Daniel Borcard
Département de sciences biologiques
Université de Montréal

Janvier 2000-2005

Régression linéaire (droites d'estimation)

• *Equation:* $\hat{Y} = aX + b$

• *Test de signification du coefficient a (pente de la droite):* les formules qui s'y rapportent sont données par Scherrer § 18.1.5 p. 637), mais il est tout à fait valide de tester en lieu et place *le coefficient de corrélation linéaire entre les deux variables:*

à l'aide d'un $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Tableau des règles de décision relatives à un test de signification de la pente d'une droite de régression: Scherrer p. 647.

• *Intervalle de confiance de la pente* (Scherrer § 18.1.6, p. 650)*:* il sert surtout à des fins d'inférence (p.ex. pour vérifier qu'une pente prédite par la théorie biologique se trouve à l'intérieur de l'intervalle de confiance calculé pour un seuil de signification donné). Il s'exprime comme suit:

$$\Pr\left[a - t_{\alpha/2}\sqrt{\text{var}(a)} < \alpha < a + t_{\alpha/2}\sqrt{\text{var}(a)}\right] = 1 - \alpha$$

où $\sqrt{\text{var}(a)} = \sqrt{\frac{s_e^2}{(n-1)s_x^2}} = \frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}$

N.B.: la racine carrée de var(*a*) est appelée "erreur type" du coefficient *a*.

Précisons aussi que l'alpha situé au milieu de l'équation désigne la pente de la population (ou pente théorique ou paramétrique), alors que l'alpha situé dans le membre de droite de l'équation désigne le seuil de probabilité choisi pour construire l'intervalle de confiance!

Intervalles de confiance ou de prédiction de la variable expliquée

Il faut distinguer ici trois notions:

1. *L'intervalle de **confiance** d'une "prédiction"* (qu'il vaudrait mieux appeler "estimation") (Scherrer § 18.1.9.2, équation 18-27) définit les limites dans lesquelles se situe probablement une valeur **individuelle** lue sur la droite de régression: lorsqu'on a construit un modèle qui se présente sous la forme d'une droite de régression, l'intervalle de confiance en question dit que, pour une valeur donnée x_i de la variable X, la vraie valeur de la variable Y devrait se situer au sein de cet intervalle de confiance:

$$\Pr\left(\hat{y}_i - t_{\alpha/2}\sqrt{\text{var}(\hat{y}_i)} < \mu_{y_i} < \hat{y}_i + t_{\alpha/2}\sqrt{\text{var}(\hat{y}_i)}\right) = 1-\alpha \quad \text{Sch. eq.18-27}$$

$$\text{où} \quad \text{var}(\hat{y}_i) = s_e^2\left[\frac{1}{n} + \frac{(x_i - x)^2}{\sum (x_i - x)^2}\right] = \frac{(n-1)s_y^2(1-r^2)}{n-2}\left[\frac{1}{n} + \frac{(x_i - x)^2}{(n-1)s_x^2}\right]$$

$$\text{ou encore} \quad \text{var}(\hat{y}_i) = \frac{s_y^2(1-r^2)}{n-2}\left[\frac{n-1}{n} + \frac{(x_i - x)^2}{s_x^2}\right]$$

2. *L'intervalle de **prédiction** d'une "prédiction"* (estimation) définit les limites dans lesquelles tombera vraisemblablement **une nouvelle** observation de Y si elle fait partie de la même population statistique que l'échantillon; la formule pour l'obtenir est la même que l'équation 18-27 de Scherrer (ci-dessus) , à ceci près que $\text{var}(\hat{y}_i)$ doit être remplacé par la quantité:

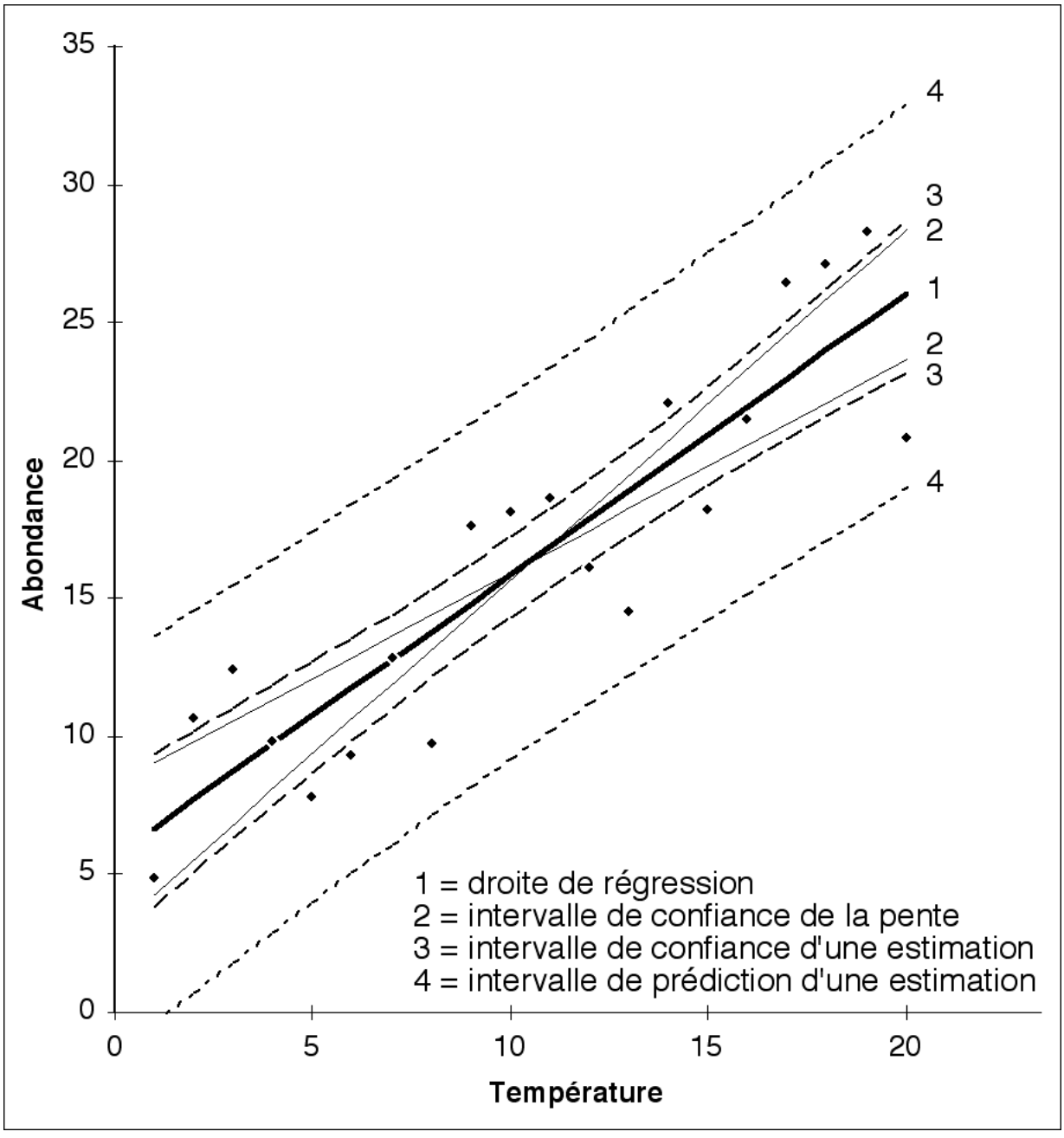
$$s_e^2\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]=\frac{(n-1)s_y^2(1-r^2)}{n-2}\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{(n-1)s_x^2}\right]$$

Cela produit un intervalle plus large que le précédent: en effet, en plus de la variance de l'échantillon qui a servi à établir l'équation, s'ajoute celle qui est associée au tirage d'un nouvel élément.

3. *L'intervalle de confiance de l'estimation de la **moyenne*** de Y pour une valeur particulière de la variable explicative lorsqu'on dispose d'une série de ***m* nouvelles valeurs de Y** pour une seule valeur de X (Scherrer § 18.1.9.3, équation 18-28); cet intervalle est constitué d'une bande plus étroite que la précédente autour de la droite de régression. En effet, au lieu de $\text{var}(\hat{y}_i)$, on utilise la variance de la moyenne estimée de ces nouveaux éléments, $\text{var}(\bar{\hat{y}}_i)$:

$$\text{var}(\hat{y}_i)=s_e^2\left[\frac{1}{m}+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]=\frac{(n-1)s_y^2(1-r^2)}{n-2}\left[\frac{1}{m}+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{(n-1)s_x^2}\right]$$

Cette formule est donc une généralisation de la précédente, où $1/m$ est là pour tenir compte de la variance des m nouveaux éléments.


35
30
25
20
15
10
5
0
0
5
10
15
20
Abondance
Température
4
3
2
1
2
3
4
1 = droite de régression
2 = intervalle de confiance de la pente
3 = intervalle de confiance d'une estimation
4 = intervalle de prédiction d'une estimation

Linéarisation des relations exponentielles

Scherrer: p.669 ; Sokal et Rohlf (1981): p. 541.

• Dans la nature, la croissance d'une population sans contrainte prend une forme exponentielle (loi de Malthus). Une telle courbe, qu'on peut noter

$\hat{y} = be^{ax}$

peut être linéarisée facilement en prenant le logarithme des valeurs de la variable *y:*

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

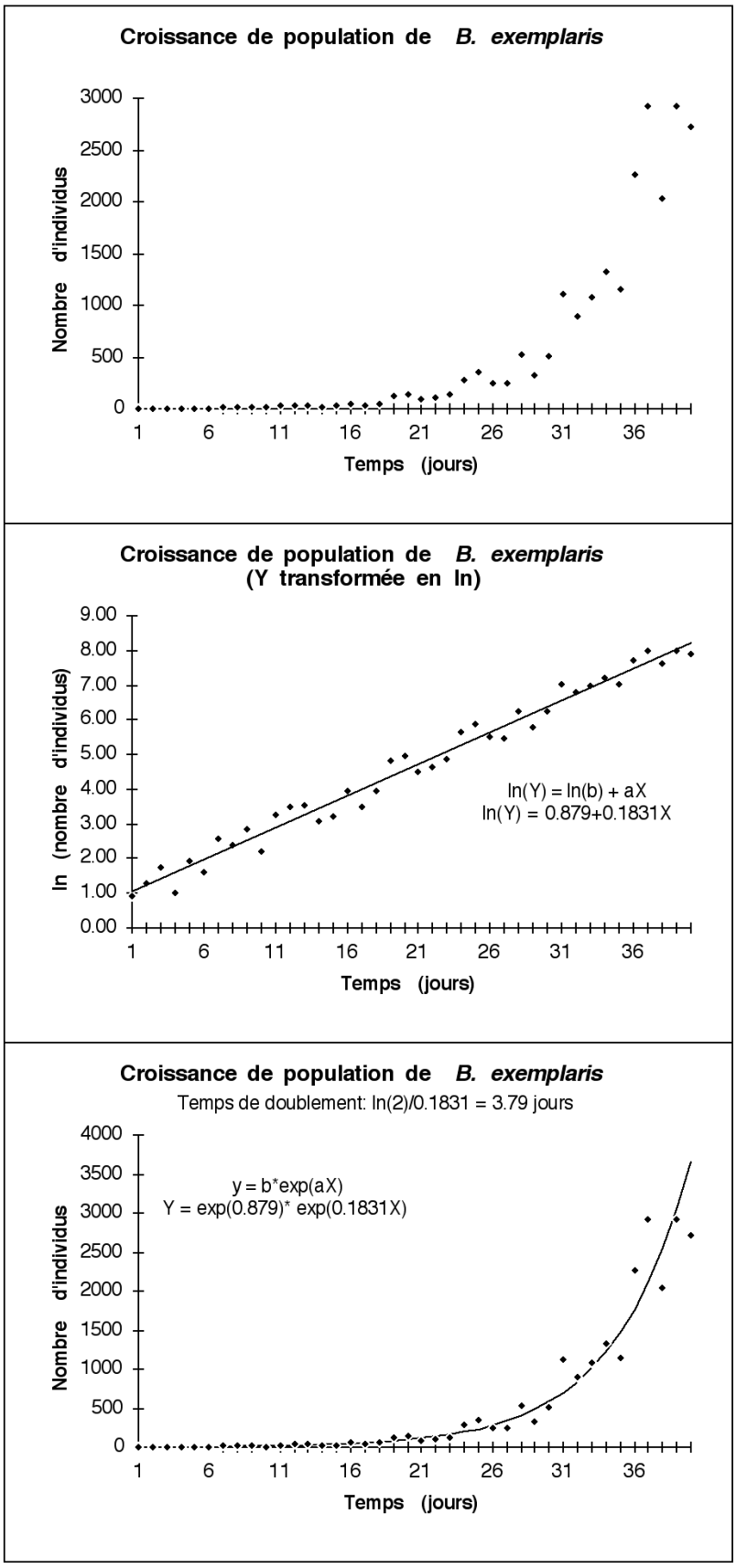
Cela permet de modéliser la relation entre *x* et *y* par régression linéaire, puis de retransformer l'équation pour l'adapter à l'échelle originale (voir graphe d'exemple).

• Attention: c'est la variable *y transformée* [ln(*y*)] qui doit être distribuée normalement, puisque c'est sur elle que s'opère la régression linéaire!

• Une valeur intéressante est la quantité ln(2)/*a*, en particulier lorsque le phénomène étudié est temporel (donc, lorsque *x* mesure le temps):
- lorsque *a* est positif, cette fraction donne le temps que met la variable *y* à doubler de valeur; on l'appelle **temps de doublement** *(doubling time)*;
- lorsque *a* est négatif, cette fraction donne le temps que met la variable *y* à perdre la moitié de sa valeur; on l'appelle **demi-vie** *(half-life)*; exemple classique: la demi-vie d'un élément radioactif.

• Une autre formulation d'équation exponentielle, soit $y = ba^x$, peut aussi être linéarisée facilement par $\log(y) = \log(b) + x \log(a)$.

Linéarisation d'une relation exponentielle



Linéarisation des relations d'allométrie

Sokal et Rohlf (1981): p. 543.

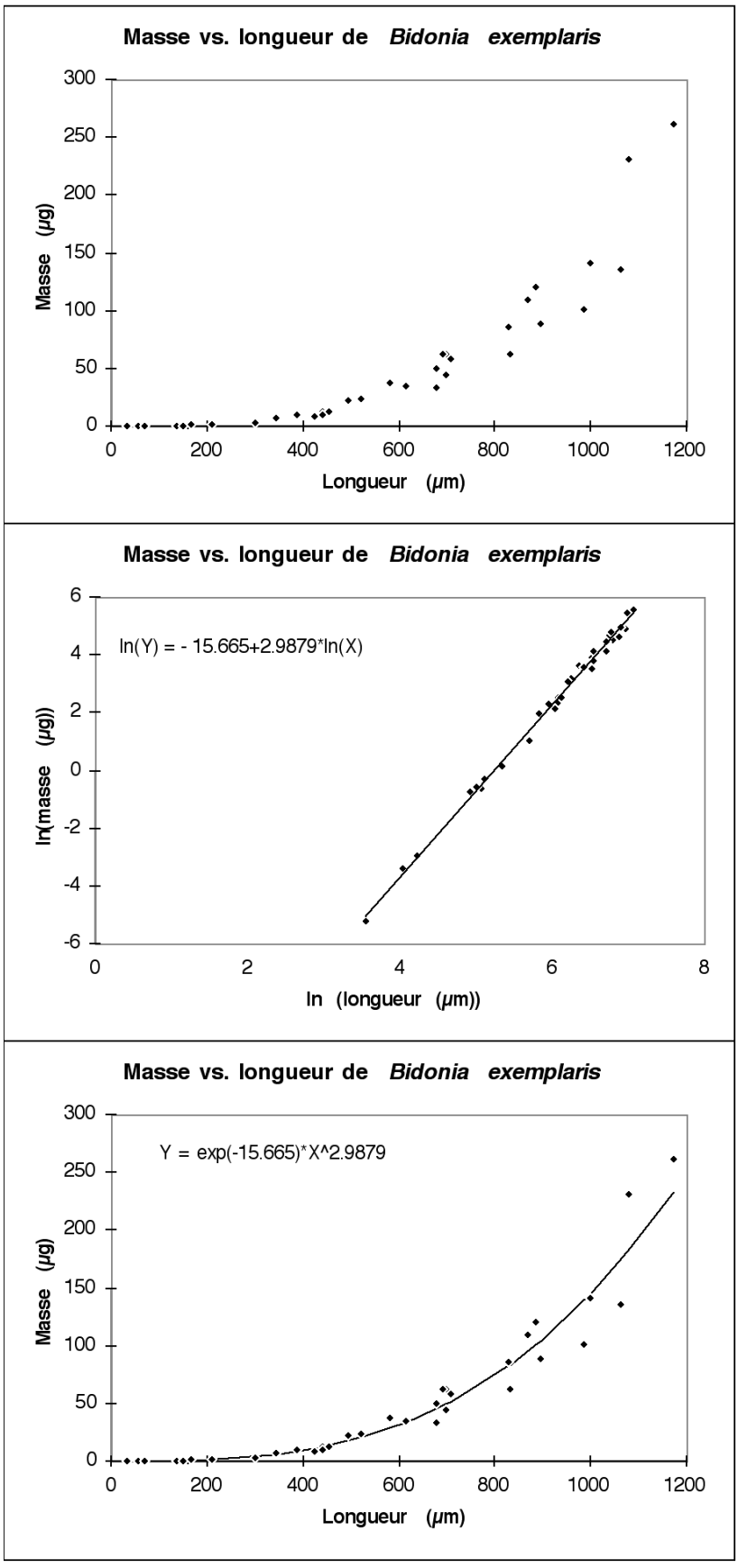
• Lors de la croissance d'un organisme, certaines parties se développent plus rapidement ou plus lentement que d'autres, ce qui altère les proportions générales. On appelle ce phénomène *allométrie.* Un cas du genre est décrit par une relation de type $\hat{y} = bx^a$, rencontrée lorsque le *rapport entre les incréments* de deux structures de tailles différentes reste constant au cours de la croissance; autrement dit, le rapport de *puissance* qui lie deux variables est constant.

• Une transformation logarithmique des *deux* variables produit une droite: $\ln(y) = a\ln(x) + \ln(b)$. Ce sont donc les *valeurs* des variables transformées qui deviennent proportionnelles entre elles [ex.: ln(surface) proportionnel aux 2/3 du ln(masse)].

• La courbe d'allométrie passe par l'origine; *a* est appelé *exposant d'allométrie*, et sa valeur renseigne sur l'orientation de la courbe:
- concave lorsque $a > 1$
- droite lorsque $a = 1$
- convexe lorsque $a < 1$

• Idéalement, les distributions des **deux** variables **transformées** suivent une courbe normale, c'est-à-dire que les distributions des deux variables originelles suivent une distribution lognormale.

Linéarisation d'une relation d'allométrie



Daniel Borcard
Département de sciences biologiques
Université de Montréal

Janvier 2000-2005

Régression linéaire simple de modèle II

Sokal & Rohlf (1981) p.547 et suiv.; 594 et suiv.; Legendre & Legendre (1998): p.504 et suiv.

La régression linéaire simple ordinaire (moindres carrés ordinaires, MCO) minimise les sommes de carrés d'écarts dans une seule direction (verticale dans le cas le plus général d'estimation de *Y* à partir de *X*). On utilise cette méthode dans les cas suivants:

- lorsque la variable explicative est contrôlée (non aléatoire) ou que sa variation aléatoire est très faible par rapport à celle de la variable à expliquer;
- lorsqu'on a une raison claire de postuler laquelle des deux variables influence l'autre;
- lorsque l'unique but du calcul est de prévoir une variable à l'aide de l'autre.

Dans les autres cas, la régression ordinaire n'est pas appropriée: il y a deux droites possibles, leurs réponses peuvent être contradictoires, et on ne sait laquelle choisir.

Lorsque les deux variables *X* et *Y* sont sujettes à des fluctuations aléatoires, on se trouve en situation dite de **régression de modèle II**, qui implique l'usage d'un groupe de méthodes dont la **régression orthogonale** (ou **régression par l'axe majeur**, AM) fait partie. Les exemples abondent en biologie: relation entre deux mesures morphométriques sur des individus de même stade de développement; comparaison des prédictions d'un modèle avec des valeurs observées; relation entre les abondances de deux espèces qui s'influencent mutuellement; ...etc.

Au lieu de minimiser les carrés d'écarts selon un axe vertical ou horizontal, la droite de régression orthogonale minimise les carrés des écarts perpendiculairement à elle-même, inpliquant par conséquent les **deux** variables dans le calcul des résidus. Cette droite de régression est le grand axe des ellipses d'égale densité de probabilité

d'une distribution normale bidimensionnelle (ou encore, l'axe majeur est la première **composante principale** du nuage de points formé par les deux variables étudiées; l'axe mineur, qui lui est orthogonal, constitue la deuxième composante principale).

On peut aussi se représenter cela comme un changement de système de référence consistant en une translation de l'origine vers le centroïde du nuage de points, suivie d'une rotation des axes. Dans le système de coordonnées "Axe majeur - axe mineur", les coordonnées des observations sur un axe sont linéairement indépendantes (= non corrélées) des coordonnées sur l'autre.

La **pente de l'axe majeur** peut être estimée de différentes manières, la plus simple étant probablement celle-ci:

$$b_{AM} = \frac{d \pm \sqrt{d^2 + 4}}{2}$$

où $d = \frac{b^2 - r^2}{br^2}$

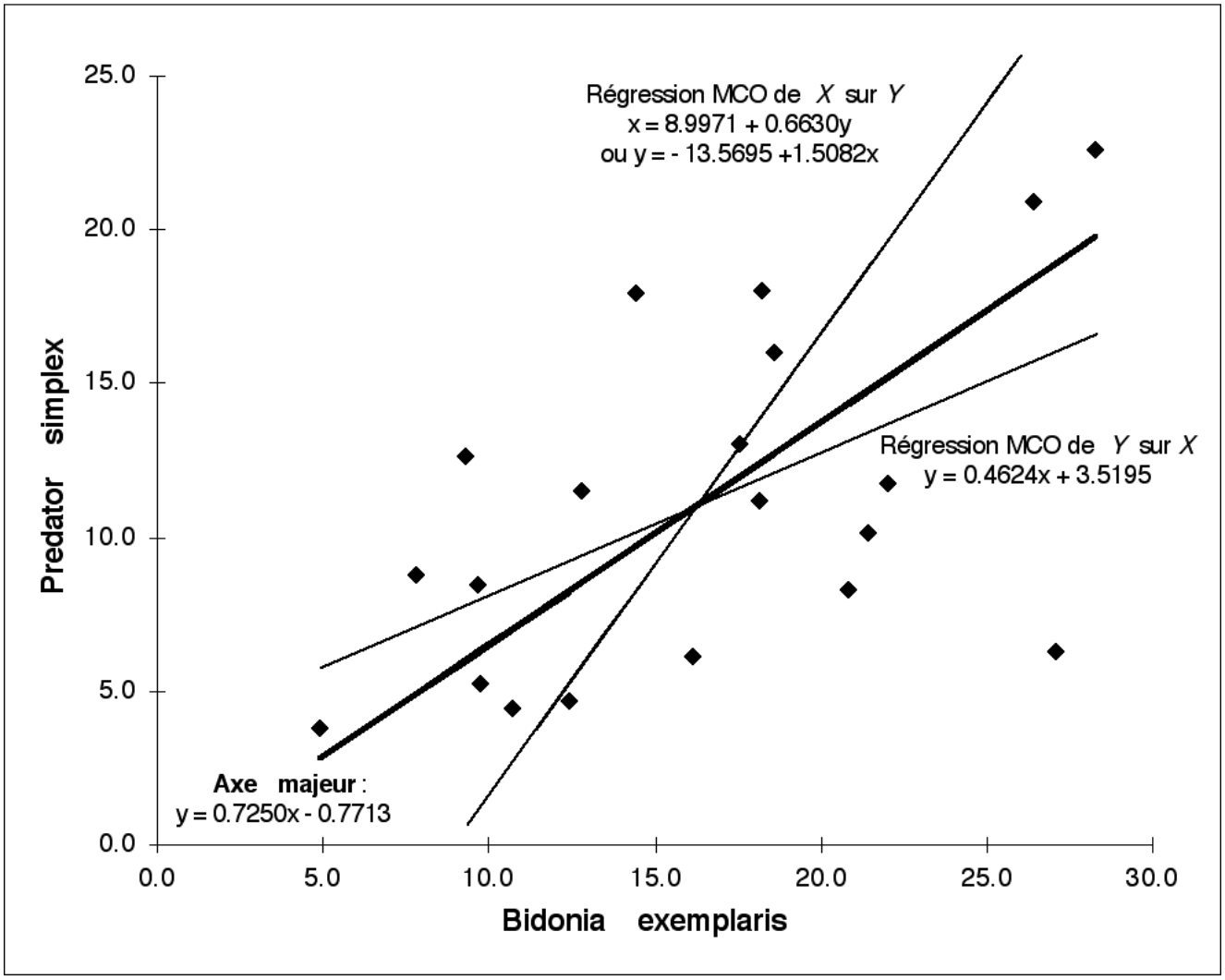
b étant le coefficient de régression linéaire ordinaire (de modèle I) de *Y* sur *X* et *r* étant le coefficient de corrélation de Pearson entre *X* et *Y*. Dans le "plus ou moins" de l'équation, le signe "+" est utilisé lorsque la corrélation *r* est positive, et inversement. **Attention**: remarquez que la convention la plus courante, utilisée ici, symbolise **la pente par b_{AM}** et **l'ordonnée à l'origine par b_0**!

L'ordonnée à l'origine (le deuxième paramètre de l'équation de régression) vaut:

$$b_0 = \bar{y} - b_{AM}\bar{x}$$

où *x* barre et *y* barre sont les moyennes des deux variables *X* et *Y*.

Régression par l'axe majeur (données brutes): AM



Intervalle de confiance

Un intervalle de confiance de la pente de l'axe majeur peut être calculé. Cet intervalle se construit sur la base d'un *F* avec 1 et *n*–2 degrés de liberté. Son calcul est assez laborieux (mais pas complexe), et se fonde sur les variances des deux variables impliquées dans la régression, symbolisées s_X^2 et s_Y^2 , et sur leur covariance s_{xy}. Les étapes de calcul, tirées de Sokal & Rohlf (1981), sont données ci-dessous, et comprennent une autre façon de calculer les pentes des axes majeur et mineur (ce dernier étant perpendiculaire à l'axe majeur).

$$D = \sqrt{\left(s_X^2 + s_Y^2\right)^2 - 4\left(s_X^2 s_Y^2 - \left(s_{XY}\right)^2\right)}$$

$$\lambda_1 = \frac{\left(s_X^2 + s_Y^2 + D\right)}{2} \qquad \lambda_2 = \frac{\left(s_X^2 + s_Y^2 - D\right)}{2}$$

Pente de l' axe majeur : $b_{(AM)} = \frac{s_{xy}}{\left(\lambda_1 - s_Y^2\right)}$

Pente de l' axe mineur = $-\frac{1}{b_{AM}}$

$$H = \frac{F_{(\alpha;1;(n-2)}}{\left(\lambda_1/\lambda_2 + \lambda_2/\lambda_1 - 2\right)(n-2)}$$

$$A = \sqrt{H/(1-H)}$$

Bornes de l' intervalle de confiance :

$$b_{AM1} = \frac{b_{AM} - A}{1 + \left(b_{AM} \times A\right)} \qquad b_{AM2} = \frac{b_{AM} + A}{1 - \left(b_{AM} \times A\right)}$$

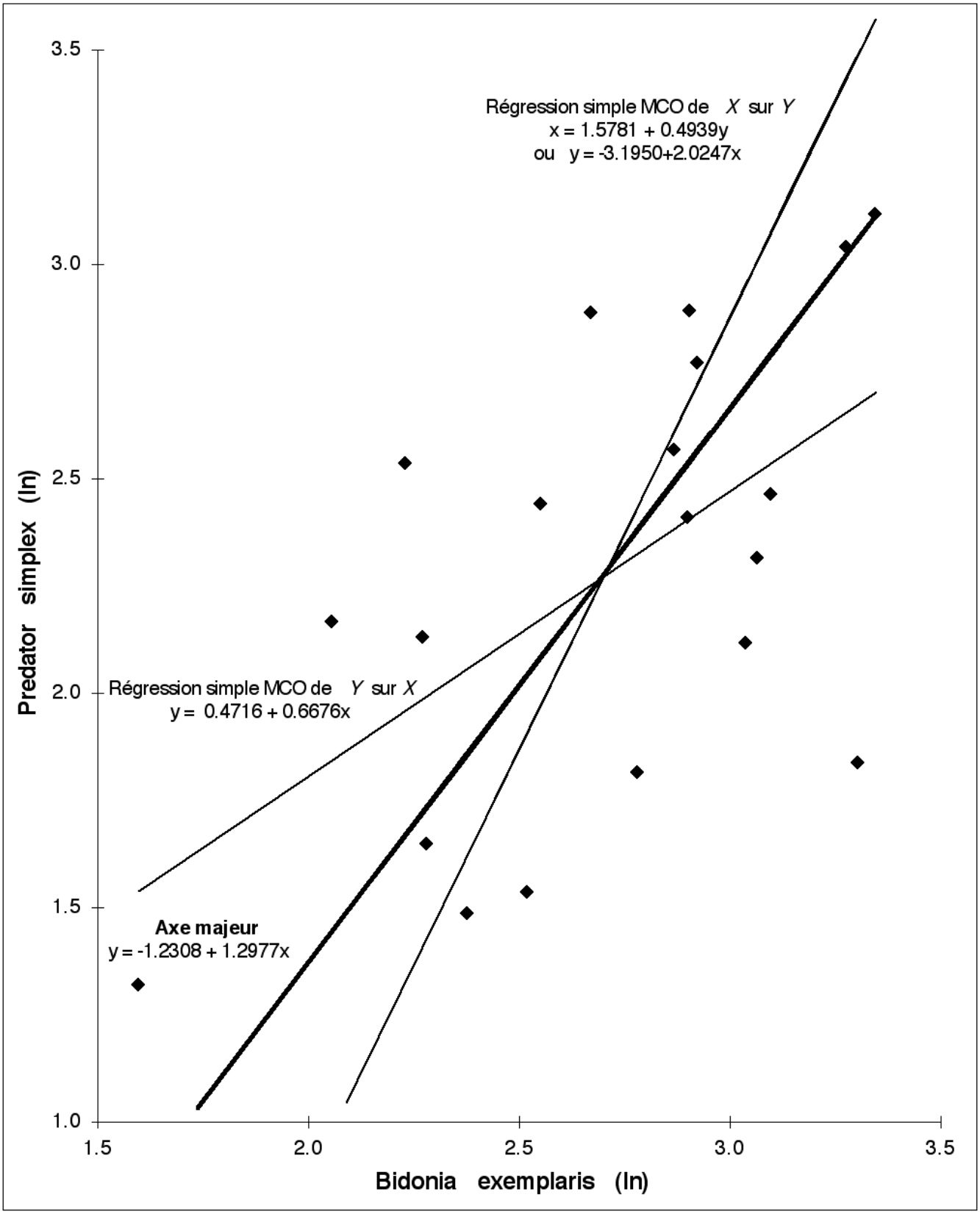
L'intervalle de confiance se comporte de manière différente suivant si la corrélation linéaire entre les deux variables est significative ou non. Lorsque *r* est significatif, les deux bornes b_{AM1} et b_{AM2} de l'intervalle de confiance se situent du même côté (positif ou négatif suivant le signe de la relation) du point zéro. Lorsque *r* n'est pas significatif, la pente peut varier de diverses manières: soit l'intervalle de confiance chevauche la valeur "pente zéro", soit il chevauche la valeur "pente verticale (infinie)" (l'intervalle de confiance est alors disjoint autour de la valeur zéro), soit la pente peut prendre n'importe quelle valeur de + ∞ à - ∞.

Axe majeur sur données transformées en ln

Un défaut de l'axe majeur est sa sensibilité aux changements linéaires inégaux d'échelle (changement d'échelle différent pour une variable que pour l'autre). La transformation en logarithmes naturels des deux variables y remédie. De plus, elle tend à minimiser la différence d'ordre de grandeur des variables, à laquelle la méthode de l'axe majeur est sensible également. Le résultat est que l'axe majeur se rapproche souvent de la bissectrice de l'angle formé par les deux droites de régression ordinaire (voir illustrations).

Cette transformation est à recommander surtout lorsque les deux variables tendent à suivre une distribution lognormale bidimensionnelle.

Régression par l'axe majeur sur données transformées en ln



Axe majeur des variables centrées-réduites ou axe majeur réduit (AMR)

Le centrage-réduction des variables X et Y rend aussi l'axe majeur invariant aux changements d'échelle et moins sensible aux différences d'ordre de grandeur. La pente de l'axe majeur calculé sur les variables centrées réduites est égale à +1 ou –1 suivant que le coefficient de corrélation est positif ou négatif. La droite passe par l'origine, de sorte qu'il n'y a pas de paramètre b_0 (ordonnée à l'origine) dans l'équation.

Dans la méthode AMR, la pente et l'ordonnée à l'origine sont ramenées à l'espace des variables brutes (c'est-à-dire des variables telles qu'elles se présentent avant le centrage et la réduction). La pente se calcule par le quotient des écarts types des deux variables:

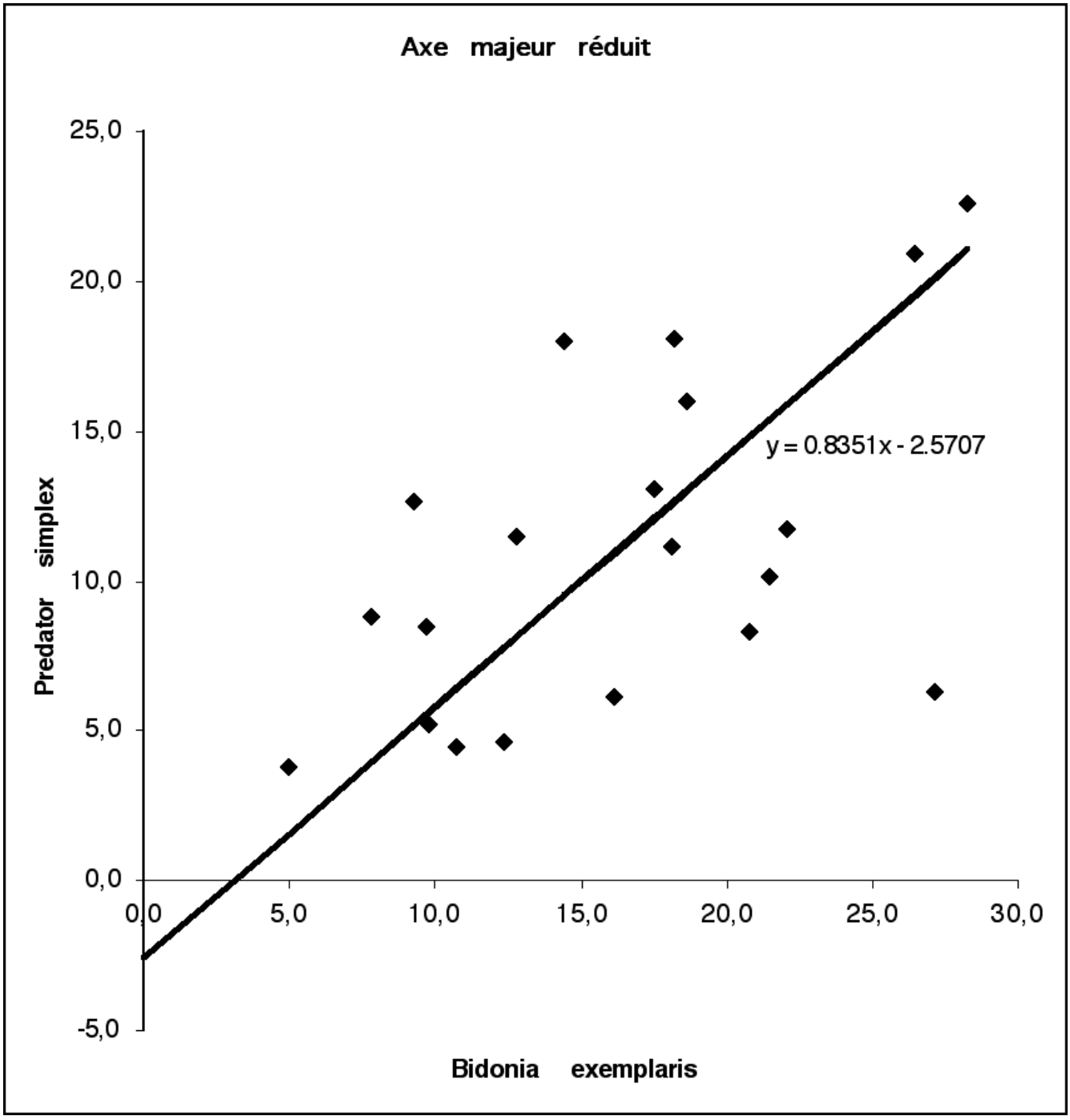
$$b_{AMR} = \pm \frac{s_y}{s_x}$$

la pente étant positive ou négative suivant le signe de la corrélation entre les deux variables.

L'ordonnée à l'origine s'obtient par

$$b_0 = \bar{y} - b_{AMR}\bar{x}$$

Attention: la méthode de l'axe majeur réduit doit être évitée lorsque la corrélation entre les variables est faible ou que le nombre d'observations est petit (l'axe majeur ordinaire est plus tolérant).


Axe majeur réduit
y = 0.8351x - 2.5707
Predator simplex
Bidonia exemplaris
25,0
20,0
15,0
10,0
5,0
0,0
-5,0
0,0
5,0
10,0
15,0
20,0
25,0
30,0

Axe majeur sur données cadrées (AMDC)

Une autre transformation utile lorsqu'on veut analyser par l'axe majeur des données qui ne sont pas exprimées dans les mêmes unités consiste à **cadrer** les deux variables entre 0 et 1:

$$y'_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \quad \text{et} \quad x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

où l'indice "*min*" signifie la valeur minimale observée dans l'échantillon, et l'indice "*max*" la valeur maximale. Remarque: dans le cas où les variables sont construites sur une échelle comportant un zéro vrai (ex.: température en degrés Kelvin ou dénombrements d'espèces), x_{min} et y_{min} valent zéro, ce qui réduit les deux formules ci-dessus à

$$y'_i = \frac{y_i}{y_{\max}} \quad \text{et} \quad x'_i = \frac{x_i}{x_{\max}}$$

L'axe majeur **ordinaire** (AM) est calculé sur les variables ainsi transformées, puis la pente et les limites de son intervalle de confiance (calculées comme pour l'axe majeur ordinaire) sont ramenées à leurs unités originelles en les multipliant par le rapport des intervalles des deux variables:

$$b_{AM} = b'_{AM} \frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}}$$

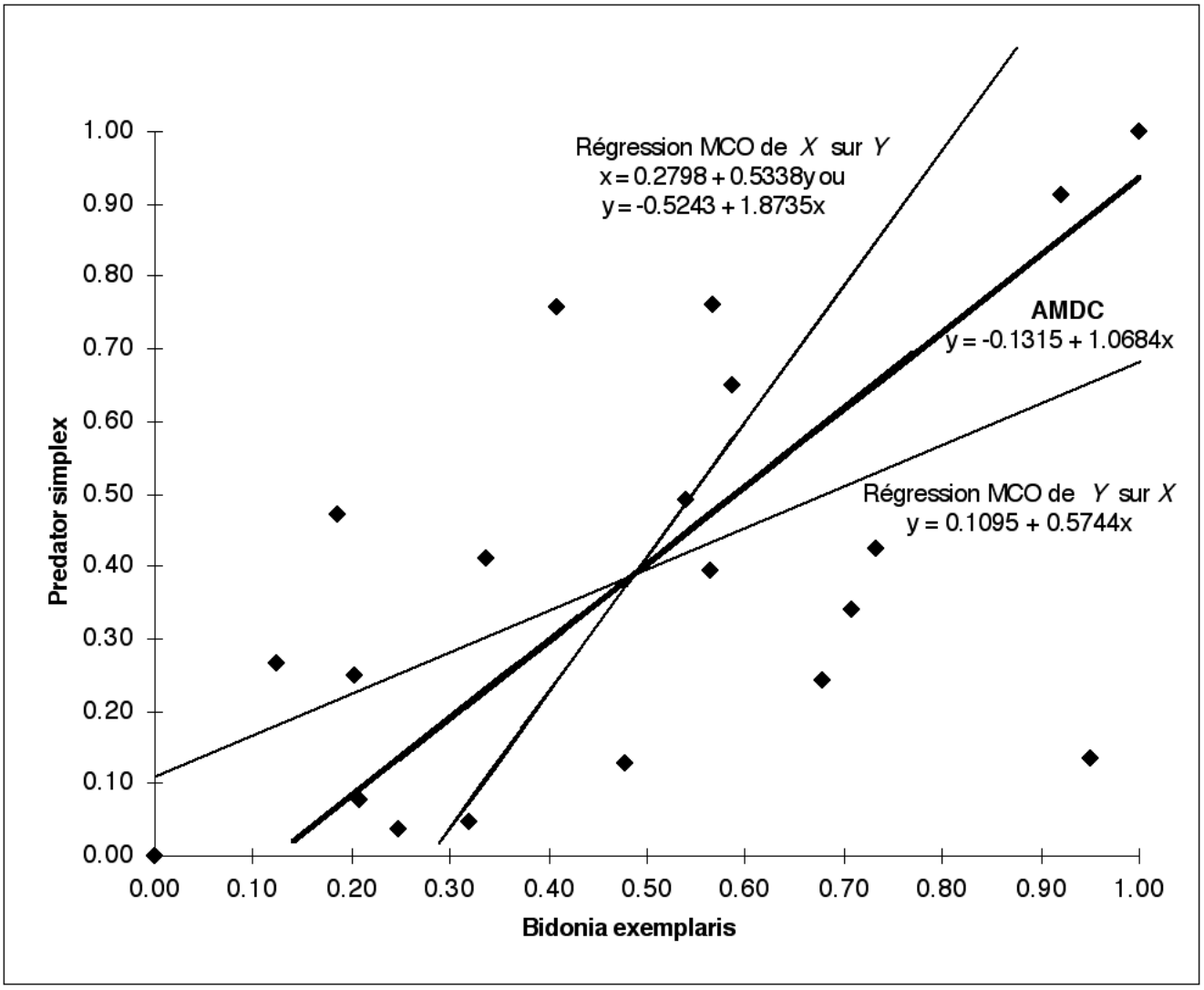
L'ordonnée à l'origine b_0 est recalculée à l'aide du centroïde originel du nuage de points $(\bar{x};\bar{y})$ ainsi que l'estimation de la pente b_{AM} (retransformée):

$$b_0 = \bar{y} - b_{AM}\bar{x}$$

L'estimateur de la pente est donc proportionnel aux unités de *X* et *Y*, mais ne perd pas sa sensibilité aux covariances (comme le fait l'axe majeur réduit).

• Attention toutefois: l'AMDC doit être évité s'il y a des valeurs aberrantes dans l'une ou l'autre variable.

Régression par l'axe majeur sur données cadrées (AMDC)



Recommandations pour l'usage de méthodes d'estimation de pentes de relations linéaires dans le cas où les deux variables sont aléatoires. Ces recommandantions sont tirées du manuel du programme de régression de modèle II de Pierre Legendre [Legendre, P. 1999. Régression de modèle II - Guide. Département de sciences biologiques, Université de Montréal, 25 p.]

Le programme de régression de type II (pour Mac et PC) et son guide (en français et anglais) sont disponibles sur le site <http://www.bio.umontreal.ca/legendre/>

Tenant compte de plusieurs études par simulation, Legendre & Legendre (1998) recommandent la démarche suivante pour choisir une méthode permettant d'estimer les paramètres de la relation fonctionnelle linéaire reliant deux variables aléatoires mesurées avec erreur (tableau 1).

1. Si la variation aléatoire (c.-à-d. la variance de l'erreur) de la variable-réponse y est beaucoup plus forte (plus de trois fois) que celle de la variable explicative *x*, utiliser la méthode des moindres carrés ordinaires (MCO) (remarque: contrairement à la variance de l'échantillon, la variance de l'erreur ne peut être estimée à partir des données. On ne peut l'estimer qu'en considérant la méthode qui a été employée pour mesurer les variables *x* et *y*). Sinon:

2. Vérifier si les données sont approximativement binormales par l'examen d'un diagramme de dispersion ou par un test de signification. Sinon, tenter une transformation pour rendre la distribution binormale. Si les données sont ou peuvent être rendues binormales, voir les recommandations 3 et 4. Sinon passer à 5.

3. Lorsque les deux variables sont exprimées dans les mêmes unités physiques ou sont sans dimension (p. ex. des variables ayant subi une transformation log) et que la variance de l'erreur est à peu près la même pour les deux variables, utiliser l'axe majeur (AM).

Lorsqu'on ne possède aucune information sur le rapport de la variance des résidus alors qu'il n'y a pas de raison de croire que ce rapport puisse différer de 1, on peut encore employer AM si les résultats sont interprétés avec prudence. La méthode AM produit des estimations non biaisées de la pente et des intervalles de confiance aux bornes exactes (Jolicoeur, 1990).

La méthode AM peut être employée avec des variables qui sont dimensionnellement hétérogènes si l'objectif de l'analyse est (1) de comparer la pente de la relation fonctionnelle entre deux variables (les mêmes) mesurées dans des conditions différentes (p. ex. à deux ou plusieurs sites d'échantillonnage), ou encore (2) de tester l'hypothèse que l'axe majeur ne diffère pas significativement d'une valeur fournie par hypothèse (p. ex. la relation $E = b_1 m$ où, selon la célèbre équation d'Einstein, $b_1 = c^2$, c étant la vitesse de la lumière dans le vide).

4. Pour des données binormales, si la régression AM ne peut pas être employée parce que les variables ne sont pas exprimées dans les mêmes unités physiques ou parce que la variance de l'erreur des deux variables diffère, il reste deux méthodes pour estimer les paramètres de l'équation fonctionnelle linéaire si on peut raisonnablement supposer que la variance de l'erreur de chaque axe est proportionnelle à la variance de la variable correspondante, i.e. si (la variance de l'erreur de *y* / la variance de *y*) ≈ (la variance de l'erreur de *x* / la variance de *x*). Cette condition est souvent satisfaite par des dénombrements d'organismes (p. ex. le nombre de plantes ou d'animaux) ou par des variables ayant subi une transformation log (McArdle, 1988).

4.1. On peut employer l'axe majeur des données cadrées (AMDC). La méthode est décrite ci-dessous. Attention, cependant, aux valeurs aberrantes ou extrêmes; on peut identifier celles-ci grâce à un diagramme de dispersion des objets.

4.2. On peut employer l'axe majeur réduit (AMR). Il faut d'abord tester la signification de la corrélation (*r*) entre les deux variables afin de déterminer si l'hypothèse de l'existence d'une liaison entre les deux variables est supportée par les données. Il n'y a pas lieu de calculer une équation AMR si la corrélation n'est pas significative.

L'AMR demeure une solution boiteuse puisqu'on ne peut pas tester la signification de sa pente. L'intervalle de confiance doit aussi être employé avec circonspection car, comme on l'a montré à l'aide de

simulations, à mesure que la pente réelle de la distribution s'éloigne de ±1, l'estimation de la pente par l'AMR devient de plus en plus biaisée et l'intervalle de confiance inclut la valeur réelle de la pente de moins en moins souvent. Même lorsque la pente de la distribution est proche de ±1 (p. ex. dans l'exemple 5), l'intervalle de confiance AMR est trop étroit si l'effectif (*n*) est très petit ou si la corrélation est faible.

5. Si la distribution des données n'est pas et ne peut être rendue binormale (p. ex. si la distribution possède deux ou plusieurs modes), il faut se demander si la pente d'une droite de régression est un modèle adéquat pour décrire la relation fonctionnelle entre les deux variables. Puisque la distribution n'est pas binormale, il est incorrect d'employer AM, AMR ou AMDC puisque ces modèles décrivent la première composante principale d'une distribution binormale. (a) Si la relation est linéaire, il vaut mieux utiliser les moindres carrés ordinaires (MCO); il faut alors tester la pente par permutation puisque la condition de binormalité n'est pas remplie. (2) Si une ligne droite ne semble pas être un modèle adéquat pour décrire la relation entre les variables, il vaut mieux employer la régression polynomiale ou la régression non-linéaire.

6. Si le but de l'étude n'est pas d'estimer les paramètres d'une relation fonctionnelle, mais plutôt de prévoir ou prédire les valeurs de *y* à partir de *x*, il faut employer la méthode MCO. C'est la seule méthode qui minimise la somme des carrés des résidus en y. La droite de régression MCO elle-même n'a aucune signification, de même que l'erreur type du coefficient de régression et son intervalle de confiance, à moins que *x* n'ait été mesuré sans erreur (Sokal & Rohlf, 1995: 545, tableau 14.3); cet avertissement s'applique en particulier à l'intervalle de confiance de la pente MCO calculé par le programme de Pierre Legendre.

7. Dans certaines recherches on désire comparer des observations aux prédictions d'un modèle. Si le modèle contient des variables sujettes à fluctuation aléatoire, il faut utiliser l'axe majeur (MA) pour cette comparaison puisque les valeurs observées de même que les prédictions du modèle devraient être dans les mêmes unités physiques.

Si le modèle colle bien aux données, on s'attend à trouver une pente de 1 et une ordonnée à l'origine de 0. Si la pente diffère significativement de 1, cela indique que la différence entre les valeurs observées et simulées est proportionnelle aux valeurs observées. Pour des variables à échelle de variation relative à un vrai zéro, si l'ordonnée à l'origine diffère significativement de 0, cela suggère l'existence d'une différence systématique entre les valeurs observées et simulées (Mesplé et al., 1996).

8. Attention: dans toutes ces méthodes, l'intervalle de confiance est grand lorsque le nombre d'observations est faible. Il diminue à mesure que *n* augmente jusqu'à 60 environ; après cela, il change beaucoup plus lentement. La régression de modèle II ne devrait idéalement être employée qu'avec des jeux de données de 60 observations ou plus.

Bibliographie:

Jolicoeur, P. 1990. Bivariate allometry: interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. Journal of Theoretical Biology 144: 275-285.

McArdle, B. 1988. The structural relationship: regression in biology. Canadian Journal of Zoology 66: 2329-2339.

Mesplé, F., M. Troussellier, C. Casellas & P. Legendre. 1996. Evaluation of simple statistical criteria to qualify a simulation. Ecological Modelling 88: 9-18.

Tableau I. Application des méthodes de régression de modèle II. Les numéros dans la colonne de gauche se réfèrent aux paragraphes correspondants dans le texte de recommandations.

§	Méthode	Conditions d'application	Test possible
1	MCO	L'erreur de *y* >> l'erreur de *x*	Oui
3	AM	La distribution est binormale	Oui
		Les variables sont dans les mêmes unités physiques ou sans dimension	
		La variance de l'erreur est à peu près la même pour *x* et *y*	
4		La distribution est binormale	
		La variance de l'erreur de chaque axe est proportionnelle à la variance de la variable correspondante	
4.1	AMDC	Vérifier le diagramme de dispersion: pas de valeurs aberrantes ou extrêmes	Oui
4.2	AMR	La corrélation *r* est significativement différente de zéro	Non
5	MCO	La distribution n'est pas binormale	Oui
		La relation entre *x* et *y* est linéaire	
6	MCO	On désire calculer des valeurs prévues ou prédites par l'équation	Oui
		(L'équation de régression ne peut être utilisée pour d'autres buts et les intervalles de confiance sont inutilisables)	
7	AM	Pour comparer des observations aux prédictions d'un modèle	Oui

Daniel Borcard 2001-2005
Département de sciences biologiques
Université de Montréal

Régression multiple

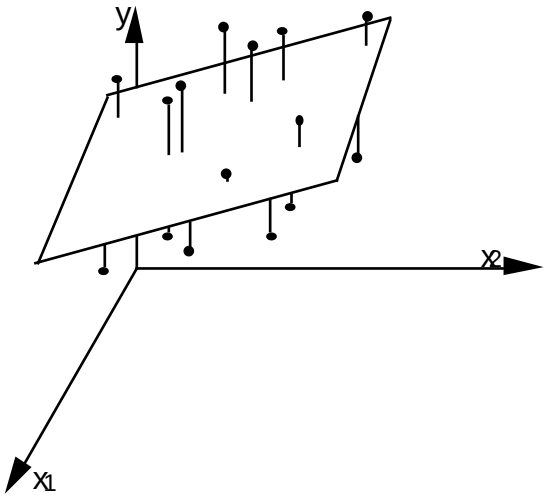
Scherrer: p.690; Sokal & Rohlf: p. 617; Legendre & Legendre (1998) p. 517

Il arrive souvent qu'on veuille expliquer la variation d'une variable dépendante par l'action de **plusieurs** variables explicatives.

Exemple: l'abondance de *Bidonia exemplaris* (y) est influencée par le taux d'humidité (x_1) et par le pourcentage de matière organique dans le sol (x_2).

Lorsqu'on a des raisons de penser que la relation entre ces variables est linéaire (faire des diagrammes de dispersion!), on peut étendre la méthode de régression linéaire simple à **plusieurs** variables explicatives; s'il y a **deux variables explicatives**, le résultat peut être visualisé sous la forme d'un **plan** de régression dont l'équation est:

$$\hat{y} = a_1 x_1 + a_2 x_2 + b$$



Le plan est ajusté selon le principe des **moindres carrés** où les sommes des carrés des erreurs d'estimation de la variable **dépendante** (on a donc affaire à une régression de modèle I) sont minimisées.

S'il y a plus que deux variables explicatives (p. ex. p–1), on peut étendre la méthode en ajoutant les variables et leurs paramètres:

$$\hat{y} = a_1 x_1 + a_2 x_2 + ... + a_j x_j + ... + a_{p-1} x_{p-1} + b$$

Cette équation est celle d'un **hyperplan** à p–1 dimensions (qu'on ne peut pas se représenter concrètement!). Les paramètres a_1, $a_2 \ldots a_{p\text{-}1}$ sont les "pentes" de l'hyperplan dans les dimensions considérées, et sont appelés "coefficients de régression".

La régression multiple peut être utilisée à plusieurs fins:

• Trouver la meilleure équation linéaire de prévision (modèle) et en évaluer la précision et la signification.

• Estimer la contribution **relative** de deux ou plusieurs variables explicatives sur une variable à expliquer; déceler l'effet complémentaire ou, au contraire, antagoniste entre diverses variables explicatives.

• Estimer l'importance relative de plusieurs variables explicatives sur une variable dépendante, en relation avec une théorie causale sous-jacente à la recherche (attention aux abus: une corrélation n'implique pas toujours une causalité; cette dernière doit être postulée *a priori*).

Le calcul des coefficients de régression est détaillé par Scherrer (p. 693-699). Il se base sur un système de p–1 équations à p–1 inconnues (au bas de la p. 695) qui permet dans un premier temps d'obtenir les "coefficients de régression centrés et réduits" (voir plus bas: c'est comme si on calculait la régression sur les variables centrées-réduites). **Attention: dans la notation de Scherrer, la *p*-ième variable est la variable dépendante (*y*)**. Les valeurs des coefficients de régression pour les variables brutes (non centrées-réduites) sont ensuite obtenues par multiplication par le rapport des écarts-types de la variable dépendante et de la variable explicative considérée (voir bas p. 698). Finalement, on calcule la valeur de l'ordonnée à l'origine. Voir aussi en fin de ce document.

Exemple d'une équation de régression multiple à deux variables explicatives x_1 et x_2:

$$\hat{y} = 0.5543x_1 + 0.7211x_2 - 41.6133$$

Si on remplace les symboles des variables par leur nom dans le "monde réel", on a:

Abond. *Bidonia* = 0.5543×Humid. + 0.7211×M.O. – 41.6133

Les signes des paramètres a_1 et a_2 sont tous deux positifs, ce qui montre que *Bidonia* réagit positivement à une augmentation du taux d'humidité et de la teneur en matière organique.

Cette équation peut servir à estimer l'abondance de *B. exemplaris* en fonction des deux descripteurs "Humidité" et " Matière organique" (exprimés en % dans cet exemple).

Pour une humidité de 80% et un taux de matière organique de 30%, on estime l'abondance de *B. exemplaris* à

Abond. *B.ex.* = 0.5543×80 + 0.7211×30 – 41.6133 = 24.3637 ind.

Comme en régression linéaire simple, on mesure la **variance expliquée** par la régression à l'aide du **coefficient de détermination multiple R^2**:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{SCER}}{\text{SCET}}$$ (et **non** comme dans Scherrer p.699!!)

Remarques:

- Scherrer (paragr. 18.3.3 p. 699) appelle le R^2 "coefficient de corrélation multiple". C'est faux. Le coefficient de corrélation multiple est défini comme la **racine carrée** du coefficient de détermination multiple;

- l'équation du R^2 donnée par Scherrer dans le même paragraphe est fausse. C'est celle ci-dessus qui est la bonne. Elle est conforme aux définitions correctes de SCER et SCET données par Scherrer p. 635.

Le R^2 peut aussi se calculer à partir des coefficients de régression centrés-réduits a'_j et des coefficients de corrélation entre la variable dépendante y et chacune des variables explicatives x_j. Voir plus loin.

Test de signification du modèle de régression multiple

La **signification** du modèle de régression multiple peut être **testée** par une variable auxiliaire F_{RMc} qui, sous H_0, est distribuée comme un F de Fisher à (p–1) et (N–p) degrés de liberté. Rappelons que dans cette notation (celle de Scherrer), p désigne le nombre de variables explicatives **plus une**, c'est-à-dire le nombre de paramètres de l'équation: coefficients de régression plus l'ordonnée à l'origine.

Les hypothèses du test sont:

H_0: la variable y est linéairement indépendante des variables x_j

H_1: la variable y est linéairement liée à au moins une des variables x_j

L'expression la plus commode de la variable auxiliaire F est basée sur le coefficient de détermination:

$$F_{RM_C} = \frac{R^2(n-p)}{(1-R^2)(p-1)}$$

En ce qui concerne les conditions d'application du test, la régression multiple est soumise aux mêmes contraintes que la régression linéaire simple:

- distribution normale de la variable dépendante
- équivariance

- indépendance des résidus

- linéarité des relations entre la variable dépendante *y* et chacune des variables explicatives *x*.

La liaison entre la variable à expliquer *y* et *l'ensemble* des variables explicatives peut se mesurer par un coefficient de "**corrélation multiple**" défini comme la racine carrée du coefficient de détermination R^2. Par définition (puisqu'on prend la racine carrée d'un nombre réel), la corrélation multiple obtenue ne peut pas être négative. De ce fait, la notion de corrélation multiple a une interprétation douteuse et doit être manipulée avec beaucoup de prudence: par exemple, même dans un cas où une variable dépendante *y* serait influencée négativement par toutes les variables explicatives $x_{p\text{-}1}$, le coefficient de corrélation multiple serait positif.

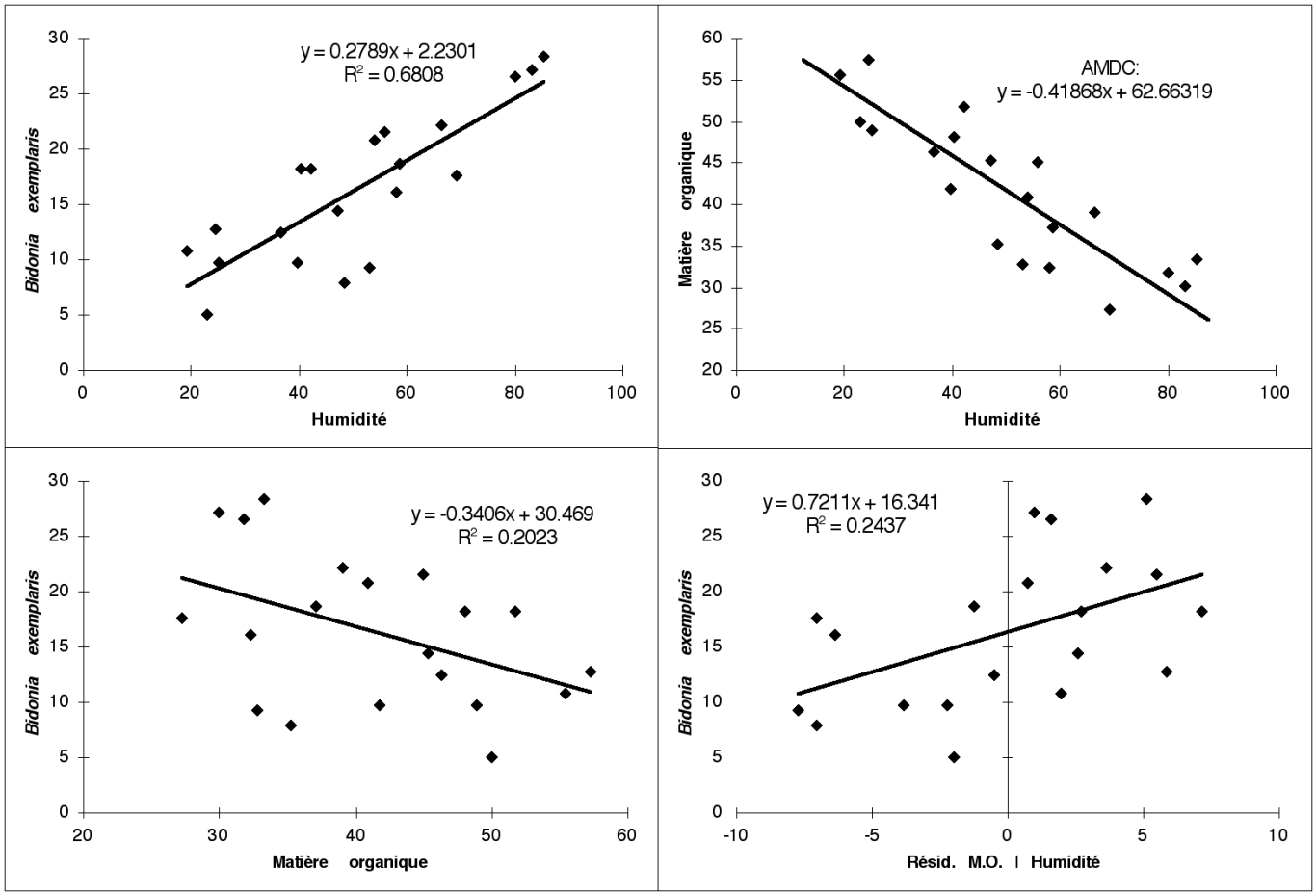
Point important, les coefficients de régression obtenus par régression multiple sont en fait des coefficients de **régression partielle**, en ce sens que chacun mesure l'effet de la variable explicative concernée sur la variable dépendante **lorsque la ou les autres variables explicatives sont tenues constantes**.

Cette propriété est très intéressante. En effet, si on désire connaître l'influence d'un groupe de facteurs sur une variable-cible (=dépendante) donnée, en contrôlant l'effet d'un autre groupe (p. ex. l'effet de la teneur en matière organique du sol sur l'abondance de *Bidonia exemplaris*, en ôtant l'effet de l'humidité), on peut calculer une régression intégrant toutes les variables explicatives, et examiner les coefficients de régression du groupe de variables voulu, en sachant que ces coefficients expliquent la variance de la variable dépendante en contrôlant pour l'effet de l'autre groupe.

Cette démarche n'est pas triviale. En effet, les influences combinées des diverses variables en jeu aboutissent quelquefois à des **effets apparents contraires à ceux qui sont en jeu**.

Dans notre exemple, en régression simple, *Bidonia* a l'air de réagir négativement à l'augmentation de la teneur en matière organique (voir figure ci-dessous). Par contre, si l'on tient constant l'effet de l'humidité,

le coefficient de régression partielle de la matière organique est positif (0.7211). Cela tient à ce que dans l'échantillonnage, les prélèvements les plus humides sont aussi ceux où le taux de matière organique est le plus faible. Or, *Bidonia* réagit fortement (et positivement) à l'humidité. Il réagit aussi positivement à une augmentation de la matière organique, mais pas de façon aussi forte que vis-à-vis de l'humidité.



En haut à gauche: régression linéaire simple de *B. exemplaris* sur l'humidité. En bas à gauche: régression linéaire simple de *B. exemplaris* sur le taux de matière organique (réaction apparemment négative). En haut à droite: relation entre humidité et matière organique. En bas à droite: régression partielle de *B. exemplaris* sur la matière organique, en maintenant l'humidité constante (la variable explicative est le résidu d'une régression de la matière organique sur l'humidité).

On voit donc qu'il est indispensable, lorsqu'on dispose de plusieurs variables explicatives, de les intégrer **ensemble** dans une analyse plutôt que d'avoir recours à une série de régressions simples. Non seulement on peut alors mesurer leur effet combiné sur la variable dépendante, mais on peut aussi tester globalement cet effet (à l'aide de la statistique *F* présentée plus haut).

Régression sur variables centrées-réduites

Une pratique courante en régression consiste à **interpréter les coefficients de régression centrés-réduits**, c'est-à-dire ceux qu'on obtient en centrant-réduisant toutes les variables (y compris la variable dépendante). En exprimant toutes les variables en unités d'écart-type, on rend les coefficients de régression insensibles à l'étendue de variation des variables explicatives, leur permettant ainsi d'être interprétés directement en termes de "poids" relatif des variables explicatives. Notez aussi que la plupart des logiciels courants fournissent de toute manière les "coefficients de régression centrés-réduits" (*standardized regression coefficients*) en plus des coefficients calculés pour les variables brutes.

On peut remarquer aussi que, si on fait le calcul à l'aide de la méthode montrée par Scherrer (p. 696 et suivantes), on obtient de toute manière d'abord les coefficients centrés-réduits (sans avoir à centrer-réduire les variables pour faire le calcul!).

Le centrage-réduction n'affecte pas la corrélation entre les variables, ni les coefficients de détermination (R^2) des régressions simples et multiples.

L'exemple de *Bidonia* exposé plus haut devient ainsi:

Abondance $Bidonia_{cr} = 1.6397 \times \text{Hum.}_{cr} + 0.9524 \times \text{M.O.}_{cr}$

L'ordonnée à l'origine vaut 0 puisque toutes les variables sont centrées.

Dans ce contexte, mentionnons que le coefficient de détermination peut aussi s'exprimer (équation 18-46 p.699 de Scherrer):

$$R^2 = \sum_{j=1}^{p-1} a'_j r_{jp}$$

Les a'_j sont les coefficients de régression des variables centrées-réduites. Donc, chaque élément $a'_j r_{jp}$ représente la **contribution** de la variable x_j à l'explication de la variance de y. Dans notre exemple, la contribution de l'humidité et celle de la matière organique s'élèvent à

$1.6397 \times 0.8251 = 1.3529$ et $0.9524 \times -0.4498 = -0.4284$

$R^2 = 1.3529 - 0.4284 = 0.9245$

Voir aussi l'exemple 18.17 de Scherrer (p. 700).

Remarque: en régression linéaire **simple** (uniquement!), lorsque les deux variables sont centrées-réduites, le coefficient de régression a (=la pente) est égal à la corrélation r entre les deux variables x et y.

R^2 ajusté

Une des propriétés de la régression multiple est que l'ajout de chaque variable explicative au modèle permet d'expliquer plus de variation, et cela même si la nouvelle variable explicative est complètement aléatoire. Cela vient du fait que si l'on compare deux variables aléatoires, les fluctuations aléatoires de chacune d'entre elles produisent de très légères corrélations: y et chacune des x_j ne sont pas strictement indépendantes (orthogonales) même s'il n'y a aucune relation entre elles. Par conséquent, le R^2 calculé comme ci-dessus comprend une composante déterministe, et une composante aléatoire d'autant plus élevée que le nombre de variables explicatives est élevé dans le modèle de régression.

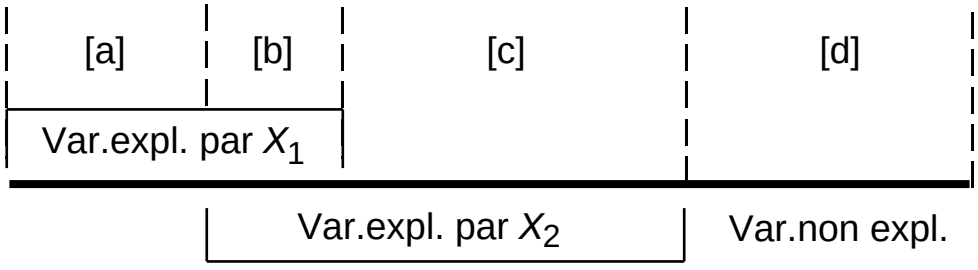
Pour contrer cet effet, et donc éviter de surestimer le R^2, certains auteurs ont proposé un **R^2 ajusté**, qui tient compte du nombre de variables explicatives du modèle de régression:

$$R^2_{aj} = 1 - \frac{(n-1)}{(n-m-1)}(1-R^2)$$

où n = nombre d'observations et m = nombre de variables explicatives
Voir p.ex. Zar (1999) p. 423.

Partitionnement de la variation (Legendre & Legendre (1998) p. 531)

Dans la grande majorité des cas, les variables explicatives intégrées à une régression multiple ne sont pas linéairement indépendantes entre elles (orthogonales). Le R^2 total de la régression multiple n'est donc pas la somme des r^2 d'une série de régressions simples impliquant tour à tour toutes les variables explicatives, mais une valeur généralement inférieure à cette somme:



Dans cet exemple, la barre grasse représente toute la variation de la variable dépendante. Comme les variables x_1 et x_2 ne sont pas linéairement indépendantes, une partie de leur pouvoir explicatif va expliquer la même part de variation de y. Cette fraction commune est appelée fraction [b]. La contribution unique de la variable x_1 est la fraction [a], et la contribution unique de la variable x_2 est la fraction

[c]. La fraction [d] constitue la partie non expliquée, soit le résidu de la régression multiple.

On peut obtenir les valeurs de chacune de ces fractions de la manière suivante:

(1) Régression linéaire simple de y sur x_1: le r^2 vaut [a]+[b].

(2) Régression linéaire simple de y sur x_2: le r^2 vaut [b]+[c].

(3) Régression linéaire multiple de y sur x_1 et x_2: le R^2 vaut [a]+[b]+[c].

(4) La valeur de [a] peut donc être obtenue en soustrayant le résultat de l'opération (2) de celui de (3).

(5) La valeur de [c] peut donc être obtenue en soustrayant le résultat de (1) de celui de (3).

(6) La valeur de [b] s'obtient de diverses manières, p. ex. (1) – (4), ou (2) – (5).

(7) La fraction [d] (variation non expliquée) s'obtient en faisant 1– ([a]+[b]+[c]).

Remarque: on ne peut pas ajuster de modèle de régression sur la fraction [b], dont la valeur ne peut être obtenue que par soustraction. Elle peut même être négative s'il y a antagonisme entre les effets de certaines variables explicatives (c'est le cas dans notre exemple de *Bidonia* montré plus haut). C'est pourquoi on parle ici de variation et non de variance au sens strict.

Voir aussi la boîte 4.1 de la future nouvelle édition du manuel de Legendre et Legendre, fournie en pdf sur la page web du cours.

Autre remarque: pour permettre la **comparaison de variables explicatives qui ne sont pas toutes mesurées dans les mêmes unités**, ou qui ont des intervalles de variation très différents, on a souvent recours au **centrage-réduction des variables explicatives**. Dans ce cas-là, il n'est pas nécessaire de centrer-réduire la variable dépendante.

Le problème de la multicolinéarité

Lorsque plusieurs, voire toutes les variables explicatives sont fortement corrélées entre elles (r = 0.8 et plus), les estimations des coefficients de régression deviennent instables (fluctuent beaucoup d'un échantillon à l'autre). Leur interprétation devient donc dangereuse. Il y a plusieurs solutions possibles:

- créer une nouvelle variable synthétique (combinant les variables interreliées) et l'utiliser à la place des autres;

- choisir une seule des variables très interreliées et s'en servir comme indicatrice des autres;

- utiliser d'autres méthodes (régression à partir des composantes principales, régression pseudo-orthogonale);

Remarque: si le seul but de la régression multiple est la prédiction (maximisation du R^2), la multicolinéarité ne dérange pas.

La corrélation partielle

Au contraire du coefficient de "corrélation multiple" évoqué plus haut, on peut définir un coefficient de corrélation partielle qui a le même sens que le coefficient de corrélation r de Pearson ordinaire.

Un coefficient de corrélation partielle mesure la liaison entre deux variables lorsque l'influence d'une troisième (ou de plusieurs autres) est gardée constante *sur les deux variables comparées*. Cela rappelle donc l'interprétation des coefficients de régression partielle montrés plus haut. On rappellera cependant qu'une corrélation ne mesure que la liaison entre deux variables, sans se préoccuper de modèles fonctionnels ou de capacité de prédiction ou de prévision.

Le calcul d'une corrélation partielle fait intervenir les corrélations ordinaires entre les paires de variables considérées. L'exemple ci-dessous vaut dans le cas où on a deux variables explicatives x_1 et x_2 (équ. 18-50 de Scherrer, p. 704). La formule décrit le calcul de la corrélation partielle de y et x_1 en tenant x_2 constant:

$$r_{y,x_1|x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1-r^2_{yx_2})(1-r^2_{x_1x_2})}}$$

Ce coefficient se teste à l'aide d'un *F* obéissant sous H_0 à une loi de Fisher-Snedecor à 1 et *n-p* degrés de liberté (rappel: *p* désigne ici tous les paramètres de l'équation de régression multiple: coefficients de régression **plus** ordonnée à l'origine). La construction du test et les règles de décision figurent aux pages 705 et 706 de Scherrer.

Le carré du coefficient de corrélation partielle $r^2_{y,x_1|x_2,x_3...}$ mesure la proportion de la variation de *y* expliquée par x_1 par rapport à la variation non expliquée par x_2, x_3, etc. Cela correspond donc au rapport des fractions de variation [a]/([a]+[d]) dans le cadre du partitionnement expliqué plus haut. Les composantes de variation [b] et [c], liées à l'autre ou aux autres variables explicatives, sont donc absentes du calcul.

L'exemple de *Bidonia* et de sa relation avec l'humidité et la teneur en matière organique du sol est assez parlant:

Correlation Matrix

	B. exemplaris	Humidité	M.O.
B. exemplaris	1.000	.825	**-.450**
Humidité		1.000	-.855
M.O.			1.000

Partial Correlation Matrix

	B. exemplaris	Humidité	M.O.
B. exemplaris	1.000	.951	**.874**
Humidité		1.000	-.959
M.O.			1.000

Un chercheur qui se contenterait d'une matrice de corrélations simples (à gauche) penserait que la relation entre *Bidonia* et la teneur en M.O. est négative. Par contre, s'il prenait la précaution de calculer une matrice de corrélations partielles, il verrait que cette illusion est due à l'effet masquant de l'humidité dans l'échantillon. La corrélation partielle forte et positive entre *Bidonia* et la M.O. mesure la relation entre *Bidonia* et la partie de la variation de la matière organique qui n'est pas expliquée par l'humidité.

Régression pas à pas

On rencontre parfois des situations dans lesquelles on dispose de *trop* de variables explicatives, soit parce que le plan de recherche était trop vague au départ (on a mesuré beaucoup de variables "au cas où elles auraient un effet"), soit parce que le nombre d'observations (et donc de degrés de liberté) est trop faible par rapport au nombre de variables explicatives intéressantes.

Une technique est parfois employée pour "faire le ménage" et sélectionner un nombre réduit de variables qui explique pourtant une quantité raisonnable de variation. Cette régression, dite "pas à pas" (*stepwise regression* en anglais) est expliquée par Scherrer (paragr. 18.3.6, p. 708). Il en existe plusieurs variantes.

1. Méthode rétrograde *(backward selection)*

Cette méthode consiste à partir du modèle de régression complet (intégrant toutes les variables explicatives), et à en retirer une par une les variables dont le *F* partiel est non significatif. Inconvénient: une fois qu'une variable a été retirée, elle ne peut plus être réintroduite dans le modèle, même si, à la suite du retrait d'autres variables, elle redevenait significative. Cette approche est néanmoins assez libérale (elle a tendance à garder un nombre plus élevé de variables dans le modèle final que les autres approches ci-dessous).

2. Méthode progressive *(forward selection)*

Approche inverse de la précédente: elle sélectionne d'abord la variable explicative la plus corrélée à la variable dépendante. Ensuite, elle sélectionne, parmi celles qui restent, la variable explicative dont la corrélation partielle est la plus élevée (en gardant constantes la ou les variables déjà retenues). Et ainsi de suite tant qu'il reste des variables candidates dont le coefficient de corrélation partiel est significatif. Inconvénient: lorsqu'une variable est entrée dans le modèle, aucune procédure ne contrôle si sa corrélation partielle reste significative après l'ajout d'une ou de plusieurs autres variables. Cette technique est en

général plus conservatrice que la précédente, ayant tendance à sélectionner un modèle plus restreint (moins de variables explicatives) que la sélection rétrograde.

3. Sélection pas à pas proprement dite *(stepwise regression)*

Cette procédure, la plus complète, consiste à faire entrer les variables l'une après l'autre dans le modèle (selon leur corrélation partielle) par sélection progressive et, à chaque étape, à vérifier si les corrélations partielles de l'ensemble des variables déjà introduites sont encore significatives (une variable qui ne le serait plus serait rejetée). Cette approche tente donc de neutraliser les inconvénients des deux précédentes en les appliquant alternativement au modèle en construction.

Quelle que soit sa variante, la régression pas à pas présente des **dangers**. En particulier, lorsqu'on a fait entrer une variable donnée dans le modèle, elle conditionne la nature de la variation qui reste à expliquer. De ce fait, rien ne garantit qu'on a choisi au bout du compte la combinaison de variables qui explique le plus de variation. De plus, le modèle devient hautement instable en présence de (multi-)colinéarité entre les variables explicatives, ce qui veut dire que les paramètres déterminés par la méthode (les poids attribués aux variables retenues), et même la liste des variables retenues elle-même, peuvent varier fortement si on change (même très peu) les données. L'utilisation la plus recommandée de la régression pas à pas se fait dans le cadre de la régression polynomiale.

Annexe: calcul des paramètres d'une régression multiple

Principe:

On peut calculer les coefficients de régression et l'ordonnée à l'origine d'une régression multiple en connaissant:

• les coefficients de corrélation linéaire simple de toutes les paires de variables entre elles (y compris la variable dépendante): r_{12}, r_{13}...r_{1p}, r_{23} ..etc.;

• les écarts-types de toutes les variables: s_1, s_2, s_3...s_p;

• les moyennes de toutes les variables.

Remarque: dans cette notation, la p-ième variable est la variable dépendante.

Étapes de calcul (principe):

1. On calcule d'abord les coefficients de **régression centrés-réduits** a_1', a_2', ... a'_{p-1} en résolvant un système de p–1 équations normales à p – 1 inconnues (p–1 = nombre de variables explicatives).

2. On trouve les coefficients de régression pour les variables originales a_1, a_2, ... a_{p-1} en multipliant chaque coefficient centré-réduit par l'écart-type de la variable dépendante, et en divisant le résultat par l'écart-type de la variable explicative considérée.

3. On trouve l'ordonnée à l'origine en posant la moyenne de la variable dépendante, et en lui soustrayant chaque coefficient obtenu au point 2, multiplié par la moyenne de la variable explicative correspondante.

Formules:

Cette technique est exposée par Scherrer (1984), p. 697 et suivantes, avec un exemple numérique.
Les formules ci-dessous sont données pour 3 variables explicatives.

1. Equations normales :

$$r_{1p} = a'_1 + r_{12}a'_2 + r_{13}a'_3$$

$$r_{2p} = r_{21}a'_1 + a'_2 + r_{23}a'_3$$

$$r_{3p} = r_{31}a'_1 + r_{32}a'_2 + a'_3$$

Ce système se résoud par substitutions successives.
1e étape:

$$a'_1 = r_{1p} - r_{12}a'_2 - r_{13}a'_3$$

est placé dans les équations 2 et 3. On isole ensuite a'_2 ou a'_3 dans l'une des équations. Dès lors, on peut trouver l'une des valeurs, et, en remontant la filière, on trouve les deux autres.

2. Coefficients pour variables brutes :

$$a_1 = a'_1 \frac{s_y}{s_{x1}} \qquad a_2 = a'_2 \frac{s_y}{s_{x2}} \qquad a_3 = a'_2 \frac{s_y}{s_{x3}}$$

3. Ordonnée à l'origine :

$$b = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 - a_3\bar{x}_3$$

Remarque: il existe une méthode différente pour calculer les coefficients de régression multiple, basée sur le calcul matriciel. On trouvera cette technique chez Legendre et Legendre (1998) pp. 79 et 517, et dans Zar (1999) p. 413 et suivantes.

On peut aussi trouver la contribution de chacune des variables explicatives à l'explication de la variance de la variable dépendante.

Par exemple, pour la variable explicative x_1:

$$\text{Contribution} = a_1' r_{yx_1}$$

Attention: cette contribution n'est pas égale au R^2 partiel. Elle n'est pas non plus égale à la fraction [a] d'un partitionnement de variation si les variables explicatives sont (même très peu!) corrélées entre elles!

Le coefficient de détermination multiple R^2 de l'équation (= pourcentage de variance expliquée par l'ensemble des variables explicatives) peut s'obtenir en faisant la somme des termes ci-dessus:

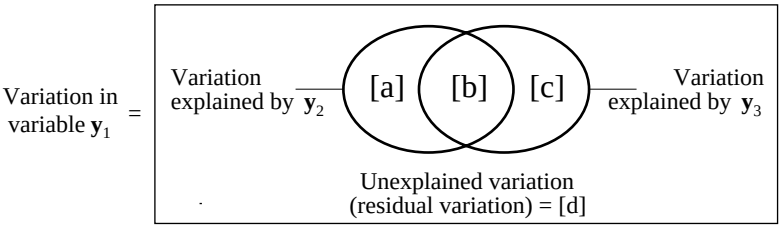
$$R^2 = \sum_{j=1}^{p-1} a_j' r_{jp}$$

Voir aussi le document "r2partiel.pdf", qui met en lumière, avec des exemples, la différence entre r^2 partiel, fraction [a] d'un partitionnement de variation et contribution d'une variable à l'explication de la variance en régression multiple.

Variation partitioning — Box 4.1

Anticipating upon Subsection 10.3.5 on partial linear regression, which should be referred to for details, the same graphic representation as in Fig. 10.10 is used here to illustrate, in a unified framework, the similarities and difference between partial correlation coefficient, coefficient of multiple determination, and *F*-statistics.

Consider three variables $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$. In the following graph, the rectangle represents the total sum of squares of variable $\mathbf{y}_1$:



The partial correlation of $\mathbf{y}_1$ with $\mathbf{y}_2$ while controlling for the effect of $\mathbf{y}_3$ is:

$$r_{12.3} = \sqrt{\frac{[a]}{[a+d]}} \qquad \text{with} \quad F = \frac{[a]/1}{[d]/(n-3)}$$

The partial correlation of $\mathbf{y}_1$ with $\mathbf{y}_3$ while controlling for the effect of $\mathbf{y}_2$ is:

$$r_{13.2} = \sqrt{\frac{[c]}{[c+d]}} \qquad \text{with} \quad F = \frac{[c]/1}{[d]/(n-3)}$$

The coefficient of partial correlation receives the same sign as the corresponding coefficient of partial regression. In the multiple regression of $\mathbf{y}_1$ on $\mathbf{y}_2$ and $\mathbf{y}_3$,

$\hat{y}_1 = b_0 + b_2 y_2 + b_3 y_3$ (this is an application of eq. 10.15),

the coefficient of multiple determination, which is the square of the coefficient of multiple correlation, is:

$$R^2_{1.23} = \frac{[a+b+c]}{[a+b+c+d]} \qquad \text{with} \quad F = \frac{[a+b+c]/2}{[d]/(n-3)}$$

The test of a partial regression coefficient, b_2 or b_3, is the same (i.e., it has the same *F*-statistic) as the test of the corresponding partial correlation coefficient, $r_{12.3}$ or $r_{13.2}$. The *F*-statistic is always the ratio of two *independent* portions of the variation of $\mathbf{y}_1$, each one divided by its degrees of freedom (given in eqs. 4.39 and 4.40).

Régression multiple:

calcul des paramètres par inversion matricielle

Tableau de données :

y	$\mathbf{x}_1$	$\mathbf{x}_2$	-----	$\mathbf{x}_m$
y_1	x_{11}	x_{12}	-----	x_{1m}
y_2	x_{21}	x_{23}	-----	x_{2m}
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
y_n	x_{n1}	x_{n4}	-----	x_{nm}

Notation matricielle de l'équation de régression : $\mathbf{y} = \mathbf{X}\,\mathbf{b}$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \dots \\ b_m \end{bmatrix}$$

L'ajout d'une colonne de "1" à la matrice **X** permettra d'estimer l'ordonnée à l'origine.

Raisonnement (*Numerical ecology* 1998, p. 79) :

$$\mathbf{y} = \mathbf{X}\,\mathbf{b}$$

$$\mathbf{X'y} = \mathbf{X'X}\,\mathbf{b} \quad \Rightarrow \quad [\mathbf{X'X}]^{-1}[\mathbf{X'y}] = [\mathbf{X'X}]^{-1}\,[\mathbf{X'X}]\,\mathbf{b}$$

$$\Rightarrow \quad [\mathbf{X'X}]^{-1}[\mathbf{X'y}] = \mathbf{b}$$

Calcul des valeurs estimée par l'équation de régression :

$$\hat{\mathbf{y}} = \mathbf{Xb}$$

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

$$\hat{\mathbf{y}} = \boxed{\mathbf{X}\,[\mathbf{X'X}]^{-1}\mathbf{X'}}\,\mathbf{y}$$

Projecteur

Au cours du test par permutation, on n'aura pas à recalculer le projecteur. On calculera simplement :

$$\hat{\mathbf{y}}_{\text{permuté}} = \boxed{\mathbf{X}\,[\mathbf{X'X}]^{-1}\mathbf{X'}}\,\mathbf{y}_{\text{permuté}}$$

$$\hat{\mathbf{y}}_{\text{permuté}} = \textbf{Projecteur}\;\mathbf{y}_{\text{permuté}}$$

Régression polynomiale

2001-2005

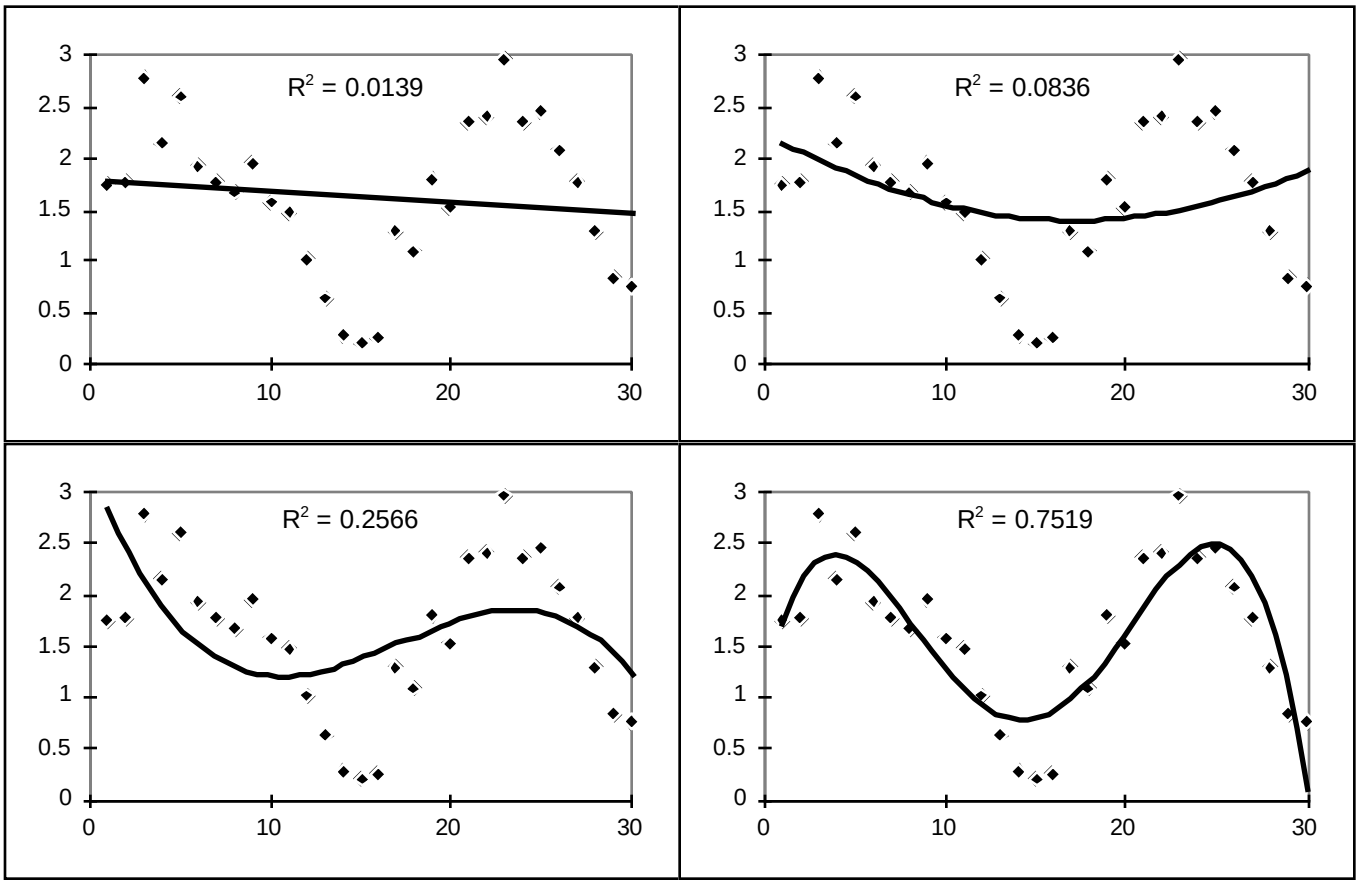
Daniel Borcard, Dép. de sciences biologiques, Université de Montréal

Référence:Legendre et Legendre (1998) p. 526

Une variante de la régression multiple peut quelquefois être appliquée pour ajuster une variable explicative *x* (ou plusieurs variables explicatives x_j) à une variable dépendante *y* de manière non-linéaire. Cette méthode consiste à ajouter à la variable explicative *x* de nouvelles variables construites en mettant *x* au carré, au cube, etc. L'équation devient donc, pour un polynôme du *k*-ième ordre:

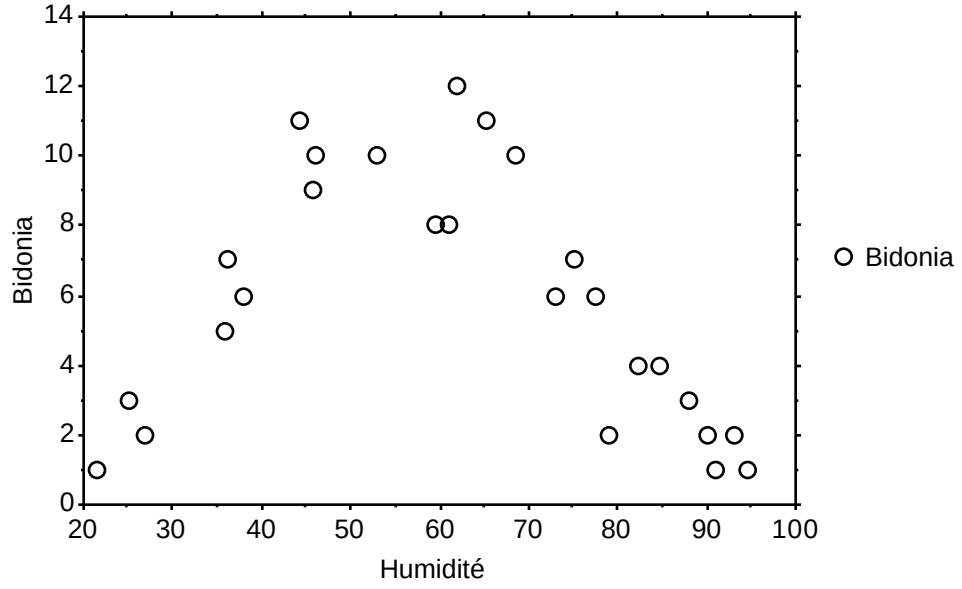
$$\hat{y} = a_1 x + a_2 x^2 + a_3 x^3 + ... + a_k x^k + b$$

Une façon simple de se représenter l'effet de l'ajout d'un ordre est la suivante: **chaque nouvel ordre permet d'ajouter un pli à la courbe**. Une équation du premier ordre est celle d'une droite, du deuxième ordre une parabole; le troisième ordre donne un S couché, etc.:



Cette figure montre comment l'ajout d'éléments de souplesse permet de mieux ajuster le modèle de régression aux données. Cependant, comme toujours, il faut savoir trouver un compromis entre un modèle trop rudimentaire et mal ajusté et un modèle ajustant bien les données, mais au prix d'un nombre excessif de termes (et donc de paramètres). Legendre et Legendre (1998, p. 526) recommandent de partir d'un modèle d'un ordre volontairement trop élevé (p. ex. du 6^{e} ordre). Une première étape consiste alors à retirer un par un et dans l'ordre décroissant les termes des ordres supérieurs (en général non significatifs) jusqu'à ce qu'on rencontre un terme significatif. Les termes restants pourront ensuite si nécessaire être "élagués" à l'aide d'une régression pas à pas.

La régression à l'aide d'un polynôme du deuxième ordre (parabole) est d'un intérêt particulier pour les biologistes. Prenons pour exemple la distribution de *Bidonia exemplaris* le long d'un gradient d'humidité:



On constate que *Bidonia* semble préférer les conditions d'humidité moyennes. L'ajustement d'une droite n'aurait à l'évidence aucun sens ici: la droite serait quasiment horizontale et ne rendrait aucunement compte du phénomène biologique qu'on cherche à modéliser. Incidemment, *la corrélation linéaire serait nulle elle aussi*, ce qui

montre qu'il ne suffit pas de calculer une corrélation linéaire entre n'importe quoi et n'importe quoi pour avoir tout dit!

Pour modéliser une telle situation, nous allons faire usage d'un artifice. Apparemment, il suffirait d'ajouter au modèle un terme en x^2 pour obtenir une courbe adéquate: on fabriquerait une équation du *deuxième degré*, ou, en d'autres termes, on ajusterait une *parabole* au nuage de points:

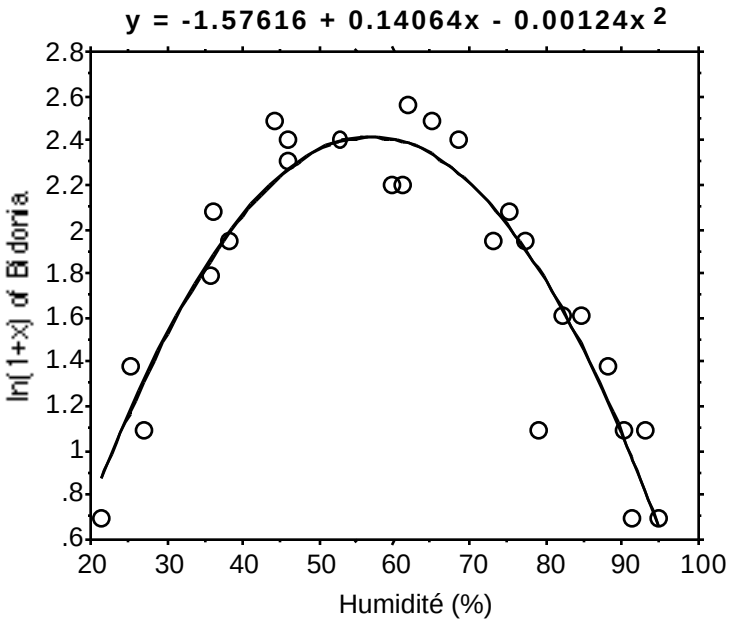
$$\hat{y} = a_1 x + a_2 x^2 + b$$

Cette opération fournirait effectivement un bon ajustement, mais les paramètres de l'équation seraient difficilement interprétables biologiquement. De plus, une parabole prédirait des abondances négatives aux extrémités du spectre de l'espèce.

Il existe heureusement une manière très élégante de s'en sortir. Avant d'ajuster la parabole, on transforme les abondances d'espèces en logarithmes naturels [$y' = \ln(y+1)$]. En effet, *ajuster une parabole à des données d'abondances d'espèces logarithmiques revient à ajuster une courbe de Gauss sur les données brutes!* Et la courbe de Gauss (difficile à ajuster directement), en plus d'être biologiquement réaliste, permet facilement le calcul de tous les paramètres recherchés par le biologiste: **optimum** et **tolérance** de l'espèce.

Remarque: il faut être très attentif à prendre suffisamment de décimales pour les calculs qui suivent, car les paramètres de l'équation de la parabole (surtout a_2) sont souvent très petits. Au besoin, refaire la régression avec des logs multipliés par 10 ou 100, puis diviser les coefficients obtenus.

Pour *Bidonia*, l'ajustement d'une équation du deuxième ordre (parabole) aux données transformées en logs naturels donne ceci:



A l'évidence, ce modèle parabolique (dont l'équation figure au-dessus du graphe) donne une image assez fidèle de la situation. Le pourcentage de variance expliquée par la parabole (le R^2) est de 0.875. On peut lire graphiquement l'optimum *u* de l'espèce: environ 57% d'humidité. Mais on peut aussi le calculer à partir des paramètres a_1 et a_2 de la régression, par la formule suivante:

$$u = \frac{-a_1}{2a_2}$$

La tolérance *t*, quant à elle, définie comme *une unité d'écart-type*, s'obtient en faisant:

$$t = \frac{1}{\sqrt{-2a_2}}$$

Finalement, on peut encore calculer la valeur du sommet *c* de la courbe (en abondances brutes), en faisant:

$$c = e^{\left(b + a_1 u + a_2 u^2\right)}$$

Voivi les valeurs obtenues dans notre exemple:

$$u = \frac{-0.14064}{2 \times 0.00124} = 56.7$$

$$t = \frac{1}{\sqrt{-2 \times 0.00124}} = 20.1$$

$$c = e^{[1.57616 + (0.14064 \times 56.7) + (0.00124 \times 56.7^2)]} = 11.2$$

Bidonia exemplaris a donc son optimum à 56.7% d'humidité, elle se tient de préférence entre 36.6 et 76.8% d'humidité, et sa densité à l'optimum est de 11 (individus par cm^2, si c'est dans cette unité que les données brutes ont été utilisées).

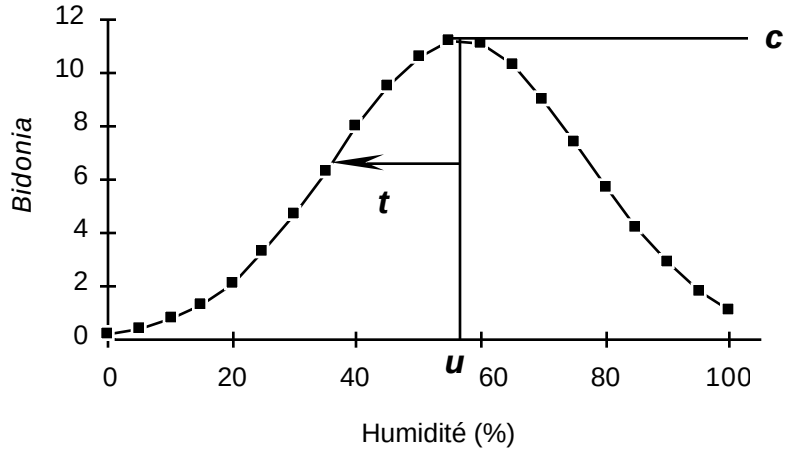
Avec ces résultats, on peut écrire l'équation de la courbe de Gauss ajustant les données brutes:

$$z = c \times \exp\left[\frac{-0.5(x-u)^2}{t^2}\right]$$

z étant la valeur d'abondance (brute) de *Bidonia* et *x* la valeur d'humidité. Donc, dans notre exemple:

$$z = 11.2 \times \exp\left[\frac{-0.5(x-56.7)^2}{20.1^2}\right]$$

Le graphe de la courbe de Gauss aurait l'allure suivante (avec l'illustration des divers paramètres que nous avons calculés):



Faut-il insister sur l'élégance de cette démarche? Apparemment oui, car il semble qu'elle soit méconnue... Les équations présentées ici sont tirées du manuel de Jongman, ter Braak et van Tongeren (1995).

Retour à la régression polynomiale au sens général

Remarque: **lorsque la variable explicative *x* est une coordonnée spatiale, on la centre** (= on lui soustrait sa moyenne) **avant de construire le polynôme** et de faire la régression, pour éviter que les termes successifs soient trop corrélés entre eux. Idem pour l'analyse des surfaces théoriques exposée ci-dessous.

Autre remarque: il est tout à fait permis de combiner régression multiple et polynomiale. Par exemple, si on étudiait la relation de *Bidonia exemplaris* avec l'humidité et le taux de matière organique MO), et que des diagrammes de dispersion *Bidonia* × humidité et *Bidonia* × MO nous montrent une relation linéaire de l'animal avec l'humidité, mais unimodale avec la MO, on pourrait modéliser le tout avec une équation du type

Abond. *Bidonia* = a_1 × humidité + a_2 × MO + a_3 × MO^2 + b

Le calcul de la régression polynomiale est exactement le même que celui de la régression multiple "ordinaire".

Analyse des surfaces polynomiales théoriques

La régression polynomiale forme aussi la base de la méthode la plus simple d'analyse spatiale: l'analyse des surfaces théoriques (*trend surface analysis* en anglais). Cette technique consiste à modéliser la distribution spatiale d'une variable dépendante *z* à l'aide d'un polynôme des coordonnées *x* et *y* des observations. L'idée est d'estimer la valeur d'une variable sur la base de sa localisation.

La technique est la même que celle de la régression polynomiale ci-dessus, à ceci près que l'ajout d'un ordre doit se faire pour les deux dimensions spatiales représentées par les coordonnées *x* et *y*:

Ordre 1: $\hat{z} = a_1 x + a_2 y + b$

Ordre 2: $\hat{z} = a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2 + b$

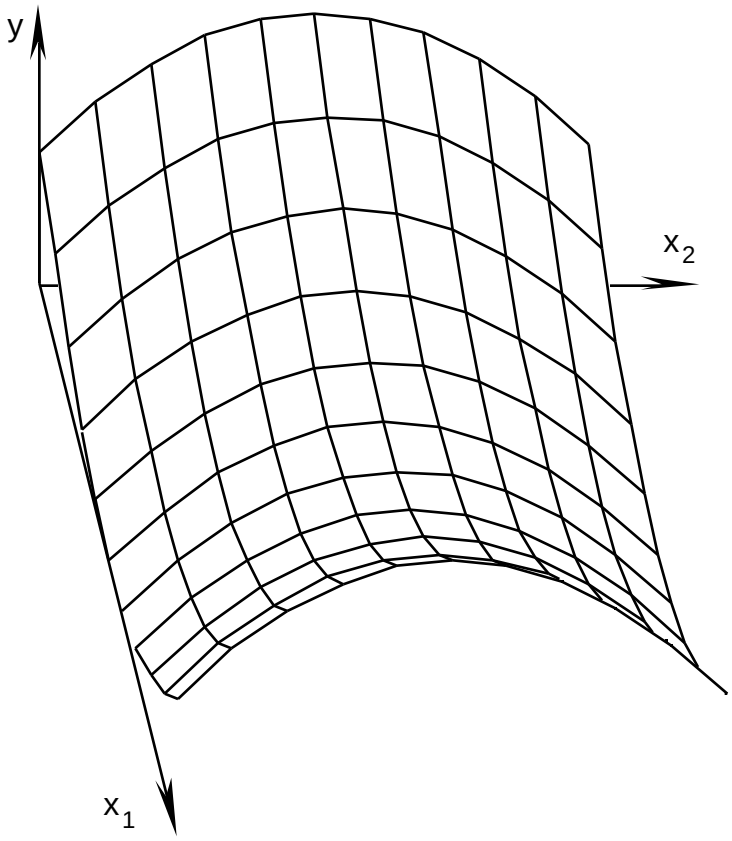
Ordre 3:

$$\hat{z} = a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2 + a_6 x^3 + a_7 x^2 y + a_8 xy^2 + a_9 y^3 + b$$

On peut construire des polynomes d'ordre supérieur, mais le nombre de variables (et donc de paramètres) devient rapidement très élevé, ce qui interdit l'usage de telles équations sur de petits échantillons.

Comme dans le cas unidimensionnel, il est recommandé de centrer les variables *x* et *y* avant de construire le polynôme, afin de réduire la colinéarité des termes de l'équation.

Voici un exemple de surface de surface théorique du deuxième ordre (une parabole est ajustée sur chacune des deux variables explicatives):


y
x2
x1

Daniel Borcard 2001-2005
Département de sciences biologiques
Université de Montréal

Régression logistique

Jongman, ter Braak & Van Tongeren (1995): Chap. 3 *in*: Data analysis in community and landscape ecology. Cambridge University Press, Cambridge.

La régression linéaire simple ordinaire (MCO) a pour but de modéliser la relation entre une variable dépendante **quantitative** et une variable explicative quantitative.

Lorsque la variable à expliquer est **binaire** (oui-non, présence[1] - absence[0]), il faut avoir recours à la régression logistique. Nous prendrons pour exemple ci-dessous une série de relevés de végétation dans lesquels on a consigné la présence ou l'absence de *Gentiana kochiana* L. (une fleur alpine) en fonction du pH du sol.

Dans notre exemple, la régression logistique permet de calculer la **probabilité** de présence de l'espèce pour une valeur de pH donnée. Les probabilités prennent des valeurs comprises entre 0 et 1, raison pour laquelle une équation de régression **linéaire** simple (MCO) est inutilisable: une telle droite peut prédire des valeurs négatives, et aussi plus grandes que 1.

Cette difficulté pourrait être surmontée en prenant l'**exponentielle**

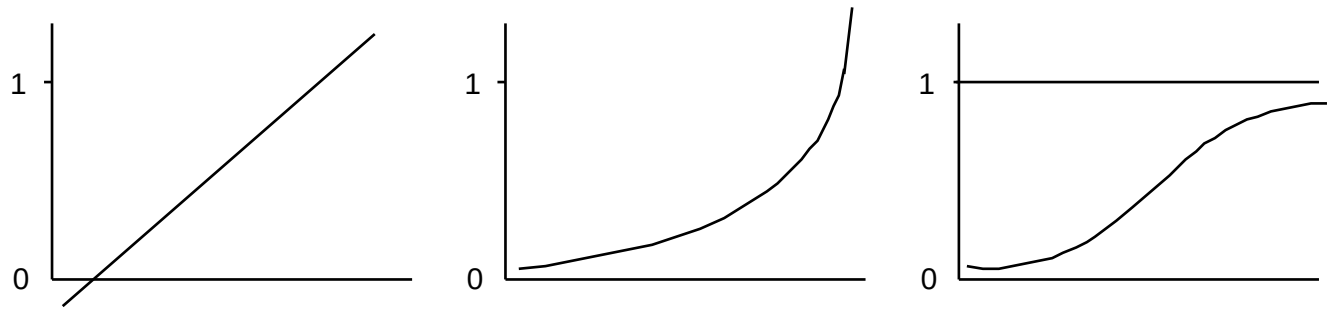
$$\hat{y} = e^{(ax+b)}$$

Cependant, cette fonction peut encore prendre des valeurs supérieures à 1, de sorte qu'il faut la borner par:

$$\hat{y} = p = \frac{e^{(ax+b)}}{1+e^{(ax+b)}}$$

Cette équation représente une **courbe sigmoïde**.

La figure ci-dessous montre ces trois courbes:



Linéaire Exponentielle Sigmoïde

Ces courbes ont toutes trois deux paramètres: *a* et *b*.

Ce qui précède constitue la partie systématique (déterministe) du modèle. L'erreur, elle, ne suit pas une distribution normale puisque la variable réponse est binaire. La distribution de l'erreur est la binomiale avec un total de 1. La variance de *y* est égale à $p(1–p)$.

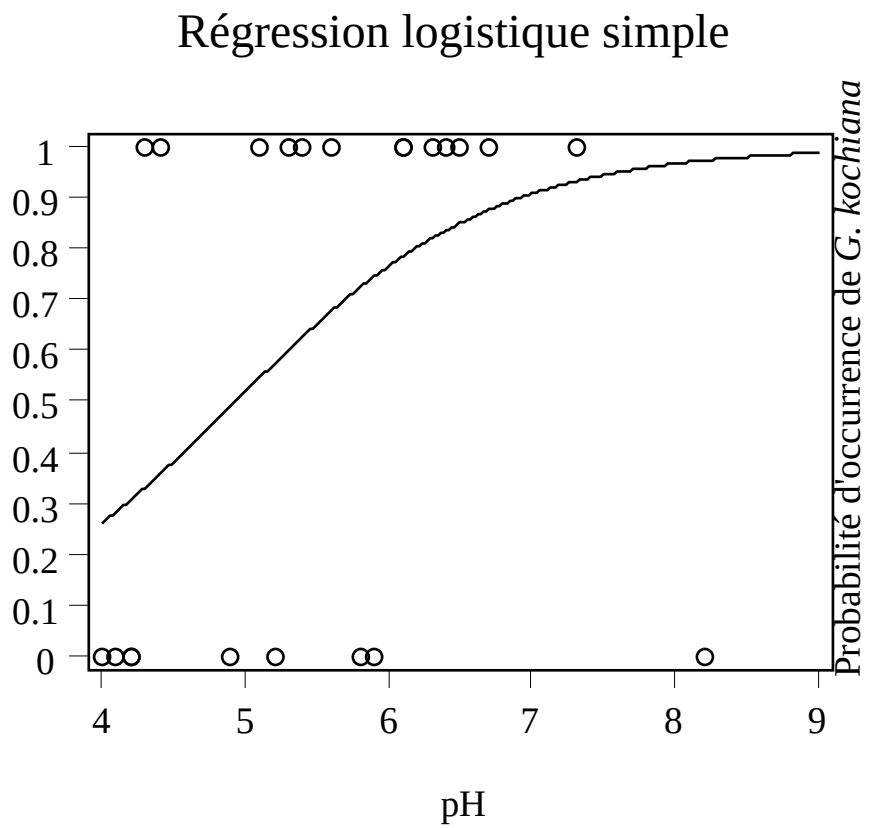
En anglais on utilise parfois le terme de "logit regression" à la place de "logistic regression". Cette expression se réfère à la transformation logit, qui est la transformation de *p*:

$$\ln [p / (1 – p)] = ax + b$$

$ax + b$ est l'élément explicatif linéaire (*linear predictor* en anglais), qui peut être adapté au besoin et prendre une forme plus complexe (voir plus loin).

Pour estimer les paramètres à partir des données, on ne peut utiliser la régression basée sur les moindres carrés ordinaires, puisque l'erreur n'est pas distribuée normalement et n'a pas une variance constante. Les paramètres sont ajustés selon le principe du maximum de vraisemblance, sur laquelles nous ne nous étendrons pas ici. De plus, ces paramètres doivent généralement être ajustés de manière itérative, à l'aide d'un programme auquel on fournit des valeurs initiales, et qui optimise ces valeurs de manière récurrente. Il existe des programmes très simples à utiliser qui permettent ce calcul, comme CurveFit (PC) (MacCurveFit sur Mac). StatView™ offre aussi un module de régression logistique.

Pour reprendre notre exemple, voici une illustration de données fictives et de la courbe ajustée à l'aide de la régression logistique simple:



Cette courbe donne, par exemple, une probabilité de 0.5 de rencontrer *Gentiana kochiana* si le pH du sol est de 5, et une probabilité de 0.9 si le pH vaut 7.

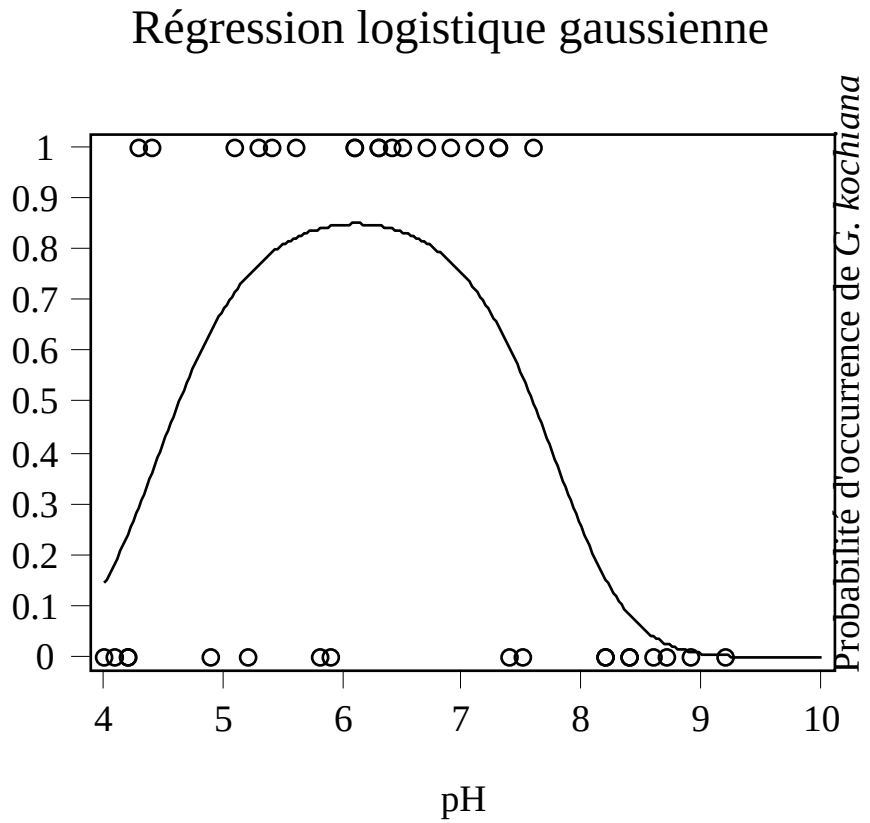
La régression logistique offre une grande souplesse grâce à la possibilité de modifier le contenu de l'élément explicatif linéaire. Un exemple très utile en écologie est de remplacer l'équation d'une droite par celle d'une **parabole**:

$$\hat{y} = p = \frac{e^{(a_1 x + a_2 x^2 + b)}}{1 + e^{(a_1 x + a_2 x^2 + b)}}$$

Le résultat de cette opération s'appelle **régression logistique gaussienne**, en ce sens que la courbe de probabilité résultante est une courbe de Gauss. Cette fonction est particulièrement utile

lorsqu'on cherche à connaître l'**optimum écologique** et la **tolérance** d'une espèce pour un facteur écologique donné, sur la base de données de présence-absence.

Reprenons l'exemple de *Gentiana kochiana*, en imaginant cette fois que le chercheur a échantillonné son espèce sur une plage plus étendue de pH. *G. kochiana* ne pousse pas sur des sols trop acides, mais pas non plus sur des sols trop alcalins. Son optimum se situe entre ces deux extrêmes, et peut être estimé sur le graphe suivant résultant de l'ajustement d'une régression logistique gaussienne sur les données:



Ce graphe montre que l'optimum écologique de *G. kochiana* pour le pH se situe vers 6.1, et que la probabilité de rencontrer cette espèce sur un sol d'un tel pH est de 0.85 (pour autant, bien sûr, que les autres conditions écologiques nécessaires à sa survie soit réunies). On peut ajouter que la probabilité de rencontrer *G. kochiana* dépasse 0.5 entre les pH de 4.6 et 7.7 . Ce serait une définition possible de la tolérance de l'espèce face au pH.

Régression logistique: un exemple

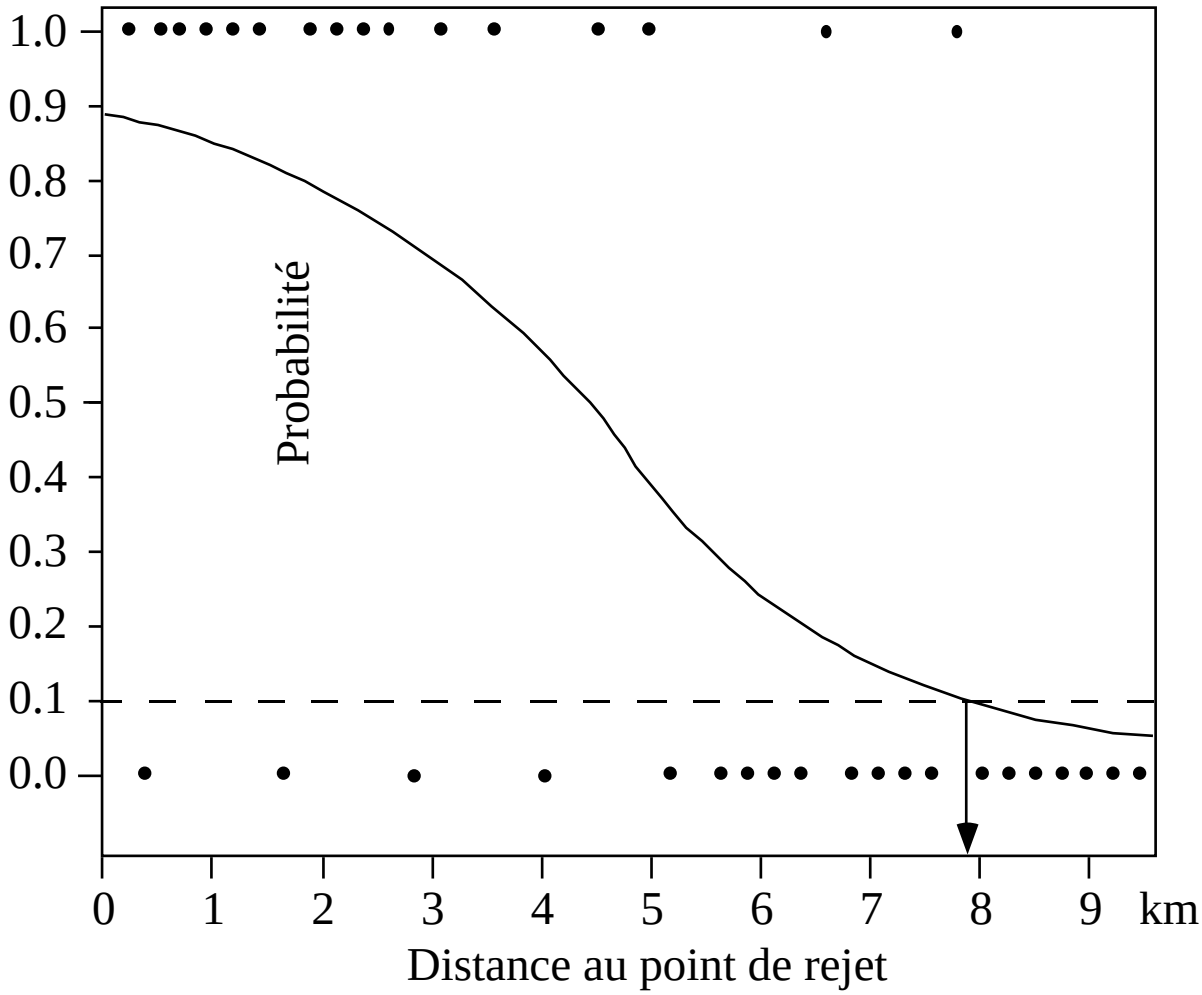
Site: un lac.

Source de contamination par la bactérie *Escherichia coli*: un point de rejet d'eaux usées.

Question: à partir de quelle distance du point de rejet la probabilité de trouver *E. coli* dans dix litres d'eau tombe-t-elle en-dessous de 10%?

Données:

Un ensemble de prélèvements dont on sait s'ils contiennent *E. coli* (oui - non) et dont on connaît la distance au point de rejet (*x*).



Solution:

On pose

$$p = \frac{e^{ax+b}}{1+e^{ax+b}}$$

L'ajustement aux données (par méthode itérative) a permis d'obtenir les paramètres suivants:

$a = -0.4888$ $b = 1.6752$

Donc:

$$p = \frac{e^{-0.4888x+1.6752}}{1+e^{-0.4888x+1.6752}}$$

On pose:

$$\ln[p/(1-p)] = -0.4888x+1.6752$$

Le seuil de probabilité désiré est de 10% (p = 0.1). Donc:

$$\ln[0.1/(1-0.1)] = -0.4888x+1.6752$$

On extrait x: $$x = -\frac{\ln\left(\frac{0.1}{1-0.1}\right) - 1.6752}{0.4888}$$

Soit: $$x = -\frac{-2.1972 - 1.6752}{0.4888} = 7.922$$

La probabilité de trouver *E. coli* dans l'eau est égale ou inférieure à 0.1 à partir de 7.922 km du point de rejet des eaux usées.

hypertension."[1] The authors specifically note, "The only patient with proteinuria . . . had a glomerular disease before her pregnancy." Thus, Benigni et al do not provide any information on eicosanoid change in preeclamptic women. In fact, only 3 of the 33 women in their "at risk for pregnancy-induced hypertension" group actually developed pregnancy-induced hypertension.

The Italian Study of Aspirin in Pregnancy[2] is one of several well-executed trials showing that aspirin does not prevent pregnancy-induced hypertension. Space limitations did not permit us to reference all the published trials.

Drs van der Weiden and Helmerhorst make the interesting point that PGI_2 metabolites increase as early as 40 days after embryo transfer in normal pregnancies, increasing the ratio of PGI_2 to TxA_2. In contrast, women who developed preeclampsia in our study showed significantly decreased levels of PGI_2 and had significantly decreased ratios of PGI_2 to TxA_2 virtually throughout the second and third trimesters of pregnancy. We agree with van der Weiden and Helmerhorst that PGI_2 production is an important factor in preeclampsia. We hope that future investigations will explore the therapeutic potential of correcting PGI_2 deficiency.

James L. Mills, MD, MS
Richard J. Levine, MD
National Institutes of Health
Bethesda, Md

Jason D. Morrow, MD
Vanderbilt University
Nashville, Tenn

1. Benigni A, Gregorini G, Frusca T, et al. Effect of low-dose aspirin on fetal and maternal generation of thromboxane by platelets in women at risk for pregnancy-induced hypertension. *N Engl J Med.* 1989;321:357-362.
2. Italian Study of Aspirin in Pregnancy. Low-dose aspirin in prevention and treatment of intrauterine growth retardation and pregnancy-induced hypertension. *Lancet.* 1993;341:396-400.

RESEARCH LETTER

Is Miss America an Undernourished Role Model?

To the Editor: The obsession with thinness in contemporary society has been cited as a contributing factor for the increase in eating disorders, particularly in young women.[1] Recent studies have found that as many as 50% to 75% of adolescent girls are dissatisfied with their weight and their body image.[2] Professions in which there are strong pressures to control body weight, such as athletics and dance, exhibit higher rates of eating disorders.[3] Beauty pageants are another tradition through which society defines its ideal of beauty, including body weight and shape.

Methods. We compiled data on weight and height of winners of the Miss America Pageant, from 1922 to 1999, obtained from the Miss America Archives.[4] The pageant was not held from 1927-1933, and data from a few other years are unavailable. We determined the body mass index (BMI), calculated as weight in kilograms divided by the square of height in meters, for each winner, and fit the BMI data to a linear regression model.

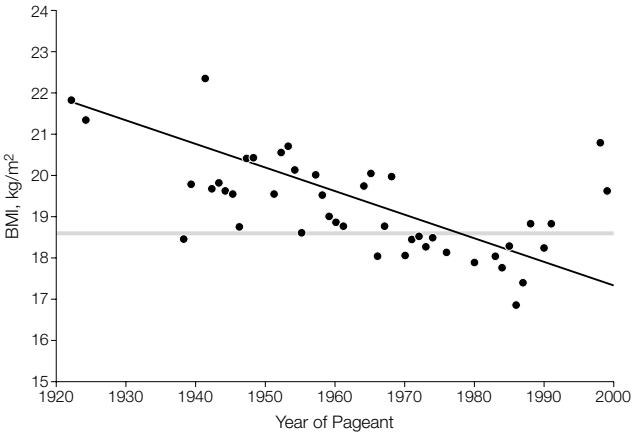
Results. We found a significant time-dependent decline in BMI ($P<.0001$), as shown in the FIGURE. In the 1920s, contestants had BMIs within the range that is now considered normal (between 20 and 25).[5] But the decline in BMI over the years has put an increasing number of winners in the range of undernutrition (defined by the World Health Organization as BMI<18.5),[5] with some having a BMI as low as 16.9.

Comment. Our finding cannot be explained simply by the secular upward trend in stature that occurred in the US population during this time. Pageant winners' height increased less than 2%, whereas body weight decreased by 12%. The actual influence of pageant competitions on young women's decisions about diet and lifestyle is not well documented, but it is likely to have a strong, if indirect, effect. Although considered politically incorrect by some, the 1999 Miss America pageant had an audience of more than 10 million, placing it 11th among prime-time programs according to the Nielsen Research service.[6] We chose the Miss America pageant for this analysis because of the completeness of its database. Its contestants, however, are chosen from local pageants, which are likely to promote a similar ideal of female undernutrition.

Sharon Rubinstein, MHS
Benjamin Caballero, MD, PhD
Johns Hopkins School of Public Health
Baltimore, Md

1. Moore DC. Body image and eating behavior in adolescents. *J Am Coll Nutr*. 1993;12:505-510.
2. Field AE, Cheung L, Wolf AM, Herzog DB, Gortmaker SL, Colditz GA. Exposure to mass media and weight concerns among girls. *Pediatrics*. 1999;103:E36.
3. Nattiv A, Agostini R, Drinkwater B, Yeager KK. The female athlete triad: the interrelatedness of disordered eating, amenorrhea, and osteoporosis. *Clin Sports Med*. 1994;13:405-418.
4. Miss America Organization Web site. Available at: http://www.pressplus.com/missam/pastwinners. Accessed March 2, 2000.
5. World Health Organization. Obesity: preventing and managing the global epidemic. Report of a WHO Consultation presented at: the World Health Organization; June 3-5, 1997; Geneva, Switzerland. Publication WHO/NUT/NCD/98.1.
6. Nielsen Media Research Rating Web site. Available at: http://www.ultimatetv.com/news/nielsen/archive/. Accessed February 16, 2000.

Trend in Body Mass Index (BMI) of Miss America Pageant Winners, 1922 to 1999



The horizontal line represents the World Health Organization's BMI cutoff point for undernutrition (18.5).[5]