STATISTIQUE APPLIQUEE

Chap. III : Régression linéaire simple

Rachid MCHICH

I. Modèle de régression linéaire simple

Le modèle de régression utilisé dans une régression linéaire simple s'écrit :

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y : variable à expliquer (dépendante)

x : variable explicative (indépendante)

β_0 *et* β_1 correspondent aux paramètres du modèle

ε est une variable aléatoire appelée: terme d'erreur.

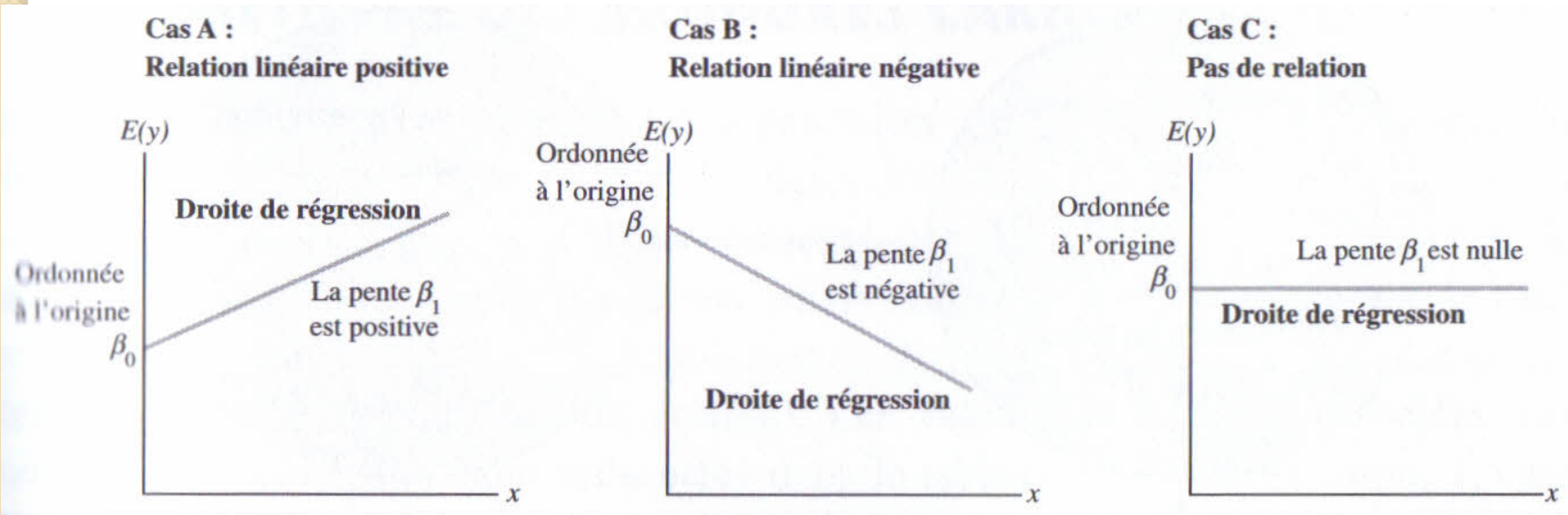
Ce terme prend en compte la variabilité de *y* qui n'est pas expliquée par la relation linéaire entre *x* et *y*.

Le terme ε mesure la différence entre les valeurs réellement observées et les valeurs qui auraient été observées si la relation spécifiée avait été rigoureusement exacte. Ce terme regroupe donc trois erreurs :

- Une erreur de **spécification** : le fait que la seule variable explicative n'est pas suffisante pour rendre compte de la totalité du phénomène expliqué;
- une erreur de **mesure**, les données ne représentent pas exactement le phénomène;
- une erreur de **fluctuation d'échantillonnage** : d'un échantillon à l'autre les observations, et donc les estimations, sont légèrement différentes.

Chaque distribution des valeurs de y a sa propre moyenne. L'équation qui décrit comment la moyenne de y est liée à x est appelée : **équation de la régression linéaire simple**. Elle est donnée par :

$$E(y) = \beta_0 + \beta_1 x$$



En pratique, la valeur des paramètres n'est pas connue et doit être estimée en utilisant les données d'un échantillon. D'où l'**équation estimée de la régression** linéaire simple :

$$\hat{y} = b_0 + b_1 x$$

(droite de régression estimée)

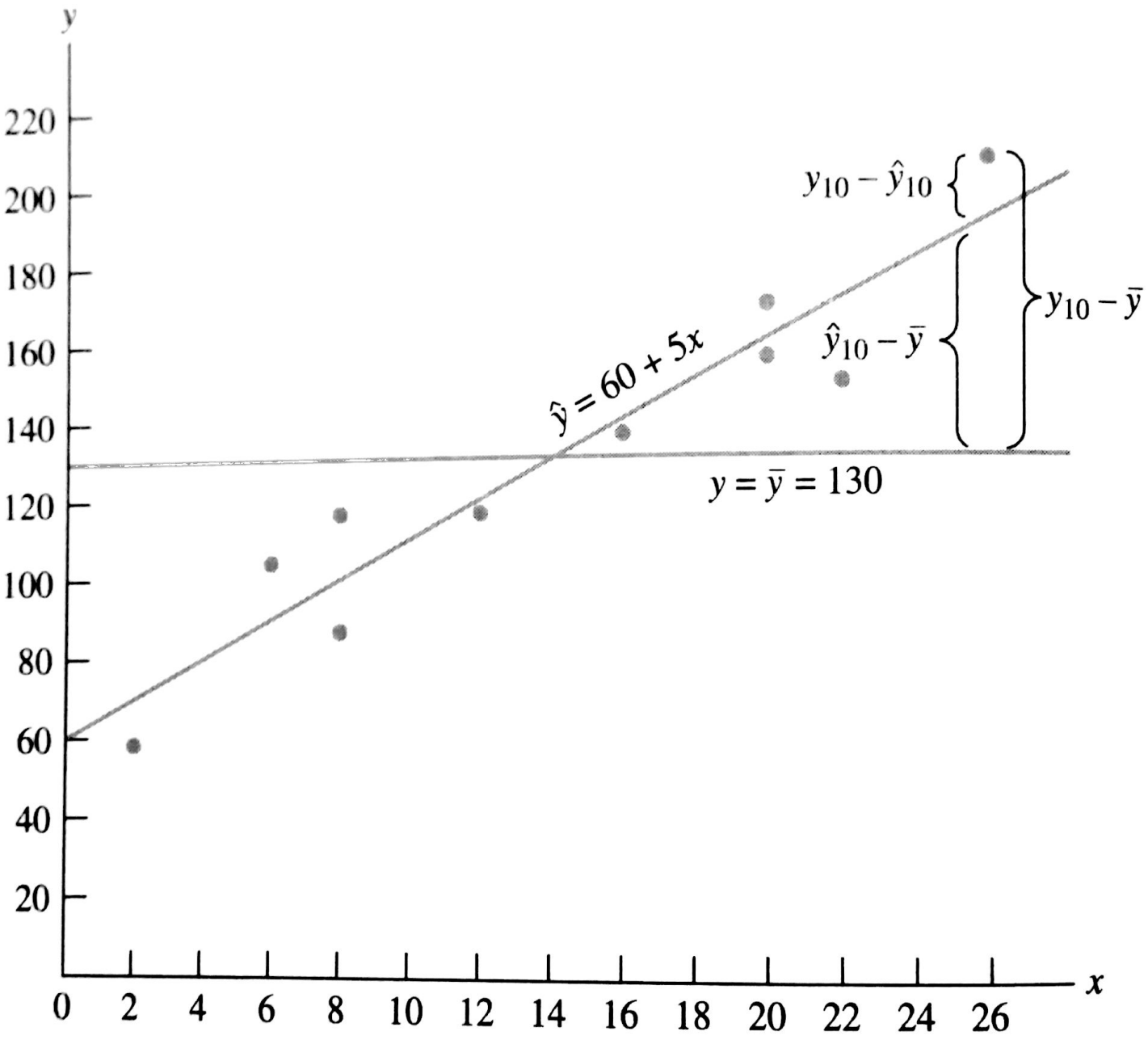
II. La méthode des moindres carrés :

La méthode des moindres carrés est une procédure qui permet d'utiliser les données de l'échantillon pour estimer l'équation de la régression (b_0 et b_1). Elle consiste à *minimiser la somme des écarts au carré* :

$$\min \sum (y_i - \hat{y}_i)^2$$

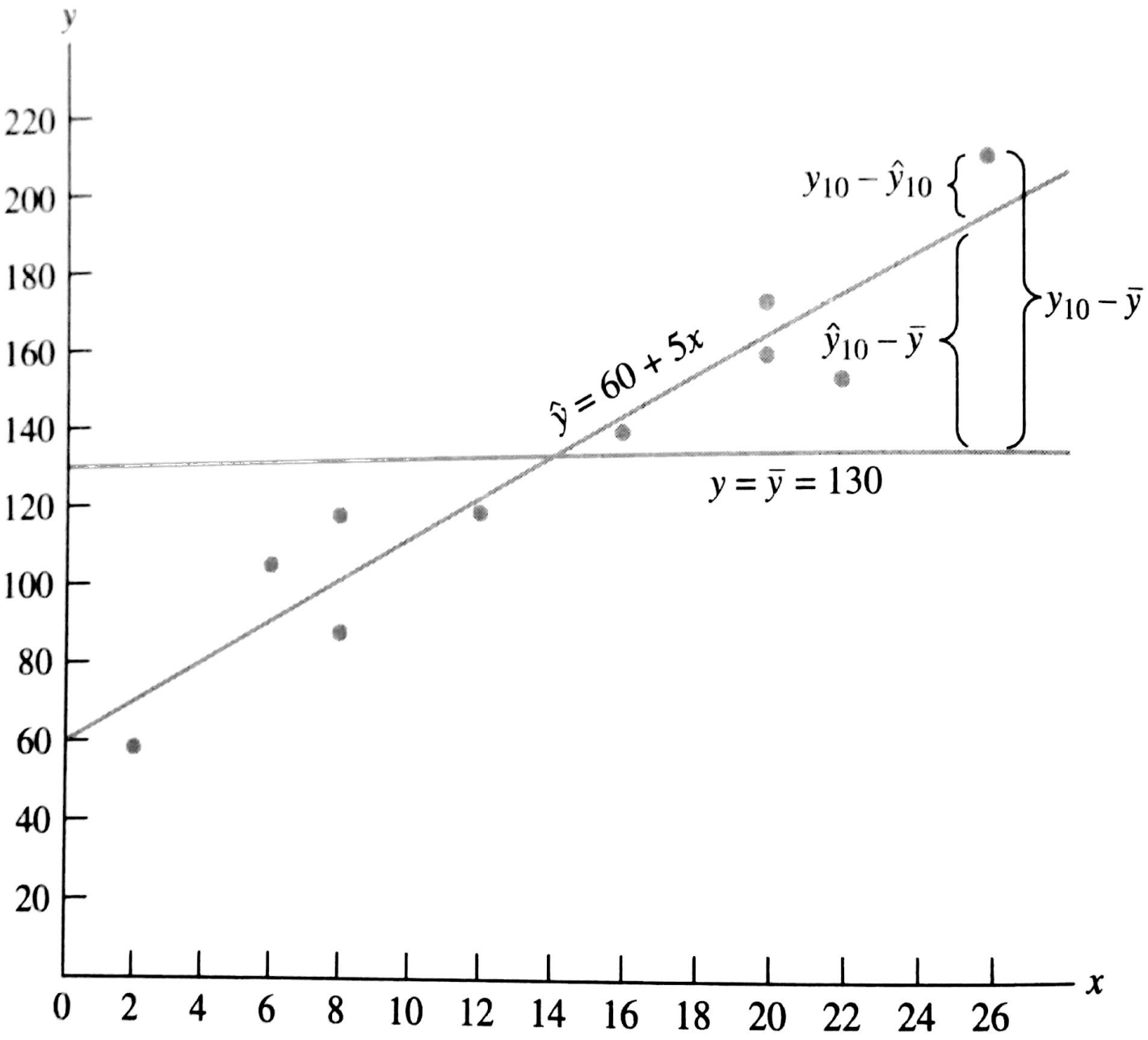
Ainsi, la pente et l'ordonnée à l'origine de l'équation estimée de la régression sont données par :

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$


y
220
200
180
160
140
120
100
80
60
40
20
0
2
4
6
8
10
12
14
16
18
20
22
24
26
x
$y_{10} - \hat{y}_{10}$
$\hat{y}_{10} - \bar{y}$
$y_{10} - \bar{y}$
$\hat{y} = 60 + 5x$
$y = \bar{y} = 130$

Exemple : Considérons les données collectées sur les ventes mensuelles d'un échantillon de 10 restaurants d'une chaîne de restaurants, par-rapport à la population locale :

Restaurant i	Population (en milliers : x_i)	Ventes mensuelles (en milliers de dh)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202


y
220
200
180
160
140
120
100
80
60
40
20
0
2
4
6
8
10
12
14
16
18
20
22
24
26
x
$y_{10} - \hat{y}_{10}$
$\hat{y}_{10} - \bar{y}$
$y_{10} - \bar{y}$
$\hat{y} = 60 + 5x$
$y = \bar{y} = 130$

III. Coefficient de détermination :

Une fois l'équation estimée de la régression établie, la question qui se pose est : Dans quelle mesure cette équation s'ajuste-t-elle aux données?

Nous montrerons ainsi que le **coefficient de détermination** fournit une mesure de l'adéquation de l'équation estimée de la régression.

Pour la *i*ème observation, le ième résidu (ou erreur commise) est donné par :

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

La somme de ces résidus, ou erreurs, au carré correspond à la quantité minimisée par la méthode des moindres carrés. Cette quantité, aussi appelée : **somme des carrés des résidus**, est donnée par :

$$SCres = \sum (y_i - \hat{y}_i)^2$$

D'autre part, pour estimer les y_i sans utiliser les x_i, on utilise $\bar{y}$ la moyenne des y_i. Ainsi, pour la $i^{ème}$ observation, $y_i - \bar{y}$ fournit une mesure de l'erreur commise en utilisant $\bar{y}$ pour estimer les ventes.

D'où **la somme des carrés totale** donnée par:

$$SCT = \sum (y_i - \bar{y})^2$$

Enfin, pour déterminer dans quelle mesure les valeurs $\hat{y}$ de la droite de régression dévient de la moyenne $\bar{y}$, une autre somme des carrés est calculée.

Cette somme est appelée **somme des carrés de la régression**, et elle est donnée par :

$$SCreg = \sum (\hat{y}_i - \bar{y})^2$$

Ainsi, on a :

$$SCT = SCreg + SCres$$

D’autre part, le **coefficient de corrélation** de l’échantillon peut être calculé de la façon suivante :

$$r_{xy} = (signe\ de\ b_1)\sqrt{r^2}$$

IV. Hypothèses du modèle :

Même avec une valeur du coefficient de détermination assez élevé, une analyse approfondie de la robustesse du modèle supposé doit être faite.

Pour cela, des **tests de significations** doivent être effectués et sont basés sur les hypothèses suivantes, concernant le terme d'erreur ε :

Hypothèses sur le terme d'erreur ε dans le modèle de la régression : $y = \beta_0 + \beta_1 x + \varepsilon$

(H1) $E(\varepsilon) = 0$ (donc $E(y) = \beta_0 + \beta_1 x$)

(H2) La variance de ε notée σ^2 est la même pour toutes les valeurs de *x*.

(H3) Les valeurs de ε sont indépendantes entre elles.

(H4) Le terme d'erreur ε est une v. a. normalement distribuée (et donc *y* aussi).

V. Test de signification :

Pour l'équation de régression simple, on a :

$$E(y) = \beta_0 + \beta_1 x$$

Ainsi, si $\beta_1 = 0$ alors $E(y) = \beta_0$; càd x et y ne sont pas liées; sinon, les deux variables le sont (si $\beta_1 \neq 0$).

Il faudrait donc effectuer un test d'hypothèses pour déterminer si $\beta_1 = 0$.

Estimation de σ^2 :

- La **moyenne des carrés des résidus** fournit une estimation de σ^2 :

$$s^2 = MCres = \frac{SCres}{n-2}$$

((n-2) ddl)

MCres fournit une estimation sans biais de σ^2.

- Erreur type de l'estimation :

$$s = \sqrt{MCres} = \sqrt{\frac{SCres}{n-2}}$$

V-1 Test *t* de Student :

On teste les hypothèses suivantes concernant β_1 *:*

$$H_0 : \quad \beta_1 = 0$$
$$H_a : \quad \beta_1 \neq 0$$

Notons d'abord que b_0 et b_1 sont des statistiques d'échantillonnage qui ont leur propre distribution d'échantillonnage; ainsi :

- Les propriétés de la distribution d'échantillonnage pour b_1 *sont données par* :

Espérance : $$E(b_1) = \beta_1$$

Ecart type de b_1 : $$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Forme de la distribution: Normale.

- Comme σ n'est pas connue, alors on calcule :

Ecart type estimé de b_1 : $$s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- Test de signification de Student dans le cadre d'une régression linéaire simple :

$$H_0 : \quad \beta_1 = 0$$
$$H_a : \quad \beta_1 \neq 0$$

- Statistique de test : $t = \dfrac{b_1}{s_{b_1}}$

- Règle de rejet :
  - Approche par la valeur p : Rejet de H_0 si $p \leq \alpha$
  - Approche par la valeur critique : Rejet de H_0 si

$$t \leq -t_{\alpha/2} \; ou \; t \geq t_{\alpha/2}$$

où $t_{\alpha/2}$ est basé sur la distribution de Student à (n-2) ddl.

Intervalle de confiance pour β_1 :

L'intervalle de confiance pour β_1 est :

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

au coefficient de confiance $(1-\alpha)$ et à (n-2) ddl.

Au seuil de 99%, l'intervalle de confiance nous offre une alternative pour conclure le test d'hypothèses dans le cadre de notre exemple.

En effet, puisque 0, la valeur hypothétique de β_1 n'appartient pas à l'intervalle de confiance, nous pouvons alors rejeter H_0 et conclure qu'une relation statistiquement significative existe entre la taille de la population et les ventes mensuelles.

V-2 Test *F* de Fisher :

Si $\beta_1 = 0$, alors la **moyenne des carrés de la régression** fournit une autre estimation de σ^2 :

$$MCreg = \frac{SCreg}{Nbr\ ddl}$$

Pour les modèles de régression considérés ici, le nombre ddl = nbr de var. indépendantes; càd. :

$$MCreg = \frac{SCreg}{Nbr\ de\ \text{var}.\ indépendantes} = \frac{SCreg}{1}$$

Test *F* de Fisher :

Statistique du test de Fisher :

$$F = \frac{MCreg}{MCres}$$

$MCreg/MCres$ suit une loi de Fisher avec 1 ddl au numérateur et n-2 ddl au dénominateur

- Test de signification de Fisher:

$$H_0 : \quad \beta_1 = 0$$
$$H_a : \quad \beta_1 \neq 0$$

- Statistique de test : $F = \dfrac{MCreg}{MCres}$

- Règle de rejet :
  - Approche par la valeur p : Rejet de H_0 si $p \leq \alpha$
  - Approche par la valeur critique : Rejet de H_0 si

$$F \geq F_\alpha$$

où F_α est basé sur la distribution de Fisher à 1 ddl au numérateur et (n-2) ddl au dénominateur.

Tableau ANOVA :

Source de la variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	F
Régression	SCreg	1	$MCreg = \frac{SCreg}{1}$	$F = \frac{MCreg}{MCres}$
Résidu	SCres	n - 2	$MCres = \frac{SCres}{n-2}$	
Totale	SCT	n - 1		

Inférence sur β_0 :

Distribution d'échantillonnage

La distribution d'échantillonnage de l'estimateur b_0 est une distribution normale :

$$b_0 \rightsquigarrow \mathcal{N}\left(\beta_0 ; \sigma\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{1/2}\right)$$

$$z = \frac{b_0 - \beta_0}{\sigma(b_0)} \rightarrow z \rightsquigarrow \mathcal{N}(0,1)$$

Inférence sur b

Remarque : Dans le cas d'un petit échantillon, **l'écart réduit** suit une **loi de Student** :

$$t = \frac{b_0 - \beta_0}{s(b_0)} \rightsquigarrow \mathcal{T}(n-2)$$

$$s(b_0) = s\left[\frac{1}{n} + \frac{\bar{x}^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]^{1/2}$$

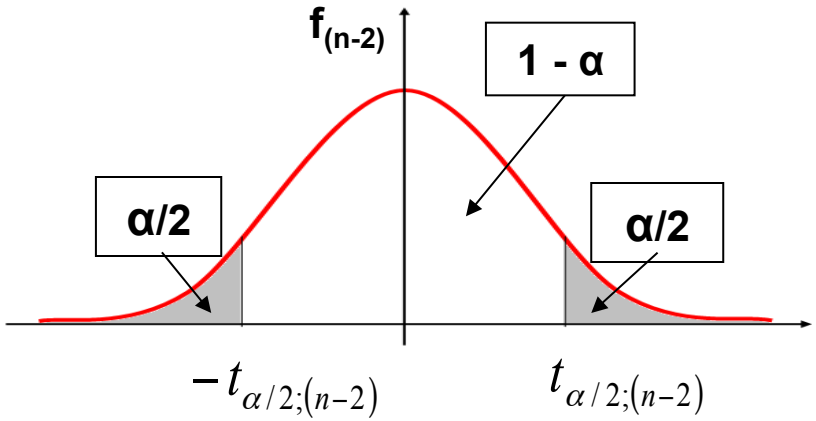
Remarque : Dans le cas où la taille de l'échantillon est grande, la distribution de l'écart réduit est

$$t = \frac{b_0 - \beta_0}{s(b_0)} \rightsquigarrow \mathcal{N}(0;1)$$

Intervalle de confiance pour β_0 :

L'intervalle de confiance pour β_0 est :

$$b_0 - s(b_0) \times t_{\alpha/2;(n-2)} \leq \beta_0 \leq b_0 - s(b_0) \times t_{\alpha/2;(n-2)}$$



Densité de probabilité de la loi Student avec (n-2) degré de liberté

Exercice :

Considérons le tableau d'observations suivant:

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Représenter le nuage de points associé à ces données.

b) Développer l'équation estimée de la régression.

c) Utiliser l'équation estimée de la régression pour prévoir la valeur de y lorsque x=4.

d) Calculer la MCres et l'erreur type de l'estimation.

e) Calculer l'écart type estimé de b_1.

f) Utiliser le test de Student et de Fisher pour tester les hypothèses suivantes :

$$H_0 : \quad \beta_1 = 0$$
$$H_a : \quad \beta_1 \neq 0$$

g) Présenter les résultats sous forme d'un tableau ANOVA.

VI. Utilisation de l'équation estimée de la régression pour estimer et prévoir :

Lorsqu'on utilise un modèle de régression linéaire simple, on fait une hypothèse sur la relation entre x et y. On détermine alors l'équation estimée de la régression linéaire simple.

Ainsi, si les résultats prouvent l'existence d'une relation statistique significative entre x et y, et si le coefficient de détermination indique que l'équation estimée de la régression semble bien adaptée aux données, l'équation estimée de la régression peut servir à faire des estimations et des prévisions.

VI.1 Estimation ponctuelle et estimation par intervalle :

Les estimations ponctuelles ne fournissent aucune information sur la précision de l'estimation. Pour cela il faut effectuer des estimations par intervalle:

- Estimation par intervalle de confiance : estimation par intervalle de la *valeur moyenne de y* pour une valeur donnée de x.

- Estimation par intervalle de prévision : utilisé lorsqu'on souhaite obtenir une estimation par intervalle *d'une seule valeur de y* correspondant à une valeur donnée de x.

Intervalle de confiance de la valeur moyenne de y :

En posant :

x_p : valeur particulière de la variable indépendante x

y_p : la valeur de y correspondant à x_p

$E(y_p)$: la moyenne ou espérance mathématique de la var. dep. y correspondant à x_p

$\hat{y}_p = b_0 + b_1 x_p$: correspond à l'estimation de $E(y_p)$ lorsque $x = x_p$

Alors, on a:

Intervalle de confiance de la valeur moyenne de y :

Variance de $\hat{y}_p$:

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

Ecart type estimé de $\hat{y}_p$:

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Intervalle de confiance de $E(y_p)$:

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

Student à n-2 ddl

Intervalle de prévision d'une valeur individuelle de y :

Estimation de l'écart type d'une valeur individuelle

$$s_{ind} = s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

Intervalle de prévision de y_p : $\hat{y}_p \pm t_{\alpha/2} s_{ind}$

Student à n-2 ddl

Estimation par intervalle de prévision de y_p :

$$\hat{y}_p \pm t_{\alpha/2} s_{ind}$$

VII. Analyse des résidus : Valider les hypothèses du modèle :

- Graphique des résidus en fonction de x

- Graphique des résidus en fonction de $\hat{y}$

- Graphique des résidus en fonction de x :

  Résidu de l'observation *i* : $y_i - \hat{y}_i$

- Graphique des résidus en fonction de $\hat{y}$:

  Résidu de l'observation *i* : $y_i - \hat{y}_i$