brand
1	M	24	M1
2	M	26	M3
3	M	26	M1
4	F	27	M2
5	F	27	M3
6	M	27	M2
⋮	⋮	⋮	

Age : variable quantitative

Marque : variable qualitative à 3 modalités



## Choix d'une marque par un consommateur

Age	Marque			Total
	M1	M2	M3	
24	1	0	0	1
26	2	0	0	2
27	8	0	1	9
28	10	2	3	15
29	14	5	0	19
30	13	6	4	23
31	19	14	7	40
32	110	168	55	333
33	7	32	16	55
34	10	29	25	64
35	1	8	26	35
36	10	29	46	85
37	1	6	15	22
38	1	8	23	32

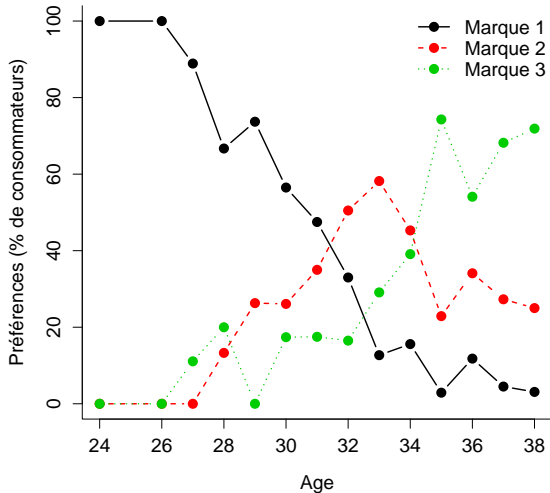


## Choix d'une marque par un consommateur

Age	Marque			Total
	M1	M2	M3	
24	100.0	0.0	0.0	100.0
26	100.0	0.0	0.0	100.0
27	88.9	0.0	11.1	100.0
28	66.7	13.3	20.0	100.0
29	73.7	26.3	0.0	100.0
30	56.5	26.1	17.4	100.0
31	47.5	35.0	17.5	100.0
32	33.0	50.5	16.5	100.0
33	12.7	58.2	29.1	100.0
34	15.6	45.3	39.1	100.0
35	2.9	22.9	74.3	100.1
36	11.8	34.1	54.1	100.0
37	4.5	27.3	68.2	100.0
38	3.1	25.0	71.9	100.0

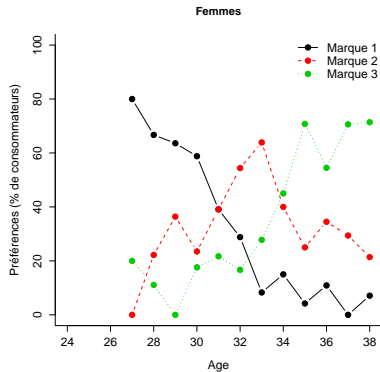
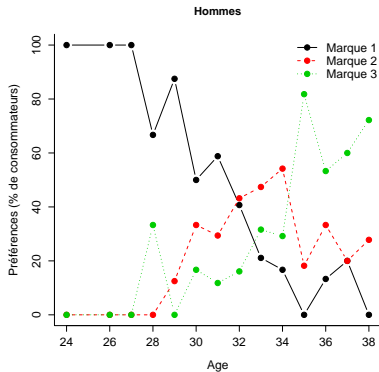


# Choix d'une marque par un consommateur





# Choix d'une marque par un consommateur





# Plan du cours

- 1 Préférence d'une marque
- 2 **Modèles pour variables qualitatives**
  - Modèles pour variables qualitatives nominales
  - Modèles pour variables qualitatives ordinales
  - Discrimination logistique
- 3 Arbres de classification
  - Règle binaire de classification
  - Arbre récursif binaire
- 4 Bilan



## Y qualitative nominale

**Y qualitative à K modalités ( $C_1, C_2, \dots, C_K$ ) :**

$$\mathbb{P}(Y = C_k) = \pi_k,$$

$$\pi_1 + \dots + \pi_K = 1$$

**Pour x fixé :**

$$\mathbb{P}_x(Y = C_k) = \pi_k(x),$$

$$\pi_1(x) + \dots + \pi_K(x) = 1$$



## Y qualitative nominale

**Y qualitative à K modalités ( $C_1, C_2, \dots, C_K$ ):**

$$\begin{aligned} \mathbb{P}(Y = C_k) &= \pi_k, & \text{Exemple : proba. que la marque k} \\ \pi_1 + \dots + \pi_K &= 1 & \text{soit choisie} \end{aligned}$$

**Pour x fixé :**

$$\begin{aligned} \mathbb{P}_x(Y = C_k) &= \pi_k(x), & \text{Exemple : proba. que la marque k} \\ \pi_1(x) + \dots + \pi_K(x) &= 1 & \text{soit choisie à l'âge x} \end{aligned}$$



## Modélisation logistique multinomiale

$Y \in \{C_1, C_2, \dots, C_K\}$  **Modèle logistique multinomial :**

$$\begin{aligned}\log [\pi_2(x)/\pi_1(x)] &= L(x; \beta^{(2)}), \\ \log [\pi_3(x)/\pi_1(x)] &= L(x; \beta^{(3)}), \\ &\vdots \\ \log [\pi_K(x)/\pi_1(x)] &= L(x; \beta^{(K)}).\end{aligned}$$



## Modélisation logistique multinomiale

$Y \in \{C_1, C_2, \dots, C_K\}$  **Modèle logistique multinomial :**

$$\log [\pi_2(x)/\pi_1(x)] = L(x; \beta^{(2)}),$$

$$\log [\pi_3(x)/\pi_1(x)] = L(x; \beta^{(3)}),$$

$$\vdots \quad \quad \quad \vdots$$

$$\log [\pi_K(x)/\pi_1(x)] = L(x; \beta^{(K)}).$$

**Exemple :** Préférence de la marque

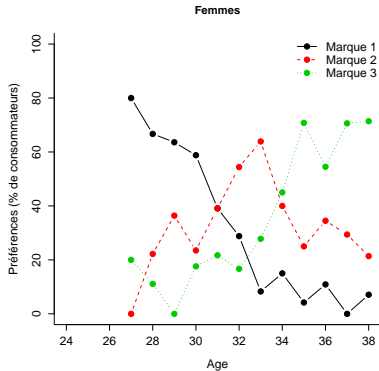
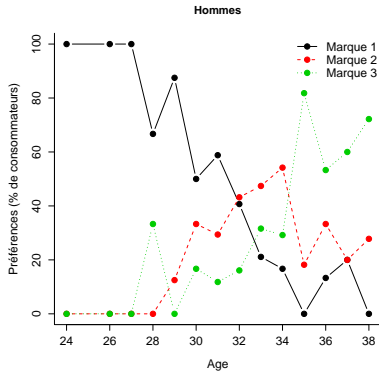
$$\log [\pi_2(x)/\pi_1(x)] = \mu^{(2)} + \alpha_i^{(2)} + [\beta^{(2)} + \gamma_i^{(2)}] x,$$

$$\log [\pi_3(x)/\pi_1(x)] = \mu^{(3)} + \alpha_i^{(3)} + [\beta^{(3)} + \gamma_i^{(3)}] x.$$

Soit 8 paramètres

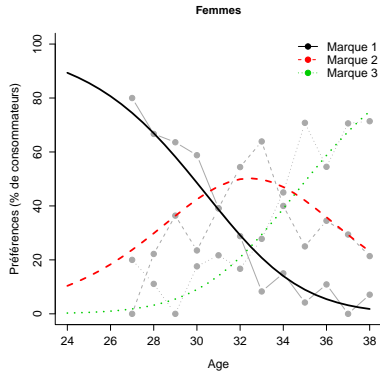
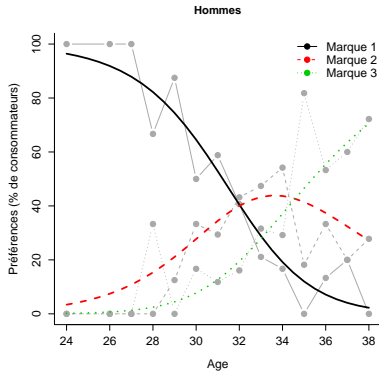


# Modélisation logistique multinomiale





# Modélisation logistique multinomiale





# Estimation par la méthode du maximum de vraisemblance

Pour chaque  $\beta$ ,

$$\begin{aligned} \pi_1(x) &= \frac{1}{1 + \exp [L(x; \beta^{(2)})] + \dots + \exp [L(x; \beta^{(K)})]}, \\ \pi_2(x) &= \frac{\exp L(x; \beta^{(2)})}{1 + \exp [L(x; \beta^{(2)})] + \dots + \exp [L(x; \beta^{(K)})]}, \\ \pi_3(x) &= \frac{\exp L(x; \beta^{(3)})}{1 + \exp [L(x; \beta^{(2)})] + \dots + \exp [L(x; \beta^{(K)})]}, \\ &\vdots \\ \pi_K(x) &= \frac{\exp L(x; \beta^{(K)})}{1 + \exp [L(x; \beta^{(2)})] + \dots + \exp [L(x; \beta^{(K)})]}. \end{aligned}$$

$$\mathcal{V}(\beta) = \mathbb{P}_{x_1}(Y = y_1) \mathbb{P}_{x_2}(Y = y_2) \dots \mathbb{P}_{x_n}(Y = y_n)$$



# Estimation par la méthode du maximum de vraisemblance

Pour chaque  $\beta$ , pour les données *marque*

$$\pi_1(x) = \frac{1}{1 + \exp [L(x; \beta^{(2)})] + \exp [L(x; \beta^{(3)})]},$$

$$\pi_2(x) = \frac{\exp L(x; \beta^{(2)})}{1 + \exp [L(x; \beta^{(2)})] + \exp [L(x; \beta^{(3)})]},$$

$$\pi_3(x) = \frac{\exp L(x; \beta^{(3)})}{1 + \exp [L(x; \beta^{(2)})] + \exp [L(x; \beta^{(3)})]}.$$

$$\mathcal{V}(\beta) = \mathbb{P}_{x_1=24}(Y = M_1) \mathbb{P}_{x_2=26}(Y = M_3) \dots \mathbb{P}_{x_n=38}(Y = M_3)$$



## Estimation par la méthode du maximum de vraisemblance

Coefficients				
	(Intercept)	Age	Gender[T.M]	Age:Gender[T.M]
M2	-10.23018	0.3364944	-3.1265847	0.08055698
M3	-22.00349	0.6772655	-0.9954803	0.01858241
Écart-types				
	(Intercept)	Age	Gender[T.M]	Age:Gender[T.M]
M2	2.263182	0.07058379	3.656095	0.1136942
M3	2.629088	0.08062280	4.205734	0.1286519
Residual deviance: 1405.142				
AIC: 1421.142				



## Estimation par la méthode du maximum de vraisemblance

Coefficients				
	(Intercept)	Age	Gender[T.M]	Age:Gender[T.M]
M2	-10.23018	0.3364944	-3.1265847	0.08055698
M3	-22.00349	0.6772655	-0.9954803	0.01858241
Écart-types				
	(Intercept)	Age	Gender[T.M]	Age:Gender[T.M]
M2	2.263182	0.07058379	3.656095	0.1136942
M3	2.629088	0.08062280	4.205734	0.1286519
Residual deviance: 1405.142				
AIC: 1421.142				



## Analyse de la déviance

Analysis of deviance							
	Model	Resid. df	Resid. dev	Test	Df	LR stat	Pr (Chi)
1	1	1468	1591.792				
2	Age	1466	1413.593	1 vs 2	2	178.1989779	0.00000000
3	Age + Gender	1464	1405.941	2 vs 3	2	7.6512358	0.02180496
4	Age * Gender	1462	1405.142	3 vs 4	2	0.7996223	0.67044664



## Analyse de la déviance

Analysis of deviance							
	Model	Resid. df	Resid. dev	Test	Df	LR stat	Pr (Chi)
1	1	1468	1591.792				
2	Age	1466	1413.593	1 vs 2	2	178.1989779	0.00000000
3	Age + Gender	1464	1405.941	2 vs 3	2	7.6512358	0.02180496
4	Age * Gender	1462	1405.142	3 vs 4	2	0.7996223	0.67044664



## Analyse de la déviance

### Maximum de vraisemblance

Coefficients			
	(Intercept)	Age	Gender[T.M]
M2	-11.25127	0.3682202	-0.5237736
M3	-22.25571	0.6859149	-0.4658215
Écart-types			
	(Intercept)	Age	Gender[T.M]
M2	1.765184	0.05500373	0.1942473
M3	2.042636	0.06262721	0.2260901
Residual deviance: 1405.941			
AIC: 1417.941			



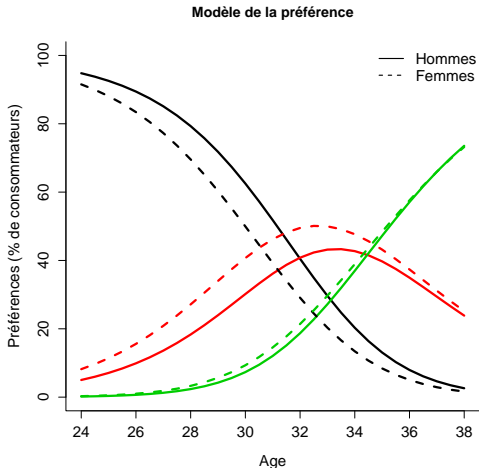
# Analyse de la déviance

## Maximum de vraisemblance

Coefficients			
	(Intercept)	Age	Gender[T.M]
M2	-11.25127	0.3682202	-0.5237736
M3	-22.25571	0.6859149	-0.4658215
Écart-types			
	(Intercept)	Age	Gender[T.M]
M2	1.765184	0.05500373	0.1942473
M3	2.042636	0.06262721	0.2260901
Residual deviance: 1405.941			
AIC: 1417.941			



# Analyse de la déviance





## Modèle pour la pression systolique

La pression sanguine dépend-t'elle du poids ?

	Gender	Weight	Sbp
1	F	71.66759	2
2	F	82.10022	2
3	M	78.01789	2
4	M	107.50139	3
5	F	77.11070	2
6	M	74.84274	1
⋮	⋮	⋮	



## Modèle pour la pression systolique

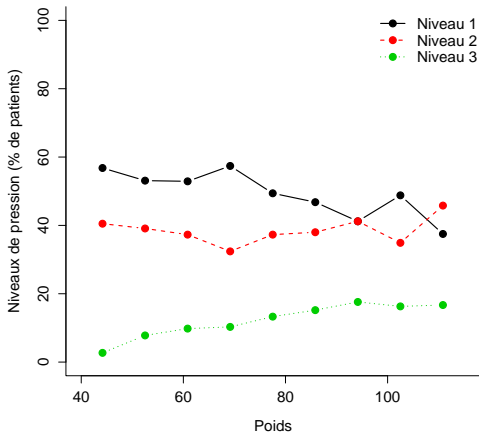
La pression sanguine dépend-t'elle du poids ?

Weights	Marque			Total
	1	2	3	
(40,48.3]	56.8	40.5	2.7	100.0
(48.3,56.7]	53.1	39.1	7.8	100.0
(56.7,65]	52.9	37.3	9.8	100.0
(65,73.3]	57.4	32.4	10.3	100.1
(73.3,81.7]	49.4	37.3	13.3	100.0
(81.7,90]	46.8	38.0	15.2	100.0
(90,98.3]	41.2	41.2	17.6	100.0
(98.3,107]	48.8	34.9	16.3	100.0
(107,115]	37.5	45.8	16.7	100.0



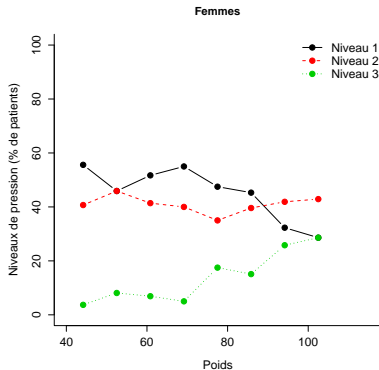
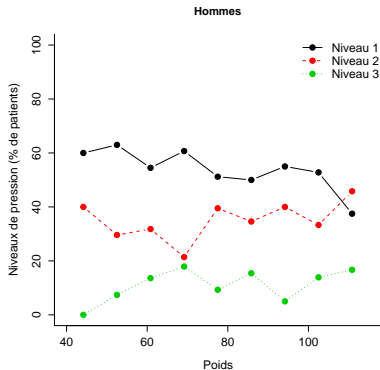
# Modèle pour la pression systolique

La pression sanguine dépend-t'elle du poids ?





# Modèle pour la pression systolique





## Y qualitative ordinale

$$Y \in \{C_1 \leq C_2 \leq \dots \leq C_K\}$$

### 3 types de modèles du risque :

- Risques adjacents : relatifs à la modalité suivante.

$$\log \left[ \frac{\pi_k(x)}{\pi_{k+1}(x)} \right] = L(x; \beta^{(k)})$$



## Y qualitative ordinale

$$Y \in \{C_1 \leq C_2 \leq \dots \leq C_K\}$$

### 3 types de modèles du risque :

- Risques adjacents : relatifs à la modalité suivante.

$$\log \left[ \frac{\pi_k(x)}{\pi_{k+1}(x)} \right] = L(x; \beta^{(k)})$$

- Risques séquentiels : relatifs aux modalités supérieures.

$$\log \left[ \frac{\pi_k(x)}{\pi_{k+1}(x) + \dots + \pi_K(x)} \right] = L(x; \beta^{(k)})$$



## Y qualitative ordinale

$$Y \in \{C_1 \leq C_2 \leq \dots \leq C_K\}$$

### 3 types de modèles du risque :

- Risques adjacents : relatifs à la modalité suivante.

$$\log \left[ \frac{\pi_k(x)}{\pi_{k+1}(x)} \right] = L(x; \beta^{(k)})$$

- Risques séquentiels : relatifs aux modalités supérieures.

$$\log \left[ \frac{\pi_k(x)}{\pi_{k+1}(x) + \dots + \pi_K(x)} \right] = L(x; \beta^{(k)})$$

- Risques cumulatifs : le risque que la variable à expliquer soit inférieure à une modalité est calculé par rapport aux modalités supérieures.

$$\log \left[ \frac{\pi_1(x) + \dots + \pi_k(x)}{\pi_{k+1}(x) + \dots + \pi_K(x)} \right] = L(x; \beta^{(k)})$$



## Modèle à risques cumulatifs

$K - 1$  modèles logit :

$$\log \left[ \frac{\pi_1(\mathbf{X})}{\pi_2(\mathbf{X}) + \dots + \pi_K(\mathbf{X})} \right] = L(\mathbf{X}; \beta^{(1)}),$$

$$\log \left[ \frac{\pi_1(\mathbf{X}) + \pi_2(\mathbf{X})}{\pi_3(\mathbf{X}) + \dots + \pi_K(\mathbf{X})} \right] = L(\mathbf{X}; \beta^{(2)}),$$

⋮

$$\log \left[ \frac{\pi_1(\mathbf{X}) + \dots + \pi_{K-1}(\mathbf{X})}{\pi_K(\mathbf{X})} \right] = L(\mathbf{X}; \beta^{(K-1)}).$$



## Modèle à risques cumulatifs

$K - 1$  modèles logit :

$$\begin{aligned} \log \left[ \frac{\pi_1(\mathbf{X})}{\pi_2(\mathbf{X}) + \dots + \pi_K(\mathbf{X})} \right] &= L(\mathbf{x}; \beta^{(1)}), \\ \log \left[ \frac{\pi_1(\mathbf{X}) + \pi_2(\mathbf{X})}{\pi_3(\mathbf{X}) + \dots + \pi_K(\mathbf{X})} \right] &= L(\mathbf{x}; \beta^{(2)}), \\ &\vdots \\ \log \left[ \frac{\pi_1(\mathbf{X}) + \dots + \pi_{K-1}(\mathbf{X})}{\pi_K(\mathbf{X})} \right] &= L(\mathbf{x}; \beta^{(K-1)}). \end{aligned}$$

Données *Pression sanguine*

$$\begin{aligned} \log \left[ \frac{\pi_{1i}(\mathbf{X})}{\pi_{2i}(\mathbf{X}) + \pi_{3i}(\mathbf{X})} \right] &= \mu^{(1)} + \alpha_i^{(1)} + [\beta^{(1)} + \gamma_i^{(1)}] \mathbf{x}, \\ \log \left[ \frac{\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})}{\pi_{3i}(\mathbf{X})} \right] &= \mu^{(2)} + \alpha_i^{(2)} + [\beta^{(2)} + \gamma_i^{(2)}] \mathbf{x}. \end{aligned}$$



## Modélisation de la pression sanguine

$$\log \left[ \frac{\pi_{1i}(\mathbf{X})}{\pi_{2i}(\mathbf{X}) + \pi_{3i}(\mathbf{X})} \right] = \mu^{(1)} + \alpha_i^{(1)} + [\beta^{(1)} + \gamma_i^{(1)}] \mathbf{X},$$
$$\log \left[ \frac{\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})}{\pi_{3i}(\mathbf{X})} \right] = \mu^{(2)} + \alpha_i^{(2)} + [\beta^{(2)} + \gamma_i^{(2)}] \mathbf{X}.$$



## Modélisation de la pression sanguine

$$\log \left[ \frac{\pi_{1i}(X)}{\pi_{2i}(X) + \pi_{3i}(X)} \right] = \mu^{(1)} + \alpha_i^{(1)} + [\beta^{(1)} + \gamma_i^{(1)}] X,$$

$$\log \left[ \frac{\pi_{1i}(X) + \pi_{2i}(X)}{\pi_{3i}(X)} \right] = \mu^{(2)} + \alpha_i^{(2)} + [\beta^{(2)} + \gamma_i^{(2)}] X.$$

**Attention : problème si les prédicteurs se croisent**

$$\log \left[ \frac{\pi_{1i}(X)}{\pi_{2i}(X) + \pi_{3i}(X)} \right] > \log \left[ \frac{\pi_{1i}(X) + \pi_{2i}(X)}{\pi_{3i}(X)} \right]$$



## Modélisation de la pression sanguine

$$\log \left[ \frac{\pi_{1i}(\mathbf{X})}{\pi_{2i}(\mathbf{X}) + \pi_{3i}(\mathbf{X})} \right] = \mu^{(1)} + \alpha_i^{(1)} + [\beta^{(1)} + \gamma_i^{(1)}] \mathbf{X},$$

$$\log \left[ \frac{\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})}{\pi_{3i}(\mathbf{X})} \right] = \mu^{(2)} + \alpha_i^{(2)} + [\beta^{(2)} + \gamma_i^{(2)}] \mathbf{X}.$$

**Attention : problème si les prédicteurs se croisent**

$$\log \left[ \frac{\pi_{1i}(\mathbf{X})}{\pi_{2i}(\mathbf{X}) + \pi_{3i}(\mathbf{X})} \right] > \log \left[ \frac{\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})}{\pi_{3i}(\mathbf{X})} \right],$$

$$\pi_{1i}(\mathbf{X})\pi_{3i}(\mathbf{X}) > [\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})][\pi_{3i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})]$$



## Modélisation de la pression sanguine

$$\log \left[ \frac{\pi_{1i}(\mathbf{X})}{\pi_{2i}(\mathbf{X}) + \pi_{3i}(\mathbf{X})} \right] = \mu^{(1)} + \alpha_i^{(1)} + [\beta^{(1)} + \gamma_i^{(1)}] \mathbf{X},$$

$$\log \left[ \frac{\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})}{\pi_{3i}(\mathbf{X})} \right] = \mu^{(2)} + \alpha_i^{(2)} + [\beta^{(2)} + \gamma_i^{(2)}] \mathbf{X}.$$

**Attention : problème si les prédicteurs se croisent**

$$\log \left[ \frac{\pi_{1i}(\mathbf{X})}{\pi_{2i}(\mathbf{X}) + \pi_{3i}(\mathbf{X})} \right] > \log \left[ \frac{\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})}{\pi_{3i}(\mathbf{X})} \right],$$

$$\pi_{1i}(\mathbf{X})\pi_{3i}(\mathbf{X}) > [\pi_{1i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})][\pi_{3i}(\mathbf{X}) + \pi_{2i}(\mathbf{X})]$$



## Modèle à risques proportionnels

$$\log \left[ \frac{\pi_{1i}(X)}{\pi_{2i}(X) + \pi_{3i}(X)} \right] = \mu^{(1)} + \alpha_i + [\beta + \gamma_i] X,$$
$$\log \left[ \frac{\pi_{1i}(X) + \pi_{2i}(X)}{\pi_{3i}(X)} \right] = \mu^{(2)} + \alpha_i + [\beta + \gamma_i] X.$$

**Risques proportionnels : odds-ratio**

$$\text{OR}_{1 \rightarrow [2-3]} = \text{OR}_{[1-2] \rightarrow 3}$$



# Estimation par la méthode du maximum de vraisemblance

Données **Pression sanguine**

Coefficients			
	Estimation	Écart-type	Stat. t
weight	0.019670849	0.007139504	2.7552124
genderM	0.212015362	0.754463596	0.2810147
weight:genderM	-0.007645852	0.009648443	-0.7924441
Constantes			
	Estimation	Écart-type	Stat. t
1 2	1.2934	0.5235	2.4708
2 3	3.3040	0.5443	6.0702
Déviance résiduelle : 958.2028			
AIC : 968.2028			



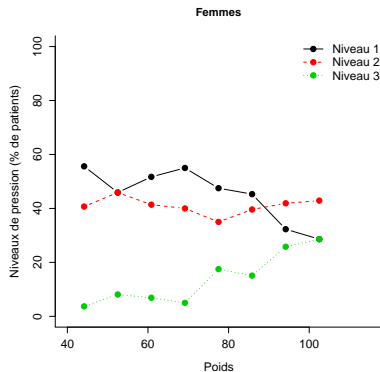
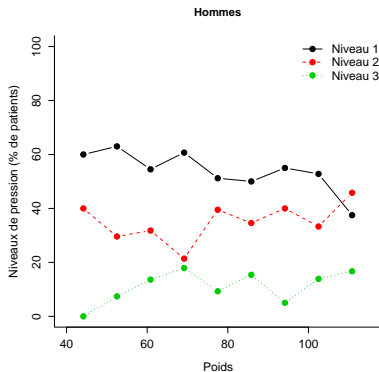
# Estimation par la méthode du maximum de vraisemblance

Données **Pression sanguine**

		Coefficients		
	Estimation	Écart-type	Stat. t	
weight	0.019670849	0.007139504	2.7552124	
genderM	0.212015362	0.754463596	0.2810147	
weight:genderM	-0.007645852	0.009648443	-0.7924441	
		Constantes		
	Estimation	Écart-type	Stat. t	
1 2	1.2934	0.5235	2.4708	
2 3	3.3040	0.5443	6.0702	
Déviance résiduelle : 958.2028				
AIC : 968.2028				

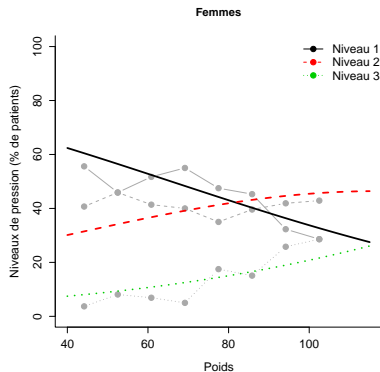
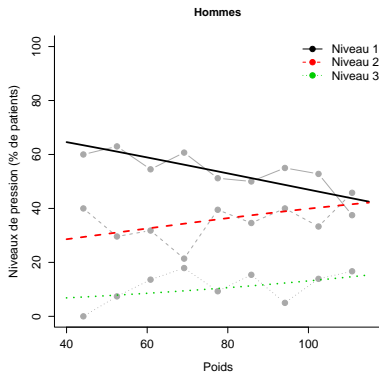


# Estimation par la méthode du maximum de vraisemblance





# Estimation par la méthode du maximum de vraisemblance





## Analyse de la déviance

Analysis of deviance							
	Model	Resid. df	Resid. dev	Test	Df	LR stat	Pr (Chi)
1	1	498	970.9707				
2	weight	497	963.0900	1 vs 2	1	7.8807009	0.004996511
3	weight + gender	496	958.8318	2 vs 3	1	4.2581469	0.039062510
4	weight * gender	495	958.2028	3 vs 4	1	0.6290315	0.427710781



## Analyse de la déviance

Analysis of deviance							
	Model	Resid. df	Resid. dev	Test	Df	LR stat	Pr (Chi)
1	1	498	970.9707				
2	weight	497	963.0900	1 vs 2	1	7.8807009	0.004996511
3	weight + gender	496	958.8318	2 vs 3	1	4.2581469	0.039062510
4	weight * gender	495	958.2028	3 vs 4	1	0.6290315	0.427710781



# Analyse de la déviance

## Données Pression sanguine

Coefficients			
	Estimation	Écart-type	Stat. t
weight	0.01552393	0.004826471	3.216415
genderM	-0.36920906	0.179606541	-2.055655
Constantes			
	Estimation	Écart-type	Stat. t
1 2	0.9970	0.3630	2.7463
2 3	3.0047	0.3880	7.7436
Déviance résiduelle : 958.8318			
AIC : 966.8318			



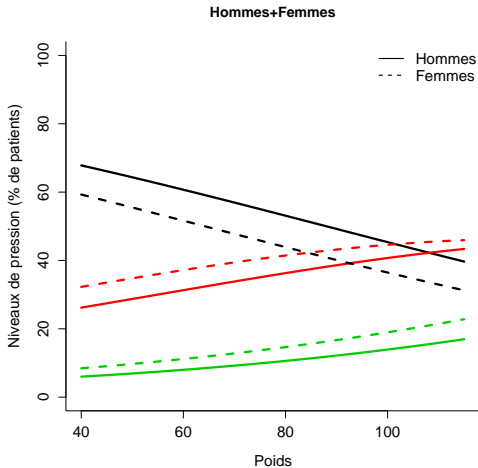
# Analyse de la déviance

Données **Pression sanguine**

Coefficients			
	Estimation	Écart-type	Stat. t
weight	0.01552393	0.004826471	3.216415
genderM	-0.36920906	0.179606541	-2.055655
Constantes			
	Estimation	Écart-type	Stat. t
1 2	0.9970	0.3630	2.7463
2 3	3.0047	0.3880	7.7436
Déviance résiduelle : 958.8318			
AIC : 966.8318			



# Analyse de la déviance





## Règle de discrimination

Estimation du modèle à partir d'un échantillon

$$\hat{\pi}_{1i}(x) = \frac{\exp(\hat{\mu}^{(1)} + \hat{\alpha}_i + \hat{\beta}x)}{1 + \exp(\hat{\mu}^{(1)} + \hat{\alpha}_i + \hat{\beta}x)},$$

$$\hat{\pi}_{1i}(x) + \hat{\pi}_{2i}(x) = \frac{\exp(\hat{\mu}^{(2)} + \hat{\alpha}_i + \hat{\beta}x)}{1 + \exp(\hat{\mu}^{(2)} + \hat{\alpha}_i + \hat{\beta}x)},$$

$$\hat{\pi}_{1i}(x) + \hat{\pi}_{2i}(x) + \hat{\pi}_{3i}(x) = 1$$

Règle d'affectation pour un individu (*Sexe* =  $i$ , *Poids* =  $x$ )

Si  $\hat{k} = \arg \max \{ \hat{\pi}_{ki}(x) \}$ , alors  $\hat{Y} = k$ .



## Validation

Matrice de confusion (validation interne)

	Prédit			
Observé	1	2	3	Total
1	85.6	14.4	0	100
2	79.4	20.6	0	100
3	73.8	26.2	0	100

Taux d'erreurs très élevés



# Plan du cours

- 1 Préférence d'une marque
- 2 Modèles pour variables qualitatives
  - Modèles pour variables qualitatives nominales
  - Modèles pour variables qualitatives ordinales
  - Discrimination logistique
- 3 Arbres de classification
  - Règle binaire de classification
  - Arbre récursif binaire
- 4 Bilan



## Motivations

### Inconvénients des méthodes de type *Scoring*

- Complexes si les variables explicatives sont nombreuses et de natures différentes  
Données **Pression sanguine** : Age, **Sexe**, Poids, **Statut marital**, **Fumeur ou non**, ...
- Peu intuitives par rapport à des **règles récursives binaires** :

Si  $\text{Age} < 35$ , **alors** Pression sanguine = 1,

Si  $\text{Age} \geq 35$  **et** Fumeur = Non, **alors** Pression sanguine = 1, ...



## Règles binaires

Règle binaire : Si  $C(x)$  alors  $\hat{Y} = y$

- Si  $x$  est quantitative,  $C(x)$  est de la forme  $x < s$
- Si  $x$  est qualitative nominale  $[x \in \{A, B, C\}]$   
 $C(x)$  est de la forme  $x \in \{A, C\}$
- Si  $x$  est qualitative ordinale  $[x \in \{A \leq B \leq C\}]$   
 $C(x)$  est de la forme  $x \leq B$



## Règles binaires

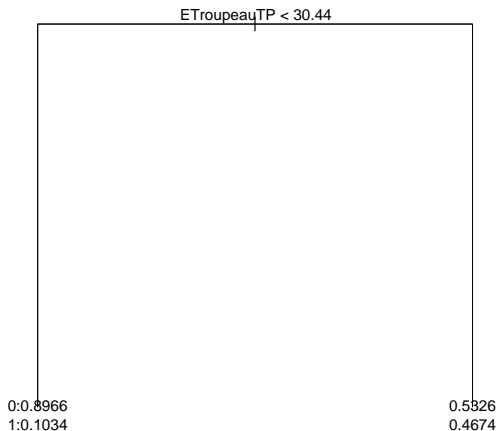
Règle binaire : Si  $C(x)$  alors  $\hat{Y} = y$

- Si  $x$  est quantitative,  $C(x)$  est de la forme  $x < s$
- Si  $x$  est qualitative nominale  $[x \in \{A, B, C\}]$   
 $C(x)$  est de la forme  $x \in \{A, C\}$
- Si  $x$  est qualitative ordinale  $[x \in \{A \leq B \leq C\}]$   
 $C(x)$  est de la forme  $x \leq B$

Nœuds d'une règle binaire : sous-ensemble des individus vérifiant  $C(x)$  et sous-ensemble des individus vérifiant  $\bar{C}(x)$



# Règles binaires





## Règles binaires

Règle binaire : Si  $C(x)$  alors  $\hat{Y} = y$

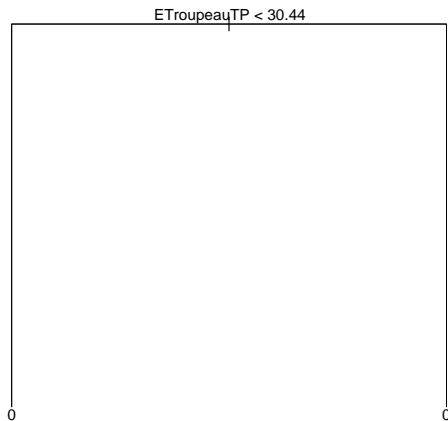
- Si  $x$  est quantitative,  $C(x)$  est de la forme  $x < s$
- Si  $x$  est qualitative nominale  $[x \in \{A, B, C\}]$   
 $C(x)$  est de la forme  $x \in \{A, C\}$
- Si  $x$  est qualitative ordinale  $[x \in \{A \leq B \leq C\}]$   
 $C(x)$  est de la forme  $x \leq B$

Nœuds d'une règle binaire : sous-ensemble des individus vérifiant  $C(x)$  et sous-ensemble des individus vérifiant  $\bar{C}(x)$

Affectation des individus d'un même nœud au groupe majoritaire dans ce nœud



# Règles binaires





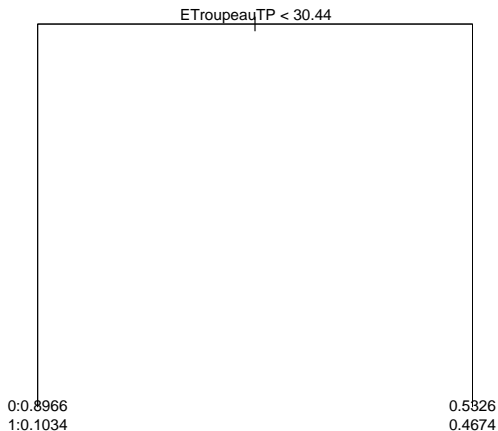
# Qualité d'une règle binaire

## Critères de qualité

- Taux de bien classés [Uniquement qualité de classification]

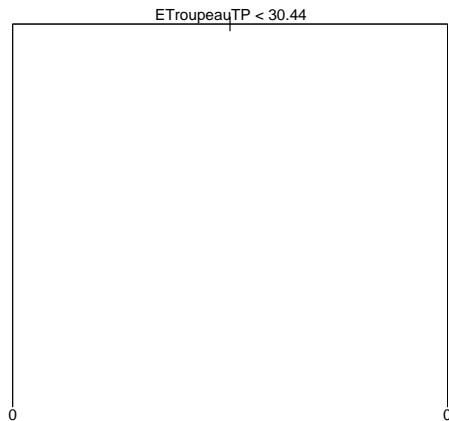


## Qualité d'une règle binaire





## Qualité d'une règle binaire





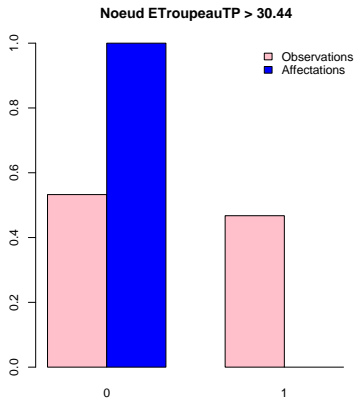
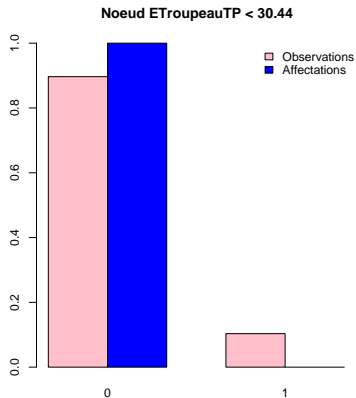
## Qualité d'une règle binaire

### Critères de qualité

- Taux de bien classés [Uniquement qualité de classification]
- Déviance des noeuds de la règle

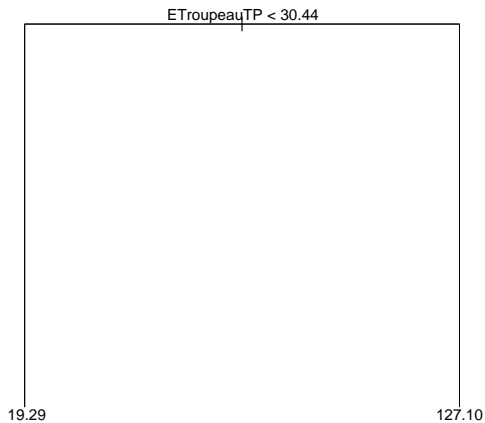


## Qualité d'une règle binaire





## Qualité d'une règle binaire





## Qualité d'une règle binaire

### Critères de qualité

- Taux de bien classés [Uniquement qualité de classification]
- Déviance des noeuds de la règle

$$19.29 + 127.10 = 146.39 \quad [\text{Déviance de base} = 160.70]$$





## Qualité d'une règle binaire

### Critères de qualité

- Taux de bien classés [Uniquement qualité de classification]
- Déviance des noeuds de la règle
- Règle binaire optimale : minimisation de la déviance



## Règle multivariée de classification

Arbre optimal : séquence de règles binaires optimales définies par récursivité

- Point de départ : l'échantillon complet [Nœud 1 : racine]



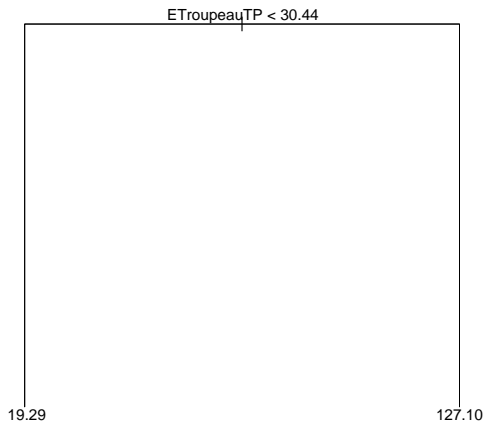
## Règle multivariée de classification

Arbre optimal : séquence de règles binaires optimales définies par récursivité

- Point de départ : l'échantillon complet [Nœud 1 : racine]
- 1ère étape : choix de la meilleure règle binaire [Nœuds  $N_2$  et  $N_3$ ]



## Règle multivariée de classification





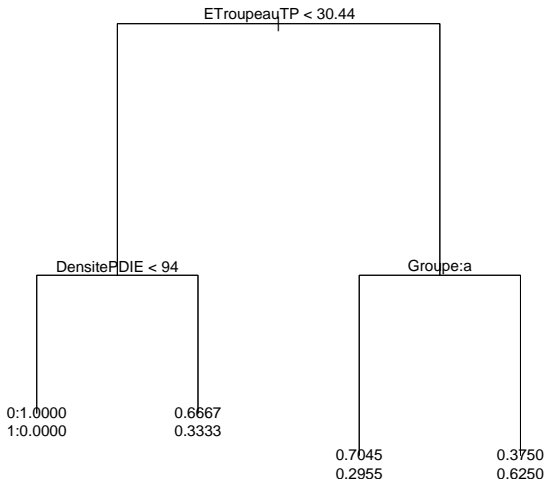
## Règle multivariée de classification

Arbre optimal : séquence de règles binaires optimales définies par récursivité

- Point de départ : l'échantillon complet [Nœud 1 : racine]
- 1ère étape : choix de la meilleure règle binaire  
[Nœuds  $N_2$  et  $N_3$ ]
- 2ème étape : choix de la meilleure règle binaire dans  $N_2$  et  $N_3$   
[Nœuds  $N_4$ ,  $N_5$ ,  $N_6$  et  $N_7$ ]

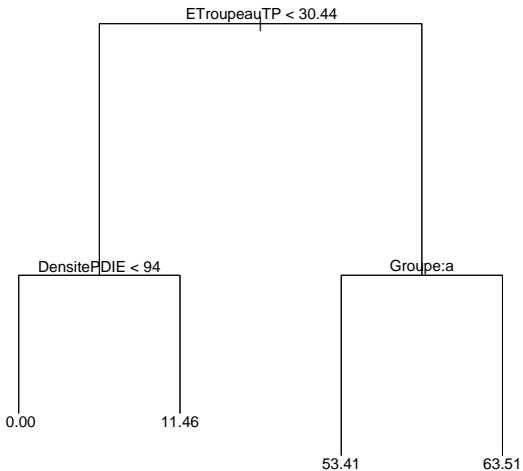


## Règle multivariée de classification





## Règle multivariée de classification





## Règle multivariée de classification

Arbre optimal : séquence de règles binaires optimales définies par récursivité

- Point de départ : l'échantillon complet [Nœud 1 : racine]
- 1ère étape : choix de la meilleure règle binaire  
[Nœuds  $N_2$  et  $N_3$ ]
- 2ème étape : choix de la meilleure règle binaire dans  $N_2$  et  $N_3$   
[Nœuds  $N_4$ ,  $N_5$ ,  $N_6$  et  $N_7$ ]
- 3ème étape : choix de la meilleure règle binaire dans  $N_4$ ,  $N_5$ , ...



## Règle multivariée de classification

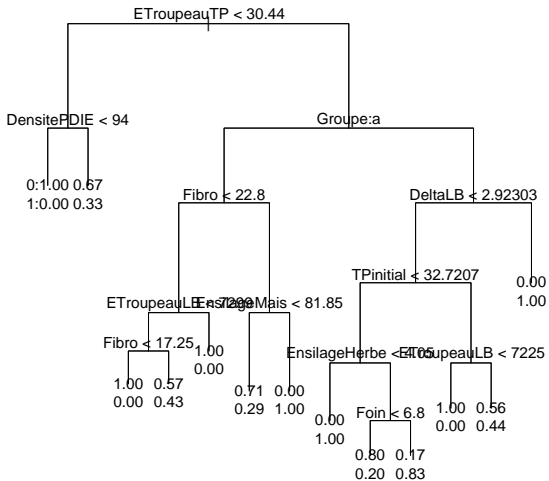
Arbre optimal : séquence de règles binaires optimales définies par récursivité

- Point de départ : l'échantillon complet [Nœud 1 : racine]
- 1ère étape : choix de la meilleure règle binaire  
[Nœuds  $N_2$  et  $N_3$ ]
- 2ème étape : choix de la meilleure règle binaire dans  $N_2$  et  $N_3$   
[Nœuds  $N_4$ ,  $N_5$ ,  $N_6$  et  $N_7$ ]
- 3ème étape : choix de la meilleure règle binaire dans  $N_4$ ,  $N_5$ , ...

Critères d'arrêt : déviance nulle ou nombre d'individus trop faibles dans un nœud terminal

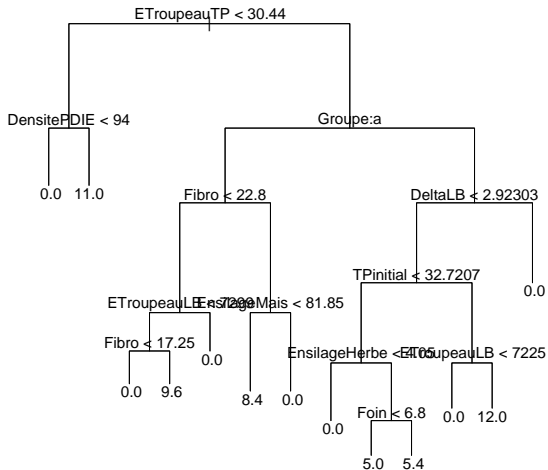


# Règle multivariée de classification





# Règle multivariée de classification





## Règle multivariée de classification

Arbre optimal : séquence de règles binaires optimales définies par récursivité

- Point de départ : l'échantillon complet [Nœud 1 : racine]
- 1ère étape : choix de la meilleure règle binaire  
[Nœuds  $N_2$  et  $N_3$ ]
- 2ème étape : choix de la meilleure règle binaire dans  $N_2$  et  $N_3$   
[Nœuds  $N_4$ ,  $N_5$ ,  $N_6$  et  $N_7$ ]
- 3ème étape : choix de la meilleure règle binaire dans  $N_4$ ,  $N_5$ , ...

Critères d'arrêt : déviance nulle ou nombre d'individus trop faibles dans un nœud terminal

Taille d'un arbre : nombre de nœuds terminaux [feuilles]

Déviance d'un arbre : somme des déviances de ses feuilles



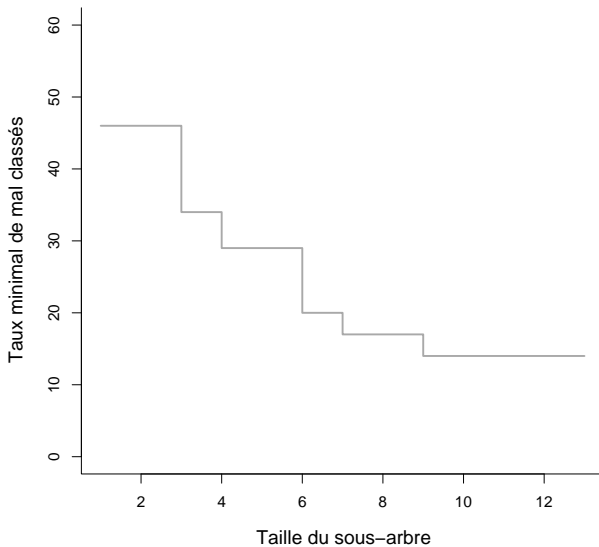
## Élagage de l'arbre

Soit  $K$  la taille de l'arbre optimal,

l'arbre optimal de taille  $k \leq K$  est un sous-arbre de l'arbre optimal

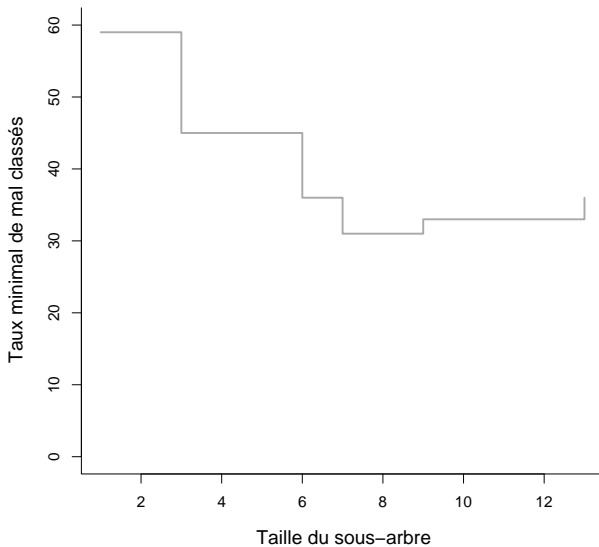


## Élagage de l'arbre



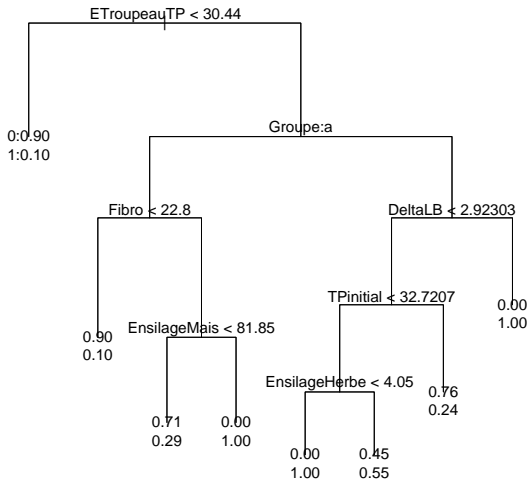


## Élagage de l'arbre





# Élagage de l'arbre





## Élagage de l'arbre

Soit  $K$  la taille de l'arbre optimal,

l'arbre optimal de taille  $k \leq K$  est un sous-arbre de l'arbre optimal

Matrice de confusion

Observations	Affectations		Total
	0	1	
0	70	5	75
1	12	34	46



## Élagage de l'arbre

Soit  $K$  la taille de l'arbre optimal,

l'arbre optimal de taille  $k \leq K$  est un sous-arbre de l'arbre optimal

Matrice de confusion

	Affectations		
Observations	0	1	
0	93.3	6.7	100
1	26.1	73.9	100



# Plan du cours

- 1 Préférence d'une marque
- 2 Modèles pour variables qualitatives
  - Modèles pour variables qualitatives nominales
  - Modèles pour variables qualitatives ordinales
  - Discrimination logistique
- 3 Arbres de classification
  - Règle binaire de classification
  - Arbre récursif binaire
- 4 Bilan



## Bilan

	2 modalités	>2 modalités	$\geq 2$ modalités
	Modèle logistique	Modèle logistique logistique multinomial	Arbre de classification
Explication	•	•	
Prédiction	•	•	•
Mise en œuvre			•