HEALTH SCIENCE ECOSYSTEMS
GRADUATE SCHOOL AND RESEARCH

Risques
Epidémiologie
Territoire
INformations
Education et
Santé

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

INTRODUCTION AUX MODELES MULTIVARIES

P Staccini

Plan du cours

- Rappels
- La régression linéaire simple
- La régression logistique
- La régression linéaire multiple
- La régression logistique multiple
- Méthodes particulières
  - Analyse en composante principale
- Stratégie d'analyse

Rappels

Rappels

- "La statistique" := méthode qui consiste à observer et étudier une/plusieurs propriétés communes chez un groupe d'être, de choses ou d'entités.
- "Une statistique" := un nombre calculé à partir d'une population (d'êtres, de choses, ou d'entités).
- "Population" := la collection (d'être, de choses, ou d'entités) ayant des propriétés communes. Terme hérité d'une des premières applications de la statistique, la démographie ; e.g. un ensemble parcelles de terrain étudiées, une population d'animaux, un groupe de patients présentant une maladie définie, l'ensemble des plantes d'une espèce donnée, une population d'humains habitants un lieu particulier…
- "Individu" := élément de la population ; e.g. un patient, un insecte, une plante...
- "Variable" := une des propriétés communes aux individus que l'on souhaite étudier. Peut-être :
  - **qualitative** : appréciation de la parcelle, l'état de santé de l'insecte, couleur des pétales, appartenance religieuse
  - **quantitative** [numérique] **continue** [pouvant prendre n'importe quelle valeur réelle] : le taux d'acidité du sol, la longueur de l'insecte, la longueur de la tige, l'indice de masse corporelle.
  - **quantitative** [numérique] **discrète** [dès qu'il y a un saut minimum obligatoire entre deux valeurs successives, e.g. les nombres entiers] : la somme (sur tous les jours) du nombre de vaches présentes sur la parcelle, l'âge de l'insecte (en jours), le nombre de pétales sur la fleur, le nombre d'année d'études (réussies) depuis la petite école.

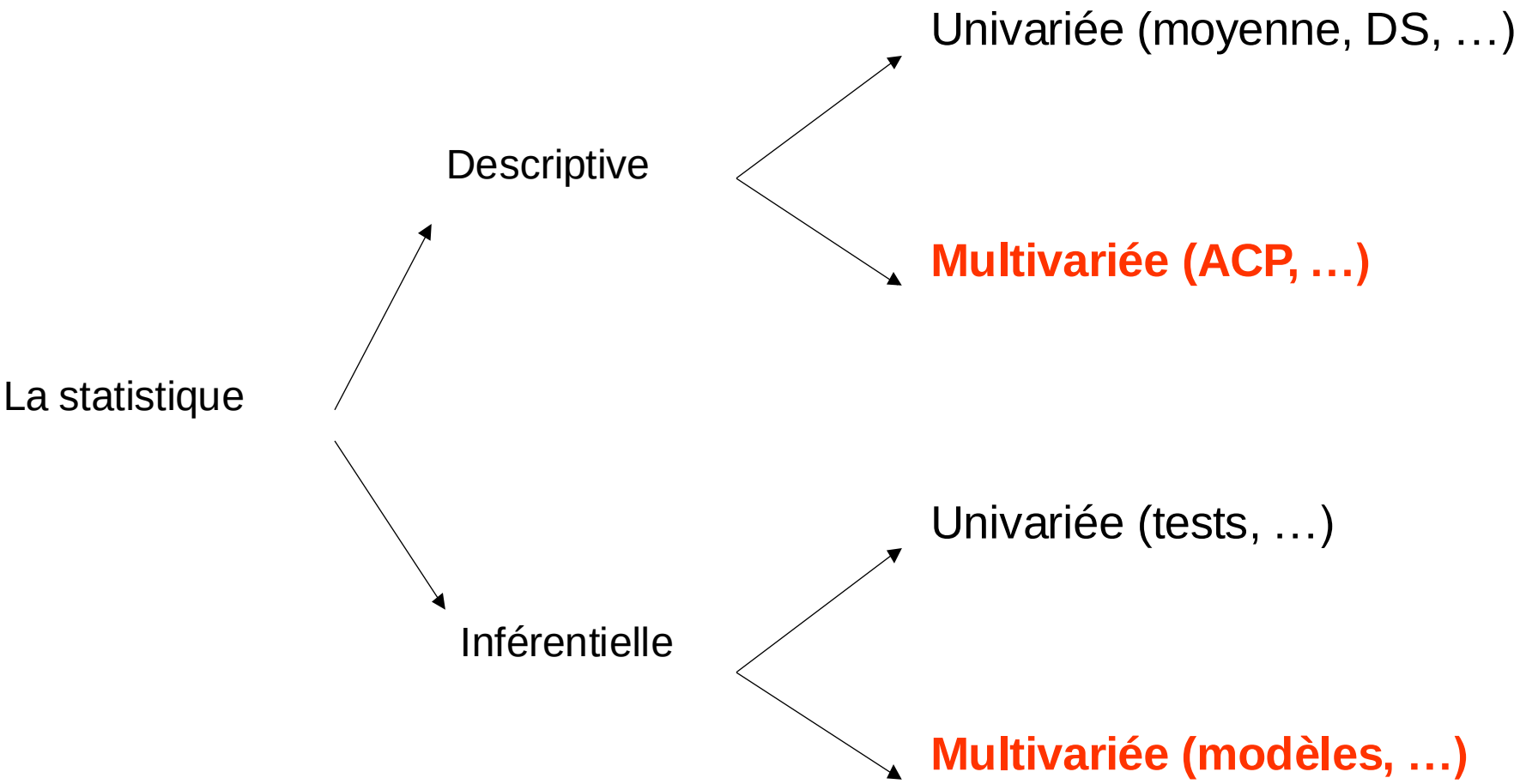
Deux directions en statistique

- **Statistique descriptive** : son but est de décrire, c'àd. de résumer ou représenter par des statistiques les données disponibles quand elles sont nombreuses. Questions types :
  - a. Représentation graphique.
  - b. Paramètres de position et dispersion.
  - c. Divers question liées aux grands jeux de données.
- **Statistique inférentielle** : les données sont considérées incomplètes et elle a pour but de tenter de retrouver l'information sur la population initiale. La prémisse est que chaque mesure est une variable aléatoire suivant la loi de probabilité de la population. Questions types :
  - a. Estimations de paramètres.
  - b. Intervalles de confiance.
  - c. Tests d'hypothèse.
  - d. Modélisation (e.g. régression linéaire).

La statistique peut être

- **Univariée** : il n'y a qu'une seule variable qui rentre en jeu.
- **Multivariée** : plusieurs variables rentrent en ligne de compte.
  - Deux variables entre elles : analyse **bivariée**
  - Plusieurs variables : analyse **multivariée**
    - Une variable expliquée
    - Plusieurs variables explicatives indépendantes deux à deux

Au final

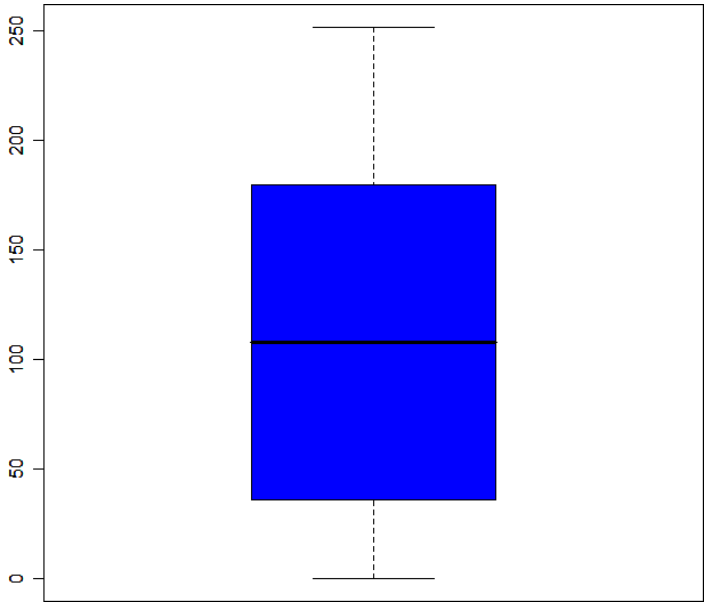


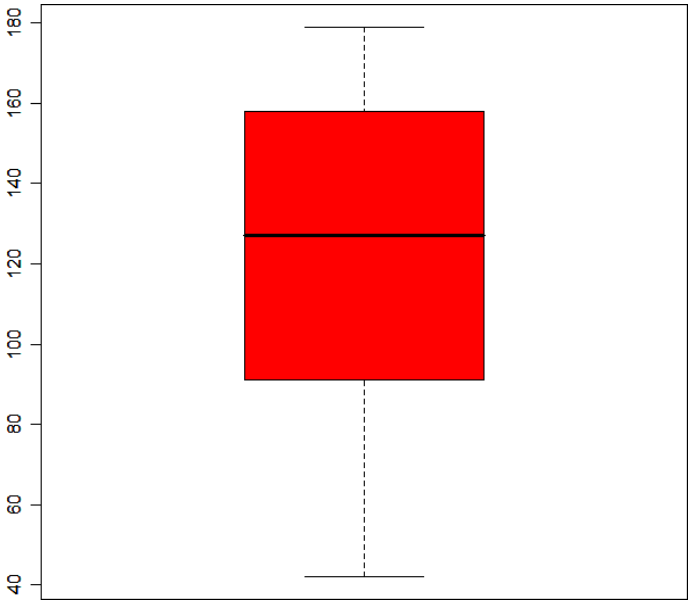
La régression linéaire

La régression linéaire

- Exemple : étude du lien entre la taille et l'âge des filles (en mois), Echantillon de 637 filles
- Questions :
  - Existe-t-il un lien entre la taille et l'âge ?
  - Quand l'âge l'augmente
    - Est-ce que la taille augmente aussi ?
  - Connaissant l'âge, peut-on prédire la taille?
    - But médical : détecter les retards de croissance ?
    - Ou le contraire, quand les médecins légistes retrouvent un « os humain » (complet ou fragment) dans la nature, est-il possible de déterminer l'âge et le sexe ?

La régression linéaire



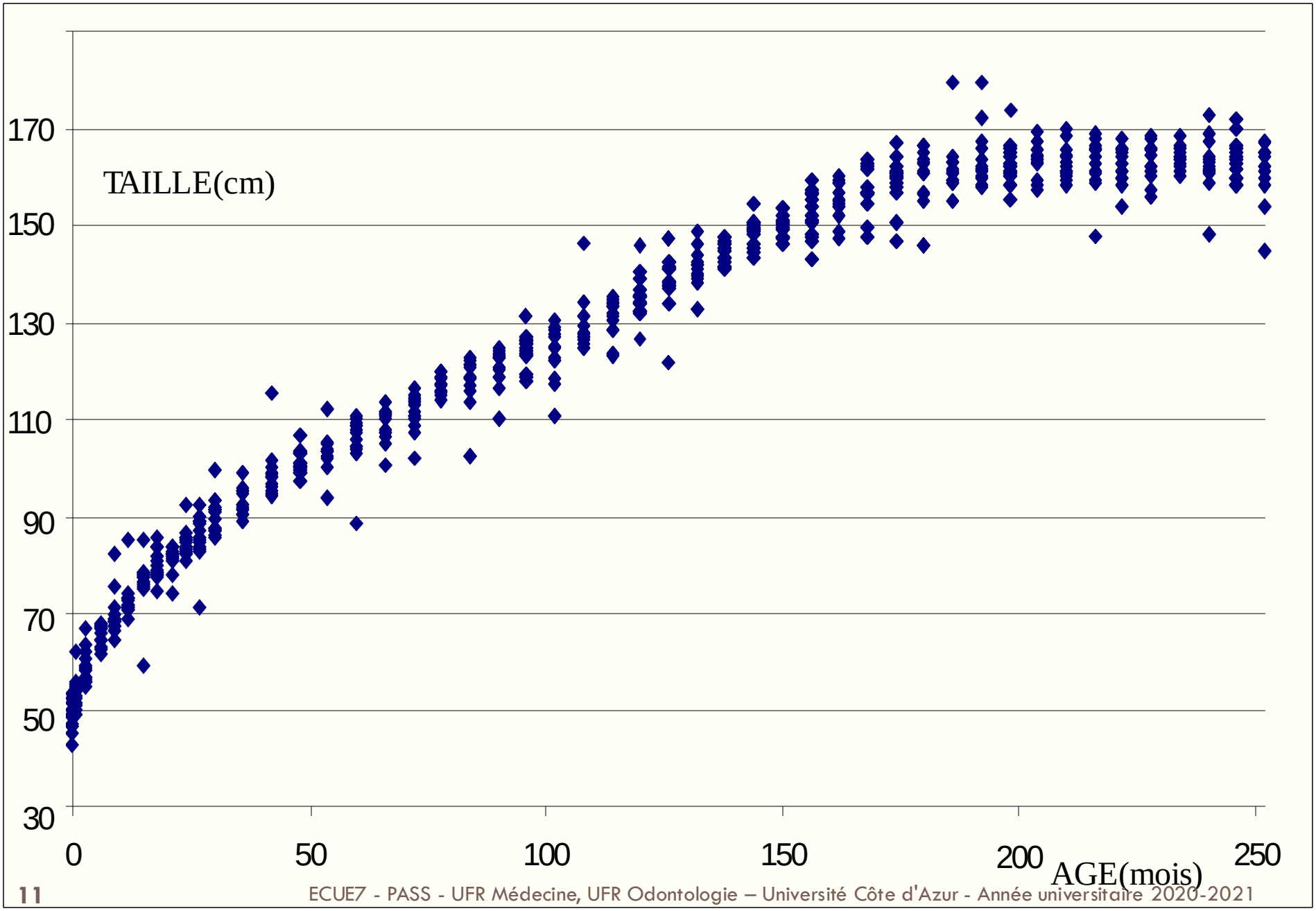


m = **112,12** mois

s^2 = **6265,86** $mois^2$

m= **122,83** cm

s^2 = **1317,43** cm^2

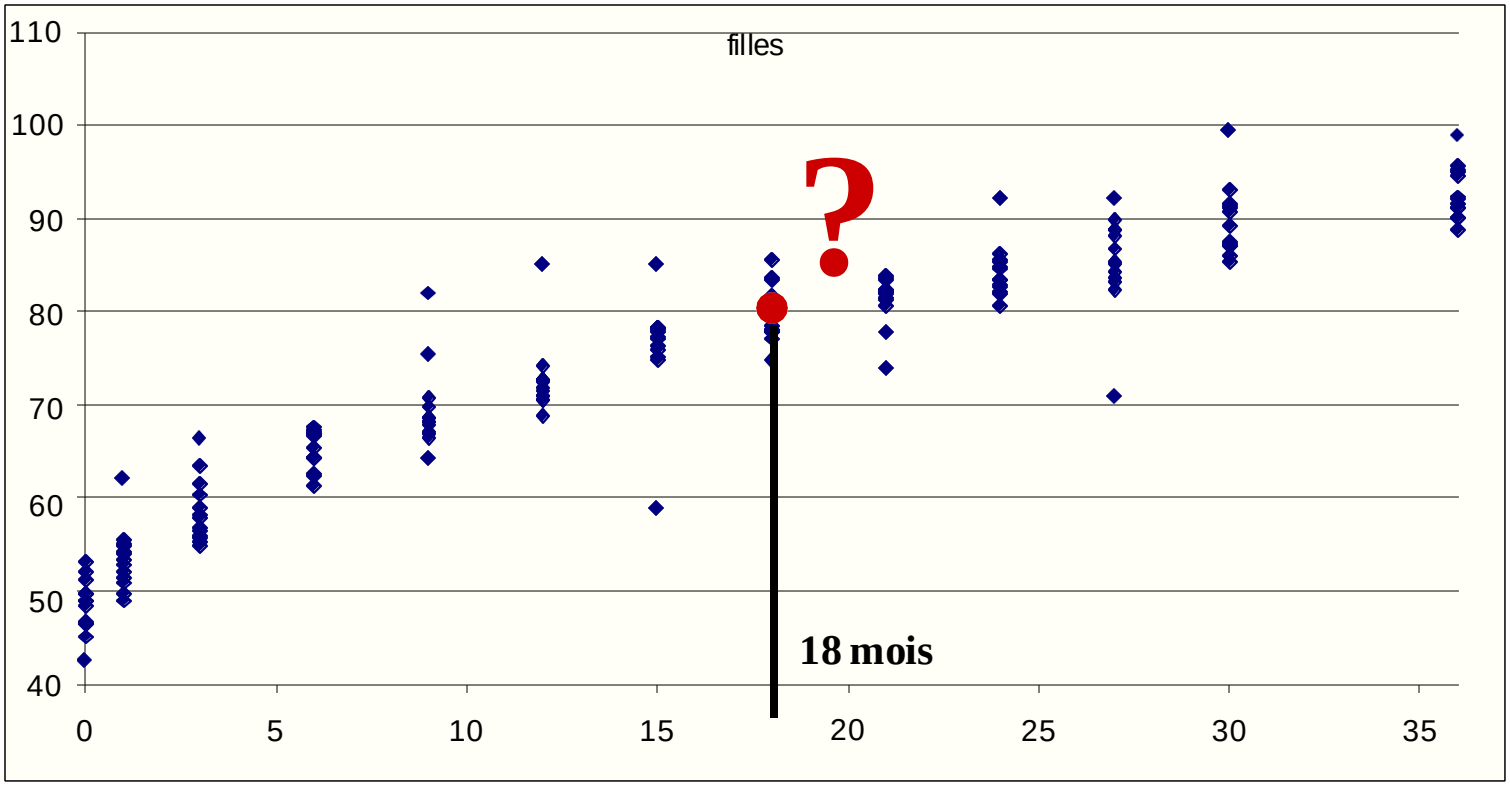

TAILLE(cm)
170
150
130
110
90
70
50
30
0
50
100
150
200
250
AGE(mois)

La régression linéaire

- Comment la Taille évolue en fonction de l'Age ?
  - Taille = f(Age)
  - Autrement dit, pour une variation de Y quelle est la variation de Y ?
  - On parle de Régression de Y en X :
    - Y= taille (cm)
    - X= âge (mois)
  - Comment évolue la Taille en fonction de l'âge ?
    - Pour chaque valeur d'âge (équation)
  - Ou quelle est la taille pour un âge donné (valeur et IC) ?

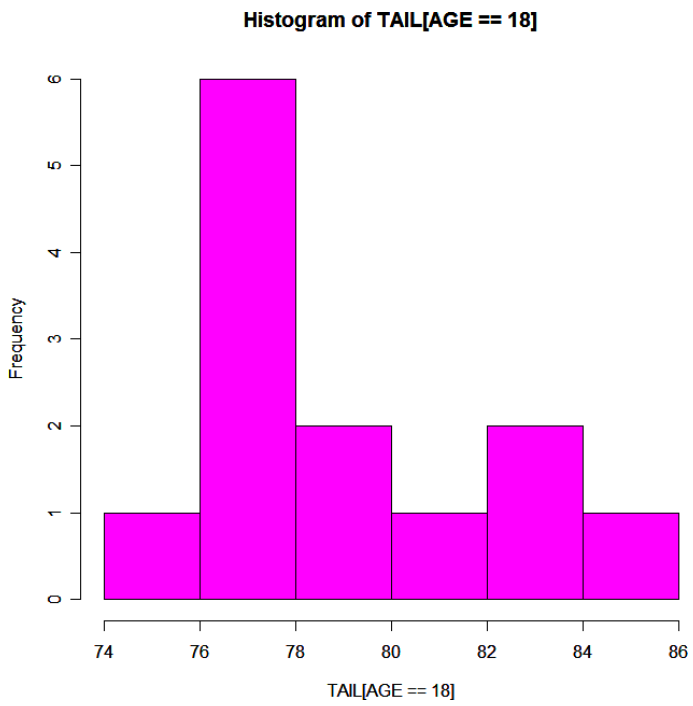
La régression linéaire

- Exemple au sein d'un groupe de filles



La régression linéaire

- Chez les filles de 18 mois,
  - Quelle est la taille moyenne ?
  - Quelle est la variance de la taille ?
  - Quelle est la distribution ?
- Données stratifiées pour 18 mois
  - Moyenne observée :
    - M(T/A=18) = 79,23 cm
  - Variance observée :
    - V(T/A=18) = 9,36 cm^2
  - On parle d'une distribution conditionnelle : valeur de la taille sachant l'âge

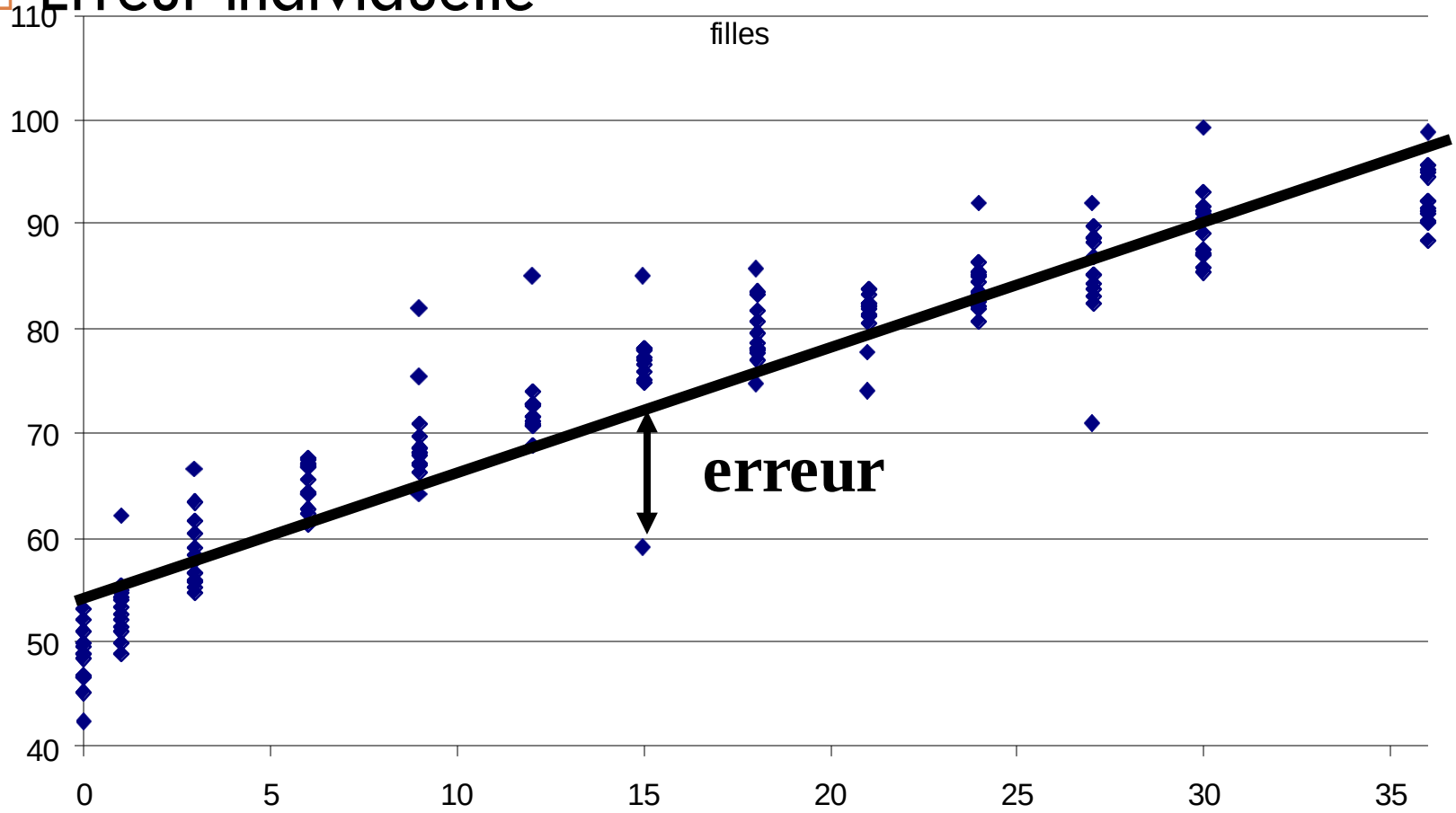


Régression linéaire

- Fonction de régression
  - Taille fonction de l'âge :
    - Moyenne(Taille/Age)= f(Age)
  - Fonction f() : droite affine (de type $y = ax + b$)
    - Espérance(Taille / Age) = $\alpha + \beta \times$ Age
  - Pour chaque sujet
    - Taille = $\alpha + \beta \times$ Age + ε
    - ε = erreur individuelle

La régression linéaire

- Erreur individuelle



La régression linéaire

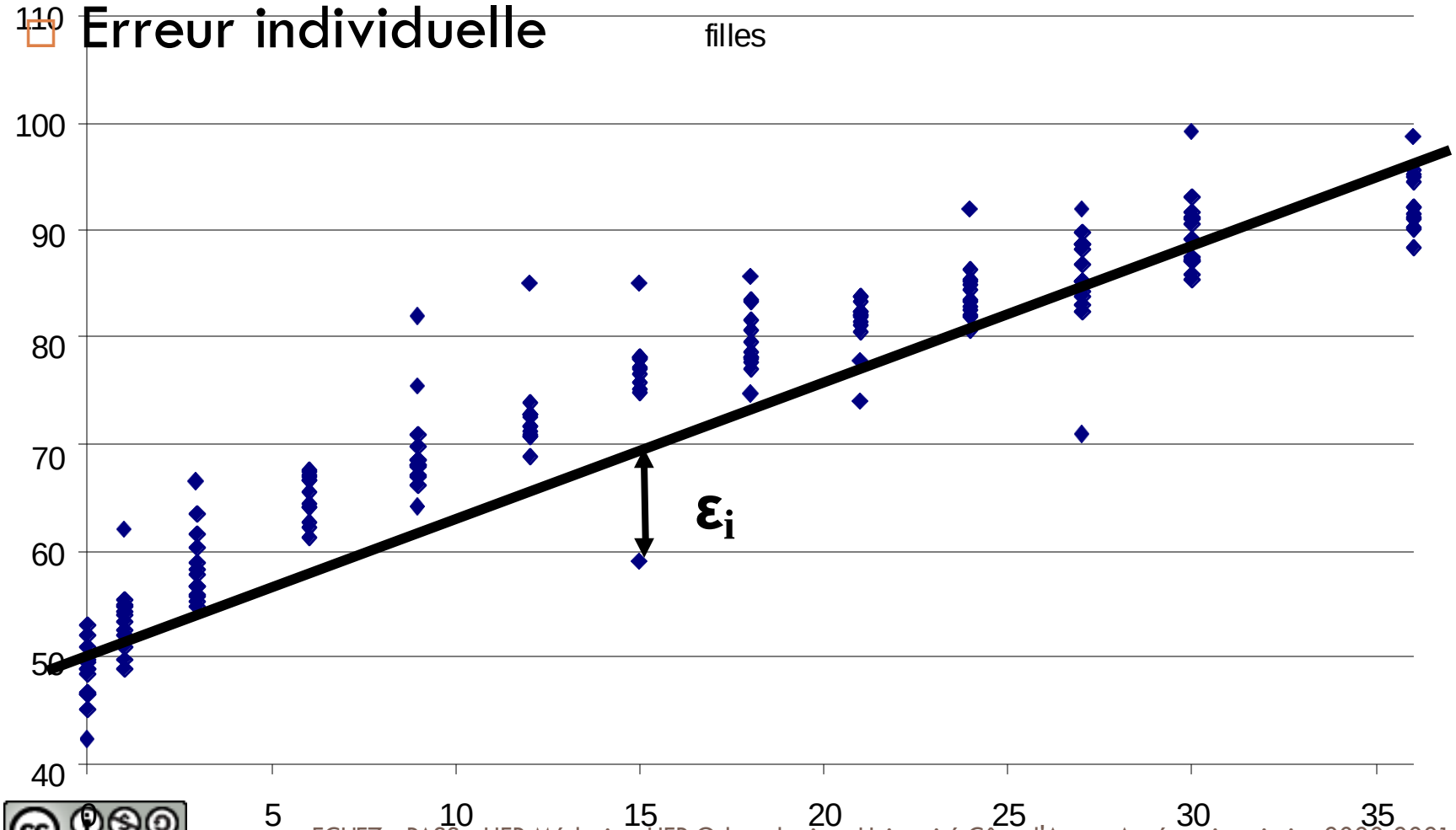
- Est le modèle le plus simple pour permettre
  - une interprétation (lien ou non entre les deux variables) (valeur du « coefficient de régression » qui englobe dans son calcul la pente de la droite donc la valeur de bêta)
  - une estimation de alpha et bêta pour que la droite d'ajustement « minimise » l'erreur individuelle
- Cette droite d'ajustement est appelée aussi droite de régression : on dit qu'elle « résume » le mieux le nuage de point
- Elle permet aussi la prédiction et l'extrapolation

La régression linéaire

- Principe de l'estimation
  - Estimer α et β tel que ε petits +++
  - ε_i : écart entre la droite et le point i
  - Pour chaque valeur de X on a $y_i = \alpha + \beta \times x_i + \varepsilon_i$
  - Or $E(Y / X) = \alpha + \beta \times X$

$$\Rightarrow \varepsilon_i = y_i - E(Y / X)$$

La régression linéaire



La régression linéaire

- Principe de l'estimation
  - Calcul de la somme des carrés des écarts

$$SCE = \sum_{i=1}^{n} (\varepsilon_i)^2$$

  - Estimer α et β tel que SCE soit la plus petite

La régression linéaire

- Estimation de la pente β
- $\beta = \frac{cov(XY)}{var(X)}$
- Covariance de la taille et de l'âge :
  - cov(TAIL,AGE) = 2742.587
- Variance de l'âge
  - var(AGE)
- Estimation de β
  - β = cov(TAIL,AGE)/var(AGE) β =0.437703

La régression linéaire

- Estimation de α :
  - La droite passe par mY et mX
  - $mY = \alpha + \beta mX$
  - $\alpha = mY - \beta mX$
  - $\alpha = 73.729$
- L'équation s'écrit donc :
  - Taille = 73.73 + 0.44 Age + ε

ou

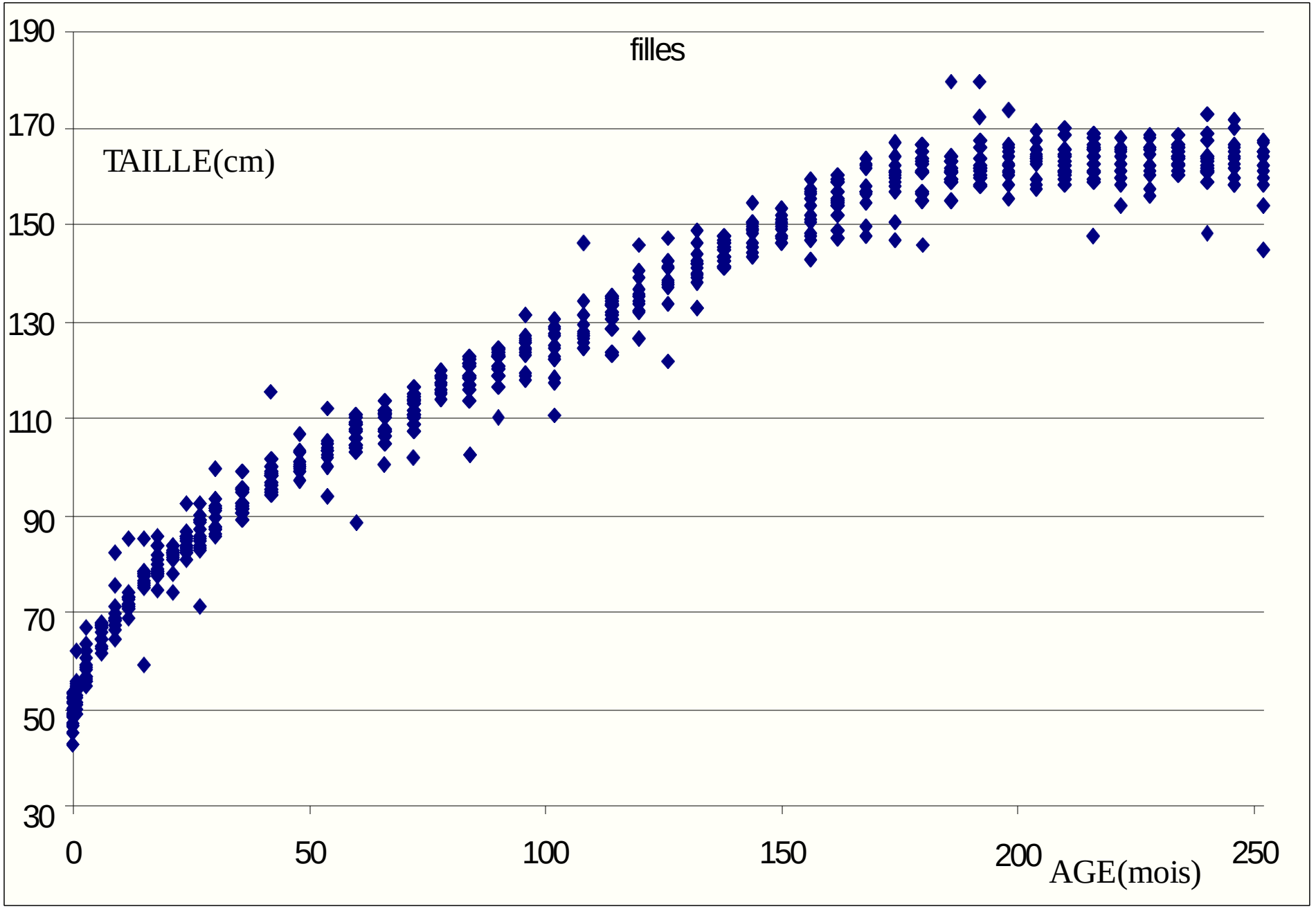
  - E(Taille/Age) = 73.73 + 0.44 Age

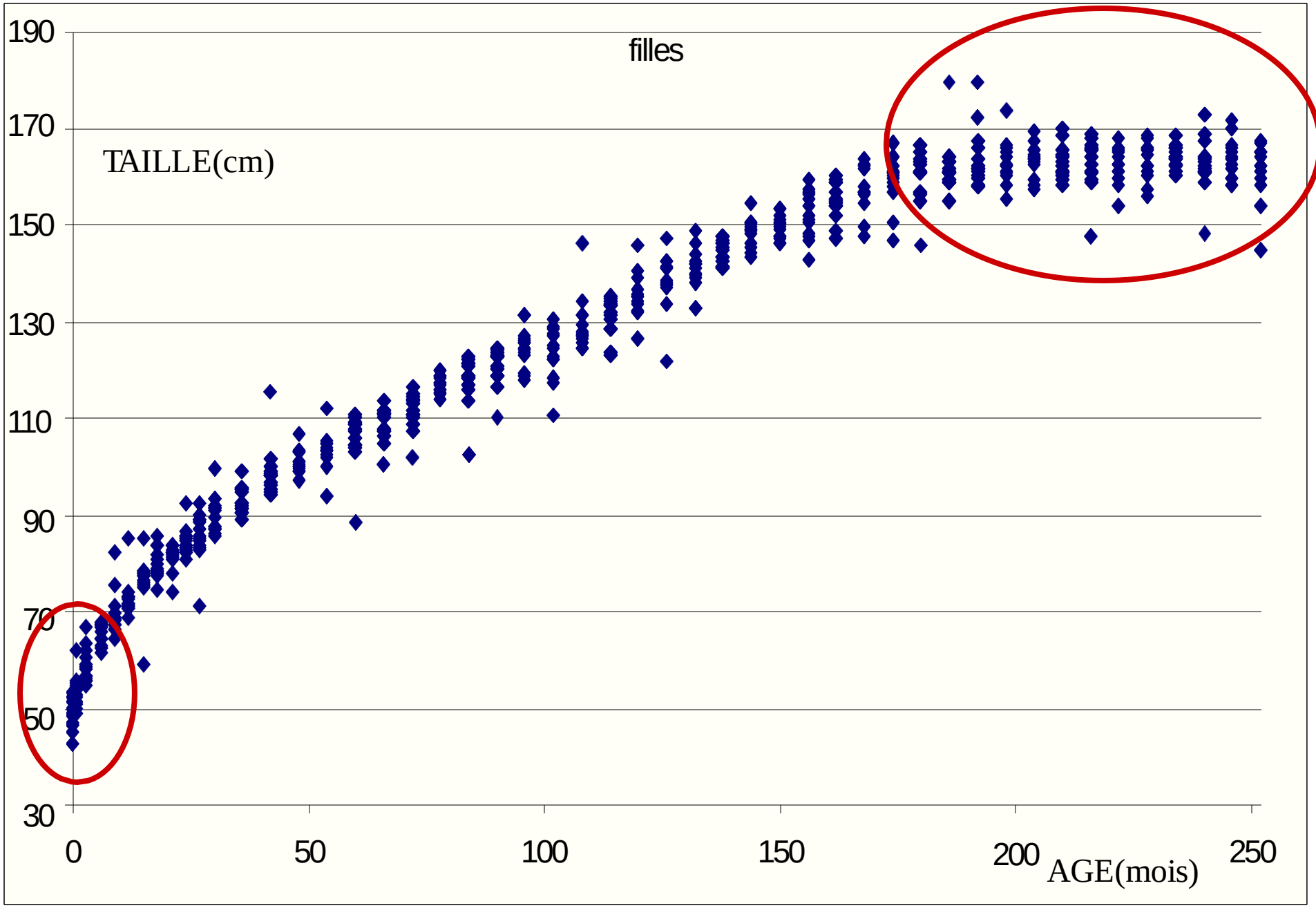
La régression linéaire

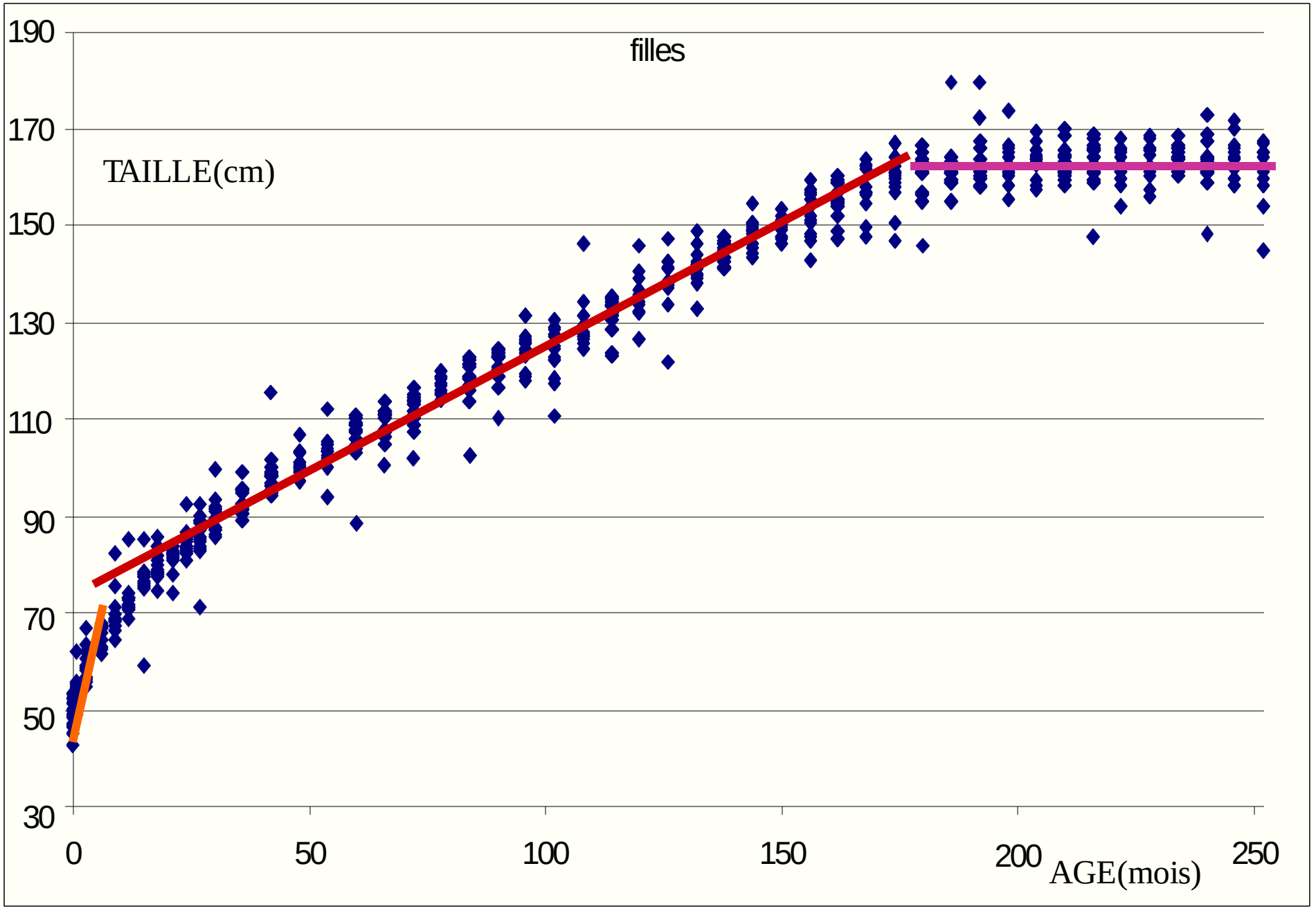
- Interprétation
  - Pente β:
    - β=0: pas de lien, évolutions indépendantes
    - β<0 : évolutions en sens contraire
    - β>0 : évolutions dans le même sens
  - Ordonnée à l'origine
    - $E(Y/X=0) = \alpha$

La régression linéaire

- Test de la pente à 0
  - Si β=0 => pas de lien entre Y et X
- Lien entre Y et X est-il significatif?
  - C'est-à-dire est-ce que β≠0?
  - b estimation de β
  - Hasard => fluctuation de b observé
- H_0 : β=0, il n'y a pas de lien entre X et Y
- H_1: β≠0, il y a un lien entre X et Y
- Sous H_0 et si les conditions d'application sont respectées,
  - la statistique $t_0 = \frac{b-\beta}{\sqrt{{s_b}^2}}$ suit une loi de Student à n-2 ddl
    - L(Y/X)~>N
    - V(Y/X) constantes pour tout X
    - à X donné, Y_i indépendants
    - La régression est linéaire


filles
TAILLE(cm)
190
170
150
130
110
90
70
50
30
0
50
100
150
200
250
AGE(mois)


filles
TAILLE(cm)
190
170
150
130
110
90
70
50
30
0
50
100
150
200
250
AGE(mois)


filles
TAILLE(cm)
190
170
150
130
110
90
70
50
30
0
50
100
150
200
250
AGE(mois)

La régression linéaire

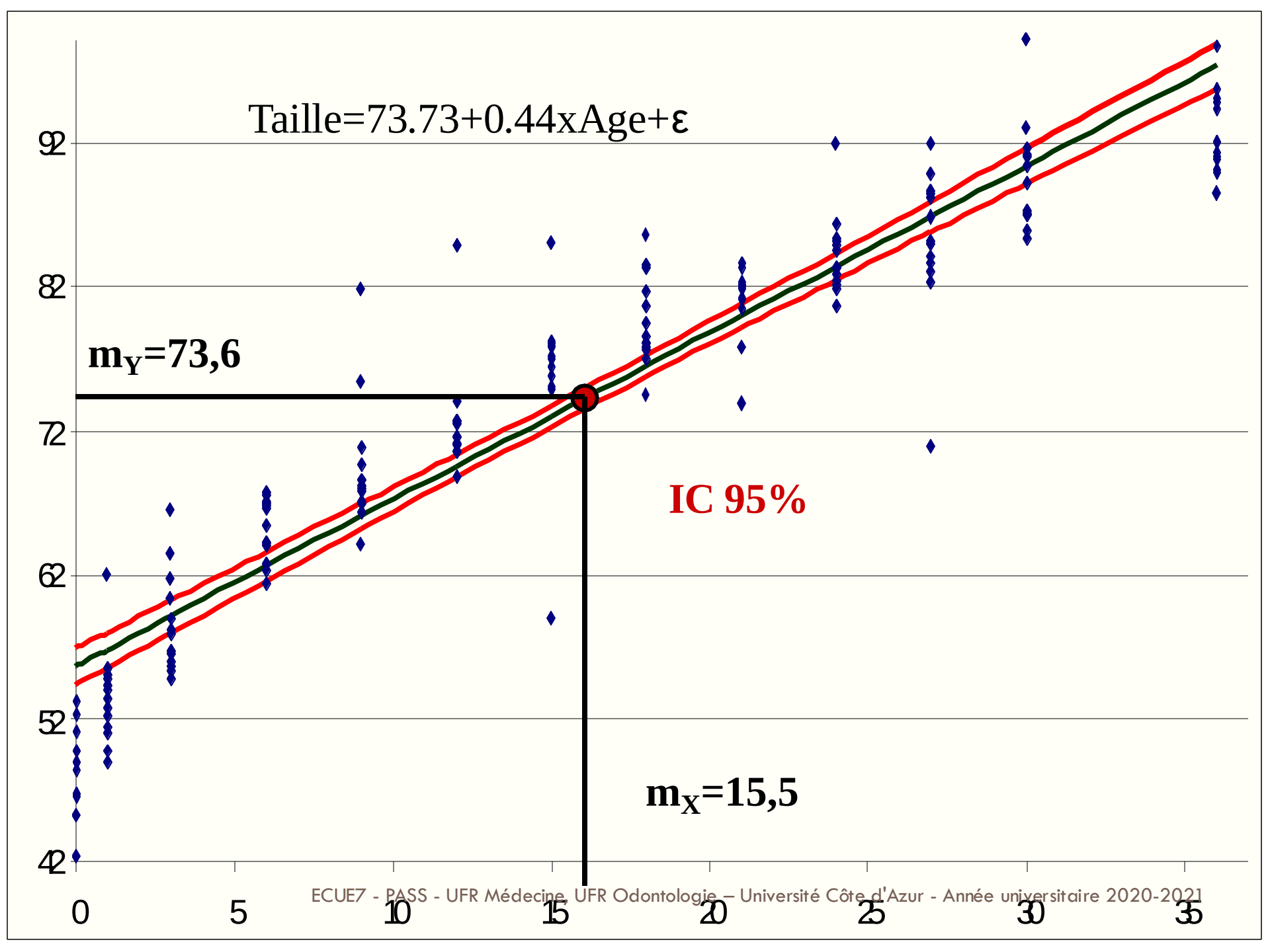
- Hasard => fluctuation de b
- Intervalle de confiance de la pente
- b ~> t_{n-2}
- $b \pm t_{n-2,\frac{\alpha}{2}} \times \sqrt{{s_b}^2}$
- Pour ce qui est de l'exemple :

	2.5 %	97.5 %
(Intercept)	72.2707108	75.1872989
AGE	0.4270751	0.4483309

- L'intervalle de confiance à 95% de b ne contient pas la valeur 0
  ➔ b ≠ 0 au risque de 5% de se tromper

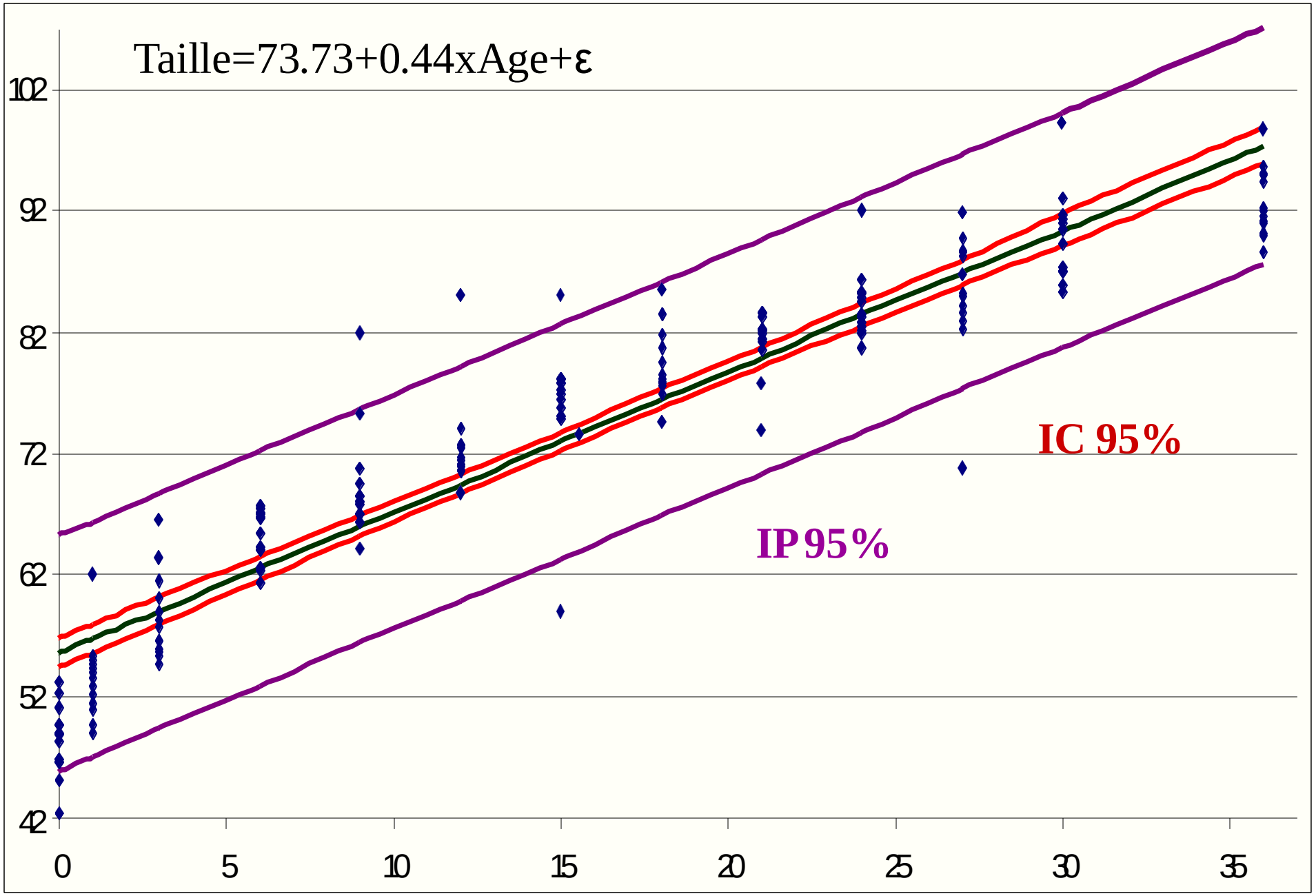
La régression linéaire

- Intervalle de confiance de la droite
  - E(Y / X) = α + βX
- Estimé par : $m_{Y/X} = a + bX$
- $m_{Y/X} \pm t_{n-2,\frac{\alpha}{2}} \times \sqrt{s_{m_{Y/X}}{}^2}$


Taille=73.73+0.44xAge+ε
92
82
72
62
52
42
0
5
10
15
20
25
30
35
m_Y=73,6
IC 95%
m_X=15,5

La régression linéaire

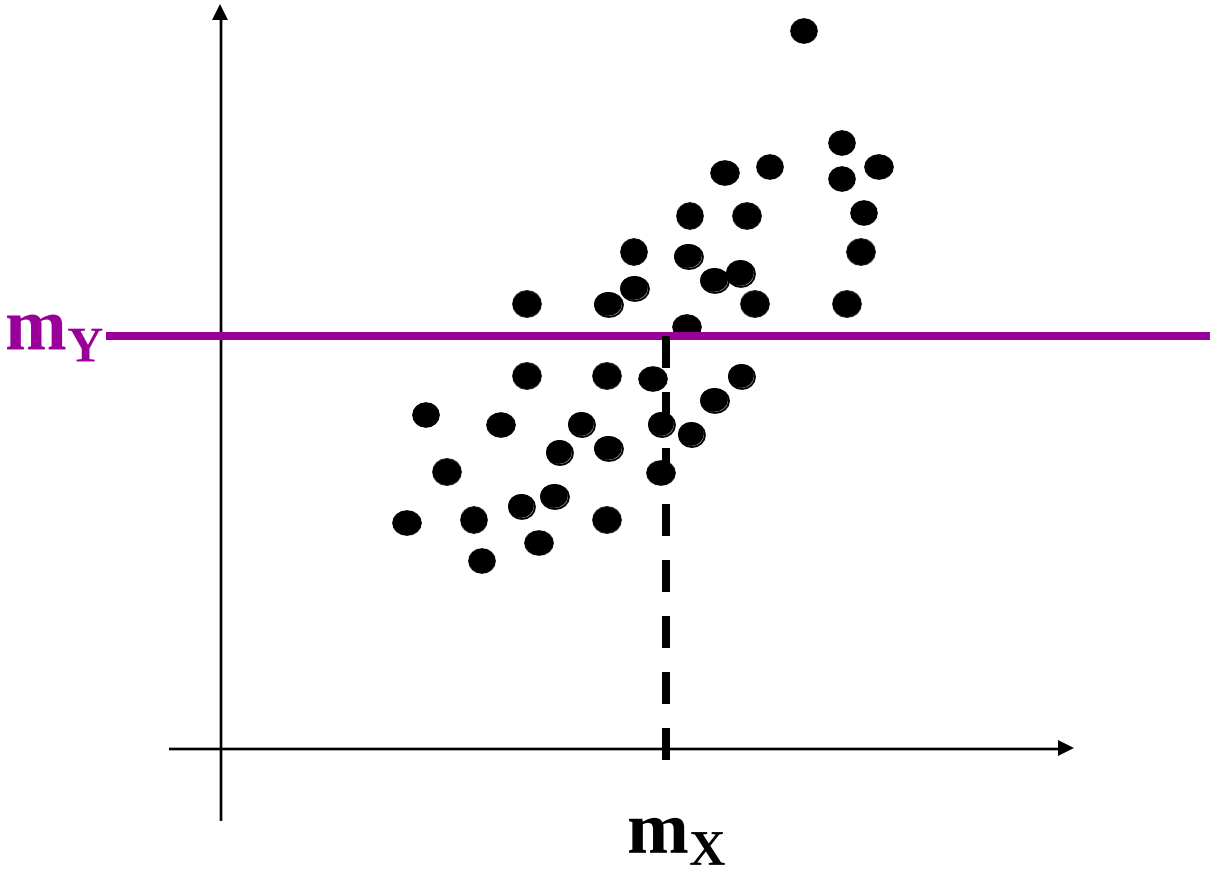
- Intervalle de prédiction
- Pour un Age (X) fixé, prédiction de la Taille (Y)
  - Yp = a + bX
  - $\text{Taille}_p = 73{,}73 + 0{,}44\ \text{Age}$
- Précision de la prédiction
  - $y_p \pm t_{n-2,\frac{\alpha}{2}} \times \sqrt{{s_{y_p}}^2}$


Taille=73.73+0.44xAge+ε
102
92
82
72
62
52
42
0
5
10
15
20
25
30
35
IC 95%
IP 95%

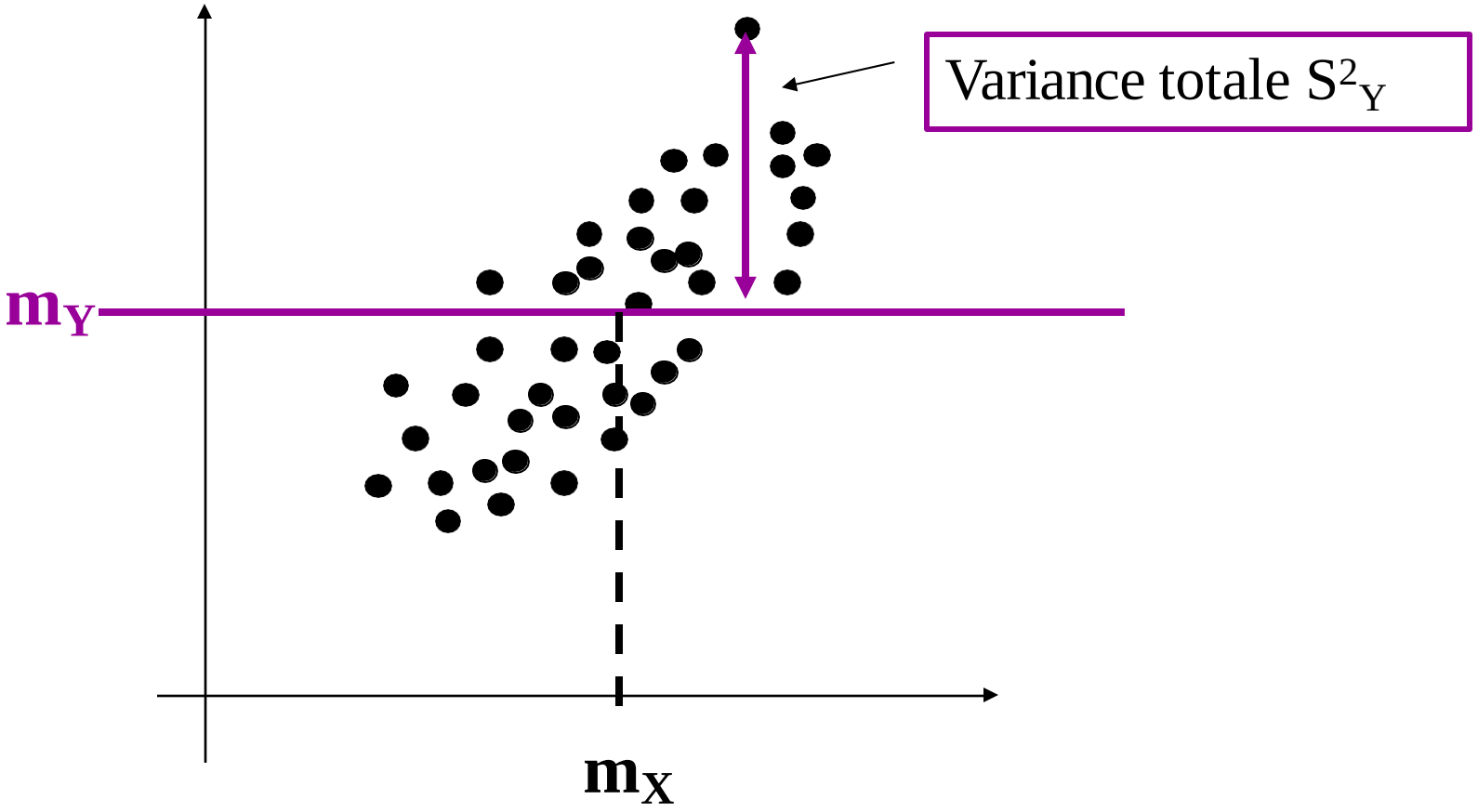
La régression linéaire

- Adéquation du modèle : le modèle est-il un bon résumé des observations ?
  - Calcul du pourcentage de variance expliquée R^2
    - $R^2 = \frac{Part\ de\ variance\ expliquée\ par\ la\ régression}{Variance\ totale}$
    - Variance totale $= S^2{}_Y$

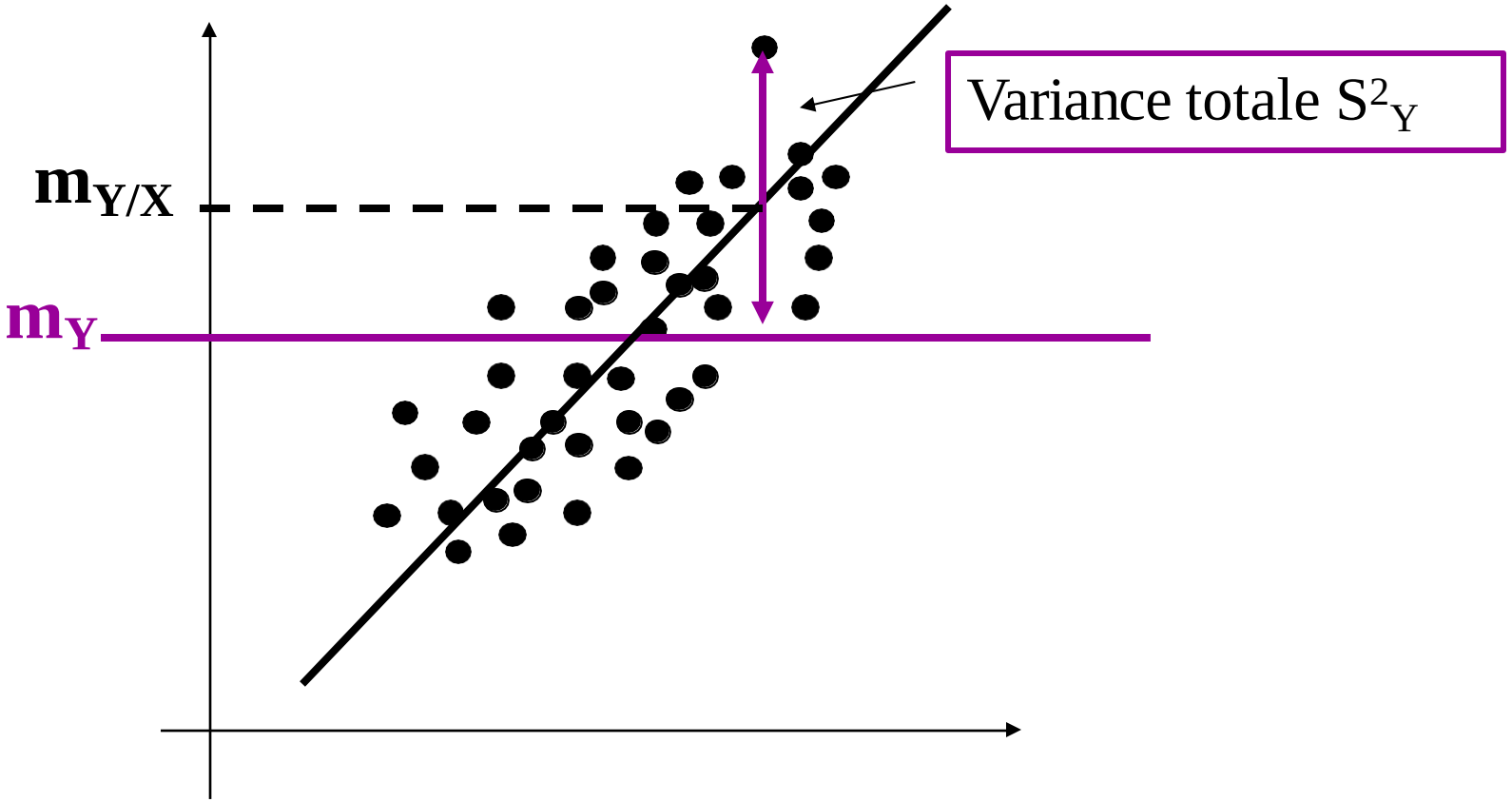
La régression linéaire



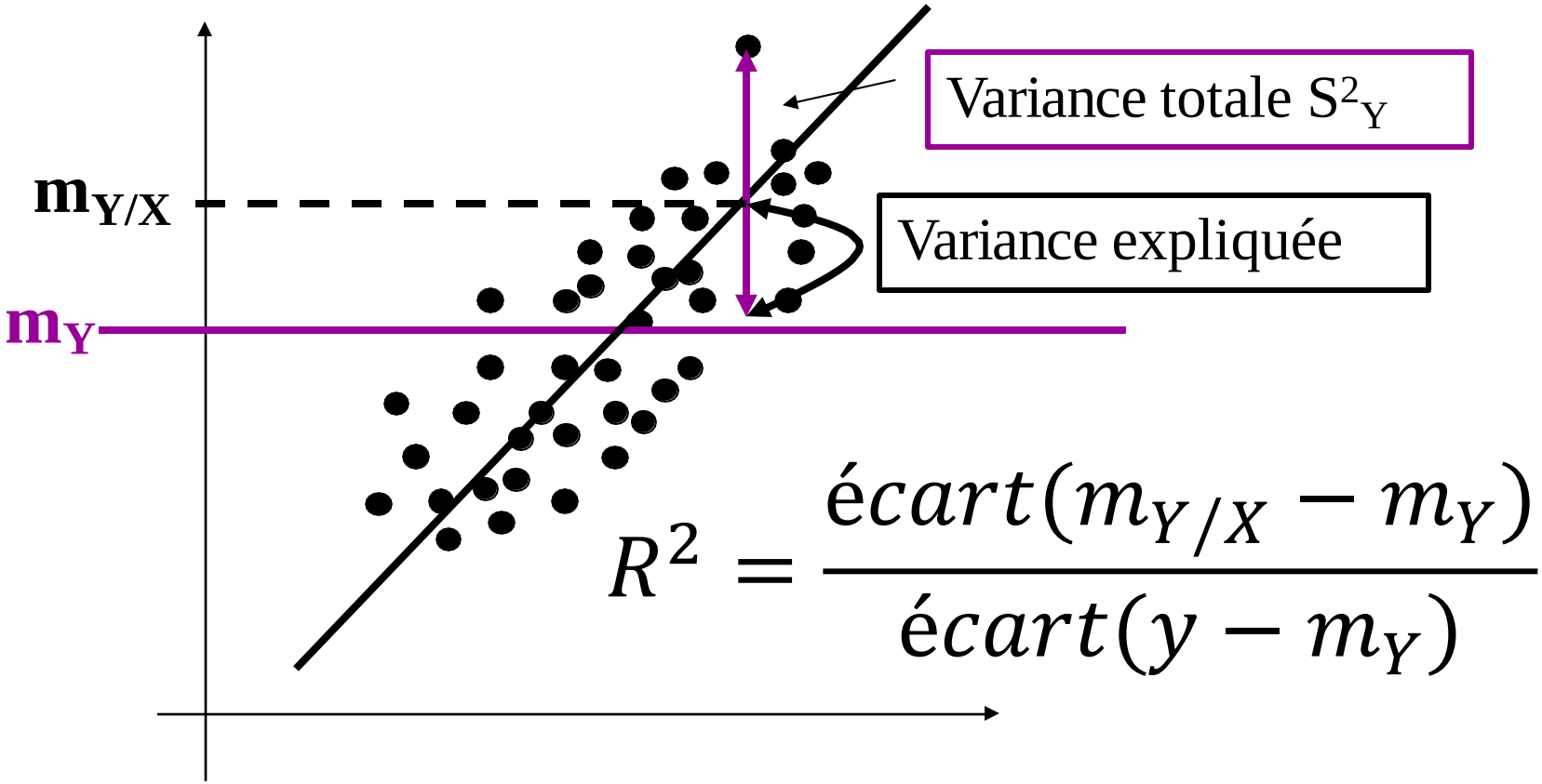
La régression linéaire



La régression linéaire



La régression linéaire



La régression linéaire

- Pourcentage de variance expliquée

$$R^2 = \frac{\sum(m_{Y/x_i} - m_Y)^2}{\sum(y_i - m_Y)^2}$$

- Exemple : $R^2 = 88\%$
- Remarque :
  - $\sqrt{R^2}$ = estimation du coefficient de corrélation entre X et Y

La régression logistique

La régression logistique

- Variable à expliquer Y : binaire (Malade oui/non)
- ⇒ Conditions d'application de la régression linéaire non remplies
- Variables explicatives X : quantitatives ou qualitatives
- $Y = f(X_1; X_2; \ldots X_n)$
- Expliquer Y ⇒ Quantifier l'association $Y \leftrightarrow x_i$
- Prédire Y à partir de nouvelles observations de x_i

La régression logistique

- Exemple : décès en fonction d'une dose de toxique

Dose X	0	10	30	50	70	90	100
effectif	30	30	30	30	30	1	29
décès	0	3	3	13	25	1	28
p	0	0.1	0.1	0.43	0.83	1	0.97

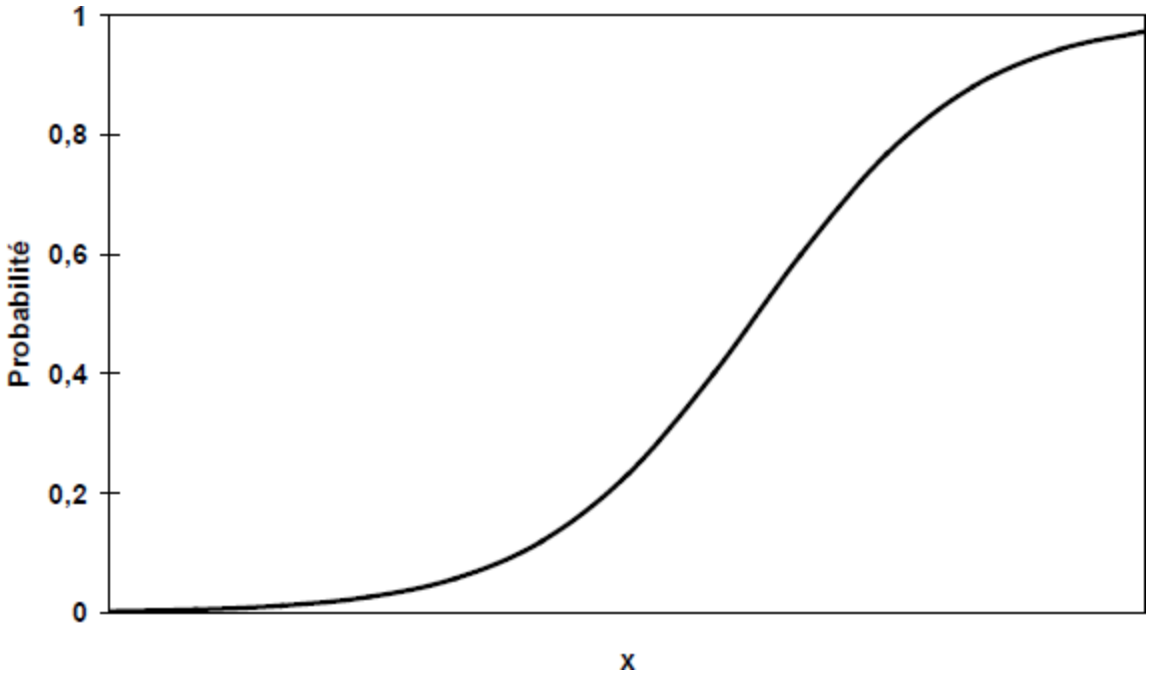
- Comment varie la proportion de décès en fonction de la dose toxique?

$$logit(p)=ln(p/1-p)=\alpha+\beta X$$

La régression logistique

- Rappel : l'estimation d'une probabilité est un rapport
- Pour pouvoir « transformer » un rapport en somme, on passe par la fonction logarithme
  - Log(A/B) = LogA – LogB
- La fonction logit donne le log népérien de la cote d'un événement, cad le rapport p/1-p
  - logit(p) = ln(p / (1-p))

La régression logistique



$$p = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

La régression logistique

- Chez les exposés
  - E=1
  - Probabilité d'être malade : p_+

$$p_+ = p(M^+/E = 1) = \frac{1}{1 + e^{-(\alpha+\beta)}}$$

  - Probabilité de ne pas être malade : 1- p_+

$$1 - p_+ = p(M^-/E = 1) = \frac{e^{-(\alpha+\beta)}}{1 + e^{-(\alpha+\beta)}}$$

- Chez les non-exposés
  - E=0
  - Probabilité d'être malade : p_-

$$p_- = p(M^+/E = 0) = \frac{1}{1 + e^{-\alpha}}$$

  - Probabilité de ne pas être malade : 1- p_-

$$1 - p_- = p(M^-/E = 0) = \frac{e^{-\alpha}}{1 + e^{-\alpha}}$$

La régression logistique

- L'OR = force du lien entre X et Y
- OR = Rapport de cotes
- Déterminé à partir de l'estimation des paramètres

$$OR = \frac{\frac{p+}{(1-p+)}}{\frac{p-}{(1-p-)}} = e^{\beta}$$

La régression logistique

- Conditions d'application
  - Relation linéaire entre logit(p) et X
- Y binomial ou multinomial
- Codage « intelligent » des X catégoriels pour pouvoir interpréter les coefficients
- Indépendance des individus

La régression logistique

- Exemple: facteurs d'hypotrophie à la naissance

Id	hypo	tabac	prema	hta	visite	poidsbb	age	poidsmer
1	0	0	0	0	0	2523	18	82,27
2	0	0	0	0	3	2551	34	70
3	0	1	0	0	1	2557	19	47,73
4	0	1	0	0	2	2594	22	49,09
5	0	1	0	0	0	2600	18	48,18
6	0	0	0	0	0	2622	21	56,82
7	0	0	0	0	1	2637	22	53,64

La régression logistique

- Le poids de la mère est-il un facteur de risque d'hypotrophie?

Logit(p)=α+β.POIDSMER

```
glm(formula = hypo ~ poidsmer, family = "binomial")
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.108  -0.914  -0.800   1.348   1.982
Coefficients:
             Estimate   Std. Error   z value Pr(>|z|)
(Intercept)  1.06467    0.78426      1.358   0.1746
poidsmer     -0.03183   0.01358      -2.344  0.0191 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 236.99 on 189 degrees of freedom
Residual deviance: 230.63 on 188 degrees of freedom
AIC: 234.63
Number of Fisher Scoring iterations: 4
```

$OR=e^{-0.03}=0.97$

CI95% β:
b+-1.96.0.014
⇒CI95% OR:
⇒[0.94 ; 0.99]

La régression logistique

- Interprétation
  - $p<0.05$
  - L'OR est significativement différent de 1
  - Il existe un lien significatif entre le poids de la mère et l'hypotrophie
  - Dans le sens : quand le poids de la mère augmente, le risque d'hypotrophie diminue
    - OR=0.97 (95% CI [0.94;0.99])

La régression logistique

La régression logistique

- Interprétation de l'OR
  - Pour chaque unité de poids maternel, le risque d'hypotrophie diminue de 0.97
  - Hypothèse d'un OR constant quelque soit le poids maternel
  - ⇒relation linéaire entre le risque d'hypotrophie et le poids maternel
  - Si non => Modification du codage du poids maternel

La régression linéaire multiple

Régression linéaire multiple

- Plusieurs causes dans l'évolution de la taille Y :
  - Age (X_1)
  - Facteurs socio-économiques (X_2)
  - Taux d'hormones de croissance (X_3)
- $E(Y / X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- Estimation
  - $\alpha, \beta_1, \beta_2, \beta_3$ estimés en tenant compte des 3 VA X_1, X_2, X_3
- On parle d'ajustement
- On peut envisager des interactions
  - $E(Y / X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$

Régression linéaire multiple

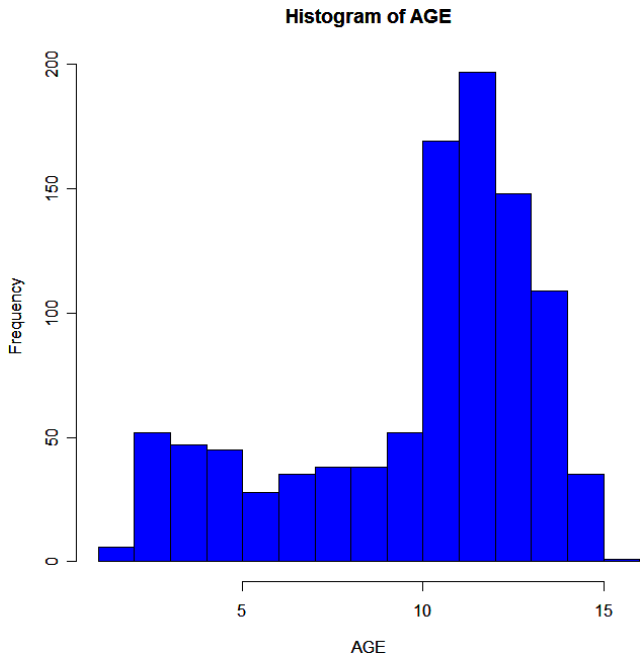
- Tests des β_1, β_2, β_3 à 0
- Interprétation identique
- Adéquation identique
- Approche pas à pas
- Choix des variables : notion de modèle
- Variables très corrélées

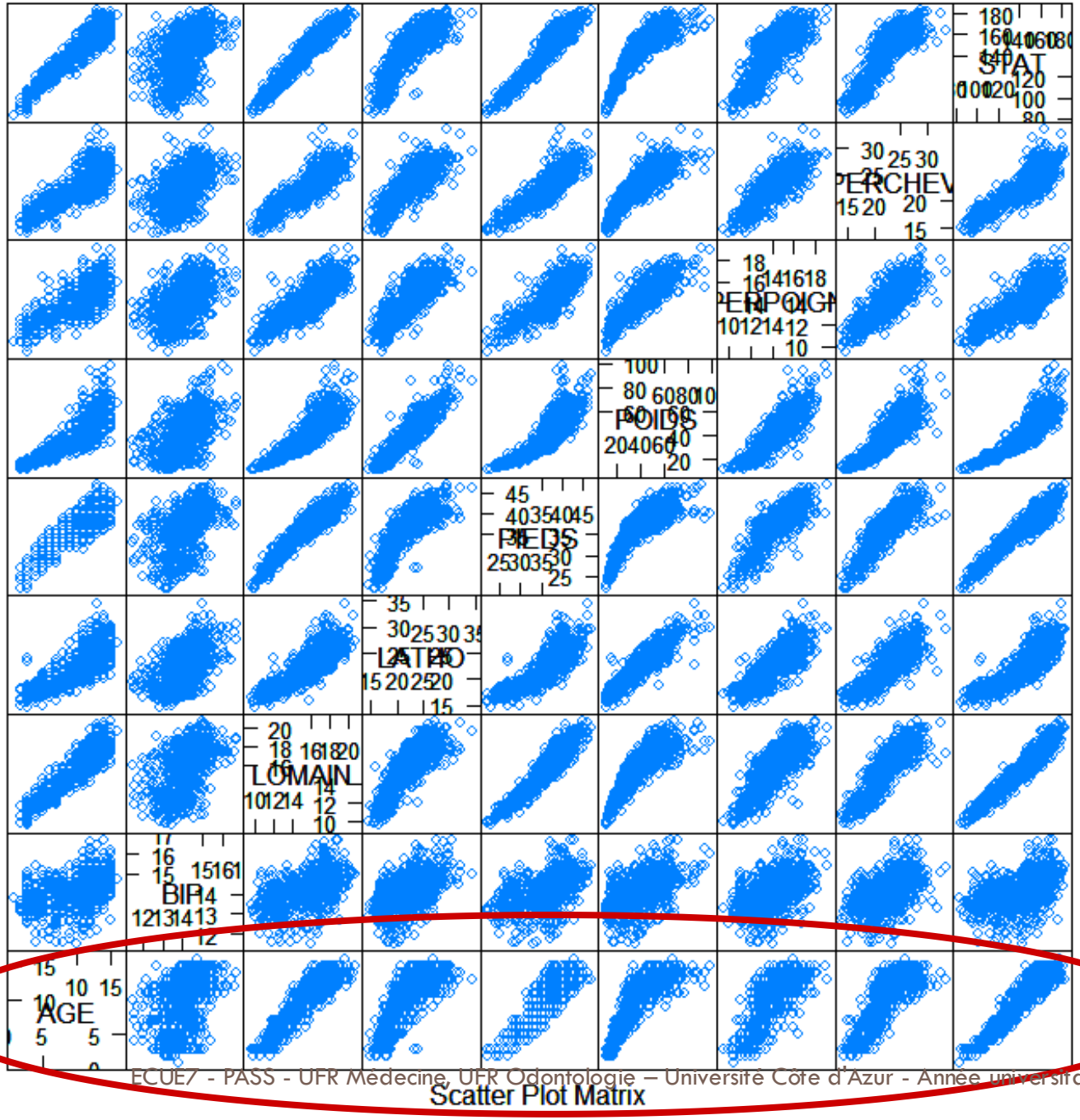
Régression linéaire multiple

- Exemple : Prédire l'âge en fonction de 8 mesures
  - Crâne (BIP)
  - Tronc (LATHO)
  - Membres supérieurs et inférieurs (LOMAIN, PERPOIGN, PERCHEV, PIEDS)
  - Globales (STAT, POIDS)
- Echantillon de 1000 enfants de 2 à 16 ans

Régression linéaire multiple

- En moyenne :
  - $AGE = \alpha + \beta_1 xBIP + \beta_2 xLATHO + \beta_3 xLOMAIN + \beta_4 xPERPOIGN + \beta_5 xPERCHEV + \beta_6 xPIEDS + \beta_7 xSTAT + \beta_8 xPOIDS$
- Statistiques descriptives
  - mean(AGE) = 10.373
  - var(AGE) = 11.53541




STAT
PERCHEV
PERPOIGN
POIDS
PIEDS
LATPAD
LONGMAIN
BIP
AGE
Scatter Plot Matrix

Call: glm(formula = AGE ~ 1 + BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV + PIEDS + STAT + POIDS, family = gaussian)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.12658	-0.72416	-0.04954	0.67239	4.36643

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.300e+01	8.684e-01	-14.966	< 2e-16 ***
BIP	**3.312e-02**	**5.423e-02**	**0.611**	**0.54156**
LATHO	**1.219e-01**	**2.659e-02**	**4.583**	**5.17e-06 *** **
LOMAIN	**1.013e-01**	**5.947e-02**	**1.704**	**0.08877 .**
PERPOIGN	**-1.370e-01**	**4.695e-02**	**-2.917**	**0.00361 ****
PERCHEV	**-4.654e-02**	**2.597e-02**	**-1.792**	**0.07341 .**
PIEDS	**7.823e-04**	**2.612e-02**	**0.030**	**0.97611**
STAT	**1.546e-01**	**7.263e-03**	**21.289**	**< 2e-16 *** **
POIDS	**-2.047e-02**	**7.153e-03**	**-2.861**	**0.00431 ****

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AGE=-13+0,03BIP+0,1LATHO+0,01LOMAIN-0,14PERPOIGN-0,05PERCHEV +0,001PIEDS+0,2STAT-0,02POIDS

AIC: 3010.6

Number of Fisher Scoring iterations: 2

Régression linéaire multiple

- Que faut-il regarder ensuite ?
  - conditions d'application
  - intervalles de confiance des paramètres
  - adéquation : R^2

Régression linéaire multiple

- Intervalle de confiance des paramètres et R^2

	2.5 %	97.5 %
(Intercept)	-14.70058351	-11.292408725
BIP	-0.07330209	0.139535454
LATHO	0.06968244	0.174040764
LOMAIN	-0.01538828	0.218001320
PERPOIGN	-0.22908876	-0.044831392
PERCHEV	-0.09750881	0.004420695
PIEDS	-0.05047023	0.052034764
STAT	0.14037312	0.168879663
POIDS	-0.03450573	-0.006430739

Adéquation: R^2 0.8989102

Sélection des variables du modèle

- Guillaume d'Ockham, 1285-1349
- « Les multiples ne doivent pas être utilisés sans nécessité »
  - = principe de parcimonie
  - => ne pas ajouter de nouvelles variables tant que celles présentes suffisent
  - => balance entre explication / prédiction
    - trop de variables : explication + / prédiction –
- overfitting~hyperadéquation
- Sélection de variables : pas à pas (stepwise)
  - Ascendant: on ajoute les variables une à une
  - Descendant: on retire les variables une à une
  - Double sens

Sélection des variables du modèle

- Critère de sélection
  - Calcul d'un « score » AIC (Akaike Information Criterion)
  - $AIC = 2p - 2\ln(L)$
    - p = nombre de paramètres
    - L = vraisemblance du modèle
  - On veut le AIC le plus petit possible
  - Dans l'exemple précédent :
    - AGE ~ 1 + STAT : AIC = 3039.4
    - AGE ~ 1 + BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV + PIEDS+ STAT + POIDS : AIC = 3010.55
    - AGE ~ BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV + STAT + POIDS : AIC = 3008.55

La régression logistique multiple

Régression logistique multiple

- L'hypotrophie à la naissance dépend-elle du tabagisme, de l'HTA, de l'âge maternel et du poids maternel?
- Attention **interactions** entre : hta et tabac, hta et poids
- **Analyse univariée**
  - HTA : p=0.054 (OR=3.28 [0.85; 13.72] *test exact de Fisher*
  - TABAC : p=0.02 (OR=2.08) [1.07;4.08] *test du Chi2 de Pearson*
  - AGE : p= 0.08 (D=1.03[-0.17; 2.89]) *test t de Student, après vérif. Bartlett, n>30*
  - POIDSMAT : p=0.01 (D=5,18[1.24;9.15]) *test t de Student, après vérif. Bartlett, n>30*
- **Interactions**
  - HTA.TABAC : p=1 (OR= 1.1 [0.26;4.21] *test exact de Fisher*
  - HTA.POIDSMAT : p=0.02 (D=-13,4[-27.1; 0.35]) *test de Wilcoxon*

Régression logistique multiple

- **Analyse multivariée, pas à pas**
  - Choix des variables testées : tabac, hta, poids maternel, âge maternel et interactions

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.11473	0.83245	1.339	0.18054	
tabac1	0.71479	0.32935	2.170	0.02998	*
hta1	1.81197	0.68613	2.641	0.00827	**
poidsmer	-0.04024	0.01442	-2.791	0.00525	**

2.04[1.07;3.9]
6.12[1.6;23.5]
0.96[0.93;0.99]

- **Modèle final**

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.15083	0.87132	1.321	0.18657	
tabac1	0.70123	0.34250	2.047	0.04062	*
hta1	1.73686	0.85786	2.025	0.04291	*
poidsmer	-0.04078	0.01494	-2.729	0.00635	**
tabac1:hta1	0.19475	1.34929	0.144	0.88524	

2.01[1.03;3.95]
5.68[1.06;30.52]
0.96[0.93;0.99]
1.22[0.09;17.1]

Méthodes particulières

Analyse en composante principale

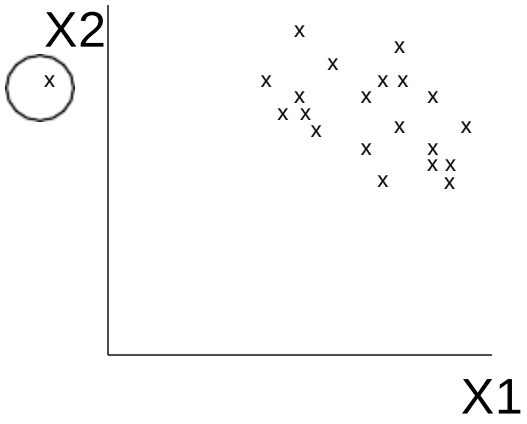
Méthodes particulières

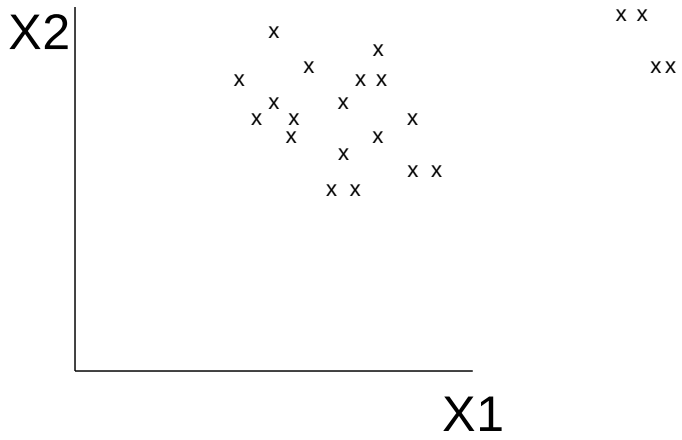
- Données de comptages : régression de Poisson
  - Nombre d'événements dans le temps
- Régression non-linéaire
- Données censurées (survie) :
  - Estimation de Kaplan Meier ou Actuarielle
  - Test du Log Rank (univariée), Modèle de Cox (multivarié)
- Séries temporelles (Box-Jenkins)
- Variabilité spatiale
- Analyse factorielle de Données
  - ACP, ACM, Arbres, CHA, Kmeans…

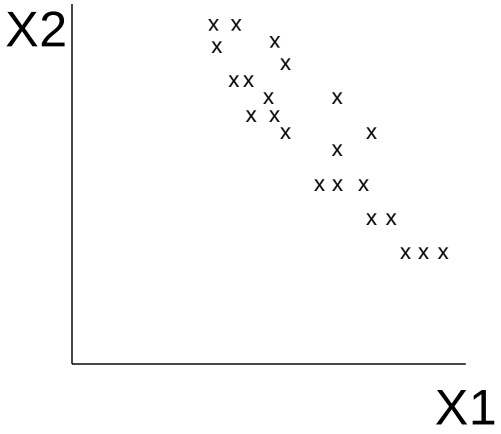
Analyse en composantes principales (ACP)

- Les variables sont toutes quantitatives
- Les moyennes, variances, corrélations ont un sens
- Examiner la structure des données
  - Les individus se ressemblent-ils tous ?
  - Existe-t-il des sous-groupes d'individus ?
  - Individus aberrants ?
- Quelles sont les variables corrélées entre elles ?
  - interpréter facilement la matrice de corrélation
  - (p variables, $p * (p + 1)/2$ corrélations possibles !)

Principes de l'ACP







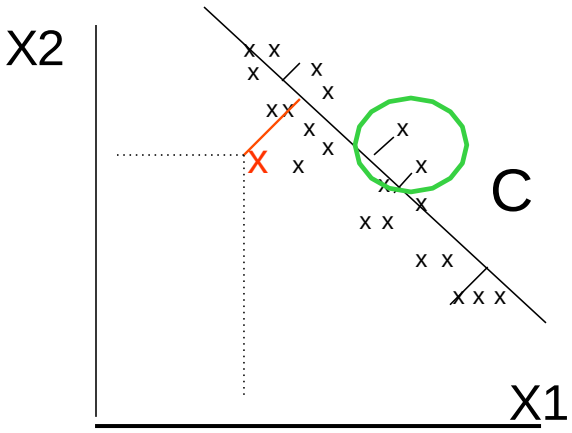
Si les données ne comportaient que 2 variables : une représentation graphique suffirait pour répondre aux objectifs.

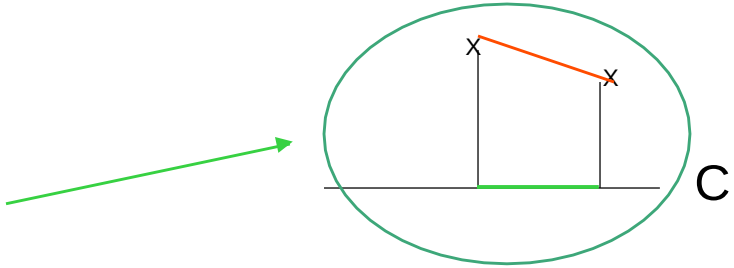
Mais, en général il y a p variables (espace à p dimensions) : représentation impossible

⇒ L'idée est donc d'obtenir des représentations approchées en dimension 2

Principes de l'ACP

- p variables ⇒ dimension p (R^p)
- Obtenir des représentations en dimension 2 les plus fiables possibles
- Critère : **conservation de la variance** = conservation de la distance entre les individus
- Construction de **nouvelles variables Cj** qui maximisent la variance
- Contraintes de simplicité : **combinaisons linéaires des variables initiales**
- $C1 = A^1_1X_1 + A^1_2X_2 + \ldots + A^1_pX_p$
- Géométriquement





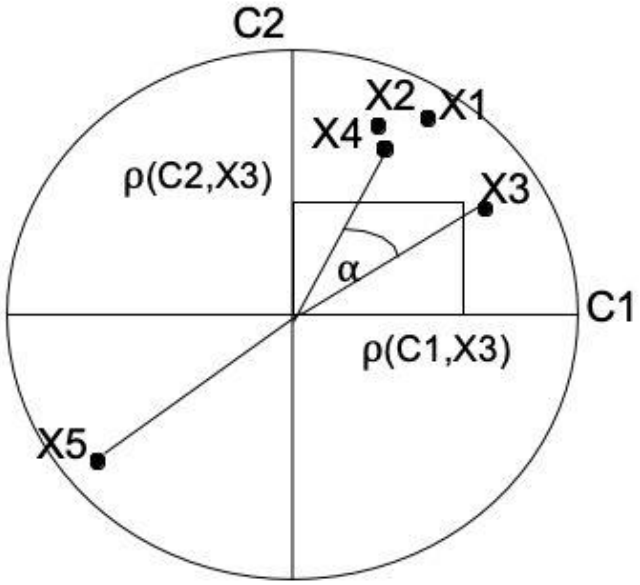
Si on considère la nouvelles variable C, l'information est reconstituée de la manière la plus fiable possible au sens de la variance

Principes de l'ACP

- Première composante principale C_1 = combinaison linéaire des variables initiales qui maximise la variance
- Deuxième composante principale : maximise la variance et est non-corrélée à la première composante (orthogonalité)
- Et ainsi de suite . . .
- Au plus p composantes principales
- En réalité, si liaisons entre les variables, l'essentiel de l'information (la variance) est contenu dans les (2 ou 3) premières composantes principales

Principes de l'ACP

- Analyse des liaisons entre variables
  - Matrice de corrélation
  - p variables → p(p + 1)/2 corrélations
  - Liaison 2 à 2, pas de liaisons multivariées
- ACP : représentation des variables : cercle des corrélations (C_1 et C_2 sont les deux premières composantes principales)
- On peut alors montrer que si des variables sont proches de la circonférence alors le cosinus de l'angle α est proche du coefficient ρ de corrélation entre ces 2 variables.



Exemple d'ACP

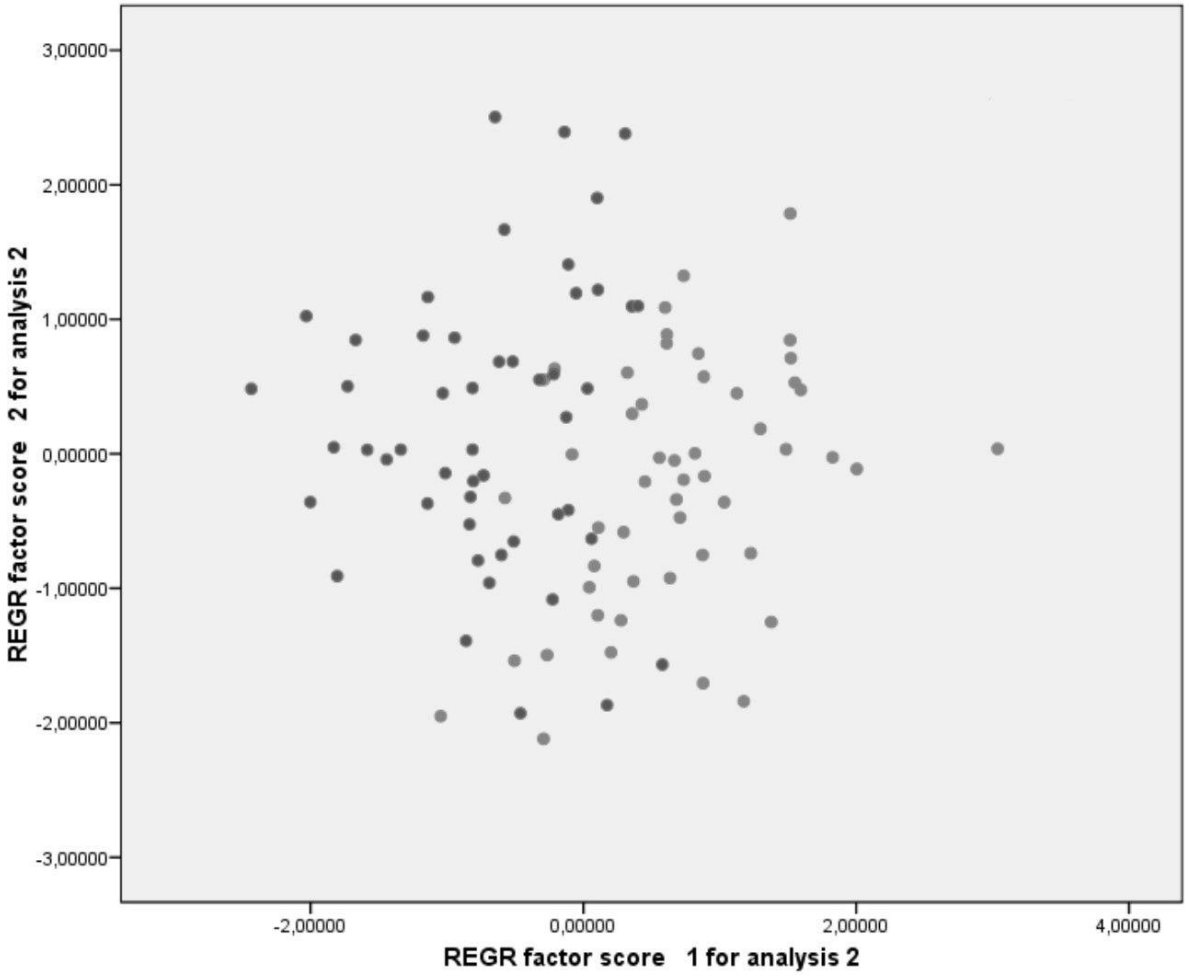
Infarctus du myocarde

- Variables numériques :
  - Fréquence cardiaque
  - Index cardiaque
  - Index systolique
  - Pression diastolique
  - Pression artérielle pulmonaire
  - Pression ventriculaire
  - Résistance pulmonaire
- Variable qualitative
  - décès

- Objectifs
  - Vérifier la cohérence des données
  - Recherche d'individus exceptionnels (en multivarié)
  - Existence de profils d'individus différents (sur p variables = multivarié)
  - Utilisation de la variable décès comme variable « illustrative »

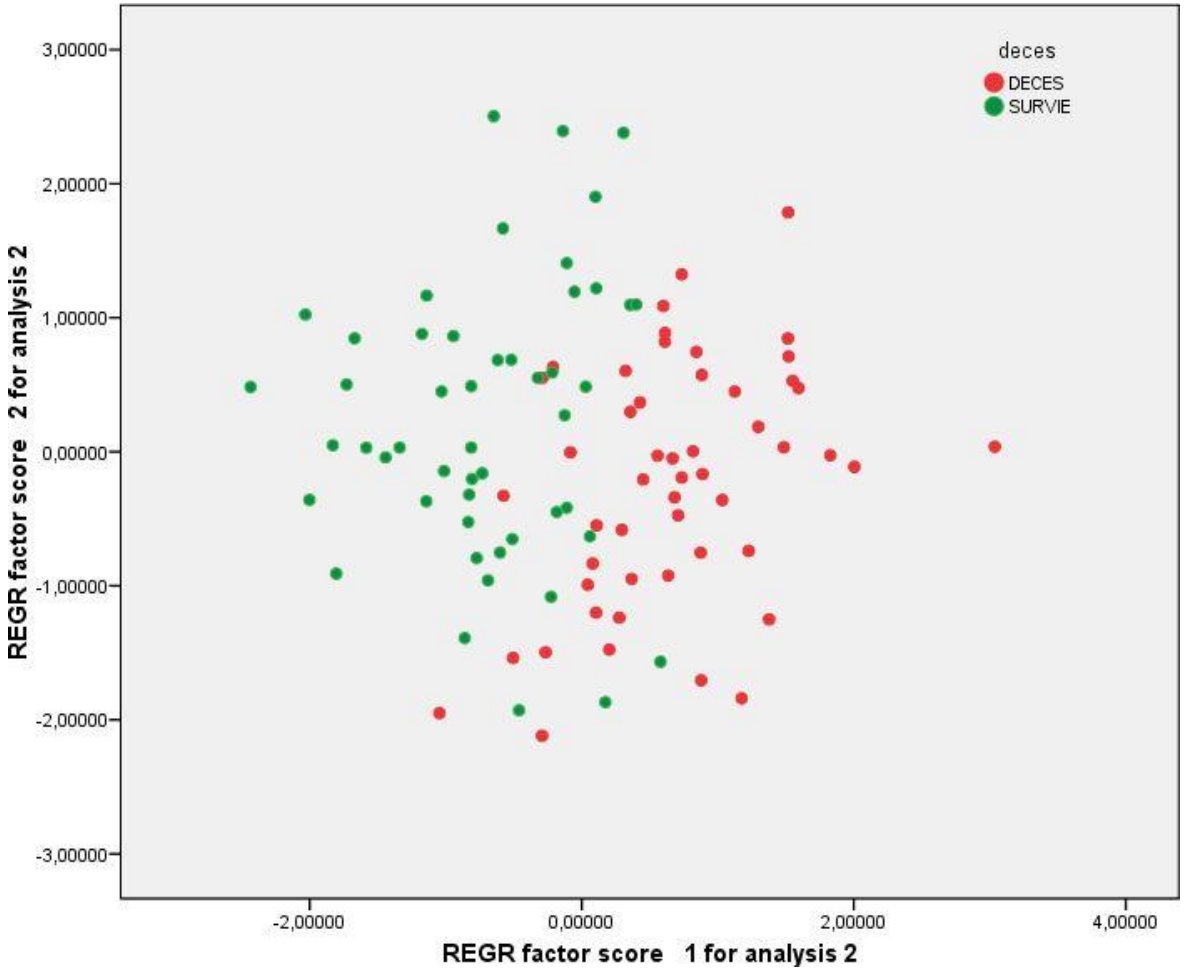
Exemple d'ACP

Nuage des individus



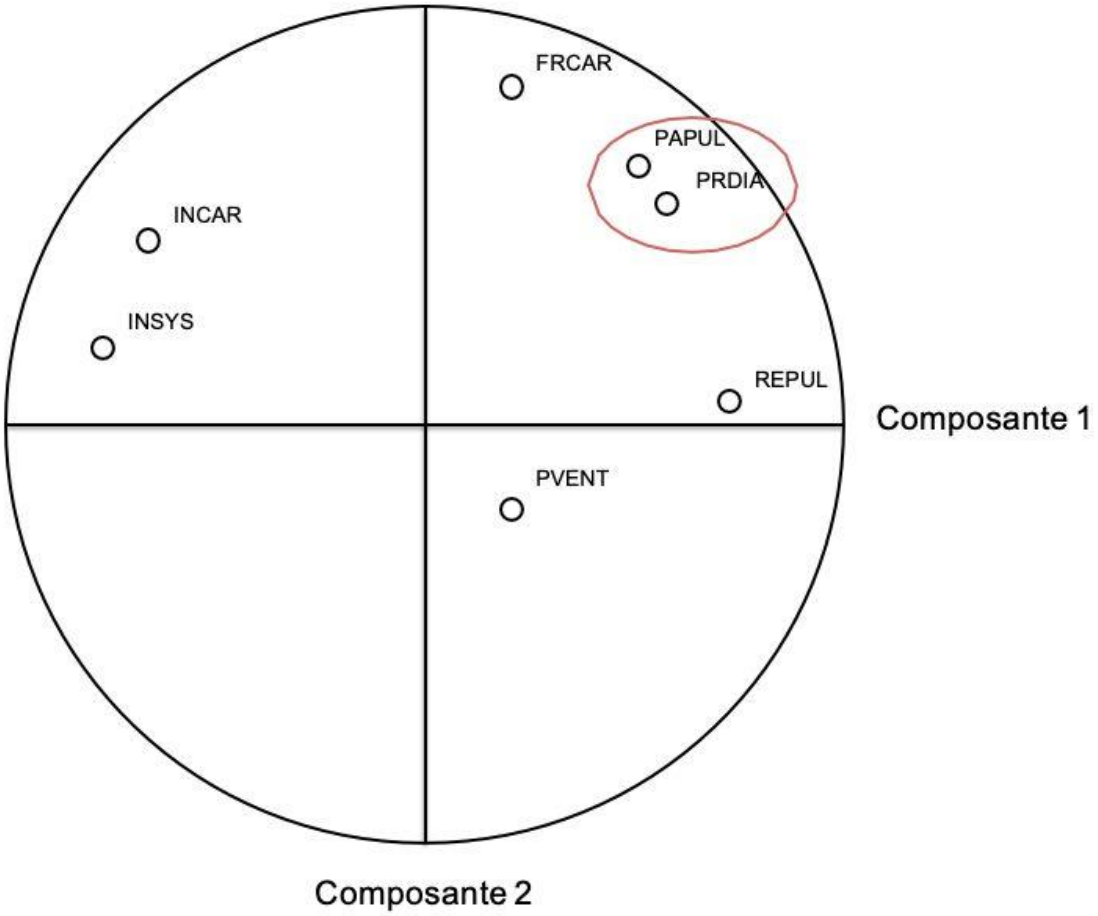
Exemple d'ACP

Nuage des individus : ajout d'une variable illustrative (vers l'inférentiel)



Exemple d'ACP

Cercle des corrélations entre variables



Stratégie d'analyse

Stratégie d'analyse

- Statistiques descriptives
  - Moyennes, pourcentages, intervalles de confiances, médianes
  - Graphiques (boxplot, histogrammes)
- Analyses univariées
  - Descriptives
    - statistiques et graphiques par groupes, survie (Kaplan- Meier)
  - Tests statistiques (+- séries appariées)
    - Pourcentages : Chi2, Fisher exact test
    - Moyennes : Student, ANOVA, Wilcoxon, Kruskal-Wallis
    - Corrélation de Pearson ou de Spearman
    - Log Rank (survie)
    - Interactions en fonction de la biologie
    - Séries chronologiques, corrélations spatiales…

Stratégie d'analyse

- Analyse multivariée
  - Choix de la méthode (R linéaire, R logistique, Modèle de Cox…)
  - Choix des variables initiales
    - Variables connues dans la littérature
    - Variables avec un sens biologique
    - Variables p<0,2 ou p<0,25 pour les tests univariés
  - Méthode pas à pas, avec les interactions, choix du critère statistique
  - Garder
    - Les variables sélectionnées par la méthode pas à pas
    - Les variables biologiquement pertinentes
  - Vérification de la qualité du modèle
  - Interprétation du modèle final

Références

- Jean Bouyer. Méthodes statistiques, Médecine-Biologie, ed INSERM
- Jean Bouyer. Epidémiologie quantitative, ed INSERM
- CIMES. Biostatistiques, ed Omnisciences
- Jean Gaudart. Introduction générale aux tests statistiques. SESSTIM, Marseille
- Michaël Genin, Alain Duhamel, Patrick Devos. Les statistiques dans la recherche médicale. Méthodes statistiques multivariées. Université de Lille 2
- Antoine Gournay. Analyse statistique multivariée. Université Neuchâtel, Suisse.

Mentions légales

- L'ensemble de ce document relève des législations française et internationale sur le droit d'auteur et la propriété intellectuelle.
- Tous les droits de reproduction de tout ou partie sont réservés pour les textes ainsi que pour l'ensemble des documents iconographiques, photographiques, vidéos et sonores.
- Ce document est interdit à la vente ou à la location par un tiers autre que l'Université Cote d'Azur.
- La diffusion, la duplication, la mise à disposition du public (sous quelque forme ou support que ce soit), la mise en réseau, de tout ou partie de ce document, sont strictement réservées à l'Université Cote d'Azur.
- L'utilisation de ce document est strictement réservée à l'usage privé des étudiants inscrits aux cours et au tutorat organisés par l'UFR de Médecine de l'Université Cote d'Azur, et non destinée à toute autre utilisation privée ou collective, gratuite ou payante.