Bio 2041

Droites d'estimation: la régression linéaire

Pierre Legendre & Daniel Borcard, Université de Montréal
Référence: Scherrer (2007), sections 18.1.1 à 18.1.5

1. Introduction

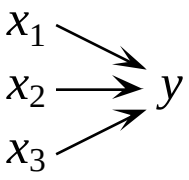
Objectif de l'étude

La méthode de la régression a pour but **de décrire** la relation entre une *variable aléatoire dépendante* (*y*) et un ensemble de *variables indépendantes* ou *prédictives x*. Le modèle obtenu permet **d'estimer** la valeur de *y* à l'aide des variables prédictives $x_1, x_2, \ldots, x_m$.

• Si les variables *x* sont contrôlées, on parle de régression de modèle I.

• Si les variables *x* sont aléatoires, on parle de régression de modèle II.

Lorsque l'estimation est fondée sur plusieurs variables prédictives, le problème en est un de *régression multiple.*

Modèle: 

Si le problème n'implique qu'une seule variable prédictive, utilisée simplement au premier degré (et non pas sous la forme x^2, x^3, etc.), il s'agit de *régression linéaire simple*. Modèle: $x \rightarrow y$

Seule la régression linéaire simple de modèle I est étudiée au Bio 2041.

Le terme *régression* a une origine curieuse. Il remonte à l'étude du physiologiste et anthropologue Sir Francis Galton (1886, 1889) sur la relation entre la taille des parents et celle des enfants. Galton était un cousin de Charles Darwin. Il observa que les enfants de parents courts de taille, lorsqu'ils étaient rendus à l'âge adulte, avaient tendance à être de petite taille eux aussi, mais pas autant que leurs parents. Ils avaient plutôt une taille les rapprochant de la moyenne de la population. Il en était de même des enfants de parents de grande taille: leurs enfants semblaient *régresser* vers la moyenne (dans le sens de "retourner vers un état antérieur"), comme semblait le montrer le diagramme de dispersion. Galton appela "rapport de régression filiale" la pente de la relation graphique linéaire entre la taille des parents et celle des enfants.

On parle de *régression linéaire* lorsqu'on désire calculer une fonction du premier degré liant les variables *y* et *x*. Cette fonction linéaire, de forme

$\hat{y} = b_0 + b_1 x$ [qui peut aussi être notée $\hat{y} = ax + b$, ou encore $\hat{y} = a + bx$],

correspond à l'équation d'une *ligne droite* traversant le nuage de points et permettant de calculer une valeur estimée $\hat{y}$ pour chaque point de l'axe des *x*, correspondant à la variable prédictive (Scherrer, Fig. 18.1).

Cette droite porte le nom de *droite d'estimation* ou *droite de régression de y en x* (ou *de y sur x*).

Vocabulaire

• *Paramètre*: quantité fixe dans une expression (modèle) mathématique.

• *Variable*: quantité qui prend ou peut prendre plusieurs valeurs distinctes dans une expression mathématique. Exemple: *x* et $\hat{y}$ dans $\hat{y} = b_0 + b_1 x$.

• *MCO* = moindres carrés ordinaires (*OLS = ordinary least squares*).

• *Modèle mathématique*: représentation simplifiée, à l'aide d'une équation, de relations empiriques ou posées par hypothèse.

La régression est une forme de modélisation. Elle peut avoir plusieurs objectifs:

• Description: trouver le meilleur modèle fonctionnel liant la variable dépendante *y* à la (aux) variable(s) indépendante(s) *x*. Estimer la valeur la plus probable des paramètres du modèle, ainsi que leur intervalle de confiance.

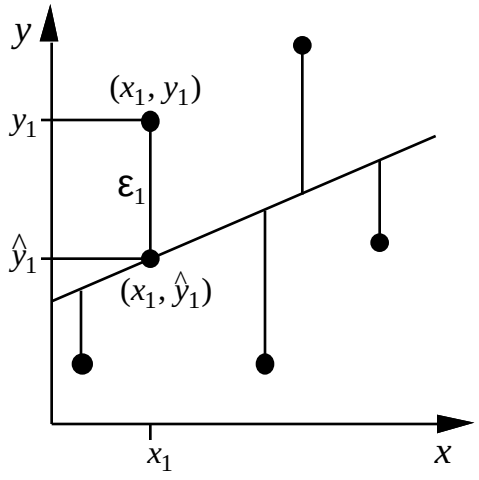
• Inférence: tester des hypothèses précises se rapportant aux paramètres du modèle dans la population statistique: ordonnée à l'origine, pente(s).

• Prédiction: prévoir ou prédire les valeurs de la variable dépendante pour de nouvelles valeurs de la (des) variable(s) indépendante(s).

2. Principe des moindres carrés

La découverte de la méthode des moindres carrés est attribuée à deux mathématiciens, Adrien Marie Le Gendre (France) et Karl Friedrich Gauss (Prusse). Tous deux travaillaient à des problèmes d'astronomie.

On désire faire passer la droite d'estimation, à travers le nuage de points, de façon à ce que les différences $(y - \hat{y})$ soient les plus faibles possible pour l'ensemble des points.



$\hat{y}_1$ est l'estimation de y_1 faite par l'équation de régression pour la valeur $x = x_1$.

On peut montrer que de minimiser la somme des carrés des écarts aux estimations $\Sigma (y_i - \hat{y}_i)^2$ conduit à une solution minimale unique, alors que ce n'est pas le cas avec les fonctions $\Sigma|y_i - \hat{y}_i|$ ou $\Sigma (y_i - \hat{y}_i)$.

La différence $\varepsilon_i = (y_i - \hat{y}_i)$ porte le nom de *résidu* pour l'observation *i*.

On peut aussi montrer que la droite satisfaisant au critère des moindres carrés ordinaires passe par le centre de masse $(\bar{x}, \bar{y})$ du nuage de points.

3. Calcul de la droite de régression par moindres carrés ordinaires

Comment peut-on, à partir du critère des moindres carrés, estimer les paramètres de l'équation linéaire (modèle) $\hat{y} = b_0 + b_1 x$?

Rappel: Dans cette équation, (Notation de Scherrer)

- $\hat{y}_i$ est l'estimation de y_i faite à partir de x_i (valeur ajustée, *fitted value*).
- b_0 = ordonnée de la droite d'estimation à l'origine (i.e., quand $x = 0$).
- b_1= pente de la droite d'estimation.

Puisque $\hat{y} = b_0 + b_1 x$

par conséquent chaque résidu $(y_i - \hat{y}_i) = [y_i - (b_0 + b_1 x_i)]$

donc la somme des carrés des résidus à minimiser est

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

Rappelons la procédure classique du calcul différentiel: on trouve le minimum d'une fonction en égalant sa dérivée première à zéro. Dans le cas qui nous occupe, il suffit de calculer successivement les dérivées partielles par rapport aux deux paramètres, b_0 et b_1. Les valeurs recherchées des paramètres sont celles qui satisfont simultanément ces deux équations, égales à zéro. Développement dans Scherrer, p. 694.

Pour la pente b_1, on trouve l'équation

$$b_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma\left(x_i^2\right) - \left(\Sigma x_i\right)^2} \quad \text{(formule développée)}$$

or nous savons que la covariance a pour formule

$$s_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\,(n-1)} \qquad \text{Scherrer, éq. 4.41}$$

et que la variance a pour formule

$$s_x^2 = \frac{n\Sigma\left(x_i^2\right) - \left(\Sigma x_i\right)^2}{n\,(n-1)} \qquad \text{Scherrer, éq. 4.18}$$

donc $b_1 = s_{xy} / s_x^2$ Scherrer, éq. 18.10

De plus, nous savons que $r_{xy} = s_{xy}/s_x\, s_y$ (éq. 17-3), donc $s_{xy} = r_{xy}\, s_x\, s_y$.

Par conséquent, $b_1 = \frac{r_{xy} s_x s_y}{s_x^2}$ et donc $b_1 = r_{xy}\,\frac{s_y}{s_x}$ et $r_{xy} = b_1 \frac{s_x}{s_y}$.

Pour l'ordonnée à l'origine b_0, on trouve $b_0 = \bar{y} - b_1 \bar{x}$ éq. 18.11

On peut donc calculer b_0 à partir de la pente, b_1, ainsi que du centre de masse $(\bar{x}, \bar{y})$ (centroïde) du nuage de points.

Langage R: la fonction *lm* ("linear model") estime les paramètres l'équation de régression par moindres carrés. Les statistiques associées (R^2, etc.) ainsi que les tests de significations sont fournis par *summary.*

4. Calcul des paramètres par inversion matricielle

[Cette section n'est pas matière à examen.]

Tableau de données:

y	x
-----	-----
-----	-----
-----	-----
-----	-----
-----	-----
-----	-----
-----	-----

Posons l'équation **y** = **X Param** avec la notation suivante:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \\ 1 & x_n \end{bmatrix} \quad \mathbf{Param} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

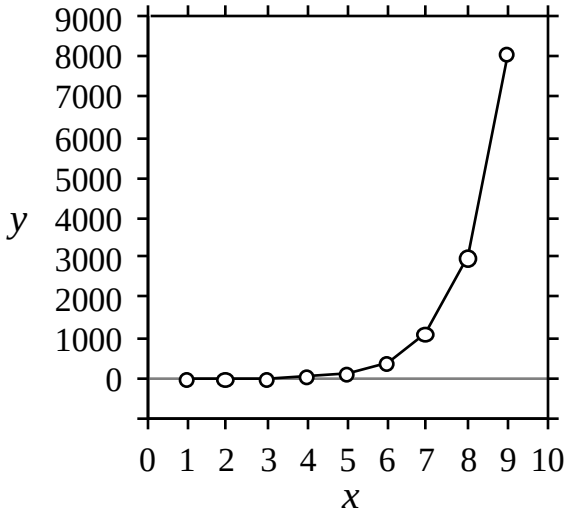
X'y = **X'X Param** $\qquad$ $[\mathbf{X'X}]^{-1}[\mathbf{X'y}] = [\mathbf{X'X}]^{-1}\,[\mathbf{X'X}]\,\mathbf{Param}$

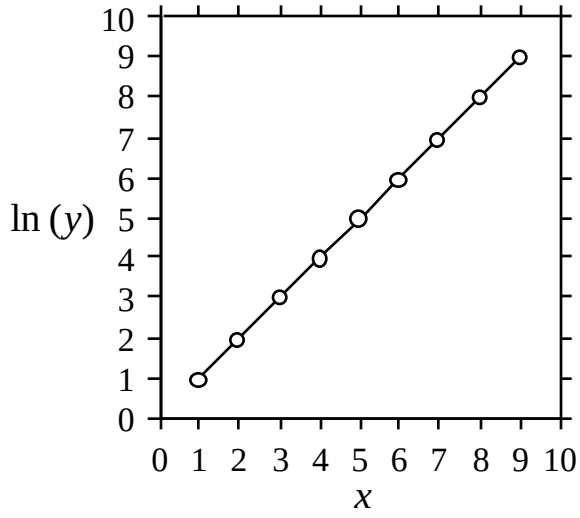
On estime les paramètres b_0 (ordonnée à l'origine) et b_1 (pente) de l'équation de régression en calculant **Param** = $[\mathbf{X'X}]^{-1}$ $[\mathbf{X'y}]$ (référence: Legendre & Legendre 1984 ou 1998, chapitre 2).

5. Linéarité

La méthode de calcul ci-dessus (régression *linéaire*) n’a de l’intérêt que si la relation entre la variable dépendante et la variable indépendante est *linéaire.* Si tel n’est pas le cas, trois méthodes s’offrent à nous.

• On peut tenter de linéariser la relation. Par exemple:





• Si on ne connaît pas la forme algébrique de la relation, on peut utiliser la *régression polynomiale*. Cette méthode consiste à ajuster à y un polynôme en x d’un degré approprié, de la forme:

$$\hat{y} = b_0 + b_1x + b_2x^2 + \ldots + b_kx^k$$

• Si on connaît la forme mathématique (i.e., l’équation) de la relation recherchée, on peut recourir à la *régression non-linéaire.* Cette méthode permet de spécifier l’équation dont on désire estimer les paramètres.

Ces formes plus complexes de régressions [Scherrer (1984) p. 669 et 722] seront étudiées aux cours Bio 2042 et Bio 6077.

6. Quelques propriétés [Scherrer, section 18.1.3]

1) Les deux droites d'estimation

La droite d'estimation de y en x, $\hat{y} = b_0 + b_1 x$, n'est pas la même (et n'a pas les mêmes paramètres) que la droite d'estimation de x en y, $\hat{x} = cy + d$. Dans le premier cas, la pente est $b_1 = s_{xy} / s_x^2$; dans le second cas, elle se calcule par $c = s_{xy} / s_y^2$. Note: $b_1 \neq 1/c$ mais $b_1 = c\left(s_y^2 / s_x^2\right)$.

Illustration: Scherrer, Fig. 18.7.

L'angle entre les deux droites de régression (α dans Scherrer, Fig. 18.7) s'amenuise à mesure qu'augmente la corrélation entre les deux variables. À la limite, quand la corrélation est 1, les deux droites sont confondues.

À l'autre extrême, lorsque $r = 0$, les deux droites d'estimation sont perpendiculaires entre elles et parallèles aux axes. On dit alors que les variables sont *linéairement indépendantes* l'une de l'autre.

La relation entre le coefficient de corrélation et l'angle α entre les deux droites de régression (Fig. 18.7, corriger!) est décrite par la formule:

$$\alpha = 90° - \left[\tan^{-1}(r\, s_x / s_y) + \tan^{-1}(r\, s_y / s_x) \right] \text{ donc}$$

$$r = \tan\left[\frac{(90° - \alpha)}{2} \right]$$

(démonstration dans Legendre & Legendre 1998: 503).

2) Partition de la variance

On peut diviser la variance totale de la variable dépendante *y* en deux parts, eu égard à la droite d'estimation:

• une partie expliquée par cette estimation;

• une partie résiduelle.

Pour y arriver, considérons d'abord un seul point (x, y) et son estimation (x, $\hat{y}$) par la droite de régression (Fig. 18.4, flèches 1, 2 et 5).

• La contribution du point (x, y) à la variance globale est ($y - \bar{y}$).

Cet écart est formé de deux parties identifiables à la Fig. 18.4:

$(y - \bar{y}) = (y - \hat{y}) + (\hat{y} - \bar{y})$ Dans la figure, "1" = "5" + "2"

Ceci reste vrai pour tous les valeurs *y,* qu'elles soient supérieures ou inférieures à $\bar{y}$. Donc, on peut effectuer la somme de ces expressions pour tous les points du nuage de points et obtenir:

$$\Sigma (y_i - \bar{y}) = \Sigma [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]$$

Pour obtenir la somme des carrés des écarts totale (SCT), qui forme le numérateur de l'expression de la variance, il suffit de mettre les deux membres de cette équation au carré:

$\Sigma (y_i - \bar{y})^2 = \Sigma [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$ Forme: $[a + b]^2$

$$\Sigma (y_i - \bar{y})^2 = \Sigma (y_i - \hat{y}_i)^2 + 2 \Sigma (y_i - \hat{y}_i) (\hat{y}_i - \bar{y}) + \Sigma (\hat{y}_i - \bar{y})^2$$

Le terme central, $\Sigma (y_i - \hat{y}_i) (\hat{y}_i - \bar{y})$, est un terme de covariance sans ses degrés de liberté. Or la solution des moindres carrés nous garantit l'indépendance des composantes "estimation" $(\hat{y}_i - \bar{y})$ et "résidu"

$(y_i - \hat{y}_i)$; démonstration dans Scherrer, p. 697. Nous voyons une démonstration du même type dans le cours sur l'Anova. Ce terme central a donc la valeur 0, ce qui simplifie l'équation décrivant la partition de la somme des carrés des écarts:

$$\Sigma (y_i - \bar{y})^2 = \Sigma (y_i - \hat{y}_i)^2 + \Sigma (\hat{y}_i - \bar{y})^2$$

Dispersion totale = disp. résiduelle + disp. expliquée par la régression

SCT = SCE + SCR Scherrer, Fig. 18.23

Dans cette équation, SCE mesure la quantité de variabilité qui n'a pas été expliquée par l'estimation, soit la dispersion des valeurs observées de part et d'autre de la droite de régression.

Comme dans le chapitre portant sur l'analyse de variance, nous pouvons disposer les éléments de variabilité que nous avons identifiés dans un tableau d'analyse de variance: Scherrer, tableau 18.1.

3) Coefficient de détermination r^2

Cette partition de la variance nous permet de dériver une mesure évidente de la justesse de la prédiction. Il s'agit du rapport de la variation expliquée par l'estimation sur la variation totale (la variation est mesurée par les termes SC ["sommes de carrés"]):

$$\text{Coeff. de détermination} = \frac{\text{SCR}}{\text{SCT}} = \frac{r^2 \Sigma (y_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2} = r^2_{xy} \qquad \text{éq. 18.41}$$

On constate que le *coefficient de détermination*, qui mesure la proportion de la variation de *y* expliquée par *x*, est égal au carré du coefficient de corrélation linéaire. Contrairement à *r*, r^2 est toujours positif ou nul.

Note: la valeur $(1 - r^2)$ = SCE/SCT mesure la *non-détermination*.

7. Test de signification de la pente b_1

On a vu à la section 6 que, si *x* et *y* sont des variables linéairement indépendantes (non corrélées), les droites d'estimation de *y* en *x* et de *x* en *y* sont perpendiculaires entre elles et parallèles aux axes de coordonnées. En corollaire, si la droite de régression de *y* en *x* est parallèle à l'abscisse ($b_1 = 0$), on ne peut pas identifier de droite de régression particulière représentant une tendance du nuage de points.

On peut étudier la signification du paramètre b_1, décrivant la pente de la droite d'estimation, par l'examen de son intervalle de confiance (section 8) ou, ce qui est plus simple, par un test *F* du rapport des deux fractions indépendantes de la variance que nous avons isolées au tableau 18.1:

$$F = \frac{\text{Variance expliquée par la régression}}{\text{Variance due aux erreurs}} = \frac{CMR}{CME} \text{ avec 1 et } (n-2) \text{ d.l.}$$

• H_0: $\beta_1 = 0$, où β_1 représente le paramètre *pente* de la population statistique d'où est extrait l'échantillon (b_1 dans l'échantillon). Si H_0 est vraie, la variance expliquée par la régression (*CMR*) calculée pour l'échantillon est à peu près égale à la variance de l'erreur (*CME*). La probabilité d'erreur de type I est égale à α (le seuil de signification).

• H_1: $\beta_1 \neq 0$. Si H_1 est vraie, la variance expliquée par la régression (*CMR*) devrait être supérieure à la variance des erreurs (*CME*).

Le rapport de variances *F* peut se calculer directement à partir du tableau d'analyse de variance.

On constate par ailleurs que ce rapport *F* peut être exprimé en termes de corrélation, en le transformant à l'aide des expressions dérivées plus haut:

$$F = \frac{CMR}{CME} = \frac{r^2 \Sigma (y_i - \bar{y})^2}{\left(\Sigma (y_i - \bar{y})^2 - r^2 \Sigma (y_i - \bar{y})^2 \right) / (n-2)} = \frac{r^2 (n-2)}{1 - r^2}$$

Le numérateur provient du tableau 18.1.

Le dénominateur provient du raisonnement suivant:

CME = SCE/(n – 2). Or SCE = SCT – SCR. Nous savons par ailleurs que SCT = $\Sigma (y_i - \bar{y})^2$ et que SCR = $r^2 \Sigma (y_i - \bar{y})^2$.

Ceci est exactement l'expression utilisée pour le test de signification d'un coefficient de corrélation linéaire. On constate donc que le test de la corrélation r (H_0: $\rho_{xy} = 0$) est équivalent au test développé ici pour le paramètre b_1 décrivant la pente de la droite d'estimation (H_0: $\beta_1 = 0$).

En régression linéaire simple, il n'y a qu'un degré de liberté au numérateur de la statistique F. On peut donc employer plutôt la table de t pour tester la statistique $t = \sqrt{F}$ avec $\nu = (n - 2)$:

$$t = \sqrt{F} = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}}$$

• Conditions d'application du test: ce sont les mêmes que pour le test de la corrélation. Si des problèmes de distribution se posent, on peut tester la signification du paramètre *pente* par la méthode des permutations.

• Règles de décision: Tableau 18.2. Le test de signification de la pente peut être unilatéral ou bilatéral selon le problème.

Exemples: Voir Scherrer, p. 695, 699; test par permutations possible.

8. Intervalle de confiance de la pente b_1

Scherrer (p. 697) montre que la statistique-test

$t = \dfrac{b_1 - \beta_1}{\sqrt{\text{Var}\,(b_1)}}$ obéit à une loi de Student à $(n - 2)$ d.l. éq. 18.19

L'*erreur type* de b_1 est estimée par

$$\sqrt{\text{Var}\,(b_1)} = \sqrt{\frac{CME}{(n-1)\,s_x^2}} = \sqrt{\frac{\text{SCE}}{(n-2)\;(n-1)\,s_x^2}} = \frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}$$

L'erreur type de b_1 permet de calculer l'intervalle de confiance de ce paramètre dans la population statistique:

$$\Pr\left[b_1 - t_{\alpha/2}\sqrt{\text{Var}\,(b_1)} < \beta_1 < b_1 + t_{\alpha/2}\sqrt{\text{Var}\,(b_1)}\;\right] = 1 - \alpha \quad \text{éq. 18.18}$$

Objectif = inférence — L'intervalle de confiance de b_1 permet de vérifier, par exemple, si une valeur de pente prédite par la théorie biologique (par exemple, $b_1 = 0$ ou encore $b_1 = 1$) se trouve effectivement à l'intérieur de l'intervalle de confiance calculé pour un seuil de signification prédéterminé.

L'intervalle de confiance peut aussi servir d'équivalent d'un test de signification. Si H_0: $\beta_1 = 0$ est vraie, alors la pente 0 est incluse dans l'intervalle de confiance dans une proportion de $1 - \alpha$ des cas. Si la pente 0 n'est pas comprise dans l'intervalle de confiance, on rejette H_0 avec un risque α de se tromper.

9. Intervalle de confiance de l'ordonnée à l'origine b_0

Scherrer (p. 701) montre également que la statistique-test

$$t = \frac{b_0 - \beta_0}{\sqrt{\text{Var}(b_0)}}$$ obéit à une loi de Student à $(n - 2)$ d.l. éq. 18.32

β_0 représente le paramètre *ordonnée à l'origine* de la population statistique d'où est extrait l'échantillon. L'*erreur type* de b_0 est estimée par

$$\sqrt{\text{Var}(b_0)} = \sqrt{\frac{CME \ \ \Sigma x_i^2}{n\Sigma (x - \bar{x})^2}} = \frac{s_y\sqrt{(1 - r^2)\,\Sigma x_i^2}}{s_x\sqrt{n\,(n-2)}}$$ éq. 18.31

L'erreur type de b_0 permet de calculer l'intervalle de confiance de ce paramètre dans la population statistique:

$$\Pr\left[b_0 - t_{\alpha/2}\sqrt{\text{Var}(b_0)} < \beta_0 < b_0 + t_{\alpha/2}\sqrt{\text{Var}(b_0)}\right] = 1 - \alpha$$ éq. 18.33

Objectif = inférence — L'intervalle de confiance de b_0 permet de vérifier, par exemple, si une valeur d'ordonnée à l'origine (par exemple, $b_0 = 0$) prédite par la théorie biologique se trouve effectivement à l'intérieur de l'intervalle de confiance calculé pour un seuil de signification prédéterminé. On peut aussi s'en servir comme d'un test, de la même façon que l'I.C. de la pente.

10. Intervalle de confiance de la valeur estimée $\hat{y}$

Cet intervalle de confiance sera étudié au cours Bio 2042.

11. Prévision et prédiction

L'un des objectifs d'une analyse de régression peut être la prévision (souvent appelée "prédiction") de la valeur de *y* correspondant à des valeurs de *x* pour lesquelles *y* n'a pas été observée.

• Une prévision est un type d'interpolation linéaire parmi les valeurs observées de *x*. Par exemple, pour estimer la valeur la plus vraisemblable de la concentration en pesticides dans la chair de brochets âgés de 3,5 ans (exemple 18.1), on peut utiliser le modèle estimé par régression linéaire pour calculer la valeur prévue de $\hat{y}$ pour $x = 3{,}5$ ans:

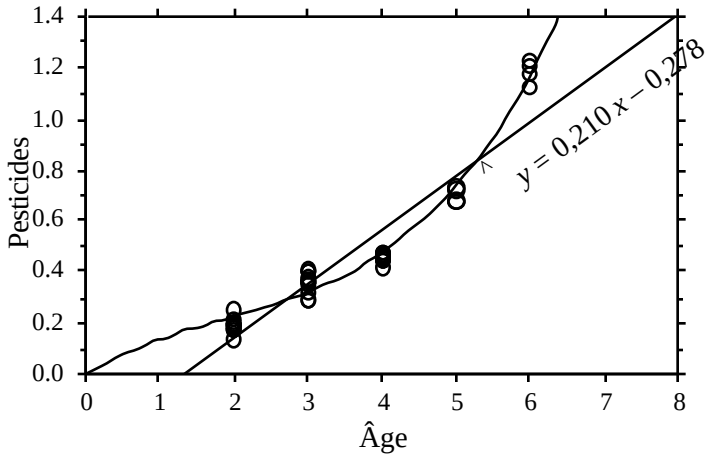
$$\hat{y}_{\text{âge = 3,5 ans}} = (0{,}210 \times 3{,}5) - 0{,}278 = 0{,}457 \text{ µg/g}$$

• Une prédiction au sens strict (de Neufville & Stafford 1971; Legendre & Legendre 1998) implique une bonne compréhension du processus biologique que reflète le modèle. Par exemple, on peut désirer prédire la concentration en pesticides dans la chair de brochets âgés de 8 ans. Si, comme ci-dessus, on utilise le modèle linéaire estimé par régression pour calculer la valeur prévue de $\hat{y}$ pour $x = 8$ ans, on obtient:

$$\hat{y}_{\text{âge = 8 ans}} = (0{,}210 \times 8) - 0{,}278 = 1,402 \text{ µg/g}$$

Avant d'extrapoler au delà des limites pour lesquelles le modèle a été calibré, il aurait fallu s'assurer que le modèle linéaire demeure valide au delà de ces limites. Le diagramme de dispersion peut nous renseigner:

L'extrapolation était-elle justifiée dans ce cas?



12. Conditions d'application du modèle I de régression

Modèle I (moindres carrés ordinaires)

• *x* et *y* sont des variables quantitatives.

• Modèle recherché: relation linéaire.

• Variable *x* contrôlée (sans erreur de mesure).

Conditions additionnelles d'application des tests de signification en régression par moindres carrés ordinaires

• Pour toute valeur de *x*, les résidus ($y - \hat{y}$) sont distribués normalement.

• Pour toute valeur de *x*, les résidus ($y - \hat{y}$) ont la même variance (homoscédasticité). *Examiner les résidus:* section 18.1.8, p. 705.

• Indépendance des observations (absence d'autocorrélation).

Modèle II (cours Bio 2042)

• Variables quantitatives.

• Modèle recherché: relation fonctionnelle linéaire entre deux variables.

• *x* et *y* sont des variables aléatoires.

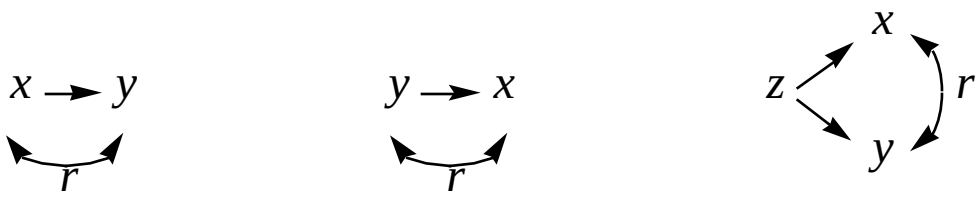
Si la variable *x* est aléatoire mais a été mesurée de façon beaucoup plus précise que *y*, le modèle II approprié est MCO (McArdle, 1988).

En pratique, les chercheurs utilisent souvent les Moindres Carrés Ordinaires (MCO, régression de modèle I) même si certaines conditions ne sont pas respectées, parce que les méthodes de modèle II ne sont pas disponibles dans les logiciels statistiques courants. Ceci peut être acceptable comme première approximation, lorsque l'objectif est uniquement descriptif (i.e., pas de test de signification des paramètres).

13. Corrélation ou régression?

La régression est une forme de modélisation d'une *variable dépendante*. Son but est d'estimer les paramètres du modèle, de tester des hypothèses se rapportant aux paramètres, ou encore de prévoir ou prédire *y* à l'aide de *x*. Modèle: $x \rightarrow y$.

La corrélation est plutôt une statistique descriptive de l'*interdépendance* (linéaire) entre deux variables, sans aucune supposition de relation fonctionnelle ou causale entre elles. On n'exprime aucune préférence entre les trois modèles suivants:



On peut également employer la corrélation dans la phase exploratoire d'une recherche lorsqu'on croit qu'une relation de dépendance peut exister mais sans être certain du modèle de causalité qui s'applique.

La confusion entre corrélation et régression est due au coefficient de détermination r^2. Le r^2 est essentiel pour juger du pouvoir prévisionnel de l'équation de régression et il est algébriquement l'équivalent du (*r* de Pearson)^2. Cependant, les objectifs des deux méthodes diffèrent.

Lorsque la *variable indépendante* est contrôlée, si on cherche à établir que la variable dépendante est fonction de la variable indépendante, il est clair que c'est la régression de modèle I qui doit être utilisée. La corrélation (*interdépendance*) n'a pas de sens dans ce cas.

Lorsque les deux variables sont aléatoires, la distinction est plus difficile. Le choix de la méthode dépend alors de l'intention du chercheur: on peut modéliser par la régression de modèle II, qui sera étudiée au cours Bio 2042, ou mesurer simplement l'interdépendance par la corrélation.

14. Relation entre l’analyse de variance et la régression

Toute analyse de variance peut être calculée par la technique de la régression linéaire (simple ou multiple). La relation est basée sur le fait que tout critère de classification de l’ANOVA peut être recodé en une ou plusieurs variables factices binaires ou variables muettes (“*dummy variable*” en anglais).

Exemple 1 — Le critère de classification “capacité de rouler la langue” de la base de données Bio 2041 (un jeu de données constitué de caractéristiques d’étudiants d’une volée précédente). Il s’agit d’une variable binaire:
0 = incapable de rouler la langue
1 = capable de rouler la langue

Si on utilise cette variable binaire comme variable indépendante (*x*) d’une régression linéaire simple contre la variable dépendante (*y*) “taille”, par exemple, le tableau d’analyse de variance de la régression sera identique à celui d’une ANOVA de la variable “taille” par le critère de classification “capacité de rouler la langue”. Le nombre de variables muettes nécessaires pour coder tous les niveaux d’un critère de classification est égal au nombre de ces niveaux moins 1 (voir l’exemple suivant). Donc, le nombre de variables muettes nécessaires est égal au nombre de *degrés de liberté* du facteur.

Exemple 2 — Reprenons les exemples 2.38 (p.68) et 14.8 de Scherrer (p. 536) portant sur la concentration en pesticides dans les chairs de brochets appartenant à différents groupes d'âge. Le critère de classification est constitué de 5 classes d'âge. Celles-ci peuvent être recodées en variables muettes comme suit:

	Variables muettes				
Âge	âge = 2	âge = 3	âge = 4	âge = 5	âge = 6
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6	0	0	0	0	1

En fait, l'une de ces variables muettes (par exemple la dernière) est inutile puisqu'elle constitue une combinaison linéaire de toutes les autres. En pratique, on élimine l'une des variables muettes de l'analyse — au choix de l'utilisateur. Le résultat restera le même quelle que soit la variable éliminée. Il reste donc 5 – 1 = 4 classes, ce qui correspond au nombre de degrés de liberté du facteur "âge" dans une ANOVA.

À chaque valeur *y* (concentration en pesticides) on attribue la liste des variables muettes correspondant à sa classe d'âge.

La régression multiple de la variable *y* (concentration en pesticides) contre le tableau des variables muettes (variables indépendantes *x*) produit un tableau d'analyse de variance équivalent à celui obtenu par ANOVA (p. 536-537; corriger les valeurs de Scherrer).

Références additionnelles

De Neufville, R. & J. H. Stafford. 1971. *Systems analysis for engineers and managers.* McGraw-Hill, New York. xiii + 353 pp.

Galton, F. 1886. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute* 15: 246-263.

Galton, F. 1889. *Natural inheritance.* Macmillan & Co., London. ix + 259 pp.

Gauss, K. F. 1809. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Frid. Perthes et I. H. Besser, Hamburg.

Le Gendre, A. M. 1805. *Nouvelles méthodes pour la détermination des orbites des comètes.* Courcier, Paris.

Legendre, P. & L. Legendre. 1998. *Numerical ecology, 2nd English edition.* Elsevier Science BV, Amsterdam. xv + 853 pp.

McArdle, B. 1988. The structural relationship: regression in biology. *Can. J. Zool.* **66**: 2329-2339.

Dérivation de certaines formules du tableau 18.1 de Scherrer (2007: 698)

* À partir de tout point estimé (x_i, $\hat{y}_i$), on peut calculer la pente *a* de la droite de régression à l'aide de l'équation

$$b_1 = \frac{\hat{y}_i - \bar{y}}{x_i - \bar{x}} \rightarrow (\hat{y}_i - \bar{y}) = b_1 (x_i - \bar{x})$$

En sommant les carrés des valeurs pour toutes les observations *i*, on obtient $\Sigma (\hat{y}_i - \bar{y})^2 = b_1{}^2 \Sigma (x_i - \bar{x})^2$.

$$** \; b_1 = r_{xy} \frac{s_y}{s_x} \rightarrow b_1{}^2 = r_{xy}^2 \frac{s_y^2}{s_x^2} \rightarrow b_1{}^2 = r_{xy}^2 \frac{\Sigma (y_i - \bar{y})^2}{\Sigma (x_i - \bar{x})^2}$$

$$\rightarrow b_1{}^2 \Sigma (x_i - \bar{x})^2 = r_{xy}^2 \; \Sigma (y_i - \bar{y})^2$$

*** Le calcul de la variance due aux erreurs *CME* (s_e^2 dans l'édition 1984 du manuel) requiert qu'on ait calculé d'abord les deux paramètres b_0 et b_1 de l'équation de régression, car l'équation est nécessaire pour pouvoir calculer les valeurs $\hat{y}_i$. Ceci fait perdre deux degrés de liberté.