La régression logistique

Loïc Desquilbet

Département des **S**ciences **B**iologiques et **P**harmaceutiques

Ecole **N**ationale **V**étérinaire d'**A**lfort

– UE Libre « Préparation à la thèse expérimentale » –

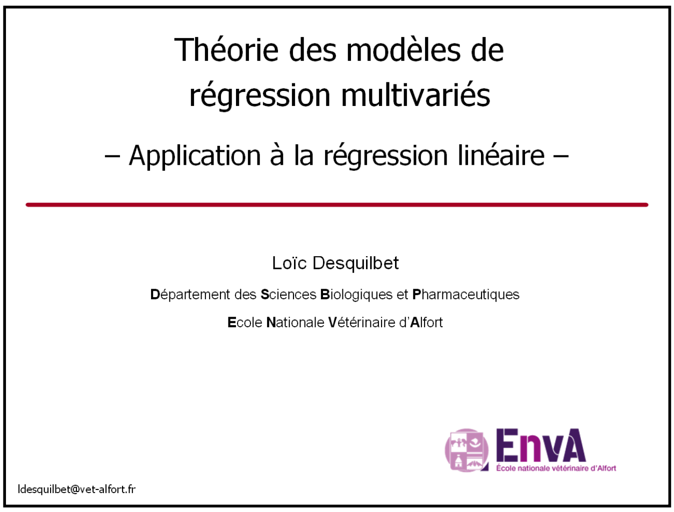
(2013-2014)

ldesquilbet@vet-alfort.fr

v1

Remarque avant de commencer ce cours

Pour être compris, ce cours nécessite d'avoir compris le cours intitulé « Théorie des modèles de régression multivariés – Application à la régression linéaire »



Plan

I. Présentation mathématique de la régression logistique

II. La régression logistique univariée en théorie

III. La régression logistique multivariée en théorie

IV. Le codage des variables avant introduction dans un modèle de régression logistique

Rappel : calcul de l'Odds Ratio

		Maladie 1	Maladie 0	
Exposition	1	a	b	e_1
	0	c	d	e_0
		m_1	m_0	

$$\begin{cases} Odds_{e1} = \dfrac{a}{c} \\ Odds_{e0} = \dfrac{b}{d} \end{cases} \qquad \begin{cases} P(M^+/E=1) = \dfrac{a}{e_1} = P_{E=1} \\ P(M^+/E=0) = \dfrac{c}{e_0} = P_{E=0} \end{cases}$$

- $OR = \dfrac{Odds_{e1}}{Odds_{e0}} = \dfrac{a/c}{b/d} = \dfrac{ad}{bc}$ = rapport de l'Odds de l'exposition chez les malades par celui chez les non malades

- $OR = \left(\dfrac{P(M^+/E=1) / 1 - P(M^+/E=1)}{P(M^+/E=0) / 1 - P(M^+/E=0)} \right) = \left(\dfrac{P_{E=1} / 1 - P_{E=1}}{P_{E=0} / 1 - P_{E=0}} \right)$

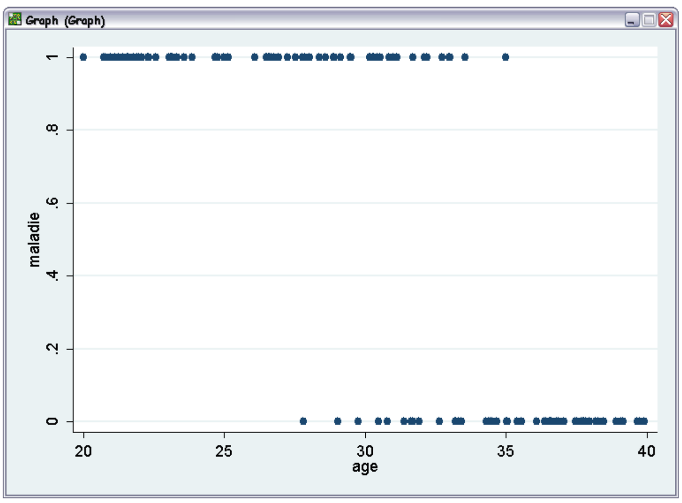
Contexte d'utilisation de la régression logistique

- Un modèle de régression multivarié permet de quantifier une association entre une exposition E et un état de santé Y ajusté sur des facteurs de confusion potentiels
- Il peut s'écrire de façon générale

  $Y = \alpha + \beta.E + \gamma_1.X_1 + \gamma_2.X_2 + \ldots + \gamma_p.X_p$
- La régression logistique doit être utilisée lorsque l'état de santé est binaire, non associé à un délai de survenue

Problématique

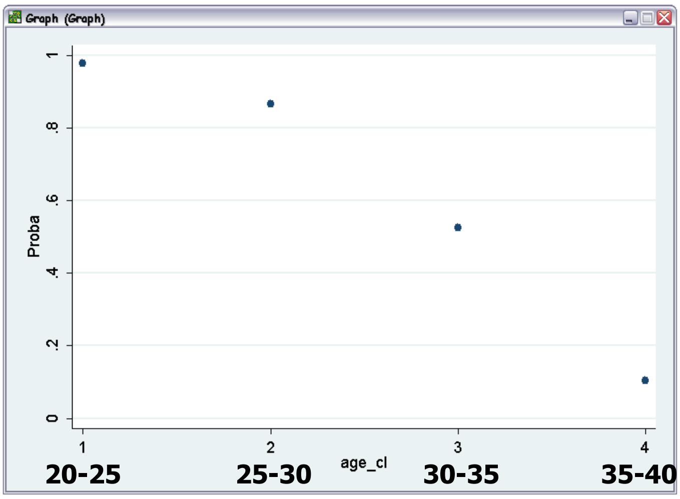
- Dans le cas où Y est binaire (de type « malade », « non malade »), écrire un modèle $Y = \alpha + \beta.E$ n'a pas de sens car la distribution de Y n'est pas du tout normale !



Problématique

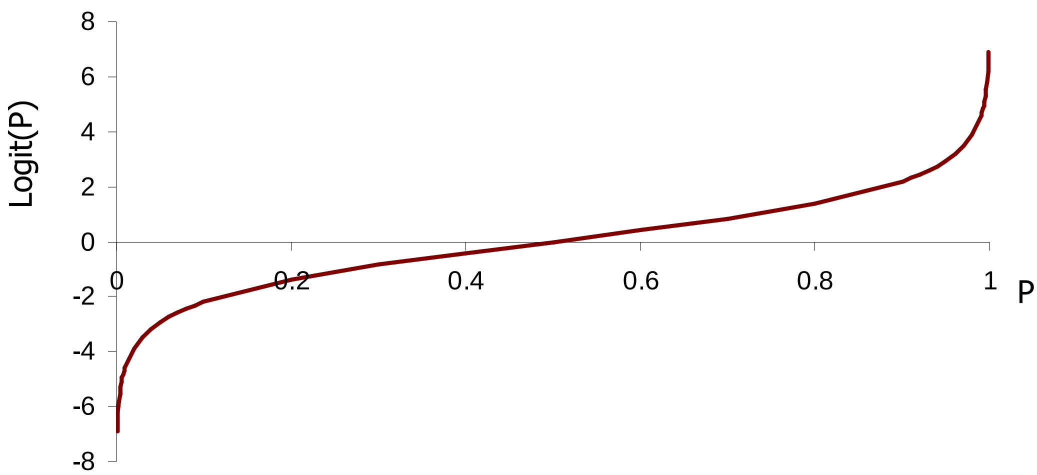
- Au lieu de Y, on pourrait modéliser la probabilité d'être malade en fonction de l'exposition

- Mais là encore, la probabilité d'être malade n'est pas distribuée normalement



Solution : utiliser la fonction *logit*

- $\text{Logit}(P) = \text{Ln}\left(\frac{P}{1 - P}\right)$ pour $P \in]0 ; 1[$, $\text{Logit}(P) \in]-\infty ; +\infty[$



Ecriture du modèle de régression logistique multivarié

- Soit $P = P(M^+ / E, X_1, X_2, \ldots, X_P)$, la probabilité d'être malade en fonction des expositions $E, X_1, X_2, \ldots, X_P$
- $\text{Logit}(P) = \alpha + \beta.E + \gamma_1.X_1 + \gamma_2.X_2 + \ldots + \gamma_p.X_p$
- β quantifie l'association entre l'exposition E et la probabilité P d'être malade, ajustée sur les variables X_i incluses dans le modèle, via la fonction Logit
- α n'a pas d'interprétation facile car c'est la valeur de Logit(P) pour E = 0

Plan

I. Présentation mathématique de la régression logistique

II. La régression logistique univariée en théorie

III. La régression logistique multivariée en théorie

IV. Le codage des variables avant introduction dans un modèle de régression logistique

Illustration avec Epi Info

Présentation du fichier de données (cf. cours sur la régression linéaire)

- La maladie : *faible_poids_naiss* = « 1 » si un chaton pèse moins de 85 grammes à la naissance ; « 0 » sinon
- Expositions (*variable*, et codage dans le fichier de données)
  - *male* : « 1 » si c'est un mâle ; « 0 » si c'est une femelle
  - *cesarienne* : « 1 » si nécessité de césarienne ; « 0 » si mise-bas naturelle
  - *race* : « 1 » pour British short/long hair ; « 2 » pour Main Coon ; « 3 » pour Chartreux ; « 4 » pour autres races
  - *age_mere* : âge de la mère en années
  - *taille_portee* : taille de la portée

Illustration avec Epi Info

LOGISTIC faible_poids_naiss = male

$\text{Logit}(P) = \alpha + \beta_{male}.\text{male}$

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S.E.	Z-Statistic	P-Value
male	0.4570	0.2857	0.7310	-0.7830	0.2396	-3.2675	0.0011
CONSTANT	*	*	*	-0.7720	0.1477	-5.2264	0.0000

$\beta_{smale} = -0,78$

$\alpha = -0,77$

$\text{Logit}(P) = -0,77 - 0,78.\text{male}$

Interprétation du paramètre β

Écriture du modèle pour E, exposition binaire (0/1), incluse seule dans le modèle

- Chez les sujets non exposés à E (E = 0) :

  $\text{Logit}(P) = \text{Logit}\,(P(M^+ / E=0)) = \text{Logit}\,P_{E=0} = \alpha + \beta \times 0 = \alpha \qquad (0)$

- Chez les sujets exposés à E (E = 1) :

  $\text{Logit}(P) = \text{Logit}\,(P(M^+ / E=1)) = \text{Logit}\,P_{E=1} = \alpha + \beta \times 1 = \alpha + \beta \qquad (1)$

$$(1) - (0) \Leftrightarrow \text{Logit}\,P_{E=1} - \text{Logit}\,P_{E=0} = (\alpha + \beta) - (\alpha) = \beta$$

$$\Leftrightarrow \text{Ln}\left(\frac{P_{E=1}}{1 - P_{E=1}}\right) - \text{Ln}\left(\frac{P_{E=0}}{1 - P_{E=0}}\right) = \beta$$

$$\Leftrightarrow \text{Ln}\left(\frac{P_{E=1} / 1 - P_{E=1}}{P_{E=0} / 1 - P_{E=0}}\right) = \beta$$

$$\Leftrightarrow \text{Ln}\,(OR_{E=1\ \textit{versus}\ E=0}) = \beta \qquad \boxed{\Rightarrow OR_{E=1\ \textit{versus}\ E=0} = e^{\beta}}$$

Interprétation du paramètre β

Écriture du modèle pour E, exposition quelconque, incluse seule dans le modèle

- Chez les sujets exposés à E = e_0 :

  $\text{Logit}(P) = \text{Logit}\,(P(M^+ / E=e_0)) = \text{Logit}\,P_{E=e0} = \alpha + \beta.e_0 \qquad (0)$

- Chez les sujets exposés à E = e_1 :

  $\text{Logit}(P) = \text{Logit}\,(P(M^+ / E=e_1)) = \text{Logit}\,P_{E=e1} = \alpha + \beta.e_1 \qquad (1)$

$$(1) - (0) \Leftrightarrow \text{Logit}\,P_{E=e1} - \text{Logit}\,P_{E=e0} = (\alpha + \beta.e_1) - (\alpha + \beta.e_0) = \beta(e_1 - e_0)$$

$$\Leftrightarrow \text{Ln}\left(\frac{P_{E=e1}}{1 - P_{E=e1}}\right) - \text{Ln}\left(\frac{P_{E=e0}}{1 - P_{E=e0}}\right) = \beta(e_1 - e_0)$$

$$\Leftrightarrow \text{Ln}\left(\frac{P_{E=e1} / 1 - P_{E=e1}}{P_{E=e0} / 1 - P_{E=e0}}\right) = \beta(e_1 - e_0)$$

$$\Leftrightarrow \text{Ln}\,(OR_{E=e1\ \textit{versus}\ E=e0}) = \beta(e_1 - e_0) \qquad \boxed{\Rightarrow OR_{E=e1\ \textit{versus}\ E=e0} = e^{\beta(e1 - e0)}}$$

Interprétation du paramètre β

Interprétation de β pour une variable quelconque incluse seule dans le modèle

- $e^{\beta.(e1 - e0)}$ = OR brut quantifiant l'association entre E et M, comparant des sujets exposés à E = e_1 par rapports aux sujets exposés à E = e_0
- Une différence de $(e_1 - e_0)$ sur l'exposition E entre 2 sujets se traduit par un OR de $e^{\beta.(e1 - e0)}$
- Si $(e_1 - e_0) = 1$ entre 2 sujets, l'OR comparant ces deux sujets = e^{β}

Interprétation du paramètre β

Conséquences

- Pour une exposition binaire E (en « 0/1 »), l'OR des sujets exposés (E=1) par rapport aux sujets non exposés (E=0) vaut e^{β}
- Pour une exposition qualitative ou quantitative, e^{β} = OR pour l'augmentation de l'exposition de 1 unité

Interprétation du paramètre β

Illustration

Logit(P) = α + β.male

LOGISTIC faible_poids_naiss = male

Term	Odds Ratio	95%	C.I.	Coefficient	S.E.	Z-Statistic	P-Value
male	0.4570	0.2857	0.7310	-0.7830	0.2396	-3.2675	0.0011
CONSTANT	*	*	*	-0.7720	0.1477	-5.2264	0.0000

l'OR est significativement différent de 1 ⇒ dans l'échantillon, le sexe du chaton était significativement associé à un faible poids à la naissance (OR=0,46 [0,29 ; 0,73] ; $p <$ 0,01), et on observait que les mâles avaient eu moins fréquemment un faible poids à la naissance que les femelles

Plan

I. Présentation mathématique de la régression logistique

II. La régression logistique univariée en théorie

III. La régression logistique multivariée en théorie

IV. Le codage des variables avant introduction dans un modèle de régression logistique

Interprétation du paramètre β

Ecriture du modèle logistique multivarié (≥ 2 variables incluses dans le modèle)

- Soit E et X deux expositions quelconques, de valeurs e_0, e_1, x_0, et x_1
- Soit P = P(M^+ / E,X), la probabilité d'être malade en fonction des expositions E et X
- Le modèle de régression logistique multivarié (ici, bivarié) s'écrit

  $\text{Logit}(P) = \alpha + \beta.E + \gamma.X$

Interprétation du paramètre β

Ecriture du modèle avec E et X incluses dans le modèle, pour $X=x_0$

- Chez les sujets exposés à $E = e_0$, et à $X = x_0$:

  $\text{Logit } P(M^+ / E=e_0,X=x_0) = \text{Logit } P_{e0,x0} = \alpha + \beta.e_0 + \gamma.x_0 \qquad (0)$

- Chez les sujets exposés à $E = e_1$, et à $X = x_0$:

  $\text{Logit } P(M^+ / E=e_1,X=x_0) = \text{Logit } P_{e1,x0} = \alpha + \beta.e_1 + \gamma.x_0 \qquad (1)$

(1) - (0) $\Leftrightarrow$ **$\ln (OR_{E=e1 \textit{ versus } E=e0,\, X=x0}) = \beta.(e_1 - e_0)$**

Interprétation du paramètre β

Ecriture du modèle avec E et X incluses dans le modèle, pour $X=x_1$

- Chez les sujets exposés à $E = e_0$, et à $X = x_1$:

  $\text{Logit } P(M^+ / E=e_0, X=x_1) = \text{Logit } P_{e0,x1} = \alpha + \beta.e_0 + \gamma.x_1$ (0')

- Chez les sujets exposés à $E = e_1$, et à $X = x_1$:

  $\text{Logit } P(M^+ / E=e_1, X=x_1) = \text{Logit } P_{e1,x1} = \alpha + \beta.e_1 + \gamma.x_1$ (1')

(1') - (0') ⇔ **$\text{Ln}(OR_{E=e1 \text{ } versus \text{ } E=e0, X=x1}) = \beta.(e_1 - e_0)$**

Interprétation du paramètre β

Conséquences

- Lorsque E et X sont incluses dans un même modèle, on fait l'hypothèse que l'OR quantifiant l'association entre E et M est le même quelle que soit la valeur de X

  ⇒ $e^{\beta.(e1 - e0)}$ est donc bien l'OR quantifiant l'association entre E et M **ajusté** sur X, comparant des sujets exposés à e_1 à des sujets exposés à e_0 pour E

- Si en réalité, cette hypothèse n'est pas vraie, cela signifie qu'il existe une interaction entre E et X

  ⇒ Le modèle contenant E et X seules ne sera pas adapté à la réalité, et le modèle fournira des estimations biaisées

  Solution : il faudra inclure dans le modèle : E, X, ainsi qu'un terme d'interaction entre E et X

Test statistique des paramètres

Illustration

Logit(P) = $\alpha + \beta$.male + γ.cesarienne

LOGISTIC faible_poids_naiss = male cesarienne

Term	Odds Ratio	95%	C.I.	Coefficient	S.E.	Z-Statistic	P-Value
male	0.4887	0.3017	0.7915	-0.7161	0.2460	-2.9105	0.0036
cesarienne	4.0196	2.1726	7.4366	1.3912	0.3139	4.4319	0.0000
CONSTANT	*	*	*	-1.0192	0.1635	-6.2322	0.0000

l'$OR_{cesarienne}$ est significativement différent de 1 $\Rightarrow$ indépendamment du sexe du chat, la présence d'une césarienne était significativement associée à un faible poids de naissance du chaton (OR=4,02 [2,27 ; 7,44] ; $p < 0,01$), et on observait dans l'échantillon que les chatons nés par césarienne avaient eu plus fréquemment un faible poids à la naissance que les chatons nés par mise-bas naturelle

Plan

I. Présentation mathématique de la régression logistique

II. La régression logistique univariée en théorie

III. La régression logistique multivariée en théorie

IV. Le codage des variables avant introduction dans un modèle de régression logistique

Codage des variables qualitatives nominales et interprétation de β

Rappel

- Il faut toujours recoder une variable qualitative nominale à *k* classes en *k* variables indicatrices

- Il faut inclure dans le modèle de régression logistique *k*-1 variables indicatrices parmi les *k* créées

  La variable indicatrice absente du modèle sera la **variable (classe) de référence** (chacune des *k-1* classes sera comparée à cette classe de référence)

Codage des variables qualitatives nominales et interprétation de β

Illustration du recodage avec Epi Info

- La variable *race_4cl* est une variable qualitative nominale à 4 classes

  ⇒ il faut la recoder en 4 variables indicatrices pour étudier l'association entre la race du chaton et son poids à la naissance

```
define race1
define race2
define race3
define race4

if race_4cl = 1 then
      assign race1 = 1
      assign race2 = 0
      assign race3 = 0
      assign race4 = 0
end

if race_4cl = 2 then
      assign race1 = 0
      assign race2 = 1
      assign race3 = 0
      assign race4 = 0
end

if race_4cl = 3 then
      assign race1 = 0
      assign race2 = 0
      assign race3 = 1
      assign race4 = 0
end

if race_4cl = 4 then
      assign race1 = 0
      assign race2 = 0
      assign race3 = 0
      assign race4 = 1
end
```

Variable initiale

Variables indicatrices créées

enne	race_4cl	age_mere	taille_portee	race1	race2	race3	race4
	3	7.9	3	0	0	1	0
	3	7.3	5	0	0	1	0
	3	4.2	5	0	0	1	0
	3	3.9	3	0	0	1	0
	3	2.9	4	0	0	1	0
	1	3.6	3	1	0	0	0
	1	4.2	6	1	0	0	0
	1	8.3	4	1	0	0	0
	1	5.6	2	1	0	0	0
	1	7.5	5	1	0	0	0
	3	5.4	2	0	0	1	0
	1	2	6	1	0	0	0
	1	0.9	7	1	0	0	0

Codage des variables qualitatives nominales et interprétation de β

Illustration du recodage avec Epi Info

LOGISTIC faible_poids_naiss = race2 race3 race4

Term	Odds Ratio	95%	C.I.	Coefficient	S.E.	Z-Statistic	P-Value
race2	$\boxed{0.0970}$	0.0126	0.7440	-2.3331	1.0395	-2.2445	$\boxed{0.0248}$
race3	0.6877	0.2571	1.8394	-0.3743	0.5019	-0.7458	0.4558
race4	1.8270	1.0649	3.1346	0.6027	0.2754	2.1881	0.0287
CONSTANT	*	*	*	-1.3304	0.2345	-5.6743	0.0000

Interprétation :

Codage des variables qualitatives nominales et interprétation de β

Illustration du recodage avec Epi Info

LOGISTIC faible_poids_naiss = race2 race3 race4

Term	Odds Ratio	95%	C.I.	Coefficient	S.E.	Z-Statistic	P-Value
race2	0.0970	0.0126	0.7440	-2.3331	1.0395	-2.2445	0.0248
race3	0.6877	0.2571	1.8394	-0.3743	0.5019	-0.7458	0.4558
race4	$\boxed{1.8270}$	1.0649	3.1346	0.6027	0.2754	2.1881	$\boxed{0.0287}$
CONSTANT	*	*	*	-1.3304	0.2345	-5.6743	0.0000

Interprétation :

Codage des variables qualitatives ordinales et interprétation de β

Rappel

- Il faut toujours vérifier la linéarité de l'association avec une variable qualitative ordinale avant de l'introduire telle quelle dans le modèle de régression logistique

- En procédant ainsi :
  - Recoder E en *k* variables indicatrices
  - Inclure dans le modèle *k*-1 variables indicatrices parmi les *k* créées (en général, on retire la première variable indicatrice)
  - Vérifier la linéarité des β associés à chacune des variables indicatrices

Codage des variables qualitatives ordinales et interprétation de β

Hypothèse de la linéarité de l'association avec une variable qualitative ordinale

Rappel du cas général

$$e^{\beta.(e1 - e0)} = OR_{E = e1\ versus\ E = e0} \Rightarrow Ln\left[e^{\beta.(e1-e0)}\right] = Ln\left[OR_{E = e1\ versus\ E = e0}\right]$$

$$\Rightarrow \beta.(e_1 - e_0) = Ln\left[OR_{E = e1\ versus\ E = e0}\right]$$

Interprétation numérique de l'hypothèse de la linéarité de l'association avec la variable

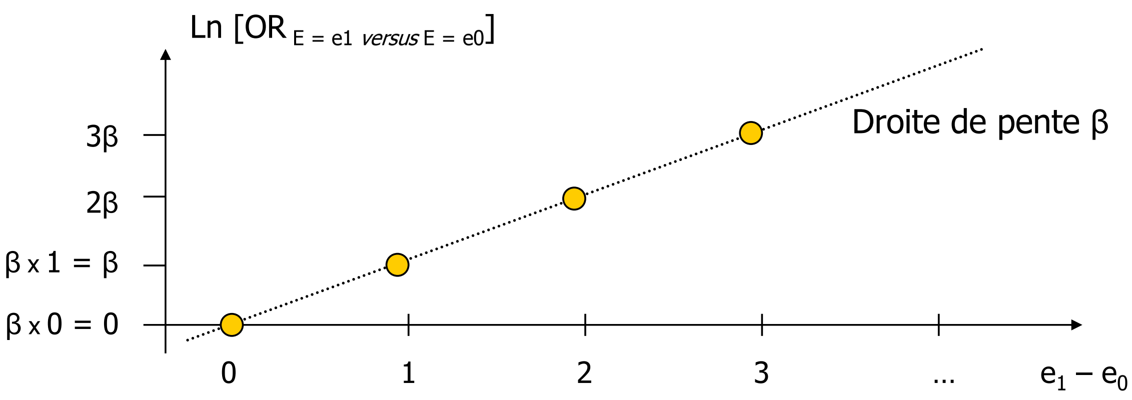
$e_1 - e_0$	$Ln[OR_{E = e1\ versus\ E = e0}]$
0	0
1	β
2	2β
3	3β
…	…

En incluant E telle quelle dans le modèle, on fait l'hypothèse que Ln(OR) augmente linéairement avec l'augmentation de l'écart entre deux valeurs d'exposition E

Codage des variables qualitatives ordinales et interprétation de β

Hypothèse de la linéarité de l'association avec une variable qualitative ordinale

Interprétation graphique de l'hypothèse de la linéarité de l'association avec la variable

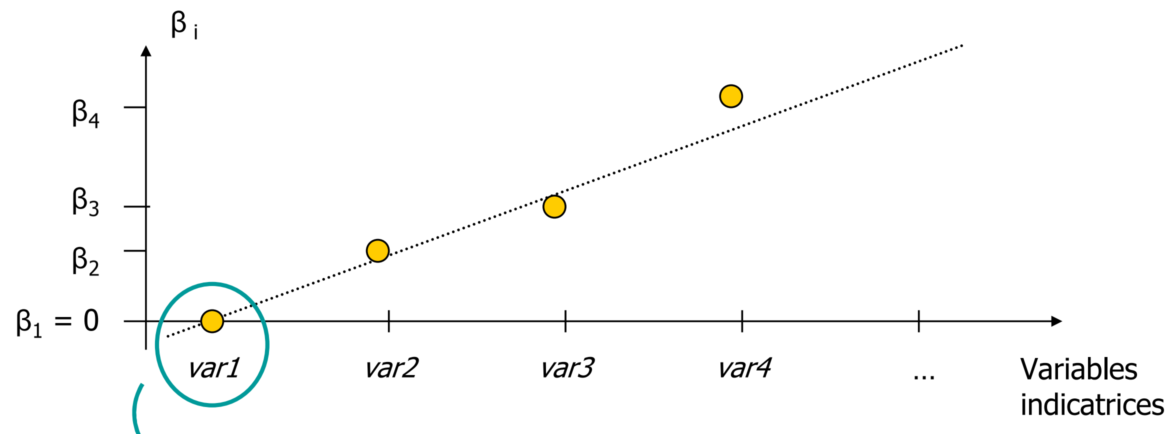


Codage des variables qualitatives ordinales et interprétation de β

Vérification graphique de l'hypothèse de la linéarité de l'association

Après avoir recodé *E* en *k* variables indicatrices, et après avoir « fait tourner » le modèle avec les *k*-1 variables indicatrices, le logiciel donne les β_i de chaque variable

Si β_i sont « relativement bien » alignés ⇒ l'association avec *E* est alors linéaire :



var1 = classe de référence ⇒ n'appartient pas au modèle ⇒ $\beta_1 = 0$

Codage des variables quantitatives et interprétation de β

Rappel

- Il faut vérifier l'hypothèse de la linéarité de l'association avec une variable quantitative avant de l'inclure telle quelle dans le modèle de régression logistique
- Démarche
  - Recoder la variable quantitative en variables qualitative ordinale à *k* classes (en fonction de seuils qui ont un sens, ou en fonction des quartiles)
  - Vérifier la linéarité de cette nouvelle variable qualitative ordinale
  - Si la linéarité n'est pas vérifiée, on ne peut pas inclure la variable quantitative telle quelle

    ⇒ Laisser les *k*-1 variables indicatrices dans le modèle, ou recoder la variable initiale en 2 classes
  - Si la linéarité est vérifiée, on peut inclure la variable quantitative telle quelle

Codage des variables quantitatives et interprétation de β

Illustration avec Epi Info sur la taille de la portée

- Vérifions la linéarité de l'association entre la taille de la portée et la présence de faible poids à la naissance
- Choix de recoder la taille de la portée en 4 classes selon les quartiles dans l'échantillon : 1-2, 3-4, 5, 6-8 chatons
  (Soit *taille_cl* cette variable qualitative ordinale)
- Création des 4 variables indicatrices issues de *taille_cl*

taille_portee	*taille_cl*	*taille1*	*taille2*	*taille3*	*taille4*
1-2	1	**1**	0	0	0
3-4	2	0	**1**	0	0
5	3	0	0	**1**	0
6-8	4	0	0	0	**1**

Codage des variables quantitatives et interprétation de β

Illustration avec Epi Info sur la taille de la portée

- Programmation sous Epi Info

```
define taille1
define taille2
define taille3
define taille4

if taille_portee <= 2 then
      assign taille1   = 1
      assign taille2   = 0
      assign taille3   = 0
      assign taille4   = 0
end

if 3 <= taille_portee and taille_portee <= 4 then
      assign taille1   = 0
      assign taille2   = 1
      assign taille3   = 0
      assign taille4   = 0
end
```

```
if taille_portee = 5 then
      assign taille1   = 0
      assign taille2   = 0
      assign taille3   = 1
      assign taille4   = 0
end

if taille_portee >= 6 then
      assign taille1   = 0
      assign taille2   = 0
      assign taille3   = 0
      assign taille4   = 1
end

LOGISTIC faible_poids_naiss = taille2 taille3 taille4
```

Logit(P) = $\alpha + \beta_2$.taille2 + β_3.taille3 + β_4.taille4

(*taille1* choisie comme la classe de référence)

Codage des variables quantitatives et interprétation de β

Illustration avec Epi Info sur la taille de la portée

- Résultats de la régression logistique

LOGISTIC faible_poids_naiss = taille2 taille3 taille4

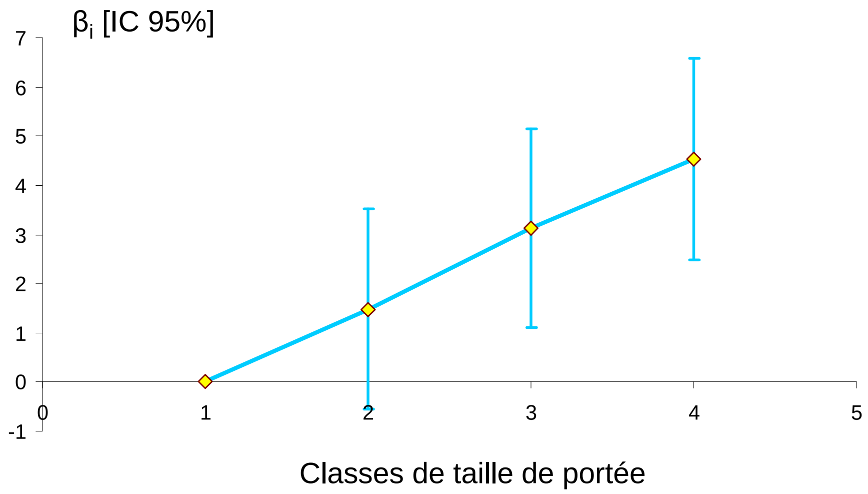
Term	Odds Ratio	95%	C.I.	Coefficient	S.E.	Z-Statistic	P-Value
taille2	4.4011	0.5727	33.8250	1.4819	.0405	1.4242	0.1544
taille3	22.5797	2.9869	170.6933	3.1171	.0321	3.0202	0.0025
taille4	92.7895	11.8755	725.0122	4.5303	.0489	4.3190	0.0000
CONSTANT	*	*	*	-3.7136	1.0121	-3.6692	0.0002

Ce sont les coefficients qu'il faut regarder pour vérifier la linéarité (avec leur écart-type), et non pas les OR !

Codage des variables quantitatives et interprétation de β (5)

Illustration avec Epi Info sur la taille de la portée

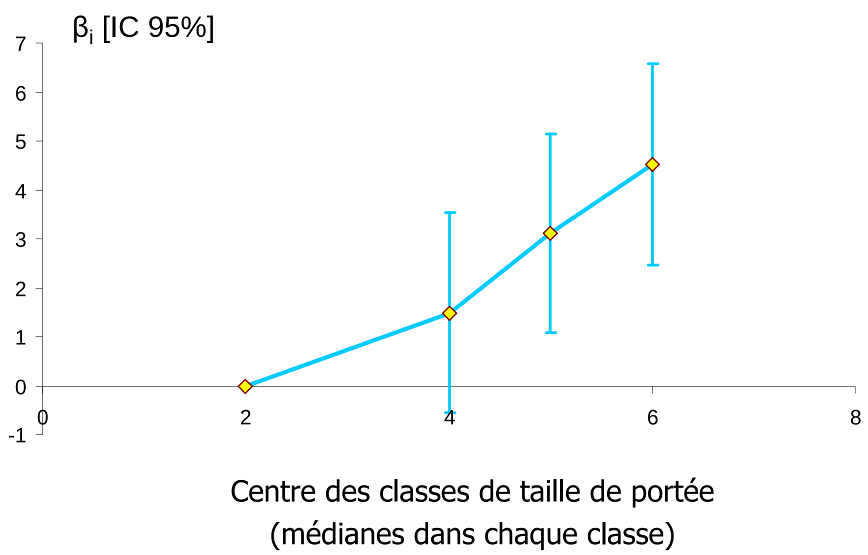
- Résultats de la régression logistique



Codage des variables quantitatives et interprétation de β (5)

Illustration avec Epi Info sur la taille de la portée

- Résultats de la régression logistique



Codage des variables quantitatives et interprétation de β

Conclusion sur l'association entre la taille de la portée et le faible poids à la naissance du chaton

- Les β_i peuvent tout à fait être considérés comme alignés
- L'association entre le faible poids à la naissance et la taille de la portée (en classes) est linéaire
- La variable *taille_portee* pourra être incluse telle quelle dans le modèle de régression logistique

Codage des variables quantitatives et interprétation de β

Illustration avec Epi Info sur le modèle multivarié incluant la taille de la portée

- Résultats de la régression logistique

LOGISTIC faible_poids_naiss = cesarienne male taille_portee

Term	Odds Ratio	95%	C.I.	Coefficient	S.E.	Z-Statistic	P-Value
cesarienne	2.7681	1.3668	5.6063	1.0182	0.3601	2.8277	0.0047
male	0.6587	0.3806	1.1400	-0.4175	0.2799	-1.4919	0.1357
taille_portee	2.8956	2.2113	3.7916	1.0632	0.1376	7.7290	0.0000
CONSTANT	*	*	*	-5.9543	0.6895	-8.6358	0.0000

Indépendamment du sexe du chaton, du type de mise-bas, la taille de la portée était significativement associée à un faible poids à la naissance ($OR_{ajusté}$=2,90 [2,21 ; 3,79] ; *P* < 0,01), et on observait que plus la taille de la portée augmentait, plus le faible poids à la naissance était fréquent

L'$OR_{ajusté}$ ici traduit la multiplication de la fréquence de faible poids à la naissance pour une augmentation de 1 chat en taille de portée