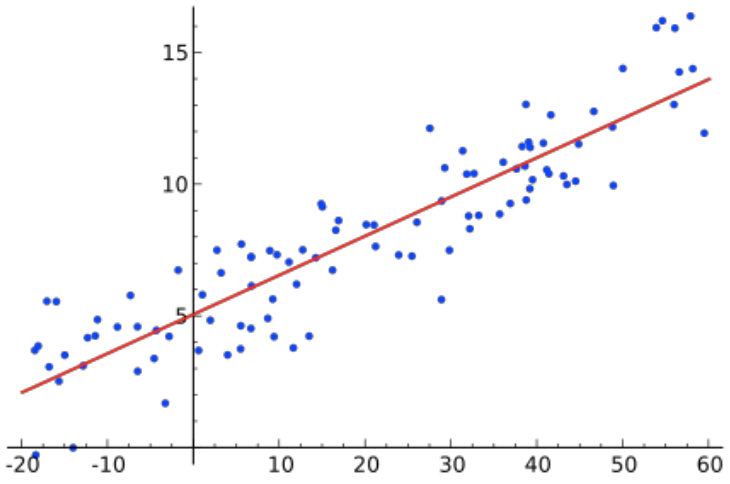
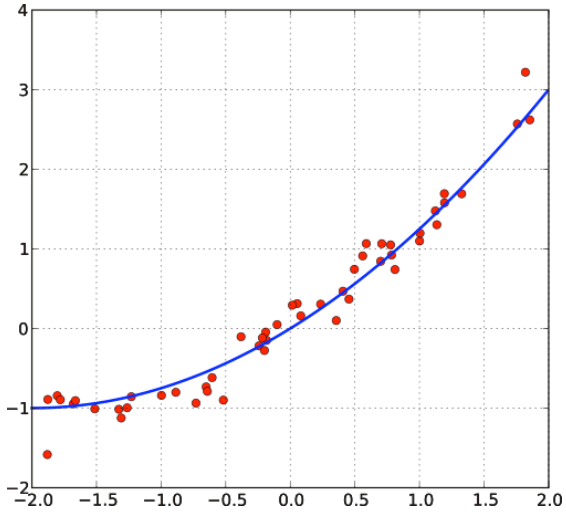
Régression linéaire

Modèle de régression linéaire

- En apprentissage automatique (machine learning), un modèle de **régression linéaire** est un modèle de régression qui cherche à établir une relation linéaire entre une variable *Y*, dite expliquée, et une ou plusieurs variables *X*, dites explicatives. $Y = X\beta + \epsilon$
- Exemple: Ajustement affine
  - consiste à rechercher la droite permettant d'expliquer le comportement d'une variable statistique *y* comme étant une fonction affine d'une autre variable statistique *x*.



Comportement linéaire d'un nuage de points



Comportement non linéaire d'un nuage de points

Présentation formelle

- Notation scalaire
  - Pour chaque individu *i*, la variable expliquée s'écrit comme une fonction linéaire des variables explicatives.

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_K x_{i,K} + \varepsilon_i$$

où y_i et les $x_{i,j}$ sont fixes et ε_i représente l'erreur.

- Notation vectorielle
  - Similaire à la notation scalaire
  - β le vecteur des paramètres du modèle (β_0 , … , β_K) et x_i' le vecteur des variables explicatives pour l'individu *i* (1, $x_{i,1}$, … , $x_{i,K}$)

$$y_i = x_i'\beta + \varepsilon_i$$

- Notation matricielle

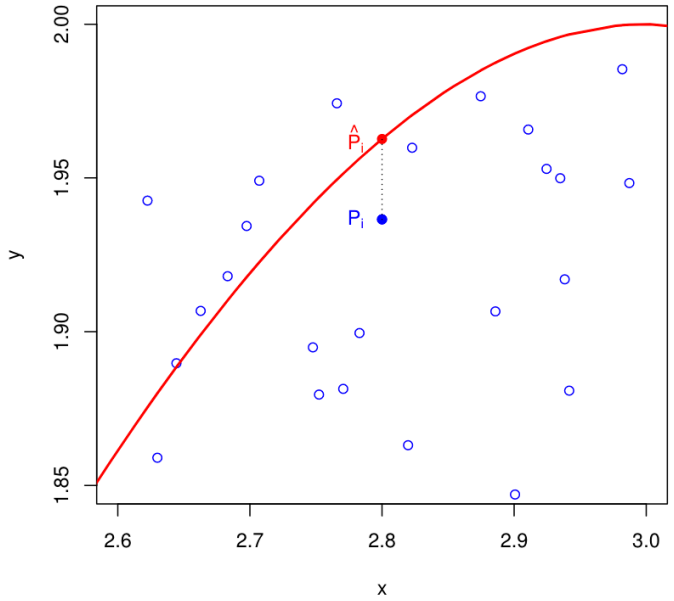
$$Y = X\beta + \varepsilon$$

avec

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Estimation des paramètres

- **La méthode des moindres carrés** consiste à minimiser la somme des carrés des écarts entre les points observés P_i et les points ajustés P_i'.
- Dans le cas d'une régression de Y par rapport à X, il s'agit d'écarts verticaux.
- Dans le cas d'une régression de X par rapport à Y, il s'agit des écarts horizontaux.
- Notons y = f(x) = ax + b l'équation de la droite $D_{Y|X}$ de régression de Y par rapport à X. Il faut trouver a et b.



Les quantités $\hat{y}_i = a\,x_i + b$, sont les valeurs ajustées des ordonnées y_i. L'écart entre les points P_i et $\hat{P}_i$ est $e_i = y_i - \hat{y}_i$.

La somme des carrés des écarts est donc

$$S = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - a\,x_i - b)^2$$

Cette somme S dépend de a et de b. On pourrait la noter $S(a, b)$. On cherche à la minimiser :

$$\min_{a,b} S(a, b) \Longleftrightarrow \quad \min_{a,b} \sum_{i=1}^{N} (y_i - a\,x_i - b)^2$$

On est donc ramenés à un problème mathématique simple : trouver des nombres a et b qui minimisent la fonction $S(a, b)$.

Il faut bien noter que dans l'expression de $S(a, b)$ les quantités x_i et y_i sont des données fixes. Ce sont a et b qui sont susceptibles de varier.

Pour trouver le minimum, il faut annuler les dérivées partielles de $S(a, b)$ par rapport à a et b. On pose donc :

$$\begin{cases} \dfrac{\partial S}{\partial a} & = 0 \\ \\ \dfrac{\partial S}{\partial b} & = 0 \end{cases}$$

Calculons, pour commencer, la dérivée partielle de $\sum_{i=1}^{N}(y_i - a\,x_i - b)^2$ par rapport à a. Il suffit de dériver chaque terme $(y_i - a\,x_i - b)^2$. On obtient donc, en appliquant la formule $(u^2)' = 2u\,u'$:

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{N} 2 \times (y_i - a\,x_i - b) \times (-x_i) = -2\sum_{i=1}^{N}(x_i y_i - a\,x_i^2 - b x_i)$$

En annulant cette dérivée, on obtient l'équation :

$$\begin{aligned} & \sum x_i y_i - a \sum x_i^2 - b \sum x_i = 0 \\ \iff \quad & a \sum x_i^2 + b \sum x_i = \sum x_i y_i \end{aligned} \tag{1}$$

Les sommes figurant dans cette dernière relation sont toutes facilement calculables car les x_i et les y_i sont fixes et connus.

Calculons, maintenant, la dérivée partielle de $\sum_{i=1}^{N}(y_i - a\,x_i - b)^2$ par rapport à b. On obtient :

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{N} 2 \times (y_i - a\,x_i - b) \times (-1) = -2\sum_{i=1}^{N}(y_i - a\,x_i - b)$$

En annulant cette dérivée, on obtient l'équation :

$$\begin{aligned} & \sum y_i - a \sum x_i - b \sum 1 = 0 \\ \iff \quad & a \sum x_i + N\,b = \sum y_i \end{aligned} \tag{2}$$

On a donc obtenu les deux équations (1) et (2) en a et b :

$$\begin{cases} a \sum x_i^2 + b \sum x_i = \sum x_i y_i \\ a \sum x_i + N\,b = \sum y_i \end{cases}$$

En divisant par N, on reconnaît différentes moyennes :

$$\begin{cases} a\,\overline{x^2} + b\,\bar{x} = \overline{xy} \\ a\,\bar{x} + b = \bar{y} \end{cases}$$

En multipliant la deuxième équation par $\bar{x}$ et en la soustrayant à la première, on obtient la valeur de a :

$$\boxed{a = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}}$$

Une fois que a est calculé, on obtient b grâce à la deuxième équation :

$$\boxed{b = \bar{y} - a\,\bar{x}}$$

On remarque que le numérateur de a est la formule développée de la covariance tandis que le dénominateur est la formule développée de la variance. Autrement dit, on a :

$$\boxed{a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}}$$

La deuxième équation peut s'écrire $\bar{y} = a\,\bar{x} + b$. Sous cette forme, elle s'interprète en disant que le barycentre G du nuage de points, c'est-à-dire le point de coordonnées $(\bar{x}, \bar{y})$, se trouve sur la droite de régression.

L'équation de $\mathcal{D}_{Y|X}$ s'écrit finalement :

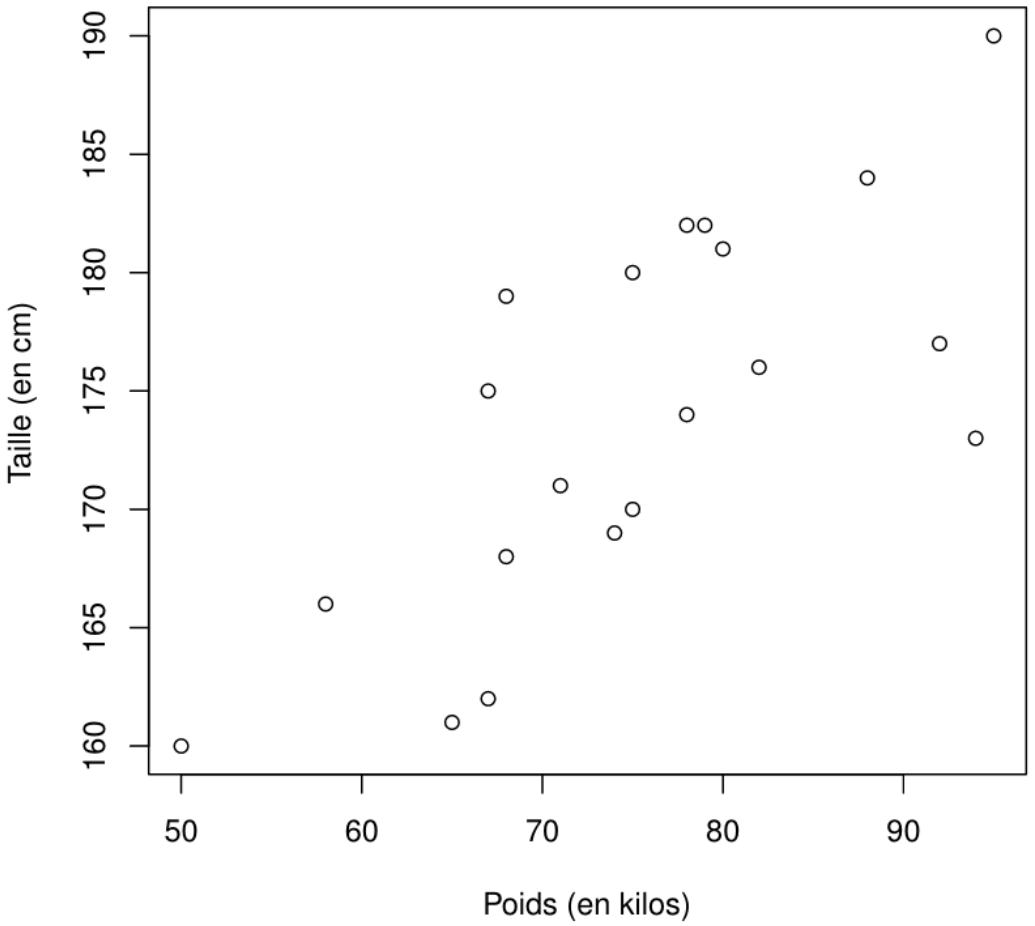
$$\boxed{y = \bar{y} + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\left(x - \bar{x}\right)}$$

Exemple

On a relevé le poids X (en kilos) et la taille Y (en centimètres) de 20 individus dans le tableau suivant :

Poids	67	71	92	74	75	94	79	58	65	68
Taille	162	171	177	169	170	173	182	166	161	179
Poids	88	78	78	75	67	95	80	50	82	68
Taille	184	174	182	180	175	190	181	160	176	168

On va calculer la droite de régression de la taille par rapport au poids.
Voici le diagramme de dispersion correspondant à cette double distribution



On commence par calculer les moyennes des deux variables :

$$\bar{x} = 75.2 \qquad \bar{y} = 174$$

Pour trouver le coefficient a, on a besoin de calculer la covariance des deux distributions et la variance de X (le poids). On trouve :

$$\mathrm{Cov}(X, Y) = 65.75 \quad \text{et} \quad \mathrm{Var}(X) = 129.16$$

d'où

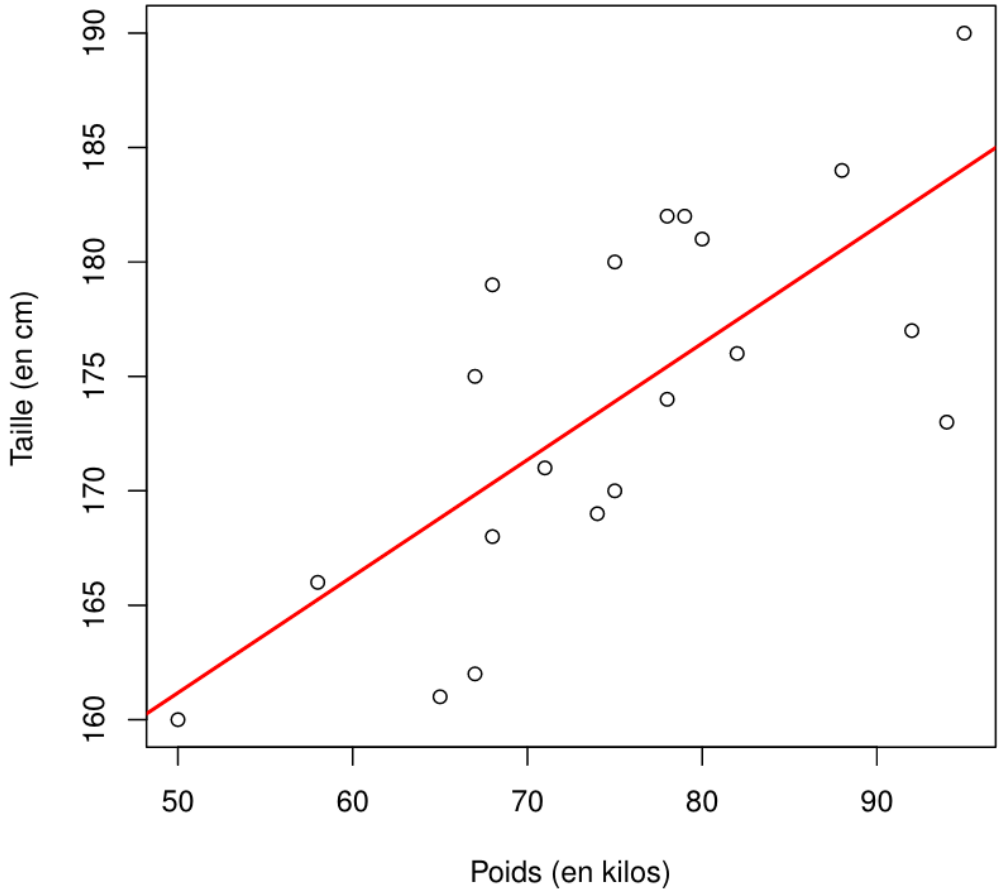
$$a = \frac{65.75}{129.16} = 0.509$$

On calcule ensuite le coefficient b :

$$b = \bar{y} - a\,\bar{x} = 174 - 0.509 \times 75.2 = 135.72$$

L'équation de la droite $\mathcal{D}_{Y|X}$ de régression est finalement :

$$\boxed{y = 0.509\,x + 135.72}$$



On peut facilement calculer les valeurs ajustées

$$\hat{y}_i = a\, x_i + b = 0.509\, x_i + 135.72$$

On obtient :

169.83	171.86	182.55	173.39	173.90	183.57	175.93
165.24	168.81	170.33	180.52	175.43	175.43	173.90
169.83	184.08	176.44	161.17	177.46	170.33	

De même, les écarts e_i sont obtenus en effectuant la différence entre les valeurs observées y_i et les valeurs ajustées $\hat{y}_i$:

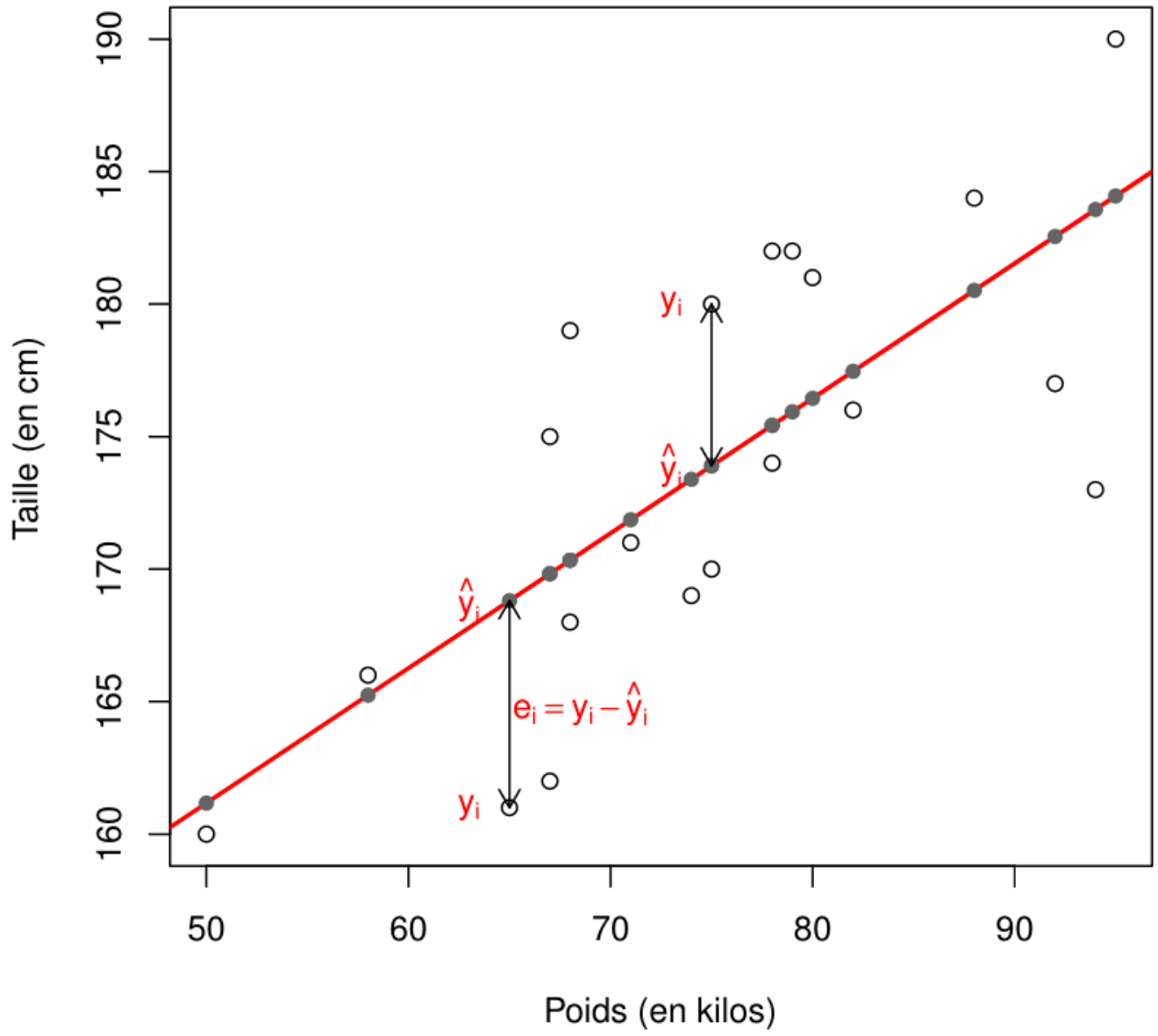
$$e_i = y_i - \hat{y}_i$$

-7.83	-0.86	-5.55	-4.39	-3.90	-10.57	6.07	0.76	-7.81	8.67
3.48	-1.43	6.57	6.10	5.17	5.92	4.56	-1.17	-1.46	-2.33

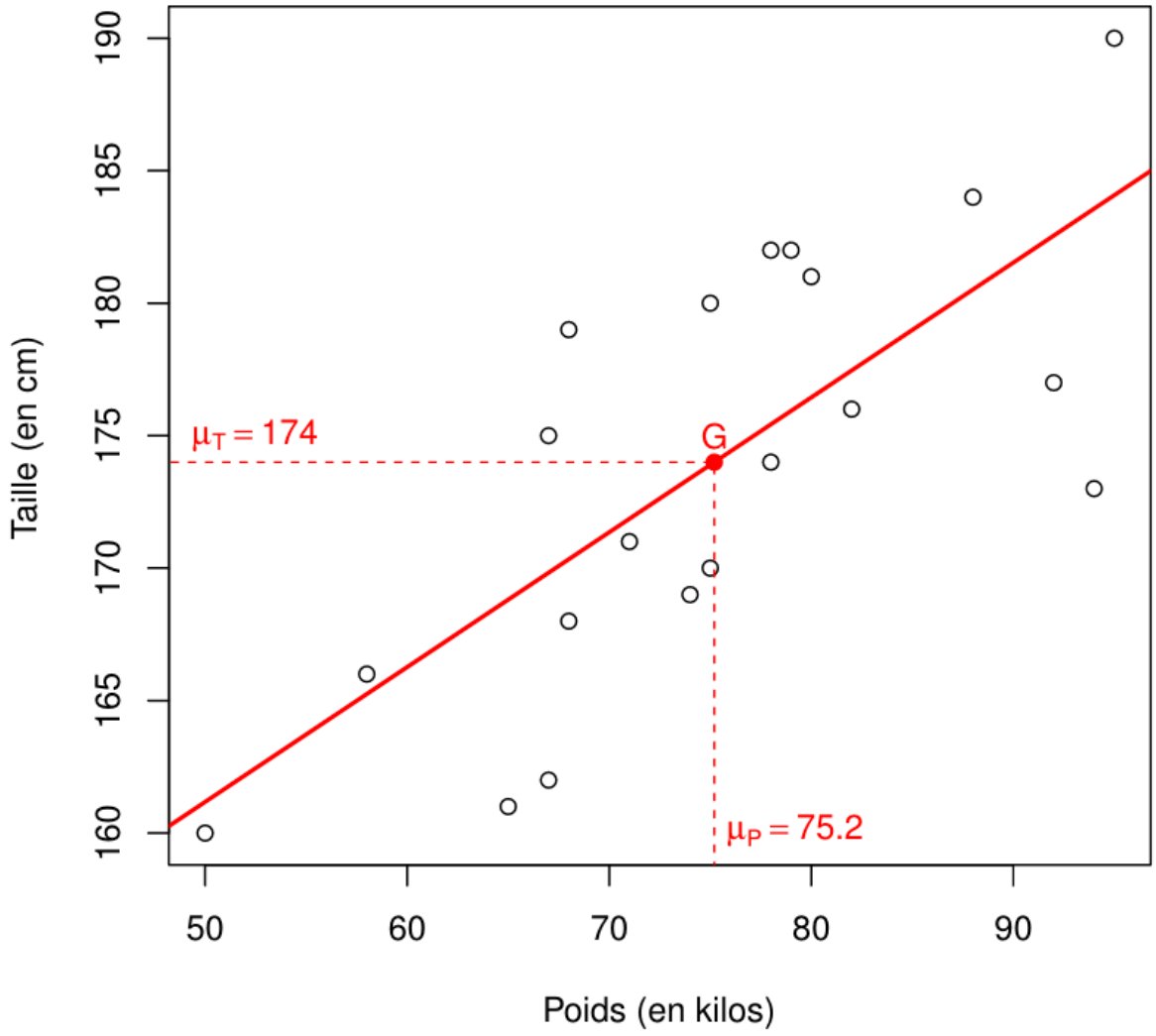
On peut enfin calculer la somme des carrés des résidus :

$$SCR = \sum e_i^2 = 598.59$$

Les écarts e_i sont parfois appelés des *résidus*. Les valeurs ajustées sont toujours sur la droite de régression.


Taille (en cm)
160
165
170
175
180
185
190
50
60
70
80
90
Poids (en kilos)
y_i
$\hat{y}_i$
$\hat{y}_i$
$e_i = y_i - \hat{y}_i$
y_i

On vérifie que la droite de régression passe par le barycentre du nuage de points.



Propriétés des résidus

- La moyenne des résidus est toujours nulle.

$$\begin{aligned}
\bar{e} &= \frac{1}{N}\sum e_i \\
&= \frac{1}{N}\sum (y_i - ax_i - b) \\
&= \frac{1}{N}\sum y_i - \frac{1}{N}\sum ax_i - \frac{1}{N}\sum b \\
&= \bar{y} - a\bar{x} - b \\
&= b - b = 0
\end{aligned}$$

La droite de régression de Y par rapport à X fait en sorte que les écarts des points situés au-dessus compensent les écarts des points situés en dessous. Il en est de même pour les écarts horizontaux dans le cas de l'autre droite de régression.

Vérifions cette propriété sur l'exemple précédent. On a :

$$\begin{aligned}
\sum e_i = &-7.83 - 0.86 - 5.55 - 4.39 - 3.90 - 10.57 \\
&+ 6.07 + 0.76 - 7.81 + 8.67 + 3.48 - 1.43 + 6.57 \\
&+ 6.10 + 5.17 + 5.92 + 4.56 - 1.17 - 1.46 - 2.33 \\
&= 0
\end{aligned}$$

Intervention des variables

Pour effectuer une régression de X par rapport à Y (plutôt que Y par rapport à X), il suffit de reprendre les formules trouvées et d'échanger le rôle des deux variables.

Notons $x = a'\,y + b'$ l'équation de la droite $\mathcal{D}_{X|Y}$ de régression de X par rapport à Y. On a les formules suivantes :

$$\begin{cases} a' &= \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} \\ b' &= \bar{x} - a'\,\bar{y} \end{cases}$$

L'équation de $\mathcal{D}_{X|Y}$ s'écrit finalement :

$$\boxed{x = \bar{x} + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\,(y - \bar{y})}$$

La pente a de $\mathcal{D}_{Y|X}$ s'appelle le coefficient de régression de Y en X. De même la pente a' de $\mathcal{D}_{X|Y}$ s'appelle le coefficient de régression de X en Y.

Reprenons les données précédentes pour calculer la droite de régression du poids par rapport à la taille.

Pour trouver le coefficient a', on a besoin de la covariance des deux distributions et de la variance de Y (la taille). On a :

$$\mathrm{Cov}(X, Y) = 65.75 \quad \text{et} \quad \mathrm{Var}(Y) = 63.4$$

d'où

$$a' = \frac{65.75}{63.4} = 1.037$$

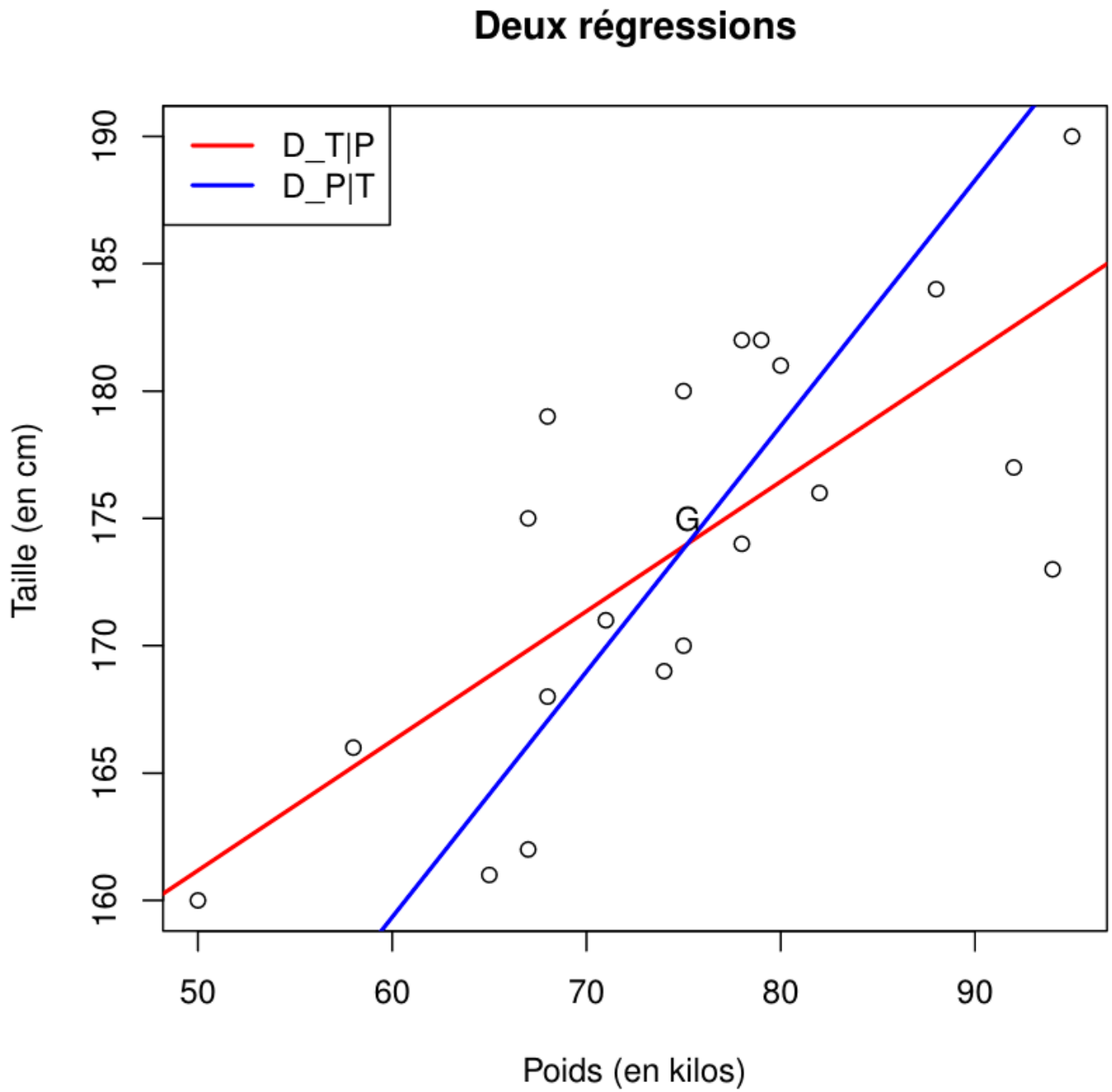
On calcule ensuite le coefficient b' :

$$b' = \bar{x} - a'\,\bar{y} = 75.2 - 1.037 \times 174 = -105.24$$

L'équation de la droite $\mathcal{D}_{Y|X}$ de régression est finalement :

$$x = 1.037\,y - 105.24 \Longleftrightarrow y = 0.96\,x + 101.49$$

Les droites de régression se croisent au barycentre du nuage de points.


Deux régressions
D_T|P
D_P|T
G
Taille (en cm)
Poids (en kilos)
160
165
170
175
180
185
190
50
60
70
80
90

Qualité de la prédiction: Coefficient de corrélation linéaire

Lorsque les variables X et Y sont indépendantes, leur covariance est nulle. Par conséquent, on a dans ce cas $a = a' = 0$ et donc $b = \bar{y}$ et $b' = \bar{x}$. Les équations des droites de régression se simplifient pour devenir $y = \bar{y}$ et $x = \bar{x}$. Elles sont parallèles aux axes et perpendiculaires entre elles. On note que, dans ce cas, $a\,a' = 0$.

Supposons maintenant qu'il y a dépendance totale et que Y est un multiple de X, par exemple $Y = \alpha X$. Dans ce cas, on trouve

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}(X, \alpha X)}{\mathrm{Var}(X)} = \alpha \frac{\mathrm{Cov}(X, X)}{\mathrm{Var}(X)} = \alpha$$

De la même manière on obtient $a' = \dfrac{1}{\alpha}$. On note que dans ce cas $a\,a' = \alpha \times \dfrac{1}{\alpha} = 1$.

On en déduit que le produit des pentes $a\,a'$ semble être un bon indicateur de l'indépendance ou de la dépendance des deux variables. Ce produit est toujours positif (car on a vu que a et a' sont toujours de meme signe) et on peut donc poser :

$$\boxed{r^2 = a\,a'}$$

Si on remplace les coefficients a et a' par leur formule, on trouve que

$$r^2 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \times \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} = \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X) \times \mathrm{Var}(Y)}$$

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

En reprenant les données précédentes, on calcule :

$$r^2 = a\,a' = 0.509 \times 1.037 = 0.5278$$

d'où $r = \sqrt{0.5278} = 0.7265$ (car les variables sont ici corrélées positivement).

Exercice

Le tableau qui suit provient du jeu de données *cars* du logiciel R. Il comporte 20 observations (parmi 50 à l'origine) correspondant à la vitesse d'automobiles et à la distance qui est nécessaire pour les stopper. Ce sont des données qui ont été relevées dans les années 1920.

Les vitesses sont en miles par heure (mph) et les distances en pieds (ft).

Vitesse	4	7	8	9	10	11	11	12	12	13
Distance	2	4	16	10	26	17	28	20	28	26
Vitesse	14	15	15	16	17	18	19	20	24	25
Distance	36	26	54	40	50	76	46	48	92	85

1-1) Calculer la moyenne arithmétique des distances et celle des vitesses.

1-2) Calculer la variance des distances et celle des vitesses.

1-3) Calculer la droite de régression $\mathcal{D}_{D|V}$ de la distance par rapport à la vitesse par la méthode des moindres carrés.

1-4) Calculer la droite de régression $\mathcal{D}_{V|D}$ de la vitesse par rapport à la distance.

1-5) Donner une représentation graphique de ces deux droites. Calculer l'angle qu'elles font entre elles.

1-6) Calculer le coefficient de corrélation linéaire. Que conclure sur le degré de dépendance entre les variables ?