Mme BOTTOLLIER

partie 7:REGRESSION LINEAIRE CORRELATION

I - Ajustement entre 2 caractères quantitatifs

On a souvent besoin d'examiner la façon dont une variable quantitative est reliée à une autre variable quantitative.

On étudiera le cas où Y est reliée linéairement à une seule variable X, il s'agit de régression linéaire simple. Y est la **variable dépendante** (Cf V-1; tableau 1 colonne 1) à expliquer ou variable de réponse et X est la **variable explicative** ou **variable indépendante** ou encore **régresseur** (Cf V-1; tableau 1 colonne 2).

I -1 - La covariance

La covariance est la moyenne du produit des écarts de 2 variables statistiques à leurs moyennes respectives.

formule de définition: **cov (x, y) = SPE (x,y) / n**

$$\text{SPE } x,y = \sum_{i=1}^{i=n} x_i y_i - n\bar{x}\bar{y}$$

formule de calcul manuel: cov (x,y) = $\frac{\sum_{i=1}^{i=n} x_i y_i}{n} - \bar{x}\bar{y}$

dans l'exemple V – 1:	covariance = 33,16
	SPE x,y = 1824,01
	SP = 62897,77
	n = 55

Si 2 variables sont indépendantes leur covariance est nulle. Une covariance positive indique que les variables X et Y varient dans le même sens, une covariance négative indique que les variables X et Y varient dans le sens contraire.

I - 2 - Nuage de points

(Cf V1; fig 1)
Il permet de visualiser l'allure du nuage (relation linéaire ou non) et l'existence de données « out ».

I – 3 - Coefficients de la régression

Equation de la droite de régression de y en fonction de x (droite de y pour x fixé ; Dy/x) (Cf V; fig 1):

$y_i = bx_i + a + e_i$ (Cf V-1; tableau 1 colonne2)
$\hat{y}_i = bx_i + a$ (Cf V-1; tableau 1 colonne3)
$e_i = \hat{y}_i - y_i$ (Cf V-1; tableau 1 colonne 4)

b = SPE (x,y) / SCE x **b = cov (x,y) / var x** $a = \bar{y} - b\bar{x}$

Les coefficients **a** et **b** sont les **coefficients de régression** (Cf V-1;tableau 3 colonne 1). Ils sont déterminés par la méthode des moindres carrés ordinaires (MCO). Un tel critère entraîne que la **somme des résidus soit nulle**.
La droite de régression passe nécessairement par le **centre de gravité du nuage** $(\bar{x}\,;\bar{y})$.

Dans l'exemple V - 1	Température moyenne = 7,34 insolation moyenne = 151,36

b exprime la variation de Y pour une variation de X = +1, dans le domaine de X étudié
a exprime la valeur estimée de Y pour x = 0, elle n'a pas toujours de sens concret

I – 4 - Analyse de la qualité de la régression

I – 4 – 1 - Coefficient de détermination r^2

Il mesure la part de la variance totale qui est expliquée par la régression:

SCE totale des y = SCE des Y expliqués par la régression + SCE résiduelle
(Cf V-1; tableau 2 colonne 2)

$$SCE(y_i) = SCE(\hat{y}_i) + SCE(ei)$$

$$r^2 = SCE(\hat{y}_i) / SCE(y_i)$$

mais aussi:

$$r^2 = cov(x,y)^2 / (s_x^2\, s_y^2) \text{ ou alors } r^2 = SPE(x,y)^2 / (SCE(x) * SCE(y))$$

r^2 est un nombre compris entre 0 et 1

Dans l'exemple V - 1	SCE y = 107,55 SCE x = 38124,75 R^2 = 0,8114

Interprétation du coefficient de détermination: il exprime le pourcentage de la variation de Y expliquée par celle de X.

I – 4 – 2 - Coefficient de corrélation r

Le carré du coefficient de corrélation est égal au coefficient de détermination.

Dans l'exemple V - 1	r = 0,9008

r est un nombre compris entre –1 et +1, son signe est le même que celui de la covariance, de la SPE et de b.

I – 4 – 3 - Analyse des résidus

Pour valider le modèle de relation proposé il vaut s'assurer de:

- *La normalité des résidus* (Cf V-1; tableau 1 colonne 4 et fig2), la série des résidus peut être regroupée en classes , l'adéquation à une loi Normale peut se faire à l'aide d'un test du X^2
- *L'absence de données aberrantes : on calcule la liste des résidus standardisés,* (Cf V-1; tableau 1 colonne 5) , à toute valeur ei/se n'appartenant pas à l'intervalle [-2 ;+2] correspondra à un point (xi,yi) « out »
- *L'indépendance des résidus ei (ou ei standardisés) et des xi* (Cf V-1; fig 3), il s'agit d'un *graphe* , le nuage ne doit présenter aucune allure particulière (pas de modèle mathématique spécifique).

II - Les estimations: application à la régression.

II – 1 - Estimations ponctuelles

Coefficients de régression	$\hat{\alpha} = a$ $\hat{\beta} = b$
Variance résiduelle	$\sigma^2(\varepsilon) = SCE\ e / (n - 2)$ (Cf V-1; tableau 2 colonne 3, se note aussi CMe).
Variances des coefficients de régression	$\sigma^2(a) = \sigma^2(\varepsilon)\left(\frac{1}{n}+\frac{\bar{x}^2}{SCEx}\right)$ $\sigma^2(b) = \frac{\sigma^2(\varepsilon)}{SCE_x}$ (Cf V-1; tableau 3 colonne 2)
Valeur de yi	$\hat{y}i = bx_i + a$ (Cf V-1; tableau 1 colonne 3)

II – 2 - Estimations par intervalle de confiance

II – 2 –1 - Les coefficients de régression

Considérons l'étude de la régression linéaire simple mettant en jeu une seule variable explicative X selon le modèle :

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

où ε est une erreur aléatoire normale, de moyenne nulle et de variance $\sigma^2(\varepsilon)$, β et α sont les coefficients de la régression dans la population concernée et qui sont estimés à partir d'un échantillon.
Dans les n paires (xi ; yi) d'observations aléatoires indépendantes, les xi sont des valeurs exactes et certaines de X, les yi sont des réalisations de la variable aléatoire Y. Il résulte que a et b sont des variables aléatoires qui suivent une loi de Student à (n – 2) ddl avec:

$$E(a) = \alpha$$
$$E(b) = \beta$$

On peut ainsi déduire les intervalles de confiance bilatéraux suivants (Cf V-1; tableau 3 colonnes 5 et 6):

$$P\left[a - t_{1-\alpha/2}\,(n-2) * \sigma(a) < \alpha < a + t_{1-\alpha/2}\,(n-2) * \sigma(a)\right] = 1 - \alpha$$

L'intervalle a α% de risque de ne pas contenir la vraie valeur β.
Quand X augmente de 1 unité, Y a α% de risque de ne pas être modifiée d'une valeur comprise entre les limites supérieure et inférieure de l'intervalle d'estimation.

$$P\left[b - t_{1-\alpha/2}\,(n-2) * \sigma(b) < \beta < b + t_{1-\alpha/2}\,(n-2) * \sigma(b)\right] = 1 - \alpha$$

L'intervalle a α% de risque de ne pas contenir la vraie valeur α.
Quand X est égal à 0, Y a α% de risque de ne pas être égale à une valeur comprise entre les limites supérieure et inférieure de l'intervalle d'estimation.

Remarque: on pourra appliquer ces formules à la régression linéaire multiple

II – 2 – 2 - Prévisions d'une valeur moyenne et d'une valeur individuelle

Considérons l'étude de la régression linéaire simple mettant en jeu une seule variable explicative X selon le modèle :

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

Dans ce cas ddl = n - 2
Après avoir étudié la régression entre 2 variables X et Y, on peut déterminer, pour un nouvel individu caractérisé par une valeur xo les valeurs de yo auxquelles on peut s'attendre si :
- le nouvel individu est un élément de la population ayant fait l'objet de l'étude de la régression
- la valeur xo est dans le domaine de l'étude

L'estimation de la **valeur moyenne** (Cf V-1; tableau 4 colonnes 4 et 5) par intervalle de confiance bilatéral, au niveau (1-α) est :

$$(a + b\,x_o) \pm t_{1-\alpha/2}\,(n-2) * \sigma(\varepsilon) * \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SCEx}}$$

L'intervalle de prédiction est à l'intérieur de la région de prédiction, au niveau (1 – α), délimitée par 2 branches d'hyperboles (Cf V-1; fig 4).

L'estimation de la **valeur individuelle** (Cf V-1; tableau 4 colonnes 2 et 3) par intervalle de confiance bilatéral, au niveau (1-α) est:

$$(a + b\,x_o) \pm t_{1-\alpha/2}\,(n-2) * \sigma(\varepsilon) * \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SCEx} + 1}$$

L'intervalle de prédiction est à l'intérieur d'une nouvelle région de prédiction, au niveau (1 – α), délimitée par 2 nouvelles branches d'hyperboles (Cf V-1; fig 4).

II – 2 – 3 - Coefficient de corrélation

La loi suivie pas r n'étant pas une loi Normale, on utilise la transformée de Fisher qui suit une loi Normale.

$$z = \frac{1}{2} \text{Ln} \frac{1+r}{1-r}$$

$$L(z) = N(E(z); \sigma(z))$$

$$\text{où } E(Z) = \frac{1}{2} \text{Ln} \frac{1+\rho}{1-\rho}$$

$$\text{et } \sigma(Z) = \frac{1}{\sqrt{n-3}}$$

remarques :

- si la variance résiduelle est estimée la transformée de Fisher suit une loi de Student à (n-2) ddl.
- on repasse de z à r par la relation: $r = \frac{e^{2z}-1}{e^{2z}+1}$

$$P[z - t_{1-\alpha/2} (n-2) * \sigma(z) < Z < z + t_{1-\alpha/2} (n-2) * \sigma(z)] = 1 - \alpha$$

$z - t_{1-\alpha/2} (n-2) * \sigma(z) = z \text{ inf}$ on repasse en r inf

$z + t_{1-\alpha/2} (n-2) * \sigma(z) = z \text{ sup}$ on repasse en r sup

$$P[r \text{ inf} < \rho < r \text{ sup}] = 1 - \alpha$$

L'intervalle a α% de risque de ne pas contenir la vraie valeur du coefficient de corrélation ρ.

On peut mettre les limites au carré:

$$P[r^2 \text{ inf} < \rho^2 < r^2 \text{ sup}] = 1 - \alpha$$

L'intervalle a α% de risque de ne pas contenir la vraie valeur du coefficient de détermination ρ^2.

Dans l'exemple V - 1	
r = 0,9008	r inf = 0,8351
z = 1,48	r sup = 0,9412
$\sigma(z)$ = 0,14	r^2 inf = 0,6973
$t_{0,975}$ = 1,96	r^2 sup = 0,8858
$z-t_{0,975} \sigma(z)$ = 1,20	
$z+t_{0,975} \sigma(z)$ = 1,75	

III - Les tests appliqués à la régression et à la corrélation

III – 1 - Test de signification du coefficient β

L (b) = S (n-k-1)(E(b); σ(b)) (pour k coefficients dans le modèle)

où **E(b)** = β

et **σ²(b)** = $\frac{\sigma^2(\varepsilon)}{SCE_x}$

H_0: β = 0 l'effet du régresseur n'est pas mis en évidence
H_1: β ≠ 0 effet du régresseur mis en évidence (test bilatéral)

Critère statistique: (Cf V-1; tableau 3 colonne 3)

t calculé = $\frac{b-\beta}{\sigma(b)} = \frac{b}{\sigma(b)}$

Pour α fixé: si t calculé < $t_{1-\alpha/2}$ (n – 2) on conserve l'hypothèse nulle, ce qui signifie que les variations de X n'expliquent pas de façon significative les variations de Y compte tenu de la valeur du risque de première espèce retenue.
Pour α non fixé: Si on rejette H_0 on a un risque égal à α de le faire à tort ,
α = 2* P(T > t calculé) (Cf V-1; tableau 3 colonne 4)

On peut aussi réaliser le test sur la constante:
H_0: α = 0
H_1: α ≠ 0 le modèle ne passe pas par l'origine

Critère statistique: t calculé = $\frac{a-\alpha}{\sigma(a)} = \frac{a}{\sigma(a)}$ (Cf V-1; tableau 3 colonne 3)

Remarque: ce test s'appliquera à chaque coefficient de régression dans le cas de la régression linéaire multiple.

III – 2 - Analyse de la variance

Cette analyse permet de conclure si la part de variation des Y expliquée par le modèle de régression est supérieure à la part non expliquée (liée aux fluctuations naturelles).

H_0: $CM\hat{y}_i / CMe = 1$

H_1 : $CM\hat{y}_i / CMe > 1$ la part expliquée par le modèle est plus grande que la non expliquée.

Quand les coefficients de la régression (a et b) ont été déterminés, pour chaque valeur de xi on calcule la liste des estimations de yi en remplaçant xi dans l'équation, on établit également la liste des résidus ei. Puis on détermine les SCE associée à chacune de ces listes. On construit le tableau de l'AOVA (Cf V-1; tableau 2):

origine de la variation	SCE	ddl	CM =SCE / ddl	F calculé
Régression	SCE $\hat{y}_i = \Sigma (\hat{y}_i - \bar{y})^2$	1	CM $\hat{y}_i$ = SCE $\hat{y}_i$ / 1	CM $\hat{y}_i$ /CMe
Résiduelle	SCE e = $\Sigma (y_i - \hat{y}_i)^2$	n – 2	CM e = SCE e/(n – 2)	
Total	SCE y	n - 1		

Pour α fixé: Si F calculé < $F_{1-\alpha}$ (1; n – 2) on conserve l'hypothèse nulle
Pour α non fixé: Si on rejette H_0 on a un risque égal à α de le faire à tort , si ce risque est inférieur à 5% on pourra considérer que le modèle est satisfaisant (Cf V-1; tableau 2 colonne 5).

III – 3 - Test de signification du coefficient de corrélation

H_0: $\rho = 0$
H_1: $\rho \neq 0$ la corrélation linéaire entre X et Y n'est pas nulle

Critère statistique: t calculé = $\sqrt{\frac{r^2(n-2)}{1-r^2}}$

Pour α fixé: si t calculé < $t_{1-\alpha/2}$ (n – 2) on conserve l'hypothèse nulle, ce qui signifie que l'on n'a pas pu mettre en évidence l'existence d'un lien linéaire entre X et Y.
Pour α non fixé: Si on rejette H_0 on a un risque égal à α de le faire à tort.

Dans l'exemple V - 1	
r =	0,9008
t calculé =	15,10
alpha =	0,0000

Remarques: dans une régression simple les t calculés pour β et ρ sont égaux et sont reliés au critère F calculé par la relation :

$$t^2_{1-\alpha/2}(n-2) = F_{1-\alpha}(1; (n-2))$$

IV - La régression multiple

La régression multiple généralise la régression simple en étudiant la liaison stochastique entre une variable aléatoire **Y** (variable dépendante, variable à expliquer) et **k** variables certaines **X1, X2, Xj, Xk** (variables indépendantes, régresseurs, variables explicatives) au sein d'une population donnée dont on observe un échantillon aléatoire . On suppose en outre que les variables indépendantes X_j sont mesurées sans erreur. Le fait d'introduire des variables supplémentaires peut diminuer la valeur de la variance résiduelle et par là améliorer l'analyse.

IV – 1 - Estimations des paramètres du modèle

dans l'échantillon de n observations et k régresseurs: (i = 1.....n et j = 1 .....k))

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_j x_{ij} + \dots + b_k x_{ik} + e_i$$

dans la population:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_j x_{ij} + \dots + \beta_k x_{ik}$$

En notation matricielle (Cf V – 2; matrices 1 et 2): **Y = XB + e**

y **[n,1]**

y_1
y_2
....
y_i
....
....
y_n

X **[n,k+1]**

1	x_{11}	x_{12}	....	x_{1j}	....	x_{1k}
1	x_{21}	x_{22}	....	x_{2j}	....	x_{2k}
....	....	....	....	....	....	....
1	x_{i1}	x_{i2}	....	x_{ij}	....	x_{ik}
....	....	....	....	.	....	....
....	....	....	....	.	....	....
1	x_{n1}	x_{n2}	....	x_{nj}	....	x_{nk}

B **[k+1,1]**

b_0
b_1
....
....
b_j
....
....
b_k

e **[n,1]**

e_1
e_2
....
....
e_i
....
....
e_n

$$B = (X'X)^{-1} X'Y$$
$$Var(B) = \sigma^2_{\varepsilon} (X'X)^{-1}$$
$$\sigma^2_{\varepsilon} = SCEe/ddl\ résiduel$$
$$ddl = n - k - 1$$

où X' est la matrice transposée de X et $(X'X)^{-1}$ la matrice inverse du produit matriciel (X'X) (Cf V – 2; matrices 3 à 7).

- $L (\varepsilon_i) = N (0 ; \sigma_{\varepsilon})$
- il n'existe aucune corrélation entre les erreurs
- les variables X_j sont des grandeurs certaines indépendantes entre elles
- le nombre d'observations n doit être supérieur à (k + 1)

bj = variation de y consécutive à une variation d'une unité de xj, les autres facteurs de régression restant constants.

Matrice **X'X**: (comprendre la construction de cette matrice: c'est la matrice d'information indispensable pour le cours de 3ème année)

n	Σx_1	Σx_2	.....	Σx_k
Σx_1	Σx_1^2	$\Sigma x_1 x_2$.	....	$\Sigma x_1 x_k$.
Σx_2	$\Sigma x_2 x_1$.	Σx_2^2.	....	$\Sigma x_2 x_k$.
....	....	....	....	....
Σx_k	$\Sigma x_k x_1$.	$\Sigma x_k x_2$.	....	Σx_k^2

Matrice X'Y(Cf V – 2; matrice 6):

Σy
$\Sigma x_1 y$
$\Sigma x_2 y$
....
$\Sigma x_k y$

Var B est la **matrice des variances covariances** selon le principe (Cf V – 2; matrice 8):

var b_0	cov b_0b_1	cov b_0b_2	....	cov b_0b_k
cov b_0b_1	var b_1	cov b_1b_2	....	cov b_1b_k
cov b_0b_2	cov b_2b_1	var b_2	....	cov b_2b_k
....	....	....	....	....
cov b_0b_k	....	covb_2b_k	....	var b_k

Remarque : la matrice $(X'X)^{-1}$ (appelée aussi **matrice de dispersion**) se calcule sur Excel, elle vous sera donnée chaque fois que cela sera nécessaire.

IV – 2 - Test de signification du modèle

Il s'agit de répondre à la question : la régression est-elle significative dans son ensemble?
On calcule les valeurs de Y estimées d'après le modèle puis on établit la liste des résidus.
On détermine les SCE des Y estimés et des résidus (Cf V-2; tableau 5).

H_0: $CM\,\hat{y}_i / CMe = 1$

H_1 : $CM\,\hat{y}_i / CMe > 1$ la part expliquée par le modèle est plus grande que la non expliquée. Le modèle statistique (**structurel**) est satisfaisant.

tableau de l'analyse de variance (Cf V-2; tableau 6)

origine de la variation	SCE	ddl	CM	F calculé
Régression	$SCE\,\hat{y}_i$	k(*)	$CM\,\hat{y}_i = SCE\,\hat{y}_i$ / k	$CM\,\hat{y}_i / CMe$
Résiduelle	$SCEe = \Sigma (yi - \hat{y}_i)^2$	n – k - 1	CMe = SCEe/(n – k - 1)	
Total	SCEy	n - 1		

k(*): il s'agit du nombre de coefficients dans le modèle - 1

Si F calculé > $F_{1-\alpha}$ (k; n – k - 1) on rejette l'hypothèse nulle avec α % de risque d'erreur et on considère que le modèle structurel est satisfaisant. Sinon, on conserve H_0 et on conclue que l'on n'a pas pu mettre en évidence que le modèle structurel était valide.

IV – 3 - Test de signification sur chaque paramètre βj

H_0: $\beta_j = 0$ l'effet du régresseur n'est pas mis en évidence
H_1: $\beta_j \neq 0$ effet du régresseur mis en évidence (test bilatéral)

Critère statistique:

$$t \text{ calculé} = \frac{b-\beta}{\sigma(b)} = \frac{b}{\sigma(b)}$$ (Cf V-2; tableau 7 colonne 2)

On trouve la valeur de $\sigma(bj)$ en prenant la racine de la variance du coefficient dont il est question dans la diagonale de la matrice des variances-covariances (Cf V-2; matrice 8).

Pour α fixé: si t calculé < $t_{1-\alpha/2}$ (n – k - 1) on conserve l'hypothèse nulle, ce qui signifie que les variations de Xj n'expliquent pas de façon significative les variations de Y compte tenu de la valeur du risque de première espèce retenue.
Pour α non fixé: Si on rejette H_0 on a un risque égal à α de le faire à tort (Cf V-2; tableau 7 colonne 3).

IV – 4 - Coefficient de détermination

On définit **un coefficient de détermination multiple** égal au rapport entre la variance expliquée par l'ensemble des régresseurs et la variance totale de Y (Cf V-2; sous le tableau 6):

$$R^2 = SCE\,\hat{y}_i \,/\, SCE\,y$$

On peut définir un coefficient de détermination **partiel** pour un régresseur j qui permet d'évaluer la réduction de la variation non expliquée, en terme de proportion, lorsqu'on introduit cette nouvelle variable explicative j dans l'équation de régression, compte tenu de l'influence des autres régresseurs déjà retenu dans le modèle (Cf V-2; tableau 12).

IV – 5 - Méthodologie

Une régression linéaire est toujours délicate à interpréter car les régresseurs ne sont généralement pas indépendants, on peut donc procéder de différentes façons , par :

- régression avec le modèle complet où l'équation contiendra alors les k régresseurs (Cf V-2; tableau 8),

- régression progressive (régression pas à pas) ascendante ou descendante qui consiste à ajouter ou supprimer une variable explicative dans la mesure où les coefficients de détermination des 2 régressions sont significativement différents. Dans le procédé ascendant on introduit la variable Xj la plus corrélée avec Y (Cf V-2; tableau 9), puis on introduit comme seconde variable celle qui, après Xj, augmente le plus R^2 à condition que sa contribution soit significative (Cf V-2; tableau 10). Pour déterminer l'ordre d'introduction des régresseurs, on examine la matrice des corrélations (Cf V-2; matrice 9). Avant de poursuivre, on examine si le premier régresseur reste significatif et s'il l'est, on peut introduire un troisième régresseur ….etc. Si tous les variables Xj sont introduites, le résultat est identique à celui obtenu avec le modèle complet.

Avec cette méthode on peut déterminer par le calcul la part des variations des Y expliquées par chacun des régresseurs, et par là faire une analyse de variance pour chacun des régresseurs (Cf V-2; tableau 11).

V - Exemples

V – 1 - Régression linéaire simple

Tableau 1	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5
Villes	T (°C)	insolation (h)	y estimé	ei	ei / σε
Abbeville	5,8	114	5,55	0,25	0,41
Agen	8,2	159,3	7,72	0,48	0,78
Ajaccio	10,1	187,7	9,07	1,03	1,66
Ambérieu	6,2	130,5	6,34	-0,14	-0,22
Angers	7,4	150,5	7,30	0,10	0,17
Aurillac	7	149,3	7,24	-0,24	-0,38
Besançon	5,8	129,1	6,27	-0,47	-0,76
Bordeaux	8,5	161,9	7,84	0,66	1,07
Bourg-st Maurice	6,8	150	7,27	-0,47	-0,76
Bourges	6,7	137,8	6,69	0,01	0,02
Brest	7,4	126,5	6,15	1,25	2,03
Bron	6,9	147,5	7,15	-0,25	-0,41
Caen	6,6	137,7	6,68	-0,08	-0,13
Carcassonne	8,7	164,7	7,97	0,73	1,17
Carpentras	8,9	210,3	10,16	-1,26	-2,03
Chartres	6,1	134,4	6,52	-0,42	-0,69
Châteauroux	6,5	140,8	6,83	-0,33	-0,54
Clermont-ferrand	6,6	142	6,89	-0,29	-0,47
Cognac	8,3	153,3	7,43	0,87	1,41
Colmar	5,7	127,8	6,21	-0,51	-0,82
Dijon	6,2	140,3	6,81	-0,61	-0,98
Dinard	7,2	142,8	6,93	0,27	0,44
Gourdon	7,6	155,3	7,52	0,08	0,12
Grenoble	5,8	142,6	6,92	-1,12	-1,81
Ile de Bréhat	7,8	134,5	6,53	1,27	2,05
La Rochelle	8,6	173,8	8,41	0,19	0,31
Le Mans	7	141,5	6,86	0,14	0,22
Lille	5,6	115	5,60	0,00	0,01
Limoges	6,6	136,9	6,64	-0,04	-0,07
Lorient	7,6	148	7,18	0,42	0,69
Mâcon	6,5	146,5	7,10	-0,60	-0,98
Marseille	9,9	215,1	10,39	-0,49	-0,79
Metz	5,5	122,4	5,95	-0,45	-0,73
Mont Aigoual	8	186,9	9,04	-1,04	-1,68
Mont-de-Marsan	8,6	161,4	7,82	0,78	1,27
Montélimar	8,4	187,6	9,07	-0,67	-1,08
Montpellier	9,7	208,1	10,05	-0,35	-0,57
Nancy	5,2	120,8	5,87	-0,67	-1,09
Nantes	7,9	148,4	7,19	0,71	1,14
Nice	10,6	202,3	9,77	0,83	1,34
Nîmes	9,6	203,1	9,81	-0,21	-0,34
Orléans	6,4	134,7	6,54	-0,14	-0,23
Paris	7,5	134,3	6,52	0,98	1,58
Pau	8,5	150,7	7,30	1,20	1,93
Poitiers	7	148,1	7,18	-0,18	-0,29
Reims	5,7	129,4	6,29	-0,59	-0,95

St-Etienne	5,9	140,9	6,84	-0,94	-1,51
St-Giron	7,5	160,8	7,79	-0,29	-0,47
St-Quentin	5,4	119,6	5,82	-0,42	-0,67
Strasbourg	5,6	122,5	5,96	-0,36	-0,57
Tarbes	7,6	155,6	7,54	0,06	0,10
Toulon	11	217,6	10,51	0,49	0,80
Toulouse	8,3	164,7	7,97	0,33	0,53
Tours	7	137	6,65	0,35	0,57
Troyes	6	120,5	5,86	0,14	0,23

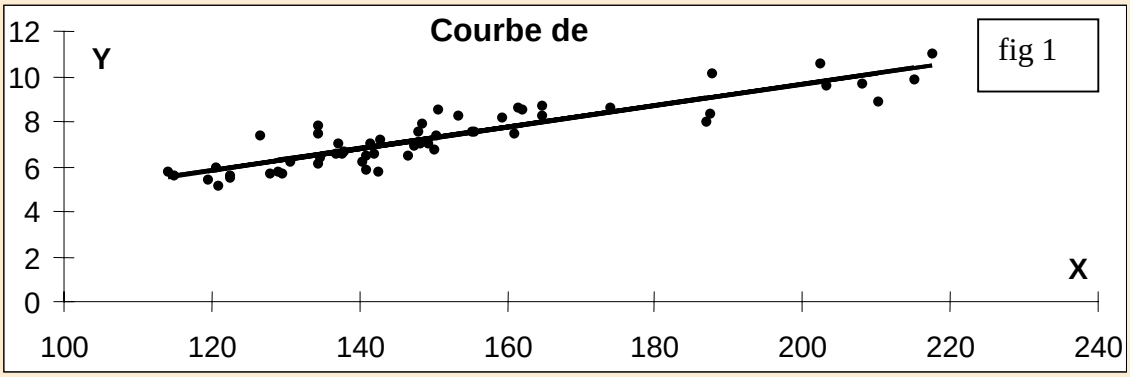
Tableau 2 : ANALYSE DE VARIANCE

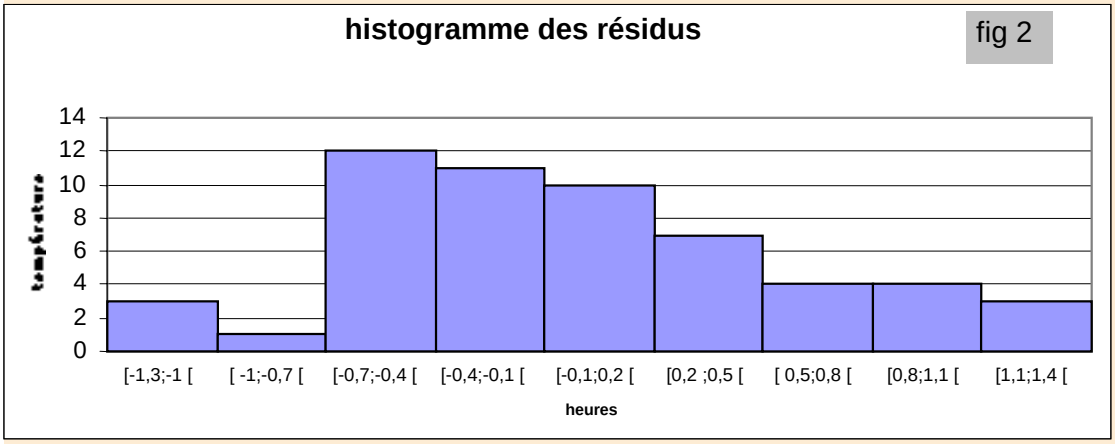
	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5
	ddl	SCE	CM	F calc	alpha
Régression	1	87,27	87,27	228,05	0,0000
Résidus	53	20,28	0,38		
Total	54	107,55			

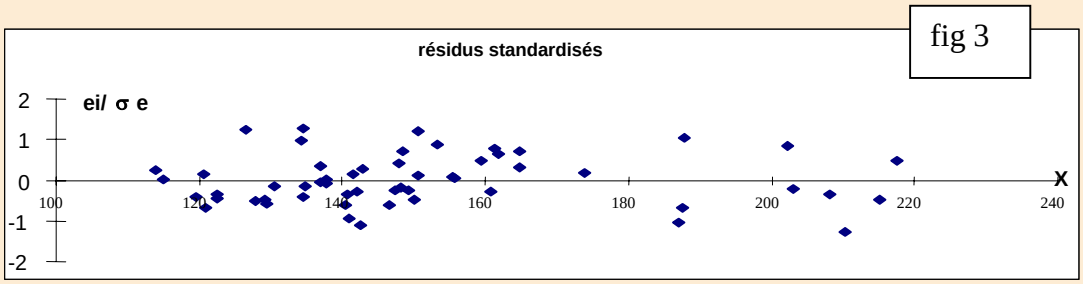
Tableau 3	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5	Colonne 6
	Coefficients	Erreur-type	t calculé	alpha	Lim inf (5%)	Lim sup(5%)
Constante	0,095	0,487	0,19	0,8463	-0,881	1,071
coefficient b	0,048	0,003	15,10	0,0000	0,041	0,054

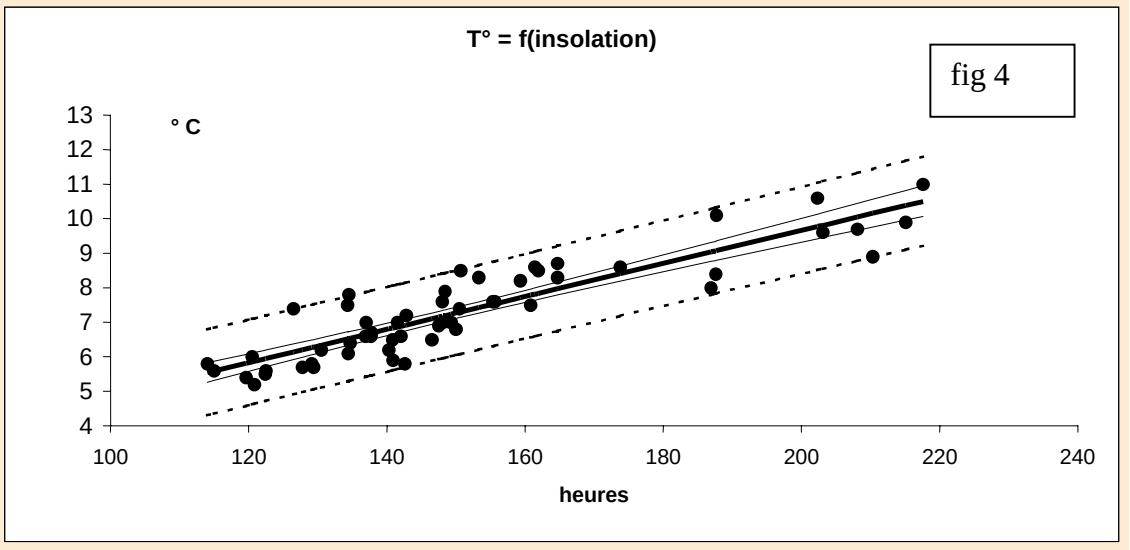
Tableau 4	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5
x	y	y estimé	y ind lim inf	y ind lim sup	y moy lim inf	y moy lim sup
114	5,8	5,55	4,30	6,79	5,27	5,83
115	5,6	5,60	4,35	6,84	5,32	5,88
119,6	5,4	5,82	4,58	7,06	5,56	6,07
120,5	6	5,86	4,62	7,10	5,61	6,11
120,8	5,2	5,87	4,64	7,11	5,62	6,12
122,4	5,5	5,95	4,71	7,19	5,71	6,19
122,5	5,6	5,96	4,72	7,19	5,71	6,20
126,5	7,4	6,15	4,91	7,38	5,92	6,37
127,8	5,7	6,21	4,98	7,44	5,99	6,43
129,1	5,8	6,27	5,04	7,50	6,06	6,49
129,4	5,7	6,29	5,05	7,52	6,07	6,50
130,5	6,2	6,34	5,11	7,57	6,13	6,55
134,3	7,5	6,52	5,29	7,75	6,33	6,71
134,4	6,1	6,52	5,30	7,75	6,33	6,72
134,5	7,8	6,53	5,30	7,76	6,34	6,72
134,7	6,4	6,54	5,31	7,77	6,35	6,73
136,9	6,6	6,64	5,42	7,87	6,46	6,83
137	7	6,65	5,42	7,88	6,46	6,84
137,7	6,6	6,68	5,46	7,91	6,50	6,87
137,8	6,7	6,69	5,46	7,91	6,50	6,87
140,3	6,2	6,81	5,58	8,03	6,63	6,98
140,8	6,5	6,83	5,61	8,06	6,65	7,01
140,9	5,9	6,84	5,61	8,06	6,66	7,01
141,5	7	6,86	5,64	8,09	6,69	7,04
142	6,6	6,89	5,66	8,11	6,72	7,06
142,6	5,8	6,92	5,69	8,14	6,74	7,09
142,8	7,2	6,93	5,70	8,15	6,75	7,10
146,5	6,5	7,10	5,88	8,33	6,94	7,27

147,5	6,9	7,15	5,93	8,38	6,99	7,32
148	7,6	7,18	5,95	8,40	7,01	7,34
148,1	7	7,18	5,96	8,40	7,02	7,35
148,4	7,9	7,19	5,97	8,42	7,03	7,36
149,3	7	7,24	6,01	8,46	7,07	7,40
150	6,8	7,27	6,05	8,49	7,11	7,44
150,5	7,4	7,30	6,07	8,52	7,13	7,46
150,7	8,5	7,30	6,08	8,53	7,14	7,47
153,3	8,3	7,43	6,21	8,65	7,27	7,59
155,3	7,6	7,52	6,30	8,75	7,36	7,69
155,6	7,6	7,54	6,32	8,76	7,37	7,70
159,3	8,2	7,72	6,49	8,94	7,55	7,89
160,8	7,5	7,79	6,56	9,01	7,61	7,96
161,4	8,6	7,82	6,59	9,04	7,64	7,99
161,9	8,5	7,84	6,62	9,07	7,66	8,02
164,7	8,7	7,97	6,75	9,20	7,79	8,16
164,7	8,3	7,97	6,75	9,20	7,79	8,16
173,8	8,6	8,41	7,18	9,64	8,20	8,62
186,9	8	9,04	7,79	10,28	8,76	9,31
187,6	8,4	9,07	7,83	10,31	8,79	9,35
187,7	10,1	9,07	7,83	10,32	8,80	9,35
202,3	10,6	9,77	8,51	11,04	9,42	10,13
203,1	9,6	9,81	8,55	11,08	9,45	10,17
208,1	9,7	10,05	8,78	11,32	9,66	10,44
210,3	8,9	10,16	8,88	11,43	9,76	10,56
215,1	9,9	10,39	9,10	11,67	9,96	10,81
217,6	11	10,51	9,21	11,80	10,06	10,95


Courbe de
fig 1
Y
X
12
10
8
6
4
2
0
100
120
140
160
180
200
220
240


histogramme des résidus
fig 2
température
14
12
10
8
6
4
2
0
[-1,3;-1 [
[-1;-0,7 [
[-0,7;-0,4 [
[-0,4;-0,1 [
[-0,1;0,2 [
[0,2 ;0,5 [
[0,5;0,8 [
[0,8;1,1 [
[1,1;1,4 [
heures


résidus standardisés
fig 3
ei/ σ e
X
2
1
0
-1
-2
100
120
140
160
180
200
220
240


T° = f(insolation)
fig 4
° C
13
12
11
10
9
8
7
6
5
4
100
120
140
160
180
200
220
240
heures

V – 2 - Régression linéaire multiple

Matrice 1 = matrice X			
1	84	9	7
1	71	6	4
1	73	5	3
1	78	9	11
1	69	5	0
1	81	11	6
1	68	5	2
1	71	8	5
1	80	9	11
1	75	7	5

Matrice 2 = mat Y
38,01
29,35
29,19
35,06
28,14
38,53
28,06
30,85
36,28
32,15

Matrice 3 = matrice X'									
1	1	1	1	1	1	1	1	1	1
84	71	73	78	69	81	68	71	80	75
9	6	5	9	5	11	5	8	9	7
7	4	3	11	0	6	2	5	11	5

Matrice 4 = produit X'X			
10	750	74	54
750	56522	5638	4181
74	5638	588	451
54	4181	451	406

Matrice 5 = matrice $(X'X)^{-1}$			
51,882	-0,829	1,108	0,404
-0,829	0,014	-0,023	-0,005
1,108	-0,023	0,097	-0,017
0,404	-0,005	-0,017	0,022

Matrice 6 = produit X'Y
325,62
24616,91
2482,88
1858,08

Matrice 7 = matrice B
-7,887
0,456
0,878
-0,045

Tableau 5	Colonne 1	Colonne 2	Colonne 3
	y observé	y estimé	ei
	38,01	38,00	0,012
	29,35	29,57	-0,222
	29,19	29,65	-0,460
	35,06	35,08	-0,024
	28,14	27,96	0,179
	38,53	38,43	0,098
	28,06	27,42	0,644
	30,85	31,28	-0,433
	36,28	36,00	0,284
	32,15	32,23	-0,079
SCE =	**149,971**	148,978	**0,993**

Tableau 6 = ANOVA				
	SCE	ddl	CM	F
régression	148,978	3	49,66	300
aléatoire	**0,993**	**6**	0,166	
totale	**149,971**	9		

Alpha = 6 E-07

R^2 = 148,978/**149,971**=0,9934

Matrice 8 = Var B			
8,612	-0,138	0,184	0,067
-0,138	**0,002**	-0,004	-0,001
0,184	-0,004	**0,016**	-0,003
0,067	-0,001	-0,003	**0,004**

Tableau 7	Colonne 1	Colonne 2	Colonne 3	
val abs bj	**σ(bj)**	**t calculé**	**alpha**	
7,887	2,93	2,69	0,0362	*
0,456	0,048	9,56	0,0001	***
0,878	0,127	6,93	0,0004	***
0,045	0,061	0,73	0,4908	NS

Tableau 8 = Fonction « droite reg » sur Excel			
-0,045	0,878	0,456	-7,887
0,061	0,127	0,048	2,93
0,9934	0,41	#N/A	#N/A
300	6	#N/A	#N/A
148,978	0,993	#N/A	#N/A

Matrice 9 = Matrice des R^2			
	X_1	X_2	X_3
Y	0,9361	0,8866	0,5797
X_1	1	0,705	0,5515
X_2	0,7047	1	0,5716
X_3	0,5515	0,5716	1

Méthode ascendante:

X_1	mat Y
84	38,01
71	29,35
73	29,19
78	35,06
69	28,14
81	38,53
68	28,06
71	30,85
80	36,28
75	32,15

Tableau 9	
0,72	-21,32
0,066	4,990
0,9361	1,09
117,17	8
140,39	9,59
SCEy/X_1	

X_1	X_2	mat Y
84	9	38,01
71	6	29,35
73	5	29,19
78	9	35,06
69	5	28,14
81	11	38,53
68	5	28,06
71	8	30,85
80	9	36,28
75	7	32,15

Tableau 10		
0,84	0,45	-7,08
0,11	0,04	2,63
0,9928	0,39	#N/A
481,27	7	#N/A
148,89	1,08	#N/A

SCEy/X_1X_2

148.89 – 140.39 = 8.5

148,98 **-148.89** = 0.09

Tableau 11 = ANOVA					
	SCE	ddl	CM	F	alpha
X_1	140,39	1	140,39	848	0,0000
X_2/ X_1	8,5	1	8,50	51	0,0004
X_3 /X_2 /X_1	0,09	1	0,09	0,54	0,4903
aléatoire	0,99	6	0,166		
totale	149,97	9			

Tableau 12	
	R^2 partiel
X_1	0,9361
X_2/ X_1	0,0567
X_3 /X_2 /X_1	0,0006
total	0,9934