Chapitre 7

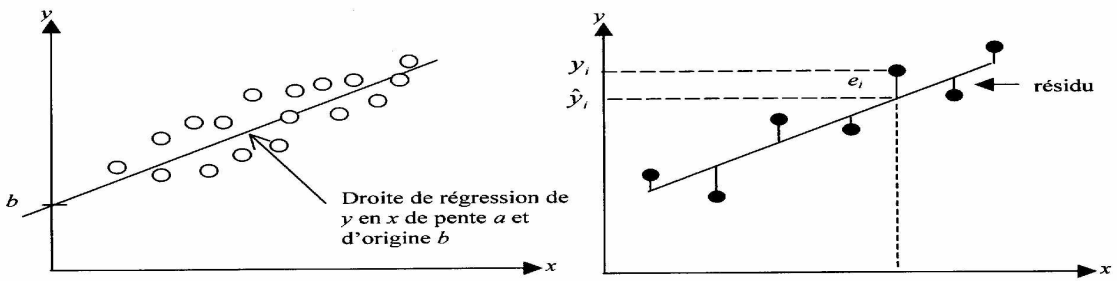
Méthode des moindres carrées

Une situation courante en sciences biologiques 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 de 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* .

7.1 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), 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 d'une propriété remarquable : c'est 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

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



Un calcul montre que ces valeurs, notées $\hat{a}$ et $\hat{b}$, sont égales à $\hat{a} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$ et $\hat{b} = \overline{y} - \hat{a}\overline{x}$. On exprime souvent $\hat{a}$ au moyen de la *variance* et de la *covariance* des variables aléatoires

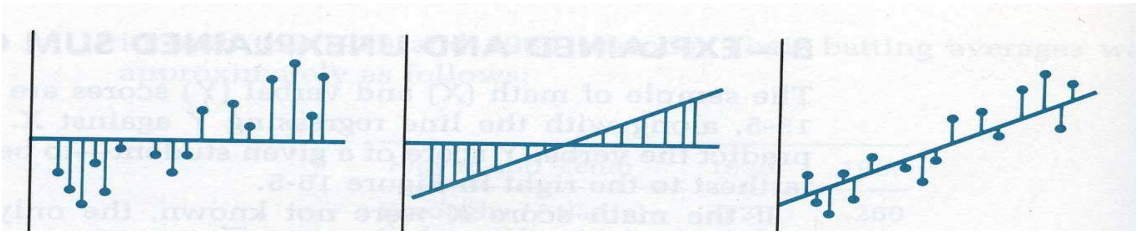


FIG. 7.1 – Illustration de la formule DT=DA+DR. La droite horizontale passe par le centre de gravité du nuage ; la première figure représente la dispersion totale DT, la seconde la dispersion due à la regression DR (nulle si la pente de la droite des moindres carrés est nulle et importante si cette pente est forte) et la troisième la dispersion autour de la droite, ou dispersion résiduelle.

X et Y par $\hat{a} = cov_{xy}/s_x^2$. En effet, on a :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{et} \quad cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

7.2 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 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. En pratique on estime souvent la régression acceptable lorsque $|r_{xy}| \geq \frac{\sqrt{3}}{2}$.

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 7.2) :

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

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 $R^2 = \frac{DR}{DT}$, 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 indépendant et distribués aléatoirement autour de 0. Ces hypothèses ne sont pas forcément faciles à vérifier. Un tracé 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).

7.3 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 prédites* de y *par le modèle* les valeurs $\hat{y}_i := \hat{a}x_i + \hat{b}$.

On utilise notamment ces valeurs pour faire des prévisions : si 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 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.

7.4 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 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 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 $E = \sum_{i=1}^{n}(x_i - (a'y_i + b'))^2$, et ce n'est plus le même et la droite que l'on obtient en général. 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 considérer 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. Mais on ne doit pas s'attendre à obtenir les mêmes résultats.

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.

7.5 Exercices

Exercice 1 : On possède 6 spécimens fossiles d'un animal disparu et ces spécimens sont de tailles différentes. On estime que si ces animaux appartiennent à la même espèce il doit exister une relation linéaire entre la longueur de deux de leurs os, le fémur et l'humérus. Voici les données de ces longueurs en cm pour les 5 spécimens possédant ces deux os intacts :

fémur	38	56	59	64	74
humérus	41	63	70	72	84

1. Tracer le nuage de point correspondant à ces données. Pensez-vous que les 5 spécimens peuvent appartenir à la même espèce et ne différer en taille que parce que certains sont plus jeunes que d'autres ?
2. Calculer à l'aide de votre calculette m_x, m_y, s_x, s_y et cov_{xy}. En déduire l'équation de la droite des moindres carrés. Contrôler vos calculs en superposant son graphe au nuage de points.
3. Calculer le coefficient de corrélation linéaire r. Qu'en concluez-vous ?
4. Reprenez les 2 questions précédentes en effectuant directement la regression linéaire au moyen de votre calculette. Vérifier que vos résultats sont identiques.

Exercice 2 :
1. Simuler au moyen de la fonction Random de votre calculette une suite de $n = 15$ nombres aléatoires $(\eta_i)_{i=1,..n}$ compris entre 0 et 1. Puis calculer les nombres $\varepsilon_i := 2\eta_i - 1$.
2. Calculer la moyenne m_ε des ε_i et les remplacer par $\varepsilon_i - m_\varepsilon$ si nécéssaire pour avoir une suite centrée, puis calculer l'écart type de cette suite. Pouviez-vous deviner sa valeur approximative ?
3. On choisit pour (x_i) la suite $0\,;0,25\,;0,5\,;0,75\,;1\,;1,25\,;1,5\,;1,75\,;2\,;2,25\,;2,5\,;2,75\,;3\,;3,25\,;3,5$ et pour (y_i) la suite $y_i = -2x_i + 3 + \varepsilon_i$. Calculer la droite de regression du nuage (x_i, y_i). Commentez.
4. Représenter les résidus et calculer la moyenne des carrés des résidus.
5. Représenter l'histogramme des résidus.

Exercice 3 : Pour étudier les problèmes de malnutrition dans un pays pauvre, on a calculé le poids moyen par age d'un échantillon de 2400 enfants répartis uniformément en 12 classes d'age. On a obtenu les données suivantes :

age	1	2	3	4	5	6	7	8	9	10	11	12
poids	4,3	5,1	5,7	6,3	6,8	7,1	7,2	7,2	7,2	7,2	7,5	7,8

1. Un statisticien pressé a fait calculer par sa machine la droite des moindres carrés pour ces données et a trouvé la relation $poid = 4,88 + 0,267age$. S'est-il trompé ?
2. A votre avis, quelle est la pertinence de son modèle ?
3. Calculer puis tracer les résidus. Vous constaterez que deux résidus successifs sont beaucoup plus souvent du même signe que du signe opposé. Ceci n'est pas compatible avec le fait qu'ils soient supposés indépendants. On dit que les résidus sont *autocorrélés*. C'est une raison de rejeter le modèle.

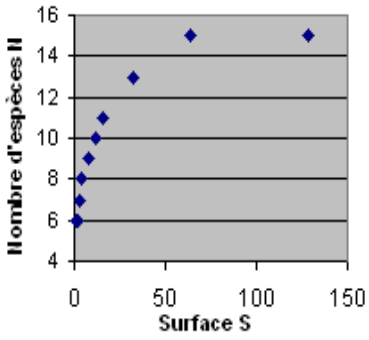
Exercice 4 : L'une des rares lois que l'on a pu mettre en évidence en Ecologie est la relation existant entre le nombre N d'espèces présentes dans un habitat donné (bien délimité) et la surface S de cet habitat. On considère généralement que cette relation est de la forme

$$N = AS^B \tag{7.1}$$

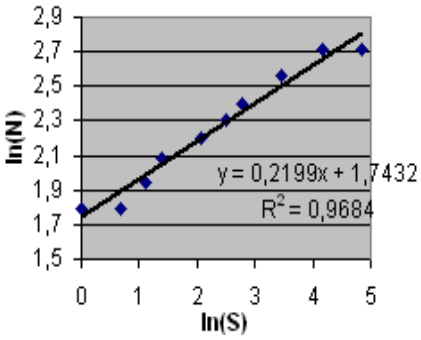
où A et B sont deux constantes. Afin de vérifier cette relation pour les plantes présentes dans une prairie (pissenlit, paquerettes, orties, boutons d'or, ...), on a effectué les mesures indiquées dans le premier tableau ci-dessous. On a représenté sur la première figure ci-dessous les valeurs de N en fonction de celles de S et sur la deuxième les valeurs de $\tilde{N} = \ln(N)$ en fonction de celles de $\tilde{S} = \ln(S)$. On voit que la regression linéaire de $\tilde{N}$ sur $\tilde{S}$ a donné :

$$\tilde{N} = 0,2199\tilde{S} + 1,7432 \text{ avec } R^2 = 0,9684 \tag{7.2}$$

S	N
1	6
2	6
3	7
4	8
8	9
12	10
16	11
32	13
64	15
128	15



ln(S)	ln(N)
0,00	1,79
0,69	1,79
1,10	1,95
1,39	2,08
2,08	2,20
2,48	2,30
2,77	2,40
3,47	2,56
4,16	2,71
4,85	2,71



1. Pourquoi n'a-t-on pas effectué directement une régression linéaire de N sur S ? Expliquez l'intéret de cette transformation des données.
2. Que représente R^2 et que peut-on déduire de sa valeur ?
3. A partir de la régression linéaire (7.2), calculer les constantes A et B de la relation (7.1).
4. Quelle valeur $\tilde{N}$ ce modèle linéaire prédit-il pour $\tilde{S} = ln(128)$? En comparant avec la valeur de $\tilde{S}$ observée, calculer le résidu ε en ce point.
5. Quelle valeur $\tilde{N}$ ce modèle linéaire prédit-il pour $\tilde{S} = ln(100)$? En déduire le nombre d'espèces pouvant coexister dans un habitat de surface $S = 100$, selon ce modèle.

Exercice 5 : On a mesuré sur un peuplement de bouleau blanc (Betula alba) dans le Massif Central les circonférences des troncs de 21 individus à la hauteur de 1.3 mtres du sol (indice DBH). Dans le même temps, un carottage des arbres a permis d'estimer leurs ages respectifs. De cet ensemble de données on a extrait les données des arbres d'ages 1 à 120 par pas de 20 ans. Par ailleurs on a constaté sur le terrain que les arbres se répartissent en trois catégories : les arbres les plus hauts (dominants), les arbres moyens (codominants) et les arbres plus petits, sous le couvert des autres : les dominés.

1. Tracez sur un même graphique les trois courbes représentant la circonférence des troncs en fonction de l'age. Que constate-t-on et comment interprétez-vous les différences constatées ? Que pensez-vous de l'allure des courbes ? Quel type de fonction peut-on envisager d'ajuster ?
2. On souhaite vérifier que la croissance en circonférence des troncs peut être modélisé par une *exponentielle saturée* de la forme $y(t) = y_{max}(1 - \exp(rt))$ où $y(t)$ est la circonférence l'instant t, y_{max} la valeur maximale que la circonférence peut prendre, r un taux de croissance en circonférence et t le temps. Les valeurs de y_{max} ont été estimées empiriquement à 86.4 cm, 65.43 cm et 36.00 cm pour chacune des trois catégories d'arbres. En remarquant que, d'après l'expression de $y(t)$, la quantité $\ln(y(t) - y_{max})$

dépend de façon linéaire de t, estimez au moyen d'une regression linéaire le paramètre r pour chacun des trois modèles. Vérifiez sur l'un des trois résultats la bonne qualité de l'ajustement des données.

Ages	1	20	40	60	80	100	120
Dominants	$1,26$	$22,29$	$40,09$	$56,15$	$63,49$	$71,69$	$81,08$
Dominés	$1,27$	$16,02$	$29,42$	$31,61$	$35,61$	$35,69$	$35,93$
Codominants	$1,29$	$22,14$	$35,69$	$49,23$	$56,88$	$60,43$	$63,74$