

Régression, corrélation

Régression, corrélation.....	2
7.1 Régression simple.....	2
7.1.1 Données individuelles et couple de variables.....	2
7.1.2 Représentations graphiques.....	3
7.1.3 La question de l'ajustement.....	4
7.1.4 Modèle versus contingence.....	5
7.1.5 Choix d'une fonction.....	6
7.1.6 Choix d'un critère d'ajustement : les moindres carrés.....	7
7.1.7 Interprétation de Galton : la régression.....	9
7.1.8 La synthèse de Yule.....	10
7.1.9 Décomposition de la variance.....	10
7.1.10 Corrélation linéaire.....	11
7.2 Généralisation à plusieurs variables.....	12
7.3 Rendement et corrélation partielle.....	14
7.4 Interprétation probabiliste.....	15
7.4.1 Modèle linéaire à erreurs aléatoires.....	15
7.4.2 Hypothèse de normalité des erreurs.....	16
7.4.3 Test de significativité d'un coefficient.....	16
7.4.4 Test de significativité globale de la régression.....	17
7.5 Exemple.....	18

Régression, corrélation

7.1 Régression simple

7.1.1 Données individuelles et couple de variables.

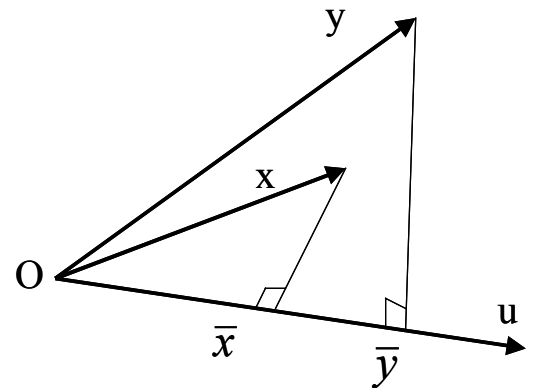
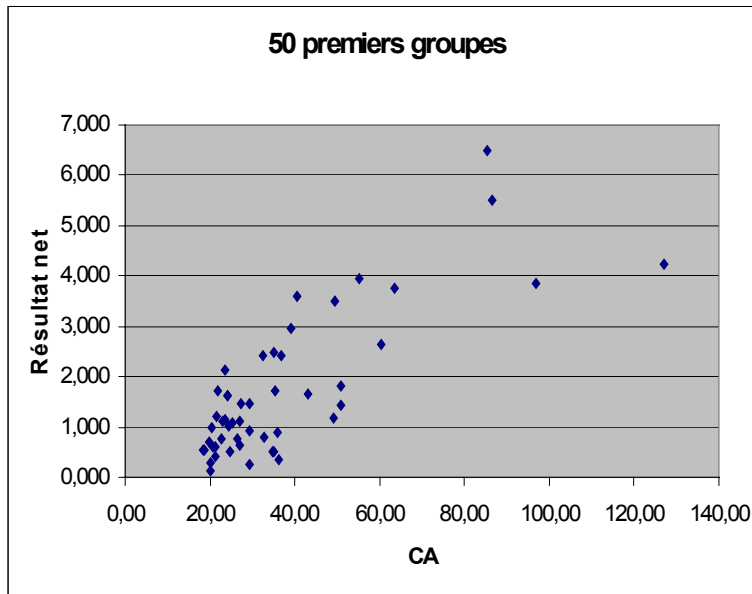
L'objectif de cette leçon est d'étudier la liaison statistique entre deux, puis plusieurs, variables quantitatives, observées sur un échantillon de n individus. Commençons par deux variables. Ces deux variables correspondent à deux colonnes du tableau Individus-Variables selon lequel on peut se représenter le fichier des résultats d'une enquête. Pour simplifier nous rebaptisons X et Y ces deux variables : par exemple $X = X_j$ et $Y = X_k$. Pour l'individu i , ces deux variables X et Y prennent les valeurs x_i et y_i .

Variables→ Individus↓	X_1	X_2	...	$X = X_j$...	$Y = X_k$	X_p
1							
2							
.....							
i				x_i		y_i	
.....							
n							

Si les variables X et Y du tableau de données individuelles sont numériques, on peut étudier leur liaison directement à partir des n couples de valeurs (x_i, y_i) , sans les réduire sous forme d'une table des fréquences conjointes, comme nous l'avons fait dans la leçon précédente.

7.1.2 Représentations graphiques

On peut d'abord *représenter* les données par un graphique. Le plus souvent on utilisera la représentation cartésienne des couples (x_i, y_i) par un ensemble de n points de R^2 , appelé *nuage de points* (*scatter diagram*), chaque point correspondant à une ligne du tableau de données, du type de celui qui figure ci-dessous, correspondant à notre



exemple introductif :

Mais on peut aussi se représenter les données comme deux vecteurs x et y de R^n dont les composantes sont les n observations, et qui correspondent à deux colonnes du tableau de données : $x = (x_1, x_2, \dots, x_n)$ et $y = (y_1, y_2, \dots, y_n)$. Tant qu'il n'y a que deux variables, ces deux vecteurs définissent un plan et la représentation graphique reste dans R^2 . Dans notre figure de droite nous avons également représenté le vecteur particulier $u = (1, 1, 1, \dots, 1)$ dit vecteur des constantes, et la figure est donc à 3 dimensions. Cela permet d'introduire les vecteurs moyennes $\bar{x}u$ (projection orthogonale de x sur u) et $\bar{y}u$ (projection de y sur u) et les vecteurs des données centrées $(x - \bar{x}u)$ et $(y - \bar{y}u)$ dont les longueurs sont les écarts-types de x et y . En effet $V(x) = \sum (x_i - \bar{x})^2 / n = \|x - \bar{x}u\|^2$, l'expression $\|z\|$ signifiant « norme » ou « longueur » du vecteur z .

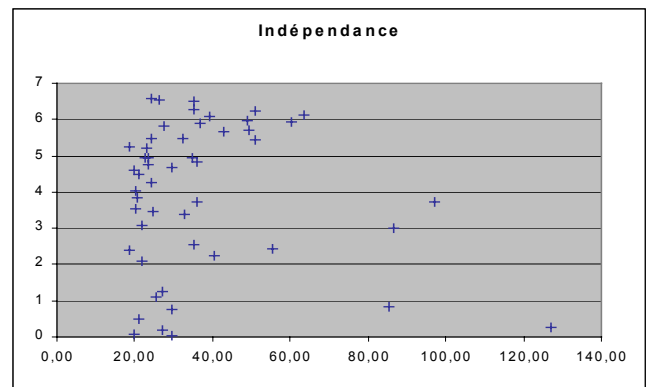
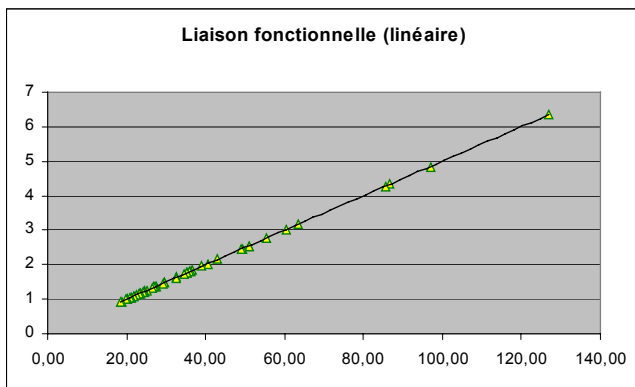
Variable	X	U	$X - \bar{x}U$	Y	$Y - \bar{y}U$
Composante 1	x_1	1	$x_1 - \bar{x}$	y_1	$y_1 - \bar{y}$
Composante 2	x_2	1	$x_2 - \bar{x}$	y_2	$y_2 - \bar{y}$
	...				
Composante n	x_n	1	$x_n - \bar{x}$	y_n	$y_n - \bar{y}$
Moyenne des composantes	$\bar{x} = \sum x_i / n$	1	0	$\bar{y} = \sum y_i / n$	0
Moyenne des carrés	$\sum x_i^2 / n$	1	$V(x) = \sum (x_i - \bar{x})^2 / n$	$\sum y_i^2 / n$	$V(y) = \sum (y_i - \bar{y})^2 / n$

7.1.3 La question de l'ajustement

Au vu de la représentation du couple (X,Y) sous forme d'un nuage de points, la question qui se pose est de caractériser le type de liaison qui existe entre X et Y. En général le graphique fait état d'une liaison statistique qui est intermédiaire entre deux situations limites :

- la *relation fonctionnelle* dans laquelle toute valeur de x s'accompagne d'une et une seule valeur de y qui est déterminée en fonction de x : $y = f(x)$. La fonction f pouvant être n'importe quelle fonction mathématique continue ou non, monotone (croissante ou décroissante) ou non. Par exemple linéaire, exponentielle.... Le nuage de point est alors concentré sur la courbe représentant la fonction f.

- la situation *d'indépendance statistique*, pour laquelle à chaque valeur de x correspond n'importe quelles valeurs de y et vice versa. Plus précisément, la loi de distribution de y pour x fixé ne semble pas dépendre du x fixé; et vice-versa. Le nuage de point a la forme d'une boule de neige, faite de points pris au hasard dans le plan.



Dès lors, la question de la liaison des variables X et Y se décompose en trois problèmes :

1. Chercher quel est le type de relation fonctionnelle la plus proche de la situation observée (problématique de l'ajustement).
2. Trouver les paramètres de la fonction choisie en 1 pour que les écarts entre modèle et données soient les plus petits possibles (problématique de l'estimation).
3. Trouver un indicateur de la qualité de l'ajustement, c'est à dire du degré de confiance que l'on peut avoir dans la relation fonctionnelle substituée aux données.

7.1.4 Modèle versus contingence

Il se peut que les observations résultent d'un modèle que l'on connaît déjà pour des raisons théoriques. C'est souvent le cas dans les sciences physiques (astronomie, mécanique...) et parfois dans les sciences économiques où la théorie permet d'écrire a priori que la relation entre y et x est de telle forme. Par exemple la longueur d'un degré de méridien varie linéairement avec le sinus² de la latitude (ellipsoïde de la géode). Le volume d'un gaz parfait varie à l'inverse de sa pression à température constante (loi de Mariotte). Le prix d'une marchandise est une fonction décroissante des quantités demandées....Auquel cas la forme de la fonction est connue a priori et les écarts des points du nuage à la courbe correspondante s'interprètent comme des *erreurs* de mesure ou de spécification (on n'est pas dans les conditions où le modèle s'applique bien, d'autres facteurs perturbent la relation fondamentale...).

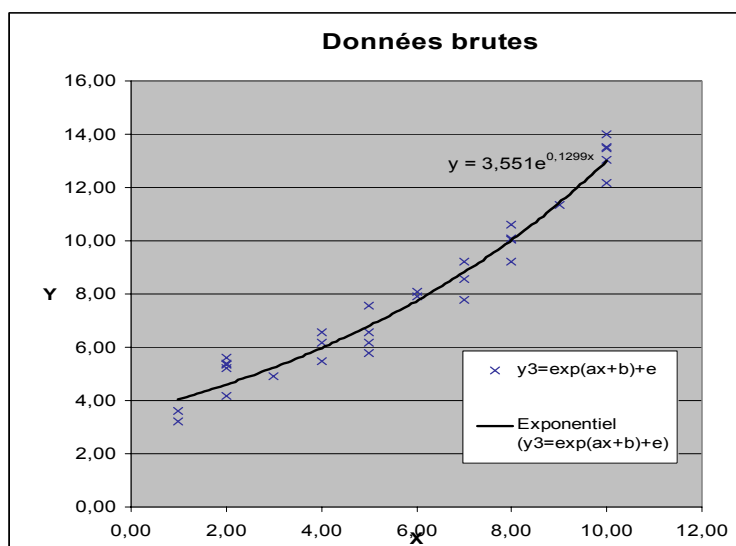
Dans les sciences sociales où de telles lois déterministes connues a priori sont rares, le nuage de point ne traduit au mieux qu'une *contingence*, selon l'expression de Karl Pearson, c'est à dire une relation qui est d'abord une co-occurrence constatée ici et maintenant, mais non nécessaire, de phénomènes. C'est dans ce cas le nuage de point et la distribution de fréquences sur les valeurs possibles de X et de Y qui forment la première trace des phénomènes, et c'est dans un second temps que l'on peut chercher à y repérer une régularité, voire une loi, dont le résumé « sténographique » peut être la courbe d'une fonction. Celle ci n'est donc pas connue ou supposée a priori, mais elle émerge de l'observation, de la mesure, et de sa trace que représente le nuage de point. C'est alors que l'on doit chercher la forme de la fonction la plus susceptible de résumer la liaison statistique.

La première approche fut plutôt celle des géomètres de la Théorie des erreurs dans le cadre de la « philosophie naturelle », telle que l'ont développée par exemple Laplace et Gauss au début du XIX^{ème} siècle. La seconde approche fut plutôt celle des biométriciens, sociologues et psychologues de la fin du XIX^{ème} siècle, à la suite des travaux de Galton et Pearson. Après 1910, c'est une sorte de synthèse hybride construite par Yule et Bowley, qui domine les usages en économie.

7.1.5 Choix d'une fonction

La première des trois phases de l'ajustement consiste à choisir la forme de la fonction à ajuster. Elle n'existe que dans la seconde approche où aucun modèle ne s'impose a priori, mais aussi dans le cas où le modèle a priori est faiblement spécifié : on sait par exemple qu'il s'agit d'une fonction décroissante mais sans plus.

Il n'existe pas malheureusement de méthode mathématique complète permettant de choisir le meilleur modèle mathématique parmi tous les modèles possibles. On est donc obligé de procéder par la méthode des essais et des erreurs, le recours aux mathématiques n'intervenant que pour estimer les paramètres de la fonction choisie, ou de comparer les qualités de deux ajustements par des fonctions différentes.



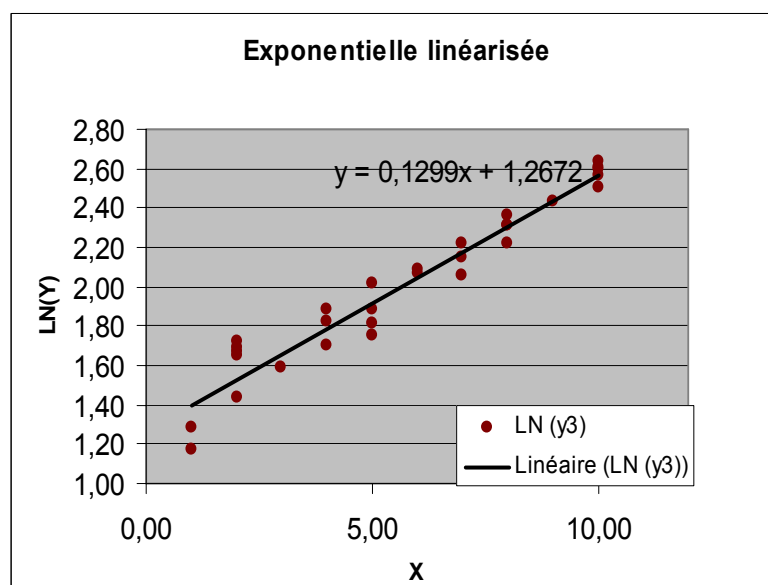
Dans l'exemple ci contre, le nuage de points est visiblement en forme de banane, et l'on doit rejeter un ajustement linéaire. Mais on peut hésiter par exemple entre l'ajustement proposé par une exponentielle, et un ajustement par une fonction puissance. Seule la comparaison des deux ajustements en terme de mesure de l'écart entre nuage et courbe pourra nous permettre de trancher. A condition d'avoir choisi le critère d'écart (cf. infra).

Supposons maintenant que sur ce nuage de point, nous ayons choisi d'ajuster une fonction

exponentielle. Deux méthodes s'offrent alors à nous pour passer à la seconde phase du programme qui est l'estimation des paramètres a et b de cette courbe :

- chercher directement les valeurs de a et b qui rendent la courbe la plus proche du nuage de point (en un sens qu'il faudra définir) : nous parlerons d'ajustement non linéaire.

- trouver une transformation de X ou de Y (ou des deux) qui permette de linéariser leur relation, puis ajuster une droite sur le nouveau nuage de points. Dans



notre exemple la transformation consiste à prendre le logarithme népérien de y . En effet :

$$y = \exp(ax+b) \Rightarrow \ln(y) = ax+b$$

puisque les fonctions logarithme népérien et exponentielle sont inverse. Et ceci permet de passer d'une relation non linéaire à une relation linéaire : si dans le plan (x,y) le nuage s'étire le long d'une courbe exponentielle, dans le plan $(x, \ln y)$ il s'étire le long d'une droite.

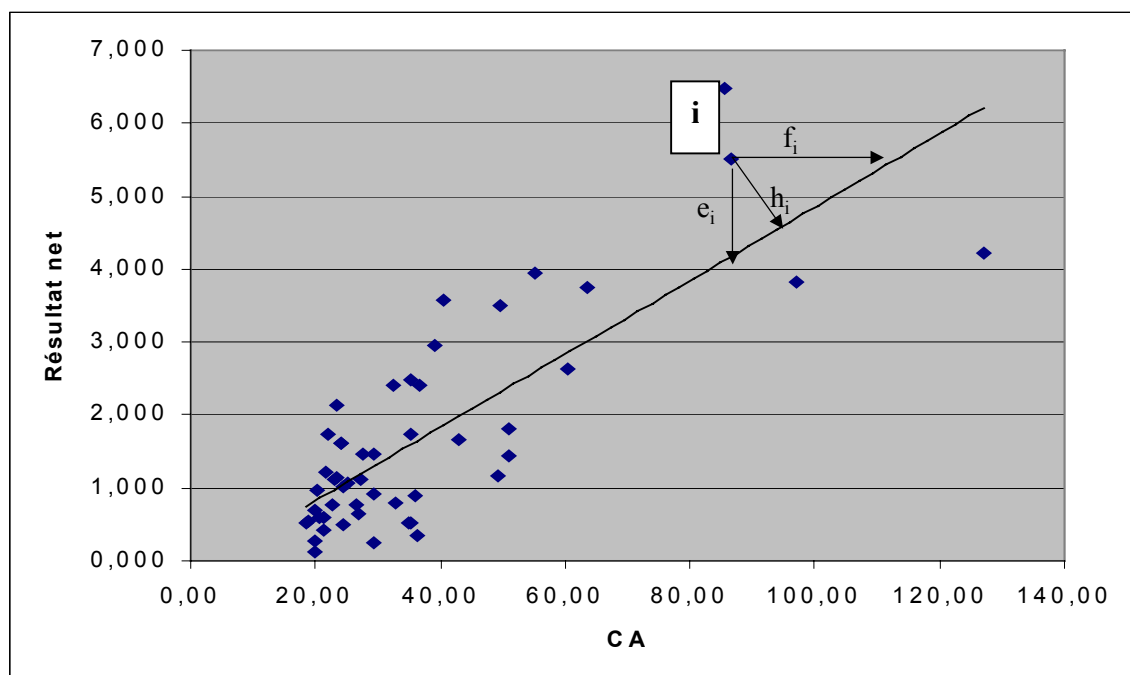
L'ajustement d'un nuage par une droite devient donc la technique de base puisque on l'utilisera aussi bien si la liaison fonctionnelle est linéaire que si elle ne l'est pas puisqu'on sait alors s'y ramener par une transformation. A ce titre la transformation logarithmique est la plus employée, car elle permet de couvrir 4 grands types de liaisons :

Fonction	Expression analytique	Forme linéaire
linéaire	$Y = ax + b$	$Y = ax + b$
Semi log	$Y = a (\text{Log}x) + b$	$Y = a (\text{Log}x) + b$
Exponentielle	$Y = \text{Exp}(ax+b)$	$\text{Log } Y = ax + b$
Puissance	$Y = \text{Exp}(b) x^a$	$\text{Log } Y = a \text{ Log } x + b$

7.1.6 Choix d'un critère d'ajustement : les moindres carrés

Nous sommes donc ramené à la question suivante : trouver la droite *la plus proche* d'un nuage de points. Soit encore, si la droite a pour équation $y = ax + b$ et si les points sont de coordonnées (x_i, y_i) , quel critère d'optimisation devons nous appliquer aux écarts entre les points et la droite pour choisir celle-ci ?

a) Pour répondre à cette question, il faut d'abord choisir une façon de mesurer les écarts entre *un point* et la droite. Dans l'exemple ci-dessous, cette distance pour le point i peut être mesurée de trois façons différentes :



- en choisissant la distance verticale e_i nous nous intéressons à l'écart entre valeur observée et valeur théorique de y pour une valeur donnée x_i .
- en choisissant la distance horizontale f_i nous nous intéressons à l'écart entre valeur observée et valeur théorique de x pour une valeur donnée y_i .
- en choisissant la distance h_i menée perpendiculairement à la droite, nous prenons la plus petite des distances, mais qui n'a pas d'interprétation concrète, et pas d'unité claire si X et Y ont des unités différentes.

Cette dernière approche nous conduirait à une technique d'analyse factorielle qui sera abordée dans une autre leçon. Choisir entre les deux premières solutions, c'est choisir quelle sera la variable explicative (dite exogène) et quelle sera la variable expliquée (dite endogène). Supposons que nous souhaitions expliquer y par x , alors c'est l'écart vertical e_i entre valeur observée et valeur expliquée de y qu'il faut chercher à réduire. Continuons avec cette hypothèse.

b) Nous voici maintenant avec n points et n distances e_i dont il faut minimiser une certaine fonction. Laquelle ? Les mathématiciens qui ont travaillé sur la théorie des erreurs entre 1750 et 1820 se sont cassé les dents sur ce problème. Ils ont proposé plusieurs solutions différentes, ayant chacune des propriétés intéressantes, dont celles-ci :

- Faire deux groupes de données et prendre la droite qui joint les deux points moyens (Mayer, 1750), ce qui revient à prendre la droite qui annule $\sum e_i$ sur deux sous-groupes.
- Prendre la droite qui passe par le point moyen ($\sum e_i = 0$) et qui a pour pente la moyenne des $n(n-1)/2$ paires de points du nuage (Boscovich 1755).
- Prendre la droite qui passe par le point moyen du nuage et qui minimise la somme des valeurs absolues des e_i (Boscovich 1770, Laplace 1793).
- Prendre la droite qui maximise la vraisemblance de l'échantillon observé (x_i, y_i) (D. Bernoulli 1778).
- Prendre la droite qui minimise l'erreur e_i maximale (Laplace 1786).
- Prendre la droite qui minimise la somme des carrés des e_i (Legendre 1805).

Pour des raisons techniques et historiques, en particulier ses liens avec le choix de la moyenne et de la loi normale, c'est cette méthode qui l'a emporté au début du XIX^{ème} siècle et s'est imposée après les travaux de Gauss et Laplace sous le nom de *droite des moindres carrés*.

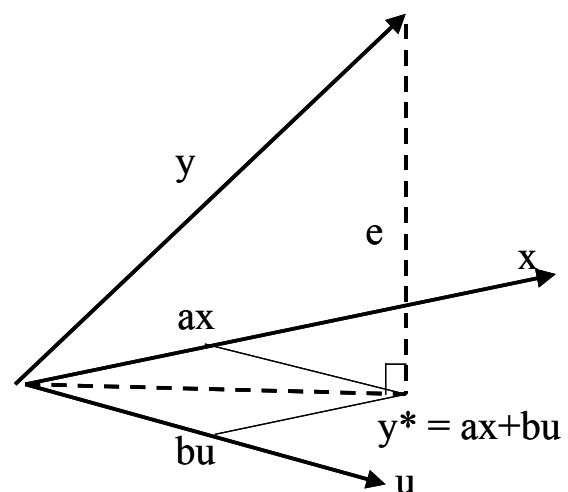
Minimiser la somme des carrés des écarts conduit par dérivation à un système de deux équations dites « normales » à deux inconnues a et b que l'on résout facilement :

$$\text{Min} \sum_i e_i^2 = \text{Min} \sum_i (y_i - ax_i - b)^2$$

$$\Leftrightarrow \begin{cases} \sum_i (y_i - ax_i - b) = 0 \\ \sum_i (y_i - ax_i - b)x_i = 0 \end{cases}$$

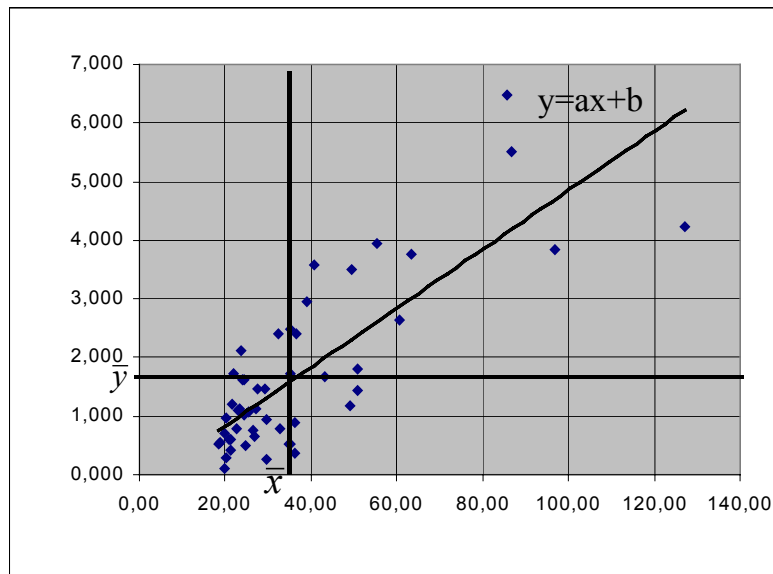
$$\Leftrightarrow \begin{cases} \sum_i y_i - a \sum_i x_i - nb = 0 \\ \sum_i (y_i - \bar{y})(x_i - \bar{x}) = a \sum_i (x_i - \bar{x})^2 \end{cases}$$

$$\Leftrightarrow \begin{cases} \bar{y} = a\bar{x} + b \Leftrightarrow \\ a = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{\text{var}(x)} \end{cases}$$



Notons que cette solution est encore plus évidente d'un point de vue géométrique : pour y , x , et u trois vecteurs donnés à n composantes, quelle est la combinaison linéaire de x et de u (que l'on peut écrire $ax+bu$) qui est la plus proche de y ? C'est la projection orthogonale $y^* = ax+bu$ de y sur le plan (x,u) . En écrivant que le

vecteur $e = y - y^*$ (dont les composantes sont les écarts e_i) est orthogonal à x et à u (produits scalaires nuls) nous retrouvons les deux équations normales et leurs solutions a et b .



La première équation indique que la droite des moindres carrés passe par le point moyen du nuage de coordonnées (\bar{x}, \bar{y}) .

La seconde équation donne la pente a de la droite des moindres carrés. La covariance qui apparaît au numérateur de la formule est une moyenne des produits des écarts de x et y à leur moyenne. Pour chaque point i , la quantité $(x_i - \bar{x})(y_i - \bar{y})$ est positive

si le point i est dans le quadrant Nord-Est ou Sud-Ouest du graphique ou négative si le point i est dans l'un des deux autres quadrants. La covariance apparaît comme un bilan de ces quantités : elle est positive, et donc la pente de la droite l'est aussi, si les points des quadrants NE et SW l'emportent.

7.1.7 Interprétation de Galton : la régression

Vers 1880, Francis Galton, qui ne connaissait rien des travaux de Gauss et Laplace, eut une autre approche : étudiant les tailles des enfants (y) en fonction des tailles moyennes des deux parents (x) il s'intéressa aux tailles médianes des enfants nés de parents de même taille. Découpant le nuage de points en bandes (classes de valeurs de x), il prit la médiane des valeurs de y dans chaque bande et fit le constat que ces médianes étaient des fonctions linéaires de x . Plus tard, il transposa cette découverte aux moyennes liées de la forme $\{\bar{y}_i = \bar{y} / (x = x_i) = f(x_i)\}$ qui caractérisent les sous population de même valeur de x . Les fonctions f sont en général quelconques et prennent le nom de *courbes de régression* de y en x . Il reconnut ensuite que l'on pouvait s'intéresser de façon symétrique aux moyennes liées de x en fonction de y et obtenir une courbe de régression de x en y . Pour des observations extraites d'une distribution bi-normale, Galton retrouvait un résultat de Laplace et Bravais, à savoir que ces fonctions étaient linéaires. Le terme de *droite de régression* qu'il leur attribuait venait du contexte de ses études sur l'hérédité et l'eugénisme : étudiant comment les fils (ou filles) héritent leur taille de leurs pères et mères, il obtint des tailles moyennes des enfants croissant linéairement avec les tailles moyennes des pères et mères, mais avec une pente inférieure à un, qui traduisait une transmission héréditaire incomplète : les parents grands avaient des enfants grands mais moins grands qu'eux. L'hérédité était contrebalancée par un retour vers les mesures moyennes ou typiques de l'espèce, que Galton baptisa « régression vers la médiocrité ». Karl Pearson (1896) établit les formules de ces droites de régression dans le cas normal, retrouvant la formule de a et b que nous avons donnée.

7.1.8 La synthèse de Yule

L'économiste Yule (1897) a fait une synthèse hybride des deux approches de Laplace-Gauss et de Galton-Pearson, en montrant que, dans le cas normal, et même dans le cas plus général d'une régression linéaire, la droite de régression (lieu des moyennes de y liées par x) de Galton et Pearson et la droite des moindres carrés de Laplace et Gauss coïncident. C'est donc par la méthode des moindres carrés qu'on procède désormais pour obtenir la droite de régression.

7.1.9 Décomposition de la variance.

Reprenons la relation $y_i = ax_i + b + e_i$ dans laquelle e_i désigne l'écart entre y_i observé et $y^* = y_{th}$, donné par la droite théorique d'équation $y_{th} = ax + b$.

Comme $\bar{y} = a\bar{x} + b$, nous obtenons par différence :

$$y_i - \bar{y} = a(x_i - \bar{x}) + e_i$$

Si nous élevons cette égalité au carré et prenons la moyenne, nous obtenons :

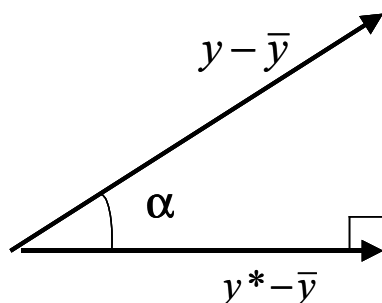
$$\sum_i (y_i - \bar{y})^2 = a^2 \sum_i (x_i - \bar{x})^2 + \sum_i e_i^2 + 2 \sum_i (x_i - \bar{x})e_i$$

Mais le dernier terme est nul (équations normales) donc :

$$\sum_i (y_i - \bar{y})^2 / n = a^2 \sum_i (x_i - \bar{x})^2 / n + \sum_i e_i^2 / n$$

$$\Leftrightarrow \|y - \bar{y}\|^2 = a^2 \|x - \bar{x}\|^2 + \|e\|^2$$

\Leftrightarrow Variance totale = variance expliquée par x + variance résiduelle.



Cette relation est dite *décomposition de la variance de y* . En effet elle montre que la variance de y est faite de deux parties : une partie théorique due à l'ajustement de y par $ax+b$ et une partie résiduelle. Cette égalité peut aussi être comprise géométriquement comme une relation de Pythagore dans le triangle rectangle des 3 vecteurs $\{y - \bar{y}, y^* - \bar{y}, e\}$: carré de l'hypoténuse = somme des carrés des côtés de l'angle droit.

Il est tentant d'évaluer le rapport « variance expliquée/variance totale » :

$$\frac{V_{expl}}{V_{tot}} = \frac{\|y^*\|^2}{\|y\|^2} = \frac{a^2 \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} = \frac{a^2 \text{var } x}{\text{var } y} = \frac{\text{cov}^2(x, y)}{\text{var } x \cdot \text{var } y} = \cos^2 \alpha = r^2$$

Cette quantité r^2 qui varie entre 0 et 1, comme une part, mesure la qualité de l'ajustement théorique. Une valeur de 1 ne sera possible que si la variance résiduelle est nulle, c'est à dire tous les points du nuage sur la droite ajustée. La valeur 0 correspond à l'indépendance entre x et y : droite de régression horizontale ($a=0$).

7.1.10 Corrélation linéaire

La quantité $r = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$ (dont le carré r^2 s'appelle coefficient de détermination)

est elle même une mesure d'intensité de la relation linéaire entre les variables x et y , que l'on appelle *coefficient de corrélation linéaire* entre x et y (dit aussi coefficient de Bravais-Pearson).

$$r = -1$$

Liaison décroiss. Corrél. <0 Inc

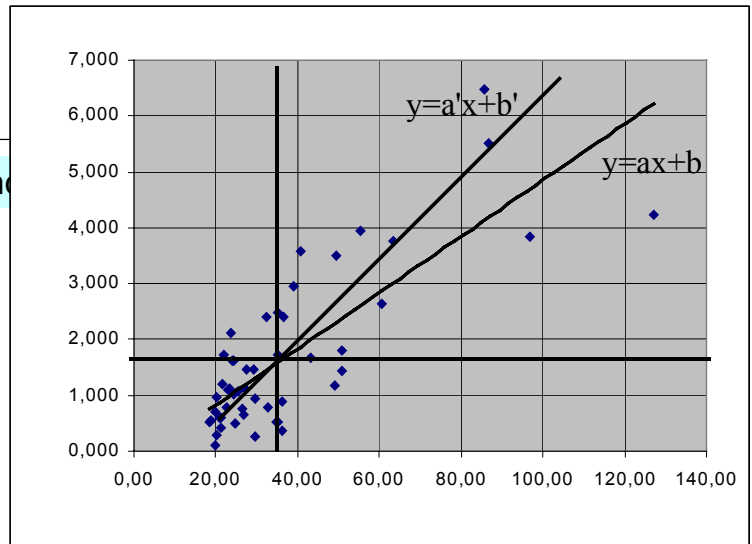
Cette quantité varie de -1 à 1 (et son carré de 0 à 1) :

Rappelons que l'on peut construire deux droites de régression de y en x et de x en y , à partir des moyennes liées de y pour x fixé (et respectivement de x pour y fixé), ou encore en minimisant la somme des e_i^2 (respectivement des f_i^2). Soient $y = ax + b$ et $y = a'x + b'$ les équations de ces deux droites, qui passent toutes les deux par le point moyen. On peut vérifier que :

$$a = r \frac{\sigma_y}{\sigma_x}, a' = r \frac{\sigma_x}{\sigma_y},$$

$$\text{et } aa' = r^2$$

La mesure de corrélation r a le même signe que a , la pente de la droite de régression de y en x , et que a' , la pente de la régression de x en y . Si la corrélation est proche de 1 les deux droites de régression sont confondues. Si elle est proche de 0 elles sont orthogonales.



7.2 Généralisation à plusieurs variables

Nous pouvons généraliser les relations obtenues à plusieurs variables « explicatives » X_1, X_2, \dots, X_p s. Mais il est alors nécessaire de recourir au calcul matriciel. Pour ceux que cela rebuterait, il suffira de savoir interpréter correctement les résultats fournis par le tableur. Dans le cas de p variables explicatives, le modèle théorique prend la forme :

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \dots + a_p x_{ip} + e_i = y_{ith} + e_i$$

$$\mathbf{y} = \mathbf{a}_1 \mathbf{x}_1 + \mathbf{a}_2 \mathbf{x}_2 + \dots + \mathbf{a}_p \mathbf{x}_p + \mathbf{e} \quad (\text{écriture vectorielle})$$

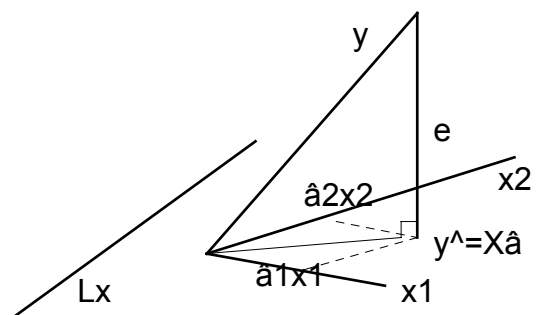
$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e} \quad (\text{écriture matricielle})$$

dans lequel \mathbf{X} est la matrice $[n,p]$ des vecteurs $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ correspondant aux variables explicatives, \mathbf{y} le vecteur $[n,1]$ des valeurs de la variable à expliquer, \mathbf{e} le vecteur $[n,1]$ des résidus, et \mathbf{a} le vecteur $[p,1]$ des coefficients :

$$\begin{array}{c} \begin{bmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_n \end{bmatrix} \\ [n,1] \end{array} = \begin{array}{c} \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \\ [n,p] \end{array} \begin{array}{c} \begin{bmatrix} a_1 \\ \dots \\ a_j \\ \dots \\ a_p \end{bmatrix} \\ [p,1] \end{array} + \begin{array}{c} \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_i \\ \dots \\ \varepsilon_n \end{bmatrix} \\ [n,1] \end{array}$$

Le cas particulier $\mathbf{x}_p = \mathbf{u}$, vecteur unitaire, correspond à un modèle avec constante dont un cas particulier est, pour $p=2$, la droite de régression que nous venons d'étudier. On peut, dans ce cas, se ramener à un modèle linéaire en variables centrées à $(p-1)$ variables explicatives.

Que l'on envisage alors une régression linéaire sous la forme d'une moyenne \bar{y}_i liée par les x_{ij} qui soit une fonction linéaire de ces p valeurs de x_{ij} , que l'on parle d'un ajustement par les moindres carrés de l'hyperplan y_{ith} sur le nuage des n points dans R^p , ou que l'on préfère parler de la projection du vecteur y_{th} sur le sous espace $(x_1 \dots x_p)$ de R^n revient encore au même.



Les équations normales s'écriront (avec $\langle x, y \rangle$ = produit scalaire de x et de y) :

$$\begin{aligned}
\begin{cases} \langle \mathbf{x}_1, \mathbf{e} \rangle = 0 \\ \dots \\ \langle \mathbf{x}_j, \mathbf{e} \rangle = 0 \\ \dots \\ \langle \mathbf{x}_p, \mathbf{e} \rangle = 0 \end{cases} &\Leftrightarrow \begin{cases} \langle \mathbf{x}_1, \mathbf{y} - \sum_j \hat{a}_j \mathbf{x}_j \rangle = 0 \\ \dots \\ \langle \mathbf{x}_j, \mathbf{y} - \sum_j \hat{a}_j \mathbf{x}_j \rangle = 0 \\ \dots \\ \langle \mathbf{x}_p, \mathbf{y} - \sum_j \hat{a}_j \mathbf{x}_j \rangle = 0 \end{cases} \Leftrightarrow \begin{cases} \langle \mathbf{x}_1, \mathbf{y} \rangle - \sum_j \hat{a}_j \langle \mathbf{x}_1, \mathbf{x}_j \rangle = 0 \\ \dots \\ \langle \mathbf{x}_j, \mathbf{y} \rangle - \sum_j \hat{a}_j \langle \mathbf{x}_j, \mathbf{x}_j \rangle = 0 \\ \dots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle - \sum_j \hat{a}_j \langle \mathbf{x}_p, \mathbf{x}_j \rangle = 0 \end{cases} \\
&\Leftrightarrow \mathbf{X}'\mathbf{e} = \mathbf{0} \quad \Leftrightarrow \mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}) = \mathbf{0} \quad \Leftrightarrow \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\mathbf{a}} = \mathbf{0}
\end{aligned}$$

Equation dans laquelle $\hat{\mathbf{a}}$ est le vecteur solution des moindres carrés pour \mathbf{a} .

On voit que l'équation numéro j contient tous les coefficients inconnus \hat{a}_j et les produits scalaires de \mathbf{x}_j associée à toutes les autres variables. La matrice $\mathbf{X}'\mathbf{X}$ de ces produits scalaires est en fait la matrice $[p, p]$ symétrique de tous les moments d'ordre deux non centrés des p variables exogènes. La matrice $\mathbf{X}'\mathbf{y}$ est la matrice $[p, 1]$ de ces mêmes moments pour les p couples $\mathbf{x}_j\mathbf{y}$.

Ce système de p équations à p inconnues possèdera une et une seule solution (un vecteur $\hat{\mathbf{a}}$) si le déterminant de la matrice $\mathbf{X}'\mathbf{X}$ est non nul. Il suffit pour cela que La matrice \mathbf{X} soit de rang p , c'est à dire régulière. La solution en $\hat{\mathbf{a}}$ de ce système des équations normales s'obtient en multipliant la dernière équation à gauche par $(\mathbf{X}'\mathbf{X})^{-1}$:

$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
[p,1] [p,p] [p,1]

Dans le cas particulier du modèle avec constante (si $\mathbf{x}_p = \mathbf{u}$) nous pouvons appliquer les formules précédentes aux variables centrées. Les matrices $\mathbf{X}'\mathbf{X}$ et $\mathbf{X}'\mathbf{y}$ sont alors à remplacer par \mathbf{V}_{xx} et \mathbf{V}_{xy} qui désignent les matrices des moments centrés (variances-covariances). La formule ci-dessous donne alors les estimations des coefficients par les moindres carrés :

$ \begin{cases} \hat{\mathbf{a}}_c = \mathbf{V}_{xx}^{-1}\mathbf{V}_{xy} \\ \hat{a}_p = \bar{y} - \sum_i \hat{a}_i \bar{x}_i \end{cases} $
--

Nous retrouvons le triangle rectangle des 3 vecteurs $\{\mathbf{y} - \bar{y}, \mathbf{y}_{th} - \bar{y}, \mathbf{e}\}$ qui permet d'écrire la *décomposition de la variance totale* :

Variance totale = variance expliquée par x + variance résiduelle

$$\|\mathbf{y} - \bar{y}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\|^2 + \|\mathbf{e}\|^2$$

Mais cette fois-ci, la variance expliquée s'entend comme expliquée simultanément par les p variables X_1, X_2, \dots, X_p .

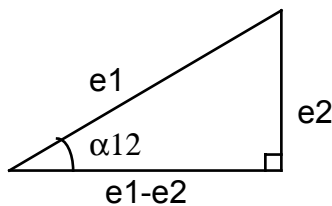
Et le rapport variance expliquée/variance totale est le carré du coefficient de *corrélation multiple* ::

$$R^2 = \frac{\|\hat{\mathbf{y}}_c\|^2}{\|\mathbf{y}_c\|^2} = \frac{\hat{\mathbf{a}}'\mathbf{V}_{xy}}{\mathbf{V}_{yy}} = \cos^2 \alpha$$

Le coefficient de corrélation multiple $R = \cos \alpha$ mesure la corrélation linéaire entre y (observé) et y (théorique) égal à $a_1x_1 + a_2x_2 + \dots + a_px_p$.

7.3 Rendement et corrélation partielle

Quand nous passons d'un modèle [M1] à r variables à un modèle [M2] à $p = r+s$ variables, nous avons ajouté s variables explicatives. On peut d'abord montrer que l'on y gagne toujours au sens où la part de la variance expliquée (R^2) augmente forcément, quelles que soient les variables rajoutées. Mais d'une façon qui n'est pas toujours significative. Le *rendement* $\eta_{1/2}$ est défini comme un gain relatif sur la variance de la variable endogène, quand nous passons de l'un à l'autre des modèles. Si nous appelons \mathbf{e}_1 et \mathbf{e}_2 les vecteurs des résidus de [M1] et [M2] nous avons :



$$\eta_{2/1} = \cos^2 \alpha_{12} = \frac{\|\mathbf{e}_1 - \mathbf{e}_2\|^2}{\|\mathbf{e}_1\|^2} = \frac{\|\mathbf{e}_1\|^2 - \|\mathbf{e}_2\|^2}{\|\mathbf{e}_1\|^2}$$

$$\eta_{2/1} = \frac{R_2^2 - R_1^2}{1 - R_1^2} \Leftrightarrow 1 - \eta_{2/1} = \frac{1 - R_2^2}{1 - R_1^2}$$

Dans le cas particulier où l'on passe du modèle [M1] au modèle [M2] par l'ajout d'une seule variable \mathbf{x}_p ($s=1$), ce rendement apparaît comme le carré d'un *coefficient de corrélation partielle* entre \mathbf{y} et \mathbf{x}_p , que Yule proposa le premier sous le nom de *coefficient de corrélation nette*, parce qu'il élimine les effets sur \mathbf{y} et \mathbf{x}_p des autres variables exogènes $\mathbf{x}_1, \dots, \mathbf{x}_{p-1}$, et s'interprète comme une liaison entre \mathbf{y} et \mathbf{x}_p *toutes choses égales par ailleurs*. En effet il s'obtient également par le calcul du coefficient de corrélation entre les résidus de la régression de \mathbf{y} et \mathbf{x}_p sur $\mathbf{x}_1, \dots, \mathbf{x}_{p-1}$. En réécrivant \mathbf{x}_0 la variable \mathbf{y} , ce coefficient est alors noté $r_{0p.12\dots(p-1)}$.

$$r_{0p.12\dots(p-1)} = \sqrt{\eta_{2/1}} = \cos(\mathbf{e}_1, \mathbf{e}_1 - \mathbf{e}_2) = \cos(\mathbf{e}_1, \mathbf{e}_{21})$$

Dans le cas très usuel des modèles avec constante, le calcul du coefficient de corrélation partielle ne nécessite pas cependant de faire la régression de \mathbf{x}_p sur $\mathbf{x}_1, \dots, \mathbf{x}_p$ mais peut se déduire des corrélations multiples des modèles [M1] et [M2] par la relation :

$$(1 - R_{0.1\dots p}^2) = (1 - R_{0.1\dots p-1}^2) (1 - r_{0p.12\dots(p-1)}^2)$$

par exemple pour $p = 2$:

$$(1 - R_{0.12}^2) = (1 - R_{0.1}^2) (1 - r_{02.1}^2)$$

Traduction : ce qui n'est pas expliqué par x_1 et x_2 c'est ce qui n'est pas expliqué par x_1 que multiplie ce qui n'est pas expliqué par x_2 dans ce qu'il restait à expliquer !

7.4 Interprétation probabiliste

7.4.1 Modèle linéaire à erreurs aléatoires

Les estimations des coefficients du modèle linéaire que fournit la méthode des moindres carrés ne sont que des estimations et pas les valeurs de la liaison entre x et y dans la population. Si nous recommençons à tirer 20 logements dans le fichier de la Chambre des Notaires, nous aurons un autre échantillon et d'autres valeurs de a , le prix marginal du mètre carré obtenu en expliquant le prix par la surface. Il convient donc de considérer l'estimateur des moindres carrés de a comme une variable aléatoire, et de se soucier de sa variance.

Si dans le cas de la régression simple (une seule variable explicative) on écrit le modèle théorique :

$$Y_i = ax_i + b + \varepsilon_i$$

dans lequel ε_i est l'erreur aléatoire sur la i ème observation, alors il faudra distinguer la vraie valeur de a et son estimateur noté \hat{a} (ou a^*) ainsi que b et son estimateur b^* . Il faudra distinguer le modèle théorique que l'on vient d'écrire du modèle ajusté (estimé) :

$$Y_i = a^*x_i + b^* + e_i$$

dans lequel e_i est l'écart ou résidu entre y_i observé et y_i^* ajusté dont nous avons minimisé la somme des carrés.

Sur les variables aléatoires ε_i on fait en général les hypothèses suivantes :

1. $E(\varepsilon_i) = 0$: l'erreur est parfois négative, parfois positive et centrée sur 0.
2. $V(\varepsilon_i) = \text{constante} = \sigma^2$: la variance de l'erreur est la même pour toutes les observations.
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$: les erreurs pour deux observations i et j ne sont pas corrélées.

Les principales conséquences sont alors les suivantes :

a) σ^2 est en général inconnu mais peut être estimé par $\hat{\sigma}^2 = \frac{\sum_i e_i^2}{n-p} = \frac{n.V_{\text{res}}}{n-p}$, p étant le nombre de variables explicatives (y compris la constante, donc $p=2$ pour le modèle simple).

b) l'estimateur de a a pour variance :

$V(\hat{a}) = \frac{\sigma^2}{nV(x)}$ et peut être estimée par $V(\hat{a}) = \frac{\hat{\sigma}^2}{nV(x)}$ dans le cas du modèle linéaire simple.

$V(\hat{a}) = \frac{\sigma^2}{n} V_{xx}^{-1}$ est la matrice des variances et covariances des coefficients estimés dans le cas d'un modèle à p variables explicatives.

Il n'est pas nécessaire de retenir ces formules, le tableur fournissant directement ces valeurs. Le calcul de ces variances permet d'assortir les estimations numériques d'une mesure de précision. Il serait encore mieux de proposer des intervalles de confiance ou de pouvoir tester certaines valeurs. Mais pour cela nous avons besoin d'une hypothèse supplémentaire.

7.4.2 Hypothèse de normalité des erreurs

L'hypothèse la plus classique faite sur les erreurs ε_i et qui se rajoute aux trois précédentes consiste à dire qu'elles suivent une loi normale. Cela est justifié parfois par la théorie des erreurs d'observation, parfois par le théorème central limite (additions de nombreuses petites erreurs élémentaires), parfois par rien du tout !

Nous posons donc : $\varepsilon_i \rightarrow \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \Leftrightarrow \boldsymbol{\varepsilon} \rightarrow \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$: le vecteur $\boldsymbol{\varepsilon}$ suit une loi normale à n dimensions. Cela permet d'en déduire quelques conséquences utiles :

$\Rightarrow \mathbf{y} \rightarrow \mathbf{N}(\mathbf{X}\mathbf{a}, \sigma^2 \mathbf{I})$: le vecteur des observations \mathbf{y} suit une loi normale.

$\Rightarrow \hat{\mathbf{a}} \rightarrow \mathbf{N}_p(\mathbf{a}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$: le vecteur des estimateurs des coefficients suit une loi normale.

$\Rightarrow (\hat{a}_j - a_j) / \hat{\sigma}_{\hat{a}_j} \rightarrow \mathbf{N}(\mathbf{0}, 1)$: l'estimateur de chaque coefficient suit une loi normale.

$\Rightarrow T = (\hat{a}_j - a_j) / \hat{\sigma}_{\hat{a}_j} \rightarrow \text{Student } (n-p)$.

Chacun des estimateurs des coefficients, s'il est centré par la vraie valeur, et réduit par l'estimation de son écart-type suit une loi de Student. Le fait de remplacer la vraie valeur de cet écart-type par son estimation nous a fait passer d'une loi normale à la loi de Student (pseudonyme de William S. Gosset, un ingénieur des brasseries Guinness, qui établit en 1908 cette distribution pour la moyenne et pour le coefficient de corrélation linéaire).

7.4.3 Test de significativité d'un coefficient

Cette dernière propriété permet de construire un intervalle de confiance pour la vraie valeur d'un coefficient ou d'une combinaison linéaire des coefficients. Dans le premier cas on obtient :

$$\alpha = \text{prob}(-t_\alpha < T = (\hat{a}_j - a_j) / \hat{\sigma}_{\hat{a}_j} < t_\alpha)$$

$$\Leftrightarrow \alpha = \text{prob}(\hat{a}_j - t_\alpha \hat{\sigma}_{\hat{a}_j} < a_j < \hat{a}_j + t_\alpha \hat{\sigma}_{\hat{a}_j})$$

Le choix d'un niveau de confiance pour α (90% à 99%) conduit à une valeur de t_α lue dans la table de Student ($n-p$) et donc, lorsque $\hat{\sigma}_{\hat{a}_j}$ a été calculé, à un encadrement de la vraie valeur inconnue de a_j entre deux estimations numériques, par un intervalle dont le centre est l'estimation ponctuelle \hat{a}_j .

Les mêmes formules peuvent être utilisées pour tester une hypothèse sur un coefficient. Dans ce dernier cas soit $\mathbf{H}_0 : a_j = a_0$ cette hypothèse.

Sous cette hypothèse $T_0 = (\hat{a}_j - a_0) / \hat{\sigma}_{\hat{a}_j}$ suit une loi de Student ($n-p$) et si nous supposons comme hypothèse alternative $\mathbf{H}_1 : a_j \neq a_0$, la région critique du test, celle dans laquelle on rejette H_0 , est de la forme $|T_0| > t_\alpha$. Le choix d'une valeur (en général 5%) pour le risque de première espèce $\alpha = \text{prob}(|T_0| > t_\alpha / H_0)$ permet de calculer les valeurs critiques $-t$ et $+t$, et d'appliquer la règle de décision :

$$|T_0|(\text{calculé}) > t_\alpha \Rightarrow \text{On décide que } H_0 \text{ est vraie.}$$

L'extrait suivant de la table de t_α pour $\alpha = 5\%$ montre que l'on peut bien souvent prendre en première approximation $t_\alpha = 2$ dès que l'échantillon a plus de 10 observations.

$n-p$	1	5	10	12	15	20	25	30	100
$\alpha = 5\%$	12,7	2,57	2,23	2,18	2,13	2,09	2,06	2,04	1,98
$\alpha = 1\%$	63,66	4,03	3,17	3,06	2,95	2,85	2,79	2,75	2,62

Les économètres préfèrent donner la *probabilité critique* du test, c'est à dire la probabilité que la valeur observée de la statistique (ici T_0) soit dépassée, quitte à la comparer ensuite à un risque choisi. Une *faible* probabilité critique (par ex. $<2\%$) nous conduit à rejeter H_0 , donc à considérer que a_0 n'est pas compatible avec les observations.

La valeur a_0 que l'on teste systématiquement est 0. En effet cette nullité du coefficient signifie que la variable qui lui est associée n'a aucun pouvoir explicatif sur la variable endogène (à expliquer), du moins dans le modèle étudié. La spécification de ce modèle est donc à revoir en éliminant ou modifiant cette variable. Le test s'appelle alors un test de significativité de a_j : en effet il indique si la valeur estimée est *significativement différente de zéro* compte tenu de l'écart-type de l'estimateur. Puisque $a_0=0$, la statistique du test est dans ce cas $T_0 = \hat{a}_j / \hat{\sigma}_{\hat{a}_j}$ et il suffit de la comparer à la valeur critique t_{α} pour prendre la décision.

7.4.4 Test de significativité globale de la régression

Dans un modèle à plusieurs variables explicatives, il est intéressant de compléter le test de significativité de chaque coefficient par un test de significativité globale de la régression qui répond à la question suivante : est-ce que la part de variance expliquée globalement par les p variables est significative (H_1) ou est-elle due au hasard (H_0) ? On utilise la statistique de Fisher qui suit une loi du même nom :

$$F = \frac{(\mathbf{y}'\mathbf{y} - \mathbf{e}'\mathbf{e})/(p-1)}{\mathbf{e}'\mathbf{e}/(n-p)} = \frac{(R_2^2)/(p-1)}{(1-R_2^2)/(n-p)} \rightarrow F(p-1, n-p)$$

C'est cette valeur de F qui est calculée et affichée par les logiciels, et qui permet par comparaison avec les valeurs critiques d'une table à 5% ou 1%, de tester l'hypothèse H_0 qui signifie alors la nullité de *tous* les coefficients du modèle, c'est à dire le caractère non explicatif de l'ensemble des variables exogènes de ce modèle, ou encore la non significativité de R^2 .

Dans le cas de la régression simple ($p=2$), ce test se confond avec un test de validité du coefficient de corrélation ordinaire entre \mathbf{y} et \mathbf{x} , la seule variable exogène du modèle. Et il est équivalent au test de Student du seul coefficient a , car $F(1, n-p) = T^2(n-p)$. Un des deux tests suffit donc dans ce cas.

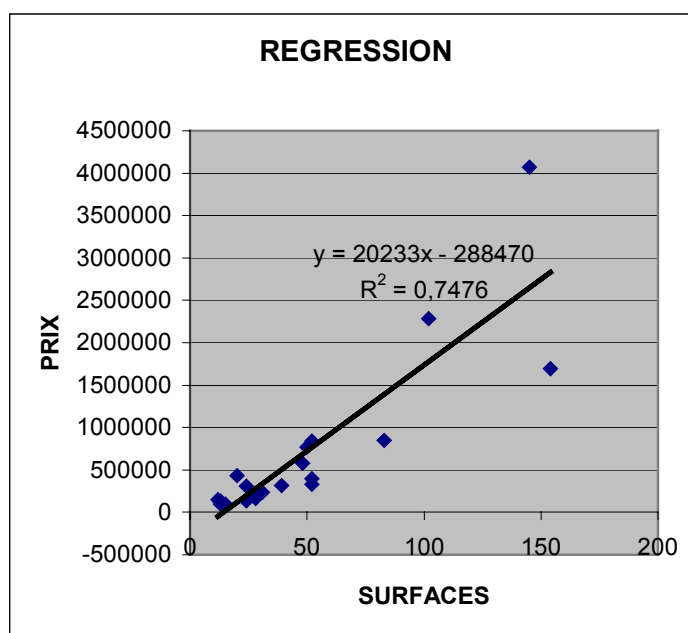
7.5 Exemple

Prenons dans le fichier *logement* les prix et surfaces des 19 premiers logements. Et cherchons à expliquer linéairement les prix par les surfaces.

Dans *Excel*, la fonction *Droitereg* fournit les valeurs des coefficients de la droite de régression. La représentation graphique de la série double est obtenue sous forme de « nuage de points » (*scatterdiagram*). Dans le menu *Graphique* la commande « Ajouter une courbe de tendance » permet d'obtenir la droite de régression et son équation. L'utilitaire d'analyse « Régression linéaire » donne des résultats plus détaillés. Nous représentons les principaux.

L'interprétation des résultats est la suivante : l'équation de régression étant de la forme $\text{PRIX} = 20\,232 \text{ SURF} - 288\,470$, le prix marginal du m^2 est de 20 232F. Le prix moyen du m^2 est croissant avec la taille du logement puisque l'ordonnée à l'origine est négative. Il serait stupide d'interpréter celle-ci comme le prix d'un logement de 0m^2 . D'ailleurs l'équation de la droite ajustée n'est valide que dans l'intervalle des données observées (entre 12 et 154 m^2). De plus l'équation de la droite serait fort différente si on retirait de l'échantillon les deux appartements très exceptionnels de 145 et 154 m^2 . Ce coefficient 20 232 est significativement différent de zéro puisque la statistique de Student est égale à 7. Mais la qualité de l'ajustement n'est que moyenne puisque la part de variance expliquées est $R^2 = 0,75$. Bien d'autres facteurs expliquent le prix d'adjudication (quartier, étage, état, ensoleillement, ascenseur...cf étude de cas).

Fichier Logements		
SURF	PRIX	
50	764600	
154	1696156	
15	94140	
24	137616	
83	848841	
52	396708	
28	162456	
13	134979	
39	317421	
12	145104	
48	575424	
20	434600	
145	4071165	
24	308832	
13	100178	
52	334152	
102	2281638	
52	839852	
31	237801	



	Coefficients	Erreur-type	Statistique t	Probabilité
Constante	-288470,1643	185554,3075	-1,554640085	0,13845045
SURF	20232,59783	2851,132156	7,096338128	1,7929E-06

ANALYSE DE VARIANCE

	Degré de liberté	Somme des carrés	Moyenne des carrés	F
Régression	1	1,3211E+13	1,3211E+13	50,3580148
Résidus	17	4,4598E+12	2,62341E+11	
Total	18	1,76708E+13		