Statistiques appliquées à la biologie : regréssion.

Alexandre Popier

L3 BBTE, Université du Maine, Le Mans

Semestre 1

Plan du cours

EXEMPLE 1 : PRESSION SYSTOLIQUE ET DIASTOLIQUE

14 DONNÉES :

Systolique	Diastolique
138	82
130	91
135	100
140	100
120	80
125	90
120	80
130	80

QUESTION : y a-t-il une relation entre les deux pressions cardiaques ? Peut-on la modéliser ?

Exemple 2 : lien poids-taille.

Échantillon : médaillés (or) masculins à Sydney :

Noms	Taille	Poids
Andrieux	192	97
Asloum	165	63
Bette	186	70
Douillet	196	125
Dumoulin	171	64
Estanguet	182	75
Ferrari	187	83
Gané	176	79
Martinez	164	50
Rousseau	182	85

Question : y a-t-il une relation entre le poids et la taille ?

Plan

Corrélation linéaire.

On a deux jeux de données, un jeu noté X, l'autre Y, avec n données dans chaque

$$X = (X_1, \ldots, X_n), \quad Y = (Y_1, \ldots, Y_n).$$

Définition

Le coefficient de corrélation linéaire définit l'intensité et la direction d'une relation linéaire entre X et Y. On le note r.

- *Il est toujours compris entre -1 et 1.*
- *Son signe indique le sens de la relation ; sa valeur absolue exprime l'intensité de la relation.*
- *Proche de 1 (resp. -1) les deux variables sont comonotones (resp. contramonotones).*
- *Proche de 0, il exprime l'absence de relation linéaire.*

Coefficient de détermination.

Définition

Le coefficient de détermination est le carré r^2 du coefficient de corrélation linéaire. Ce coefficient évolue de 0 (pas de relation linéaire) à 1 (relation parfaitement linéaire, positive ou négative).

Problème :

- Que signifie proche de zéro ?
- Comment déterminer s'il y a une relation significative entre les variables ?

Test d'hypothèses :

- Hypothèse nulle (H_0) : $r^2 = 0$;
- Hypothèse alternative (H_1) : $r^2 \neq 0$.

Sous R, utiliser `cor.test`

Coefficient de détermination.

Sur les deux exemples :

1. Pressions cardiaques : $n = 14$, $r = 0,66$, $r^2 = 0,43$.

   Résultat du test sous R :

   - `true correlation is not equal to 0`
   - intervalle de confiance à 95 % sur r : $[0,19; 0,88]$.

2. Taille-poids : $n = 10$, $r = 0,86$, $r^2 = 0,74$.

   Résultat du test sous R :

   - `true correlation is not equal to 0`
   - intervalle de confiance à 95 % sur r : $[0,51; 0,97]$.

VITESSE AU SPRINT ET LACTATES : CAS LIMITE.

On teste 36 sprinteurs sur 100m en mesurant leur vitesse et leur taux de lactates dans le sang trois minutes après (mesure la contribution de la glycolyse anaérobie).

Vitesse	Lactate
9.33	12.1
9.05	14.5
8.79	15.2
9.17	14
⋮	⋮

RÉSULTATS : $r = 0,36$, $r^2 = 0,13$. Résultat du test sous R :

- `true correlation is not equal to 0`
- intervalle de confiance à 95 % sur r : $[0,04; 0,62]$.

Plan

Graphique de liaison.

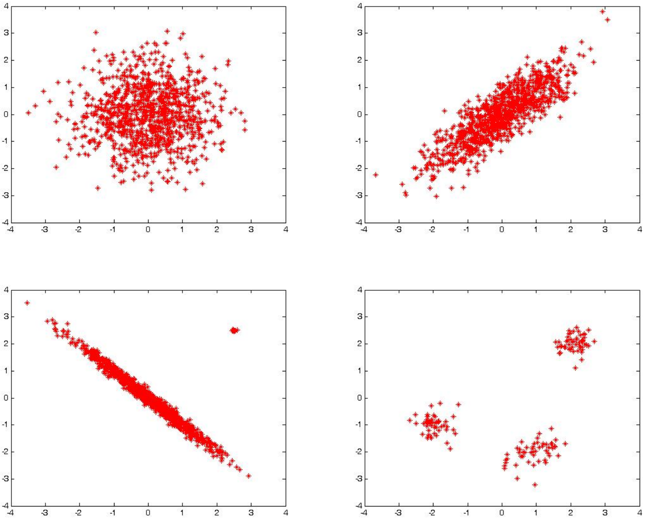
Définition

Le nuage de points décrit la relation entre les jeux de données X et Y.

- *Chaque couple de données (X_i, Y_i) apparaît comme un point dont la coordonnée horizontale est sa valeur X_i, et la coordonnée verticale sa valeur Y_i.*
- *On ajoute parfois au nuage de points une droite verticale indiquant la moyenne des données X et une droite horizontale repérant celle des Y.*

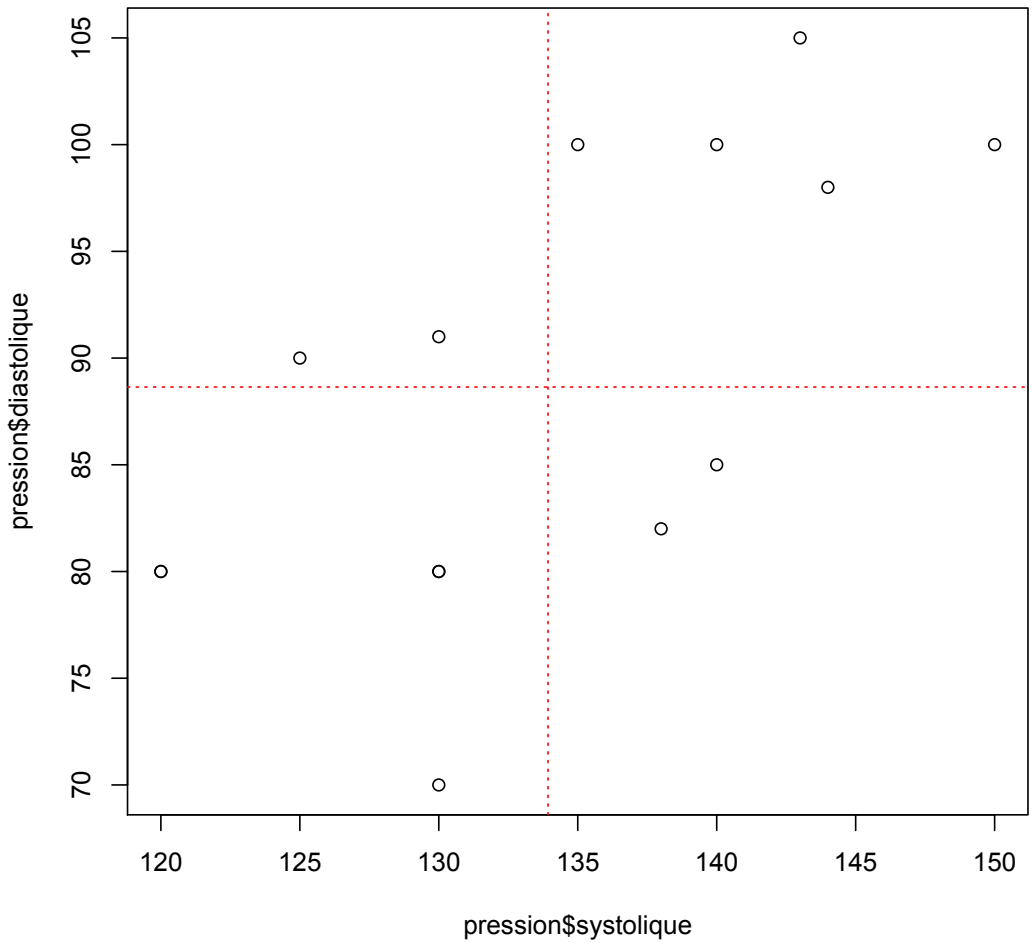
Graphique de liaison.

Différentes types de nuages :



EXEMPLES (RETOUR)

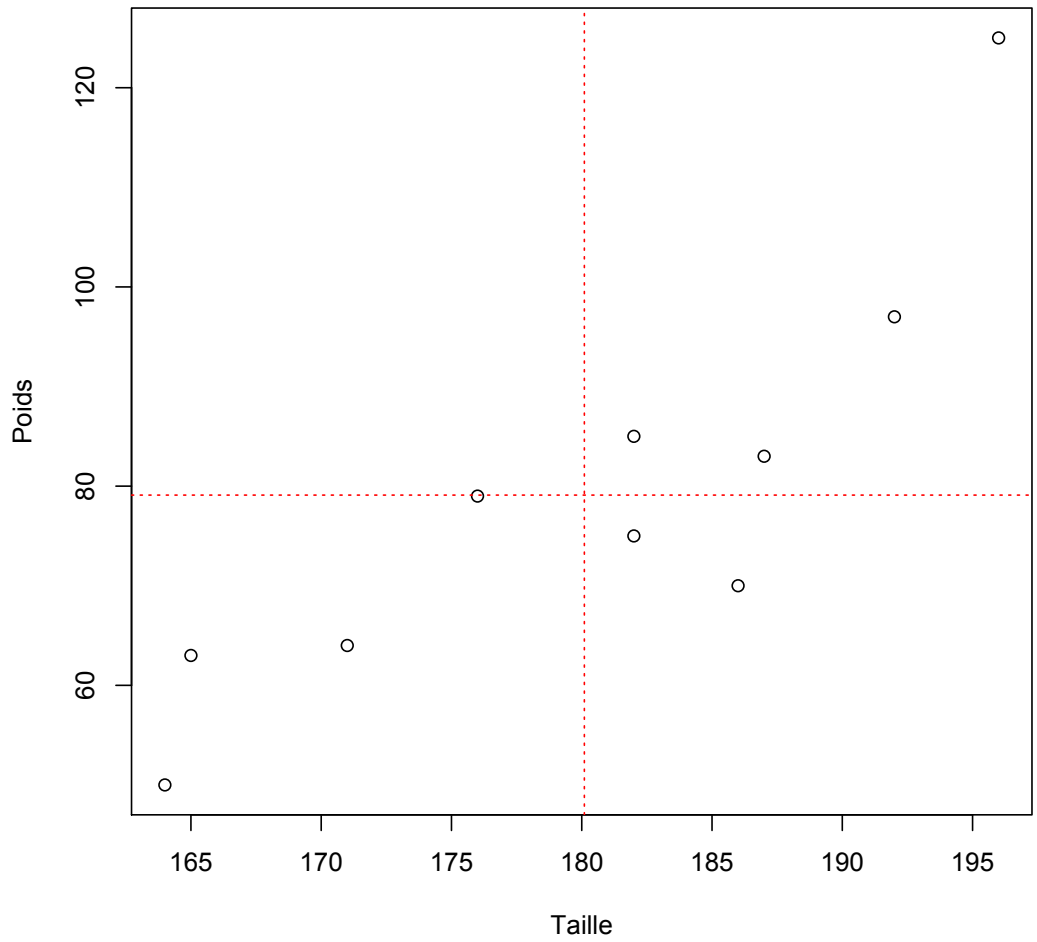
PRESSIONS CARDIAQUES : Moyenne systolique = 133,9, moyenne diastolique = 88,6.



EXEMPLES (RETOUR)

MÉDAILLÉS OLYMPIQUES :

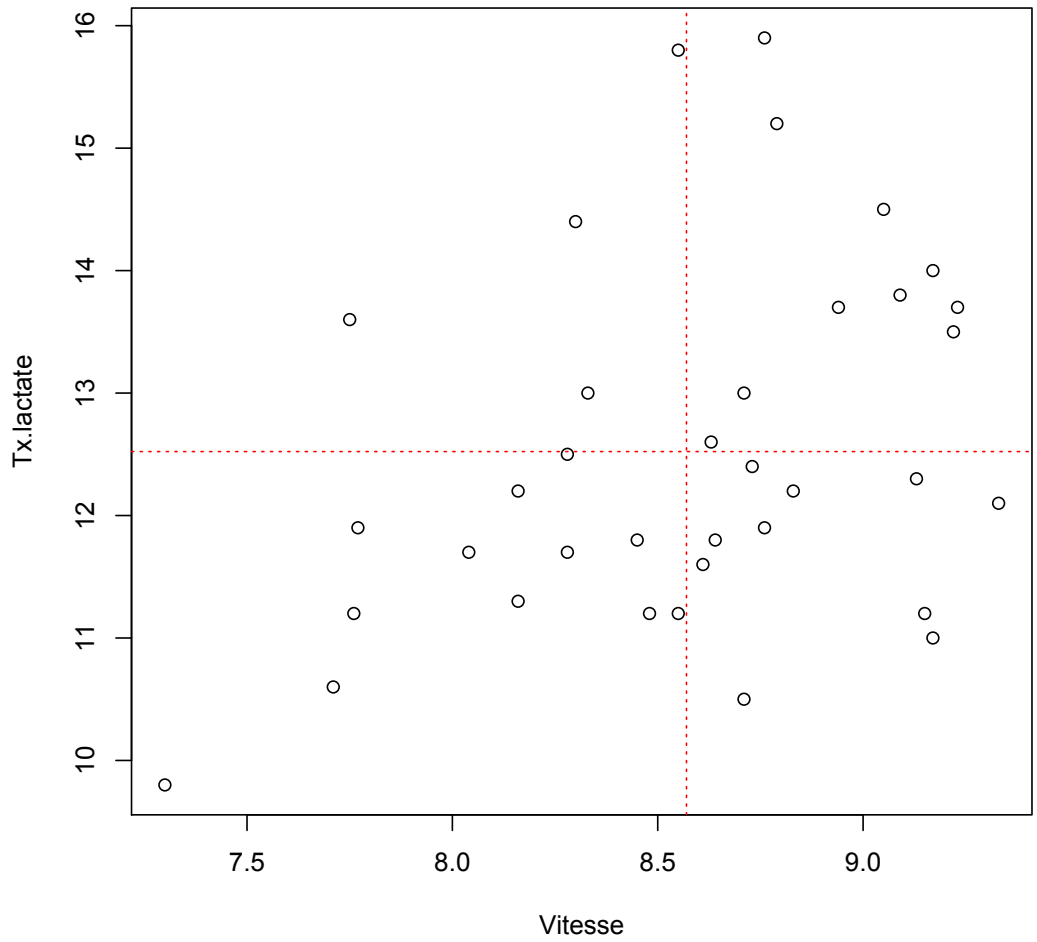
Moyenne des poids = 79,1, moyenne des tailles : 180,1.



EXEMPLES (RETOUR)

TAUX DE LACTATE :

Moyenne vitesse = 8,57, moyenne taux lactate : 12,52.

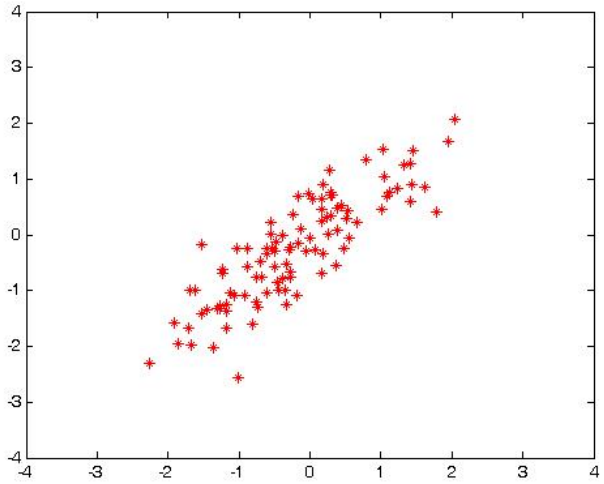


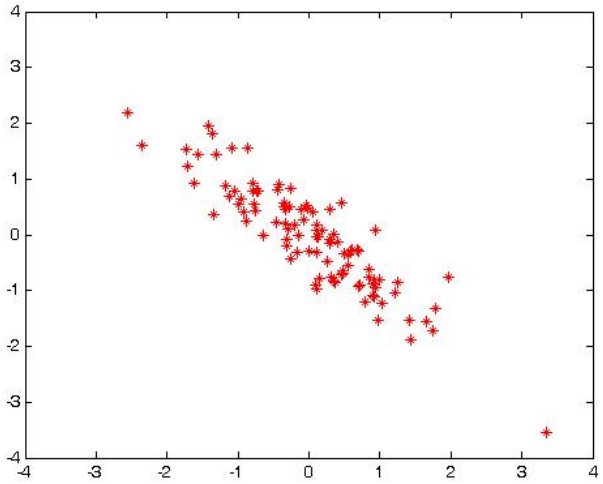
Relations linéaires.

Définition

Une relation est linéaire lorsque le nuage de points paraît étiré le long d'une droite.
On parle de

- *relation positive lorsque les coordonnées verticales Y tendent à augmenter en même temps que les coordonnées horizontales X ;*
- *relation négative si les coordonnées verticales Y tendent à diminuer quand les coordonnées horizontales X augmentent.*



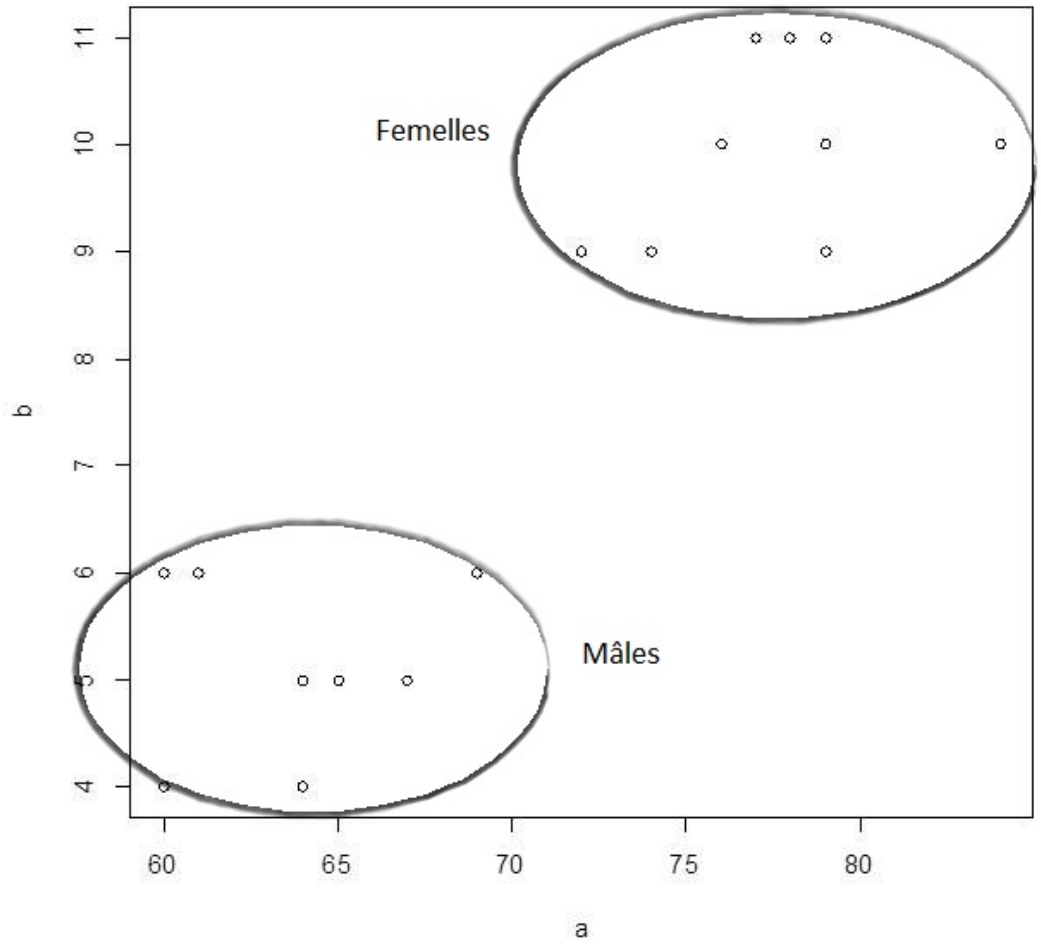


Quelques erreurs à ne pas commettre.

- Une corrélation forte n'implique pas forcément une relation de cause à effet.
- Un coefficient de corrélation faible n'implique pas qu'il n'y ait aucune relation entre les deux variables ; il peut y avoir une relation non linéaire.
- L'association de groupes hétérogènes peut créer une corrélation artificielle.

QUELQUES ERREURS À NE PAS COMMETTRE.

Exemple chez des oiseaux : relation entre la masse des individus adultes et le nombre de passages au nid pendant sa construction.



Plan

Régression.

But : essayer de prédire l'une des variables par l'autre $\rightarrow$ régression.

Définition

Une droite de régression modélise la réponse moyenne de y en tout point d'abscisse x. Son équation est de la forme $y = ax + b$, avec :

- *b : l'ordonnée à l'origine,*
- *a : la pente de la droite de régression.*

Définition

Pour les données mesurées, on peut calculer l'écart entre la prédiction $\hat{y}$ et la valeur observée Y. On parle de résidu ou d'erreur :

$$e = Y - \hat{y} = Y - aX - b.$$

Exemple : modèle de Lorentz.

Modèle de Lorentz (1929) :

$$Poids = 0,75 \times Taille - 62,5.$$

- Prédiction pour $Taille = 180$: $\widehat{Poids} = 72,5$.
- Pour D. Douillet : $Taille = 196$, $Poids = 125$:

$$\widehat{Poids} = 84,5, \qquad e = 125 - 84,5 = 40,5.$$

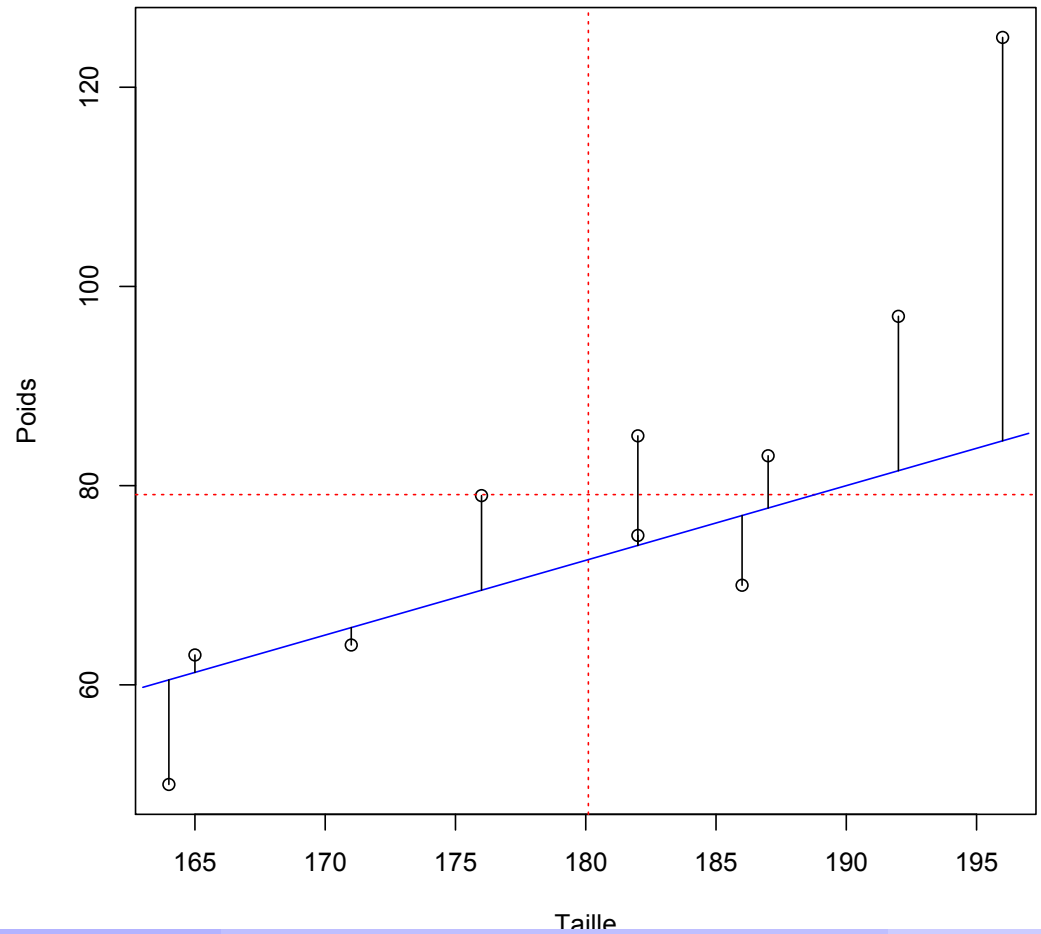
- Pour G. Drut (Or J.O. 1976) : $Taille = 189$, $Poids = 80$:

$$\widehat{Poids} = 79,25, e = 0,75.$$

Exemple : modèle de Lorentz.

Modèle de Lorentz (1929) :

$$Poids = 0,75 \times Taille - 62,5.$$



Droite des moindres carrés.

Définition

La droite de régression des moindres carrés minimise les carrés des résidus en moyenne, c'est-à-dire la quantité :

$$SS_{res} = \sum_{i=1}^{n} (Y_i - (aX_i + b))^2.$$

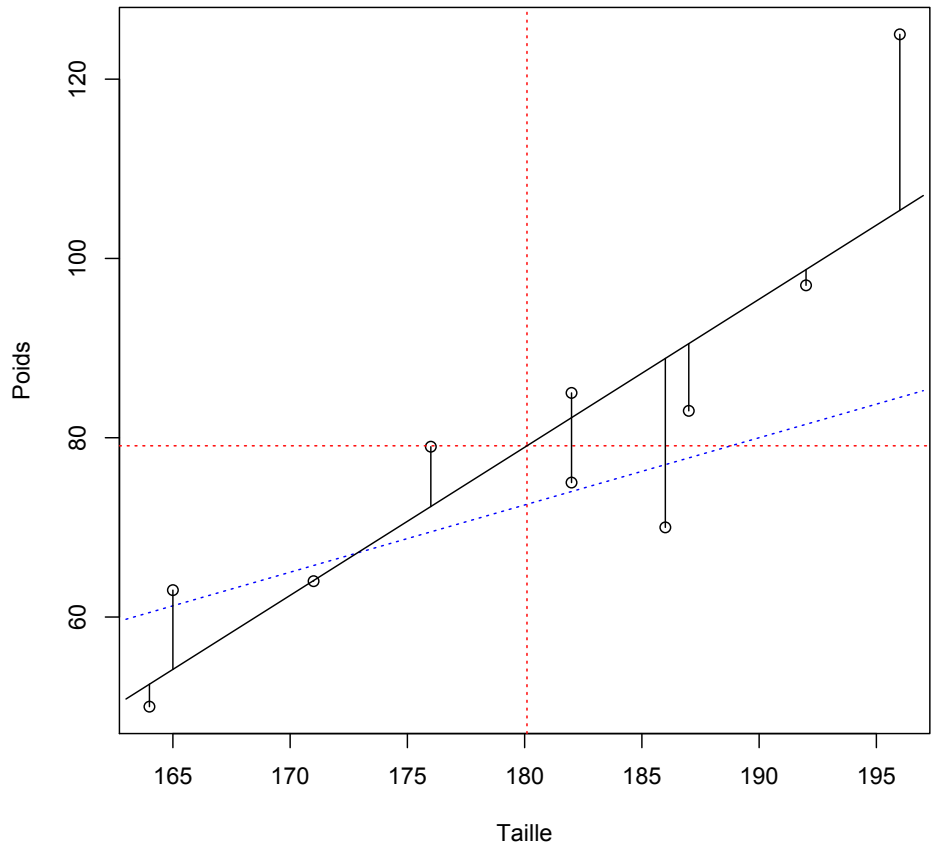
Calcul des coefficients

- Pente : $a = \dfrac{s_Y}{s_X} r = \dfrac{\text{Cov}\,(X, Y)}{\text{Var}\,(X)}$.
- Ordonnée à l'origine : $b = m_Y - a m_X$.

DROITE DES MOINDRES CARRÉS.

MÉDAILLÉS OLYMPIQUES : $m_X = 180,1$, $s_x = 10,9$, $m_Y = 79,1$, $s_y = 20,8$, $r = 0,86$.

$$a = 1,65, \quad b = -218,39 \Rightarrow \textit{Poids} = 1,65 * \textit{Taille} - 218,39$$



COMPARAISON DES MODÈLES :

- Modèle de Lorentz (1929) :

$$Poids = 0,75 \times Taille - 62,5.$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - (aX_i + b))^2 = 2285,69.$$

- Droite des moindres carrés :

$$Poids = 1,65 \times Taille - 218,39.$$

$$SS_{res} = 989,05.$$

Nouvelle interprétation de r^2.

Noms	*Poids*	$\widehat{Poids}$	$Poids - \widehat{Poids}$
Andrieux	97	98.8	-1.8
Asloum	63	54.1	8.9
Bette	70	88.8	-18.8
Douillet	125	105.4	19.6
Dumoulin	64	64	0
Estanguet	75	82.2	-7.2
Ferrari	83	90.5	-7.5
Gané	79	72.3	6.7
Martinez	50	52.5	-2.5
Rousseau	85	82.2	2.8
Moyennes	79.1	79.1	0
Variances	434.5	324.6	109.9

$$r^2 = 0.75 = \frac{324.6}{434.5}$$

Nouvelle interprétation de r^2.

De façon générale :

- la moyenne des prédictions est égale à la moyenne des données. En moyenne il n'y a pas d'erreur !

Définition

Variance expliquée par la régression = variance des valeurs prédites .

- Analyse de la régression : variance totale = variance expliquée + variance des erreurs.

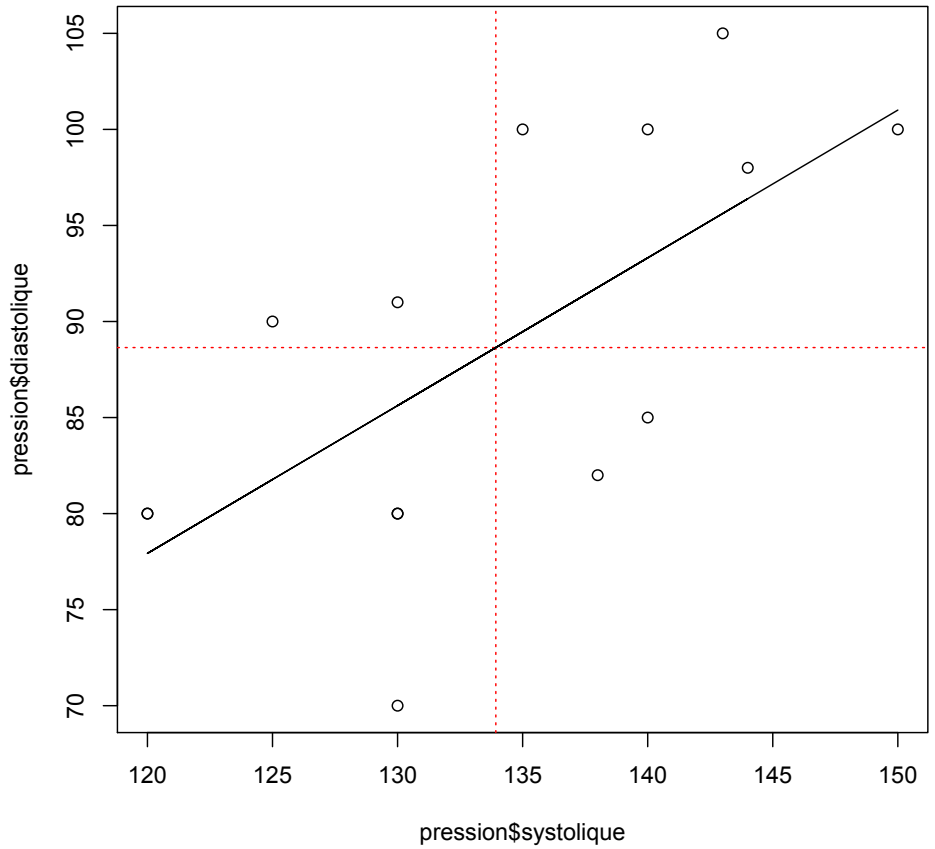
Propriété

$$r^2 = \frac{\text{Variance expliquée}}{\text{Variance totale}}.$$

Le coefficient de détermination est la proportion de variance de la réponse Y pouvant être expliquée par la régression sur X.

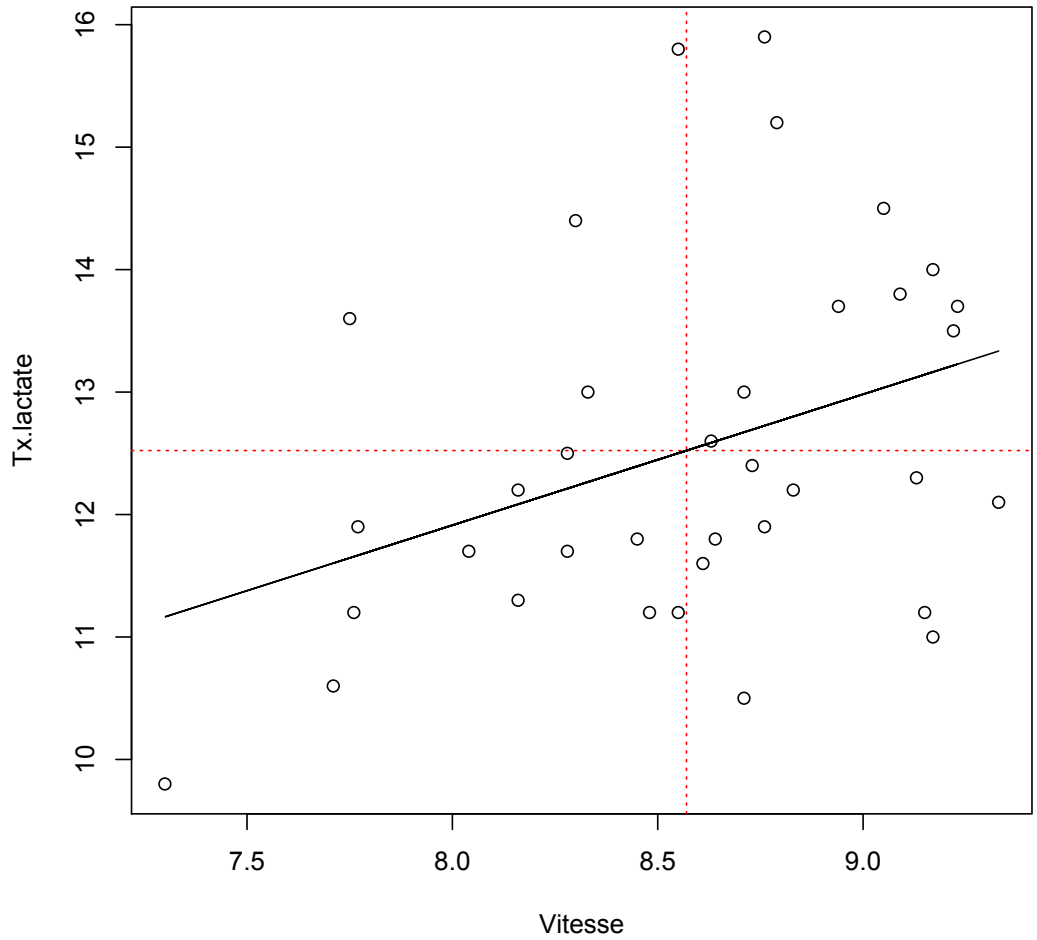
PRESSIONS CARDIAQUES.

- Moyenne systolique = 133,9, moyenne diastolique = 88,6.
- $r^2 = 0,43$. $a = 0,77$, $b = -14,38$.



VITESSE AU SPRINT ET LACTATES.

$r^2 = 0,13$, $a = 1,07$, $b = 3,36$.



Plan

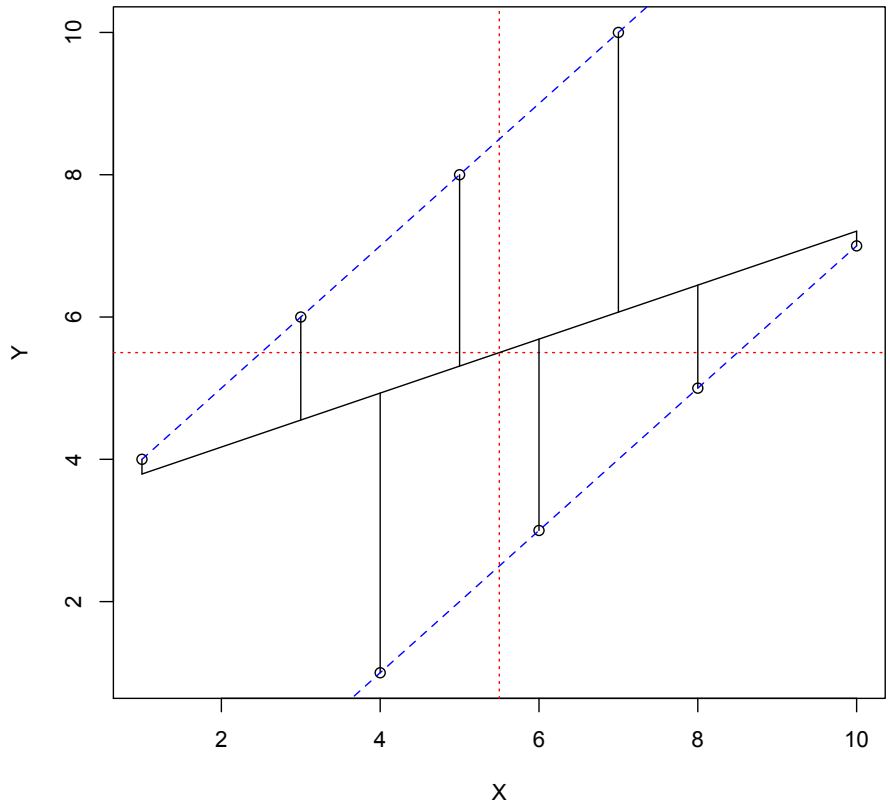
Plan

RUPTURE DE LA SYMÉTRIE.

À NOTER

La régression de Y par X n'est pas la même que de X par Y !

EXEMPLE JOUET :

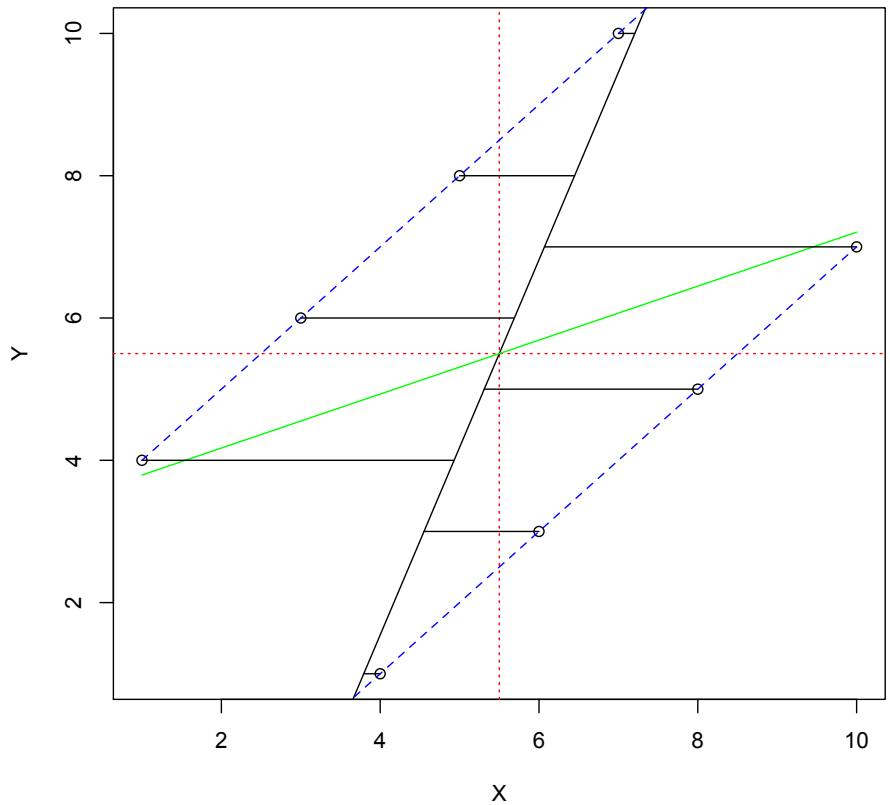


RUPTURE DE LA SYMÉTRIE.

À NOTER

La régression de Y par X n'est pas la même que de X par Y !

EXEMPLE JOUET :

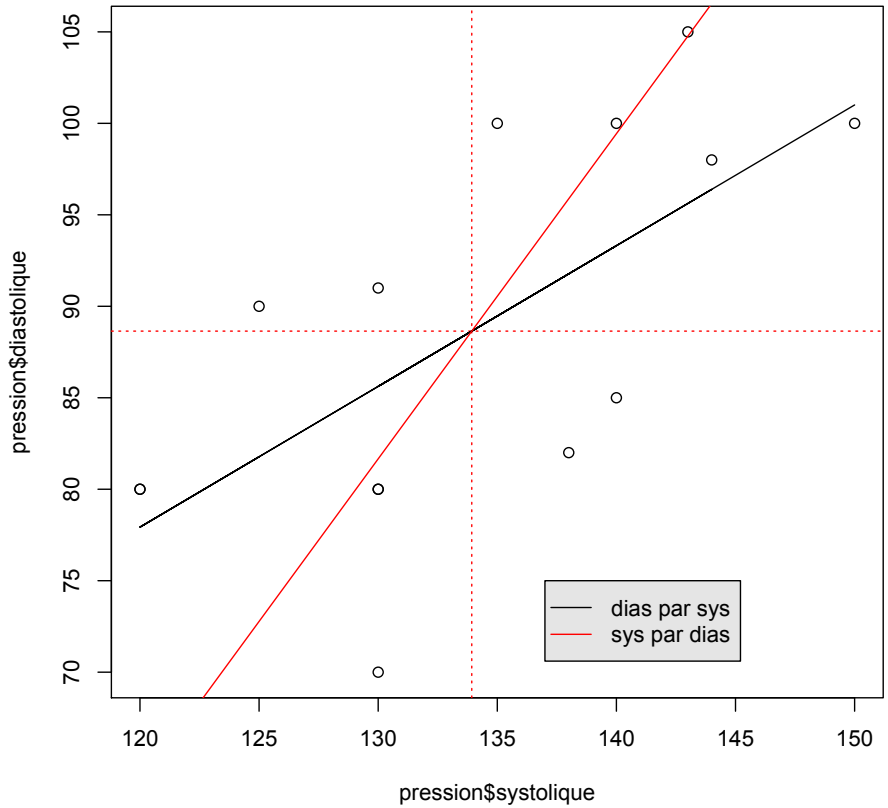


RUPTURE DE LA SYMÉTRIE.

À NOTER

La régression de Y par X n'est pas la même que de X par Y !

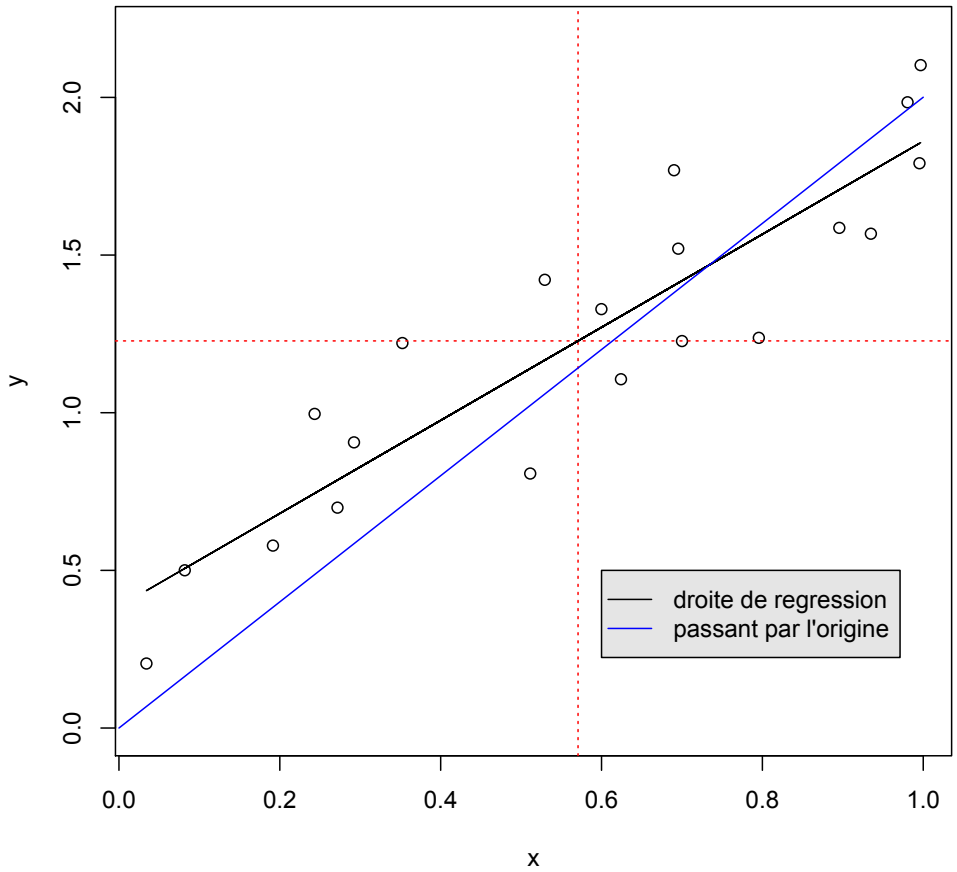
PRESSIONS CARDIAQUES.



PASSAGE PAR L'ORIGINE.

CONDITION TECHNIQUE : (0,0) seul point « certain ». Exemple : mesure de spectrophotométrie.

RÉGRESSION : droite d'équation : $y = ax$ (un paramètre).



Passage par l'origine.

Comparaison :

- Sans contrainte : $r^2 = 0,81$, $SS_{res} = 0,89$.
- Avec contrainte : $r^2 = 0,96$, $SS_{res} = 1,55$.

Exemple de biochimie

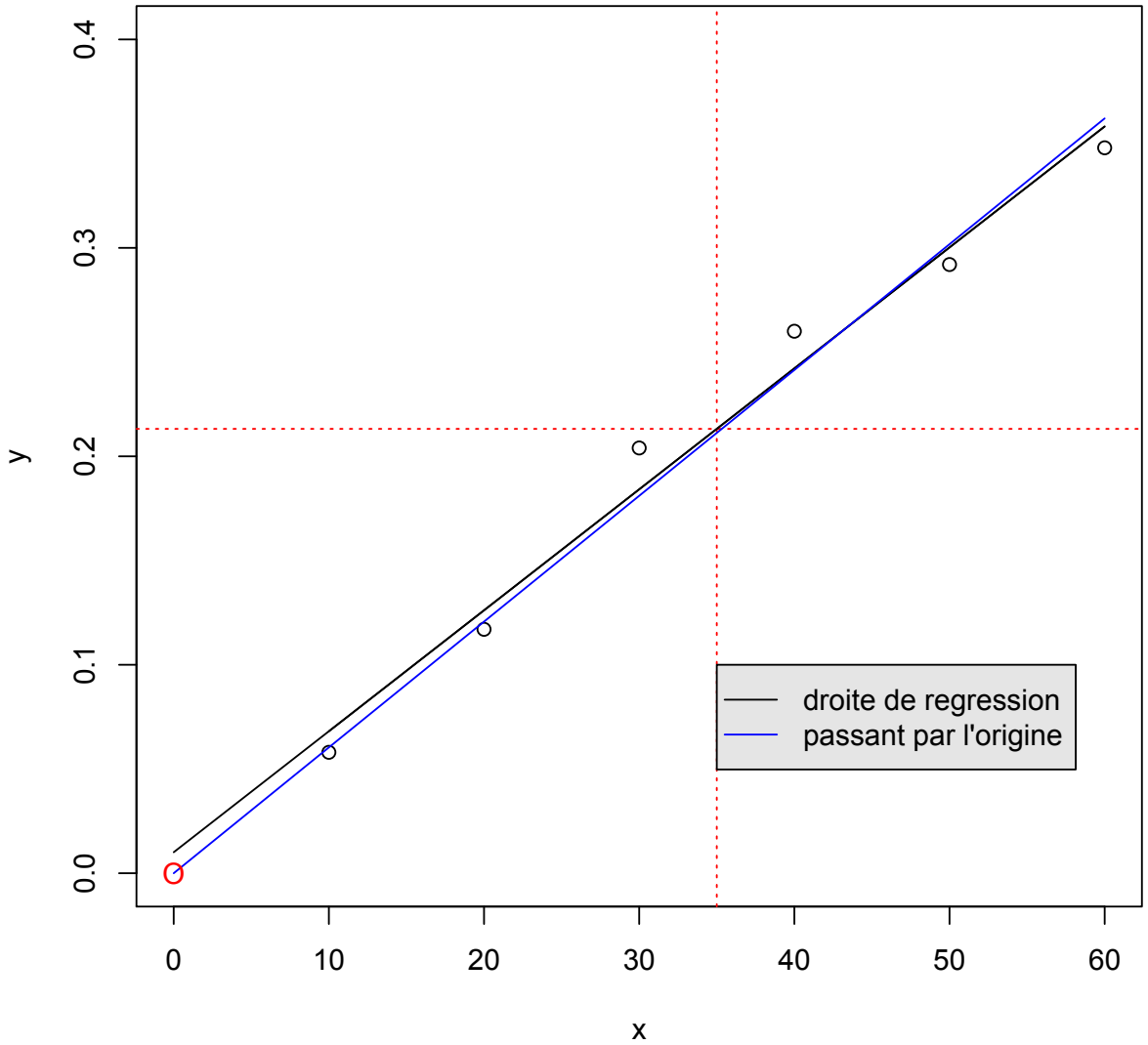
6 données :

Masse protéine	DO
10	$0,058$
20	$0,117$
30	$0,204$
40	$0,26$
50	$0,292$
60	$0,348$

Comparaison :

- Sans contrainte : $r^2 = 0,98$, $SS_{res} = 1,07 \times 10^{-3}$.
- Avec contrainte de passer par l'origine : $r^2 = 0,99$, $SS_{res} = 1,18 \times 10^{-3}$.

EXEMPLE DE BIOCHIMIE



Plan

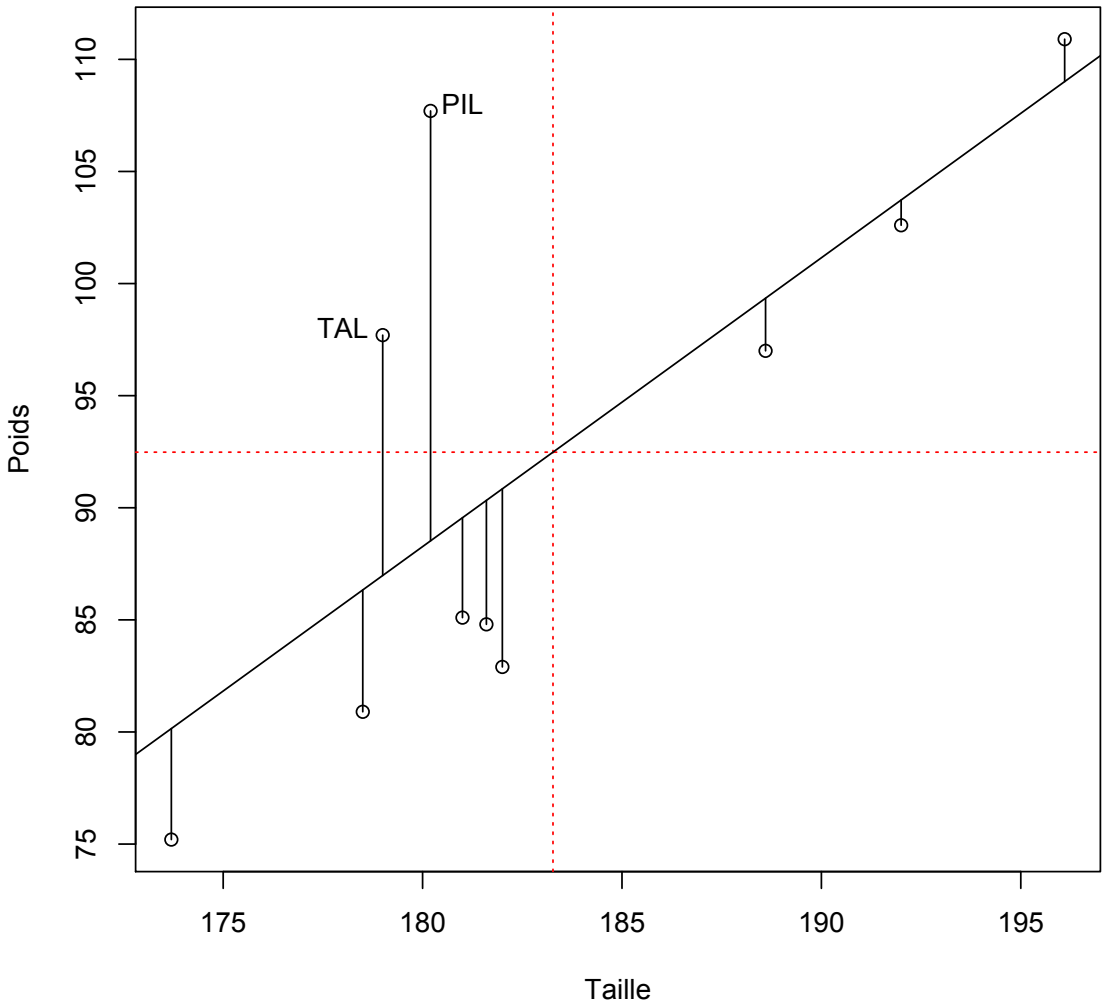
TAILLE ET POIDS MOYEN POUR LES RUGBYMEN.

DONNÉES provenant du championnat français (Top 14).

Poste	Taille	Poids
PIL	180.2	107.7
TAL	179.0	97.7
DL	196.1	110.9
TLA	188.6	97.0
TLC	192.0	102.6
DM	173.7	75.2
DO	178.5	80.9
TQC	181.0	85.1
TQA	181.6	84.8
AR	182.0	82.9

TAILLE ET POIDS MOYEN POUR LES RUGBYMEN.

$r^2 = 0,51$ (taux significatif 0,4). DEUX POINTS EXTRÊMES ?



Point extrême ?

Point extrême dans une régression :

- extrême sur Y : ordonnée très diffrente des autres points d'abscisse proche. Point non consistant.
- extrême sur X : abscisse nettement plus petite ou plus grande que celle des autres points. Phénomène de levier.

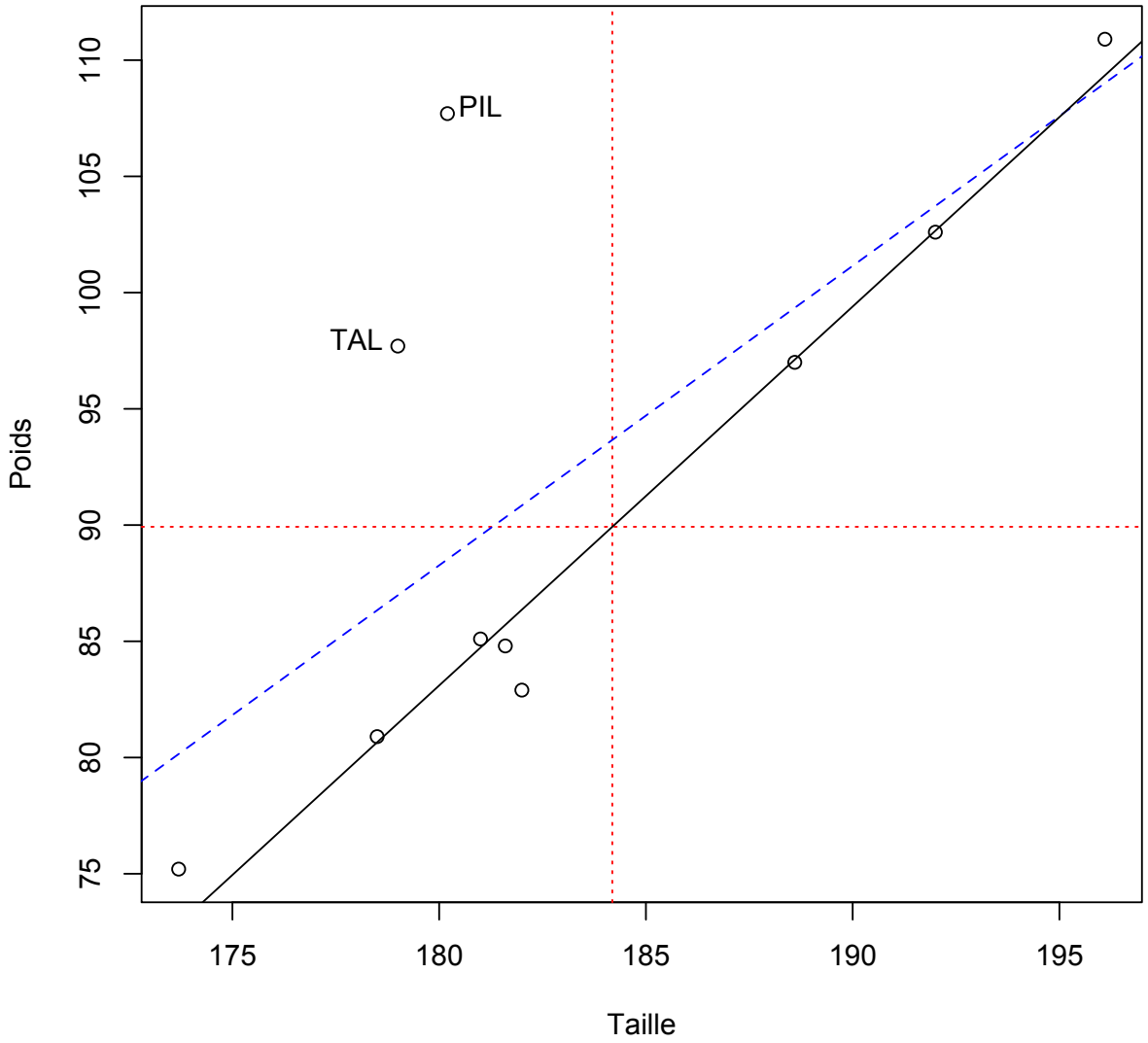
Définition

Un point est influent lorsque la régression pratiquée avec ou sans ce point conduit à des résultats très différents.

Comparaison :

- Tous postes : $r^2 = 0,51$, $SS_{res} = 660,6$.
- Sans PIL et TAL : $r^2 = 0,98$, $SS_{res} = 21$.

SUPPRESSION DES PILIERS ET TALONNEURS.

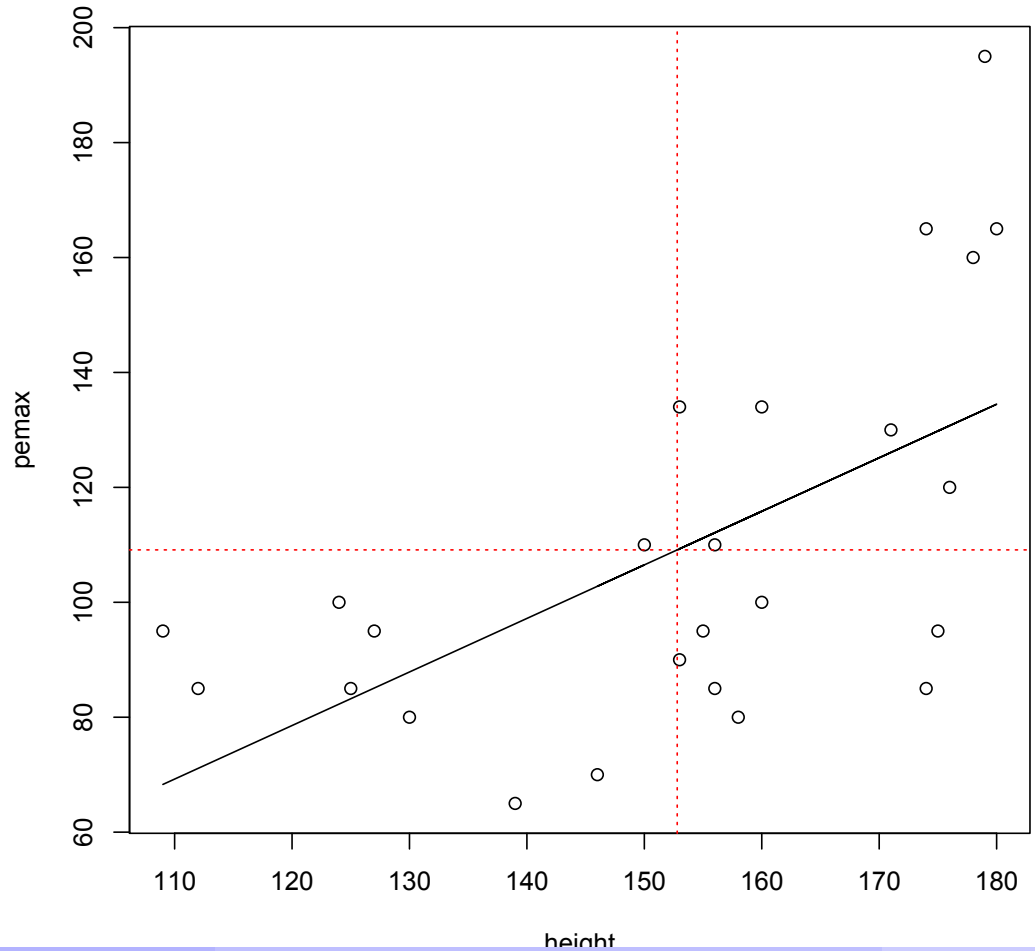


Plan

FIBROSE KYSTIQUE (OU MUCOVISIDOSE).

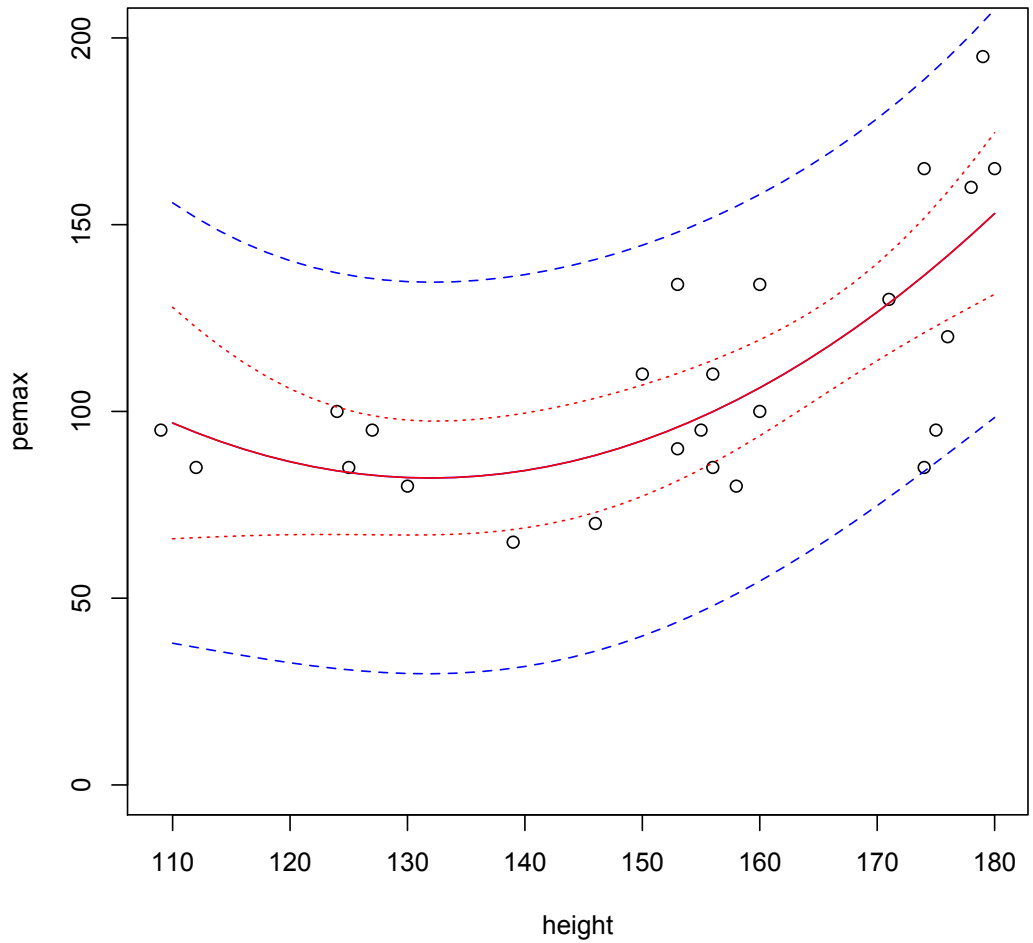
LIEN PRESSION D'EXPIRATION MAXIMALE ET TAILLE.

$$pemax = -33,28 + 0,93 \times hauteur.$$



RÉGRESSION NON LINÉAIRE.

$$pemax = 615,36 - 8,08 \times hauteur + 0,06 \times hauteur^2.$$



COMPARAISON.

RÉSULTATS :

- 25 données, $r^2 = 0,36$.
- Régression linéaire : $SS_{res} = 17198$.
- Régression non linéaire : $SS_{res} = 12866,4$.