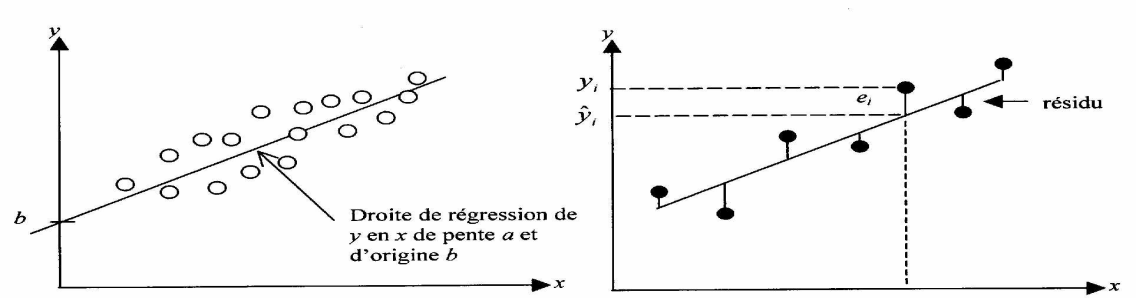
Statistiques : notes du cours 1
Régression linéaire

Une situation courante dans les applications est d'avoir à sa disposition deux ensembles de données de taille n, $\{y_1, y_2, \ldots, y_n\}$ et $\{x_1, x_2, \ldots, x_n\}$, obtenus expérimentalement ou mesurés sur une population. Le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x et les y, par exemple de la forme $y = f(x)$. Lorsque la relation recherchée est affine, c'est-à-dire lorsqu'elle peut prendre la forme $y = ax + b$, on parle de *régression linéaire*. Mais même si une telle relation est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement. Pour tenir compte dans le modèle mathématique des erreurs observées, on considère les données $\{y_1, y_2, \ldots, y_n\}$ comme autant de réalisations d'une variable aléatoire Y et parfois aussi les données $\{x_1, x_2, \ldots, x_n\}$ comme autant de réalisations d'une variable aléatoire X. On dit que la variable Y est la *variable dépendante* ou *variable expliquée* et que la variable X est la *variable explicative*.

La droite des moindres carrés : Les données $\{(x_i, y_i), i = 1, \ldots, n\}$ peuvent être représentées par un nuage de n points dans le plan (x, y), appelé parfois le *diagramme de dispersion*. Le *centre de gravité* de ce nuage peut se calculer facilement : il s'agit du point de coordonnées $(\overline{x}, \overline{y}) = (\frac{1}{n}\Sigma_{i=1}^n x_i, \frac{1}{n}\Sigma_{i=1}^n y_i)$. Rechercher une relation affine entre les variables X et Y revient à rechercher une droite qui s'ajuste le mieux possible à ce nuage de points. Parmi toutes les droites possibles, on retient celle qui jouit de la propriété suivante : celle qui rend minimale la somme des carrés des écarts des valeurs observées y_i à la droite $\hat{y}_i = ax_i + b$. Si ε_i représente cet écart, appelé aussi *résidu*, le principe des *moindres carrés ordinaire* (MCO) consiste à choisir les valeurs de a et de b qui minimisent

$$E = \sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n} (y_i - (ax_i + b))^2.$$



Nous allons voir que ces valeurs, notées $\hat{a}$ et $\hat{b}$, sont égales à

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = cov_{xy}/s_x^2 \tag{1}$$

$$\hat{b} = \overline{y} - \hat{a}\overline{x} \tag{2}$$

où $s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$ et $cov_{xy} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$ sont respectivement des estimateurs de la *variance* et de la *covariance* des variables aléatoires X et Y.

Calcul des estimateurs $\hat{a}$ et $\hat{b}$ de la pente et de l'ordonnée à l'origine. Notons tout d'abord que les lois des variables aléatoires X et Y étant inconnues, on estime leur espérance $\overline{x}$ et $\overline{y}$ ainsi que leur variance et covariance en mettant la probabilité $\frac{1}{n}$ à chacune de leurs valeurs.

Notons aussi les formules suivantes, souvent utiles dans les calculs et qui se vérifient facilement :

Lemme 1 *Pour toutes familles $\{x_1, x_2, \ldots, x_n\}$ et $\{y_1, y_2, \ldots, y_n\}$, on a :*

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$$

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

Pour calculer les valeurs de a et de b qui minimisent la somme des carrées des résidus $E(a,b) = \sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n}(y_i - (ax_i + b))^2$, il suffit de résoudre le système linéaire

$$\left\{ \begin{array}{rcl} \frac{\partial E}{\partial a} & = & 0 \\ \frac{\partial E}{\partial b} & = & 0 \end{array} \right. \tag{3}$$

car en son minimum (a,b), la fonction $E(a,b)$ a nécéssairement ses deux dérivées partielles nulles. Ce système s'écrit :

$$\left\{ \begin{array}{rcl} \sum_{i=0}^{n} 2(y_i - (ax_i + b))(-x_i) & = & 0 \\ \sum_{i=0}^{n} 2(y_i - (ax_i + b))(-1) & = & 0 \end{array} \right. \tag{4}$$

soit, en développant,

$$\left\{ \begin{array}{rcl} \sum_{i=0}^{n} x_i y_i + a\sum_{i=0}^{n} x_i^2 - b\sum_{i=0}^{n} x_i & = & 0 \\ -\sum_{i=0}^{n} y_i + a\sum_{i=0}^{n} x_i + nb & = & 0 \end{array} \right. \tag{5}$$

ce qui s'écrit encore

$$\left\{ \begin{array}{rcl} \sum_{i=0}^{n} x_i y_i + a\sum_{i=0}^{n} x_i^2 - (\overline{y} - a\overline{x})\sum_{i=0}^{n} x_i & = & 0 \\ \overline{y} - a\overline{x} & = & b \end{array} \right. \tag{6}$$

d'où les formules (1) et (2).

Evaluation de la qualité de la régression : Pour mesurer la qualité de l'approximation d'un nuage $(x_i, y_i)_{i=1..n}$ par sa droite des moindres carrés (après tout on peut toujours faire passer une droite par n'importe quel nuage !), on calcule son *coefficient de corrélation linéaire* défini par

$$r_{xy} = \frac{cov_{xy}}{s_x s_y}.$$

C'est un nombre compris entre -1 et $+1$, qui vaut $+1$ (resp. -1) si les points du nuage sont exactement alignés sur une droite de pente a positive (resp. négative). Ce coefficient est une *mesure de la dispersion du nuage*. On considère que l'approximation d'un nuage par sa droite des moindres carrés est de bonne qualité lorsque $|r_{xy}|$ est proche de 1 (donc r_{xy} proche de $+1$ ou de -1) et de médiocre qualité lorsque $|r_{xy}|$ est proche de 0.

Parfois on préfère calculer non plus r_{xy} mais son carré noté $R^2 = r_{xy}r_{xy}$ car on a la relation suivante (voir figure) :

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

qui exprime que la dispersion totale de Y (DT) est égale à la dispersion autour de la régression (DA) plus la dispersion due à la régression (DR). Or on peut vérifier que l'on a $R^2 = \frac{DR}{DA}$, c'est-à-dire que le R^2 représente la part de la dispersion totale de Y que l'on peut expliquer par la régression. Ainsi si l'on obtient une valeur de $R^2 = 0,86$ (et donc $r = \mp 0,92$), cela signifie que la modélisation par la droite des moindres carrés explique 86% de la variation totale, ce qui est un très bon résultat.

Cependant, même avec un R^2 excellent (proche de 1), notre modèle linéaire peut encore être rejeté. En effet, pour être assuré que les formules données $\hat{a}$ et $\hat{b}$ fournissent de bonnes estimations de la pente et de l'ordonnée à l'origine de la droite de régression, il est nécessaire que les résidus ε_i soient indépendants et distribués aléatoirement autour de 0. Ces hypothèses ne sont pas forcément faciles à vérifier. Un tracé de des résidus et un examen de leur histogramme permet de détecter une anomalie grossière mais il faut faire appel à des techniques statistiques plus élaborées pour tester réellement ces hypothèses (ce que nous ne ferons pas ici).

Prévisions : Si $y = \hat{a}x + \hat{b}$ est la droite des moindres carrés d'un nuage de points $(x_i, y_i)_{i=1..n}$, on appelle *valeurs* de y *prédites par le modèle* les valeurs $\hat{y}_i$ données par : $\hat{y}_i := \hat{a}x_i + \hat{b}$.

On utilise notamment ces valeurs pour faire des prévisions : si par exemple les x_i sont des dates successives, $x_1 < \ldots < x_n$, la valeur prédite pour y à une date future x_{n+1} est simplement $\hat{y}_{n+1} = \hat{a}x_{n+1} + \hat{b}$. Notons cependant que s'il peut sembler naturel d'utiliser une valeur prédite pour compléter les données initiales *dans l'intervalle* des valeurs de X, on se gardera de prédire sans de multiples précautions supplémentaires des valeurs de Y *en dehors* de cet intervalle. En effet il se peut par exemple que la relation entre X et Y ne soit pas du tout linéaire mais qu'elle nous soit apparue comme telle à tort parce que les x_i sont proches les uns des autres.

Remarques : Pour finir voici quelques remarques :

1. Certains ne manqueront pas d'être surpris du fait qu'à coté des définitions de la variance et de la covariance que nous avons données on trouve dans certains ouvrages (ou dans les calculettes) une autre définition dans laquelle le facteur $\frac{1}{n}$ a été remplacé par le facteur $\frac{1}{n-1}$. Disons que "notre" définition est la définition de la *variance* (ou la *covariance*) *théorique* alors que celle

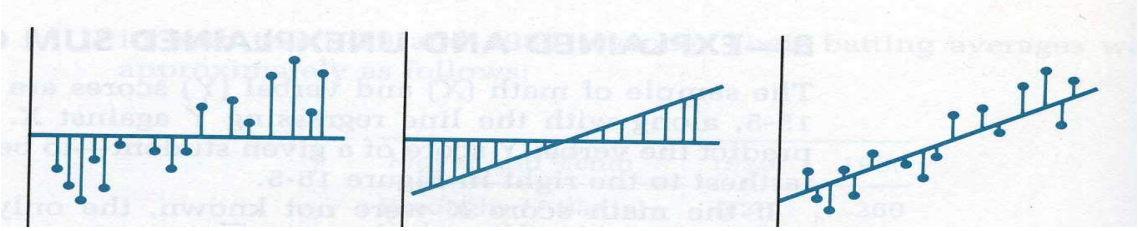


FIG. 1 – Illustration de la formule DT=DA+DR. La droite horizontale passe par le centre de gravité du nuage ; la première figure représente la dispersion totale DT, la seconde la dispersion due à la regression DR (nulle si la pente de la droite des moindres carrés est nulle et importante si cette pente est forte) et la troisième la dispersion autour de la droite, ou dispersion résiduelle.

qui comporte un facteur $\frac{1}{n-1}$ est la définition de la *variance* (ou la *covariance*) *empirique*. La première est celle que l'on utilise lorsque n est l'effectif total de la population alors que la seconde est celle que l'on utilise lorsque l'on estime la variance (ou la covariance) sur un échantillon de taille n beaucoup plus petite que la taille totale de la population dont est extrait l'échantillon. De toute façon, dans le cadre de la régression linéaire, on notera que tant pour le calcul de $\hat{a}$ que dans celui de r_{xy}, le résultat sera le même que l'on utilise l'une ou l'autre de ces formules.

2. Dans le calcul de la droite des moindres carrés, les variables X et Y ne jouent pas des rôles interchangeables. La variable dépendante Y prend, comme son nom l'indique, des valeurs qui dépendent de celles de X. D'ailleurs si l'on échange les rôles de X et de Y, on calcule une approximation linéaire de la forme $x = \hat{a}'y + \hat{b}'$, le critère des MCO est alors $E = \sum_{i=1}^{n}(x_i - (a'y_i + b'))^2$ qui n'est plus le même critère. Ainsi on ne sera pas surpris que la droite de regression de Y sur X n'est pas la même que la droite de régression de X sur Y. Cette droite, tout comme la précédente, passe par le centre de gravité du nuage de point, mais c'est leur seul point commun. C'est le problème considéré qui indique s'il faut considéré Y ou plutôt X comme variable dépendante (et l'autre comme variable explicative). Mais si l'on s'intéresse aux interactions entre deux variables X et Y dont ni l'une ni l'autre n'est clairement dépendante de l'autre, alors on pourra choisir de régresser Y en fonction de X ou bien le contraire.
3. On appelle *donnée éloignée* (*outlier*) un point du nuage situé à l'écart. S'il est éloigné dans la direction de y, il lui correspondra un important résidu. S'il est éloigné dans la direction des x, il peut présenter un très petit résidu et en même temps avoir une grande influence sur les valeurs de $\hat{a}$ et $\hat{b}$ trouvées.

   On appelle *donnée influente* un point du nuage dont l'oubli conduirait à une droite des moindres carrés bien différente. C'est souvent le cas des données éloignées dans la direction des x.
4. Attention à ne pas déduire trop hativement de la présence d'une liaison entre deux variables une relation de cause à effet ! Si quelqu'un devait suivre le degré de murissement des pêches et des abricots (par dosage de l'éthylène ou du fructose), il trouverait certainement une relation linéaire entre les deux. Mais le murissement des abricots n'influe pas sur celui des pèches ; ni l'inverse d'ailleurs. Par contre, les oscillations du niveau du lac Tchad (Afrique centrale) ont bel et bien leur source dans le cycle de 11 ans de l'activité solaire avec lequel elles sont parfaitement corrélées. Prudence donc.