Part III

Régression linéaire.

On rappelle la régression linéaire entre deux variables d'un point de vue descriptif.

7 Régression linéaire – Statistiques descriptives

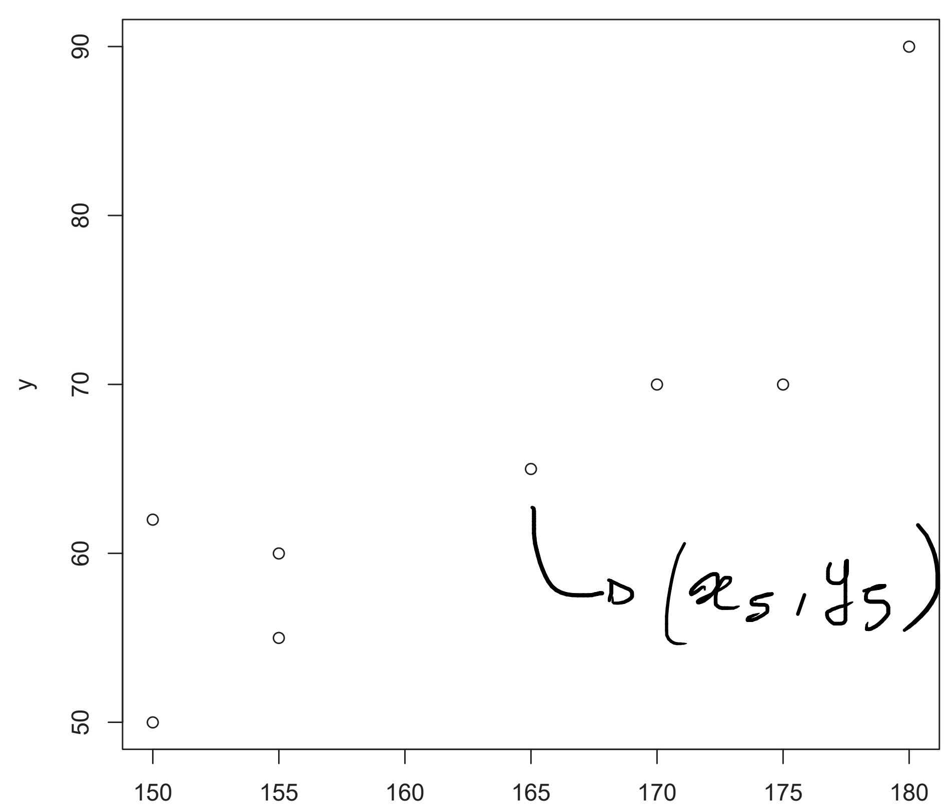
Dans certaines situations, on est amené à étudier deux caractères distincts d'une même population. On peut par exemple considérer la taille (x) et le poids (y) d'un ensemble d'individus. L'objectif principal de l'étude est de déterminer l'éventuel lien entre les deux variables x et y.

7.1 Nuage de points

On relève le couple (taille, poids) de 8 individus. On résume les données dans le tableau suivant.

	i	1	2	3	4	5	6	7	8
taille	x_i	150	155	155	150	165	175	170	180
poids	y_i	50	55	60	62	65	70	70	90

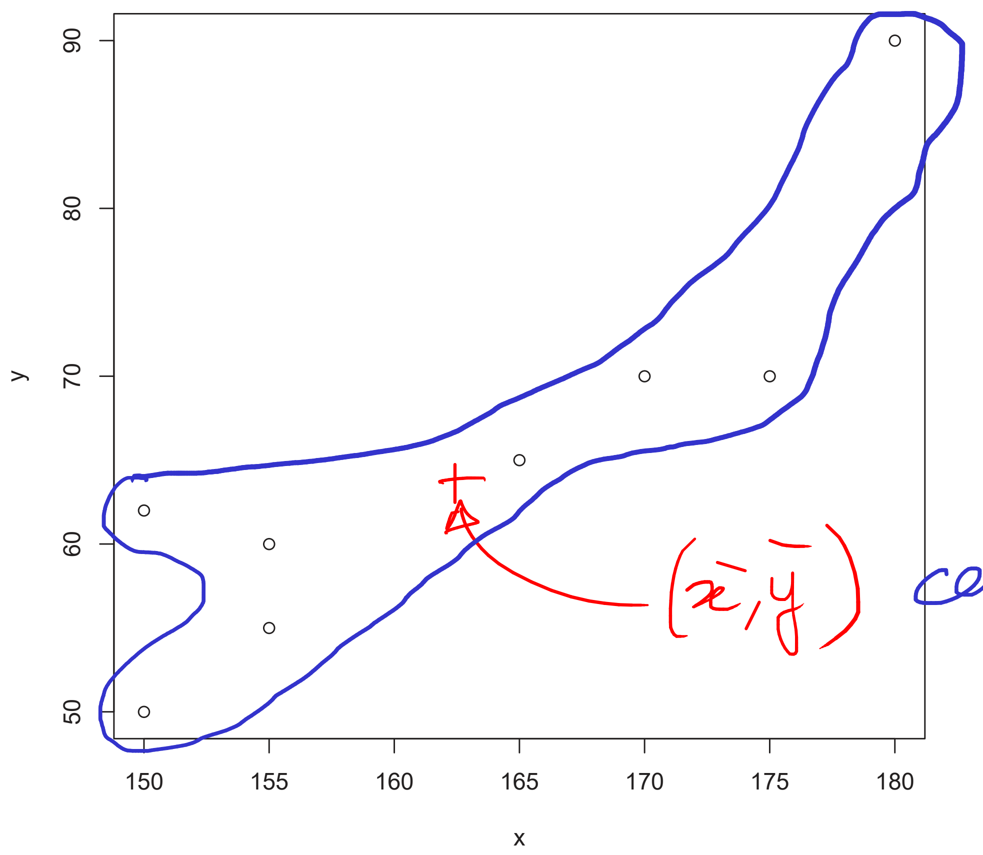
Definition 7.1 *Soit une population de N individus. Le graphe des N points* (x_i, y_i) *est appelé nuage de points de la série.*



$$\bar{x} = \frac{\sum_i x_i}{n}$$

effectif total

$$\bar{y} = \frac{\sum_i y_i}{n}$$



$(\bar{x}, \bar{y})$ centre de gravité du nuage

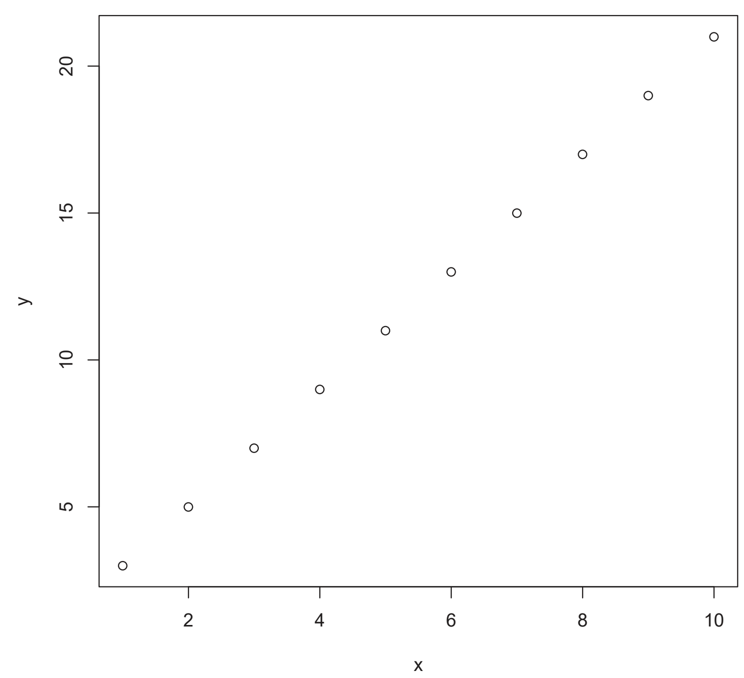
Definition 7.2 *Le point ayant pour coordonnées les moyennes* $(\bar{x}, \bar{y})$ *est appelé le point moyen.*

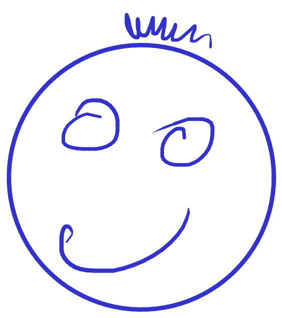
Il s'agit du centre de gravité du nuage. On rencontrera parfois cette dénomination. Dans notre exemple, le point moyen est (65.2, 162.5).

7.2 Forme du nuage de points

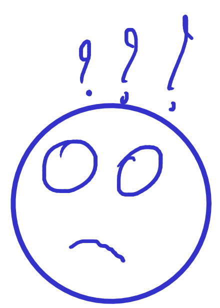
D'une manière générale, trois cas peuvent se présenter en ce qui concerne le profil du nuage :

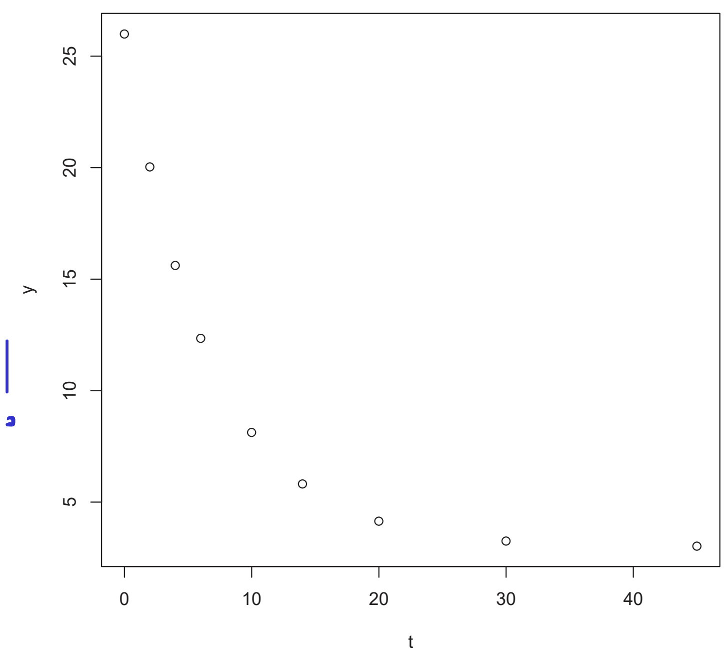
(i) forme allongée et rectiligne : les points sont plus ou moins alignés





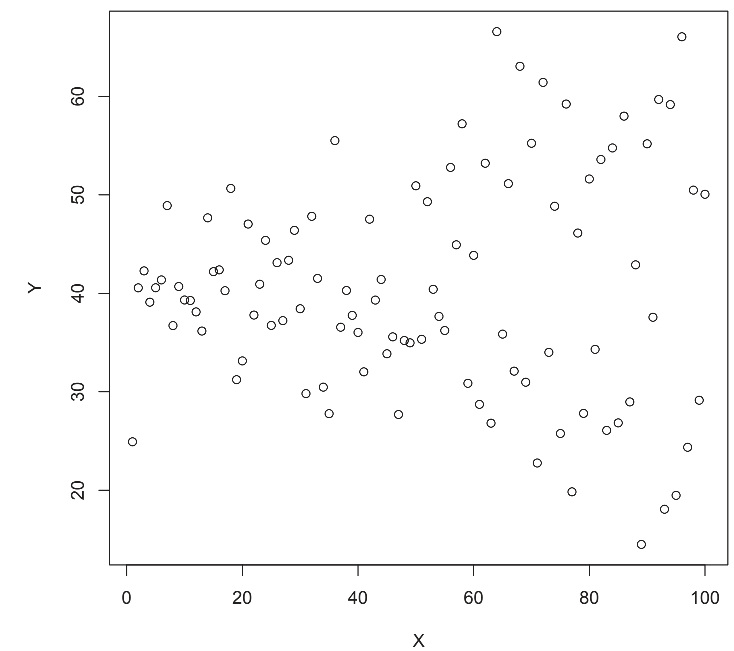
(ii) forme allongée mais non rectiligne : les points ne sont pas alignés mais ont un profil ordonné





à redresser !

(iii) forme quelconque



7.3 Ajustement affine (droite de régression linéaire)

On s'intéresse plus particulièrement au premier cas 7.2.1. Procéder à un ajustement affine revient à chercher une droite D d'équation

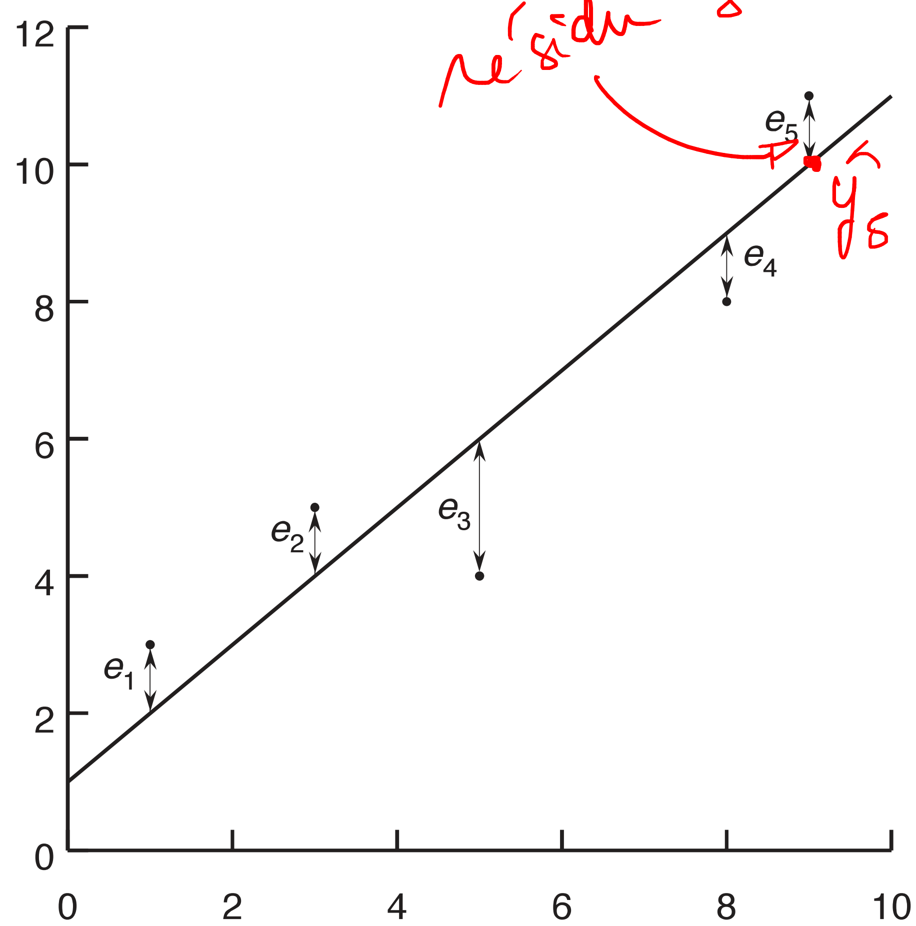
$$y = ax + b$$

7.3.1 La méthode des moindres carrés

L'idée de cette méthode est de chercher la droite qui minimise la somme des carrés des écarts verticaux entre la droite et les points du nuage, les *résidus*.

valeur sur la droite

résidu $y_5 - \hat{y}_5$



$\hat{y}_5$

On suppose qu'il n'y a pas d'erreur sur x. On minimise la somme des carrés des écarts entre y observé et son estimation sur la droite (erreur sur y = résidu)

En pratique, on détermine les coefficients de la droite $D : y = ax + b$ à l'aide de R ou d'un tableur.

La droite ainsi obtenue est unique. Cette droite s'appelle la droite de régression linéaire de y en x par la méthode des moindres carrés et on aura

$$y_i = \underbrace{ax_i + b}_{\hat{y}_i \text{ estimé par la droite.}} + e_i,$$

résidu = erreur

On suppose, pour l'instant, que l'étude se fait sur une population entière (statistiques descriptives) et non sur un échantillon (statistiques inférentielles).

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad \sigma_x = \sqrt{Var(x)}$$

$$\sigma_{x,y} = Cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

Cette quantité est nommée covariance de x et y. Si la quatité σ_x est une distance entre les valeurs de x est $\bar{x}$, on peut considérer la covariance comme un produit scalaire entre les variables x et y. Ainsi, si la covariance est proche de 0, on peut penser que les variables ont une dynamique qui n'ont rien de commun (penser à l'orthogonalité), c'est à dire le nuage 7.2.3.

On a

$$\begin{aligned} a &= cov(x,y)/\sigma_x^2, \\ b &= \bar{y} - a\bar{x}. \Longleftrightarrow \end{aligned}$$

$\bar{y} = a\bar{x} + b \Longleftrightarrow$ la droite passe par le centre de gravité.

rappel

(X,Y) observé par $(x_i, y_i)_{1 \le i \le n}$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$Var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = m(x^2) - \bar{x}^2$$

$$Var(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = m(y^2) - \bar{y}^2$$

$$Cov(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= m(xy) - \bar{x}\,\bar{y}$$

La droite de régression de y en x par la méthode des moindres carrés a pour équation :

$$y = ax + b \quad \text{avec} \quad a = \frac{Cov(x,y)}{Var(x)}$$

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

Modèle : $y_i = a x_i + b + e_i$

erreur

Preuve. On suppose qu'il n'y a pas d'erreur systématique, sinon on pourrait souhaiter translater ou modifier la pente de la droite. On suppose donc que l'erreur moyenne sur nos observations est nulle

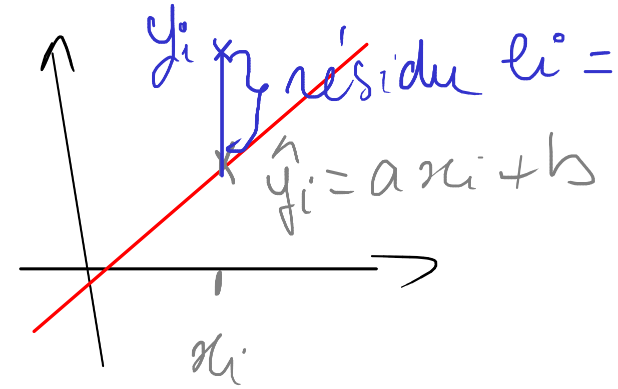
$$\bar{\varepsilon} = \frac{1}{\cancel{N}\, n} \sum_{i=1}^{\cancel{N}\, n} e_i = 0$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} (a x_i + b + e_i) = \frac{1}{n} \sum_{i=1}^{n} a x_i + \underbrace{\frac{1}{n} \sum_{i=1}^{n} b}_{nb} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} e_i}_{\bar{\varepsilon} = 0}$$

$$= a \bar{x} + b + 0$$

D'où $\bar{y} = a\bar{x} + b \iff b = \bar{y} - a\bar{x}$

Résidu = erreur



résidu $e_i = y_i - \hat{y}_i = y_i - (ax_i + b)$

$$\hat{y}_i = ax_i + \underbrace{b}_{\bar{y} - a\bar{x}} = ax_i + \bar{y} - a\bar{x} = a(x_i - \bar{x}) + \bar{y}$$

On pose la somme des carrés des résidus

$$M(a) = \sum_{i=1}^{N} (y_i - \overbrace{(a(x_i - \bar{x}) + \bar{y})}^{\hat{y}_i})^2 = \sum_{i=1}^{N} (y_i - \bar{y} - a(x_i - \bar{x}))^2.$$

Nous cherchons donc la valeur de a qui minimise cette quantité. Dérivons la en a

$$M'(a) = \sum_{i=1}^{N} (u^2(a))' = \sum_{i=1}^{N} 2u'(a)\,u(a)$$

avec $u(a) = \overbrace{y_i - \bar{y}}^{\beta} - a\overbrace{(x_i - \bar{x})}^{\alpha}$

$u'(a) = \underbrace{-(x_i - \bar{x})}_{\alpha}$

$$= 2\sum_{i=1}^{n} -(x_i - \bar{x})\left((y_i - \bar{y}) - a(x_i - \bar{x})\right)$$

$$= 2\left[-\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + a\sum_{i=1}^{n} (x_i - \bar{x})^2\right]$$

Pour minimiser, on annule $M'(a)$.

$$M'(a) = 0 \iff -\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) + a\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$$

$$\iff a = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_i(x_i-\bar{x})^2}$$

$$= \frac{Cov(x,y)}{Var(x)}$$

Vérifions que c'est un minimum :

$$M''(a) = 2\sum_{i=1}^{n}(x_i-\bar{x})^2 > 0$$

La fonction est convexe, c'est bien un minimum.

Fin de la preuve

7.3.2 Coefficient de corrélation linéaire

Notons que la méthode des moindres carrés peut être utilisée pour n'importe quelle série double. On peut tout à fait obtenir une droite de régression dans le cas 7.2.3. Pour s'assurer de façon objective (et non purement visuelle) que l'ajustement est valide, on considère un autre paramètre de la série, le coefficient de corrélation linéaire r :

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$a = \frac{Cov(x,y)}{Var(x)}$ $\quad$ $Var(x) > 0$

$r = \frac{Cov(x,y)}{\underbrace{\sigma_x \sigma_y}_{\geq 0}}$

Proposition 7.3 *On a les propriétés suivantes :*

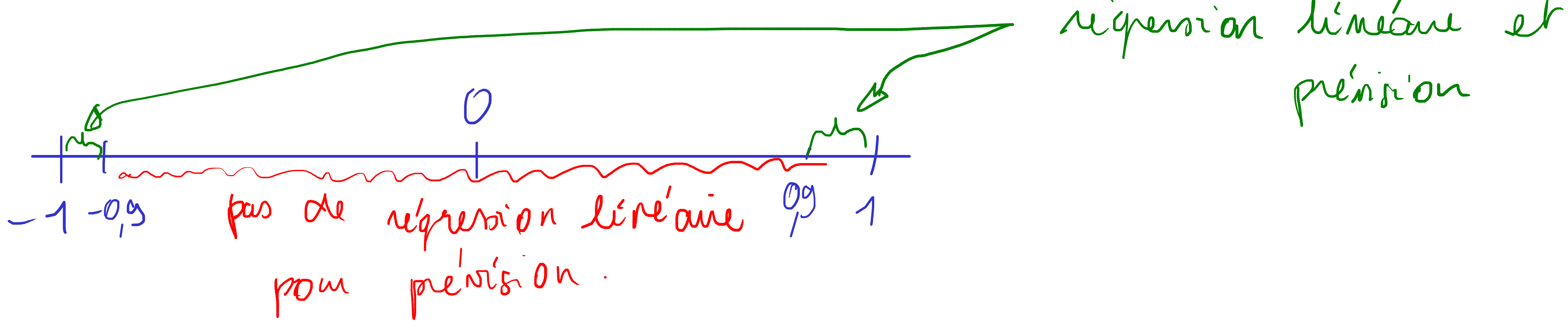
(i) *on a toujours* $-1 \leq r \leq 1$ *;*

(ii) *le coefficient directeur de la droite de régression et le coefficient de corrélation sont de même signe ;* (celui de la covariance).

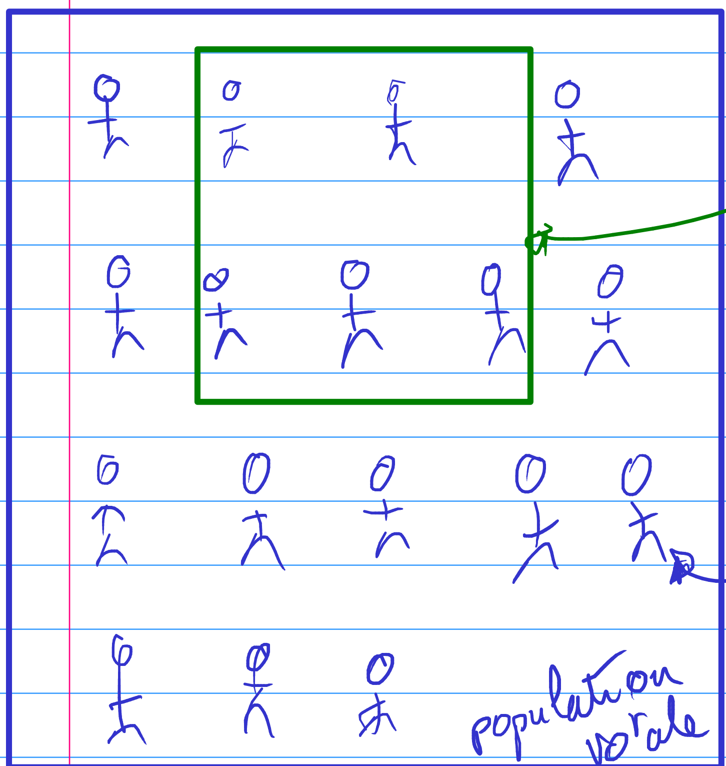
(iii) *le degré de corrélation est d'autant plus fort que* r *est proche de* 1 *ou* -1*.*

↳ plus le nuage est rectiligne

C'est l'assertion 3.*iii* qui nous permet de dire si la droite de régression est proche des points. En pratique, une régression linéaire est légitime si $r > 0.9$ ou si $r < -0.9$.



Point de vue inférentiel : On n'a plus accès à la population totale (stat descriptives) mais à un échantillon.



échantillon aléatoire : (X_i, Y_i).

Après tirage de l'échantillon, il y a une observation $(X_i = x_i, Y_i = y_i)$.

(x_i, y_i)

Estimation ponctuelle de la moyenne μ_x et μ_y :

pour μ_X : $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ en observant : $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

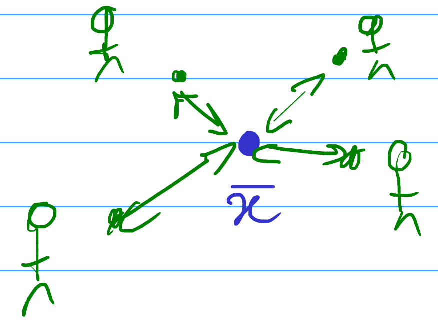
pour μ_Y : $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

$\bar{X}$ et $\bar{Y}$ sont nommés estimateurs des moyennes.

Problème de biais pour les variances: → carré de la distance moyenne

σ_x

Sur population entière, on connaît la vraie valeur de la moyenne:



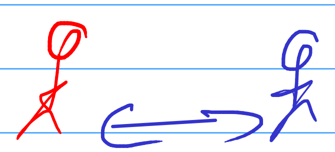
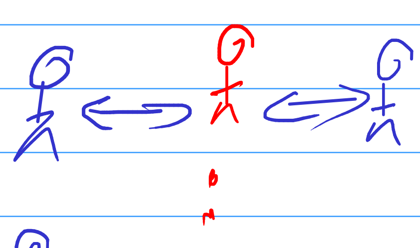
n personnes $\Longleftrightarrow$ n écarts $\Longleftrightarrow$ on divise par n.

$$Var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\sigma_x = \sqrt{Var(x)}.$$

Estimation sur échantillon: (Vu rigoureusement au S5)

Intuitivement, estimer la moyenne sur l'échantillon revient à choisir un individu comme individu moyen  (Ce sera vu par le calcul au S5)

effectif échantillon		nombre d'écarts	Estimateur de la variance
2		1	$1 \sum_{i=1}^{2} (X_i - \bar{X})^2$
3		2	$\frac{1}{2} \sum_{i=1}^{3} (X_i - \bar{X})^2$
n		n-1	$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Sur population entière

$$\sigma^2 = \mathrm{Var}\,x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\sigma = \sqrt{\mathrm{Var}\,x}$$

Sur échantillon :

$$S^2 = \mathrm{Var}\,X = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Si on divise par n dans l'estimateur de la variance, on obtient une valeur à peine plus faible. C'est un estimateur biaisé.

Ce problème d'estimation est-il gênant dans ce qu'on vient de faire ? NON

$$a = \frac{Cov(x,y)}{Var(x)} = \frac{\cancel{\frac{1}{n}} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\cancel{\frac{1}{n}} \sum_i (x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)}$$

avec estimateurs non biaisés !

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

pas de problème de biais !

$$r_{xy} = \frac{\cancel{\frac{1}{n}} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\cancel{\frac{1}{n}} \sum_i (x_i - \bar{x})^2}\sqrt{\cancel{\frac{1}{n}} \sum_i (y_i - \bar{y})^2}}$$

8 Point de vue inférentiel

On peut supposer que x et y sont les observations d'un échantillon des variables X et Y. On écrit donc le modèle $Y = aX + b + \epsilon$ ou, avec des notations plus fréquemment utilisées,

$$Y = \beta_1 X + \beta_0 + \epsilon$$

erreurs.

sur population totale (pas accès)

$Y_i = \beta_1 X_i + \beta_0 + \epsilon_i$

β_1 est estimé par $\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$

On pose

$\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0$

$\hat{\imath}$ prédite par la droite.

β_0 est estimé par $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

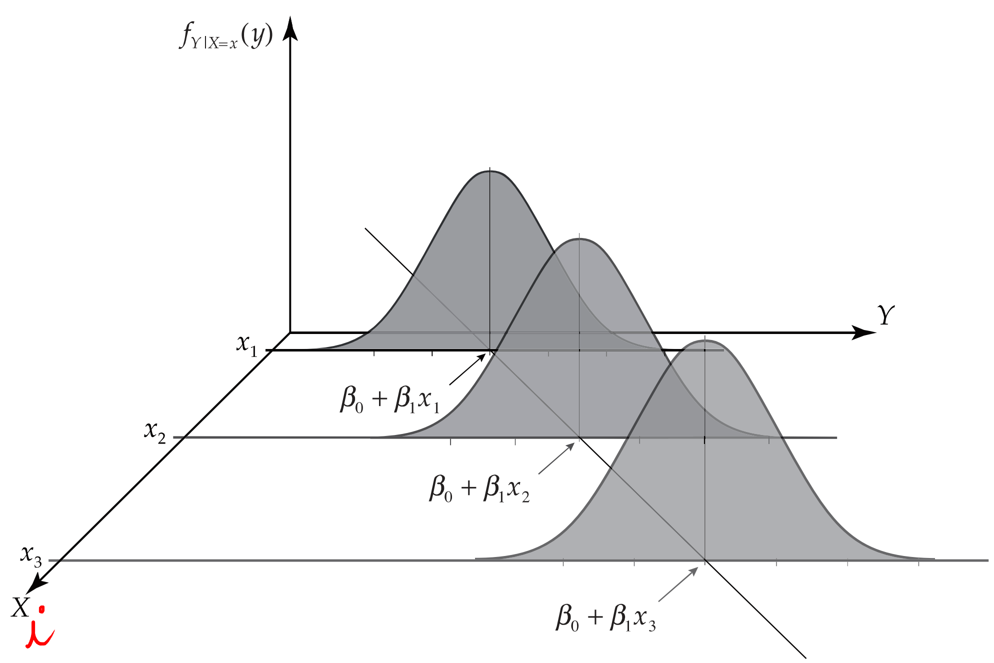
$\hat{\beta}_1$ et $\hat{\beta}_0$ sont les estimateurs des moindres carrés de β_1 et β_0.

Les valeurs β_1 et β_0 calculées ci-dessus sont en réalité les estimations $\hat{\beta}_1$ et $\hat{\beta}_0$ par la méthode des moindres carrés, *i.e.* minimisant la somme des carrés des écarts (par rapport à la droite)

$$SCE = \sum_i \varepsilon_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

8.1 Hypothèses sur les termes d'erreur ε

- Indépendance des erreurs : les $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ sont indépendants.
- Éxogénéité : les variables explicatives $(X_1, \cdots X_n)$ ne sont pas corrélées au terme d'erreur. De plus, les erreurs sont centrées $E(\varepsilon_i) = 0$
- Homoscédasticité : les termes d'erreurs sont supposés de variance constante.
- Normalité des termes d'erreur : les termes d'erreurs suivent une loi normale, centrées, de variance σ^2



$$Y_i = \overbrace{\beta_1 X_i}^{\text{à i fixé, c'est constant}} + \beta_0 + \varepsilon_i \rightarrow \mathcal{N}(0, \sigma)$$

Donc $Y_i \sim \mathcal{N}(\underbrace{\beta_1 X_i + \beta_0}_{\text{modèle}} ; \sigma)$ indépendante

écart-type de l'erreur.

indépendance entre ε_i et X_j.

8.2 Équation de la variance

$\beta_1 X_i + \beta_0$

D'après les hypothèses précédentes, il vient que

Équation de la variance

$$Var(Y) = Var(\hat{Y}) + Var(\varepsilon),$$

$\beta_1 X + \beta_0 + \varepsilon$ ε_i

grâce à l'éxogénéité. Il suit

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y} - \bar{Y})^2 + \sum_i (Y_i - \hat{Y})^2.$$

Rappels

$(X_i, Y_i)_{1 \le i \le N}$ observations sur échantillon. Modèle linéaire :

$$Y_i = \underbrace{\beta_1 X_i + \beta_0}_{\hat{Y}_i} + \varepsilon_i$$

← erreur aléatoire $\mathcal{N}(0, \sigma^2)$ ← constantes

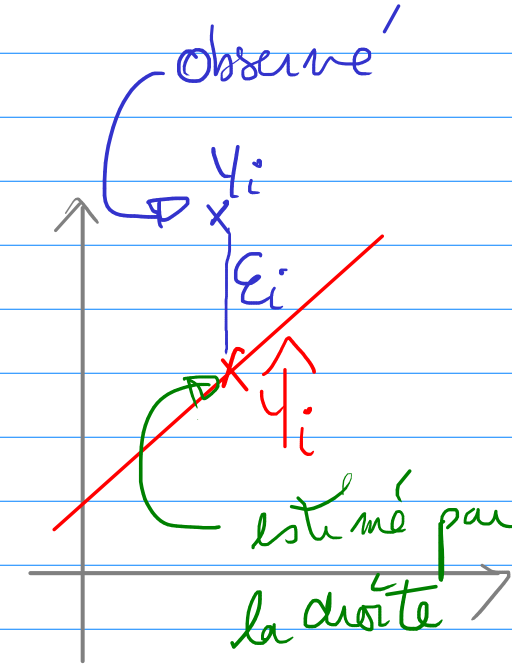
i.i.d

Hypothèses du modèle : les X_i ne portent pas d'erreur, ils ne sont pas aléatoires

$$\hat{Y}_i = \beta_1 X_i + \beta_0$$

ε_i

$$\Longrightarrow Y_i \sim \mathcal{N}(\hat{Y}_i, \sigma^2)$$ indépendantes.



somme des carrés totale → $SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

Somme des carrés due au modèle $SCM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

observé estimé par la droite

somme des carrés due à l'erreur → $SCE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

8.2 Équation de la variance

D'après les hypothèses précédentes, il vient que

$$Var(Y) = Var(\hat{Y}) + Var(\varepsilon),$$

grâce à l'éxogénéité. Il suit

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y} - \bar{Y})^2 + \sum_i (Y_i - \hat{Y})^2.$$

SCT = SCM + SCE

Notons que $SCM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (\beta_1 X_i + \cancel{\beta_0} - (\beta_1 \bar{X} + \cancel{\beta_0}))^2$

$$= \sum_{i=1}^{n} (\beta_1 (X_i - \bar{X}))^2 = \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$Var\,\hat{Y} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{N-1} = \frac{SCM}{N-1}$$

$$Var\,Y = \frac{\sum_i (Y_i - \bar{Y})^2}{N-1} = \frac{SCT}{N-1}$$

8.3 Coefficient de détermination

Le coefficient de détermination est le rapport de variance de Y expliquée par la régression :

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = \frac{SCM}{SCT}$$

Le coefficient R^2 est donc la proportion de variance de Y expliquée par le modèle.

$$r_{xy}^2 = \frac{Cov(x,y)^2}{Var(x)\,Var(y)} = \frac{\boxed{\frac{Cov^2(x,y)}{Var^2(x)}}\,\overset{\beta_1^2}{} Var(x)}{Var\,y} = \frac{\beta_1^2\,Var(x)}{Var(y)} = \frac{\beta_1^2 \frac{\sum_i (x_i - \bar{x})^2}{\cancel{N-1}}}{\frac{\sum_i (y_i - \bar{y})^2}{\cancel{N-1}}} = \frac{SCM}{SCT} = R^2$$

Estimateurs des moindres carrés de β_1 et β_0 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Si les X_i ne sont pas aléatoires et seuls Y_i et $\bar{Y}$ le sont, alors $\hat{\beta}_1$ est une combinaison linéaire des Y_i et de $\bar{Y}$ (elle même combinaison linéaire des Y_i).

On a vu que les Y_i sont des normales indépendantes

$$\Rightarrow \hat{\beta}_1 \sim \mathcal{N}(?, ?)$$

Theorem 8.1 *Sous les hypothèses du modèle de régression linéaire simple, $\hat{\beta}_1$ suit une loi normale d'espérance β_1 et de variance*

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1 ; \sqrt{\frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}\right).$$

$$\sqrt{\sigma^2_{\hat{\beta}_1}} = \sigma_{\hat{\beta}_1}$$

On fait les remarques suivantes.

- Ce théorème montre que $\hat{\beta}_1$ est sans biais pour β_1. Cela signifie que d'un échantillon à l'autre, la valeur de $\hat{\beta}_1$ oscille autour de la valeur théorique β_1.

- Ces écarts par rapport à la moyenne β_1 sont distribués selon une loi normale dont la variance est
$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\varepsilon}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$
On note, donc, que la variance de $\hat{\beta}_1$ croît avec σ^2_ε, mais qu'elle décroît lorsque $\sum_{i=1}^n (X_i - \bar{X})^2$ croît. Ainsi, plus les X_i sont nombreux et dispersés, plus notre estimation sera fiable.

- On acceptera que la distribution de $\hat{\beta}_0$ est normale et suit la loi
$$\mathcal{N}\left(\beta_0, \sigma^2_\varepsilon \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]\right).$$
(Moins indispensable).

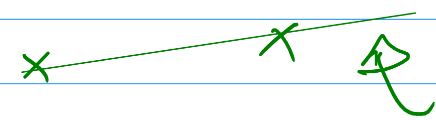
On va avoir besoin d'estimer σ_ε^2, variance des erreurs.

$$\bar{\varepsilon} = 0 \Rightarrow Var(\varepsilon_i) = \frac{\sum_i (\varepsilon_i - 0)^2}{L} = \frac{\sum_i \varepsilon_i^2}{L} = \frac{SCE}{L}$$

L : N ?, $N-1$?, $N-2$?

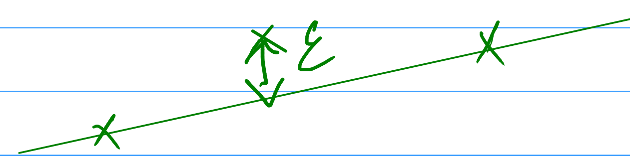
nombre d'écart observés pour N données.

Pour $N = 2$



Par 2 points, une droite, 0 erreurs.

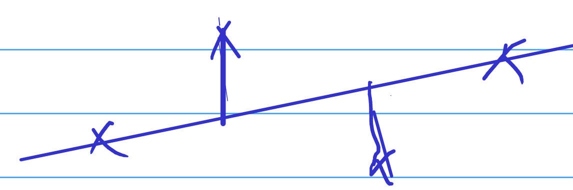
Pour $N = 3$



On suppose que la droite passe par 2 points, il y a 1 erreur

$L = 3 - 2 = 1$

Pour $N = 4$



2 erreurs : $L = 4 - 2 = 2$ ...

$Var(\varepsilon_i)$ ~~est~~ estimée par $\frac{SCE}{N-2}$

$\sum_i \varepsilon_i \longrightarrow \dfrac{SCE/_{n-2}}{\sigma^2}$ → somme de $n-2$ écarts au carré indépendants de lois normales

réduites par le carré de leur écart-type

donc $\quad S_\varepsilon^2 = \dfrac{SCE}{\sigma^2} \sim \chi^2(n-2)$.

On avait $\quad \dfrac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim \mathcal{N}(0,1)$

$$\|$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}}} \sim \mathcal{N}(0,1)$$

Si $\quad \mathcal{N}(0,1) \quad \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma_\varepsilon^2}{\sum_i (x_i-\bar{x})^2}} \times \sqrt{\dfrac{SCE/_{n-2}}{\sigma_\varepsilon^2}}} \rightarrow \sqrt{\dfrac{\chi_2(n-2)}{n-2}}$

On en déduit que $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{SCE}{(n-2)\sum_i (x_i - \bar{x})^2}}} \sim t_{n-2}$

estimation de $\hat{\sigma}^2_{\hat{\beta}_1}$

Theorem 8.3 *En estimant σ_ε par s_ε, on obtient les distributions des estimateurs suivantes*

$$(\hat{\beta}_0 - \beta_0)/s_{\hat{\beta}_0} \backsim t_{n-2}, \qquad (\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1} \backsim t_{n-2},$$

avec

$$s^2_{\hat{\beta}_1} = \frac{SCE}{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

et

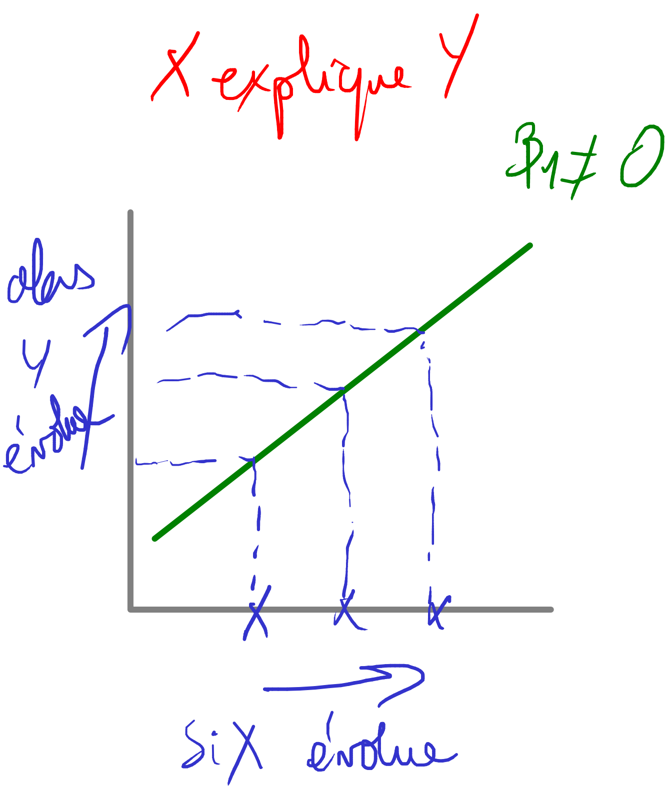
$$s^2_{\hat{\beta}_0} = \frac{SCE}{(n-2)}\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

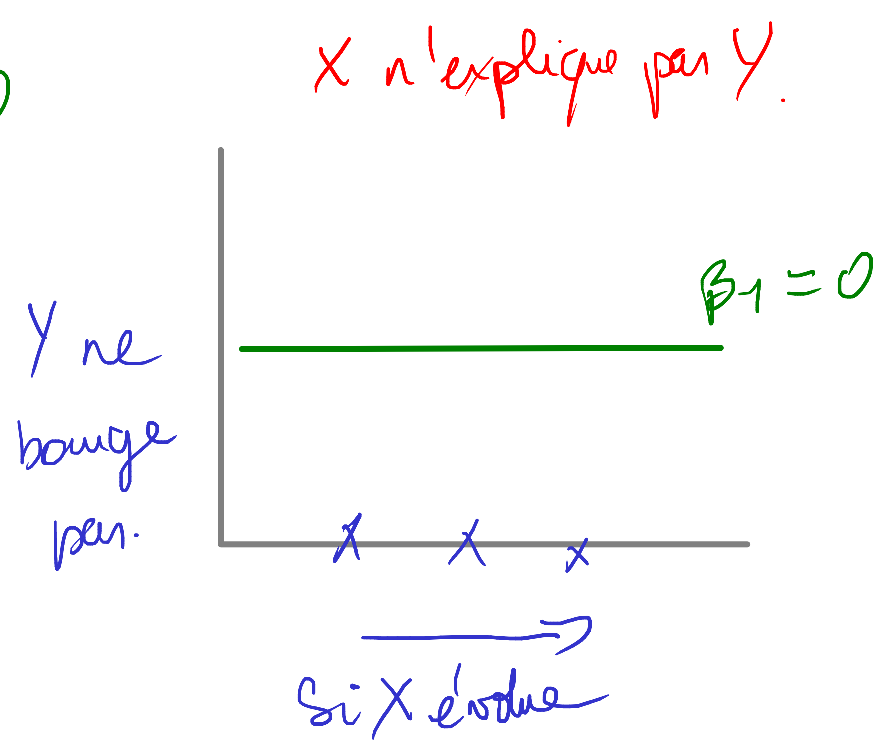
Si le nombre de degrés de liberté est assez élevé (plus de trente), on peut faire une approximation de la loi t de Student par une loi $\mathcal{N}(0,1)$.

9 Tests sur la pente de la droite

Pour faire simple, les tests F de Fisher et t de Student testent l'hypothèse $\mathcal{H}_0$ sous laquelle le coefficient β_1 est nul, contre β_1 est non nul (ce qui permet d'affirmer que X explique Y, au moins en partie).

Comprendre : Si X évolue, Y aussi (ne pas y voir un nécessaire lien de cause à effet direct).





X explique $Y \iff$ "$\beta_1 \neq 0$".

Pour les test à venir (Student / Fisher)

$\mathcal{H}_0 =$ "$\beta_1 = 0$".

9.1 Test de Student

Notons l'hypthèse nulle

$$\mathcal{H}_0 = \text{“}\beta_1 = 0\text{”},$$

autrement formulée, $\mathcal{H}_0$ est équivalente à "X n'explique pas Y". L'estimation de β_1 dans le théorème 8.3 montre que

β_1 nul sous $\mathcal{H}_0$.

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \backsim t_{n-2},$$

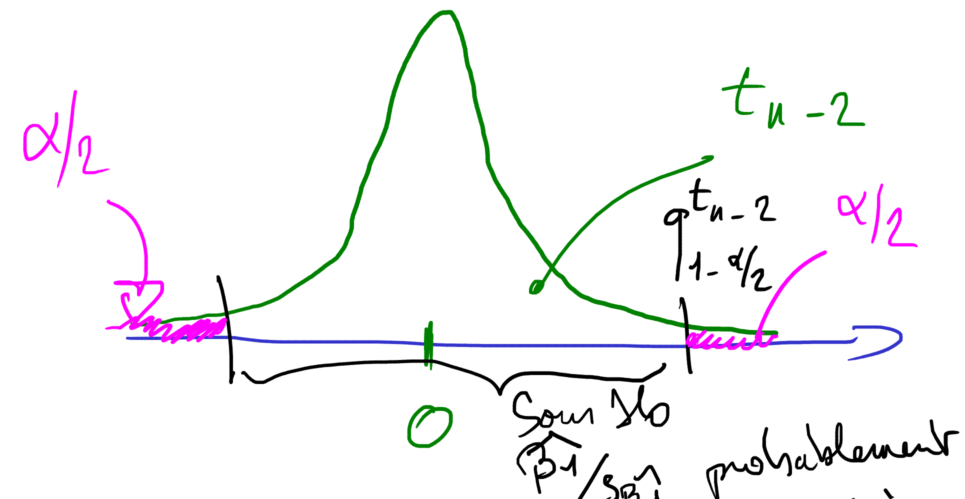
$$\sqrt{\frac{SCE}{(n-2)\sum_i (X_i - \bar{X})^2}}$$

et sous $\mathcal{H}_0$, nous garderons

$$\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \backsim t_{n-2}.$$

Rappelons l'estimation

$$s_{\hat{\beta}_1} = \sqrt{\frac{SCE}{(n-2)\sum(X_i - \bar{X})^2}}.$$



Nous ferons donc un test (bilatéral) de Student sur la statistique de test $\hat{\beta}_1/s_{\beta_1}$. On rejettera $\mathcal{H}_0$ au seuil α si

$$\frac{|\hat{\beta}_1|}{s_{\hat{\beta}_1}} \geq q_{1-\alpha/2}^{t_{n-2}}.$$

Pb : le test de Student est bilatéral et la statistique du test $\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$

n'est pas très belle ...

Rappel : $t_{n-2}^2 \sim \mathcal{F}_{1,n-2}$

9.2 Table d'ANOVA

L'analyse de la variance, souvent présentée sous forme d'un tableau, permet d'éclairer sur l'influence de la variable X sur la variable Y grâce à l'étude de la décomposition de la variance (8.1). Notons, encore, l'hypthèse nulle

$$\mathcal{H}_0 = \text{“}\beta_1 = 0\text{”}.$$

On a vu que c'est SCM

Prenons le carré de la variable du test précédent :

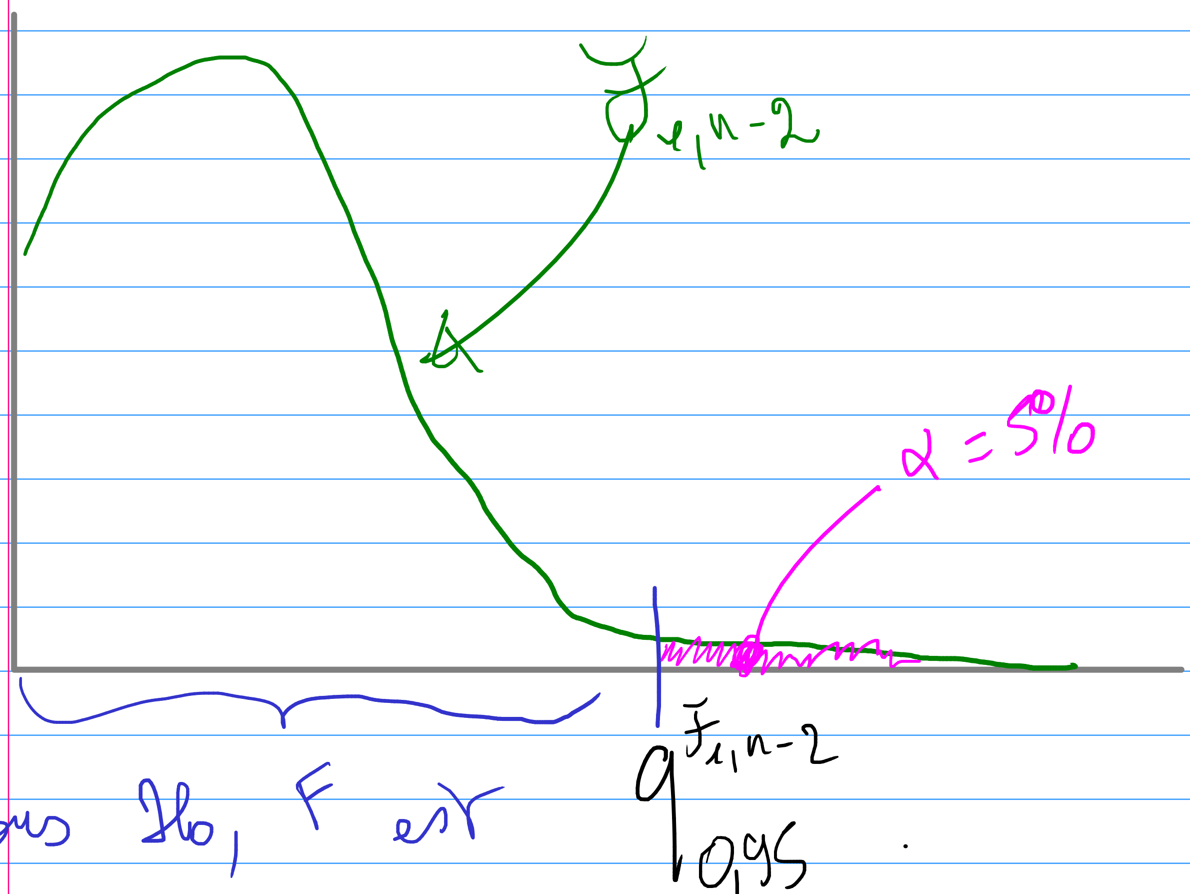
$$\frac{\hat{\beta}_1^2}{S^2_{\hat{\beta}_1}} = \frac{\hat{\beta}_1^2}{\left(\frac{SCE}{(n-2)\sum_i (X_i-\bar{X})^2}\right)} = \frac{\hat{\beta}_1^2 \sum_i (X_i-\bar{X})^2}{SCE/(n-2)} \sim (t_{n-2})^2$$

Donc $\dfrac{\hat{\beta}_1^2}{S^2_{\hat{\beta}_1}} = \dfrac{SCM}{SCE/(n-2)} \sim \mathcal{F}_{1,n-2}$

La variable est assez simple et un lien avec les sommes de carré définies plus tôt.

De plus, c'est un test unilatéral

$$F = \frac{SCM}{SCE/(n-2)}$$



Sous H_0, F est probablement ici.

Si $F \geq q^{\mathcal{F}_{1,n-2}}_{0,95}$, alors au seuil 5%, on rejette H_0. On en déduit que X explique Y.

Ainsi, si la p-value dans la table d'ANOVA est proche de zéro (ou en dessous du seuil fixé), on rejettera la nullité de $\hat{\beta}_1$.

Dans ce cadre, on voit qu'on peut utiliser le test de t de Student pour le rapport $\hat{\beta}_1/s_{\hat{\beta}_1}$ ou F de Fisher pour le carré de ce rapport, sans distinction. Il est totalement équivalent en cas de régression simple (ce n'est pas le cas sur une régression multiple). On note d'ailleurs que la statistique F est le carré de la statistique t.

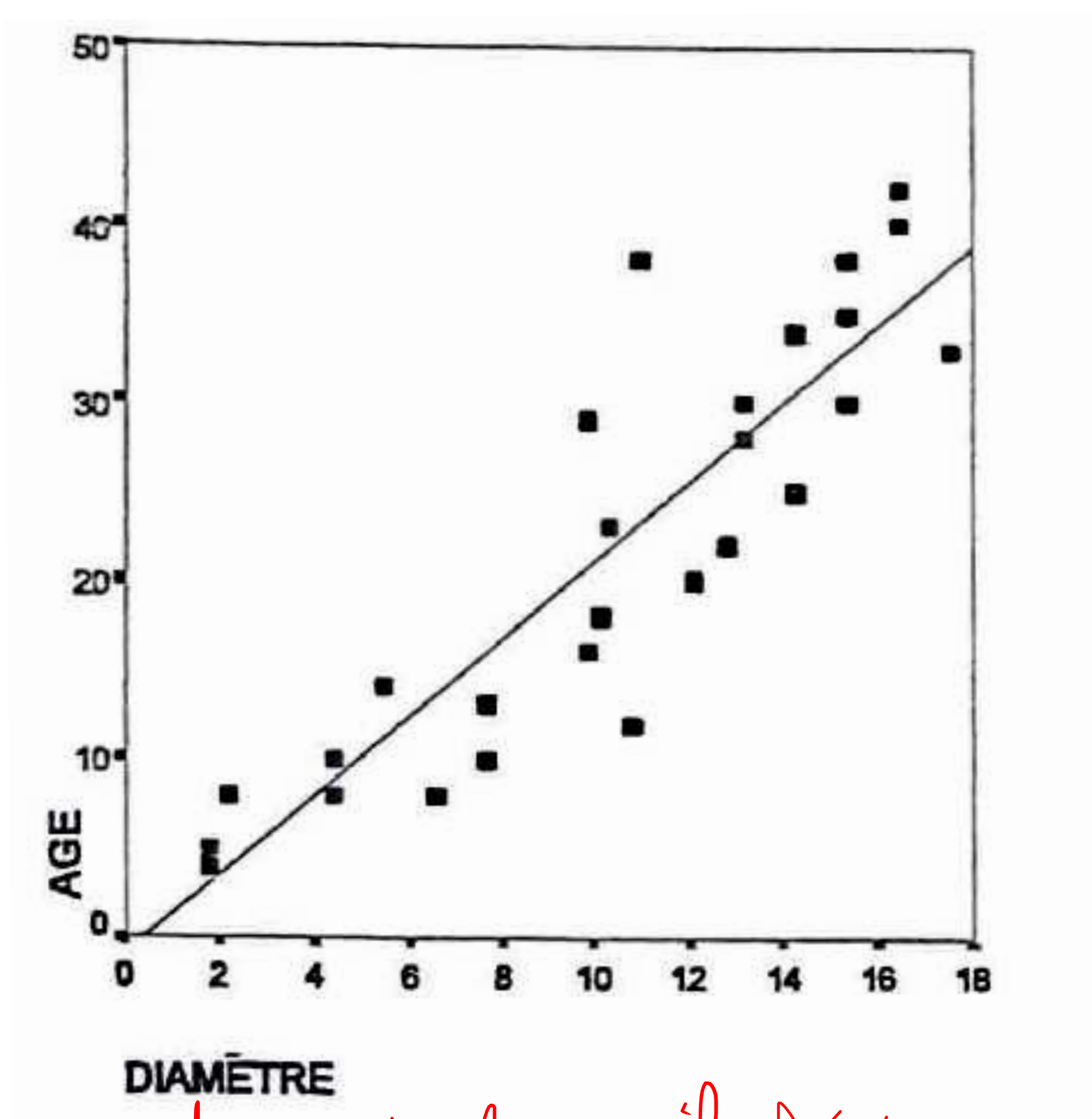
Dans le cadre d'une régression multiple (sur plusieurs variables explicatives), le test de Fischer teste l'effet global des variables sur la variable Y, les tests de Student testent l'effet de chaque variable explicative sur Y.

10 Exemple : le cas de la régression linéaire simple

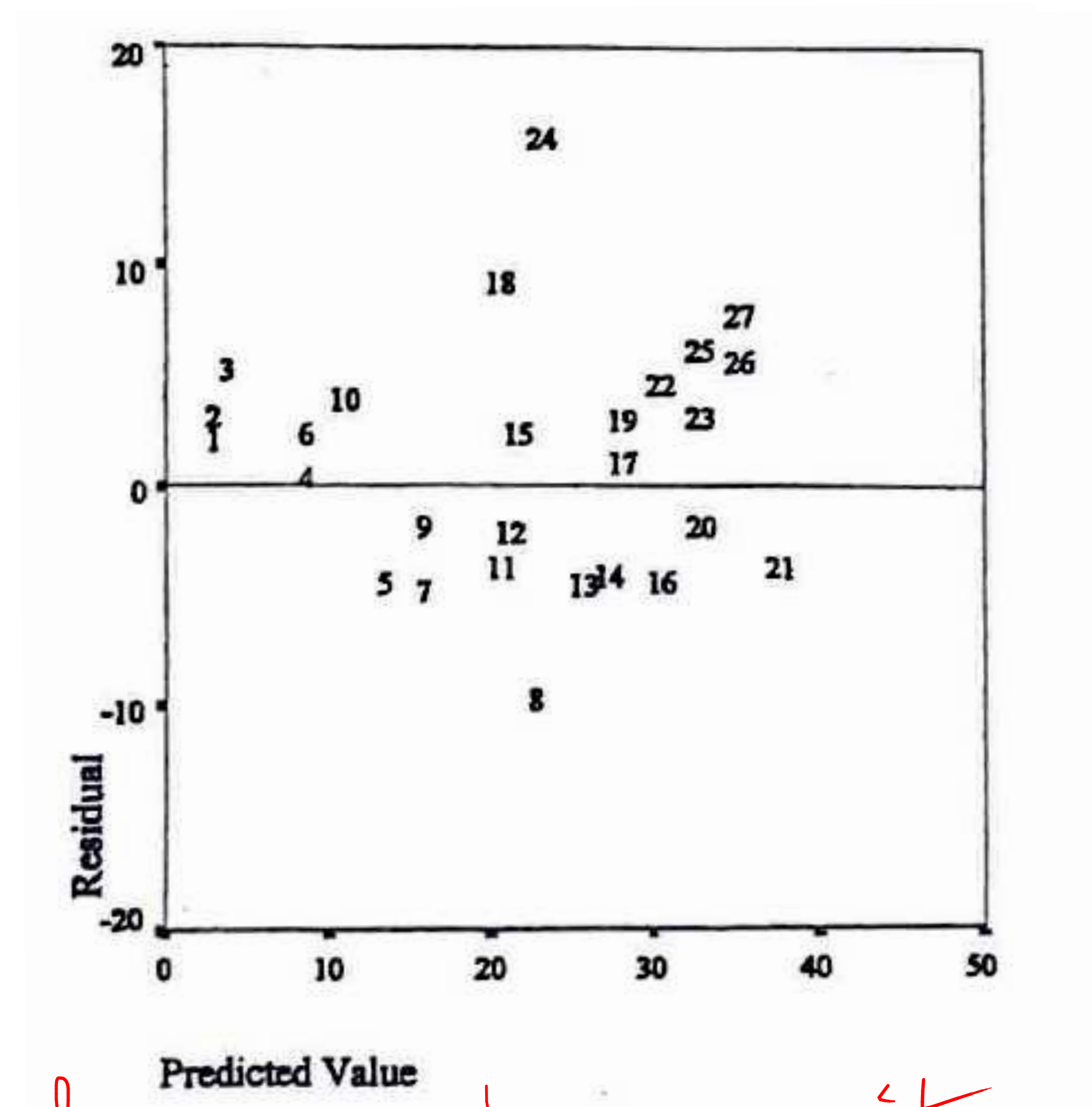
On sait que l'on mesure l'âge d'un arbre en comptant les anneaux sur une section transversale du tronc, mais cela nécessite de l'avoir abattu auparavant. Peut-on connaître l'âge à partir de la mesure de sa circonférence ?

Afin de répondre à cette question, on a effectué les mesures sur un échantillon de 27 arbres de la même espèce. À partir de ces données, on a effectué une régression de l'âge en fonction du diamètre . Les résultats ont été traités à l'aide du logiciel SPSS.

Nuage de points



Résidus



Rappel: p-value est le seuil théorique auquel on pouvait encore rejeter vu la valeur de la variable du test.

→ si $p<0{,}05$, on rejette H_0.

Analyse de variance

Test de Fisher.

P value proche de 0 - On rejette H0. ⇒ diamètre explique Age significativement

	Somme des carrés	ddl	Carré moyen	F	Signification
Régression	2905,549	1	2905,55	93,44	,000
Résidu	777,414	25	31,097		
Total	3682,963	26			

SCM/SCE/(n-2)

$9,67^2 \simeq 93$

$F \simeq t^2$

Résumé de la régression

test de Student.

	Coefficients non standardisés		Coefficients standardisés		
	B	Erreur standard	Bêta	t	Signification
(constante)	-,974	2,604		-,374	,711
DIAMETRE	2,206	,228	,888	9,67	,000

Régression

Age = 2,206 diamètre − 0,974

Dans le cas de la régression linéaire simple, le test de Student et de Fisher sont totalement équivalents. Par conséquent, il n'est pas surprenant d'observer que $\sqrt{F} = t_{diametre}$.

Dans notre exemple, nous avons, tant pour F que pour t, des $p-$values inférieures à 5%. Nous en concluons que le diamètre explique significativement l'âge des arbres.