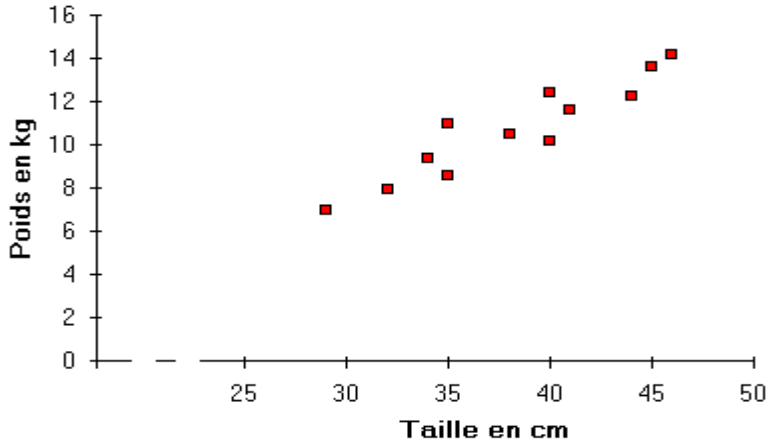
LES PETITS HOMMES VERTS

Notre première étude sur les petits hommes verts date de mai 1985.
Nous nous proposions de déterminer la nature de la relation entre : la taille X mesurée en cm et le poids Y exprimé en kg de ces petits hommes verts.
Nous disposions de données mesurées sur un échantillon de 12 petits hommes verts.

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	29	32	35	34	40	38	35	41	44	40	45	46
y_i	7	8	8.6	9.4	10.2	10.5	11	11.6	12.3	12.4	13.6	14.2

La première tâche fut bien sûr de représenter le nuage des points de coordonnées (x_i , y_i).

* Représentation graphique des données:



Compte tenu du profil du nuage, rien ne semblait s'opposer à un ajustement affine du nuage.
En d'autres termes , nous proposions la modélisation suivante:
Si $\mathbf{x_i}$ est la taille d'un petit homme vert

* son poids théorique vaut $y_i^* = \alpha \quad x_i \quad + \quad \beta$
* la différence entre le poids réel $\mathbf{y_i}$ de cet individu et le poids théorique est l'erreur notée $\boldsymbol{\varepsilon_i}$
* la répartition des erreurs ne dépend pas de $\mathbf{x_i}$
* Cette répartition est supposée Normale $\mathbf{N(0,\sigma)}$.

Modèle de paramètres α **et** β.

$$Y = \alpha \quad x \quad + \quad \beta \quad + \quad \varepsilon$$

α **et** β sont les paramètres du modèle.

La taille	**x**	est la variable ***indépendante (explicative)***et ***non aléatoire***
Le poids	**Y**	est la variable ***aléatoire dépendante(expliquée).***
L'erreur	ε	est une variable aléatoire distribuée selon la loi Normale $\mathbf{N(0,\sigma)}$.
Les tailles	$\mathbf{x_i}$	sont 12 valeurs de taille **x** connues sans erreur de mesure.
Les erreurs	$\boldsymbol{\varepsilon_i}$	sont 12 réalisations mutuellement indépendantes de la variable ε
Chaque poids	$\mathbf{y_i}$	est une réalisation de la variable $Y_i = \alpha \quad x_i \quad + \quad \beta \quad + \quad \varepsilon$

α, β et σ **sont les paramètres du modèle.**

Nous devions alors :

1. Estimer les paramètres α **et** β du modèle.
2. Apprécier l'adéquation du modèle aux données recensées et à d'éventuelles données supplémentaires.

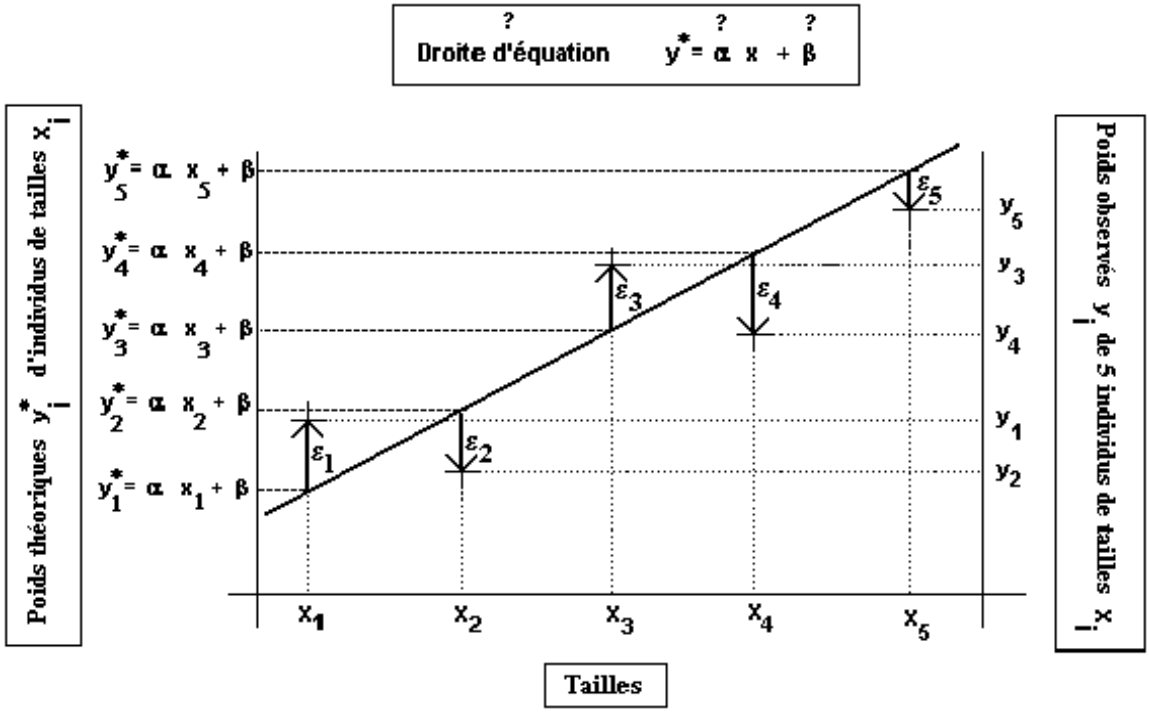
La "*meilleure*" méthode d'ajustement est la **méthode des moindres carrés.**

Droite d'ajustement selon la méthode des moindres carrés

Pour simplifier la présentation de cette méthode, nous allons dans les calculs et illustrations qui suivent, nous limiter à un nuage de 5 points.

A. Résidus et Erreurs : des notions à bien différencier.

1. Le schéma ci-dessous permet de comprendre la notion d'erreur.



Les paramètres α **et** β sont connus par le dieu des petits hommes verts, mais inaccessibles à nous, pauvres mortels.

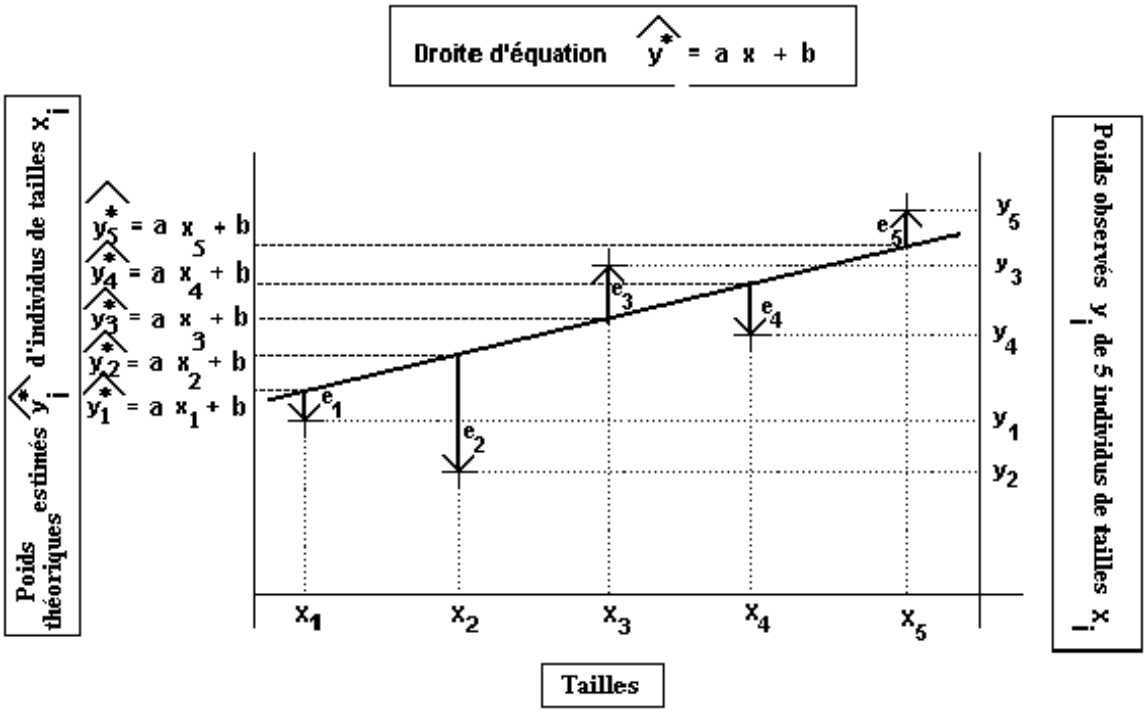
α et β nous sont à jamais inconnus

Par conséquent, les cinq erreurs sont inconnues

ε_1 , ε_2 , ε_3 , ε_4 , ε_5 **nous sont à jamais inconnues**

2. Le schéma ci-dessous permet de comprendre la notion de résidu.

Considérons une droite d'**ajustement empirique** d'équation **y = a x + b**



Les observations faites nous permettent de proposer deux nombres **a et b** comme estimations empiriques des paramètres **α et β**	$\hat{y}^*_1$, $\hat{y}^*_2$, $\hat{y}^*_3$, $\hat{y}^*_4$, $\hat{y}^*_5$ sont les estimations empiriques des poids théoriques d'individus de tailles respectives $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, $\mathbf{x_4}$, $\mathbf{x_5}$

e_1 , e_2 , e_3 , e_4 , e_5
sont
les **résidus** relatifs à l'ajustement proposé **pour l'échantillon observé**

B. Droite d'ajustement.

Quelles sont les contraintes auxquelles doit satisfaire la droite d'ajustement selon la méthode des moindres carrés ?

1. La moyenne des résidus e_i doit être nulle. ***(justesse)***

$$e_1 + e_2 + e_3 + e_4 + e_5 = 0$$

2. La variabilité des résidus e_i doit être minimale. ***(précision)***

$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$ doit être minimale

des contraintes somme toute très naturelles !!

1. Analysons la première contrainte,

Ajoutons membre à membre les 5 égalités suivantes :

$$\begin{aligned} y_1 &= a\,x_1 + b + e_1 \\ y_2 &= a\,x_2 + b + e_2 \\ y_3 &= a\,x_3 + b + e_3 \\ y_4 &= a\,x_4 + b + e_4 \\ y_5 &= a\,x_5 + b + e_5 \end{aligned}$$

Nous obtenons :

$$y_1 + y_2 + y_3 + y_4 + y_5 = a\,(x_1 + x_2 + x_3 + x_4 + x_5) + 5\,b + e_1 + e_2 + e_3 + e_4 + e_5$$

Divisons chaque membre par 5 :

$$\bar{y} = a\,\bar{x} + b + \bar{e}$$

Comme $\bar{e}$ doit valoir 0, il en résulte:

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

2. <u>Analysons la seconde contrainte</u>

$$\left.\begin{aligned} y_1 &= a\,x_1 + b + e_1 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_1 - \bar{y}) - a\,(x_1 - \bar{x}) = e_1$$

$$\left.\begin{aligned} y_2 &= a\,x_2 + b + e_2 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_2 - \bar{y}) - a\,(x_2 - \bar{x}) = e_2$$

$$\left.\begin{aligned} y_3 &= a\,x_3 + b + e_3 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_3 - \bar{y}) - a\,(x_3 - \bar{x}) = e_3$$

$$\left.\begin{aligned} y_4 &= a\,x_4 + b + e_4 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_4 - \bar{y}) - a\,(x_4 - \bar{x}) = e_4$$

$$\left.\begin{aligned} y_5 &= a\,x_5 + b + e_5 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_5 - \bar{y}) - a\,(x_5 - \bar{x}) = e_5$$

Il en résulte :

$$\begin{aligned} e_1^2 &= (y_1 - \bar{y})^2 - 2a\,(x_1 - \bar{x})(y_1 - \bar{y}) + a^2\,(x_1 - \bar{x})^2 \\ e_2^2 &= (y_2 - \bar{y})^2 - 2a\,(x_2 - \bar{x})(y_2 - \bar{y}) + a^2\,(x_2 - \bar{x})^2 \\ e_3^2 &= (y_3 - \bar{y})^2 - 2a\,(x_3 - \bar{x})(y_3 - \bar{y}) + a^2\,(x_3 - \bar{x})^2 \\ e_4^2 &= (y_4 - \bar{y})^2 - 2a\,(x_4 - \bar{x})(y_4 - \bar{y}) + a^2\,(x_4 - \bar{x})^2 \\ e_5^2 &= (y_5 - \bar{y})^2 - 2a\,(x_5 - \bar{x})(y_5 - \bar{y}) + a^2\,(x_5 - \bar{x})^2 \end{aligned}$$

Notons σ_x et σ_y les écart-types respectifs des séries statistiques (x_i) et (y_i)

$$\sigma_x^2 = \frac{\sum_{i=1}^{5} (x_i - \bar{x})^2}{5} \qquad \text{et} \qquad \sigma_y^2 = \frac{\sum_{i=1}^{5} (y_i - \bar{y})^2}{5}$$

Notons σ_{xy} la covariance de la série double :

$$\sigma_{xy} = \frac{\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y})}{5}$$

La variabilité des résidus vaut donc :

$$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 = 5\sigma_x^2\ a^2\ -10\ \sigma_{xy}\ a\ +\ 5\ \sigma_y^2$$

La variabilité des résidus est une fonction du second degré en a qu'il s'agit de minimiser en choisissant judicieusement a.

Etudions la fonction f définie par : $f(a) = 5\sigma_x^2\ a^2\ -10\ \sigma_{xy}\ a\ +\ 5\ \sigma_y^2$

Dérivée de f : $f'(a) = 10\sigma_x^2\ a\ -10\ \sigma_{xy}$

Signe de la dérivée de f :

$f'(a) > 0$ si et seulement si $a > \dfrac{\sigma_{xy}}{\sigma_x^2}$

$f'(a) = 0$ si et seulement si $a = \dfrac{\sigma_{xy}}{\sigma_x^2}$

Désignons par $\hat{\alpha}$ la quantité $\dfrac{\sigma_{xy}}{\sigma_x^2}$

Variations de f :

a	$-\infty$	$\hat{\alpha} = \dfrac{\sigma_{xy}}{\sigma_x^2}$	$+\infty$
f '(a)	$-$	0	$+$
	$+\infty$		$+\infty$
f (a)	↘	$f(\hat{\alpha})$	↗

La droite d'ajustement selon la méthode des moindres carrés a donc pour équation :

$$\mathbf{y} = \hat{\alpha}\ \mathbf{x} + \hat{\beta}$$

où $\hat{\alpha} = \dfrac{\sigma_{xy}}{\sigma_x^2}$ et $\hat{\beta} = \bar{y} - \hat{\alpha}\ \bar{x}$

En posant $\mathbf{r} = \dfrac{\sigma_{xy}}{\sigma_x\ \sigma_y}$, nous obtenons $\hat{\alpha} = \mathbf{r}\dfrac{\sigma_y}{\sigma_x}$.

La variabilité des résidus relatifs à cet ajustement vaut

$$f(\hat{\alpha}) = 5\sigma_x^2\ \hat{\alpha}^2\ -10\ \sigma_{xy}\ \hat{\alpha} + 5\ \sigma_y^2.$$

Exprimons cette variabilité en fonction de **r** :

$$f(\hat{\alpha}) = f\left(\mathbf{r}\frac{\sigma_y}{\sigma_x}\right) = 5\sigma_x^2\ \mathbf{r}^2\ \frac{\sigma_y^2}{\sigma_x^2}\ -10\ \sigma_{xy}\ \mathbf{r}\frac{\sigma_y}{\sigma_x} + 5\ \sigma_y^2$$

soit $$f(\hat{\alpha}) = 5\ \mathbf{r}^2\ \sigma_y^2\ -10\ \ \mathbf{r}^2\ \sigma_y^2\ +\ 5\ \sigma_y^2$$

soit $f\left(\hat{\alpha}\right) = 5 \quad \sigma_y^2 \quad \left(1 - \mathbf{r}^2\right)$

La variabilité des résidus relatifs à l'ajustement selon la méthode des moindres carrés vaut donc :

$$\boxed{e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 = 5 \quad \sigma_y^2 \quad \left(1 - \mathbf{r}^2\right)} \qquad \text{Formule n° 1}$$

Quelque remarques :

* La valeur de r ne dépend pas des unités choisies pour mesurer les x_i et y_i
  r est donc un coefficient.
* Compte tenu de la positivité de f : **r est compris entre -1 et 1**.
* Si $\mathbf{r}^2 = 1$, la variabilité des résidus est nulle : les points du nuage sont alors alignés.
* La variabilité des résidus est d'autant plus faible que r² est proche de 1 (à σ_y constant)

La valeur de r² permet d'apprécier la qualité de l'ajustement.

* Le signe de r est celui de la pente de la droite : **r est du signe de $\hat{\alpha}$**

r est appelé coefficient de corrélation linéaire de la série (x_i, y_i).

C. Décomposition de la variabilité totale.

* Nous désignerons par variabilité résiduelle la somme des carrés des résidus notée $\mathbf{SC_{Res}}$

$$\mathbf{SC_{Res}} = \sum_{i=1}^{5} \mathbf{e_i^2} = \sum_{i=1}^{5} \left(\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}^* \right)^2$$

* Nous désignerons par variabilité totale, celle des (y_i)

$$\mathbf{SC_{Tot}} = \sum_{i=1}^{5} \left(\mathbf{y_i} - \bar{\mathbf{y}}\right)^2$$

* Nous désignerons par variabilité expliquée la quantité notée $\mathbf{SC_{Exp}}$ définie par

$$\mathbf{SC_{Exp}} = \mathbf{SC_{Tot}} - \mathbf{SC_{Res}}$$

Compte tenu de la formule n°1

$$\mathbf{SC_{Exp}} = \mathbf{r}^2 \sum_{i=1}^{5} \left(\mathbf{y_i} - \bar{\mathbf{y}}\right)^2$$

Ainsi

$$\boxed{\mathbf{r}^2 = \frac{\mathbf{SC_{Exp}}}{\mathbf{SC_{Tot}}} = \frac{\textbf{Variabilité Expliquée}}{\textbf{Variabilité Totale}}}$$

Dans la mesure où $\mathbf{r}^2$ mesure la part de variabilité expliquée par l'ajustement

$\mathbf{r}^2$ est appelé **coefficient de détermination**

* Autre expression de $\mathbf{SC_{Exp}}$:

$$\mathbf{SC_{Exp}} = 5 \quad \mathbf{r}^2 \quad \sigma_y^2 \quad = 5 \quad \hat{\alpha}^2 \quad \sigma_x^2$$

Comme $\hat{y}_i^* = \hat{\alpha} \quad x_i + \hat{\beta}$, $\hat{\alpha}^2 \, \sigma_x^2$ est la variance de la série statistique $(\hat{y}_i^*)$

N'oublions pas d'autre part, que la moyenne des (y_i) et celle des $(\hat{y}_i^*)$ sont égales:

Il en résulte $$\boxed{SC_{Exp} = \sum_{i=1}^{5} \left(\hat{y}_i^{*} - \bar{y} \right)^2}$$

* Formule de décomposition de la variabilité

$$\sum_{i=1}^{5}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{5}\left(\hat{y}^*_i - \bar{y}\right)^2 + \sum_{i=1}^{5}\left(y_i - \hat{y}^*_i\right)^2$$

$$SC_{Tot} = SC_{Exp} + SC_{Res}$$

C. analyse des résultats obtenus pour les douze points.

$\hat{\alpha} = +0.389$

$\hat{\beta} = -4.14$

L'équation de la droite de régression de Y en x est

$$Y = +0.389\ x - 4.14$$

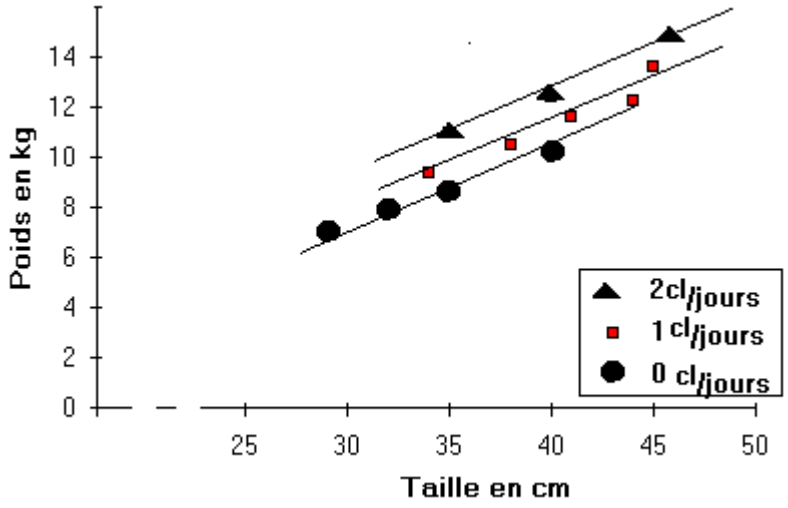
Variabilité expliquée:	47.839
Variabilité résiduelle:	6.528
Variabilité totale:	54.367

Le coefficient de détermination vaut : $r^2 = \frac{47.839}{54.367} = 0{,}88$

Ces résultats semblaient corrects !
Quelle ne fut pas notre surprise, lorsque quelques mois plus tard, furent retrouvés des renseignements égarés :

* à savoir les consommations journalières en eau de chaque individu.

Le nuage de points prenait alors une toute autre signification:



En un seul instant, notre modèle n°1 était anéanti :
Qu'allait-on faire, pour tenir compte de ces nouveaux renseignements?

- = - = - = - = - = - = - = - = -