Théorie des modèles de régression multivariés

– Application à la régression linéaire –

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

v2

Plan

Plan

Quand utilise-t-on des modèles multivariés en épidémiologie ?

- Quand on veut étudier la relation causale entre une exposition et un état de santé en prenant en compte ≥ 1 facteur de confusion à la fois

- Quand X est un facteur de confusion quantitatif que l'on veut prendre en compte dans les analyses

  En effet, le transformer en classes pour pouvoir stratifier sur ses classes peut conduire à des biais de confusion résiduel (non prise en compte de toute l'information contenue dans X)

  (Exemple, l'âge du sujet)

- Quand l'exposition d'intérêt est quantitative et que l'on souhaite caractériser la forme de l'association avec l'état de santé d'un individu (caractérisation de la relation dose-effet)

Présentation des modèles multivariés

- Notations

  Soit une variable Y représentant l'état de santé, E_1, E_2, …, E_n, *n* expositions d'intérêt principal, et X_1, X_2, …, X_p, *p* expositions dont on veut tenir compte (facteurs de confusion potentiels)

- Ecriture du modèle

  $Y = \alpha + \beta_1.E_1 + \beta_2.E_2 + \ldots + \beta_n.E_n + \gamma_1.X_1 + \gamma_2.X_2 + \ldots + \gamma_p.X_p$

  Le logiciel va estimer la valeur des paramètres α, β_i, et γ_i

- Attention !

  - Le modèle ne « tourne » que sur les individus qui ont des données pour **TOUTES** les variables incluses dans le modèle
  - Une donnée manquante sur une variable incluse ⇒ exclusion de l'individu

Présentation des modèles multivariés

Le principe de la modélisation

- L'objectif de l'épidémiologie analytique est d'estimer des associations qui soient les plus proches possibles des associations causales réelles

- La réalité causale est par définition « complexe » qu'il va falloir simplifier pour l'approcher ⇒ création de « modèles »

  ⇒ On va donc faire des hypothèses (toute simplification d'un problème impose de faire des hypothèses) pour s'approcher de la réalité causale

Avantages et inconvénients des modèles multivariés

Avantages des modèles multivariés

- Les modèles multivariés permettent de prendre en compte simultanément plusieurs facteurs de confusion
- Ils permettent d'étudier les associations avec des variables quantitatives, et de caractériser une relation dose-effet (courbe linéaire, courbe en U, effet seuil, …)

Avantages et inconvénients des modèles multivariés

« Inconvénients » des modèles multivariés

- L'écriture d'un modèle fait des hypothèses assez fortes que l'on a souvent tendance à oublier (cf. plus loin)
- La « simplicité » de l'écriture d'un modèle dans un logiciel fait oublier les 3 critères que doivent remplir les expositions pour être incluses en tant que facteur de confusion potentiel !

Variables à inclure dans un modèle multivarié

4 familles de variables (expositions) à inclure dans un modèle multivarié

- Les expositions d'intérêt principal
- Les expositions citées dans la littérature comme étant associées à la maladie étudiée
- Les variables d'appariement (en enquête cas-témoins appariée)
- Les facteurs de confusion potentiels (expositions vérifiant les 3 critères)

Rappel : ne jamais inclure une exposition qui serait une conséquence de la maladie !!

Quel modèle de régression multivarié ?

Le principal critère de choix d'un modèle multivarié est la nature de la variable correspondant à l'état de santé étudié (variable Y) :

- Y quantitative : régression **linéaire**
  $\overline{Y} = \alpha + \Sigma\beta_i.E_i + \Sigma\gamma_i.X_i$ avec $\overline{Y}$ = la valeur moyenne de Y en fonction des E_i et X_i
- Y binaire (M+ / M-) : régression **logistique**
  Logit [P] = $\alpha + \Sigma\beta_i.E_i + \Sigma\gamma_i.X_i$ avec P = probabilité moyenne d'être « malade » en fonction des E_i et X_i
- Y binaire associée à un délai de survenue d'événement : **modèle de Cox**
  Ln [$\lambda(t)$] = $\alpha(t) + \Sigma\beta_i.E_i + \Sigma\gamma_i.X_i$ avec $\lambda(t)$ = incidence instantanée moyenne de la « maladie » en fonction des E_i et X_i (modèle utilisé en analyse de survie)
- Y est un nombre entier $\in \{0, 1, \ldots k\}$ avec k « petit » : **modèle de Poisson**
  Ln [E(Y)] = $\alpha + \Sigma\beta_i.E_i + \Sigma\gamma_i.X_i$ avec E(Y) = espérance de Y en fonction des E_i et X_i

Les paramètres d'un modèle multivarié

- α, β_i, et γ_i sont les « paramètres » (« *estimates* ») du modèle
- Les paramètres d'un modèle sont en général estimés par la méthode du maximum de vraisemblance

  $\Leftrightarrow$ Estimations telles que les valeurs de $\overline{Y}$* estimées par le modèle en fonction des expositions du modèle soient les plus proches possibles des valeurs de Y observées

<u>Remarque pour toute la suite du cours concernant la régression linéaire</u>

Plutôt que de noter $\overline{Y} = \alpha + \Sigma\beta_i.E_i + \Sigma\gamma_i.X_i$, on notera « Y » sans la barre du dessus…

$\Rightarrow$ $Y = \alpha + \Sigma\beta_i.E_i + \Sigma\gamma_i.X_i$,

* Ou logit [P], Ln [λ(t)], …

Nombre de variables maxi à inclure dans un modèle multivarié

- Le nombre de variables incluses dans un modèle dépend de la taille de l'échantillon ou du nombre d'événements
- Régression linéaire : nb maxi variables ≤ (taille échantillon / 10)
- Régression logistique : nb maxi variables ≤ (nb malades / 10)
- Modèle de Cox : nb maxi variables ≤ (nb d'événements / 10)

Plan

I. Vue d'ensemble des modèles de régression multivariés

II. La régression linéaire univariée en théorie

III. La régression linéaire multivariée en théorie

IV. Le problème de la « case vide » dans un modèle de régression multivarié

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

VI. Hypothèses sur lesquelles reposent tous les modèles de régression

VII. Interprétation des résultats issus d'un modèle de régression

Pour commencer…

Ecriture du modèle linéaire univarié (1 seule variable incluse dans le modèle)

- $Y = \alpha + \beta.E$
- Y sera l'estimation de la moyenne de Y en fonction seulement de E
- $\alpha = Y(E=0)$ = l'estimation de la moyenne de Y lorsque E=0

Illustration avec Epi Info

Présentation du fichier de données

- La « maladie » : poids de naissance d'un chaton en grammes (variable *poids_naissance*)
- 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_4cl* : « 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

REGRESS poids_naissance = male

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
male	9.588	1.815	27.8930	0.000000
CONSTANT	93.269	1.257	5509.5309	0.000000

Poids_naissance = $\alpha + \beta_{male}$.male

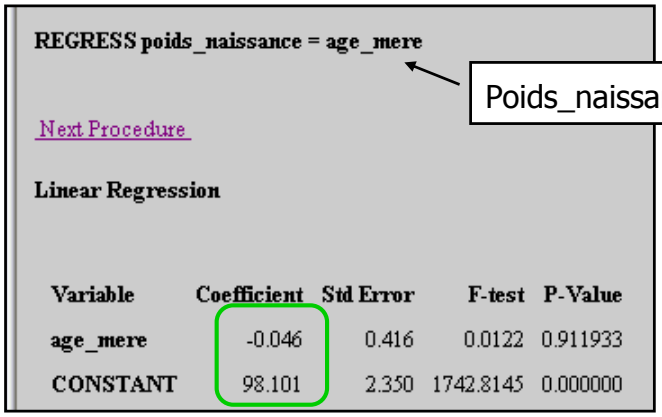
Ecart-type des paramètres estimés (SE_β)

β_{male} = 9,6

α = 93,3

Poids_naissance = 93,3 + 9,6.male

Illustration avec Epi Info



REGRESS poids_naissance = age_mere

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age_mere	-0.046	0.416	0.0122	0.911933
CONSTANT	98.101	2.350	1742.8145	0.000000

Poids_naissance = $\alpha + \beta_{age\_mere}$.age_mere

β_{age_mere} = -0,05

α = 98,1

Poids_naissance = 98,1 – 0,05.age_mere

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) :

  $Y(E=0) = Y_{E=0} = \alpha + \beta \times 0 = \alpha$ (0)

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

  $Y(E=1) = Y_{E=1} = \alpha + \beta \times 1 = \alpha + \beta \times 1 = \alpha + \beta$ (1)

(1) - (0) $\Leftrightarrow$ $Y_{E=1} - Y_{E=0} = (\alpha + \beta) - (\alpha) = \beta$

Interprétation du paramètre β

Interprétation de β pour une exposition binaire (0/1) incluse seule dans le modèle

- $Y_{E=1} - Y_{E=0} = \beta$
- β = La différence entre la moyenne de Y lorsque E=1 et la moyenne de Y lorsque E=0

  = L'écart moyen sur Y entre les individus qui ont E=1 et les individus qui ont E=0

- Retour sur l'illustration (poids naissance et sexe du chaton)

  β = 9,6 ⇒ L'écart moyen de poids la naissance entre les mâles (*male*=1) et les femelles (*male*=0) est de +9,6 grammes

Interprétation du paramètre β

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

- Chez les sujets exposés à $E=e_0$:

  $Y(E=e_0) = Y_{E=e0} = \alpha + \beta \times e_0$ (0)

- Chez les sujets exposés à $E=e_1$:

  $Y(E=e_1) = Y_{E=e1} = \alpha + \beta \times e_1$ (1)

(1) - (0) ⇔ $Y_{E=e1} - Y_{E=e0} = (\alpha + \beta.e_1) - (\alpha + \beta.e_0) = \beta.(e_1 - e_0)$

Interprétation du paramètre β

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

- $Y_{E=e1} - Y_{E=e0} = \beta.(e_1 - e_0)$
- $\beta.(e_1 - e_0)$ = La différence entre la moyenne de Y lorsque $E=e_1$ et la moyenne de Y lorsque $E=e_0$

  = L'écart moyen sur Y entre les individus qui ont $E=e_1$ et les individus qui ont $E=e_0$

- β = La différence de moyennes de Y lorsque $(e_1 - e_0) = +1$

  = L'écart moyen sur Y entre des individus qui diffèrent de +1 pour leur exposition

  = L'écart moyen sur Y pour une augmentation de 1 unité de l'exposition

Interprétation du paramètre β

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

- Retour sur l'illustration (poids naissance du chaton et l'âge de la mère)

  Supposons $e_1 = 3$ ans et $e_0 = 1,5$ ans

  $\beta.(3 - 1,5) = -0,05 \times 1,5 = -0,075$ gramme

  ⇒ L'écart moyen de poids la naissance entre les chatons dont la mère est âgée de 3 ans et ceux dont la mère est âgée de 1,5 ans est de -0,075 gramme

  β = -0,05 gramme

  ⇒ L'écart moyen de poids à la naissance entre des chatons dont l'âge des mères diffèrent de +1 an est de -0,05 gramme

Interprétation du paramètre β

En résumé

- Pour une exposition binaire (en « 0/1 »), l'écart moyen sur Y entre des sujets exposés et des sujets non exposés vaut β
- Pour une exposition qualitative ou quantitative, β = écart moyen sur Y pour l'augmentation de l'exposition de 1 unité

Plan

Pour commencer…

Ecriture du modèle linéaire 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
- $Y = \alpha + \beta.E + \gamma.X$
- Y sera l'estimation de la moyenne de Y en fonction de E et de X
- $\alpha = Y(E=0,X=0)$ = l'estimation de la moyenne de Y lorsque E=0 et lorsque X=0

Illustration avec Epi Info

```
REGRESS poids_naissance = male age_mere

Next Procedure

Linear Regression
```

Variable	Coefficient	Std Error	F-test	P-Value
male	9.605	1.820	27.8440	0.000000
age_mere	0.071	0.404	0.0307	0.860956
CONSTANT	92.894	2.480	1402.6085	0.000000

Poids_naissance = α + β.male + γ.age_mere

β = 9,6
γ = 0,07
α = 92,9

Poids_naissance = 92,9 + 9,6.male + 0,07.age_mere

Interprétation du paramètre β

Écriture du modèle avec E et X incluses toutes les deux dans le modèle, pour $X=x_0$

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

  $Y(E=e_0,X=x_0) = Y_{E=e0,X=x0} = \alpha + \beta \times e_0 + \gamma.x_0$ (0)

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

  $Y(E=e_1,X=x_0) = Y_{E=e1,X=x0} = \alpha + \beta \times e_1 + \gamma.x_0$ (1)

$$(1) - (0) \Leftrightarrow Y_{E=e1,X=x0} - Y_{E=e0,X=x0} = (\alpha + \beta \times e_0 + \gamma.x_0) - (\alpha + \beta \times e_1 + \gamma.x_0)$$
$$= \beta.(e_1 - e_0)$$

Interprétation du paramètre β

Écriture du modèle avec E et X incluses toutes les deux dans le modèle, pour $X=x_1$

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

  $Y(E=e_0,X=x_1) = Y_{E=e0,X=x1} = \alpha + \beta \times e_0 + \gamma.x_1$ (0')

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

  $Y(E=e_1,X=x_1) = Y_{E=e1,X=x1} = \alpha + \beta \times e_1 + \gamma.x_1$ (1')

$$(1') - (0') \Leftrightarrow Y_{E=e1,X=x1} - Y_{E=e0,X=x1} = (\alpha + \beta \times e_0 + \gamma.x_1) - (\alpha + \beta \times e_1 + \gamma.x_1)$$
$$= \beta.(e_1 - e_0)$$

Interprétation du paramètre β

Interprétation de β associé à E lorsque E et X sont incluses dans le modèle

- $Y_{E=e1,X=x0} - Y_{E=e0,X=x0} = \beta.(e_1 - e_0)$
- $Y_{E=e1,X=x1} - Y_{E=e0,X=x1} = \beta.(e_1 - e_0)$
- $\beta.(e_1 - e_0)$ = La différence entre la moyenne de Y lorsque $E=e_1$ et la moyenne de Y lorsque $E=e_0$ parmi les sujets pour lesquels $X=x_0$

  = L'écart moyen sur Y entre les individus qui ont $E=e_1$ et les individus qui ont $E=e_0$ parmi les sujets pour lesquels $X=x_1$

  = La différence entre la moyenne de Y lorsque $E=e_1$ et la moyenne de Y lorsque $E=e_0$ **quelle que soit la valeur de X**

Définition de l'association entre E et Y **ajustée** sur X

Interprétation du paramètre β

Interprétation de β associé à E lorsque E et X sont incluses dans le modèle

- $\beta.(e_1 - e_0)$ = La différence entre la moyenne de Y lorsque $E=e_1$ et la moyenne de Y lorsque $E=e_0$ ajustée sur X
- β = La différence de moyennes de Y lorsque $(e_1 - e_0) = +1$ après ajustement sur X

β = L'écart moyen sur Y, ajusté sur X, pour une augmentation de 1 unité de l'exposition E

Interprétation du paramètre β

Conséquences

- Lorsque E et X sont incluses dans un même modèle, et que l'on pense que ce modèle est valide, cela signifie que l'on **fait l'hypothèse** que l'association entre E et Y (quantifiée par un écart moyen sur Y) est la même quelle que soit la valeur de X

- Si en réalité, cette hypothèse n'est pas vraie, cela signifie…

  … que l'association entre E et M est différente selon les valeurs de X

  ⇒ 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

Interprétation du paramètre β

Retour sur l'illustration

Poids_naissance = α + β.male + γ.age_mere

REGRESS poids_naissance = male age_mere

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
male	9.605	1.820	27.8440	0.000000
age_mere	0.071	0.404	0.0307	0.860956
CONSTANT	92.894	2.480	1402.6085	0.000000

En faisant tourner un tel modèle, on fait l'hypothèse que l'association entre le poids de naissance du chaton et le sexe du chaton est la même quel que soit l'âge de la mère, et…

on fait l'hypothèse que l'association entre le poids de naissance du chaton et l'âge de la mère est la même quel que soit le sexe du chaton

⇔ Hypothèse de l'absence d'**interaction** entre l'âge de la mère et le sexe du chaton

Interprétation du paramètre β

Retour sur l'illustration

Poids_naissance = α + β.male + γ.age_mere

REGRESS poids_naissance = male age_mere

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
male	9.605	1.820	27.8440	0.000000
age_mere	0.071	0.404	0.0307	0.860956
CONSTANT	92.894	2.480	1402.6085	0.000000

Sous cette hypothèse, on peut dire que…

L'écart moyenne de poids à la naissance entre les mâles et les femelles, indépendamment de l'âge de la mère, est estimé à +9,6 grammes

Interprétation du paramètre β

Retour sur l'illustration

Poids_naissance = α + β.male + γ.age_mere

REGRESS poids_naissance = male age_mere

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
male	9.605	1.820	27.8440	0.000000
age_mere	0.071	0.404	0.0307	0.860956
CONSTANT	92.894	2.480	1402.6085	0.000000

Sous cette hypothèse, on peut dire que…

L'écart moyen de poids à la naissance entre des chatons dont l'âge des mères diffèrent de +1 an, indépendamment du sexe du chaton, est estimé à +0,07 gramme

Interprétation du paramètre β

Retour sur l'illustration

Poids_naissance = α + β.male + γ.age_mere

REGRESS poids_naissance = male age_mere

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
male	9.605	1.820	27.8440	0.000000
age_mere	0.071	0.404	0.0307	0.860956
CONSTANT	92.894	2.480	1402.6085	0.000000

Cette hypothèse d'absence d'interaction tient-elle la route ?

Oui, si cliniquement, il n'y a aucune raison de penser qu'il pourrait y avoir une telle interaction

Remarque

Ne pas vérifier cette hypothèse si l'on n'a pas de raison de la vérifier

Plan

Présentation du problème

- Le problème de la « case vide » concerne *tous* les modèles multivariés (régression linéaire, logistique, modèle de Cox, …)
- La « case vide » = « case » qui ne contient aucun sujet
- Supposons l'échantillon suivant, ainsi que deux expositions binaires E_1 et E_2

		Exposition E_1	
		0	1
Exposition E_2	0	a	b
	1	c	0

« case vide » : aucun sujet dans l'échantillon exposé à E_1=1 **et** à E_2=1

2 raisons d'avoir une case « vide »

- L'échantillon est trop petit, et avec un échantillon plus grand, on n'aurait pas eu de case vide ⇒ c'est la case vide **pratique**
- Même avec un échantillon de taille infinie, il n'y aurait eu aucun sujet exposé à E_1=1 **et** à E_2=1 ⇒ c'est la case vide **théorique**
  - Exemple

    E_1 = *sexe* : « 1 » pour les femelles, et « 0 » pour les mâles

    E_2 = *gestations_ant* : « 1 » ≥ 1 gestation(s) antérieure(s), et « 0 » si pas de gestation antérieure
  - Aucun animal ne peut avoir en même temps *sexe* =0 **et** *gestations_ant* =1 !

Case vide et modèle multivarié

- Supposons le modèle incluant les expositions binaires E_1 et E_2

  $Y = \alpha + \beta_1.E_1 + \beta_2.E_2$

- Par hypothèse, ce modèle estime que…

  … l'association entre M et E_1 vaut β_1, quelle que soit la valeur de E_2 (c'est-à-dire, que E_2 vaille « 0 » ou « 1 »)

  … l'association entre M et E_2 vaut β_2, quelle que soit la valeur de E_1 (c'est-à-dire, que E_1 vaille « 0 » ou « 1 »)

Case vide et modèle multivarié

- Si, dans l'échantillon, il existe une case vide (par exemple, aucun sujet exposé à E_1=1 **et** à E_2=1), le modèle estimera les paramètres β_1 et β_2 en faisant l'hypothèse que la case vide n'est *que* pratique

  ⇔ Le modèle va estimer β_1 seulement chez les sujets exposés à E_2=0, et il va faire l'hypothèse que cette association est la même chez des sujets exposés à E_2=1

  (Idem bien sûr pour l'estimation de β_2)

- Cette hypothèse ne pose pas de problème si effectivement la case vide est « pratique »

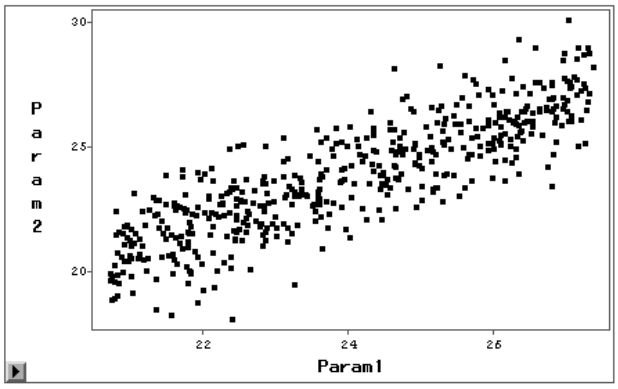
- Cette hypothèse est fausse (car sans objet) si la case vide est « théorique » !

⇒ Ne jamais faire tourner un modèle qui comporterait une case vide **théorique** !!

⇒ ~~$Y = \alpha + \beta.\text{sexe} + \gamma.\text{gestations\_ant}$~~

Case vide et modèle multivarié – illustration avec la régression linéaire

- Supposons $Parm_1$ et $Param_2$ deux paramètres biologiques quantitatifs très fortement associés



Case vide et modèle multivarié – illustration avec la régression linéaire

- Supposons que Y soit un caractère quantitatif
- Supposons que l'on veuille étudier l'association entre $Param_1$ et Y ajustée sur le $Param_2$

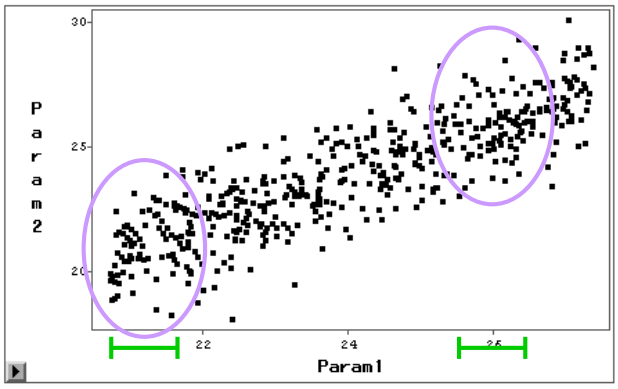
  Le modèle de régression linéaire bivariée s'écrit donc

  $Y = \alpha + \beta_1.Param_1 + \beta_2.Param_2$

- β_1 = écart moyen sur Y pour une augmentation de 1 unité du paramètre 1, quelle que soit la valeur du paramètre 2

Case vide et modèle multivarié – illustration avec la régression linéaire

- Sauf que…



« Une augmentation de 1 unité du paramètre 1, quelle que soit la valeur du paramètre 2 » n'existe pas :

- Passer de 20,5 à 21,5 pour le paramètre 1 n'est pas possible pour des valeurs du paramètre 2 > 25
- Passer de 25,5 à 26,5 pour le paramètre 1 n'est pas possible pour des valeurs du paramètre 2 < 23

Donc, le modèle de régression $Y = \alpha + \beta_1.Param_1 + \beta_2.Param_2$ n'est pas acceptable car il va fournir des paramètres β_1 et β_2 qui n'auront aucun sens

Plan

Rappel des 2 type de variables qualitatives

- Variable qualitative **nominale** à *k* classes : variable dont les chiffres affectés aux classes n'ont pas de sens particulier, et peuvent être interchangés

  Exemple sur la race du chat

  Race_4cl : « 1 » pour British short/long hair ; « 2 » pour Main Coon ; « 3 » pour Chartreux ; « 4 » pour autres races

Rappel des 2 type de variables qualitatives

- Variable qualitative **ordinale** à *k* classes : variable dont les chiffres affectés aux classes ont un sens, et ne peuvent pas être interchangés
- Parfois, les variables qualitatives ordinales proviennent d'une variable quantitative

  Exemple n°1

  Variable *montrer_dents* correspondant à la fréquence à laquelle un chien montre les dents, codée « 1 » pour jamais, « 2 » pour rarement, « 3 » pour souvent

  Exemple n°2

  Variable *tx_progest* le taux plasmatique de progesterone chez la vache, codée « 1 » pour < 2 ng/ml, « 2 » pour 2-4 ng/ml, « 3 » pour 4-6 ng/ml, et « 4 » pour > 6 ng/ml

Codage des variables et interprétation de β – Problématique

Introduction

- Les variables d'exposition doivent être numériques (par opposition à alphanumérique) pour pouvoir interpréter correctement les paramètres estimés
- L'interprétation de β dépend **totalement** du codage de la variable incluse dans le modèle
- Si deux épidémiologistes codent l'exposition avec des valeurs différentes pour chacune des classes, ils obtiendront à partir d'un même échantillon de départ une estimation de β différente !

Codage des variables et interprétation de β – Problématique

Introduction – Exemple

- Supposons que la vraie différence de poids à la naissance de chatons entre les mâles et les femelles soit égale à +10 grammes
- Supposons qu'à partir de la même base de données, 3 épidémiologistes codent la variable « « sexe » de 3 façons différentes

Codage des variables et interprétation de β – Problématique

Introduction – Exemple

- 3 codages de la variable *sexe*
  - L'épidémiologiste n°1 code les mâles « 1 » et les femelles « 0 »
  - L'épidémiologiste n°2 code les mâles « 1 » et les femelles « 2 »
  - L'épidémiologiste n°3 code les mâles « 5 » et les femelles « 2 »
- Supposons ces 3 modèles

Poids_naissance = $\alpha + \beta_1$.sexe (modèle de l'épidémiologiste n°1)
Poids_naissance = $\alpha + \beta_2$.sexe (modèle de l'épidémiologiste n°2)
Poids_naissance = $\alpha + \beta_3$.sexe (modèle de l'épidémiologiste n°3)

Codage des variables et interprétation de β – Problématique

Introduction – Exemple

- Ecart moyen de +10 grammes entre les mâles et les femelles

  $\Rightarrow$ Poids_naissance (mâles) – Poids_naissance (femelles) = +10
- Ecriture des modèles

$(\alpha + \beta_1 \times 1) - (\alpha + \beta_1 \times 0) = +10$ (1)
$(\alpha + \beta_2 \times 1) - (\alpha + \beta_2 \times 2) = +10$ (2)
$(\alpha + \beta_3 \times 5) - (\alpha + \beta_3 \times 2) = +10$ (3)

$\beta_1 = +10 \Rightarrow$ le modèle 1 conduit à une estimation de $\beta_1 = +10$
$-\beta_2 = +10 \Rightarrow$ le modèle 2 conduit à une estimation de $\beta_2 = -10$
$3\beta_3 = +10 \Rightarrow$ le modèle 3 conduit à une estimation de $\beta_3 = +3{,}33$

Associations identiques (+10 grammes entre les mâles et les femelles) mais estimations du paramètre associé au sexe différentes car codages différents !

Codage des variables et interprétation de β – En pratique

Cas des variables binaires

- Soit le modèle $Y = \alpha + \beta.E$ avec E une variable binaire
- Pour interpréter facilement β, il est préférable que…

  … les sujets exposés et les sujets non exposés diffèrent de 1 unité

  … la valeur de l'exposition des sujets exposés doit être supérieure à celle des sujets non exposés

⇒ Le codage en 0/1 est donc un codage de variable binaire qui permet une interprétation la plus facile qui soit

- Exemple

  Variable *traitement* : « 1 » pour les sujets traités, et « 0 » pour les sujets non traités ⇒ β = écart moyen sur Y entre les sujets traités et les sujets non traités

Codage des variables et interprétation de β – En pratique

Cas des variables qualitatives avec > 2 classes ou quantitatives

- Le codage de ces variables nécessite une réflexion particulière
  - une variable **qualitative nominale** ne doit jamais figurer telle quelle* dans le modèle de régression
  - une variable **qualitative ordinale** ou **quantitative** ne peut figurer dans le modèle qu'à condition d'avoir vérifié la linéarité de l'association entre cette variable et Y
- Articles pour le codage des variables quantitatives : cf. site de l'UE libre

* « variable introduite telle quelle » = « variable introduite dans le modèle sans recodage préalable »

Codage des variables qualitatives nominales et interprétation de β

Description du problème : illustration sur la race d'un chat (variable *race_4cl*)

Codage : « 1 » pour British short/long hair ; « 2 » pour Main Coon ; « 3 » pour Chartreux ; « 4 » pour autres races

Le modèle : Y = α + β.race_4cl fait les hypothèses que :

- $Y_{race\_4cl = 3 \text{ versus } race\_4cl = 2} = \beta.(3-2) = \beta$
- $Y_{race\_4cl = 2 \text{ versus } race\_4cl = 1} = \beta.(2-1) = \beta$

$\rightarrow$ $Y_{race\_4cl = 3 \text{ versus } race\_4cl = 2} = Y_{race\_4cl = 2 \text{ versus } race\_4cl = 1}$

- $Y_{race\_4cl = 4 \text{ versus } race\_4cl = 1} = \beta.(4-1) = 3\beta = 3 \times Y_{race\_4cl = 2 \text{ versus } race\_4cl = 1}$

⇒ Hypothèses fausses *a priori* !!

⇒ Ne **jamais** inclure une variable qualitative nominale telle quelle dans un modèle !

Solution : recoder la variable qualitative nominale à *k* classes en *k* variables indicatrices (« dummy variables »)

Codage des variables qualitatives nominales et interprétation de β

Illustration avec Epi Info

- Faisons tourner le modèle Poids_naissance = α + β.race_4cl

REGRESS poids_naissance = race_4cl

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
race_4cl	-3.445	0.697	24.3977	0.000001
CONSTANT	107.842	2.216	2367.9446	0.000000

Paramètre qui n'a aucun sens clinique puisque le modèle est basé sur des hypothèses **absurdes** :

- Ecart moyen identique entre deux races consécutives
- Ecart moyen entre les autres races et les British = 3 fois l'écart moyen entre Main Coon et British

Codage des variables qualitatives nominales et interprétation de β

Illustration avec Epi Info

- Faisons tourner le modèle Poids_naissance = α + β.race_4cl

REGRESS poids_naissance = race_4cl

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
race_4cl	-2.445	0.697	24.3977	0.000001
CONSTANT	107.842	2.216	2367.9446	0.000000

Ce n'est pas parce qu'un modèle « tourne » (⇔ l'ordinateur ne râle pas) que le modèle est valide !!

Codage des variables qualitatives nominales et interprétation de β

Démarche du recodage

Soit E une variable qualitative nominale à *k* classes

1) Recoder E en *k* variables « indicatrices » (variables binaires)

Variable *race_4cl* d'origine dans le fichier de données

Variables à créer dans le fichier de données

Exemple pour la variable *race_4cl*, recodée en 4 variables indicatrices :

Race_4cl	*race1*	*race2*	*race3*	*race4*
1	**1**	0	0	0
2	0	**1**	0	0
3	0	0	**1**	0
4	0	0	0	**1**

Codage des variables qualitatives nominales et interprétation de β

Démarche du recodage

2) Inclure dans le modèle *k*-1 variables indicatrices parmi les *k* créées

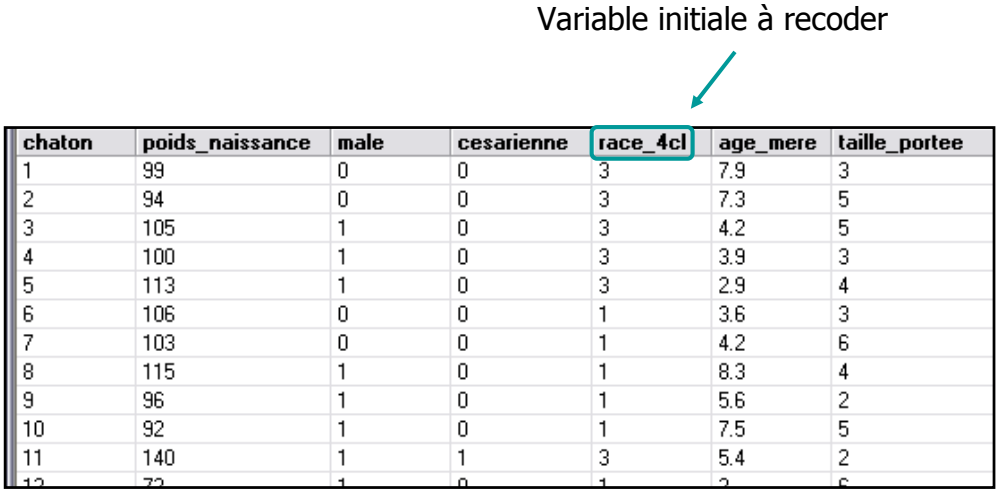
- La variable indicatrice absente du modèle sera la **variable (classe) de référence** (chacune des *k-1* classes sera comparée à cette classe de référence)
- On choisit la classe de référence selon les critères suivants
  - > Classe considérée comme « non exposée » dans la littérature / selon ses connaissances
  - > S'il y a doute, on préfèrera choisir la classe qui comprend le plus de sujets

Codage des variables qualitatives nominales et interprétation de β

Illustration du recodage avec Epi Info

- Voici une partie du fichier de données correspondant à l'étude du poids de naissance chez le chaton

Variable initiale à recoder



chaton	poids_naissance	male	cesarienne	race_4cl	age_mere	taille_portee
1	99	0	0	3	7.9	3
2	94	0	0	3	7.3	5
3	105	1	0	3	4.2	5
4	100	1	0	3	3.9	3
5	113	1	0	3	2.9	4
6	106	0	0	1	3.6	3
7	103	0	0	1	4.2	6
8	115	1	0	1	8.3	4
9	96	1	0	1	5.6	2
10	92	1	0	1	7.5	5
11	140	1	1	3	5.4	2
12	72	1	0	1	2	6

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

Ecriture du modèle linéaire avec variables indicatrices – illustration avec *race_4cl*

- Ecriture du modèle incluant les variables indicatrices issues de la variable *race_4cl* en prenant la classe « British » comme classe de référence

  Poids_naissance = $\alpha + \beta_2.race2 + \beta_3.race3 + \beta_4.race4$

- Chez les chats British : race2 = 0, race3 = 0, et race4 = 0

  Poids_naissance(race=British) = $\alpha + \beta_2 \times 0 + \beta_3 \times 0 + \beta_4 \times 0 = \alpha$ (1)

- Chez les chats Main Coon : race2 = 1, race3 = 0, et race4 = 0

  Poids_naissance(race=Main Coon) = $\alpha + \beta_2 \times 1 + \beta_3 \times 0 + \beta_4 \times 0 = \alpha + \beta_2$ (2)

Codage des variables qualitatives nominales et interprétation de β

Ecriture du modèle linéaire avec variables indicatrices – illustration avec *race_4cl*

- Ecriture du modèle incluant les variables indicatrices issues de la variable *race_4cl* en prenant la classe « British » comme classe de référence

  Poids_naissance = $\alpha + \beta_2.race2 + \beta_3.race3 + \beta_4.race4$

- Chez les chats Chartreux : race2 = 0, race3 = 1, et race4 = 0

  Poids_naissance(race=Chartreux) = $\alpha + \beta_2 \times 0 + \beta_3 \times 1 + \beta_4 \times 0 = \alpha + \beta_3$ (3)

- Chez les chats d'autres races : race2 = 0, race3 = 0, et race4 = 1

  Poids_naissance(race=autres races) = $\alpha + \beta_2 \times 0 + \beta_3 \times 0 + \beta_4 \times 1 = \alpha + \beta_4$ (4)

Codage des variables qualitatives nominales et interprétation de β

Interprétation des βi – illustration avec *race_4cl*

Poids_naissance(race=British)	$= \alpha + \beta_2 \times 0 + \beta_3 \times 0 + \beta_4 \times 0 = \alpha$	(1)
Poids_naissance(race=Main Coon)	$= \alpha + \beta_2 \times 1 + \beta_3 \times 0 + \beta_4 \times 0 = \alpha + \beta_2$	(2)
Poids_naissance(race=Chartreux)	$= \alpha + \beta_2 \times 0 + \beta_3 \times 1 + \beta_4 \times 0 = \alpha + \beta_3$	(3)
Poids_naissance(race=autres races)	$= \alpha + \beta_2 \times 0 + \beta_3 \times 0 + \beta_4 \times 1 = \alpha + \beta_4$	(4)

(2) - (1) ⇔ Ecart moyen de poids à la naissance entre des Main Coon et des British = β_2

(3) - (1) ⇔ Ecart moyen de poids à la naissance entre des Chartreux et des British = β_3

(4) - (1) ⇔ Ecart moyen de poids à la naissance entre des chats d'autres races et des British = β_4

Codage des variables qualitatives nominales et interprétation de β

Interprétation des βi associés à des variables indicatrices – cas général

- Soit une variable qualitative nominale à *k* classes, recodée en *k* variables indicatrices nommées var_1, var_2, …, var_k
- Modèle de régression linéaire avec variables indicatrices, après avoir retiré une variable indicatrice (l'épidémiologiste choisit, rien n'est pas imposé)

  $Y = \alpha + \Sigma \beta_i.var\_i$ (ici, choix d'avoir exclu var_1 ; i allant de 2 à *k*)

- Chaque β_i quantifie l'écart moyen sur Y entre la classe i et la classe de référence (ici la classe 1 car c'est celle qui a été retirée du modèle)

J'ai changé ordinale par nominale

Codage des variables qualitatives nominales et interprétation de β

Modèle linéaire avec variables indicatrices – Illustration avec Epi Info

```
REGRESS poids_naissance = male race2 race3 race4

Next Procedure

Linear Regression
```

Variable	Coefficient	Std Error	F-test	P-Value
male	8.693	1.646	27.9013	0.000000
race2	20.564	3.052	45.4045	0.000000
race3	4.502	3.084	2.1318	0.145058
race4	-6.226	1.937	10.3313	0.001414
CONSTANT	94.580	1.799	2764.1785	0.000000

Poids_naissance = α +β.male + γ_2.race2 + γ_3.race3 + γ_4.race4

Poids_naissance = 94,6 + 8,7.male + 20,6.race2 + 4,5.race3 – 6,2.race4

Codage des variables qualitatives nominales et interprétation de β

Modèle linéaire avec variables indicatrices – Illustration avec Epi Info

REGRESS poids_naissance = male race2 race3 race4

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
male	8.693	1.646	27.9013	0.000000
race2	20.564	3.052	45.4045	0.000000
race3	4.502	3.084	2.1318	0.145058
race4	-6.226	1.937	10.3313	0.001414
CONSTANT	94.580	1.799	2764.1785	0.000000

Interprétation :

Codage des variables qualitatives nominales et interprétation de β

Modèle linéaire avec variables indicatrices – Illustration avec Epi Info

REGRESS poids_naissance = male race2 race3 race4

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
male	8.693	1.646	27.9013	0.000000
race2	20.564	3.052	45.4045	0.000000
race3	4.502	3.084	2.1318	0.145058
race4	-6.226	1.937	10.3313	0.001414
CONSTANT	94.580	1.799	2764.1785	0.000000

Interprétation :

Codage des variables qualitatives ordinales et interprétation de β

Justification d'un recodage nécessaire

Exemple : soit la variable *age_4cl* l'âge d'un cheval codée en 4 classes :
« 1 » si ≤ 5 ans ; « 2 » si 5-10 ans ; « 3 » si 10-15 ans ; « 4 » si > 15 ans

Le modèle : $Y = \alpha + \beta.age\_4cl$ fait les hypothèses que :

- $Y_{age\_4cl = 2 \text{ versus } age\_4cl = 1} = \beta.(2-1) = \beta$
- $Y_{age\_4cl = 3 \text{ versus } age\_4cl = 2} = \beta.(3-2) = \beta$
- $Y_{age\_4cl = 4 \text{ versus } age\_4cl = 3} = \beta.(4-3) = \beta$

$\rightarrow$ $Y_{age\_4cl = 2 \text{ versus } age\_4cl = 1} = Y_{age\_4cl = 3 \text{ versus } age\_4cl = 2} = Y_{age\_4cl = 4 \text{ versus } age\_4cl = 3}$

- $Y_{age\_4cl = 4 \text{ versus } age\_4cl = 1} = \beta.(4-1) = 3\beta = 3 \times Y_{age\_4cl = 2 \text{ versus } age\_4cl = 1}$

⇒ Hypothèse qu'une augmentation d'une unité de l'exposition, quelle que soit sa valeur, se traduit par un **même écart moyen** sur Y

⇔ Hypothèse de la **linéarité de l'association**, hypothèse qu'il faut **toujours** vérifier avec une variable qualitative ordinale avant de l'introduire telle quelle dans le modèle

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

- Soit le modèle linéaire avec une exposition E qualitative ordinale : $Y = \alpha + \beta.E$
- $\beta.(e_1 - e_0) = Y_{E=e1} - Y_{E=e0}$

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

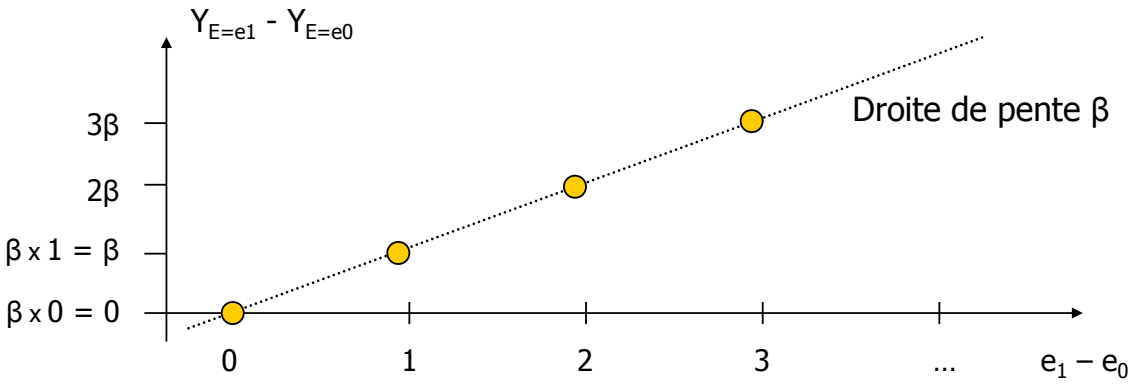
$e_1 - e_0$	$Y_{E=e1} - Y_{E=e0}$
0	0
1	β
2	2β
3	3β
…	…

En incluant E telle quelle dans le modèle, on fait l'hypothèse que l'écart moyen sur Y 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

Démarche pour vérifier l'hypothèse de la linéarité de l'association

Soit E une variable qualitative ordinale à *k* classes

- 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

Démarche pour vérifier l'hypothèse de la linéarité de l'association

- Si la linéarité n'est pas vérifiée, on ne peut pas inclure telle quelle la variable qualitative ordinale initiale

  ⇒ Laisser les *k*-1 variables indicatrices dans le modèle, ou recoder la variable initiale en 2 classes (⇒ nouvelle variable binaire qui ne pose pas de problème)

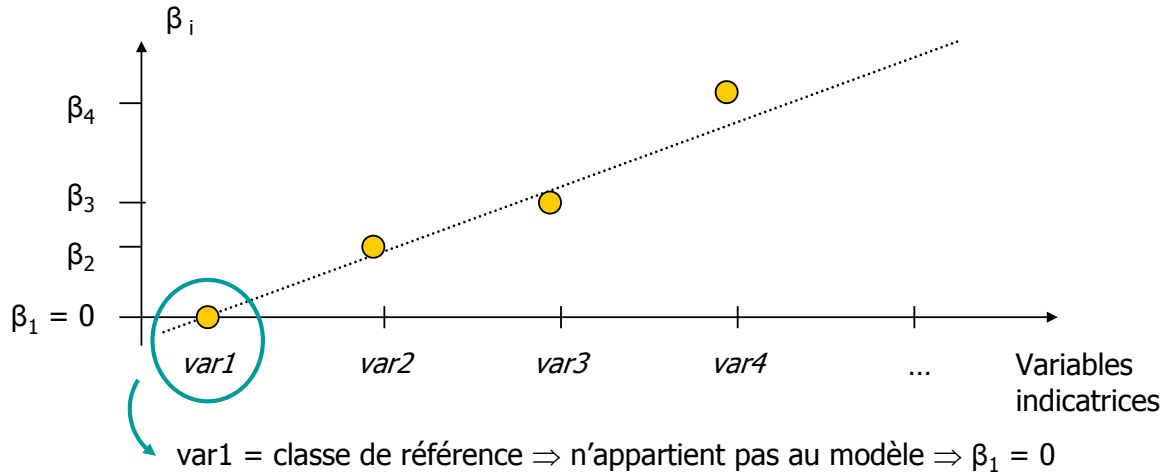
- Si la linéarité est vérifiée, on peut inclure telle quelle la variable qualitative ordinale

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 :



Codage des variables quantitatives et interprétation de β

Illustration n°1 avec Epi Info

- Reprenons l'exemple sur le poids à la naissance des chatons, et vérifions la linéarité de l'association entre l'âge de la mère et le poids à la naissance des chatons
- Choix de recoder l'âge de la mère en 4 classes selon les quartiles dans l'échantillon : ≤ 3,4 ans, 3,4-5,1 ans, 5,1-6,8 ans, > 6,8 ans
  (Soit *age_cl* cette variable qualitative ordinale)
- Création des 4 variables indicatrices issues de *age_cl*

age_mere	*age_cl*	*age1*	*age2*	*age3*	*age4*
≤ 3,4 ans	1	**1**	0	0	0
3,4-5,1 ans	2	0	**1**	0	0
5,1-6,8 ans	3	0	0	**1**	0
> 6,8 ans	4	0	0	0	**1**

Codage des variables quantitatives et interprétation de β

Illustration n°1 avec Epi Info

- Programmation sous Epi Info

```
define age_cl

define age1
define age2
define age3
define age4

if age_mere <= 3.4 then
      assign age_cl = 1
      assign age1   = 1
      assign age2   = 0
      assign age3   = 0
      assign age4   = 0
end

if 3.4 < age_mere and age_mere <= 5.1 then
      assign age_cl = 2
      assign age1   = 0
      assign age2   = 1
      assign age3   = 0
      assign age4   = 0
end
```

```
if 5.1 < age_mere and age_mere <= 6.8 then
      assign age_cl = 3
      assign age1   = 0
      assign age2   = 0
      assign age3   = 1
      assign age4   = 0
end

if age_mere > 6.8 then
      assign age_cl = 4
      assign age1   = 0
      assign age2   = 0
      assign age3   = 0
      assign age4   = 1
end

REGRESS poids_naissance = age2 age3 age4
```

$$Poids\_naissance = \alpha + \beta_2.age2 + \beta_3.age3 + \beta_4.age4$$

(*age1* choisie comme la classe de référence)

Codage des variables quantitatives et interprétation de β

Illustration n°1 avec Epi Info

- Résultats de la régression linéaire

REGRESS poids_naissance = age2 age3 age4

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age2	17.204	2.504	47.1963	0.000000
age3	10.841	2.405	20.3241	0.000009
age4	0.811	2.457	0.1090	0.741409
CONSTANT	90.829	1.716	2800.8667	0.000000

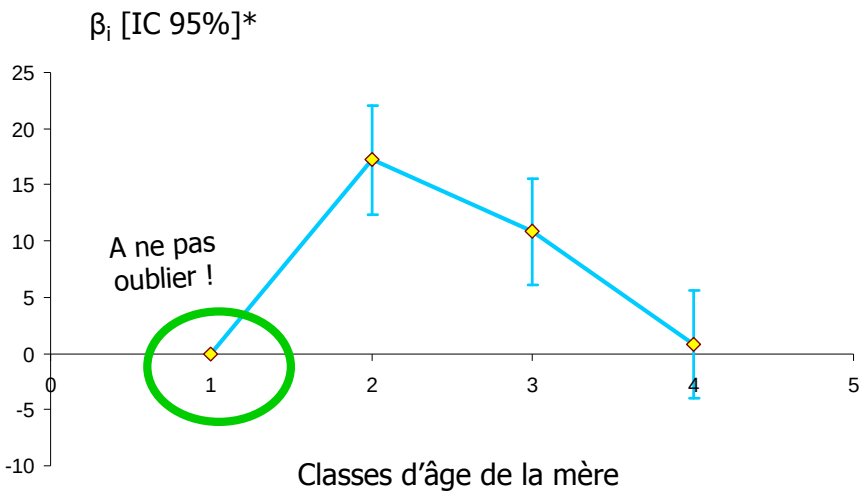
Les paramètres β_2, β_3, et β_4 laissent-ils penser que l'association entre l'âge de la mère **en classes** et le poids de naissance est linéaire ?

Poids_naissance = 90,8 + 17,2.age2 + 10,8.age3 + 0,8.age4

Codage des variables quantitatives et interprétation de β

Illustration n°1 avec Epi Info

- Résultats de la régression linéaire



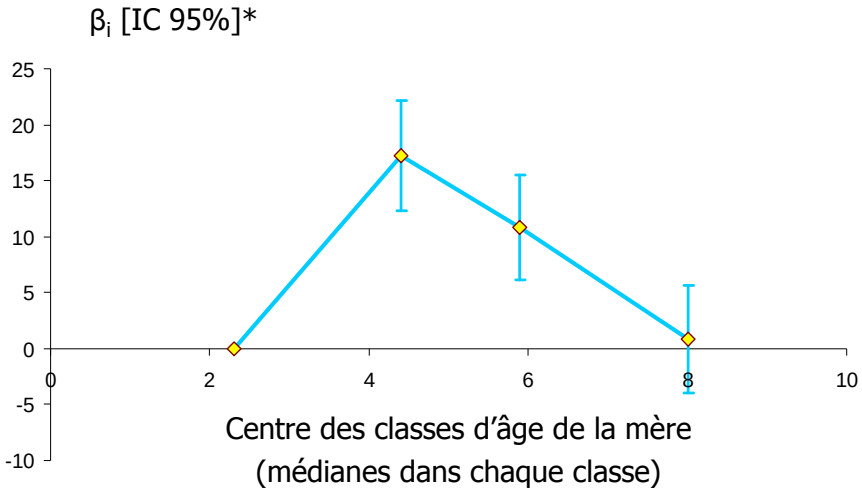
Mettre les classes en abscisses espacées de 1 unité à chaque fois n'est pas optimal pour vérifier la linéarité

* $IC_{95\%}$ de β : β ± 1,96 SE_β

Codage des variables quantitatives et interprétation de β

Illustration n°1 avec Epi Info

- Résultats de la régression linéaire



* $IC_{95\%}$ de β : β ± 1,96 SE_β

Codage des variables quantitatives et interprétation de β

Conclusion sur l'association entre l'âge de la mère et le poids de naissance du chaton

- Les β_i ne sont pas du tout alignés
- L'association entre le poids de naissance et l'âge de la mère (en classes) n'est pas linéaire
- La variable *age_cl* ne pourra pas être incluse dans le modèle
- Les variables indicatrices devront donc figurer dans le modèle

J'ai remis age_cl, puisque je ne parle de la var quant qu'après

Codage des variables quantitatives et interprétation de β

Interprétation du modèle avec variables indicatrices issues de *age_cl*

REGRESS poids_naissance = age2 age3 age4

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age2	17.204	2.504	47.1963	0.000000
age3	10.841	2.405	20.3241	0.000009
age4	0.811	2.457	0.1090	0.741409
CONSTANT	90.829	1.716	2800.8667	0.000000

Interprétation :

Codage des variables quantitatives et interprétation de β

Cas général

- Il faut vérifier l'hypothèse de la linéarité de l'association avec une variable quantitative avant de l'inclure telle quelle dans le modèle
- 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

Remarque à propos de l'illustration avec Epi Info

- La création de la variable qualitative ordinale *age_cl* n'a pas été utilisée, donc aurait pu ne pas être créée…
- Cette création de variable doit cependant être **absolument** présente dans sa **tête**, puisque c'est *age_cl* dont on vérifie la linéarité de l'association, et non pas *age_mere*
- Si l'hypothèse de la linéarité de l'association avec *age_cl* avait été vérifiée, la linéarité de l'association avec *age_mere* l'aurait été aussi
- Une fois que l'on a compris ce principe, il n'est plus utile de, **physiquement**, créer *age_cl*
- Il suffit de la créer **mentalement**, et de créer physiquement les variables indicatrices à partir de la variable (virtuelle, désormais) *age_cl*

Codage des variables quantitatives et interprétation de β

Supposons que l'on fasse tourner le modèle suivant (sur les mêmes données)

REGRESS poids_naissance = age_mere

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age_mere	-0.046	0.416	0.0122	0.911933
CONSTANT	98.101	2.350	1742.8145	0.000000

Interprétation :

Codage des variables quantitatives et interprétation de β

Supposons que l'on fasse tourner le modèle suivant (sur les mêmes données)

REGRESS poids_naissance age_mere

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age_mere	-0.046	0.416	0.0122	0.911933
CONSTANT	98.101	2.350	1742.8145	0.000000

L'interprétation précédente n'est pas correcte car l'association entre l'âge de la mère et le poids de naissance des chatons n'est pas linéaire

Pas le droit de faire tourner ce modèle !

Rappel : ce n'est pas parce qu'un modèle « tourne » (⇔ l'ordinateur ne râle pas) que le modèle est valide !!

Codage des variables quantitatives et interprétation de β

Illustration n°2 avec Epi Info

- Reprenons l'exemple sur le poids à la naissance des chatons, et vérifions la linéarité de l'association entre la taille de la portée et le poids à la naissance des chatons
- 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 n°2 avec Epi Info

- 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

REGRESS poids_naissance = taille2 taille3 taille4
```

Poids_naissance = $\alpha + \beta_2$.taille2 + β_3.taille3 + β_4.taille4

(*taille1* choisie comme la classe de référence)

Codage des variables quantitatives et interprétation de β

Illustration n°2 avec Epi Info

- 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

REGRESS poids_naissance = taille2 taille3 taille4
```

Poids_naissance = $\alpha + \beta_2$.taille2 + β_3.taille3 + β_4.taille4

(*taille1* choisie comme la classe de référence)

Codage des variables quantitatives et interprétation de β

Illustration n°2 avec Epi Info

- Résultats de la régression linéaire

REGRESS poids_naissance = taille2 taille3 taille4

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
taille2	-13.100	2.555	26.2831	0.000000
taille3	-27.988	2.736	104.6133	0.000000
taille4	-38.413	3.003	163.5881	0.000000
CONSTANT	117.381	2.319	2562.2851	0.000000

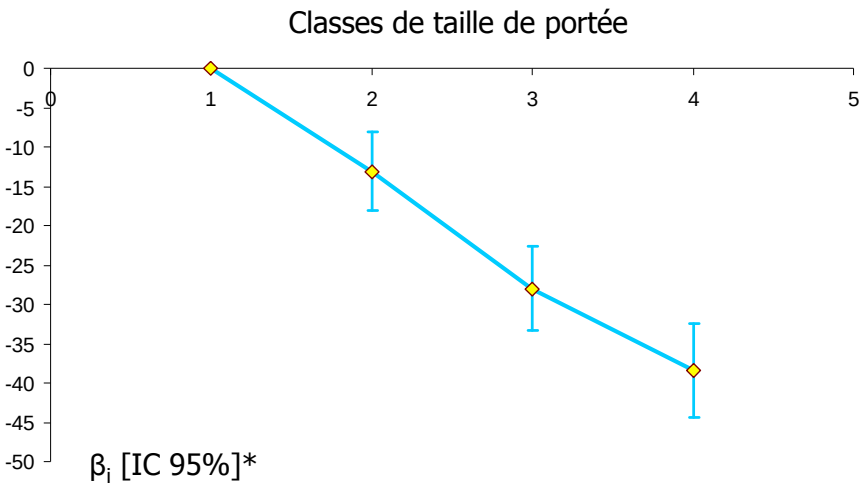
Les paramètres β_2, β_3, et β_4 laissent-ils penser que l'association entre la taille de la portée **en classes** et le poids de naissance est linéaire ?

Poids_naissance = 117,4 - 13,1.taille2 - 28,0.taille3 – 38,4.taille4

Codage des variables quantitatives et interprétation de β

Illustration n°2 avec Epi Info

- Résultats de la régression linéaire

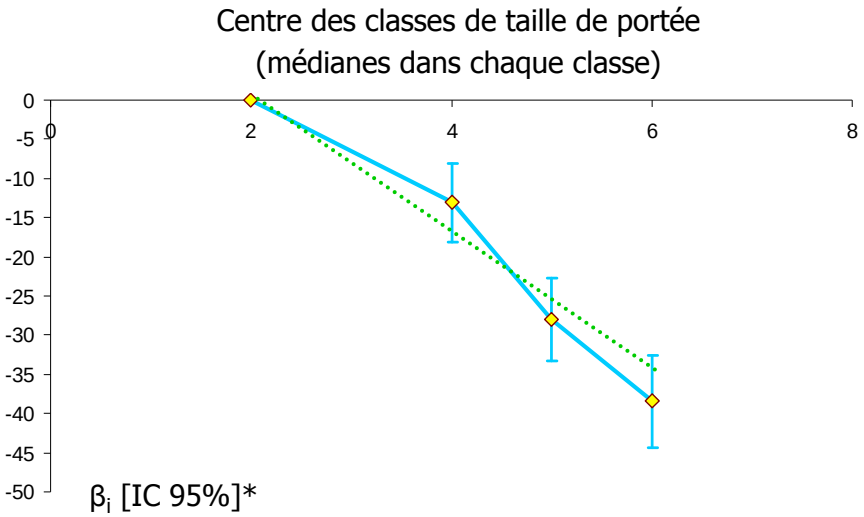


* $IC_{95\%}$ de β : β ± 1,96 SE_β

Codage des variables quantitatives et interprétation de β

Illustration n°2 avec Epi Info

- Résultats de la régression linéaire



* $IC_{95\%}$ de β : β ± 1,96 SE_{β}

Codage des variables quantitatives et interprétation de β

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

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

Codage des variables quantitatives et interprétation de β

Interprétation du modèle avec la variable *taille_portee*

REGRESS poids_naissance = taille_portee

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
taille_portee	-9.116	0.544	280.7703	0.000000
CONSTANT	135.961	2.385	3249.3533	0.000000

Interprétation :

Plan

Interactions et cases vides théoriques

Soit le modèle de régression multivarié suivant :

$Y^* = \alpha + \beta.E + \gamma_1.X_1 + \gamma_2.X_2 + \ldots + \gamma_p.X_p$

- Ajuster l'association entre E et Y sur les variables X_1, X_2, … X_n fait l'hypothèse que l'association entre E et Y est la même **quelles que soient** les valeurs de toutes les variables X_1, X_2, … X_n

  $\Leftrightarrow$ Pas d'interaction entre E et X_1, E et X_2, …, E et X_p

- Aucune case vide théorique : toutes les combinaisons entre toutes les valeurs des variables du modèle sont théoriquement possibles

  (Il est théoriquement possible d'avoir E élevée **et** X_1 faible, **et** X_2 élevée, **et** X_3 faible, etc…)

* « Y » dans le cas général (donc, Y, Ln(P), Ln(λ(t)), … (cf. diapo 10))

Linéarité et conséquences de Y

Soit le modèle de régression multivarié suivant :

$Y^* = \alpha + \beta.E + \gamma_1.X_1 + \gamma_2.X_2 + \ldots + \gamma_p.X_p$

- Si l'une des expositions est qualitative ordinale ou quantitative, la **linéarité de son association** avec Y doit être vérifiée

- Aucun des expositions telles que recueillies dans l'enquête ne doit être une conséquence de Y

* « Y » dans le cas général (donc, Y, Ln(P), Ln(λ(t)), … (cf. diapo 10))

Plan

I. Vue d'ensemble des modèles de régression multivariés

II. La régression linéaire univariée en théorie

III. La régression linéaire multivariée en théorie

IV. Le problème de la « case vide » dans un modèle de régression multivarié

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

VI. Hypothèses sur lesquelles reposent tous les modèles de régression

VII. Interprétation des résultats issus d'un modèle de régression

Hypothèse nulle dans un modèle de régression

- Tester un paramètre associé à une exposition incluse dans le modèle = tester l'association entre cette exposition et Y
- Dans tous les modèles : $H_0 : \beta = 0 \Leftrightarrow$ absence d'association
- « Tester β » = savoir si β s'éloigne significativement de « 0 »

Interprétation au niveau de l'échantillon

Poids_naissance = $\alpha + \beta_{age2}.age2 + \beta_{age3}.age3 + \beta_{age4}.age4 + \beta_{male}.male + \beta_{race2}.race2 + \beta_{race3}.race32 + \beta_{race4}.race4 + \beta_{taille\_portee}.taille\_portee$

REGRESS poids_naissance = age2 age3 age4 male race2 race3 race4 taille_portee

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age2	10.290	1.872	30.2233	0.000000
age3	9.176	1.818	25.4677	0.000001
age4	1.960	1.836	1.1389	0.286540
male	4.435	1.334	11.0494	0.000970
race2	13.093	2.473	28.0302	0.000000
race3	2.131	2.457	0.7520	0.386352
race4	-2.713	1.545	3.0850	0.079787
taille_portee	-6.967	0.539	167.2579	0.000000
CONSTANT	119.509	3.083	1502.7028	0.000000

Indépendamment du sexe du chaton, de sa race, et de la taille de la portée, les chatons dont la mère était âgée entre 5,1 et 6,8 ans avaient un poids à la naissance en moyenne significativement différent des chatons dont la mère était âgée de moins de 3,4 ans (écart de 9,2 grammes ; $p < 0,01$), et l'on observe que les chats dont la mère était âgée entre 5,1 et 6,8 ans pesaient en moyenne plus lourd que les chats dont la mère était âgée de moins de 3,4 ans

Interprétation au niveau de l'échantillon

Poids_naissance = $\alpha + \beta_{age2}.age2 + \beta_{age3}.age3 + \beta_{age4}.age4 + \beta_{male}.male + \beta_{race2}.race2 + \beta_{race3}.race32 + \beta_{race4}.race4 + \beta_{taille\_portee}.taille\_portee$

REGRESS poids_naissance = age2 age3 age4 male race2 race3 race4 taille_portee

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age2	10.290	1.872	30.2233	0.000000
age3	9.176	1.818	25.4677	0.000001
age4	1.960	1.836	1.1389	0.286540
male	4.435	1.334	11.0494	0.000970
race2	13.093	2.473	28.0302	0.000000
race3	2.131	2.457	0.7520	0.386352
race4	-2.713	1.545	3.0850	0.079787
taille_portee	-6.967	0.539	167.2579	0.000000
CONSTANT	119.509	3.083	1502.7028	0.000000

Indépendamment du sexe du chaton, de l'âge de la mère, et de la taille de la portée, les chatons de race Chartreux n'avaient pas un poids à la naissance significativement différent des chats de race British (écart de 2,1 grammes ; $p = 0,39$)

Interprétation au niveau de l'échantillon

Poids_naissance = $\alpha + \beta_{age2}.age2 + \beta_{age3}.age3 + \beta_{age4}.age4 + \beta_{male}.male + \beta_{race2}.race2 + \beta_{race3}.race32 + \beta_{race4}.race4 + \beta_{taille\_portee}.taille\_portee$

REGRESS poids_naissance = age2 age3 age4 male race2 race3 race4 taille_portee

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
age2	10.290	1.872	30.2233	0.000000
age3	9.176	1.818	25.4677	0.000001
age4	1.960	1.836	1.1389	0.286540
male	4.435	1.334	11.0494	0.000970
race2	13.093	2.473	28.0302	0.000000
race3	2.131	2.457	0.7520	0.386352
race4	-2.718	1.545	3.0850	0.079787
taille_portee	-6.967	0.539	167.2579	0.000000
CONSTANT	119.509	3.083	1502.7028	0.000000

Indépendamment du sexe du chaton, de sa race, et de l'âge de la mère, il existe une évolution linéaire significative du poids à la naissance du chaton lorsque la taille de la portée augmentait (-7,0 grammes par chaton en plus dans la portée ; $p <$ 0,01) et l'on observe que plus la taille de la portée augmentait, plus le poids à la naissance diminuait

Interprétation au niveau de la population

Inférence statistique ou inférence causale ?

- Inférence stat' quand on ne recherche pas le lien de cause à effet, mais simplement l'association
- En inférence stat' ou en inférence causale, il faut discuter de la présence de biais de classement (et, *a priori*, de sélection)
- En inférence causale, il faut discuter **en plus** de la présence de biais de confusion résiduel

Interprétation au niveau de la population

Illustration avec la taille de la portée – Inférence statistique

Sous l'hypothèse d'absence de biais de sélection et de classement, il y a de grandes chances pour que, dans la population cible, le poids à la naissance et la taille de la portée soient inversement associés, et ce, indépendamment du sexe du chaton, de sa race, et de l'âge de la mère

Illustration avec la taille de la portée – Inférence causale

Sous l'hypothèse d'absence de biais de sélection, de classement, et de confusion résiduel, il y a de grandes chances pour que, dans la population cible, une augmentation de la taille de la portée conduise à une diminution du poids à la naissance