Données binaires – introduction à la régression logistique

Emmanuelle Gilot-Fromont

Janvier 2024

Introduction

Présence / absence d'un caractère:

À l'échelle individuelle:
Probabilité d'être gestante ou non, male ou femelle…
Risque pour une personne de développer un cancer au contact du tabac, de la poussière…

A l'échelle populationnelle:
Proportion de male ou femelle…

- Variable à expliquer: présence / absence
- Variables explicatives: facteurs de risque: exposition au tabac…

VetAgro Sup

Les données

Ex: cancer et exposition professionnelle à la poussière

Données individuelles:

	exposé1	**exposé2**	**malade**
Individu 1	non/oui	0/1	0/1
Individu 2	non/oui	0/1	0/1
….			

Données groupées:

	malade	sain	
non exposé	a	b	a+b
exposé	c	d	c+d

Les données

	malade	sain	
non exposé	a	b	a+b
exposé	c	d	c+d

Risque(non exposé) = P(M/NE) = a/(a+b)
Estimé par la proportion de malades parmi les non exposés

Risque (exposé) = c/(c+d)

Risque relatif (exposé/non exposé) =
RR = [P(M/E)] / [P(M/NE)] = [c/(c+d)] / [a/(a+b)]

VetAgro Sup

Les protocoles

	malade	sain	
non exposé	a	b	a+b
exposé	c	d	c+d

Comparaison entre exposés et non exposés (études de cohorte, cohorte exposée non exposée): estimation des risques et du RR

Comparaison entre malades et non malades (études cas-témoins): pas d'estimation du risque ni du RR

Odds, odds-ratios

- Odds: a/b , c/d = P(M/x) / P(S/x) = p/1-p = odds = cote. Proche du risque lorsque la maladie est rare

- [c/d] / [a/b] = odds-ratio (OR) = rapport des chances: un indicateur symétrique du niveau de lien entre les deux variables de la table

	M	S
non Exp	a = 1	b = 9
Exp	c = 1	d = 4

Non exposés: risque = 1/10, odds = 1/9
exposés: risque = 1/5, odds = 1/4
Risque relatif = (1/5)/(1/10) = 2
OR = (1/4)/(1/9) = 2,25

VetAgro Sup

OR et RR

- L'OR est symétrique lignes/colonnes:
  [c/d] / [a/b] = [c/a] / [d/b]
  = bc/ad

	malade	sain
non expo	a	b
expo	c	d

Estime le risque des exposés/ non-exposés (d'être malade)
Et le risque des malades/non-malades (d'être exposé)

- S'utilise aussi bien dans les études cas-témoin que dans les études exposé-non-exposé

OR surestime RR, d'autant moins que le caractère est rare
(caractère rare = 1-p proche de 1 donc OR proche de RR)

VetAgro Sup

Confusion et OR ajustés, interaction

Exemple: on teste l'effet de l'exposition à la poussière, mais:

- Si les individus exposés sont aussi plus souvent fumeurs: confusion

- Si l'effet de la poussière est différent selon qu'on est fumeur ou non fumeur: interaction

Nécessité d'un modèle explicatif du risque et du risque relatif, ou de l'odds et de l'OR, de la forme $Y = b_0 + \Sigma b_i X_i$

VetAgro Sup

Le modèle logistique

1. Ecriture et dimensions
2. Ajustement
3. Interprétation
4. Test de l'effet d'une variable
5. Exemple

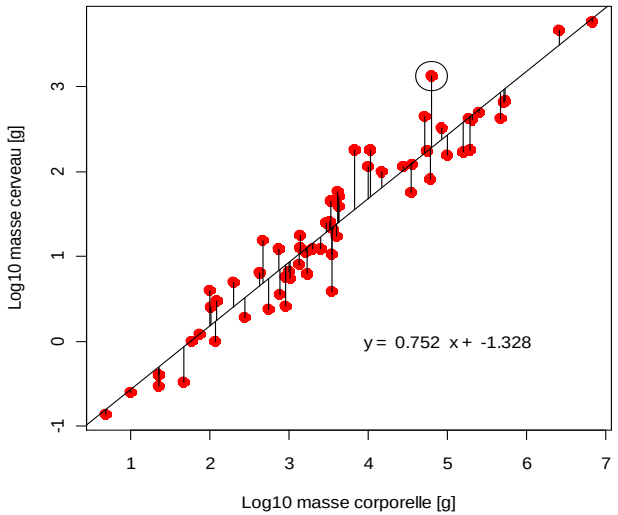
$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

VetAgro Sup

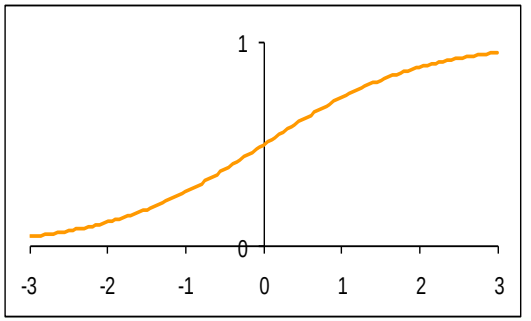
1. Ecriture et dimensions

Rappel: modèle linéaire simple:

$$Y = \beta_0 + \Sigma\beta_i X_i + \varepsilon$$



Ici Y = probabilité =>
Comprise entre 0 et 1,
Pas d'homoscédasticité
Distribution non normale des résidus



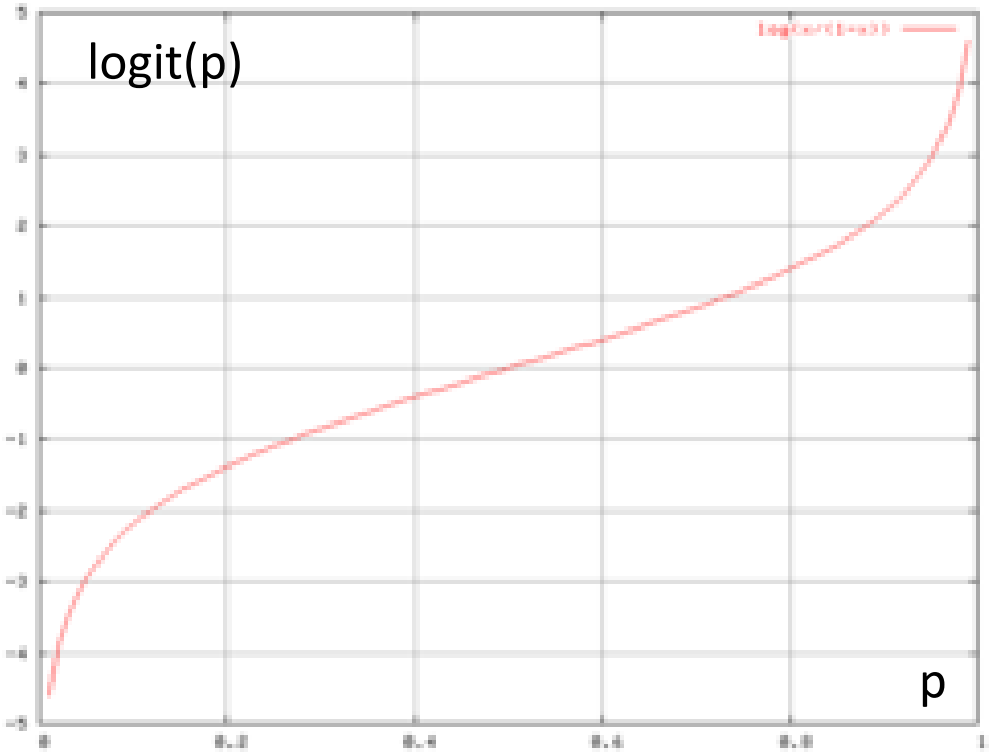
VetAgro Sup

transformation logit

Une transformation possible:

Y = ln (p / 1-p)
= logit (p)

Transformation = fonction de lien



Fonction logit

VetAgro Sup

Fonction logistique

logit p = $\beta_0 + \Sigma\beta_i X_i$ = variable explicative z

=> p/(1-p) = $e^{\beta 0 + \Sigma\beta i Xi}$

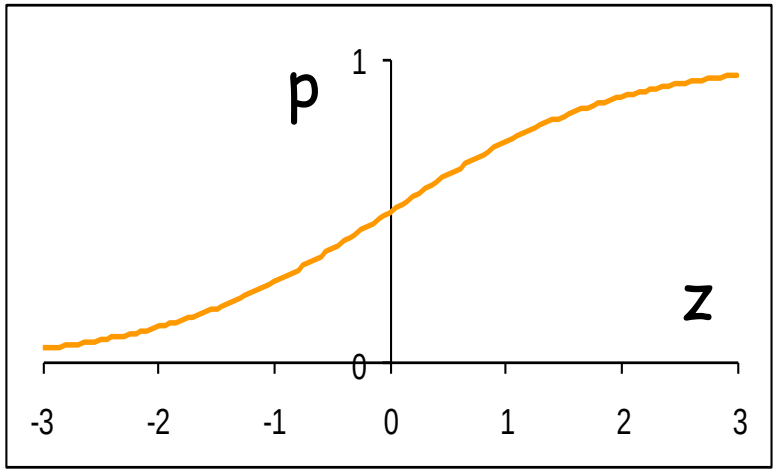
=> p/(1-p) = $e^{\beta 0 + \Sigma\beta i Xi}$

=> ...

=> $$p = \frac{e^{\beta_0 + \sum \beta_i X_i}}{1 + e^{\beta_0 + \sum \beta_i X_i}}$$

= logistique ($\beta_0 + \Sigma\beta_i X_i$)

Fonction réaliste pour p



Fonction logistique

VetAgro Sup

Dimensions du modèle logistique

$\ln (p/(1-p)) = \beta_0 + \Sigma\beta_i X_i$

$\Rightarrow \ln (odds) = \beta_0 + \Sigma\beta_i X_i$

$\Rightarrow odds = e^{\beta 0} e^{\Sigma\beta i Xi}$

Nature des X_i?

Interprétation des β_i ?

Si modèle $odds = e^{\beta 0} e^{\beta 1 X1}$, avec X_1: exposé (indicatrice)

odds (non-exposés) = $e^{\beta 0}$ => β_0 est le ln de l'odds de la catégorie où $X_i = 0$ (de référence = non expo)

odds (expo) = $e^{\beta 0} e^{\beta 1}$

$\Rightarrow e^{\beta 1}$ = odds (expo)/ odds (non expo) = OR expo

=> β_1= ln(OR(expo/non))

VetAgro Sup

Les variables explicatives Xi et les OR

Qualitative à 2 modalités (présence/absence): 1 X_i indicatrice

ex : logit p (malade) = $\beta_o + \beta_1 X$ X : non expo (0) / expo (1)

β_1 : OR des exposés / non exposés (contraste)

Continue :

ex : logit p (malade) = $\beta_o + \beta_1$. Nombre cigarettes/jour

β_1 : OR par cigarette supplémentaire / jour

VetAgro Sup

Les variables explicatives Xi et les OR

Variable qualitative à k modalités: k-1 X_i indicatrices

ex : logit p (malade) = $\beta_o+\beta_1X_1 +\beta_2X_2 +\beta_3X_3$

X_1 : peu expo, X_2 : moyennement, X_3 : très exposé

(non exposé: $X_1 = X_2 = X_3 = 0$)

β_1 : OR des exposés (peu, moyen ou très) / non exposés (contraste)

Interaction: ajout d'indicatrices

ex : logit p(malade) = $\beta_o+\beta_1X_1 +\beta_2X_2 +\beta_3X_3$

X_1 : exposé au tabac, X_2 : exposé à la poussière, X_3 : interaction

β_1 : OR des exposés conjointement / exposés à l'un et à l'autre

VetAgro Sup

[Lien avec le modèle linéaire simple

Permet de

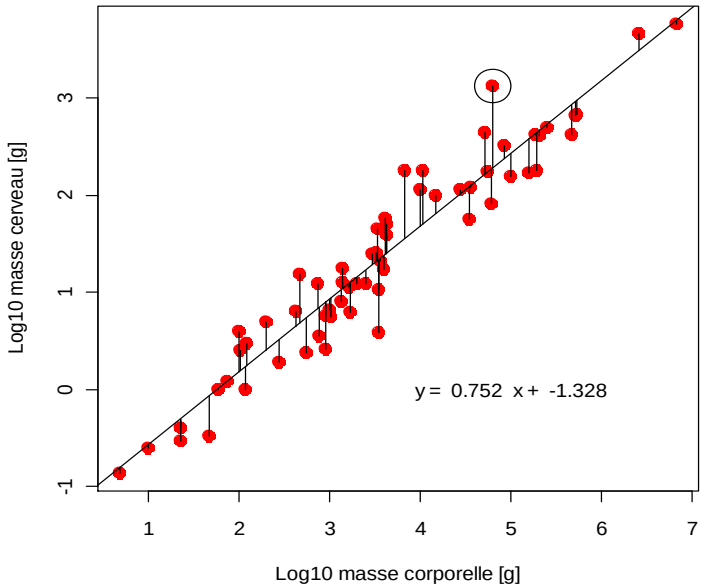
- Estimer les paramètres β_i

- Tester l'effet d'une variable X_i (= pente β_i)

- Sélectionner des variables et interactions explicatives pour Y

- Elaborer/choisir un modèle multivarié

- Estimer la part de variabilité de Y expliquée par le modèle : R^2

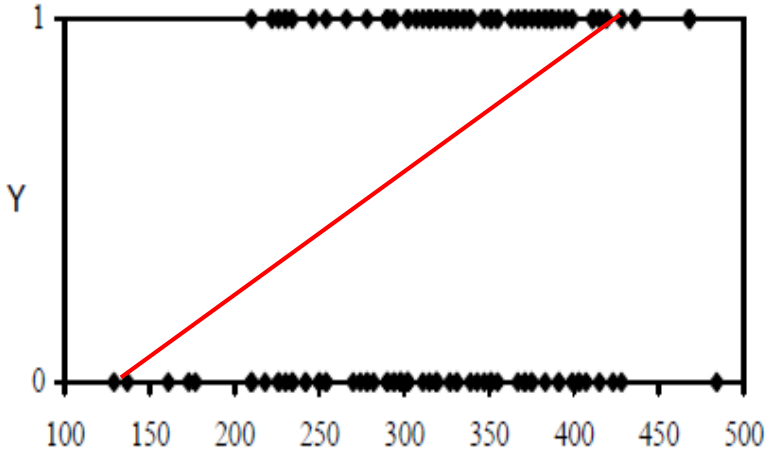


$$Y = \beta_0 + \Sigma\beta_i X_i + \varepsilon$$

VetAgro Sup

Réponse présence/absence

Variable réponse (à expliquer) = 0 ou 1, succès/échec, mâle/femelle, malade/non-malade = qualitative à deux modalités



Pas de linéarité

Variance maximale au centre
Erreurs non indépendantes

Exemple: gestation = f(masse)

Modèle linéaire simple impossible: recherche d'un autre modèle

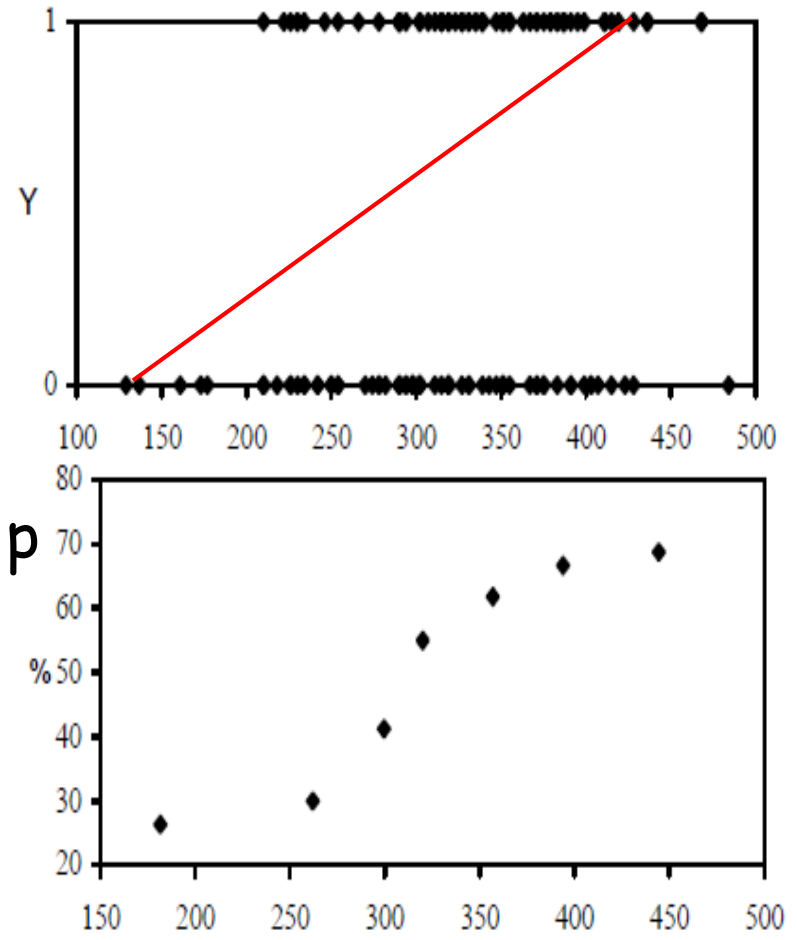
Présence/absence et probabilité

Ce qu'on peut expliquer:
Pour chaque valeur (classe) de X,
proportion de réponse = 1

Ou pour un individu :
Probabilité d'avoir 1

= p

Relation entre X et p non linéaire :
min = 0, max = 1



VetAgro Sup

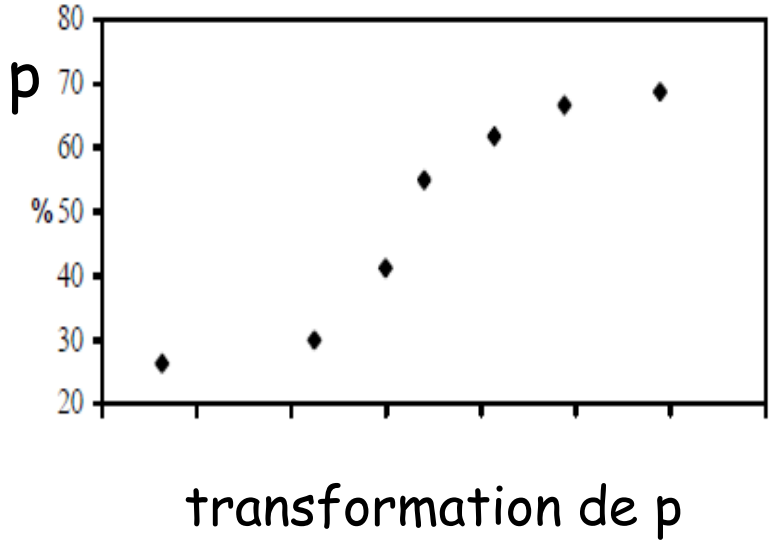
Fonction de lien

Idée = chercher une transformation de p avec min = 0, max = 1, allure sigmoide

Permettant d'écrire

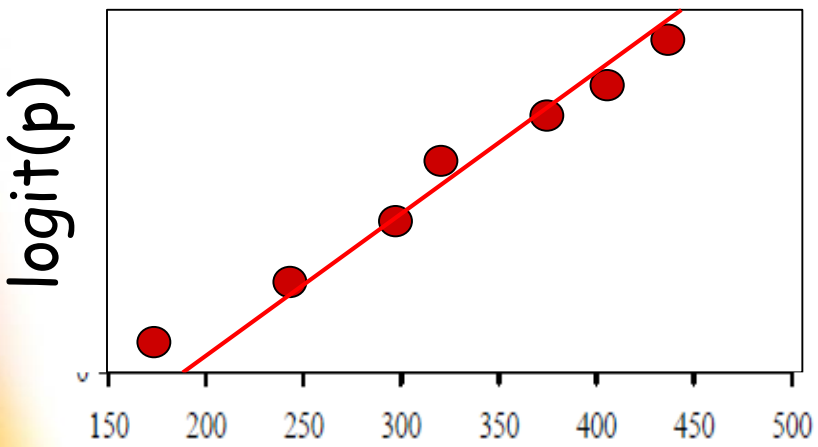
$f(p) = \beta_0 + \beta_1 X_1 + \ldots\ldots + \varepsilon$

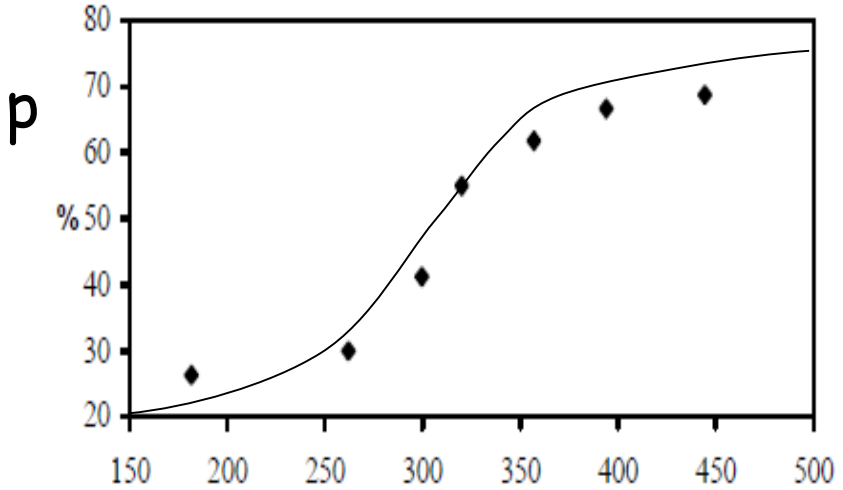
(pour retrouver les usages et propriétés du modèle linéaire simple)
= fonction de lien



VetAgro Sup

Fonction de lien logit





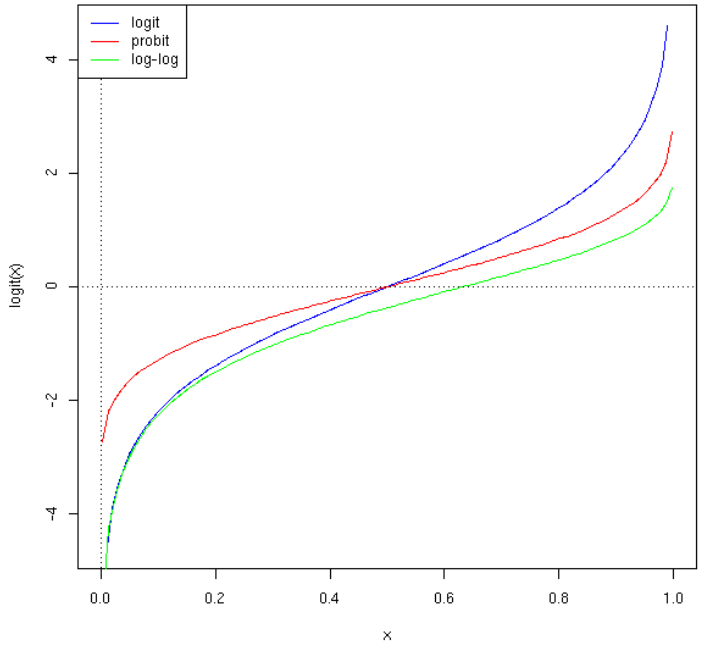
$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \ldots\ldots + \varepsilon \quad <=> \quad p = \text{logistique}(\beta_0 + \beta_1 X_1 + \ldots\ldots + \varepsilon)$$

Y = transformée de la variable à expliquer

VetAgro Sup

Autres modèles linéaires généralisés]

- Réponse qualitative à 2 modalités (0/1), autres fonctions de lien:
  - f = probit (p) = inverse de la fonction de distribution cumulative de la loi normale
  - f = « complementary log-log » (p) = ln(-ln(1-p))



- Réponse qualitative à plus de deux modalités: modèle logistique généralisé (risques adjacents, séquentiel, cumulatif)

- Réponse = énumération
  - f = ln(nombre) = modèle log-linéaire

Le modèle linéaire généralisé s'applique à ces trois types de variables

VetAgro Sup

2. Ajustement

Ajuster = estimer les paramètres β_i : technique et sorties R

- Méthodes: moindres carrés, …
  Ici: maximum de vraisemblance (maximum likelihood)
  la fonction de vraisemblance :

$$\mathcal{L}(y_1, y_2, \ldots, y_n ; \beta_i) = \Pi\, P(y_i / X_i)$$

- $\mathcal{L}$ mesure la probabilité d'obtenir les valeurs observées y_i, sachant les valeurs de paramètres β_i

  Principe de l'ajustement : maximiser $\mathcal{L}$

Maximum de vraisemblance

- Déterminer la fonction de vraisemblance $\mathcal{L}$ puis étudier ses variations en fonction des valeurs de paramètres
  la fonction est maximale lorsque sa dérivée s'annule (et sa dérivée seconde est négative):

$$\frac{d(\ln(L(\beta))}{d\beta} = \sum_{i=1}^{n} \frac{d(\ln(P(x_i;\beta))}{d\beta}$$

- Estimateurs au ML:
  - variance minimum ou asymptotiquement de variance minimum
  - de distribution normale => possibilité de tests
  - …pas toujours sans biais

VetAgro Sup

Vraisemblance, déviance

- L'ajustement produit des valeurs d'estimations ponctuelles pour les β_i, des « erreurs standard » (s.e. = erreur-type = écart-type de l'estimateur)

  + une mesure de l'ajustement :

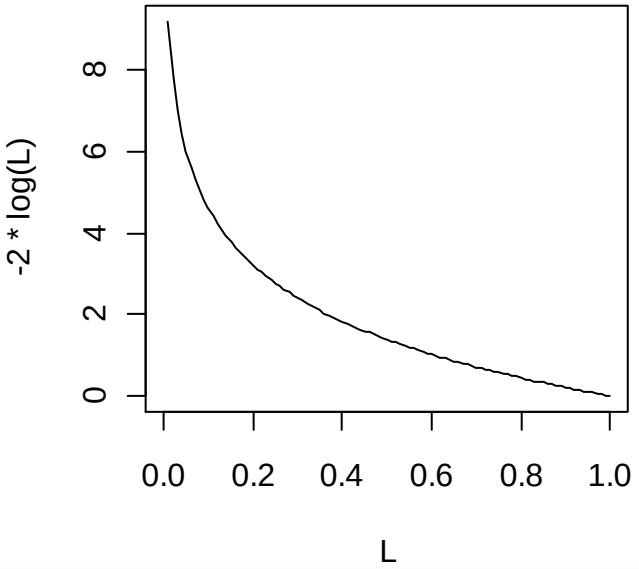
- Vraisemblance $\mathcal{L}$ (likelihood)

  $\mathcal{L}$ « proche de 1 » = bon ajustement (mais toujours compris entre 0 et 1, difficile à interpréter)

- Déviance = -2ln($\mathcal{L}$)

  Bon ajustement = faible déviance

  « Déviance résiduelle »



VetAgro Sup

3. interprétation du modèle

Exemple:

```
Coefficients:
             Estimate
(Intercept)   -1.7458   = β0
expos          0.7569   = β1
```

ln (p/(1-p)) = -1.7458 + 0.7569 * **Expo**

- Chez les non-exposés : ln (odds) = -1,7458
  Pour les exposés (par rapport aux non-exposés) : ln(OR) = 0,7569

- Soit OR(exposition) = exp(0,7569) = 2,13
  Les exposés ont un OR de 2,13 par rapport aux non exposés

Interprétation du modèle

Plus généralement:

```
Coefficients:
             Estimate
(Intercept)   -1.7458   = β0
X              0.7569   = β1
```

$\ln(p/(1-p)) = \beta_0 + \beta_1 * \mathbf{X}$

- Si X qualitative: β_1 = contraste entre la modalité de référence (par ordre alphanumérique, par ex expo = « non ») et la modalité de la ligne (expo = « oui »); plusieurs paramètres si expo > 2 modalités
- Si X quantitative: β_1 = pente = effet de l'augmentation de X de 1 unité (exemple: expo passe de 0 à 1 ou de 3 à 4)

Intervalle de confiance de l'OR

Exemple: cancer:

```
Coefficients:
             Estimate Std. Error
(Intercept)  -1.7458     0.1227
Expos         0.7569     0.1622
```

$\beta_1 = 0{,}7569 \pm 1{,}96 * 0{,}1622$

- $IC(\beta_1) = [0.44 - 1.08]$
- Donc IC(OR(exposition)) = [1,55 – 2,92]
  Ne contient pas la valeur 1 donc effet significatif

Test d'un contraste: test de Wald

Exemple: risque de cancer: « summary »:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7458     0.1227 -14.227  < 2e-16 ***
expos         0.7569     0.1622   4.667 3.06e-06 ***
```

Sous Ho : $\beta_1 = 0$,

$$z = \beta_1/\sigma(\beta_1) : \mathcal{N}(0,1)$$

- Ici, z (effet exposition) = 0,7569/0,1622 = 4,667: à comparer avec l'écart-réduit d'une loi normale centrée réduite, ici p = $3{,}06.10^{-6}$: H_0 rejetée, effet significatif de l'exposition
- Le test de Wald estime un contraste entre 2 modalités ou l'effet d'une variable quantitative (estimate: effet lors d'une augmentation de 1), il ne teste pas l'effet d'une variable

Variables à plus de deux modalités

- Variable explicative continue ou à deux modalités = 1 paramètre => test du paramètre ≈ test de l'effet du facteur

- Variable à plus de 2 modalités = au moins 2 indicatrices: le test de Wald peut être significatif seulement pour une partie des indicatrices.

ex : logit p(malade) = $b_o+b_1X_1 +b_2X_2 +b_3X_3$

X_1 : faiblement exposé, X_2 : moyennement, X_3 : très

b_1 NS différent de 0 => risque « faible expo » non différent du risque des non exposés

b_2 et b_3 différents de 0 => risques différents du risque des non exposé

4. Test de l'effet d'une variable

Tester si X_1 a un effet = comparer deux modèles :

$\text{logit } p = \beta_0$ modèle nul, si X_1 n'a pas d'effet

$\text{logit } p = \beta_0 + \beta_1 X_1$ modèle si X_1 a un effet

H_0: X_1 n'a pas d'effet, les deux modèles sont équivalents
Ha: X_1 a un effet, le deuxième modèle est « significativement » meilleur

VetAgro Sup

Test du rapport de vraisemblance

Likelihood Ratio Test LRT:

L(modèle + facteur) > L (modèle - facteur)

Ratio des vraisemblances : $2 \ln (L_+ / L_-) = 2\ln L_+ - 2\ln L_- = D_- - D_+$

sous H_0 : facteur sans effet

$D_- - D_+ : \chi^2$

ddl : différence de ddl entre M+ et M-

ddl (résiduels) d'un modèle = n – k (nombre de paramètres):

Comparer les ddl est équivalent à comparer les k

VetAgro Sup

Modèles emboîtés

Exemple, modèle Expos (k = 2) versus modèle nul (k = 1):

```
> anova(MT, M0,test="Chisq")
Analysis of Deviance Table
Model 1: malade ~ expos
Model 2: malade ~ 1
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       973     967.16
2       974     989.49 -1  -22.329 2.297e-06 ***
```

De manière générale, le LRT permet de comparer un modèle avec tous ses sous-modèles (modèles inclus) = des modèles emboîtés.

Et les modèles non emboîtés?

Bon modèle = modèle ayant une faible déviance

ET

Pas trop de paramètres, car

- problème d'estimation des paramètres
- modèle sur-paramétré = modèle inutile

ex: le modèle saturé = modèle dans lequel on a autant de paramètres que de combinaisons de facteurs

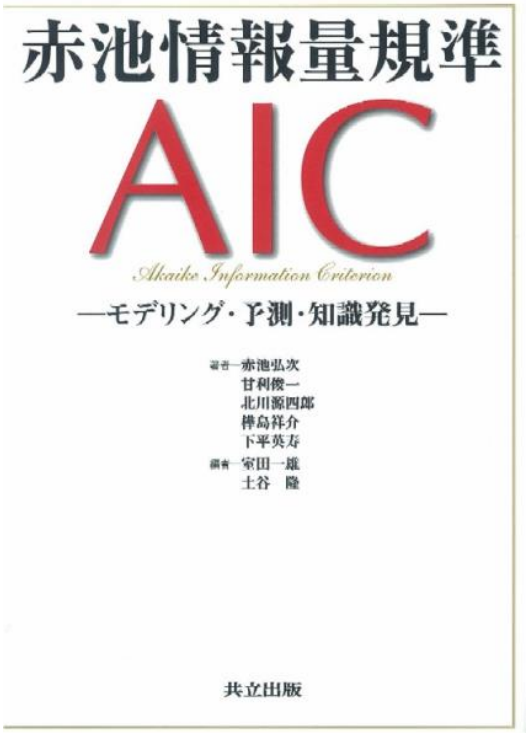
La sélection de modèle est un processus important

Critère d'Information d'Akaike

Akaike Information Criterion
(Akaike 1973):

AIC = Dev + 2 k

AIC minimal : compromis entre déviance et nombre de paramètres



5. exemple: cancer de l'oesophage

Lien avec le tabagisme? Lien avec l'exposition à la poussière?

Que se passe-t-il lorsque ces deux facteurs sont présents simultanément?

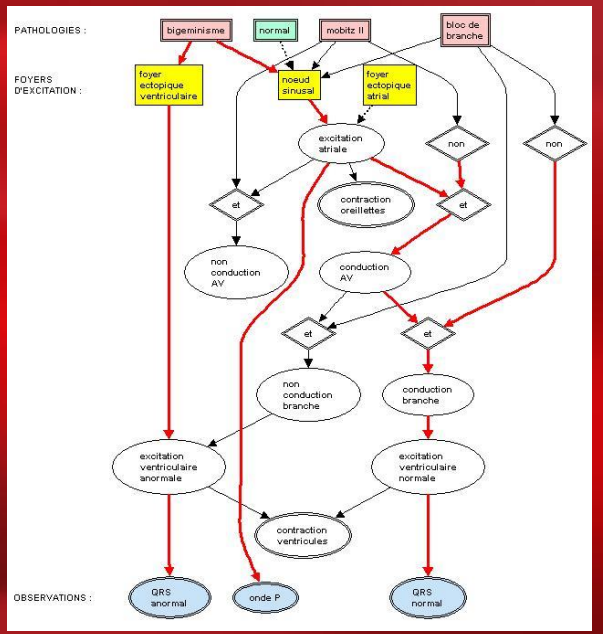
- confusion: si les personnes qui fument sont aussi exposées à la poussière => le lien avec le tabac pourrait n'être que la conséquence de cette association, la poussière étant le vrai facteur de risque et le tabac un facteur de confusion => une fois prise en compte l'effet de l'une, l'autre variable n'a pas d'effet supplémentaire

- interaction: la combinaison poussière + tabac possède un effet qualitativement différent de la somme des deux effets.

VetAgro Sup

Jeu de données complexe

1. Stratégie de construction de modèle
2. Critères et tests de la qualité d'ajustement
3. Eléments de discussion
4. Exemple



VetAgro Sup

1. Stratégies de construction de modèle

Souvent, de très nombreux modèles sont possibles, il faut faire un choix.

Avoir d'abord une réflexion préliminaire :

* Quelle démarche? Démarche analytique (recherche de causes), pragmatique (recherche d'indicateurs) ou exploratoire?

* Les variables explicatives sont-elles toutes pertinentes? Sont-elles collinéaires? Possibilité d'agréger, de simplifier, de résumer?

Dans quelle démarche se situe l'analyse?

Exploratoire?? Analyse multivariée descriptive

Explicative??

Prédictive??

Quelles variables explicatives?

Les variables explicatives sont-elles toutes pertinentes? Sont-elles collinéaires? Possibilité d'agréger, de simplifier, de résumer?

- une analyse multivariée descriptive des variables explicatives peut aider à faire le tri

Stratégies ascendantes et descendantes

Stratégie ascendante: complexifier progressivement le modèle : tester l'effet d'un X1 / s'il est significatif tester X1+X2

/ sinon tester X2, etc.

= le pire! L'effet d'un facteur peut être conditionnel à l'effet d'un autre facteur

Stratégie descendante: simplifier progressivement le modèle : tester X1*X2*X3….. Et enlever un à un les termes NS. Mais: modèle complet souvent non estimable, interactions ininterprétables, surparamétrage

Stratégies mixtes

Phase ascendante: tester l'effet des facteurs et interactions biologiquement plausibles, en considérant comme à retenir un effet libéral (ex : p = 0,2)

Construire un modèle complet comprenant les variables et interaction retenues

puis le simplifier; stratégie de simplification possible: stepAIC, ou stratégie « totale » : comparaison entre tous les sous-modèles possibles

VetAgro Sup

Théorie de l'information

Développée dans les années 1940 (codes): Kullback et Leibler
Liée à la théorie dans les années 1970 par Akaike

Information = quantité d'information perdue lorsqu'on représente la réalité par le modèle = distance entre deux espérances statistiques.
objectif = minimiser la distance

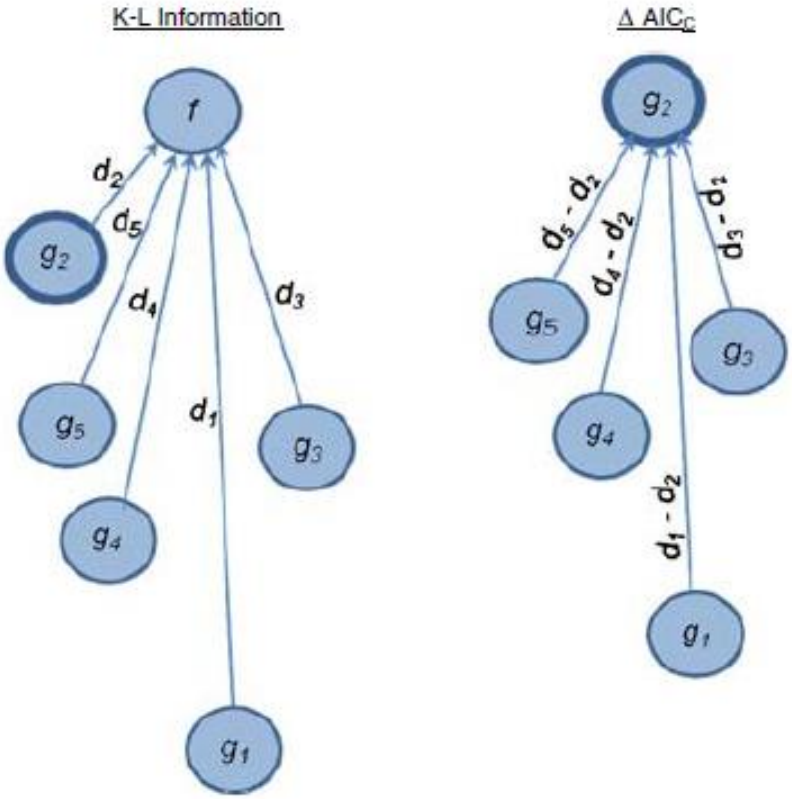
Idée = mesurer la force d'évidence (strength of evidence) pour chaque hypothèse alternative et plus seulement pour l'hypothèse nulle

Théorie de l'information

f = réalité
g_i = modèles
g_2 = meilleur modèle (perte minimale d'information)

Information de f non estimable mais constante,

information de g_i estimable / paramètres



VetAgro Sup

Critères d'information

$AIC = Dev + 2k$

$AICc = AIC + 2k(k+1) / (n-k-1)$

Dans la comparaison multimodèles, les ΔAIC(c) permettent de hiérarchiser les modèles: lorsque deux modèles ont des valeurs d'IC proches ils sont équivalents de ce point de vue. Dans ce cas choisir le modèle sur d'autres critères (parcimonie, modèle le plus simple).

VetAgro Sup

Usage des critères d'information

ΔAIC<2 : modèle supporté par les données
ΔAIC entre 2 et 7: support modéré
ΔAIC>10 : quasiment aucun support

Burnham et Anderson 2004

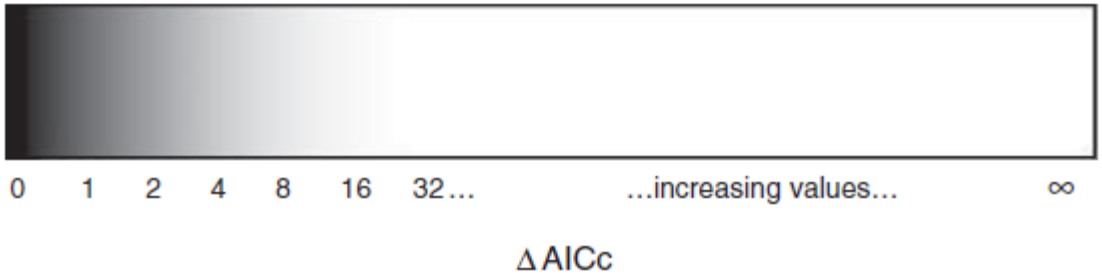


Fig. 2 Plausible hypotheses are identified by a narrow region in the continuum where Δ< perhaps four to seven (*black and dark grey*). The evidence in the light grey area is inconclusive and value judgments for hypotheses in this region are equivocal. Implausible models are shown in *white*, Δ> about 14

Burnham, Anderson, Huyvaert 2011, Behav Ecol Sociobiol 65: 23-35

VetAgro Sup

Vraisemblance relative des modèles

A partir des ΔAIC on peut estimer la vraisemblance relative de chacun des modèles testés par rapport aux données étudiées:

$\ell(M/D) = \exp(-0.5 * \Delta AIC)$

Δ_j	Evidence ratio
2	2.7
4	7.4
6	20.1
8	54.6
9	90.0
10	148.4
11	244
12	403
13	665
14	1,097
15	1,808
20	22,026
50	72 billion

VetAgro Sup

Poids d'Akaike

Le poids d'AIC pour un modèle donné est la valeur de L(M/D) divisé par la somme de ces valeurs pour tous les modèles.

$$w_i = l_i / \sum_{j=1}^{R} l_j$$

Estime la probabilité de chacun des modèles testés par rapport aux données => critère de choix

Choix de modèle, inférence multimodèle

Lorsqu'un modèle est très supérieur aux autres (ΔAIC > 5): pas de doute (relativement à ce jeu de modèles!)

Lorsque plusieurs modèles sont proches, deux stratégies:

1) choix d'un modèle = parcimonie

2) incertitude: les modèles suivants comprennent une part de l'information =>

a) "model averaging": estimation de parametres basée sur un ensemble de modèles, pas seulement sur le meilleur. Prédiction = moyenne pondérée par le poids d'AIC des predictions de chaque modèle; Attention, peu robuste si les variables explicatives sont collinéaires!

b) estimation de l'incertitude de selection de modèle

Parcimonie, variables non informatives

Choisir un modèle dans une liste :

1) parcimonie: le modèle le plus simple est plus robuste: modèle le plus parcimonieux parmi les modèles ayant DeltaAIC < x
-> il peut y en avoir plusieurs: considerer le mieux ajusté, à discuter au plan biologique

2) variables non informatives (“nesting rule”, Arnold et al. 2010): éliminer les modèles dans lesquels un meilleur modèle est emboité, pour choisir un modèle ou un ensemble de modèles sur lesquels appliquer du modèle averaging

Parcimonie, variables non informatives

Liste fictive de selection de modèle (dredge), effets age, pluviométrie, sexe :

Model selection table

	(Intrc)	age	Rain	sex	df	logLik	AICc	delta	weight
2226	**-1.7720**	**-0.1277**	**0.05999**		**3**	**-125.130**	**262.7**	**0.00**	**0.035**
6322	-1.6500	-0.1485	0.06258	+	4	-124.408	263.4	0.70	0.025
1202	**-0.7805**	**-0.1285**		**+**	**3**	**-124.706**	**264.0**	**1.29**	**0.018**
2194	**-2.0410**	**-0.1223**			**2**	**-126.961**	**264.2**	**1.54**	**0.016**

2226 = le mieux ajusté

2194 = le plus parcimonieux (et incluant les variables présentes dans tous les meilleurs modèles)

1202 = concurrent du mieux ajusté

6322 = incluant des variables non informatives = identiques à un modèle mieux ajusté, incluant en plus une variable supplémentaire (sexe); à exclure pour le modèle averaging

VetAgro Sup

2. Critères et tests de la qualité d'ajustement

Le modèle choisi

- Est-il globalement « bien ajusté »? = « assez bien ajusté » pour espérer avoir pris en compte les facteurs essentiels lié à la variable à expliquer?

- dans quel mesure explique-t-il « bien » les variations entre groupes, années….

- la fonction de lien est-elle bien choisie?

=> Analyse des résidus = écarts entre valeurs observées et valeurs prédites par le modèle

Critères et tests de la qualité d'ajustement

a* Etude des résidus = écarts entre valeurs observées et valeurs prédites par le modèle

b* Tests d'ajustement

c* R^2 et pseudo R^2

(d Etude du classement des individus*

e Analyse des relations entre variables explicatives*

f Validation du lien logit)*

Pas vus ici

VetAgro Sup

a. Trois types (principaux) de résidus

Résidus bruts

$$y_i - \hat{\pi}_i$$

Résidus de Pearson

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Résidus en déviance : contribution de chaque point à la déviance totale du modèle:

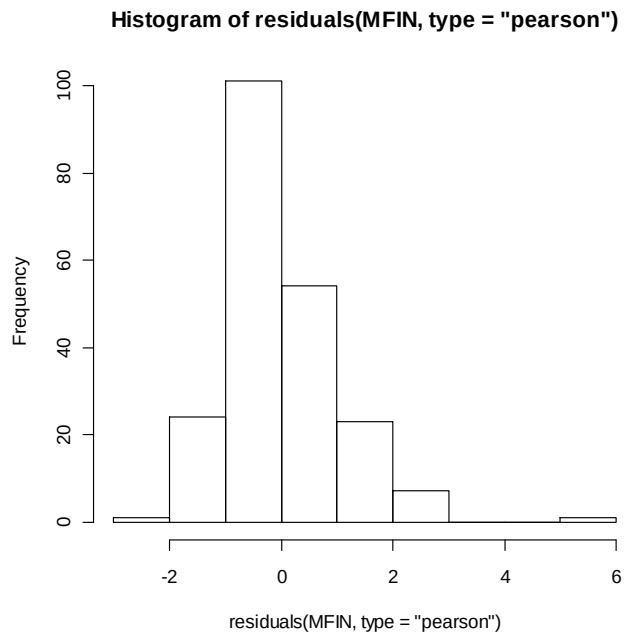
$$d_i = \text{signe}(y_i - \hat{\pi}_i)\sqrt{-2\log(\text{Prob}_{\text{estimée}}[Y = y_i / X = x_i]}$$

La somme des carrés des résidus en déviance = déviance résiduelle du modèle (en GLM, Pas en GLMM)

Distribution des résidus

Si données groupées : les résidus de Pearson et en déviance ont une distribution normale approximative

Sert plutôt à repérer les distributions aberrantes de résidus et la présence de résidus inattendus, de valeur absolue >2



VetAgro Sup

b. Ajustement global: tests de Chi2

Chi2 de déviance, Chi2 de Pearson:

Sur données groupées uniquement: n individus groupés dans j combinaisons de modalités et k paramètres au modèle, la somme des carrés des résidus (par groupe!) suit une distribution de chi2 à j-k-1 ddl

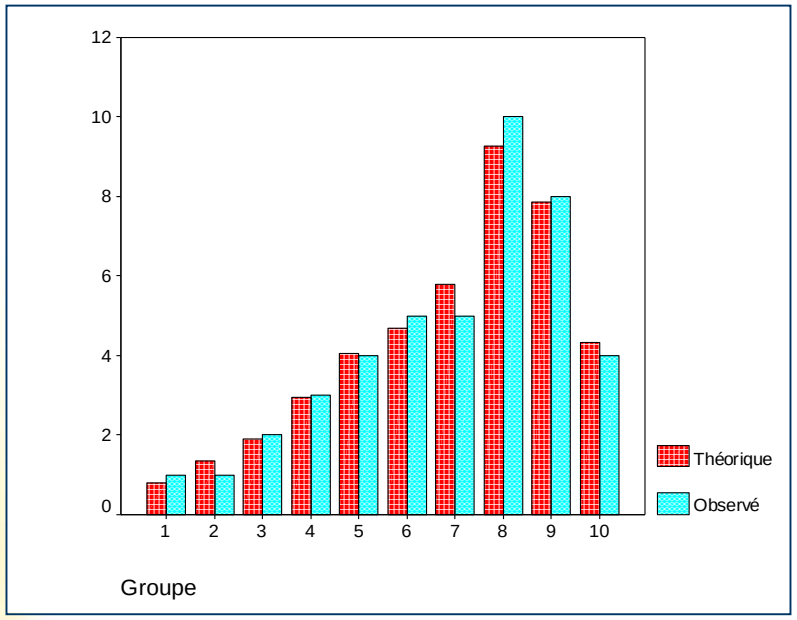
Chi2 de Pearson $$Q_P = \sum_{i=1}^{s} r_i^2$$ Chi2 de déviance $$Q_L = \sum_{i=1}^{s} d_i^2$$

Ces tests ne sont utilisables que si beaucoup d'individus, peu de variables et peu de modalités (10 fois plus d'individus que de combinaisons de variables)

VetAgro Sup

Test de Hosmer-Lemeshow

Les données sont rangées par ordre croissant des probabilités calculées à l'aide du modèle, puis partagées en p groupes. Le test du khi-deux est utilisé pour comparer les effectifs observés aux effectifs théoriques.



Nb de degrés de liberté
= Nb de groupes – 2
Plusieurs techniques pour créer les groupes (groupes de taille égale, groupe de largeur égale)

Faire le test avec 6 – 10 groupes

Test utilisé mais critiqué (peu puissant et dépendant du nombre de classes choisi)

VetAgro Sup

Test de Hosmer - Le Cessie

Utilise la somme des carrés des résidus bruts (n*score de Brier), ainsi que son espérance et son écart-type: formation d'une variable qui suit une loi normale centrée réduite sous H0 (ajustement).

« le moins contesté » actuellement

c. Mesures de type « R^2 »

Nombreuses mesures! Parmi ceux régulièrement utilisés :

Score de Brier : 1/n * somme des carrés des résidus bruts = score de Brier
Utilisé pour comparer les prévisions (météo) à la réalité observée (Brier 1950): varie de 0 (meilleur possible) à 1

* R^2 de Maddala, Cox & Snell (1989) et R^2 ajusté de Nagelkerke : pour le modèle choisi (« Full »):

$$R^2_{MCS} = 1 - \left(\frac{L(Null)}{L(Full)} \right)^{2/N},$$

Parfois > 1

$$R^2_{NK} = \frac{1 - \left(L(Null) / L(Full) \right)^{2/N}}{1 - L(Null)^{2/N}}$$

entre 0 et 1

VetAgro Sup

Pseudo-R^2 en déviance

Pseudo R^2 de McFadden (1974)

$$R^2_{MF} = 1 - \frac{LL(Full)}{LL(Null)}.$$

Mesure la déviance totale expliquée par le modèle choisi (« *Full* »).

Pseudo-R^2 de Veall-Zimmerman

Pseudo R^2 d'Aldrich et Nelson (1984) corrigé par Veall-Zimmermann

$$R^2_{AN} = \frac{G(M)}{G(M)+N}, \qquad R^2_{VZ} = \frac{R^2_{AN}}{\max\left(R^2_{AN} / \hat{p}\right)}$$

G(M) = différence de déviance entre *Mnull* et *Mfull* (chi2 du modèle),
^p = proportion de base

$$\max\left(R^2_{AN} / \hat{p}\right) = \frac{-2\left[\hat{p}\log(\hat{p}) + (1-\hat{p})\log(1-\hat{p})\right]}{1-2\left[\hat{p}\log(\hat{p}) + (1-\hat{p})\log(1-\hat{p})\right]}.$$

Le plus proche d'un R2 des moindres carrés selon une étude par simulation (Smith et McKenna 2013)

VetAgro Sup

Bilan: qualité d'un modèle

Examen graphique de la relation logistique et des résidus

Test: Hosmer-le-Cessie ou Chi-2 de Pearson si données groupées

Indicateurs: pseudo-R2

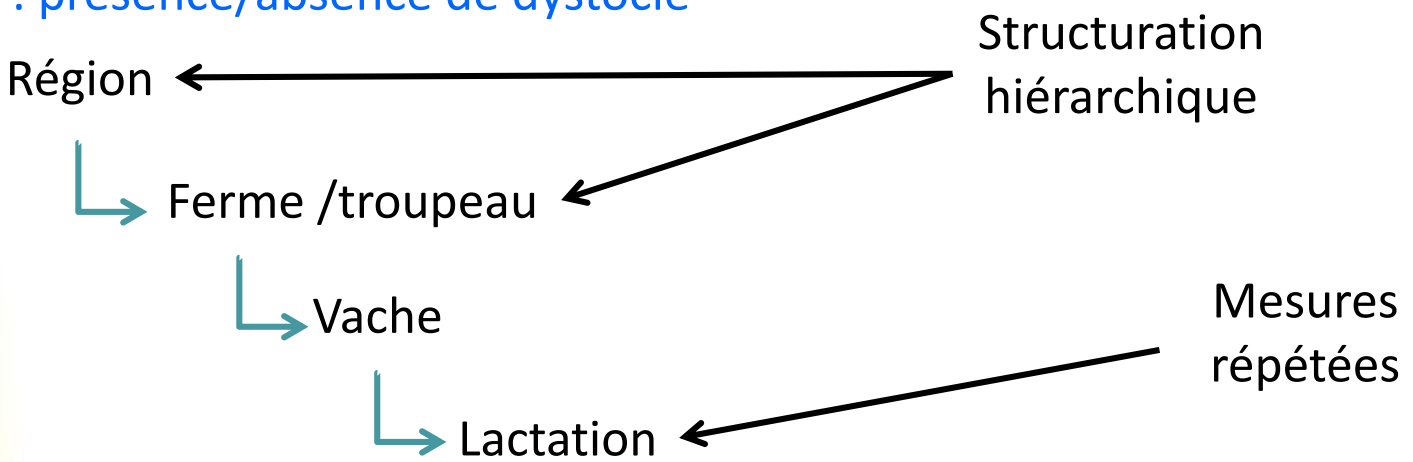
3. Eléments de discussion

Hypothèse d'indépendance:

les points-individus sont indépendants

En réalité:

Ex : présence/absence de dystocie



Eléments de discussion

La qualité du modèle est:

- Relative aux autres modèles possibles: AIC, R2…

- Absolue: On cherche soit à détecter des facteurs de variations de Y, soit à estimer leur effet (BIC)
  - Limiter les tests aux hypothèses biologiques
  - Revenir au plan d'expérimentation ou d'observation, et aux questions posées a priori

attention à ne pas surparamétrer ou tester des hypothèses juste parce que des informations sont disponibles!

VetAgro Sup