www.fundp.ac.be/biostats **Module 20**

20 Statistiques descriptives à deux dimensions

Nous allons à présent aborder les techniques relatives à la description de deux variables mesurées simultanément. Ces techniques sont utilisées lorsque l'expérimentateur s'intéresse à la relation qui pourrait exister entre deux variables qui interviennent dans un phénomène naturel. Ces études sont très fréquentes; imaginons par exemple l'étude de l'évolution de la capacité respiratoire en fonction de l'exercice, de l'évolution du rythme cardiaque en fonction de l'administration d'une drogue, du poids des individus en fonction de leur taille, du taux de cholestérol sanguin en fonction du poids du corps, de la production laitière en fonction de la teneur des aliments en protéines, etc…

L'expérimentateur mesure **deux** valeurs expérimentales pour chaque observation individuelle, ce qui produira une série statistique à deux dimensions :

x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
...	...
x_i	y_i
...	...
x_n	y_n

Tableau -20 -20-1 Représentation schématique d'un tableau de données à deux variables.

20.1 Tables et graphiques

Nous pouvons généraliser les techniques utilisées pour la description des observations à une dimension. Les variables continues seront regroupées en classes comme précédemment. Supposons que nous disposions de 464 observations individuelles pour lesquelles la masse corporelle (kg) est mesurée simultanément avec le taux de cholestérol sanguin (mg/100ml de sang). Ces observations peuvent être réparties en 12 classes de poids (intervalle de 3 kg) et 12 classes de taux de cholestérol (intervalle de 20 mg/100 ml).

On dresse alors une table de fréquences présentant les classes de cholestérol en tête de ligne et les classes de poids en tête de colonne. Chaque élément de cette table représente le nombre d'observations individuelles qui se trouvent dans la catégorie déterminée simultanément par les classes définies pour les deux variables.

S'il existe une relation entre les variables, elle apparaîtra sous la forme d'une structure diagonale dans le tableau : aux poids plus élevés correspondent généralement des taux de cholestérol plus élevés.

Cholestérol		70	75	80	85	90	
P	140	23	15	12	8	2	**60**
O	180	12	45	18	11	5	**91**
I	220	8	22	38	14	12	**94**
D	260	5	13	17	55	24	**114**
S	300	3	5	14	35	48	**105**
		51	**100**	**99**	**123**	**91**	**464**

Tableau 20 -20-2 Table à deux entrées représentant la fréquence des observations pour des catégories de poids et de taux de cholestérol. Exemple : 23 personnes ont un poids de classe 70kg et un taux de cholestérol de classe 140mg/100ml.

La table de fréquences permet d'obtenir un classement des observations expérimentales. Elle peut être représentée sous la forme d'un stéréogramme qui est une généralisation en trois dimensions de l'histogramme. Le poids et le taux de cholestérol se placent en abscisse et la fréquence (ou la fréquence relative) en ordonnée.

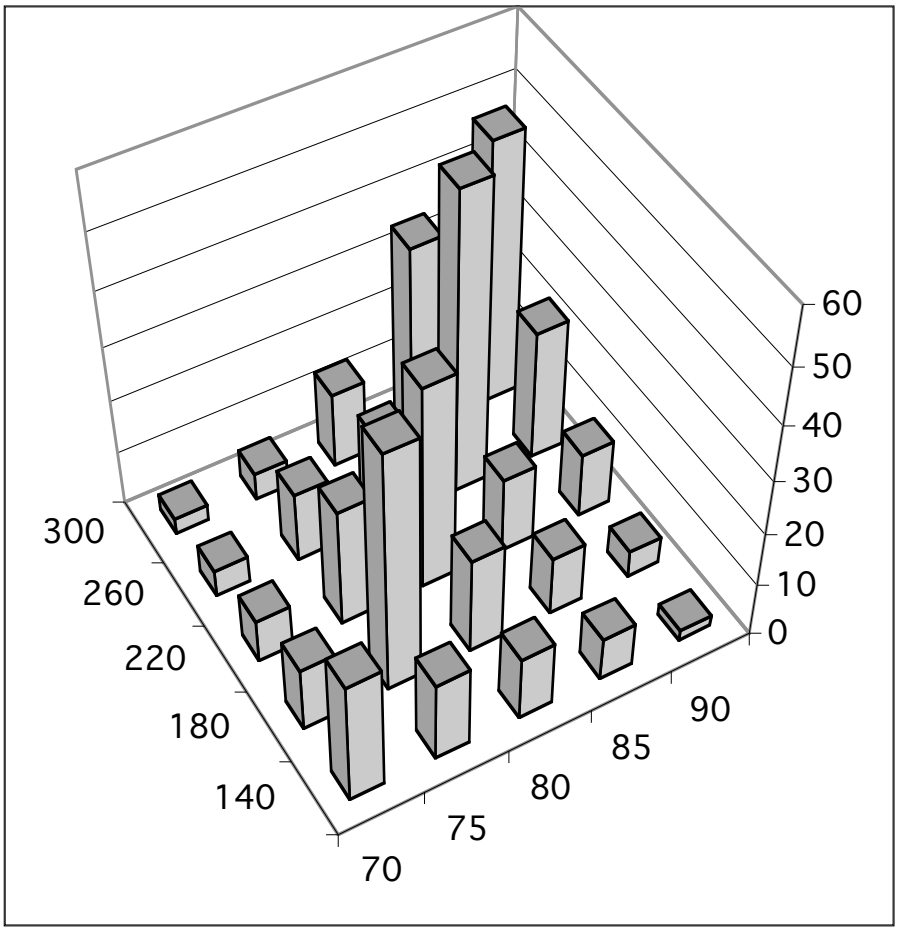

Figure 20 -1 Représentation d'un stéréogramme : histogramme à 3 dimensions : la hauteur (ordonnée) représente la fréquence et les deux abscisses, les valeurs de poids et de cholestérol.

Ce type de représentation reste toutefois relativement lourd à manipuler, et ne propose pas de valeurs synthétisant l'essentiel de l'information. Nous allons entreprendre de rechercher les valeurs caractéristiques permettant de mesurer l'intensité de la relation entre deux variables, et de caractériser cette relation. Nous envisageons cette démarche dans le cadre des variables continues.

20.2 Intensité de la relation entre deux variables

20.2.1 Modélisation de la variabilité

Envisageons la pesée de 11 poissons pris au hasard dans une population donnant la série statistique suivante (Y, en g) :

Y (g)
0,2
3,3
4,9
8,7
11
8,4
9,2
13,6
19,9
19,8
21,4

Tableau 20 -20-3 Poids de 11 poissons (en g).

La variabilité, exprimée par SCET, vaut 500 g^2 (nous laisserons tomber les unités par la suite) et représente à ce stade -ci une variabilité totale, inexpliquée.

SCET = variabilité inexpliquée

Prenons à présent en considération l'âge des poissons, (X, exprimé en semaines), puis représentons chaque poisson par un point dans un diagramme de dispersion, chaque point correspondant à une valeur de X en abscisse et de Y en ordonnée.

X semaines	Y g
0	0,2
1	3,3
2	4,9
3	8,7
4	11
5	8,4
6	9,2
7	13,6
8	19,9
9	19,8
10	21,4

Tableau 20 -20-4 Age et poids de 11 poissons.

Ce type de graphique met en évidence une relation entre l'âge et le poids, les poissons plus âgés étant plus lourds. Cette relation peut être modélisée par l'équation d'une droite :

$Y_m = B_o + B_1 X$

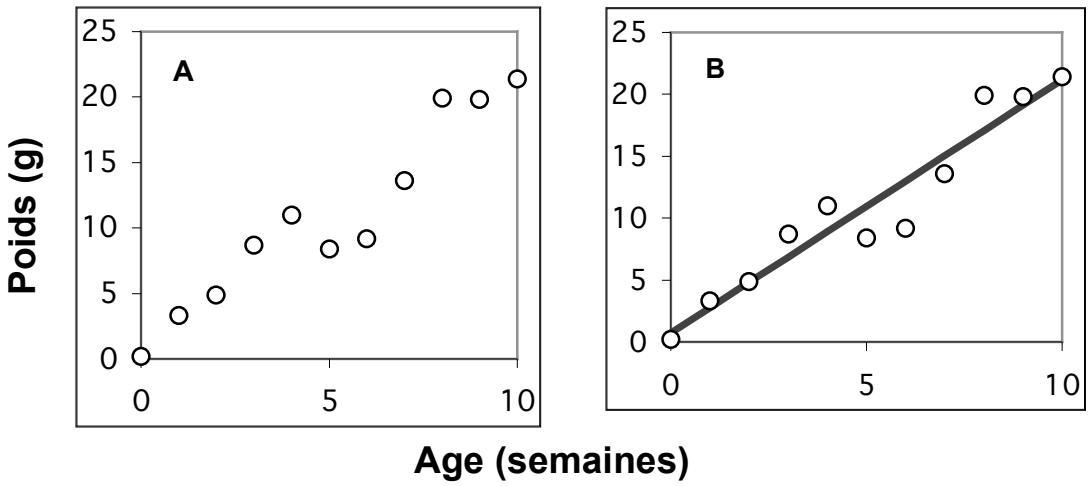

Figure 20 -2 A – B - Représentation graphique (diagramme de dispersion) du poids et de l'âge observés pour 11 poissons. B - Droite représentant le modèle linéaire de la croissance des poissons.

Nous verrons plus loin comment déterminer les paramètres b_0 et b_1.

Les points expérimentaux (appelons-les Y_o pour bien les distinguer) ne coïncidant pas exactement avec le modèle, chaque poisson s'écarte quelque peu de la droite, d'une valeur que nous appellerons Y_e.

$$Y_e = Y_o - Y_m$$

$$Y_o = Y_m + Y_e$$

$$Y_o = B_o + B_1X + Y_e$$

X semaines	Y_o g	Y_m g	Y_e g
0	0,2	0,72	-0,52
1	3,3	2,76	0,54
2	4,9	4,81	0,09
3	8,7	6,85	1,85
4	11	8,90	2,10
5	8,4	10,95	-2,55
6	9,2	12,99	-3,79
7	13,6	15,04	-1,44
8	19,9	17,08	2,82
9	19,8	19,13	0,67
10	21,4	21,17	0,23
MY	11	11	0
SCEY	500	460	40

Tableau 20 -20-5 Moyenne et somme des carrés d'écarts pour les valeurs de poids observées sur 11 poissons (Y_o), les valeurs de poids prédites par un modèle linéaire (Y_m) et l'écart de poids entre le modèle et l'observation (Y_e). Notez l'équation de l'analyse de la variance $SCEY_o = SCEY_m + SCEY_e$.

Nous pouvons constater que
$SCE_{Y_o} = 500$, $SCEY_m = 460$ et $SCE_{Ye} = 40$,
ce qui nous conduit à l'équation de la SCE :

$SCEY_o = SCEY_m + SCEY_e$
SCET = SCEF +SCER

Variabilité totale = information + bruit

20.2.2 Qualité du modèle

Le rapport entre la variabilité factorielle et la variabilité totale représente la qualité explicative du modèle :

$$R^2 = \frac{SCE_F}{SCE_T} = \frac{SCE_{Ym}}{SCE_{Yo}} = \frac{information}{variabilité\ \ totale}$$ **Équation 20-1**

Le modèle linéaire est de bonne qualité car le rapport 460/500 = 0,92 soit 92% de la variabilité du poids est expliquée par l'âge. Les 8% restants représentent la variabilité inexpliquée par ce modèle (peut -être explicable par un autre modèle). Il s'agit de la variabilité résiduelle du modèle.

Ce rapport porte le nom de coefficient de détermination. Il est noté R^2.

Nous montrerons que dans le cas du modèle linéaire (uniquement) ce coefficient est le carré du coefficient de corrélation, noté R, ce qui justifie sa notation R^2. Plus le modèle est complexe, plus la variabilité résiduelle est censée diminuer. Des modèles plus complexes sont soit non linéaires, soit font intervenir plusieurs variables explicatives (le sexe, la T° de l'eau..). Ces modèles complexes ne sont pas étudiés ici.

La figure suivante illustre la relation taille -poids chez l'adolescent, obtenue sur 37 individus pris au hasard dans la population.

Il apparaît que le poids a tendance à augmenter avec la taille : les poids les plus élevés sont observés chez les plus grands individus, et réciproquement.

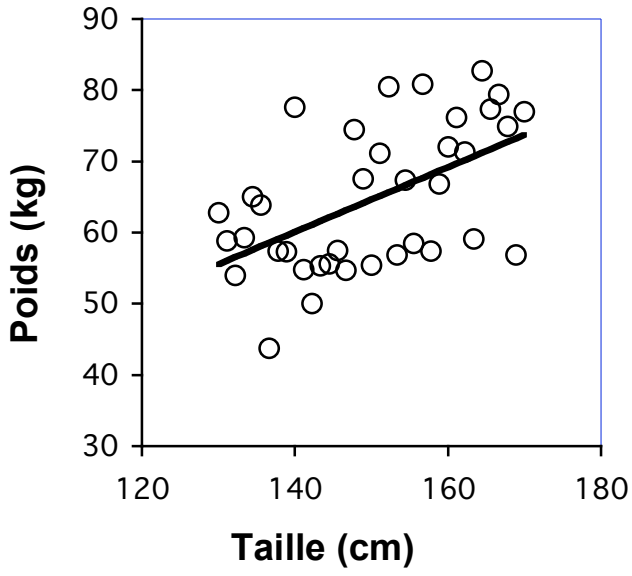

Figure 20 -3 Représentation graphique (diagramme de dispersion) du poids et de la taille observés pour 37 adolescents. Droite représentant le modèle linéaire de la relation poids / taille.

Nous considérerons qu'il s'agit d'une relation linéaire*, car il n'y a pas, manifestement, de relation non linéaire (courbe) qui exprime mieux la relation entre le poids et la taille. Toutefois, la relation apparaît plus* diffuse *que dans l'exemple de la relation âge -poids chez le poisson. Cela signifie que le poids de l'adolescent est plus variable, à taille égale, que ne l'est le poids du poisson, à âge égal.*

L'équation de la SCE est la suivante :

Variabilité totale = information + bruit

SCET = SCEF + SCER

3715,1 = 1075,4 + 2639,7

Notez bien que les valeurs ne sont pas comparables aux valeurs 500, 460, 40 trouvées pour les poissons, respectivement, En effet, les unités et le nombre d'observations sont différents. Par contre, dans le calcul du coefficient de détermination R^2 :

R^2 = 1075,4 /3715,2 = 0,29

les unités et l'effet de la taille de l'échantillon se sont annulés et cette valeur exprime la qualité du modèle, de façon comparable dans n'importe quel contexte.

Dans le cas présent, 29% de la variabilité du poids sont expliqués par la taille et 71% de la variabilité sont inexpliqués par ce modèle. Il s'agit de la variabilité résiduelle du modèle.

20.2.3 Propriétés du coefficient de détermination

De ce qui précède, nous pouvons conclure que la variabilité du poids (SCET) peut être répartie en deux types de variabilité :

1 la variabilité expliquée par la relation linéaire entre X et Y (SCEF)

Dans une certaine mesure, Y varie lorsque X varie : le poids est plus élevé lorsque la taille est plus élevée.

2 la variabilité inexpliquée par la relation linéaire entre X et Y (SCER)

La variabilité de Y n'est pas strictement liée à la variabilité de X : deux enfants de même taille ont des poids différents, parce que le poids dépend d'autres variables qui ne sont pas prises en considération (qui pourraient l'être dans un modèle plus complexe), et par effet du hasard et de l'erreur expérimentale (considérée comme inexpliquée).

Le coefficient de détermination R^2 représente la **proportion** de la variabilité de y qui est expliquée par la relation linéaire entre X et Y.

Cette affirmation ne sera démontrée qu'après le développement de la régression.

Nous pouvons en déduire que cette proportion est comprise entre deux extrêmes :

$R^2 = 0$: il n'y a aucune relation linéaire entre X et Y, et donc aucune variabilité de Y n'est expliquée par X;

$R^2 = 1$: la valeur de Y est déterminée entièrement par la valeur de X, et donc il n'y a aucune variabilité inexpliquée de Y.

$$0 \leq R^2 \leq 1 \text{ et donc } -1 \leq R \leq 1$$

R s'appelle le coefficient de corrélation linéaire. Il est toujours compris entre -1 et +1. Ces deux extrêmes correspondent à une relation parfaite, entre X et Y. Un coefficient de corrélation nul correspond à une absence de relation linéaire entre X et Y.

20.3 Mesure de l'intensité de la relation entre deux variables

20.3.1 Somme des produits d'écarts à la moyenne

Reprenons les observations obtenues pour la taille et le poids des 37 adolescents, et centrons -les en retirant de chaque observation la moyenne de l'échantillon (cfr 1.5.7) :

$$x_i'' = x_i - Mx \qquad y_i'' = y_i - My$$

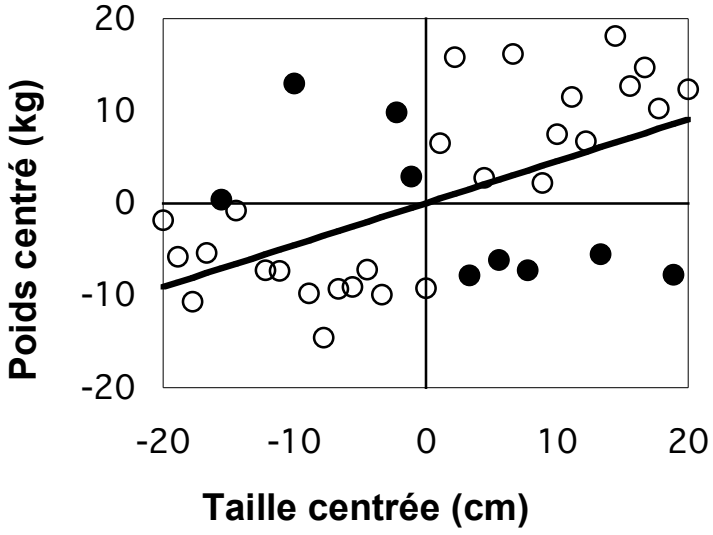

Figure 20 -4 Répartition des points dans les quatre quadrants définis par les écarts à la moyenne en X et en Y.

Cette opération fait apparaître quatre quadrants dans le graphique.

Comment les points vont -ils s'y répartir?
Les points situés dans le quadrant inférieur gauche et supérieur droit (blancs) correspondent à des individus plus petits et plus légers que la moyenne, ou plus grands et plus lourds que la moyenne, respectivement . Ces points sont majoritaires et contribuent à établir qu'il existe une relation positive entre la taille et le poids (positive signifie que lorsque l'une augmente, l'autre augmente).
Les points situés dans le quadrant inférieur droit et supérieur gauche (noirs) correspondent à des individus plus grands et plus légers que la moyenne, ou plus petits et plus lourds que la moyenne, respectivement. Ces points sont minoritaires et représentent des exceptions à la relation positive majoritaire entre la taille et le poids.
Si les points se répartissaient de façon équivalente dans les 4 quadrants, Y pourrait prendre n'importe quelle valeur pour une valeur donnée de X et nous n'aurions pas de raison d'établir une relation positive entre la taille et le poids.

Pour quantifier cette observation graphique, nous allons définir, pour le point (x_i, y_i) le produit des écarts aux moyennes :

$$(x_i - Mx)(y_i - My)$$

Ce produit a la propriété d'être positif pour les points situés dans les quadrants inférieur gauche et supérieur droit, et négatif dans les deux autres quadrants.

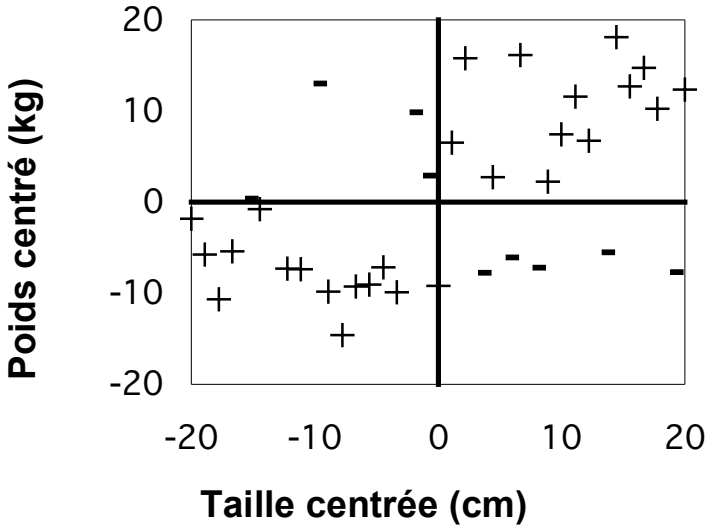

Figure 20 -5 Représentation, dans chacun des quadrants, du signe du produit des écarts, lié au sens de la relation entre X et Y (+ : Y augmente lorsque X augmente, - : Y augmente lorsque X diminue).

En sommant ce produit sur toutes les observations, nous obtenons la somme des produits des écarts à la moyenne (SPE_{xy}) :

$$SPE_{xy} = \sum_{i=1}^{n} (x_i - Mx)(y_i - My)$$

Équation 20-2

La valeur et le signe de SPE_{xy} seront influencés par le sens de la relation liant X et Y .

- Une SPE_{xy} positive indiquera une concentration des points dans les quadrants correspondant à un produit des écarts positif, et donc à une relation positive entre X et Y ;

- une SPE_{xy} nulle, ou pratiquement nulle indiquera une distribution des points équivalente dans les 4 quadrants, et donc une absence de relation ;

- une SPE_{xy} négative indiquera une concentration des points dans les quadrants correspondant à un produit des écarts négatifs, et donc à une relation négative entre X et Y.

Il faut noter que la valeur de SPE_{xy} dépend du système d'unités et du nombre d'observations dans l'échantillon. Il est nécessaire de calculer une valeur moyenne standardisée, comme nous l'avons fait pour la variance et le Zscore. Dans cet exemple, SPE_{xy} est exprimée en cm . kg.

20.3.2 Covariance

La valeur moyenne de la variabilité conjointe de X et Y est appelée covariance.

$$\mathrm{cov}_{xy} = S_{xy} = \frac{SPE_{xy}}{n}$$

Équation 20-3

La "co"variance mesure si les variables varient "ensemble" ou non (co - comme dans copain, collaborateur...).

Comme pour la variance, il existe une autre définition, dans laquelle n est remplacé par n -1.

20.3.3 Coefficient de corrélation linéaire

La dernière étape de notre démarche va être de standardiser les variables X et Y, c'est -à -dire de rendre les valeurs indépendantes du système d'unités. Pour cela, nous allons exprimer les écarts à la moyenne en "unités écart -type", ce qui correspond à la transformation :

$$Z_{xi} = \frac{x_i - Mx}{S_x} \quad et \quad Z_{yi} = \frac{y_i - My}{S_y}$$

Z étant un écart à la moyenne, le calcul de SPE se simplifie :

$$SPE_{ZxZy} = \sum_{i=1}^{n} Z_{xi} Z_{yi}$$

et la covariance devient :

$$\text{cov}_{ZxZy} = \frac{SPE_{ZxZy}}{n} = R$$

ce que l'on peut écrire aussi :

$$R = \frac{\text{cov}_{xy}}{S_x S_y} = \frac{S_{xy}}{S_x S_y}$$ **Équation 20-4**

C'est ce que l'on appelle R, le coefficient de corrélation linéaire de Spearman.

> *On peut considérer R comme étant la covariance des variables préalablement standardisées.*

Pour établir que, dans le modèle linéaire, R^2 est le coefficient de détermination nous devrons d'abord établir la mesure des paramètres de la droite de régression.

20.3.4 Interprétation de la valeur du coefficient de corrélation

Le coefficient de corrélation a une interprétation similaire à celle du coefficient de détermination (ils sont évidemment très liés, l'un étant le carré de l'autre).
Lorsque l'on s'intéresse à l'intensité de la relation, on préférera R^2 qui s'exprime directement en proportion (souvent exprimée en %) de variabilité expliquée par le modèle ; si on s'intéresse au signe de la relation, on choisira R, car R^2 n'en a pas.

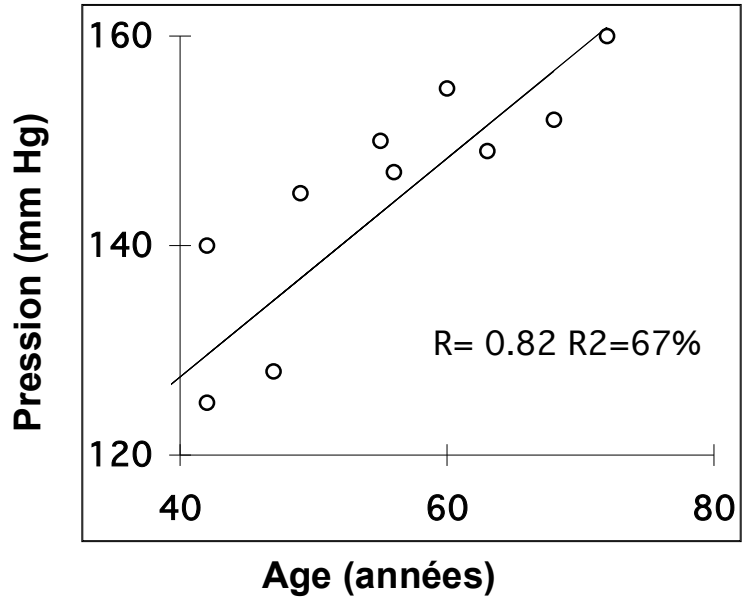

Figure 20 -6 Représentation graphique d'une relation positive entre l'âge et la pression sanguine. Il s'agit d'une relation linéaire intense.

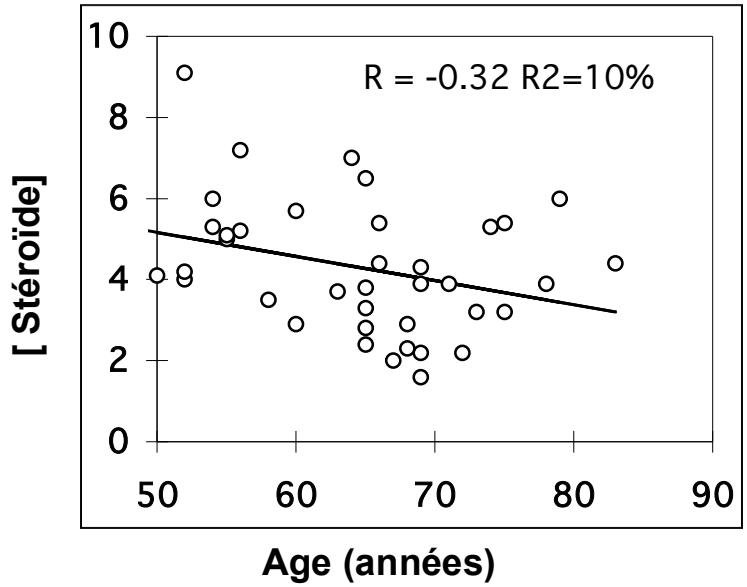

Figure 20 -7 Représentation graphique d'une relation négative entre l'âge et la concentration d'un stéroïde dans l'urine. Il s'agit d'une relation linéaire peu intense.

Remarquez que dans l'étude de la relation entre l'âge et la concentration d'un stéroïde dans l'urine, on parle de relation linéaire, de la même façon que dans l'étude entre l'âge et la pression sanguine. Seule l'intensité (mis à part le signe) change d'une relation à l'autre. On parle de relation linéaire lorsqu'il n'y a aucune raison d'utiliser un modèle non linéaire pour mieux exprimer les résultats.

A ce stade, le critère auquel nous nous référerons pour en juger est visuel. Une analyse quantitative de la non linéarité sera présentée dans le cadre de l'analyse de la variance.

20.3.5 Relations non linéaires

Ni R ni R^2 ne permettent de déterminer si une relation est linéaire ou non.

Certaines relations très étroites entre deux variables, mais non linéaires, correspondent à un coefficient de corrélation nul ou presque nul.

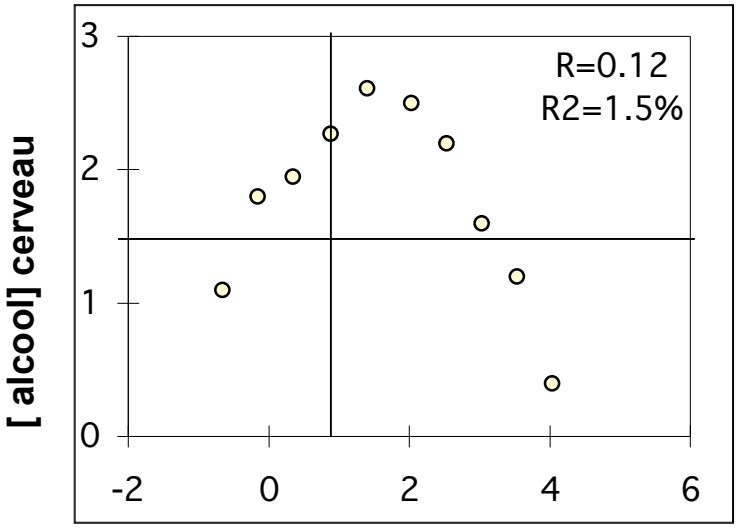

Figure 20 -8 Concentration de différents alcools dans le cerveau en fonction de leur lipophilicité (affinité pour un tissu gras).

Si l'on représente les données centrées, on observe que les valeurs positives de produit d'écarts vont contrebalancer les valeurs négatives, et par conséquent SPE_{xy} et de ce fait R et R^2 seront à peu près nuls.

Un coefficient de corrélation nul ne peut pas être interprété comme une absence de toute relation entre les variables.

Ceci montre bien que la variabilité résiduelle inexpliquée (ici 98,5%) n'est pas nécessairement inexplicable : seul un modèle non linéaire pourrait contribuer à l'expliquer.

Dans d'autres cas le modèle linéaire peut expliquer une partie de la variabilité de Y, ce qui explique que dans ce cas R^2 ne soit pas nul, et puisse même être très élevé.

Ceci s'explique par le fait que le modèle linéaire n'est pas très éloigné des points. La variabilité résiduelle est faible et R^2 est élevé. Le graphique montre manifestement qu'une courbe s'accorderait mieux au phénomène décrit.

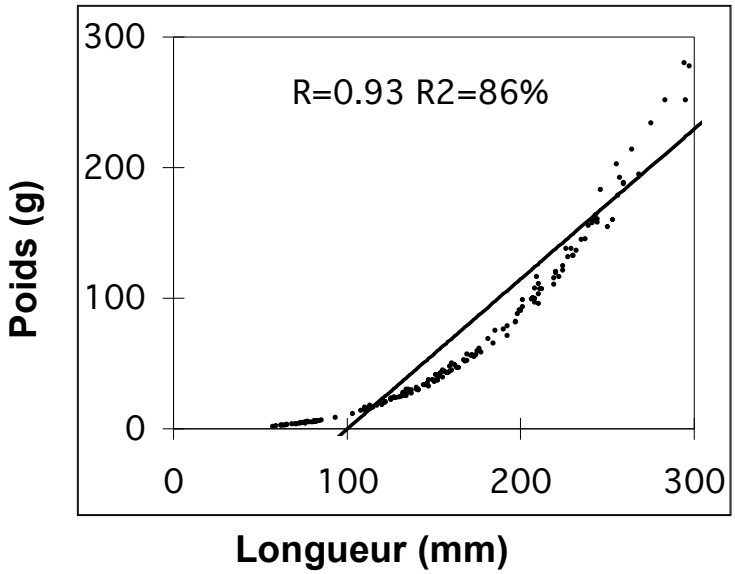

Figure 20 -9 Des écologistes ont capturé 181 truites (Salmo Trutta Fario) dans le Samson et ont mesuré la longueur à la fourche (mm) et le poids (g) de chaque animal.

Un coefficient R^2 élevé ne signifie pas nécessairement que la relation soit linéaire.

20.3.6 Corrélation et causalité

Etablir une relation entre deux variables ne signifie en aucun cas établir une relation de causalité entre elles : le fait que la variation de X soit liée à la variation de Y n'implique pas qu'elle en soit la cause.

Prenons un exemple caricatural. On peut recenser, dans un certain nombre de carrés d'un km pris au hasard sur la carte de Belgique, le nombre de GSM (X) et le nombre de cancers du poumon dans la population (Y). Ce recensement produirait très certainement un coefficient de corrélation très élevé : des couples de valeurs nulles correspondant aux carrés qui ne sont pas habités (pas de GSM, pas de cancer), et des couples de valeurs élevées correspondant aux agglomérations (beaucoup de GSM, beaucoup de cancers). Remarquez que l'on obtiendrait un résultat similaire en recensant les cabines téléphoniques au lieu des GSM : aucune relation de causalité ne peut être établie entre ces deux variables. Simplement, ces variables sont liées à la densité de population, et donc augmentent ensemble.

Il faut rechercher des preuves expérimentales de la relation de causalité.

Par exemple des expériences montrent sans équivoque la relation entre le nombre de cigarettes fumées par jour et la fréquence du cancer du poumon, entre la durée du jour et la vitesse de croissance d'une plante, entre le taux d'œstrogène et le développement de l'ovule etc… qui sont caractérisées par un R positif, mais pour lesquelles on a également montré une relation avec la dose, une réversibilité, la disparition de la relation en utilisant un placebo…

20.4 Caractérisation de la relation entre deux variables

20.4.1 Ecarts résiduels

S'il existe une relation entre deux variables, l'intérêt de l'expérimentateur sera de pouvoir prédire la valeur que devrait prendre une variable à partir de la valeur observée pour l'autre. Nous allons donc rechercher des valeurs qui nous permettront de caractériser la relation, de façon à disposer d'un outil de prédiction.

Dans le cadre des relations linéaires entre variables, on peut exprimer dans la population la relation entre X et Y par l'équation d'une droite qui exprime la relation entre X et Y :

$$\mu_{y_i} = \beta o + \beta_1 x_i$$

Dans cette expression, β_o est l'ordonnée à l'origine, qui correspond à la valeur de Y lorsque X est nul, β_1 est la pente de la droite, qui correspond à l'augmentation de Y attendue pour une augmentation d'une unité de X.

Dans un échantillon, la relation est estimée par la relation suivante :

$$Ym_i = Bo + B_1 x_i$$

Si X est exprimé en cm et Y en kg, B_0 est exprimé en kg et B_1 en kg/cm.

Dans cette expression, $B_{\square}$ est l'ordonnée à l'origine, qui correspond à la valeur estimée de Y lorsque X est nul, B_i est la pente de la droite, qui correspond à l'augmentation de Y estimée pour une augmentation d'une unité de X.

Les valeurs Yo_i observées dans l'échantillon ne seront pas égales aux valeurs Ym_i formant la droite.

$$Yo_i = Ym_i + Ye_i$$

Nous pourrions donc tracer plusieurs droites, caractérisées par différentes valeurs de B_0 et de B_1 .

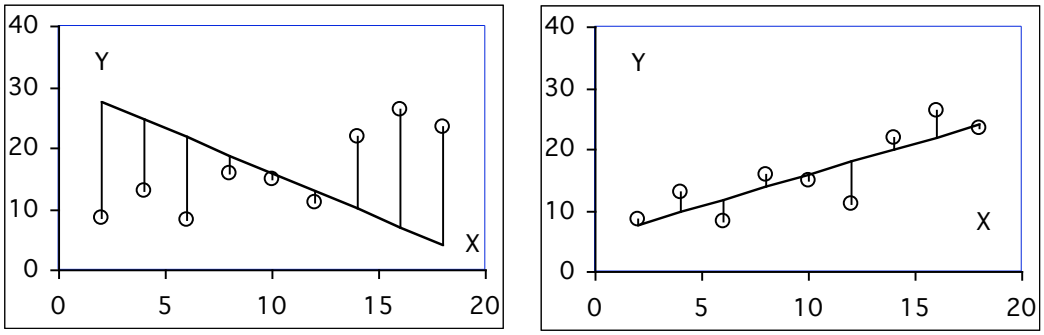

Figure 20 -10 Ajustement des paramètres B_o et B_1 pour un même ensemble de points. ***a:*** *ajustement non optimal ;* ***b*** *: ajustement optimal.*

Intuitivement, nous pouvons établir que des deux situations présentées ci - dessus, la seconde apparaît meilleure que la première. Le but de notre démarche est de rechercher *la ou les* droite(s) qui exprime(nt) au mieux la relation linéaire entre X et Y de l'échantillon.

> *La seconde solution apparaît meilleure parce que les écarts entre les points et la droite sont plus petits.*

On comprend intuitivement que la droite qui exprimera au mieux la relation entre X et Y correspondra à l'écart moyen le plus petit possible, pour l'ensemble des observations.

Cependant, en considérant les observations de chacune des situations, nous constatons que les écarts ont une somme (et dès lors une moyenne) nulle, que l'ajustement soit favorable ou non.

X	Y_o	Y_m	Y_e
sem		g	
2	8,6	26,7	-18,1
4	12,9	24,0	-11,1
6	8,1	21,3	-13,2
8	15,8	18,6	-2,8
10	15,1	15,9	-0,8
12	11,3	13,2	-1,9
14	21,8	10,6	11,2
16	26,3	7,9	18,4
18	23,5	5,2	18,3
$\sum$	143,3	143,3	0,0
My	15,9	15,9	0,0
SCE	345,7	432,6	1436,0

Tableau 20 -20-6 Observations, modèle et écarts au modèle pour un ajustement défavorable,

X	Y_o	Y_m	Y_e
sem		g	
2	8,6	7,8	0,9
4	12,9	9,8	3,1
6	8,1	11,9	-3,7
8	15,8	13,9	1,9
10	15,1	15,9	-0,9
12	11,3	18,0	-6,7
14	21,8	20,0	1,8
16	26,3	22,0	4,2
18	23,5	24,1	-0,6
∑	143,4	143,4	0
My	15,9	15,9	
SCE	345,7	248,5	95,9

Tableau 20 -20-7 Observations, modèle et écarts au modèle pour un ajustement favorable.

Le critère d'ajustement sera donc de rendre minimale SCER, ce qui revient à rendre maximale SCEF et donc le rapport SCEF/SCET = R^2.

C'est aussi dans cette seule situation que nous respecterons la relation :

SCET = SCEF + SCER : 344,1 = 249,3 + 94,8

Nous pouvons remarquer que cette équation, n'est pas satisfaite dans la situation d'ajustement défavorable :

344,1≠431,8 +1432,2

20.4.2 Droite des moindres carrés Y(X)

Déterminer la meilleure droite revient donc à minimiser SCER calculée entre les valeurs observées et les valeurs estimées. SCER est une fonction de B_0 et de B_1 : SCER se modifie lorsque l'on modifie les paramètres.

$$SCER = \sum_{i=1}^{n} (yo_i - ym_i)^2 = \sum_{i=1}^{n} (yo_i - (Bo + B_1 x_i))^2$$

Nous pouvons visualiser schématiquement la façon dont la SCER, toujours positive, varie en fonction de B_0 et de B_1

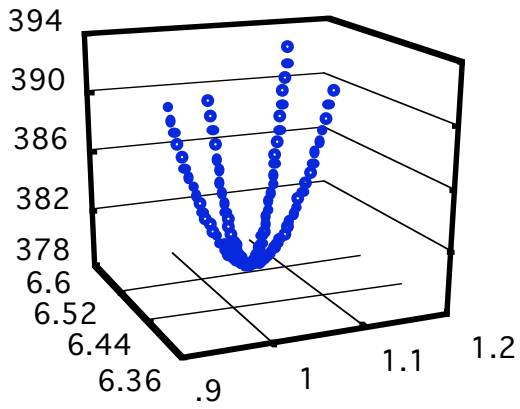

Figure 20 -11 A chaque combinaison de valeurs de B_0 et de B_1 correspond une valeur de SCER. Cette fonction présente un minimum unique.

Lorsque les valeurs de B_0 et de B_1 sont telles qu'elles ne peuvent pas être modifiées sans provoquer une augmentation de SCE, elles correspondent aux valeurs caractéristiques de la droite de régression de Y en fonction de X, caractérisant la relation linéaire entre ces variables.

On peut imaginer calculer un très grand nombre de droites et choisir celle qui présente la SCER minimale. Certains algorithmes programmés sur ordinateur procèdent de cette façon.

Il existe cependant une solution analytique à ce problème qui repose sur le principe que l'on se trouve au minimum d'une fonction de plusieurs paramètres lorsque la dérivée première par rapport à chacun de ces paramètres est nulle, et la dérivée seconde positive. Il n'est cependant pas nécessaire d'entrer dans ces détails pour bien interpréter la régression.

$$B_o = My - B_1 Mx \quad \{2.12\}$$

$$B_1 = SPE / SCEX \quad \{2.13\}$$

Remarquez bien :

1. les valeurs de B_o et de B_1 sont exprimées dans les unités des variables originales. Elles se modifient donc si les unités sont modifiées

	X	Y_o		X	Y_o
	sem	g		jours	kg
	2	4		14	0,004
	4	7,1		28	0,007
	6	8,1		42	0,008
	8	15,8		56	0,016
	10	15,1		70	0,015
	12	18		84	0,018
	14	21,8		98	0,022
	16	26,3		112	0,026
	18	35		126	0,035
M	10,0	16,8	M	70,0	0,017
SCE	240,0	785,4	SCE	11760,0	0,0008
SPE	422,4		SPE	3,0	
B_1 g/sem	1,8		B_1 kg/jour	0,00025	
B_o g	-0,8		B_o kg	-0,00080	

Tableau 20 -20-8 Relation âge/poids exprimée dans deux systèmes d'unités différents.

Notez la différence entre les valeurs des paramètres obtenus dans les deux systèmes d'unités.

2. B_0 est la valeur estimée de Y lorsque X vaut 0

Il ne faut pas confondre X = 0 avec la valeur minimale de X affichée sur le diagramme de dispersion.

Dans de nombreux graphiques, la valeur la plus à gauche de l'abscisse n'est pas zéro et l'intersection entre la droite et l'axe des Y n'est pas Bo.

La valeur de Bo est la valeur de Y qui permet d'ajuster la droite aux points dans le domaine de X qui a fait l'objet d'une mesure expérimentale.

Bo n'est pas nécessairement la valeur que l'on mesurerait pour Y si on faisait la mesure pour X = 0 (ce qui est impossible dans certains cas, si X est l'âge, la pression sanguine, la taille…). La valeur de B_0 peut donc très bien n'avoir aucun sens physiologique, par exemple produire une valeur de poids négative (ici –0,8 g). La relation linéaire qui décrit la relation dans le domaine dans lequel nous avons effectué nos mesures, ne la décrit pas nécessairement en dehors de ce domaine. Il est fréquent qu'une relation soit raisonnablement linéaire dans un certain domaine et devienne non linéaire à l'approche de certaines limites.

3. Il ne faut pas confondre somme des carrés des écarts à la moyenne et somme des carrés des écarts au modèle.

La SCE_R minimisée *pour établir les paramètres de l'équation mesure des écarts entre les points observés et les points modélisés (Y_o -Y_m). SCE_x et SPE_{XY} utilisées pour* calculer *le coefficient de pente de la droite de régression mesurent des écarts entre les observations et la moyenne de l'échantillon (x_i -M_x, Y_i -M_y).*

20.4.3 Interpolation et extrapolation

La bonne utilisation de la droite de régression pour effectuer des prédictions doit prendre en compte les éléments suivants :
- il n'y a pas d'évidence que la relation entre X et Y soit une relation non linéaire (à ce stade, se référer au diagramme de dispersion) ;
- R^2 indique une relation entre X et Y est suffisamment intense (appréciation fonction du contexte et en partie subjective).

En fait, il est permis de faire des prédictions à partir de relations dont le R^2 est faible, mais il faut en tenir compte dans l'interprétation des résultats : on doit alors attendre une distance parfois considérable entre la valeur prédite par la droite et les observations individuelles que l'on pourrait faire.

La valeur de X pour laquelle on effectue une prédiction de Y se trouve comprise entre les valeurs extrêmes de X pour lesquelles on a réalisé l'observation de Y (interpolation).

En pratique, l'expérimentateur sera pourtant souvent intéressé par une prédiction sortant du domaine de l'interpolation. Il pratiquera alors une extrapolation. Mais il doit savoir que la linéarité de relation et la valeur du R^2 ne sera pas vérifiée dans le domaine où il effectue ses prédictions. Il prend donc un risque non évalué, d'autant plus grand que la valeur de X sera éloignée de ce domaine. Cette situation est très probable dans un contexte où les relations ne peuvent pas être linéaires dans un grand domaine de X (par exemple l'effet de l'âge, de la température, du temps sur des variables physiologiques).

20.4.4 Transformations linéaires

Pour analyser une relation non linéaire, deux possibilités se présentent : la régression non linéaire et, dans certains cas, la transformation linéaire des variables Y et/ou X.

La régression non linéaire est une technique simple sur le plan statistique, mais relativement complexe sur le plan algorithmique[1]. Elle n'est abordée que plus tard dans la formation (1° licence bio).

Dans le domaine bio -médical, trois équations classiques se prêtent à la linéarisation

La fonction puissance $y = ax^b$

caractérise par exemple des relations entre le poids et la longueur. Le coefficient b permet de déterminer un facteur de condition physiologique chez le poisson.

La transformation logarithmique de X et de Y permet de recomposer l'équation d'une droite :

$$\log(y) = \log(a) + b\log(x)$$
$$y'' = a'' + bx''$$

soit graphiquement :

[1] *Algorithme : al -Kharazmi, surnom d'un mathématicien arabe. Suite finie d'opérations élémentaires constituant un schéma de calcul ou de résolution d'un problème.*

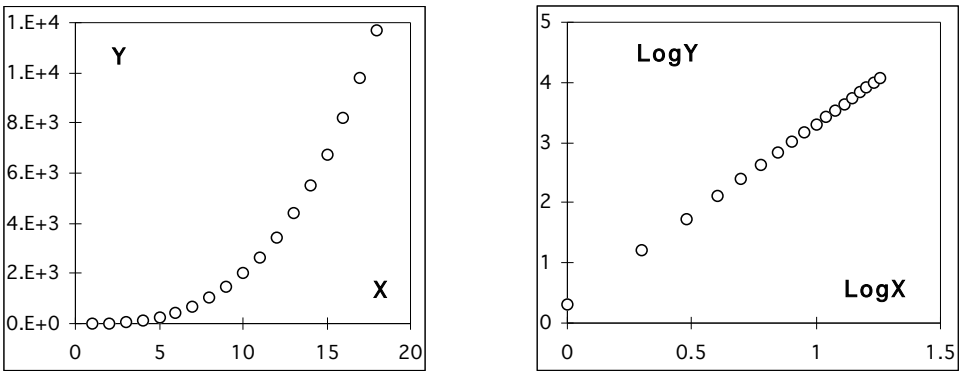

Figure 20 -12 a - Représentation graphique de la fonction $Y = 2X^3$, sans variabilité. b - Représentation graphique de Log Y vs Log X.

La fonction exponentielle $$y = ae^{bx}$$

exprime, par exemple, la croissance d'une population, comme le développement de bactéries dans un bouillon de culture. Le paramètre b détermine le temps de doublement de la population et le paramètre a représente le nombre de cellules au temps t = 0.

Pour un développement plus détaillé, voir le module « fonction exponentielle et logarithmique » sur le site www.fundp.ac.be/biostats

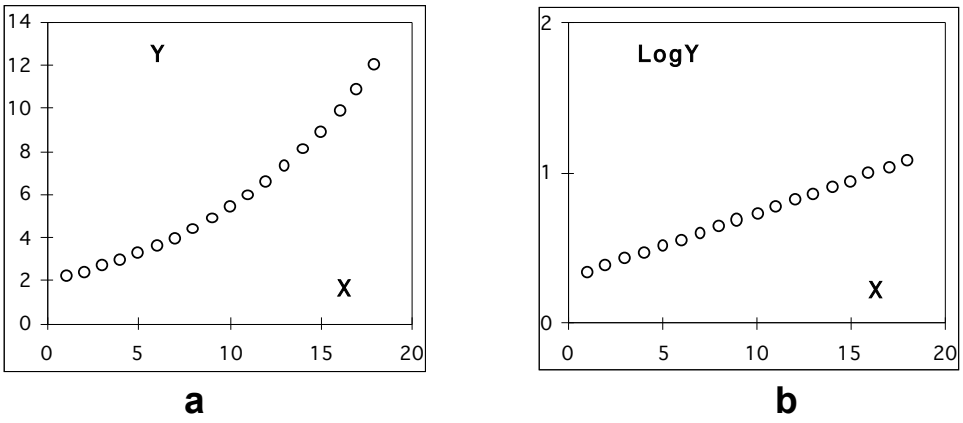

Figure 20 -13 a - Représentation graphique de la fonction $Y = 2e^{0.1X}$, sans variabilité. b - Représentation graphique de Log(Y) vs X.

Dans ce cas, c'est la transformation logarithmique de Y seulement qui produit l'équation de la droite :

$$\log(Y) = \log(a) + bX$$
$$Y" = a" + bX$$

La fonction hyperbolique $Y = \frac{aX}{b+X}$

exprime, par exemple, la vitesse initiale d'une réaction enzymatique en fonction de la concentration en substrat (relation de Michaelis -Menten).

Graphiquement :

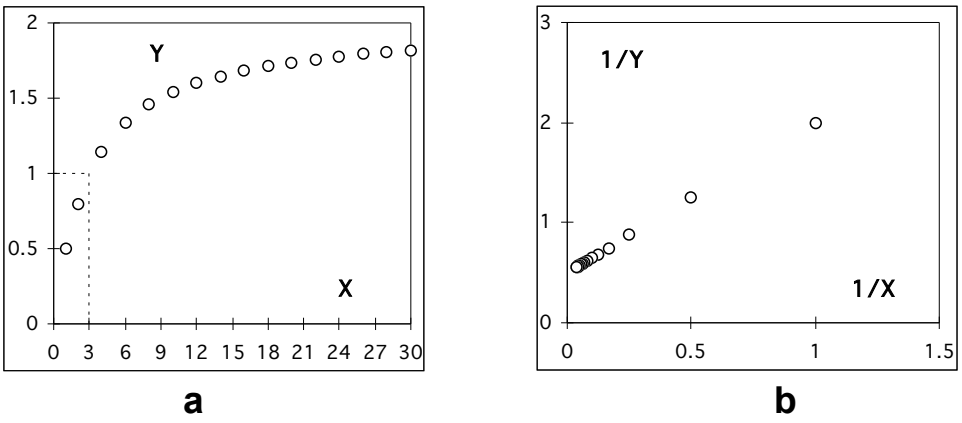

Figure 20 -14 a - Représentation graphique de la fonction $Y = \frac{2X}{3+X}$ *sans variabilité. b - Représentation graphique de l'inverse de Y vs inverse de X.*

Le paramètre a, (appelé Vmax, vitesse maximale) représente la valeur vers laquelle tend Y lorsque X tend vers l'infini. Le paramètre b, (appelé Km, constante d'affinité) représente la valeur de X correspondant à $y = V_{\max}/2$

Dans ce cas, c'est la fonction inverse de Y, inverse de X qui produit l'équation de la droite :

$$Y = \frac{aX}{b+X} \Leftrightarrow \frac{1}{Y} = \frac{b}{aX} + \frac{X}{aX} =$$

$$\frac{1}{a} + \frac{b}{a}\frac{1}{X} \Leftrightarrow Y" = a" + b" X"$$

Les transformations linéaires présentent donc l'avantage incontestable de permettre de traiter des données complexes par un modèle simple. Toutefois, elles provoquent une distorsion des données qui pose des problèmes à l'inférence statistique. On leur préférera donc un traitement par régression non linéaire.

20.4.5 Droite des moindres rectangles

On peut rencontrer des conditions expérimentales qui ne justifient pas le fait de minimiser la variabilité en Y sans minimiser la variabilité de X.

> *Le principe de la droite des moindres carrés est de rendre minimale la variabilité en Y. Ceci se justifie lorsque la variable X est une variable fixée sans erreur par l'expérimentateur et Y une mesure expérimentale affectée d'une variabilité. Dans ce cas, la variabilité minimisée est bien la variabilité résiduelle de l'expérience.*

Dans une échographie par laquelle on estimerait le poids d'un fœtus (Y) en mesurant le diamètre de la tête (X). X est dans ce cas une mesure expérimentale affectée d'une variabilité qui doit également être minimisée.

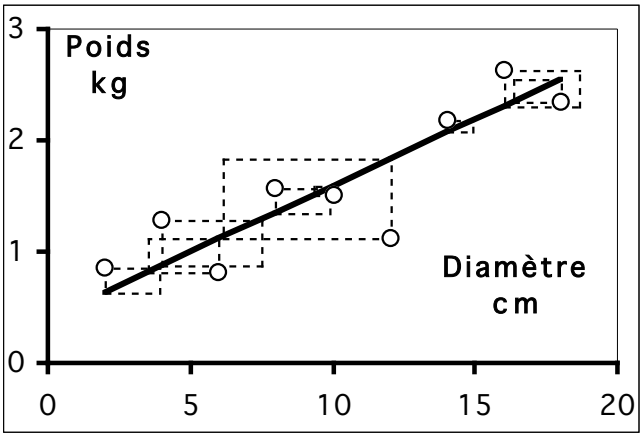

Figure 20 -15 Poids du fœtus vs[2] diamètre de la tête. En pointillés : écarts entre les observations et le modèle, en X et en Y. Relation fictive.

Dans ces conditions, nous devons minimiser simultanément les écarts en X et en Y, représentés par la surface du rectangle qui joint (x_i -x_m) et (y_i -y_m) et cela globalement pour tous les points.

[2] *Vs : abréviation de versus, qui signifie « en fonction de ».*

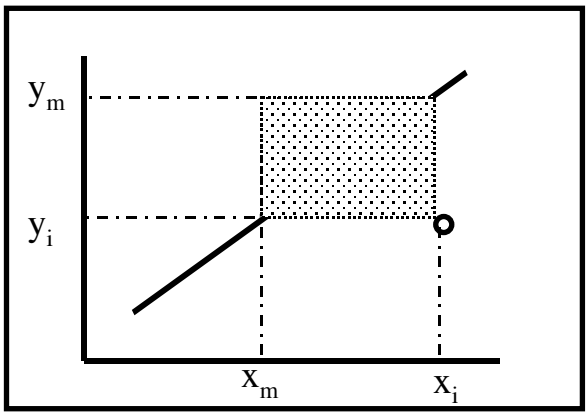

Figure 20 -16 Représentation, pour un point, des écarts (x_i -x_m) et (y_i -y_m) et du rectangle de surface (x_i -x_m).(y_i -y_m).

La quantité à minimiser est la somme sur tous les points des surfaces des rectangles de surface (x_i -x_m) (y_i -y_m), que l'on nomme « somme des produits des écarts au modèle » (SPEm) :

$$SPEm = \sum_{i=1}^{n} (x_i - x_m)(y_i - y_m)$$

cette valeur est minimale pour la valeur des paramètres :

$$B1_{m.r.} = \frac{S_y}{S_x}$$ **Équation 20-5**

$$Bo_{m.r.} = My - B1_{m.r.} Mx$$ **Équation 20-6**

La droite obtenue s'appelle droite des moindres rectangles et est d'autant plus distincte de la droite des moindres carrés que la variabilité est grande.

La droite des moindres carrés et la droite des moindres rectangles ne sont confondues que si R = ±1

Il convient de bien distinguer

- SPExy, calculée sur les écarts entre les observations et leurs moyennes ;
- SPEm, calculée sur les écarts entre les observations et leurs estimations par le modèle.

La droite des moindres rectangles décrit l'échantillon mais ne se prête pas à l'inférence statistique (voir plus loin dans ce cours). La condition « X est fixée sans erreur » est un préliminaire nécessaire à l'étude de la relation entre X et Y dans la population.

20.4.6 Le tableur Excel

propose les fonctions suivantes, dans lesquelles X et Y représentent l'ensemble des cellules contenant les x_i et y_i, respectivement :

Moyenne(X) ou (Y) ;

Var(X) ou (Y) en divisant par n -1
Ecartype(X) ou Y en divisant par n –1 ;

Var.p(X) ou Y en divisant par n
Ecartype.p(X) ou Y en divisant par n.

Somme.carres.ecarts(X) ou (Y) ;

Coefficient.correlation(Y ;X) ;

Pente(Y ;X) coefficient de pente estimé par les moindres carrés ;

La somme des produits des écarts n'est pas pré-programmée, il convient d'encoder la formule somme((X-moyenne(X)(Y-moyenne(Y))*

Ordonnee.origine(Y ;X) estimé par les moindres carrés ;

Sur un graphique XY « nuage de points » se trouve une fonction « ajouter une courbe de tendance » , type : linéaire, options : afficher l'équation, le coefficient de détermination.