Daniel Borcard 2001-2006
Département de sciences biologiques
Université de Montréal

Régression multiple

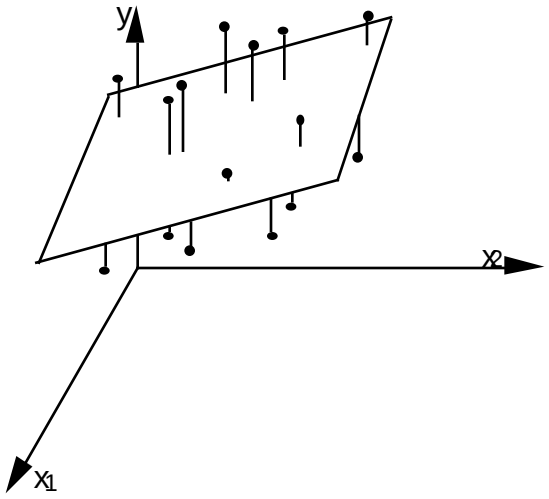
Scherrer: p.690; Sokal & Rohlf: p. 617; Legendre & Legendre (1998) p. 517

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, $a_2 \ldots a_{p-1}$ 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 la variation d'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 un résumé de la technique 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:

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

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.

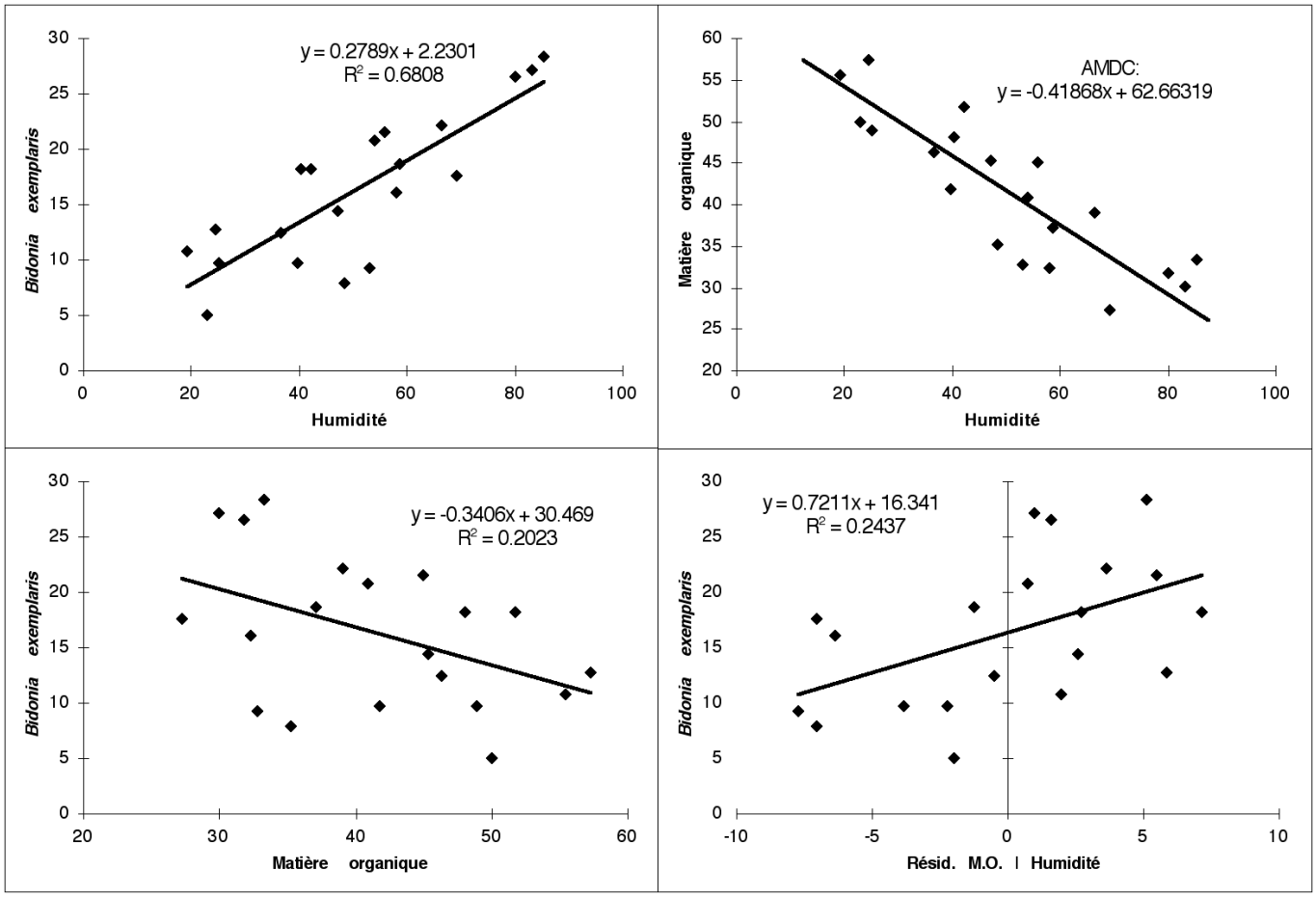
Point important, les coefficients de régression obtenus par régression multiple sont en fait des coefficients de **régression partielle**, en ce sens que chacun mesure l'effet de la variable explicative concernée sur la variable dépendante **lorsque la ou les autres variables explicatives sont tenues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

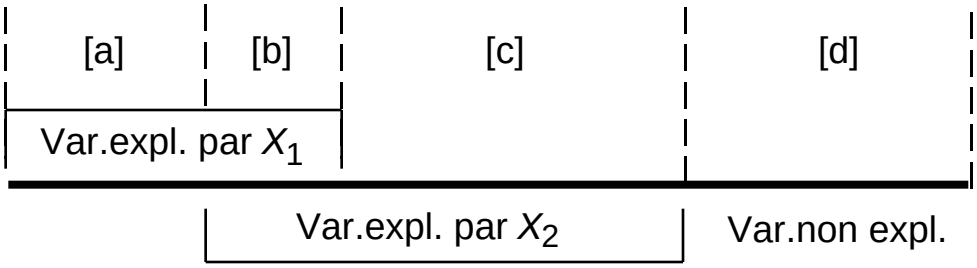
Pour contrer cet effet, et donc éviter de surestimer le R^2, plusieurs auteurs ont proposé un ***R*² ajusté**, qui tient compte du nombre de variables explicatives du modèle de régression. La formule la plus couramment utilisée est la suivante:

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

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 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]. L'explication unique de la variable x_1 est la fraction [a], et l'explication 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].

Étape intermédiaire: il faut maintenant ajuster les r^2 et R^2 ci-dessus à l'aide de la formule du R^2 ajusté (p. 9). Ensuite:

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

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

(6) La valeur de $[b]_{aj}$ s'obtient de diverses manières, p. ex. $([a]+[b])_{aj} - [a]_{aj}$, ou $([b]+[c])_{aj} - [c]_{aj}$.

(7) La fraction $[d]_{aj}$ (variation non expliquée) s'obtient en faisant $1 - ([a]+[b]+[c])_{aj}$.

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*. 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 construire un 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 (en commençant par celle qui explique le moins de variation). 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**:

1. 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.
2. Le modèle devient hautement instable en présence de (multi) colinéarité entre les variables explicatives, ce qui veut dire que les paramètres estimés par la méthode (les coefficients *a*, donc 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.
3. Il semble que quelle que soit la variante de sélection pas-à-pas utilisée, cette méthode est un peu trop libérale, c'est-à-dire qu'elle laisse souvent au moins une variable non significative dans le modèle.

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 (celle de Scherrer), 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 :

$$b = y - a_1x_1 - a_2x_2 - a_3x_3$$

Remarque: il existe une méthode différente pour calculer les coefficients de régression multiple, basée sur le calcul matriciel. C'est celle qui est utilisée dans les programmes d'ordinateur. 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 (au sens de Scherrer).

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.