SOCI1241-1 Eléments du calcul des probabilités appliquées aux sciences sociales et exercices pratiques (en ce compris les bases de statistiques inférentielles)

Transparents

Philippe Lambert

http : //www.statsoc.ulg.ac.be/proba.html

Institut des Sciences Humaines et Sociales
Université de Liège

La régression logistique

Références spécialisées

- Agresti, A. (2002, 2nd ed.) Categorical Data Analysis. Wiley.
- Hosmer, D.W and Lemeshow, S. (2001, 2nd ed.) Applied Logistic Regression. Wiley.

Motivation

- La régression logistique est un outil puissant permettant de modéliser comment la probabilité de succès change avec des variables explicatives catégorielles ou continues.

Ex Evolution du risque du développement de la violence dans un couple avec la durée de la relation, la nationalité des partenaires, l'âge de la conjointe au début de l'union, les revenus du ménage, la consommation d'alcool, etc.

- C'est une extension de la comparaison directe de 2 proportions.
- Avant de présenter, par étapes, cet outil, quelques définitions et concepts nouveaux sont nécessaires.

Probabilité et cote du succès

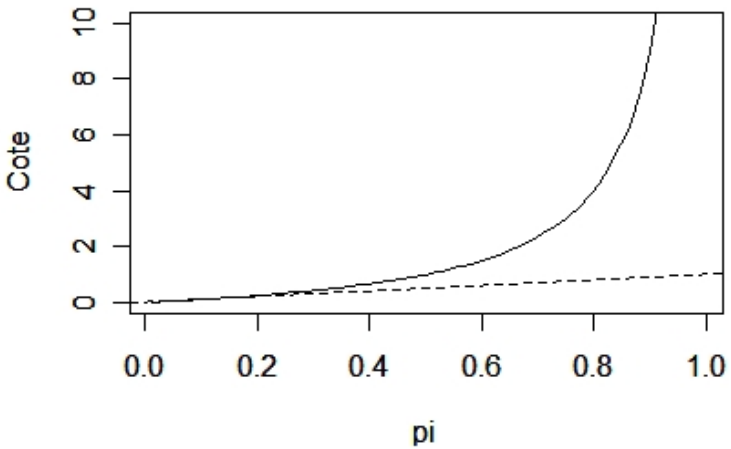
- On appelle *cote du succès* le rapport

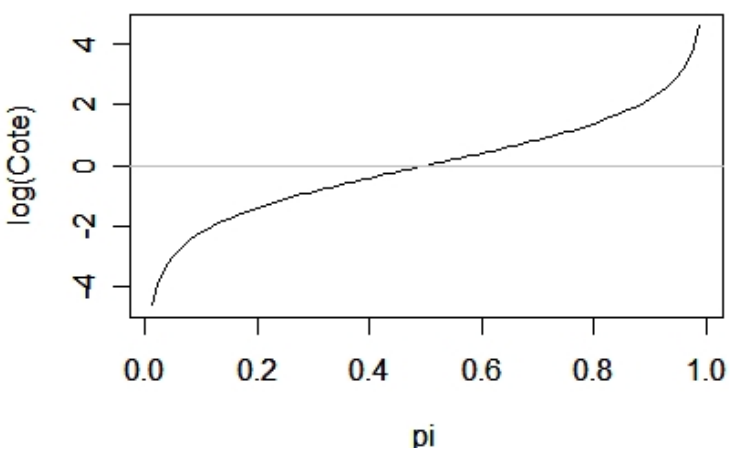
$$e^{\theta} = \frac{\pi}{1-\pi}$$

où π est la probabilité de succès (parfois également appelée le *risque*).

- Probabilité et cote sont très proches lorsque la 1ère est inférieure à 20%.

- (Le logarithme de) cette cote est
  - $(< 0) < 1$ lorsque $\pi < 0.5$,
  - $(= 0) = 1$ lorsque $\pi = 0.5$,
  - $(> 0) > 1$ lorsque $\pi > 0.5$,
  - $(\to -\infty) \to 0$ lorsque $\pi \to 0$,
  - $(\to +\infty) \to +\infty$ lorsque $\pi \to 1$.





Probabilité et cote du succès (2)

Ex Nombre de souris développant une tumeur au poumon après exposition à la fumée de cigarettes (Essenberg, Science, 1952).

Groupe	Tumeur présente	Tumeur absente	Total
Traitement	$n_{11} = 21$	$n_{12} = 2$	23
Contrôle	$n_{21} = 19$	$n_{22} = 13$	32

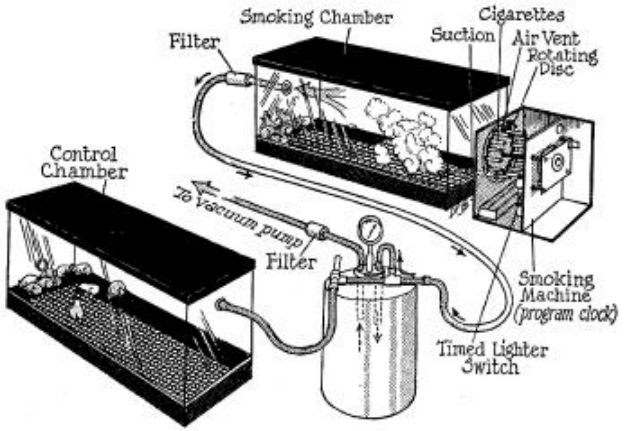


FIG. 1. The smoking machine, consisting of experimental and control units. Connection with the vacuum pump is indicated.

? L'exposition augmente-t-elle le risque de cancer?

• Si on ignore l'exposition éventuelle à la fumée et si un a priori uniforme est pris pour le risque de base, alors le mode a posteriori pour la probabilité de succès est

$$\hat{\pi} = 40/55 = 0.73 \;\Rightarrow\; \mathrm{e}^{\hat{\theta}} = \frac{\hat{\pi}}{1-\hat{\pi}} = 2.67 \;\Rightarrow\; \hat{\theta} = 0.98$$

$$\text{On a la relation } \pi = \frac{\mathrm{e}^{\theta}}{1+\mathrm{e}^{\theta}}$$

Le (logarithme du) rapport de cotes

• On peut calculer la cote de succès dans différentes conditions. Le *rapport de cotes* Ω permet alors d'évaluer l'influence du facteur considéré:

$$\Omega = \frac{e^{\theta_1}}{e^{\theta_2}} = e^{\theta_1 - \theta_2}$$

est > 1 (< 1) lorsque le succès a une cote supérieure (inférieure) pour le 1er niveau du facteur.

• Le *logarithme du rapport de cotes*, $\theta_1 - \theta_2$, est > 0 (< 0) lorsque le succès a une cote supérieure (inférieure) pour le 1er niveau du facteur.

Ex Dans notre exemple, sous un a priori unforme pour les risques dans les 2 conditions, le mode a posteriori pour la cote du succès est

$$\widehat{\text{Cote}}(\text{succès}|\text{exposé}) = e^{\hat{\theta}_1} = 21/2 = 10.50 \;\; ; \;\; \widehat{\text{Cote}}(\text{succès}|\text{contrôle}) = e^{\hat{\theta}_2} = 19/13 = 1.46$$

$$\Rightarrow \;\; \hat{\text{RC}} = \hat{\Omega} = 10.50/1.46 = 7.18 > 1 \quad \Rightarrow \;\; \log \hat{\text{RC}} = \log \hat{\Omega} = \hat{\theta}_1 - \hat{\theta}_2 = 1.97 > 0$$

Le RC indique que la cote de la tumeur est supérieure (multipliée par 7) lorsque les souris sont exposées à la fumée de cigarettes.

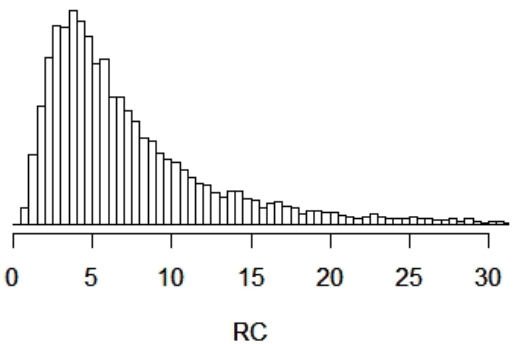
Distribution a posteriori pour le rapport de cotes Ω

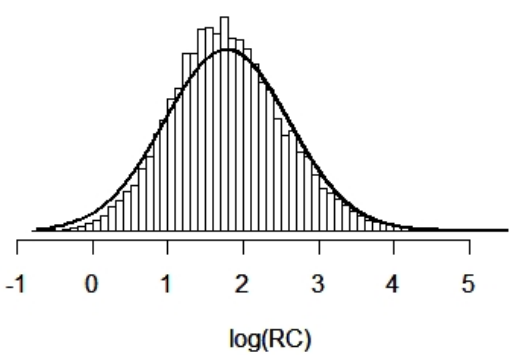
• En considérant les groupes "Contrôle" et "Traitement" séparément (cf. plan d'expérience) avec des a priori uniformes sur $(0, 1)$, on déduit les a posteriori pour les risques:

$$(\pi_1|\text{donnees}) \sim \text{Beta}(22, 3) \;\; ; \;\; (\pi_2|\text{donnees}) \sim \text{Beta}(20, 14)$$

• On ne peut pas en déduire aisément une forme analytique pour la distribution du rapport des cotes $\Omega = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ et de son logarithme.

• Cependant, comme pour l'inférence à propos d'une différence de proportions, on peut simuler un très grand nombre M (=10000 par ex.) de paires $\{(\pi_1^{(m)}, \pi_2^{(m)}) : m = 1, \ldots, M\}$ de ces distributions a posteriori et étudier les propriétés de $\left\{\Omega^{(m)} = \frac{\pi_1^{(m)}/(1-\pi_1^{(m)})}{\pi_2^{(m)}/(1-\pi_2^{(m)})} : m = 1, \ldots, M\right\}$.





Distribution a posteriori pour le rapport de cotes Ω (2)

De cette simulation , nous pouvons estimer:

- La moyenne a posteriori de Ω et $\log(\Omega)$: 8.15 et 1.79.
- la variance a posteriori de Ω et $\log(\Omega)$: 79.48 et 0.569.
- L'écart-type a posteriori de Ω et $\log(\Omega)$: 8.92 et 0.75.
- Les quantiles a posteriori 2.5% et 97.5% de Ω et $\log(\Omega)$:

$$2.5\% : \ 1.46 \text{ et } 0.38 \ ; \ 95\% : 28.76 \text{ et } 3.36$$

On en déduit les intervalles de crédibilité 95% respectifs $(1.46, 28.76)$ et $(0.38, 3.36)$.

- La probabilité a posteriori que $\Omega > 1$: elle permet d'évaluer la plausibilité que l'exposition à la fumée augmente le risque pour ces souris de développer une cancer. Elle peut être estimée par la proportion des $\Omega^{(m)}$ supérieurs à 1: 0.99.

Approximation normale pour $(\Omega|\text{donnees})$

- Le seul histogramme dont la forme se rapproche d'une cloche est celui de $\log(\Omega)$.
- Une approximation normale de la distribution a posteriori de $\alpha = \log(\Omega)$ est:

$$(\log(\Omega)|\text{donnees}) \sim \mathcal{N}(\tilde{\alpha}, \sigma^2_\alpha)$$
$$\tilde{\alpha} = \log(n_{11} + .5) + \log(n_{22} + .5) - \log(n_{21} + .5) - \log(n_{12} + .5)$$
$$\sigma^2_{\tilde{\alpha}} = n_{11}^{-1} + n_{22}^{-1} + n_{12}^{-1} + n_{22}^{-1}$$

Ex Dans notre exemple, nous avons approximativement

$$(\log(\Omega)|\text{donnees}) \sim \mathcal{N}(1.78, 0.677 = 0.82^2)$$

- Grâce à cette approximation, nous pouvons notamment approximer

→ La moyenne a posteriori de $\log(\Omega)$: 1.78.

→ L'écart-type a posteriori de $\log(\Omega)$: 0.82.

- On en déduit un intervalle de crédibilité 95% approximatif pour $\log(\Omega)$:

$$1.78 \pm 1.96 \times 0.82 = (0.17, 3.39)$$

- On approxime la probabilité a posteriori que $\log(\Omega) > 0$ par

$$\text{P}(Z > (0 - 1.78)/0.82) = 0.99$$

La régression logistique avec une explicative catégorielle

Définition

• Examinons comment la régression logisitique est définie dans le cas simple où une seule variable explicative catégorielle est disponible.

• Si Z est une variable explicative à K niveaux, le modèle logistique suppose que

$$(Y|Z = z_k) \sim \text{Bin}(n_k, \pi_k) \quad \text{avec}$$

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \theta_k = \mu + \alpha_k \;\; ; \;\; (\alpha_1 = 0) \quad \Rightarrow \pi_k = \frac{\exp(\mu + \alpha_k)}{1 + \exp(\mu + \alpha_k)}$$

• Le logarithme de la cote du succès sous le 1er niveau du facteur vaut μ.

• Le logarithme du rapport des cotes du succès du $k$ème niveau versus le 1er vaut

$$\theta_k - \theta_1 = \alpha_k$$

Par conséquent, une valeur de $\alpha_k > 0$ (< 0) indique que la cote du succès observée est plus grande (petite) sous le $k$ème niveau du facteur que sous le 1er niveau.

La régression logistique avec une explicative catégorielle (2)

Sortie logicielle typique

• Les logiciels supposent, pour la plupart, que les a priori concernant les paramètres μ et α_k $(k = 2, \ldots, K)$ sont *non informatifs*: ils n'ont donc (pratiquement) pas d'influence sur les résultats.

• Voici une sortie typique pour un tel logiciel:

Tumeur - Paramètres estimés (souris.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	0,379490	0,359937	1,111596	0,291735
Groupe	Traitement	2	1,971886	0,822897	5,742129	0,016563
Groupe	Contrôle	3	0,000000			
Echelle			1,000000	0,000000		

Tumeur - Test Vraisemblance Type 1 (souri
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-32,2274		
Groupe	1	-28,4100	7,634852	0,005725

• On retrouve les estimations déjà obtenues précédemment pour le logarithme
→ de la cote du succès du groupe de référence (contrôle): $\hat{\mu} = 0.38$,
→ du rapport des cotes du succès (traités versus contrôle): $\hat{\alpha}_2 = 1.97$.

La régression logistique avec une explicative catégorielle (3)

• Les erreurs standards (ou "écarts-types") des coefficients permettent de dériver des approximations normales aux distributions a posteriori (sans assurance qu'elles soient de qualité!):

$$(\hat{\mu}|\text{donnees}) \sim \mathcal{N}(0.38, 0.36^2) \;\; ; \;\; (\hat{\alpha}_2|\text{donnees}) \sim \mathcal{N}(1.97, 0.82^2)$$

• On peut, notamment, en déduire, pour le logarithme du rapport de cotes, une approximation

→ d'un intervalle de crédibilité 95%: $1.97 \pm 1.96 \times 0.82 = (0.36, 3.58)$.

→ de $\text{P}(\log(\Omega) > 0|\text{donnees}) \approx \text{P}(Z > (0 - 1.97)/0.82) = 0.99$.

Remarques

• Le coefficient "P" qui apparaît en regard de chaque coefficient estimé s'appelle la P-valeur. Il permet d'évaluer si 0 fait partie des valeurs plausibles pour le paramètre: la règle consensuelle est de déclarer ce fait non plausible si $P < 0.05$.

• L'approximation normale n'est pas toujours fiable. Cela arrive notamment lorsque les fréquences de la table de contingence sont petites.

La régression logistique avec une explicative catégorielle (4)

• Nous avons vu comment une approximation de l'a posteriori du log(RC) peut être utilisée pour déterminer si 0 est une valeur plausible pour α_2.

• Une méthode plus fiable consiste à comparer les fréquences observées avec les fréquences attendues sous l'hypothèse que la variable explicative n'est pas nécessaire (càd sous l'hypothèse que $\alpha_2 = 0$).

• Sous cette dernière hypothèse, $\pi_1 = \pi_2 = \pi$ et $\hat{\pi} = \frac{n_{11}+n_{21}}{n_{++}} = \frac{n_{+1}}{n_{++}} = \frac{40}{55}$

Le nombre de souris avec tumeur attendu, en moyenne, est alors estimé par
→ $n_{1+}\hat{\pi} = \frac{n_{1+} \times n_{+1}}{n_{++}} = \frac{23 \times 40}{55} = 16.73$ chez les traitées,
→ $n_{2+}\hat{\pi} = \frac{n_{2+} \times n_{+1}}{n_{++}} = \frac{32 \times 40}{55} = 23.27$ chez les contrôles.

• Nous avons donc la table suivante (avec les attendus entre parenthèses):

Groupe	Tumeur présente	Tumeur absente	Total
Traitement	$n_{11} = 21$ (16.73)	$n_{12} = 2$ (6.27)	23
Contrôle	$n_{21} = 19$ (23.27)	$n_{22} = 13$ (8.73)	32

La régression logistique avec une explicative catégorielle (5)

• La comparaison des fréquences observées et attendues permet d'évaluer si 0 est une valeur plausible pour α_2.

• Le *test de rapport de vraisemblance (ou chi-carré)* suggère de comparer ces valeurs à l'aide de:

$$G^2 = -2 \sum_i \sum_j n_{ij} \log \frac{A_{ij}}{n_{ij}}$$

où A_{ij} est la 'fréquence' attendue dans la cellule (i, j).

Ici, nous avons

$$G^2 = -2 \left\{ 21 \log \frac{16.73}{21} + 19 \log \frac{23.27}{19} + 2 \log \frac{6.27}{2} + 13 \log \frac{8.73}{13} \right\} = 7.63$$

Comme nous avons $G^2 = 7.63 > \chi^2_1(0.95) = 3.84$, on conclut que 0 ne fait pas partie des valeurs plausibles pour α_2, ($\chi^2_1(0.95)$ étant le quantile 95% de la distribution chi-carré avec un degré de liberté).

• Cette statistique G^2 peut être calculée par tout logiciel comptant la régression logistique parmi ses outils. Dans notre exemple, on parle aussi de *test d'indépendance*.

La régression logistique avec une explicative catégorielle (6)

Autre exemple

• Relation entre les habitudes tabagiques d'étudiants en Arizona et les habitudes de leurs parents (Agresti, 1990, p. 124).

# parents fumeurs	Enfant fume	Enfant ne fume pas	Total
Deux	400	1380	1780
Un seul	416	1823	2239
Aucun	188	1168	1358

• Si le succès est défini comme étant le fait de fumer pour l'enfant, le modèle logistique précédent devient

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \theta_k = \mu + \alpha_k \quad ; \quad (\alpha_1 = 0)$$

Prenons "Aucun" comme catégorie de référence.

La régression logistique avec une explicative catégorielle (7)

Enfant fumeur - Paramètres estimés (chap3_parentsEnfants.st
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-1,82661	0,078583	540,2950	0,000000
Parents	Deux	2	0,58823	0,096953	36,8105	0,000000
Parents	Un	3	0,34905	0,095539	13,3481	0,000259
Parents	Aucun	4	0,00000			
Echelle			1,00000	0,000000		

Enfant fumeur - Test Vraisemblance Type 1
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-2588,26		
Parents	2	-2569,07	38,36582	0,000000

• Comme $G^2 = 7.63 > \chi^2_2(0.95) = 5.99$, on conclut que $\alpha_2 = \alpha_3 = 0$ n'est pas une configuration plausible pour les α_k's: au moins un de ces logarithmes de rapport de cotes se distingue significativement de 0.

• Donc, $\log\Omega$(Un vs Aucun)$= \alpha_2 = 0.35$ et $\log\Omega$(Deux vs Aucun)$= \alpha_3 = 0.59$ avec intervalles de crédibilité 95% approximatifs respectivement égaux à $(0.16, 0.54)$ et $(0.40, 0.78)$.

• Ces intervalles suggèrent que la probabilité qu'un enfant fume est plus grande lorsqu'un ou deux de ses parents fument. On a approximativement:
$\rightarrow$ $\mathrm{P}(\log\Omega(\text{Un vs Aucun}) > 0) \approx \mathrm{P}(Z > (0 - 0.35)/0.096) = 1.00$
$\rightarrow$ $\mathrm{P}(\log\Omega(\text{Deux vs Aucun}) > 0) \approx \mathrm{P}(Z > (0 - 0.59)/0.096) = 1.00.$

La régression logistique avec une explicative catégorielle (8)

Remarque

• La sortie du logiciel ne permet pas d'approximer la distribution a posteriori de $\log \Omega(\text{Deux vs Un})$. On dispose juste d'une approximation du mode a posteriori

$$\begin{aligned} \log \hat{\Omega}(\text{Deux vs Un}) &= \log \hat{\Omega}(\text{Deux vs Aucun}) - \log \hat{\Omega}(\text{Un vs Aucun}) \\ &= 0.59 - 0.35 = 0.24 \end{aligned}$$

• G^2 a été comparé au quantile 95% d'une distribution chi-carré avec 2 degrés de liberté. Cette valeur de 2 correspond à $(I-1) \times (J-1)$ où $I = 3$ et $J = 2$ sont respectivement les nombres de colonnes et de lignes.

La régression logistique avec 2 explicatives catégorielles

Ex Etude de l'impact de l'inhalation de la fumée de tabac brun ou autre (blond ou mélange) sur le risque de développer un cancer de la vessie (Clavel et al., 1989, Int. Jr. Cancer, 44: 605-610).

		Cancer de la vessie	
Inhalation	Tabac	Oui	Non
Non	Autre	23	34
	Brun	86	119
Oui	Autre	32	39
	Brun	267	134

• Notation: π_{jk}: probabilité d'avoir le cancer de la vessie (CV) pour une inhalation de type j (1: non ; 2: oui) et du tabac k (1: autre ; 2: brun).

• Le modèle avec interaction s'écrit

$$\text{logit}(\pi_{jk}) = \log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \theta_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk} \quad ; \quad (\beta_1 = \alpha_1 = \gamma_{j1} = \gamma_{1k} = 0)$$

• On a donc la structure suivante pour les logarithmes de cotes, $\text{logit}(\pi_{jk})$:

	$k=1$ (T-A)	$k=2$ (T-B)
$j=1$ (I-N)	μ	$\mu + \beta_2$
$j=2$ (I-O)	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_2 + \gamma_{22}$

La régression logistique avec 2 explicatives catégorielles (2)

• Interprétation des 4 coefficients (dans notre exemple):

$\rightarrow$ μ: logarithme de la cote du cancer de la vessie (CV) pour un non-inhaleur (I-N) de tabac autre (T-A).

$\rightarrow$ α_2: log. du rapport des cotes (RC) du CV d'un I-O versus un I-N avec un T-A.

$\rightarrow$ $\alpha_2 + \gamma_{22}$: log. du RC du CV d'un I-O versus un I-N avec un T-B.

$\rightarrow$ β_2: log. du RC du CV avec un T-B versus un T-A pour un I-N.

$\rightarrow$ $\beta_2 + \gamma_{22}$: logarithme du RC du CV avec un T-B versus un T-A pour un I-O.

Remarque

• Lorsque $\gamma_{22} = 0$,

$\rightarrow$ le log. du RC du CV de I-0 versus I-N ne dépend pas du type de tabac.

$\rightarrow$ le log. du RC du CV de T-B versus T-A ne dépend pas du rejet ou de l'inhalation de la fumée.

Par conséquent, γ_{22} quantifie l'interaction entre les 2 facteurs dans leurs effets sur le log. du RC du CV.

La régression logistique avec 2 explicatives catégorielles (3)

Cancer - Paramètres estimés (vessie.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-0,390866	0,269982	2,095986	0,147686
Inhale	O	2	0,193041	0,360251	0,287135	0,592062
Inhale	N	3	0,000000			
Tabac	B	4	0,066090	0,304830	0,047006	0,828357
Tabac	A	5	0,000000			
Inhale*Tabac	1	6	0,821144	0,401273	4,187535	0,040722
Inhale*Tabac	2	7	0,000000			
Inhale*Tabac	3	8	0,000000			
Inhale*Tabac	4	9	0,000000			
Echelle			1,000000	0,000000		

Cancer - Test Vraisemblance Type 1 (vess
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-504,180		
Inhale	1	-488,034	32,29146	0,000000
Tabac	1	-484,281	7,50672	0,006147
Inhale*Tabac	1	-482,209	4,14356	0,041793

• De cette sortie logicielle, on en déduit que $\hat{\mu} = -0.39$, $\hat{\alpha}_2 = 0.19$, $\hat{\beta}_2 = 0.066$ et $\hat{\gamma}_{22} = 0.82$. L'interprétation des coefficients a deja été détaillée.

• Un intervalle de crédibilité 95% pour γ_{22} est $0.82 \pm 1.96 \times 0.401 = (0.03, 1.61)$: zero ne fait donc pas partie des valeurs plausibles pour ce paramètre. Cela suggère qu'il existe une interaction entre inhalation et tabac: l'effet de l'inhalation sur le risque de développer un cancer de la vessie dépend du tabac consommé.

• Le caractère non-significatif de l'interaction est confirmé par la statistique $G^2 = 4.14 > \chi^2_1(0.95) = 3.84$ (P-valeur= 0.04 <5%)

La régression logisitique avec 2 explicatives catégorielles (4)

• Détaillons cet effet pour chacun des 2 tabacs au travers du log du rapports des cotes des inhaleurs versus les non-inhaleurs:

→ T-A: α_2: $0.19 \pm 1.96 \times 0.360 = (-0.52, 0.90)$.

→ T-B: $\alpha_2 + \gamma_{22} = 0.19 + 0.82 = 1.01$.

Alors que nous n'avons pas d'indication que la cote du cancer de la vessie augmente avec l'inhalation de la fumée d'un tabac autre, nous avons une augmentation significative de cette cote lorsque l'inhalation concerne du tabac brun.

• De la même manière, évaluons l'effet du tabac pour chaque type de fumeur au travers du log du rapports des cotes (du CV) du tabac brun versus le tabac autre:

→ I-N: β_2: $0.07 \pm 1.96 \times 0.304 = (-0.56, 0.64)$.

→ I-0: $\beta_2 + \gamma_{22} = 0.07 + 0.82 = 0.89$.

Alors que nous n'avons pas d'indication que la cote du cancer de la vessie change d'un tabac à un autre lorsque la fumée n'est pas inhalée, nous avons une cote significativement supérieure avec du tabac brun lorsqu'il y a inhalation.

Sortie d'un logiciel bayesien *

```
Iterations = 1001:11000  Thinning interval = 1  Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,  plus standard error of the mean:

                   Mean     SD Naive SE Time-series SE
(Intercept)    -0.39134 0.2707 0.002707        0.01019
InhOui          0.19445 0.3611 0.003611        0.01553
TabBrun         0.07102 0.3103 0.003103        0.01109
InhOui:TabBrun  0.81841 0.4045 0.004045        0.01673

2. Quantiles for each variable:

                  2.5%      25%     50%     75%  97.5%
(Intercept)    -0.9551 -0.56541 -0.3799 -0.2053 0.1381
InhOui         -0.5029 -0.05022  0.1967  0.4188 0.9631
TabBrun        -0.5307 -0.13474  0.0699  0.2677 0.6936
InhOui:TabBrun  0.0043  0.54622  0.8211  1.0918 1.5905
```

• Ce logiciel travaille sous le paradigme bayesien: il ne repose pas sur des approximations normales des distributions a posteriori (qui s'avèrent être des distributions d'échantillonnage d'estimateurs dans l'approche fréquentiste).

Les résultats générés sont donc également fiables en petit échantillon.

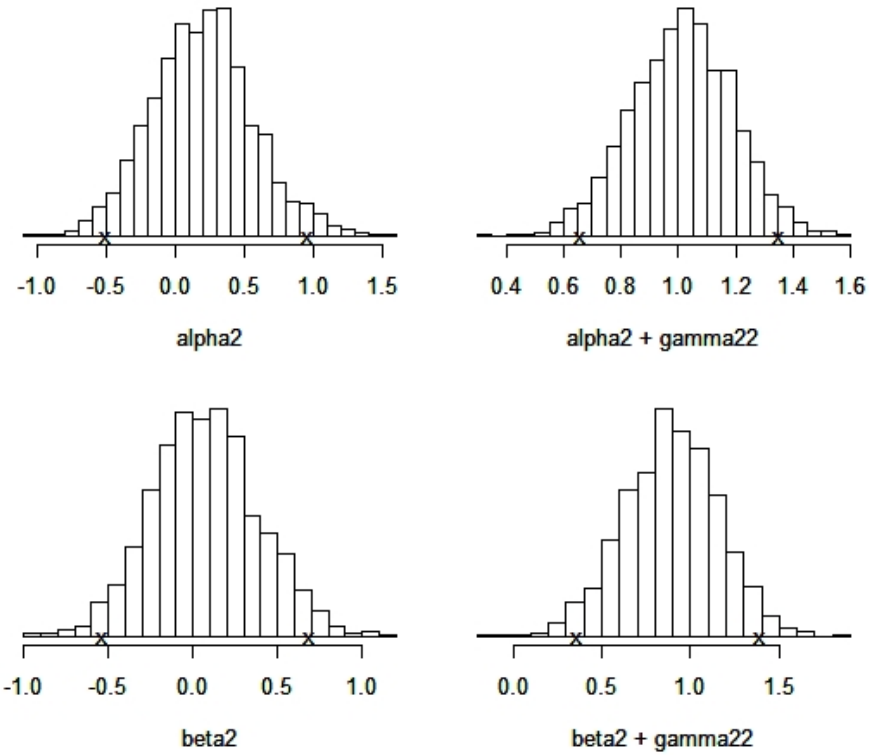
• Il permet d'étudier, sans effort supplémentaire, l'a posteriori de n'importe quelle fonction des paramètres.

Ex On peut facilement visualiser l'a posteriori et estimer un intervalle de crédibilité 95% pour les rapports de cotes évoqués ci-dessus:

$\rightarrow$ $\alpha_2 + \gamma_{22}$: $(0.66, 1.35)$.

$\rightarrow$ $\beta_2 + \gamma_{22}$: $(0.36, 1.39)$.

Tout ceci confirme l'augmentation de risque lorsque inhalation et tabac brun sont combinés.



La régression logistique avec une explicative continue

• Lorsque la variable explicative est continue, deux attitudes sont possibles:
-1- catégoriser la variable continue sur base de critères contextuels ou des quantiles observés pour cette variable.
Ex Si on souhaite expliquer comment la probabilité d'être un chômeur de longue durée change avec le nombre d'années d'études, on peut très bien décider de catégoriser cette variable par -1- études primaires, -2- au plus 3 ans après le primaire, -3- au plus 6 ans après le primaire, -4- supérieur de type court, -5- supérieur de type long non universitaire, -6- diplôme universitaire.

-2- exploiter le caractère continu de la variable continue dans la modélisation.

• On sait comment traiter les données de l'approche -1- une fois que la catégorisation a été réalisée. Nous allons donc nous concentrer sur la 2ème approche.

La régression logistique avec une explicative continue (2)

• L'European Social Survey (année 2002 pour la Belgique) demande de réagir à l'affirmation suivante (item B31): "Les homosexuels hommes et femmes devraient être libres de vivre leur vie comme ils le souhaitent." Les réponses possibles (exprimant une opinion) sont 1=Tout à fait d'accord; 2=Plutôt d'accord; 3=Ni d'accord, ni en désaccord; 4=Plutôt en désaccord; 5=Tout à fait en désaccord.

• Nous allons nous intéresser à la probabilité d'être "Tout à fait en désaccord" et étudier comment elle change avec l'âge (en années) et le sexe des 1764 répondants (parmi les 1778 interviewés).

Tout à fait en désaccord	
Non	Oui
1657	93

Femme	Homme
896	868

Age	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,100)
n_k	204	274	314	328	266	219	123	36

La régression logistique avec une explicative continue (3)

• La première chose à faire est de visualiser l'évolution du logarithme de la cote avec les variables explicatives.

	Tout à fait en désaccord	
Age	Non	Oui
(10,20)	197	7
(20,30)	266	8
(30,40)	302	12
(40,50)	312	16
(50,60)	254	12
(60,70)	203	16
(70,80)	106	17
(80,100)	30	6

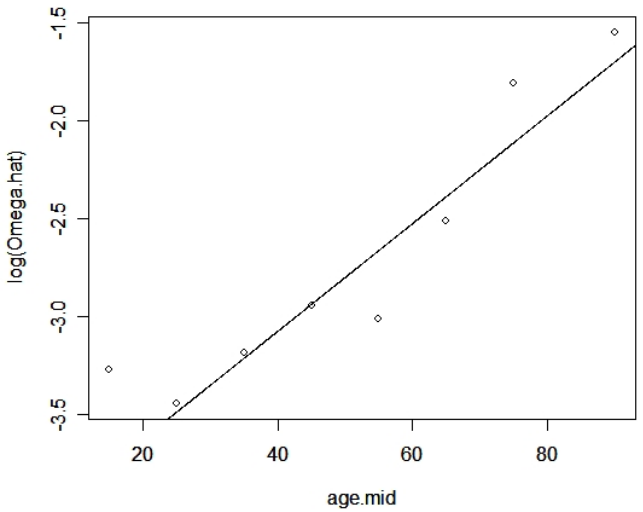
• A cette fin, on calcule le logarithme de la cote empirique

$$\tilde{\Omega}_x = \frac{y_x + 0.5}{n_x - y_x + 0.5}$$

où y_x cas ont été observés parmi n_x personnes pour une valeur x de la variable explicative.

La régression logistique avec une explicative continue (4)

• Cependant, comme la variable explicative continue d'intérêt est l'âge en années, le nombre de valeurs différentes observées est grand et ne permet pas une analyse descriptive satisfaisante. A cette fin (uniquement), nous allons catégoriser la variable âge en travaillant par dizaine d'années.



• Voici un graphe de dispersion visualisant le lien entre $\log \tilde{\Omega}$ et le score.

• Le logarithme de la cote semble augmenter linéairement avec l'âge. Cela suggère le modèle de régression logistique

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

β_0: logarithme de la cote du succès lorsque $X = 0$.
β_1: logarithme du rapport des cotes du succès de $X = x + 1$ versus $X = x$.

La régression logistique avec une explicative continue (5)

Attitude - Paramètres estimés (homo.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-4,17399	0,318122	172,1537	0,000000
Age		2	0,02752	0,005852	22,1098	0,000003
Echelle			1,00000	0,000000		

Attitude - Test Vraisemblance Type 1 (hom
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-367,062		
Age	1	-355,732	22,66039	0,000002

• Comme précédemment, les a priori pour les paramètres sont non-informatifs. Une approximation normale est faite pour les a posteriori des paramètres impliqués.

• On en déduit les modes a posteriori $\hat{\beta}_0 = -4.17$ et $\hat{\beta}_1 = 0.0275$ avec erreurs standards respectives 0.318 et 0.0059.

• Un intervalle de crédibilité 95% approximatif pour β_1 est donné par

$$0.0275 \pm 1.96 \times 0.0059 = (0.0160, 0.0390) \ .$$

Comme 0 ne fait partie des valeurs plausibles, on en déduit qu'il est très vraisemblable que la cote du succès change (ici: augmente) avec l'âge.

• Cela est confirmé par la statistique $G^2 = 22.66 > \chi^2_1(0.95) = 3.84$ (P-valeur= 0.000002 <5%)

La régression logistique avec une explicative continue (6)

Sortie d'un logiciel bayesien *

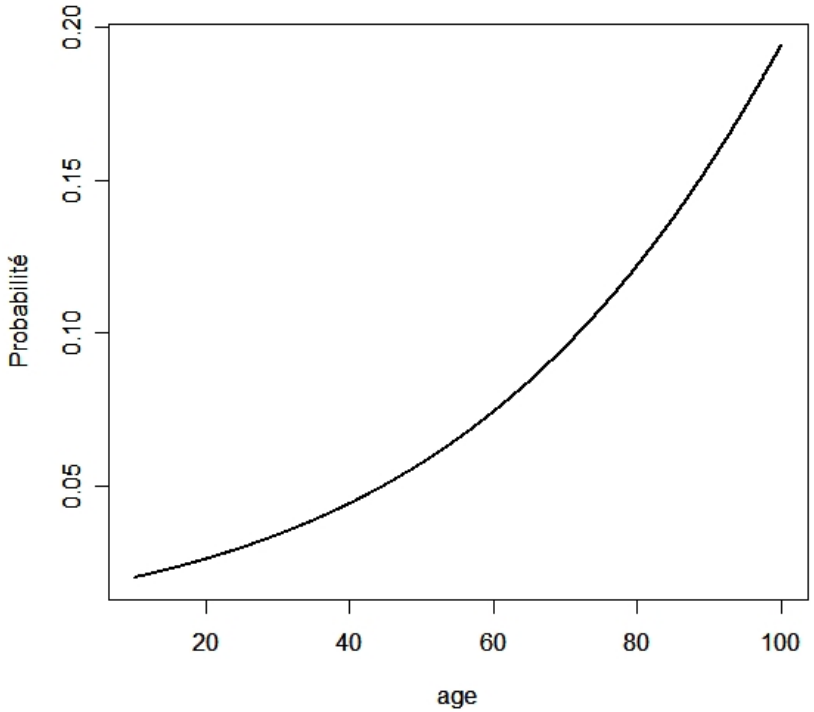
```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD  Naive SE Time-series SE
(Intercept) -4.18703 0.320126 3.201e-03      0.0085346
age          0.02767 0.005846 5.846e-05      0.0001531

2. Quantiles for each variable:

                2.5%      25%      50%      75%    97.5%
(Intercept) -4.83688 -4.40498 -4.17869 -3.96474 -3.60913
age          0.01654  0.02353  0.02756  0.03178  0.03909
```

La régression logistique avec une explicative continue (7)

• On peut facilement déduire une estimation de la plausibilité du succès pour une valeur de x donnée:

$$\pi_x = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

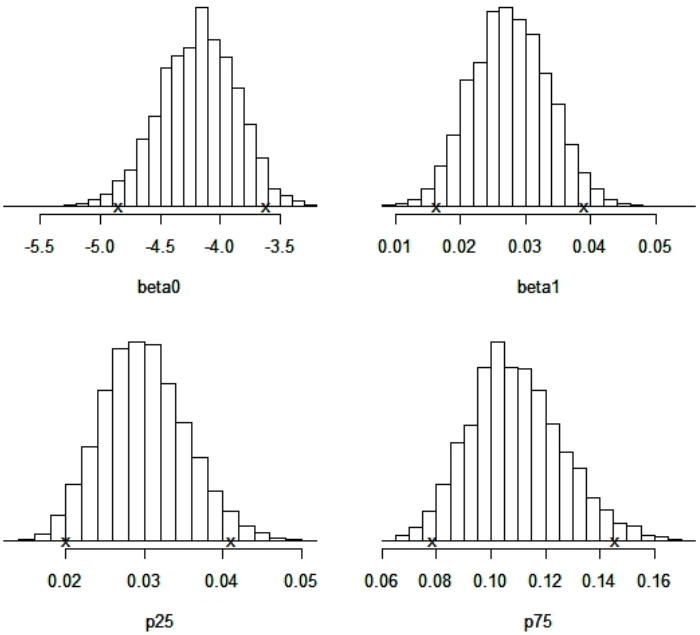
Ex Pour 25 et 75 ans, les valeurs les plus plausibles pour ces probabilités sont $\hat{\pi}_{25} = 0.03$ et $\hat{\pi}_{75} = 0.11$.



• Il est déconseillé d'utiliser cette procédure pour une valeur de x qui n'appartient pas à l'intervalle des valeurs observées pour X: le comportement du logarithme de la cote n'est peut-être pas linéaire en dehors de cet intervalle.

La régression logistique avec une explicative continue (8)

• Avec l'approche bayesienne, on peut utiliser le très grand nombre M (=10000 par ex.) de paires $\{(\beta_0^{(m)}, \beta_1^{(m)}) : m = 1, \ldots, M\}$ générées durant la procédure d'exploration des distributions a posteriori pour en déduire les valeurs de $\pi_{25}^{(m)}$ et $\pi_{75}^{(m)}$ et ainsi visualiser les valeurs plausibles pour ces quantités:

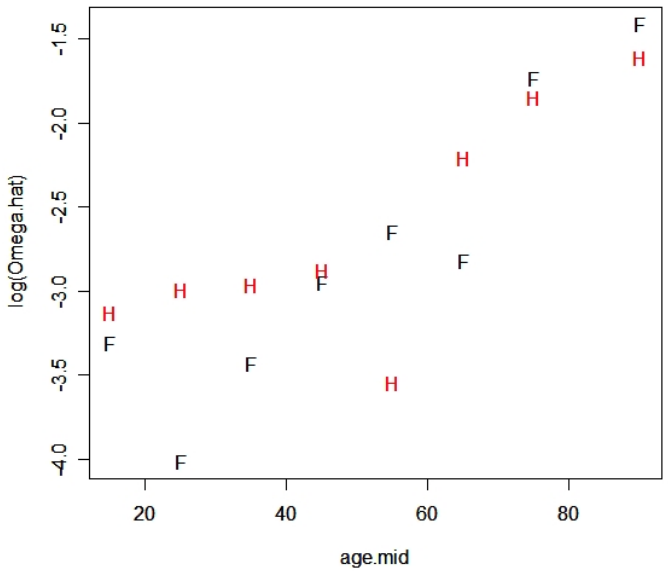


• C'est une nouvelle illustration de la puissance de l'approche bayesienne qui permet, sans grande difficulté, de visualiser les valeurs plausibles de n'importe quelle fonction des paramètres sur lesquels l'inférence a porté.

La régression logistique avec explicatives mixtes

- Utilisons également la variable sexe dans le modèle. Les données sont

	Femme		Homme	
Age	Non	Oui	Non	Oui
(10,20)	95	3	102	4
(20,30)	137	2	129	6
(30,40)	138	4	164	8
(40,50)	162	8	150	8
(50,60)	133	9	121	3
(60,70)	108	6	95	10
(70,80)	59	10	47	7
(80,100)	18	4	12	2



- Le modèle de régression logistique devient:

$$\begin{aligned} \text{logit}(\pi_x) &= \beta_0 + \beta_1 x \quad \text{pour les femmes} \\ \text{logit}(\pi_x) &= (\beta_0 + \alpha_2) + (\beta_1 + \tau_2)x \quad \text{pour les hommes} \end{aligned}$$

La régression logistique avec explicatives mixtes (2)

Attitude - Paramètres estimés (ESS2002BEbis.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-4,60942	0,486108	89,91388	0,000000
Sexe	Homme	2	0,80923	0,644324	1,57737	0,209140
Sexe	Femme	3	0,00000			
Age		4	0,03445	0,008504	16,41573	0,000051
Sexe*Age	1	5	-0,01317	0,011809	1,24436	0,264633
Sexe*Age	2	6	0,00000			
Echelle			1,00000	0,000000		

• Les estimations des modes a posteriori (erreurs standards) sont

$$\hat{\beta}_0 = -4.61\ (0.486)\ ;\ \hat{\alpha}_2 = 0.81\ (0.644)\ ;\ \hat{\beta}_1 = 0.034\ (0.0085)\ ;\ \hat{\tau}_2 = -0.013\ (0.0118)$$

• 0 fait partie des valeurs plausibles pour τ_2: il n'y a donc pas d'indication qu'il existe une interaction entre âge et sexe dans leurs effets sur (le log. de) la cote du succès.

• 0 fait partie des valeurs plausibles pour α_2: on conclut qu'il n'y a pas d'indication que sexe soit une variable utile pour expliquer (le logarithme de) la cote du succès.

La régression logistique avec explicatives mixtes (3)

• Les constatations précédentes, reposant sur des approximations normales des distributions a posteriori, sont confirmées par des tests de rapport de vraisemblances réputés plus fiables:

Attitude - Test Vraisemblance Type 1 (ES
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-367,062		
Sexe	1	-366,993	0,13704	0,711237
Age	1	-355,537	22,91228	0,000002
Sexe*Age	1	-354,912	1,25118	0,263327

Attitude - Test Vraisemblance Type 1 (ES
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-367,062		
Age	1	-355,732	22,66039	0,000002
Sexe	1	-355,537	0,38893	0,532861
Sexe*Age	1	-354,912	1,25118	0,263327

• Le tableau de gauche évalue d'abord le retrait de l'interaction ($\tau_2 = 0$): la P-valeur (0.26) confirme que 0 est bien une valeur plausible pour τ_2. Ensuite, le retrait supplémentaire de Age ($\beta_1 = 0$) dans le modèle sans interaction est évalué: la P-valeur (0.000002) indique que 0 n'est pas une valeur plausible pour β_1.

• Le tableau de droite évalue d'abord le retrait de l'interaction ($\tau_2 = 0$): la P-valeur (0.26) indique que 0 est bien une valeur plausible pour τ_2. Ensuite, le retrait supplémentaire de Sexe ($\alpha_2 = 0$) dans le modèle sans interaction est évalué: la P-valeur (0.53) indique que 0 est une valeur plausible pour α_2.

La régression logistique avec explicatives mixtes (4)

• En conclusion, seul l'âge apparaît comme explicative de la cote du succès. Il n'y a pas d'indication que le sexe apporte une information complémentaire utile.

La régression logistique avec explicatives mixtes (5)

Sortie d'un logiciel bayesien *

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                          Mean       SD  Naive SE Time-series SE
(Intercept)            -4.63166 0.473840 4.738e-03      0.0181756
sexeHomme               0.81797 0.625171 6.252e-03      0.0267469
age                     0.03455 0.008276 8.276e-05      0.0003013
sexeHomme:age          -0.01321 0.011423 1.142e-04      0.0004733

2. Quantiles for each variable:

                          2.5%      25%      50%       75%     97.5%
(Intercept)            -5.58789 -4.95727 -4.60776 -4.297692 -3.755277
sexeHomme              -0.38165  0.39981  0.79083  1.252152  2.084676
age                     0.01883  0.02877  0.03441  0.039988  0.050860
sexeHomme:age          -0.03568 -0.02105 -0.01308 -0.005441  0.008413
```

La régression logistique: une extension utile

• Lorsqu'une variable explicative est continue, il est essentiel de s'assurer que sa relation avec le logarithme de la cote du succès est linéaire.

Il est possible de s'affranchir de cette vérification en travaillant avec une spécification plus générale.

• Cette extension suppose que

$$\text{logit}(\pi) = \beta_0 + f_1(x_1) + \ldots f_K(x_K)$$

où $f_k(x_k)$ est une fonction "lisse" de x_k à estimer et où les effets des différentes variables $x_1, \ldots, x_K$ apparaissent de manière additive.

C'est un cas particulier du *modèle linéaire généralisé additif* (GAM=generalized additive model).