1.2 Le modèle linéaire simple

1.2.1 Introduction

On suppose que l'on dispose de deux suites finies d'observations (x_i) $(1 \leqslant i \leqslant n)$ et (y_i) $(1 \leqslant i \leqslant n)$, on cherche à expliquer les y_i par les x_i. Plus précisément on souhaiterait établir une relation linéaire où les variations des x_i provoquent les variations de y_i, mais d'autres facteurs, ou des erreurs, vont perturber cette relation qui ne sera qu'approchée. Au lieu de se contenter de déterminer la droite d'ajustement linéaire[3], on va modéliser l'erreur ou l'écart. On écrit alors $y_i = ax_i + b + \varepsilon_i \;\; 1 \leqslant i \leqslant n$.
où ε_i est une variable aléatoire réelle (dite erreur, résidu,....).
Comme a, b sont déterministes et les ε_i sont aléatoires, les y_i sont aussi des variables aléatoires. On écrira $y_{i,obs}$ pour l'observation numérique des y_i.
Généralement l'indice est le temps en économétrie c'est pour cela qu'on adoptera en définitive la notation suivante du modèle.

$$y_t = ax_t + b + \varepsilon_t \;\; 1 \leqslant t \leqslant n$$

Exemple 1.1. *On cherche à établir une relation entre consommation et revenu :*
$x_t = R_t$ revenu de la période t
$y_t = C_t$ consommation de la période t

R_t	85	92	99	108	116
C_t	82	88	93	102	110

Definition 1.1. *Dans le modèle $y_t = ax_t + b + \varepsilon_t \;\; 1 \leqslant t \leqslant n$.*
x_t : est la variable explicative, ou exogène (mesurée sans erreur c'est une variable certaine).

3. durant le cours de S1, le modèle étudié d'ajustement linéaire ne prévoit pas le terme d'erreur c'est juste la relation $y_i = ax_i + b$

y_t : est la variable expliquée, ou endogène. C'est une variable aléatoire.
ε_t : est la perturbation, le résidu ou l'erreur (attribuée à l'ensemble des facteurs non prise en compte).
a,b : sont des paramètres, ou des coefficients.
Toutes ces grandeurs ont des statuts différents, qu'on résume dans le tableau suivant :

	aléatoire	*non aléatoire*
observable	y_t	x_t
non observable	ε_t	*a,b*

Le but de l'étude du modèle linéaire simple est d'obtenir des informations sur la relation entre les y_t et les x_t, donc sur a et b c.a.d (estimation et tests sur a et b).

Remarque 1.1. *L'utilisation ici du modèle linéaire simple n'est pas dû à un hasard, bien au contraire, c'est quelque chose qui est imposé.En effet la modélisation mathématique la plus simple de $Y = f(X)$ est une fonction affine , toute autre formes et il en existe, quadratique, exponentielle ou logarithmique seront très difficile à modéliser.*

1.2.2 Estimation de a et b par la méthode des moindres carrés ordinaire (MCO)

On va estimer a et b (qui jouent le rôle de θ dans la théorie de l'estimation) par la méthode des MCO.
on cherche $\hat{a}, \hat{b}$ les estimateurs de a et b qui minimisent la somme des carrés des résidus.
$Q(a,b) = \sum_{t=1}^{n} \varepsilon_t^2 = \sum_{t=1}^{n} (y_t - ax_t - b)^2.$

Proposition 1. *Les estimateurs de a et b par la MOC sont données par :*

$$\hat{a} = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \quad et\ \hat{b} = \bar{y} - \hat{a}\bar{x}$$

Preuve 1.1. *Il s'agit de minimiser la fonction $Q(a,b)$, comme c'est une fonction de deux variables a et b nous devons chercher les équations normales c.a.d, les dérivées partielles par rapport à a et par rapport à b, et chercher après les points critiques. Les équations normales sont donc :*

$$\begin{cases} (1)\ \dfrac{\partial Q}{\partial a} = \displaystyle\sum_{t=1}^{n}(2(y_t - ax_t - b)(-x_t) \\ (2)\ \ = -\displaystyle\sum_{t=1}^{n}(2(y_t - ax_t - b) \end{cases}$$

On doit chercher les points critiques ceci $\Rightarrow$

$$\begin{cases} (1)\ \dfrac{\partial Q}{\partial a} = \displaystyle\sum_{t=1}^{n}(2(y_t - ax_t - b)(-x_t) = 0 \\ (2)\ \dfrac{\partial Q}{\partial b} = -\displaystyle\sum_{t=1}^{n}(2(y_t - ax_t - b) = 0 \end{cases}$$

(2) $\Rightarrow \hat{b} = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n}(y_t - \hat{a}x_t) = \bar{y} - \hat{a}\bar{x}$ *avec* $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n} y_t$ *et* $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n} x_t$.

A partir de la valeur de $\hat{b}$ *on a alors (1)* $\Rightarrow \sum_{t=1}^{n}(2(y_t - \hat{a}x_t - \hat{b})(-x_t) = 0$

$\Rightarrow \sum_{t=1}^{n} x_t(y_t - \hat{a}x_t - \bar{y} + \hat{a}\bar{x}) = 0$

$\Rightarrow \sum_{t=1}^{n} x_t(y_t - \bar{y}) = \hat{a}\sum_{t=1}^{n} x_t(x_t - \bar{x})$.

On démontre facilement que $\sum_{t=1}^{n} x_t(y_t - \bar{y}) = \sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})$ *et que*

$\sum_{t=1}^{n}(x_t(x_t-\bar{x}) = \sum_{t=1}^{n}((x_t-\bar{x})^2$.
D'où $\hat{a} = \dfrac{\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}$.

On vérifie bien qu'il s'agit d'un minimum en calculant le déterminant de la matrice Hessienne tel que $det(HessQ) = 2(n+\sum_{t=1}^{n}(x_t)^2 > 0$. □

Exemple 1.2. *Il s'agit de reprendre le tableau de l'exemple 1 sur la consommation et le revenu on a le tableau suivant :*

R_t	*85*	*92*	*99*	*108*	*116*
C_t	*82*	*88*	*93*	*102*	*110*

$$\begin{cases} \bar{x} = \bar{R} = \dfrac{1}{5}\displaystyle\sum_{t=1}^{5}(85+92+99+108+116) = 100 \\ \bar{y} = \bar{C} = \dfrac{1}{5}\displaystyle\sum_{t=1}^{5}(82+88+93+102+110) = 95 \end{cases}$$

on a aussi $\displaystyle\sum_{t=1}^{n}(x_t-\bar{x})^2 = 15^2+8^2+1^2+8^2+16^2 = 610.$

$$\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y}) = 549$$

d'où $\hat{a_{obs}} = \dfrac{\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t=1}^{n}(x_t-\bar{x})^2} = \dfrac{549}{610} = 0.9.$

$\widehat{b_{obs}} = \widehat{y_{obs}} - \widehat{a_{obs}}\bar{x} = 95 - 0.9.100 = 5.$

D'où l'équation $y_t = 0.9x_t + 5$.

1.2.3 Propriétés des estimateurs $\hat{a}$ et $\hat{b}$

On va supposer maintenant maintenant que les $(\varepsilon_t)_{1\leqslant t\leqslant n}$ forment un bruit blanc c.a.d une suite de variables aléatoires réelles telle que.
H_1 : les ε_t sont centrés c.a.d $E(\varepsilon_t) = 0 \quad \forall t$ (les erreurs sont centrées et qu'on a pas oublié un terme pertinent.)
H_2 : les ε_t sont de variance constante c.a.d $V(\varepsilon_t) = \sigma^2 \quad , \forall t$ (c'est que la variance ne varie pas en fonctions des individus).
H_3 :Les ε_t sont indépendantes (une observation n'a pas d'influence sur une autre observations) c.a.d $Cov(\varepsilon_t, \varepsilon_{t'}) = 0 \quad \forall t \neq t'$ et qu'il y a indépendance entre les erreurs et la variable x_t, $Cov(x_t, \varepsilon_t) = 0$.

Remarque 1.2. *Les précautions prises sur les hypothèses émises sur les ε_t sont très minutieuses.On prend une suites de variables aléatoires réelles indépendantes, c'est le hasard qui va intervenir pour le choix de cette suite, en plus elles sont centrées pour qu'on oublie pas une personnes importantes dans notre échantillon. La variance étant constante, c.a.d qu'en changeant l'ordre des individus on aura le même résultat de notre expérience. Avec ces hypothèses on a un très bon échantillon qui nous aidera à mieux approcher l'expérience de la réalité.En fait ce terme d'erreur et le plus important car c'est ce terme qu'on a minimisé pour trouver nos estimateurs.*

On peut alors déduire des ces hypothèses H_1, H_2 et H_3 les propriétés suivantes.

Proposition 2. *Si H_1, H_2 et H_3 sont vraies alors $\hat{a}$ et $\hat{b}$ sont des estimateurs sans biais de a et b*

Pour démontrer que les estimateurs sont sans biais rappelons tout

d'abord la définition du biais [4].

Definition 1.2. *Un estimateur T_n de θ est dit sans biais si $E(T_n) = \theta$ autrement dit $b_n(\theta) = E(T_n) - \theta = 0$*

Preuve 1.2. *On a :*
$y_t = ax_t + b + \varepsilon_t$.
$\bar{y} = a\bar{x} + b + \bar{\varepsilon}$ *avec* $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_t$ *et donc.*

$$\begin{aligned}
\hat{a} &= \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\
&= \frac{\sum_{t=1}^{n}(x_t - \bar{x})(a(x_t - \bar{x}) + \varepsilon_t - \bar{\varepsilon})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\
&= \frac{\sum_{t=1}^{n} a(x_t - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} + \frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t - \bar{\varepsilon})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\
&= a + \frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t - \bar{\varepsilon})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\
&= a + \frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \qquad (1.1)
\end{aligned}$$

On a utilisé le fait comme pour la preuve de la proposition précédente : $\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t - \bar{\varepsilon}) = \sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)$ *et donc finalement.*
$E(\hat{a}) = E(a) + E(\frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2})$.
$= a + 0$ *puisque* $E(\varepsilon_t) = 0$ *hypothèse* (H_1) *et que l'espérance de la somme est égale à la somme des espérance* $E(\sum_{t=1}^{n} x_t) = \sum_{t=1}^{n} E(x_t)$ *et que* $E(a) = a$ *(l'espérance d'une constante est égale à la constante.)*
D'où $E(\hat{a}) = a$. *et donc* $\hat{a}$ *est un estimateur sans biais de a.*

4. voir cours de S3 échantillonnage et estimation

Pour l'estimateur de b on a.
$\hat{b} = \bar{y} - a\bar{x}$ *et* $\bar{y} = a\bar{x} + b + \bar{\varepsilon}$, *donc*
$E(\hat{b}) = E(\bar{y} - \hat{a}\bar{x}) = E(a\bar{x} + b + \bar{\varepsilon} - a\bar{x}) = a\bar{x} + b + E(\bar{\varepsilon}) - a\bar{x} = b$
puisque $E(\bar{\varepsilon}) = E(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t) = \frac{1}{n}\sum_{t=1}^{n}E(\varepsilon_t) = 0$ *(d'après l'hypothèse* H_1*).*
D'où $E(\hat{b}) = b$, *et donc l'estimateur de b est sans biais.*

□

Maintenant il faut chercher les variances des estimateurs, mais avant cela, rappelons tout d'abord quelques propriétés qui vont nous être utile dans la preuve de la proposition qu'on énoncera dans la suite de ce cours.

propriétés de la variance

$Var(a) = 0$ la variance d'une constante est nulle.
$Var(X + Y) = Var(X) + Var(Y)$ si X et Y sont deux v.a.r indépendantes.
$Var(X.Y) = Var(X).Var(Y)$ si X et Y sont deux v.a.r indépendantes
$Var(aX) = a^2 Var(X)$.

Proposition 3. $1)Var(\hat{a}) = \dfrac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2})$ *et.*
$2)Var(\hat{b}) = \sigma^2(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2})$

Preuve 1.3. *D'après l'écriture (3.1) de la preuve précédente on a :*

$$\begin{aligned}
\hat{a} &= a + \frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\
Var(\hat{a}) &= Var(a + \frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2}) \\
&= Var(a) + \frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2}) \\
&= Var(\frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2}) \ \ car\ Var(a) = 0 \\
&= \frac{1}{(\sum_{t=1}^{n}(x_t - \bar{x})^2)^2} Var(\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)) \\
&= \frac{1}{(\sum_{t=1}^{n}(x_t - \bar{x})^2)^2} Var(\varepsilon_t)(\sum_{t=1}^{n}(x_t - \bar{x})^2) \\
&= \frac{1}{(\sum_{t=1}^{n}(x_t - \bar{x})^2(\sum_{t=1}^{n}(x_t - \bar{x})^2} Var(\varepsilon_t)(\sum_{t=1}^{n}(x_t - \bar{x})^2) \\
&= \frac{Var(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\
&= \frac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}
\end{aligned}$$

Pour la démonstration de 2) de la proposition précédente on admet les relations suivantes :

1) $E(\hat{a} - a)^2) = Var(\hat{a})$.

2) $E(\bar{\varepsilon}^2) = Var(\bar{\varepsilon}) = \frac{\sigma^2}{n}$.

3) $E(\hat{a} - a)\bar{\varepsilon} = E(\bar{\varepsilon}.(\frac{\sum_{t=1}^{n}(x_t - \bar{x})(\varepsilon_t)}{\sum_{t=1}^{n}(x_t - \bar{x})^2}) = \frac{\sum_{t=1}^{n}(x_t - \bar{x})E(\varepsilon_t.\bar{\varepsilon})}{\sum_{t=1}^{n}(x_t - \bar{x})^2})$

$= \frac{\sum_{t=1}^{n}(x_t - \bar{x})E(\varepsilon_t)E(\bar{\varepsilon})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}) = 0$ *car* $E(\varepsilon_t = 0)$

On a

$$\begin{aligned} Var(\hat{b}) &= E((\hat{b} - b)^2) \\ &= E(\bar{y} - \hat{a}\bar{x} - (\bar{y} - a\bar{x} - \bar{\varepsilon})^2 \\ &= E((a - \hat{a})\bar{x} + \bar{\varepsilon}) \\ &= E(\bar{x}^2((\hat{a} - a)^2) - 2\bar{\varepsilon}\bar{x}(\hat{a} - a) + \bar{\varepsilon}^2 \\ &= \bar{x}^2 E(\hat{a} - a)^2) - E(2\bar{\varepsilon}\bar{x}(\hat{a} - a)) + E(\bar{\varepsilon}^2) \end{aligned}$$

En utilisant 1) et 2) et puisque $E(\varepsilon_t) = 0$ *on a :*

$Var(\hat{b}) = \bar{x}^2 Var(\hat{a}) + \frac{\sigma^2}{n} = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2})$ □

Exemple 1.3. *Reprenons le même exemple sur la consommation revenu.*

On a $\sum_{t=1}^{5}(x_t - \bar{x})^2 = 610$ *et donc* $Var(\hat{a}) = \frac{\sigma^2}{610}$.

et $(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}) = \frac{1}{5} + \frac{1000}{610} = 16,593 \Leftrightarrow Var(\hat{b}) = \sigma^2.16,593.$

Introduisons maintenant l'hypothèse H_4.

(H_4) *: les* ε_t *suivent des lois normales.*

Remarque 1.3. *En regroupant les hypothèses* $(H_1), (H_2), (H_3)$ *et* (H_4) $\Leftrightarrow$ *les* ε_t *sont indépendantes et de lois normales* $(\mathcal{N}(0,1))$.

On note les ε_t *sont* **i.i.d** *et* $\hookrightarrow (\mathcal{N}(0,1)$ *(i :indépendantes, i :identiquement d :distribuées).*

Remarque 1.4. *Les hypothèses des* ε_t *concernant l'indépendance et la normalité sont considérées comme les conditions nécessaires pour commencer l'étude du modèle.Dans le cas de la non vérification il faut chercher à normaliser les erreurs en éliminant des fois ce qu'on appelle les*

points aberrants.(faire un nettoyage).
Il y a plusieurs moyens graphiques et numériques pour vérifier la normalité des erreurs comme la courbe d'Henry, le QQ-plot,ou l'histogramme, le test de Kolmogorov-Smirnov, le test de Shapiro-Wilk ou encore de test d'asymétrie et d'aplatissement.
Ce postulat de normalité ne se pose pas généralement, lorsque n est assez grand qui dépasse les centaines sauf peut être dans des cas rare.Il se pose naturellement quand $n \leqslant 30$.

Proposition 4. *Sous les hypothèse* $(H_1), (H_2), (H_3)$ *et* (H_4) *c.a.d les* ε_t *sont* **i.i.d** *et* $\hookrightarrow \mathcal{N}(0,1)$.

(i) $\hat{a} \hookrightarrow \mathcal{N}(a, \dfrac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2})$

(ii) $\hat{b} \hookrightarrow \mathcal{N}(b, \sigma^2(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2})$

(iii) $\widehat{\sigma^2} = \dfrac{1}{n-2}\sum_{t=1}^{n}(y_t - \hat{a}x_t - \hat{b})^2$ *est un estimateur sans biais de* σ^2

et $(n-2)\dfrac{\widehat{\sigma^2}}{\sigma^2} \hookrightarrow \chi^2_{n-2}$ *et* $\widehat{\sigma^2}$ *est indépendante de* $(\hat{a}, \hat{b})$.

Preuve 1.4. $(\hat{a}, \hat{b})$ *suivent des lois normales car sont des combinaisons linéaires de v.a indépendantes et normales. D'où (i) et (ii) avec les propositions 1 et 2 on admet (iii).*

On peut adopter les notations suivantes :
$SCT = \sum_{t=1}^{n}(y_t - \bar{y})^2$ *la somme des carrés totale.*
$V(Y) = \dfrac{SCT}{n}$ *la variance totale*
$SCE = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$ *la somme des carrés des expliquées .*
$V_e(Y) = \dfrac{SCE}{n}$ *la variance des expliquées*

$SCR = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$ *la somme des carrés résiduelles*
$V_r(Y) = \dfrac{SCE}{n}$ *la variance résiduelle*

Proposition 5. *Le théorème de la décomposition de la variance s'écrit* $V(Y) = V_e(Y) + V_r(Y)$ *et* $SCT = SCE + SCR$

En prenant ces notations on a, $SCR = \sum_{t=1}^{n}(y_t - \hat{a}x_t - \hat{b})^2$ *d'où* $\widehat{\sigma^2} = \dfrac{SCR}{n-2}$

1.2.4 Application aux tests et intervalles de confiance des paramètres de a et b

1.2.5 Application aux tests

On suppose que les ε_t *sont iid et* $\hookrightarrow (\mathcal{N}(0,1)$*.En général on ne connaît pas* σ*. Il faut donc utiliser* $\hat{\sigma}$ *pour réaliser les tests sur a et b.*
On commence par réaliser des tests sur a puis sur b.

a)Test a=0

Ce test revient à s'interroger sur l'influence réelle de x_t *sur* y_t *dans le cas où ce test est validé,c.a.d* $a = 0$ *ceci veut dire que le modèle ainsi supposé ne peut pas être écrit de cette façon,et l'hypothèse de linéarité peut être mis en cause, il faut supposer un autre modèle.*
On s'inspire du test de la moyenne vu dans le chapitre1 et on pose.

$$T = \frac{\hat{a} - a_0}{\sqrt{Var(\hat{a})}} = \frac{\hat{a}}{\dfrac{\hat{\sigma}}{\sqrt{\displaystyle\sum_{t=1}^{n}(x_t - \bar{x})^2}}}$$

Sous H_0, $T_a \hookrightarrow T_{n-2}$.
Donc on accepte H_0 *si* $\mid T_a \mid \leqslant c_\alpha$ *sinon on accepte* H_1.
avec $\frac{\alpha}{2} = P(T_{n-2} > c_\alpha) \Rightarrow P(T_{n-1} \leqslant c_\alpha) = 1 - \frac{\alpha}{2} \Rightarrow c_\alpha = t_{n-2;\,1-\frac{\alpha}{2}}$

Exemple 1.4. *On reprend toujours le même exemple de la consommation revenu.*
Il faut chercher $\widehat{\sigma^2} = \frac{SCR}{n-2} = \frac{1}{3}\sum_{t=1}^{5}(y_t - \hat{y}_t)^2 = \frac{1}{3}\sum_{t=1}^{3}(y_t - \hat{a}x_t - \hat{b})^2$.
Le calcul de $\hat{y}_t$ *se fait à partir du tableau suivant, on rappelle que* $\hat{y}_t = \hat{a}x_t + \hat{b}$

x_t	*85*	*92*	*99*	*108*	*116*
y_t	*82*	*88*	*93*	*102*	*110*
$\hat{y}_t$	*81.5*	*87.8*	*94.1*	*102.2*	*109.4*

Sous H_0 $T_a \hookrightarrow T_3$ *on accepte* H_0 *si* $\mid T_a \mid \leqslant c_\alpha$ *avec* $\frac{\alpha}{2} = P(T_3 > c_\alpha) \Rightarrow$
$P(T_{n-1} \leqslant c_\alpha) = 1 - \frac{\alpha}{2} \Rightarrow c_\alpha = t_{3;\,0.975} \Rightarrow t_{3;\,0.975} = 9.35.$
$\widehat{\sigma^2_{obs}} = \frac{1}{3}((0.5)^2 + \cdots + (0.6)^2 = \frac{1.90}{3} = 0.633.$
$T_{obs} = \frac{0.9.\sqrt{610})}{\sqrt{0.633}} = 27.93.$
On a $T_{obs} = 27.93 > t_{3;\,0.975} = 9.35$ *on accepte* H_1 *on a bien* $a \neq 0$

b) Test a>1

Un autre test important à faire ici pour le modèle consommation revenu c'est de valider l'hypothèse de Keynes sur la propension marginale à consommer et qui est comprise entre 0 et 1.On peut réaliser les tests suivants :
$H_0 :\ a \leqslant 1$ *contre* $H_1 :\ a > 1$

On pose $T_a = \dfrac{\hat{a}-1}{\sqrt{Var(\hat{a})}} = \dfrac{\hat{a}-1}{\dfrac{\hat{\sigma}}{\sqrt{\sum_{t=1}^{n}(x_t-\bar{x})^2}}}$.
SousH$_0$ *c.a.d lorsque* $a=1$ *on a* $T_a \hookrightarrow T_{n-2}$.
On accepte H_1 *si* $T \leqslant c_\alpha$.
On accepte H_0 *si* $T > c_\alpha$.
Avec $P(T_{n-2} > c_\alpha) = \alpha \Rightarrow T_{n-2} \leqslant c_\alpha) = 1-\alpha)$ *et* $c_\alpha = t_{n-2;\,1-\alpha}$

c) Test sur b

Si on veut tester par exemple $H_0;\ b = b_0$ *contre* $H_1;\ b \neq b_0$
On pose $T_b = \dfrac{\hat{b}-b_0}{\sqrt{Var(\hat{b})}} = \dfrac{\hat{b}-b_0}{\sqrt{\sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}}})$.
Sous H_0 *c.a.d lorsque* $b=b_0$ *on a* $T_b \hookrightarrow T_{n-2}$.
On accepte H_0 *si* $\mid T_b \mid \leqslant t_{n-2,1-\frac{\alpha}{2}} \Rightarrow T_b \in [-t_{n-2,1-\frac{\alpha}{2}}, t_{n-2,1-\frac{\alpha}{2}}]$

Un autre exemple de test sur b est $H_0;\ b \leqslant 0$ *contre* $H_1 :\ b > 0$ *c.a.d la consommation est t-elle positive lorsque le revenu est nul ?*

On pose toujours $T_b = \dfrac{\hat{b}}{\sqrt{Var(\hat{b})}} = \dfrac{\hat{b}}{\sqrt{\sigma^2(\dfrac{1}{n}+\dfrac{\bar{x}^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}}}$.
SousH$_0$ *c.a.d lorsque* $b=0$ *on a* $T_b \hookrightarrow T_{n-2}$.
On accepte H_1 *si* $T > c_\alpha$.
On accepte H_0 *si* $T \leqslant c_\alpha$.
$P(T_{n-2} > c_\alpha) = \alpha \Rightarrow T_{n-2} \leqslant c_\alpha) = 1-\alpha)$ *et* $c_\alpha = t_{n-2;\,1-\alpha}$.

Exemple 1.5. *Si on reprend l'exemple consommation revenu on a alors.*

$T_b = \dfrac{\hat{b}}{\sqrt{Var(\hat{b})}} = \dfrac{\hat{b}}{\sqrt{\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2})}}.$

Sous H_0; $T_b \hookrightarrow T_3$.

On a $\hat{b} = 5,\ n = 5,\ \hat{\sigma} = \sqrt{0.633},\ \sum_{t=1}^{5}(x_t - \bar{x})^2 = 610,\ \bar{x} = 100.$

$T_{b,obs} = \dfrac{5}{\sqrt{0.633}\sqrt{\frac{1}{5} + \frac{100^2}{610}}} = 1.543$ *si on prend* $\alpha = 5\%$ *on a* $c_\alpha = t_{3;\,0.95} = 7.81$ *et donc* $T_{b,obs} = 1.543 < 7.81 \Rightarrow$ *on accepte* H_0 $b<0$.

1.2.6 Application aux intervalles de confiance

a) Intervalle de confiance de a

On suit la méthode de constructions des intervalles de confiance déjà vu dans le cours de S3.

On doit construire l'intervalle à partir de la statistique utilisée dans les tests.

On sait que $\dfrac{\hat{a} - a}{\sqrt{Var(\hat{a})}} = \dfrac{\hat{a} - a}{\dfrac{\hat{\sigma}}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}}} \hookrightarrow T_{n-2}.$

Donc l'intervalle de confiance de a, de coefficient de sécurité $1 - \alpha$ *est donnée par.*

$$| \hat{a} - a | \leqslant t_{n-2;\,1-\frac{\alpha}{2}} . \frac{\hat{\sigma}}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}}$$

$$\Longrightarrow \boxed{a\in [\hat{a} - t_{n-2;1-\frac{\alpha}{2}}.\frac{\hat{\sigma}}{\sqrt{\sum_{t=1}^{n}(x_t-\bar{x})^2}}, \hat{a} + t_{n-2;1-\frac{\alpha}{2}}.\frac{\hat{\sigma}}{\sqrt{\sum_{t=1}^{n}(x_t-\bar{x})^2}}]}$$

Exemple 1.6. *Reprenons toujours l'exemple de consommation revenu. Si on prend* $\alpha = 5\%$ *n=5 on a* $t_{n-2;1-\frac{\alpha}{2}} = t_{3,0975} = 3.182$ *et l'intervalle de confiance de a est* $[0.9-3.182\sqrt{\frac{0.633}{610}}, 0.9+3.182\sqrt{\frac{0.633}{610}}] = [0.8-1]$.

b) Intervalle de confiance de b

Le même procédé est suivi comme pour l'intervalle de confiance de a. Pour construire un intervalle de confiance de b, de coefficient de sécurité $1-\alpha$, *on a.*

$$\frac{\hat{b}-b}{\sqrt{Var(\hat{b})}} = \frac{\hat{b}-b}{\sqrt{\sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}}} \hookrightarrow T_{n-2}.$$

L'intervalle de confiance de coefficient de sécurité $1-\alpha$ *est donné par :*

$$|\ \hat{b}-b\ | \leqslant t_{n-2;1-\frac{\alpha}{2}}.\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}}.$$

$$\Longrightarrow \boxed{b\in [\hat{b} - t_{n-2;1-\frac{\alpha}{2}}.\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}}; \hat{b} + t_{n-2;1-\frac{\alpha}{2}}.\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}}]}.$$

Exemple 1.7. *Reprenons toujours l'exemple de consommation revenu.*

Si on prend $\alpha = 5\%$ *n=5 on a* $t_{n-2;\,1-\frac{\alpha}{2}} = t_{3,0975} = 3.182$ *et l'intervalle de confiance de b est.*

$$[5-3.182.\sqrt{0.633}\sqrt{\frac{1}{5}+\frac{100^2}{610}}; 5+3.182.\sqrt{0.633}\sqrt{\frac{1}{5}+\frac{100^2}{610}}] = [-5.3; 15.3]$$

1.3 Notion de corrélation et Prévision

1.3.1 Coefficient de corrélation linéaire

Lorsque deux phénomènes ont une évolution commune, nous disons qu'ils sont corrélés. La corrélation simple mesure le degré de liaison existant entre ces 2 phénomènes représentés par des variables statistiques. On peut distinguer une corrélation linéaire positive, négative ou corrélation quelconque linéaire ou non.
Pour illustrer le phénomène de la corrélation on présente souvent les deux variables sur un graphique dit nuage de points et on a une idée sur la tendance de cette corrélation.
Mais le représentation graphique ne donne qu'une simple impression de la corrélation entre deux variables sans donner une idée précise de l'intensité de la liaison, c'es pourquoi nous calculons une statistique appelée coefficient de corrélation linéaire simple notée r et qui est donné par :

Definition 1.3. *Le coefficient de corrélation linéaire de* x_t *avec* y_t *est donné par :*

$$r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2 \sum_{t=1}^{n}(y_t - \bar{y})^2}}.$$ *r est à valeurs dans* $[-1; 1]$.

Remarque 1.5. *1)Par construction même, ce coefficient est compris* $-1 \leqslant r \leqslant 1$.

• *Si r est proche de 1, les variables sont corrélées positivement fortement.*
• *Si r est proche de -1, les variables sont corrélées négativement fortement.*
• *Si r est proche de 0, les variables ne sont pas corrélées .*

Dans la pratique il est des fois difficile de proposer une interprétation fiable à la simple lecture de ce coefficient. Ceci est surtout vraie en économie où les variables sont toutes plus au moins liées entre elles.De plus, il n'est calculé qu'à partir d'un échantillon d'observations et non sur l'ensemble des valeurs.
Pour y remédier à ce problème et donner une interprétation juste de la corrélation entre deux variables, on utilisera la théorie des tests.
Pour un échantillon donné ρ_{xy} *est une estimation de* r_{xy}.
Nous réalisons donc le test suivant :
$H_0;\ r_{xy} = 0$
$H_0;\ r_{xy} \neq 0$
On pose $T^* = \dfrac{\rho_{xy}}{\sqrt{\dfrac{1-\rho_{xy}^2}{n-2}}}$ *sous* $H_0,\ T^* \hookrightarrow T_{n-2}$.
On accepte H_0 *si* $\mid T^* \mid \leqslant c_\alpha$ *avec* $\dfrac{\alpha}{2} = P(T_{n-2} > c_\alpha) \Longrightarrow P(T_{n-1} \leqslant c_\alpha) = 1 - \dfrac{\alpha}{2} \Longrightarrow c_\alpha = t_{n-2;1-\frac{\alpha}{2}}$ *donc finalement on accepte* H_0 *si* $T^* \in [-t_{n-2;1-\frac{\alpha}{2}}; t_{n-2;1-\frac{\alpha}{2}}]$, *on accepte* H_1 *sinon.*

Proposition 6. *(admise) On a la statistique* $T_a = \dfrac{\hat{a}}{\dfrac{\hat{\sigma}}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}}} =$

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Remarque 1.6. *A partir de cette proposition on remarque que la statistique utilisée pour tester $a = 0$, est la même que celle utilisée pour tester $r = 0$ ce qui veut dire que les deux tests sont équivalents. Ceci peut être interprété tout simplement par le fait que lorsque r est significatif alors automatiquement $a \neq 0$ c.a.d que l'impact de x_t sur y_t existe.*

1.3.2 Le coefficient de détermination linéaire

On a vu que le coefficient de corrélation linéaire permet de mesurer l'intensité de la liaison entre x_t et y_t. Il existe un autre coefficient qui, lui nous permettra de juger sur la qualité de la regression linéaire. On l'appelle coefficient de détermination

Definition 1.4. *Le coefficient de détermination du modèle est :*

$$\boxed{R^2 = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}}.$$

et $0 \leqslant R^2 \leqslant 1$ il mesure la qualité globale de la régression, plus R^2 est proche de 1, plus le modèle est explicatif.

Remarque 1.7. *On avait dans la proposition 1.5 que la variance totale $V_t = V_e + V_r$. Cette équation permet aussi de définir le coefficient de détermination par $R^2 = \frac{V_e}{V_t}$, alors plus la variance expliquée est plus proche de la variance totale, meilleur est l'ajustement.*

Proposition 7. *On a $r^2 = R^2$*

Exemple 1.8. *Pour l'exemple consommation revenu on a.*
$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{1.90}{496} = 0.996.$

La liaison est très forte entre x_t et y_t et 99% de la variance totale V(Y) est expliquée par X.
Comme $R^2 = r^2 = 0.996 \Rightarrow r = \sqrt{r^2} = \sqrt{0.996} = 0.998$, donc on a aussi une très bonne corrélation linéaire entree x_t et y_t.

1.3.3 Prévision dans le modèle linéaire simple

L'étude du modèle linéaire simple montre un processus très rigoureux qu'on a détaillé avec les 8 étapes de l'organigramme d'une étude économétrique. On a vu à travers l'exemple consommation revenu par le bais d'un échantillon la modélisation de la variable C_t en fonction de R_t donné par $C_t = aR_t + b + \varepsilon_t$. L'estimation du paramètre a permis de savoir combien varie en moyenne la valeur de C_t lorsque celle de R_t augmente d'une unité.Ainsi, on sait que si le revenu augmente de 100 dh, la consommation va augmenter de 90 dh car la valeur estimée de $\hat{a} = 0.9$.
Le problème qui se pose maintenant est le suivant si une personne doit gagner un revenu R quelle doit être sa consommation C ?.
On peut le savoir avec précision et avec un intervalle de confiance qui nous dira de combien on pourra se tromper.On dit qu'on va prévoir, ou on fera de la prévision et c'est parmi les objectifs d'une étude économétrique et en particulier pour ce modèle linéaire simple.

Prévision

On rappelle le modèle théorique $y_t = ax_t + b + _t$. Lorsque les coefficients du modèle ont été estimés, il est possible de calculer une prévision à un horizon h.
Soit le modèle estimé $\hat{y}_t = \hat{a}x_t + \hat{b} + e_t$ sur la période $t = 1 \cdots .n$.
Si la valeur de la variable explicative x_t est connue à l'horizon h la prévision est donnée par :

$$\hat{y}_{t+h} = \hat{a}x_{t+h} + \hat{b}$$

La prévision sans biais est donc obtenue par l'application directe du modèle de régression estimé.Cependant, dans la pratique, il n'est que peu d'utilité de connaître la prévision si nous ne savons pas quel degré de confiance nous pouvons lui accorder. C'est pour cela nous donnons ici l'intervalle de confiance de cette prévision mais sans le démontrer.On note l'intervalle de prévision par :

$$IC = \hat{y}_{t+h} - t_{n-2;1-\frac{\alpha}{2}}\hat{\sigma}_\varepsilon\sqrt{\frac{1}{n} + \frac{x_{t+h} - \bar{x}}{\sum_{t=1}^{n}(x_t - \bar{x})^2 +} + 1};\ \hat{y}_{t+h} + t_{n-2;1-\frac{\alpha}{2}}\hat{\sigma}_\varepsilon\sqrt{\frac{1}{n} + \frac{x_{t+h} - \bar{x}}{\sum_{t=1}^{n}(x_t - \bar{x}}$$

Avec $\hat{\sigma}_\varepsilon = \dfrac{SCR}{n-2} = \dfrac{1}{n-2}\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$.

Exemple 1.9. *On reprend toujours l'exemple de la consommation revenu si.Si on sait que le revenu à l'époque t=8 est de 126.Quelle serait la consommation avec un intervalle de prédiction pour cette période ?.*$\alpha = 5\%$.
on sait que le modèle estimé est donné par $\hat{y}_t = 0.9x_t + 5 + e_t$ *On avait* $(1 \leqslant t \leqslant 5)$ *alors pour* $t = 8 \Rightarrow h = 3$.
La valeur de l'estimation à l'horizon h est $\hat{y}_8 = 0.9x_8 + 5 = 0.9.126 + 5 = 118.4$.
On a d'après les résultats déjà calculés : $\hat{\sigma}_\varepsilon = \sqrt{0.633} = 0.795,\ \sum_{t=1}^{5}(x_t - \bar{x})^2 = 610,\ n = 5,\ t_{3;0.975} = 3.182,\ \hat{y}_8 = 118.4$.

$$IC = [118.4 - 3.82.0.795\sqrt{\frac{126-116}{610} + 1 + \frac{1}{5}};\ 118.4 - 3.82.0.795\sqrt{\frac{126-116}{610} + 1 + \frac{1}{5}}] =$$
$$[118.4 - 2.789;\ 118.4 + 2.789] = [115.61;\ 121.189]$$

Bibliographie

[1] *Gérard.BAILLARGEON. Probabilités, Statistique et Techniques de régression. Les éditions SMG Quebec 1989.*
[2] *Jean Philippe REAU et Gérard CHAUVAUT. Probabilités et statistiques 4ème édition ARMAND COLIN 2008*
[3] *Régis.BOURBONNAIS Économétrie Cours et exercices corrigés 9ème édition DUNOD-2015*
[4]*Régis.BOURBONNAIS Exercices pédagogiques d'économétrie édition Economica-2008*
[5] *Stephen BAZEN et Mareva SABATIER. Économétrie des fondements à la modélisation Edition-Vuibert 2007.*