REGRESSION LINEAIRE MULTIPLE

Méthodes, applications, programmes

Thierry Foucart
Département de Mathématiques
Université d'Orléans

INTRODUCTION

La régression linéaire est la méthode statistique vraisemblablement la plus utilisée par les praticiens de toutes disciplines: la recherche d'une liaison entre deux ou plusieurs caractères est une démarche très courante en médecine, en psychologie, en physique, en économie etc...

Cette recherche correspond en premier lieu à la vérification d'un modèle construit pour représenter une certaine réalité: c'est le cas des sciences exactes, physique, chimie par exemple, où le modèle doit représenter très fidèlement les phénomènes étudiés. C'est aussi le cas des sciences humaines, économie, sociologie, psychologie ..., où l'on se contente par contre d'une approximation jugée suffisante compte tenu du contexte.

Inversement, les liens existant entre différents paramètres peuvent être inconnus: il s'agit alors de déterminer les liaisons qui existent et d'en déduire un modèle approprié. Cette démarche, plus difficile que la précédente, est souvent employée sans suffisamment de précautions: une liaison mise en évidence par la statistique est en fait une question posée au praticien: pourquoi cette liaison existe-t-elle? Il faut apporter une explication déterministe à un phénomène détecté par des méthodes basées sur l'analyse du hasard.

Ces deux approches de la régression multilinéaire demandent une bonne connaissance théorique de la méthode, un logiciel commode et suffisamment puissant et une grande expérience dans le traitement des données statistiques.

La plupart des ouvrages généraux de statistique disponibles actuellement contiennent un chapitre au moins consacré à la régression: on ne rencontre donc guère de difficultés à se mettre au courant de la méthode au plan théorique. Nous donnons à la fin de cet article une courte bibliographie commentée.

On pourra se procurer pour un prix dérisoire le logiciel de régression multilinéaire LORELI (LOgiciel de REgression LInéaire), qui permet d'appliquer la plupart des méthodes classiques de régression linéaire, auprès de l'IREM d'Orléans.

Nous donnons dans le texte qui suit des applications détaillées des méthodes à des données réelles publiées dans l'ouvrage de Saporta et figurant en annexe. Tous les résultats numériques ont été établis à l'aide du logiciel LORELI.

Dans le texte qui suit, on ne discute que des méthodes de base de la régression linéaire; les lecteurs intéressés par les méthodes plus complexes telles que la régression sur composantes principales et la ridge régression, pourront consulter les ouvrages donnés en bibliographie en particulier ceux de Tomassone (en français) et de Weisberg (en anglais).

Chapitre 1

REGRESSION ET MODELE LINEAIRES

1. REGRESSION ET MODELE LINEAIRES. DEFINITIONS.

La régression et le modèle linéaires sont deux méthodes statistiques souvent confondues parce que les procédures de calcul sont les mêmes.

La régression concerne un couple de variables aléatoires (X,Y). On considère un échantillon $(X_i,Y_i)i=1,...,n$ de ce couple: les v.a. X_i et Y_i vérifient donc les propriétés suivantes:

— X_i et X_j sont indépendantes, de même que Y_i et Y_j, X_i et Y_j $(i \neq j)$.

— (X_i, Y_i) est un couple aléatoire de même loi que (X,Y).

Supposons que X et Y soient liées par la relation:

$$Y = f(X) + \varepsilon$$

où les v.a. ε et X sont indépendantes. On en déduit que cette relation est vérifiée pour tous les couples (X_i, Y_i):

$$\forall i=1,...,n \quad Y_i = f(X_i) + \varepsilon_i$$

Dans la formule précédente, les suites (Y_i), (X_i), (ε_i) constituent des échantillons des v.a. Y, X, et ε.

4 *Régression et modèle linéaires* *Chap. 1*

Un modèle s'applique lorsque la variable X n'est plus aléatoire mais contrôlée par l'utilisateur, ce qui est souvent le cas en physique par exemple; il consiste à poser:

$$\forall i=1,\ldots,n \quad Y_i = f(x_i) + \varepsilon_i$$

Dans ce modèle, les v.a. Y_i ne constituent pas un échantillon d'une v.a. puisqu'elles n'ont pas la même loi. L'hypothèse fondamentale du modèle linéaire est que, par contre, les v.a. ε_i constituent un échantillon d'une v.a. ε: elles sont indépendantes et de même loi.

On raisonne toujours en régression conditionnellement aux valeurs observées x_i, ce que l'on note X = x: il n'y a plus aucune différence formelle avec le modèle, et cela explique pourquoi on utilise parfois la terminologie de la régression en étudiant un modèle.

On cherche donc à ajuster une fonction f représentant la liaison entre les deux variables et à donner une estimation de ses paramètres. Lorsque la fonction f peut s'écrire sous la forme d'une fonction linéaire de ses paramètres à estimer, on parle de régression ou de modèle linéaire.

Les v. a. ε_i, qui constituent un échantillon d'une v.a. ε, suivent donc toutes la même loi de probabilité, que l'on suppose être dans la plupart des cas la loi normale centrée de variance σ^2. Il existe des modèles plus généraux, dans lesquels les ε_i ne sont pas indépendants ou n'ont pas la même variance, mais nous nous limiterons au cas le plus simple qui est d'ailleurs le plus employé.

L'hypothèse de normalité de la v.a. ε se justifie par le fait que ce terme d'erreur représente dans le modèle toute l'information non prise en compte; en physique par exemple, toutes les erreurs de mesure. Le cumul de ces erreurs justifie alors le choix de la loi normale.

Définitions:

— la variable Y est appelée variable expliquée (ou dépendante).

— la variable X èst appelée variable explicative (ou indépendante).

— la variable ε est appelée variable résiduelle.

— la variance σ^2 de la v.a. ε est appelée variance résiduelle.

Les représentations graphiques des couples (x_i, y_i) donnent des indications sur la nature de la fonction f à calculer; lorsqu'elle n'est pas linéaire, on peut procéder à une transformation des variables, par le logarithme par exemple; on peut penser à une fonction exponentielle de la fréquence cardiaque pour ajuster la résistance pulmonaire (fig. 1.1).

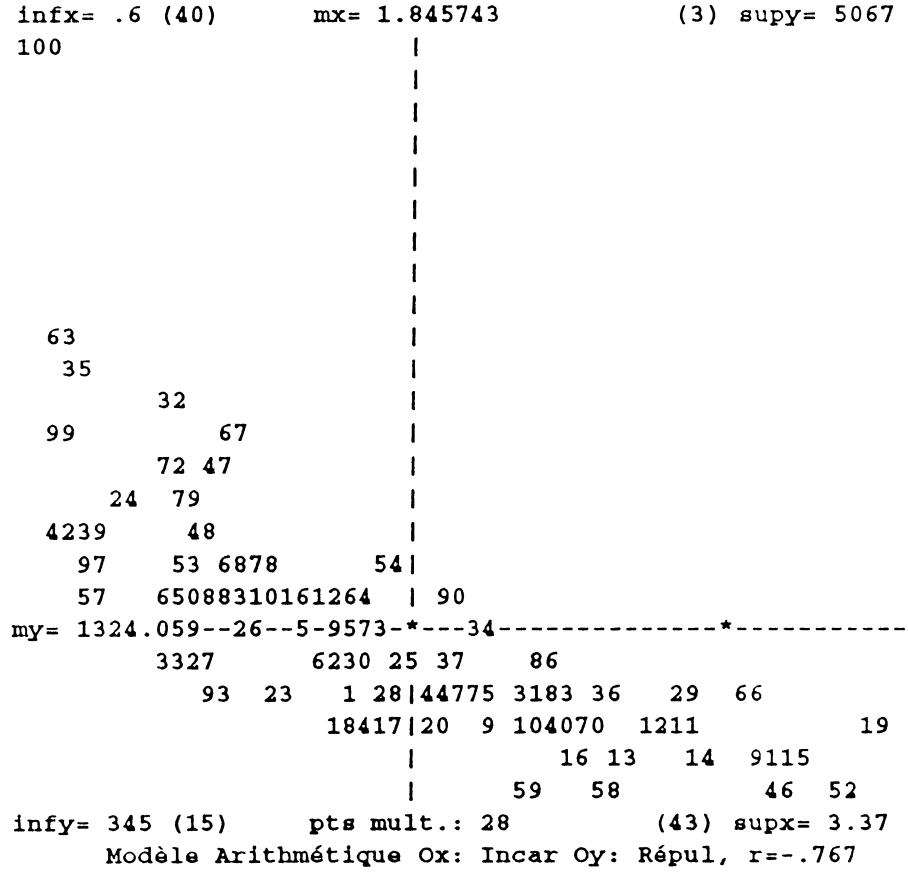

Fig.1.1: Index cardiaque x résistance pulmonaire
(repère arithmétique)

2. AJUSTEMENT D'UNE DROITE.

La courbe la plus simple à ajuster au nuage de points est la droite. Pour ajuster une fonction exponentielle, une simple transformation des données suffit pour se ramener au cas linéaire, alors que la régression polynomiale sera introduite comme une régression multilinéaire (chapitre 2).

Le modèle prend alors la forme, sur les valeurs observées:

$$\forall i=1,\ldots,n \quad y_i = a\,x_i + b + e_i$$

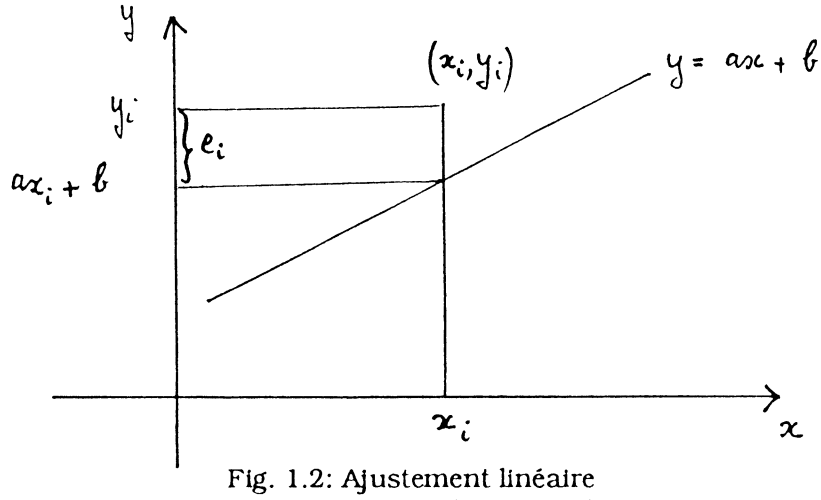

Fig. 1.2: Ajustement linéaire
(critère des moindres carrés)

La variable expliquée étant toujours placée en ordonnée, l'erreur que l'on commet en remplaçant y_i par $a\,x_i + b$ est $e_i = y_i - (a\,x_i + b)$.

On cherche à minimiser ces erreurs: le critère des moindres carrés consiste à minimiser la fonction S(a,b):

$$S(a,b) = \sum_{i=1}^{n} [y_i - (a\,x_i + b)]^2$$

Les résultats sont connus et figurent dans tous les livres; mais nous présentons la démonstration de façon à introduire le calcul matriciel que l'on

utilise en régression multiple.

On sait que le minimum de la fonction S(a,b), s'il existe, est obtenu pour les valeurs a et b qui annulent les dérivées partielles premières. Nous admettrons la condition suffisante, qui fait intervenir des dérivées partielles secondes, et calculons a et b tels que:

$$\partial S/\partial a = 0 \qquad \partial S/\partial b = 0$$

On a:

$$\partial S/\partial a = -2 \sum_{i=1}^{n} [(y_i - (a\, x_i + b)]\, x_i = 0$$

$$\partial S/\partial b = -2 \sum_{i=1}^{n} [(y_i - (a\, x_i + b)] = 0$$

On trouve le système des "équations normales":

$$\sum_{i=1}^{n} x_i^2\, a + \sum_{i=1}^{n} x_i\, b = \sum_{i=1}^{n} x_i\, y_i$$

$$\sum_{i=1}^{n} x_i\, a + n\, b = \sum_{i=1}^{n} y_i$$

Ce système se met sous la forme matricielle suivante:

$$M \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i\, y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} \quad \text{avec} \quad M = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix}$$

La résolution mathématique de ce système ne pose pas de problème. On trouve:

$$\boxed{a = \mathrm{cov}(x,y)/s^2(x) \qquad b = \bar{y} - a\,\bar{x}}$$

Ce qui nous intéresse ici, c'est la résolution matricielle, qui consiste à

inverser la matrice M (si elle est inversible). On obtient:

$$\begin{bmatrix} a \\ b \end{bmatrix} = M^{-1} \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}$$

Cette résolution matricielle du système des équations normales sera utilisée pour calculer les coefficients de régression b_j, j=0,...,p du modèle de régression multilinéaire (Cf. chap. 2).

3. ETUDE DES ESTIMATEURS ET DES RESIDUS.

Le modèle linéaire s'exprime de la façon suivante:

$$\forall i=1,\ldots,n \quad Y_i = \alpha x_i + \beta + \varepsilon_i$$

Définition: les coefficients α et β sont appelés coefficients de régression.

Nous avons calculé dans le paragraphe précédent les coefficients a et b de la droite la plus proche des points observés au sens des moindres carrés. Les valeurs que l'on obtient sont en fait des estimations des vrais coefficients, puisqu'ils dépendent de l'échantillon; nous les appellerons aussi coefficients de régression, en précisant qu'ils sont estimés en cas d'ambiguïté.

Théorème: les estimateurs A et B des coefficients α et β sont des estimateurs linéaires efficaces (sans biais et de variance minimale) indépendants de la v.a. ε. Si la variable résiduelle est gaussienne, ils sont gaussiens.

Nous ne démontrerons pas ce théorème ni les formules des variances des estimateurs A et B conditionnellement à x que nous donnons ci-dessous:

$$V(A) = \frac{\sigma^2}{n\, s^2(x)} \qquad V(B) = \frac{\sigma^2}{n} + \frac{\sigma^2\, \bar{x}^2}{n\, s^2(x)} \qquad cov(A,B) = -\frac{\sigma^2\, \bar{x}}{n\, s^2(x)}$$

Ces formules montrent que les estimateurs sont convergents (leurs variances tendent vers 0 lorsque n tend vers l'infini) et que la variance de B est une fonction croissante de $\bar{x}$. On peut montrer que le coefficient de corrélation entre A et B tend vers -1 lorsque $\bar{x}$ tend vers l'infini.

La réalisation de la v.a. ε_i est, d'après le modèle, égale à $y_i - (\alpha x_i + \beta)$. Ne connaissant pas les vraies valeurs des coefficients de régression α et β, on ne peut la calculer: la variable résiduelle n'est pas observable. On peut toutefois calculer les valeurs des erreurs commises après estimation des coefficients de régression:

$$e_i = y_i - (a\, x_i + b)$$

Définition: les termes $e_i = y_i - (a\, x_i + b)$ sont appelés résidus.

Ces résidus, qui ne constituent pas comme nous le verrons ultérieurement un échantillon d'une v.a., possèdent un certain nombre de propriétés que nous retrouverons en régression multiple et que nous admettrons:

— la série des résidus est de moyenne nulle:

$$1/n \sum_{i=1}^{n} e_i = 0$$

— elle est non corrélée avec la variable explicative:

$$1/n \sum_{i=1}^{n} e_i\, x_i = 0$$

— sa variance est égale à:

$$s^2(e) = 1/n \sum_{i=1}^{n} e_i^2 = s^2(y)\ (1-r^2)$$

où r est le coefficient de corrélation empirique:

$$r = \frac{cov\ (x,y)}{s(x)\ s(y)}$$

On trouve ici une autre interprétation du coefficient de corrélation: une valeur proche de ±1 montre que la variance de la série des résidus est faible par rapport à la variance observée de la variable expliquée. Ces propriétés ont pour conséquence:

Théorème: la quantité $s^2 = n\ s^2(e) / (n-2)$ est une estimation sans biais de la variance résiduelle σ^2.

L'estimation de la variance résiduelle est utilisée pour effectuer des prédictions de la variable expliquée en fonction de la variable explicative.

Le modèle étant donné par:

$$Y = \alpha\ x + \beta + \varepsilon$$

où ε suit la loi normale centrée, il est facile de montrer que l'espérance E(Y/X=x) de Y pour X=x, est égale à $\alpha\ x + \beta$.

La première prédiction que l'on effectue est donc celle de l'espérance conditionnelle de Y sachant X=x. On peut aussi vouloir prédire une valeur particulière: la formule est la même, seule la variance de la prévision diffère.

Théorème: L'estimateur de E(Y/X=x) est Y':

$$Y' = A\ x + B$$

dont la variance est égale à:

$$V(Y') = \frac{\sigma^2}{n} \left[1 + \frac{(x - \bar{x})^2}{s^2(x)} \right]$$

La variance de la prévision d'une valeur est obtenue en ajoutant le terme σ^2 à la variance précédente; pour estimer ces variances, il suffit de remplacer la variance résiduelle σ^2 par son estimateur sans biais s^2.

On s'efforce, dans la pratique, d'aboutir à des résidus gaussiens. En effet, dès que cette hypothèse est acceptée, on peut effectuer des tests et des estimations par intervalle de confiance sur les coefficients de régression, sur le coefficient de détermination et sur les prévisions (on trouvera des exemples numériques dans le paragraphe 4 de ce chapitre):

— Pour donner une estimation du coefficient de régression α par intervalle de confiance et tester l'égalité à une valeur spécifiée, on utilise la statistique $T = (A - a)/ S_A$, où S_A^2 est l'estimateur sans biais de la variance de A, qui suit la loi de Student de degré de liberté n-2.

— Le test de Fisher Snedecor peut être appliqué pour tester la nullité du coefficient de corrélation théorique: sous cette hypothèse, la statistique $F = (n-2)R^2/(1-R^2)$ suit la loi de Snedecor de degrés de liberté 1 et n-2.

— On peut donner des intervalles de confiance aux prévisions de la variable expliquée, en utilisant la statistique $(Y'-y)/S_{Y'}$ qui suit la loi de Student de degré de liberté n-2 si y est l'espérance de Y' et $S_{Y'}^2$ l'estimateur sans biais de sa variance.

Remarque: cas des données groupées.

Nous avons présenté dans les paragraphes précédents l'ajustement d'un nuage de points par une droite dans le cas de données individuelles. Tous les résultats que nous avons donnés peuvent être appliqués au cas de données groupées sous la forme d'un tableau de corrélation: il suffit d'introduire dans les formules un terme correspondant à l'effectif $n_{k,l}$ des valeurs observées dans la classe $C_k \times D_l$ et identifiées au couple (c_k, d_l) constitué des centres.

4. EXEMPLE NUMERIQUE.

Nous effectuons dans ce paragraphe la régression de la résistance pulmonaire (Répul) par l'index cardiaque (Incar) sur l'échantillon constitué des 101 malades observés. Nous n'appliquons ici que les définitions et résultats présentés précédemment.

Nous avons observé dans le paragraphe 1.1 que la liaison entre l'index cardiaque et la résistance pulmonaire présente un caractère exponentiel qui disparait si l'on considère le logarithme de la résistance pulmonaire au lieu de la variable initiale.

En outre, cette transformation atténue considérablement la particularité de l'unité statistique n° 100, que nous conserverons donc dans les données.

Notre première opération consiste donc à définir le modèle:

$$Y_i = \alpha\, x_i + \beta + \varepsilon$$

où Y est le logarithme népérien de la résistance pulmonaire. Pour simplifier, nous continuerons à la noter "Répul".

Les paramètres statistiques observés sur les données sont les suivants:

Variables	Moyennes	écarts-types	Variances
Incar	1.8457	.655747	.4300047
Répul	7.0499	.529936	.2808321

Le coefficient de corrélation entre la variable expliquée et la variable explicative est égal à -0.839, plus élevé en valeur absolue que lorsque l'on considère comme variable expliquée la résistance pulmonaire proprement dite (-0.767), ce qui justifie a posteriori le choix du logarithme.

L'analyse de variance nous donne la variance résiduelle estimée et le carré du coefficient de corrélation:

$$s^2 = 0.0846 \qquad r^2 = 0.7044 \qquad F(1\ ,99) = 235.942$$

La régression donne de bons résultats: le coefficient de corrélation est significativement non nul puisque la valeur observée de F appartient à la région critique [6.90, + ∝ [définie pour un risque de première espèce α égal à 0.01, et la variance résiduelle estimée est très inférieure à la variance de la variable expliquée.

Les coefficients de régression sont:

estimation	écart-type	t de Student	Interv de conf.
a = -0.67827	0.04416	-15.360	[-0.76482,-0.59172]
b = 8.30185	0.08649	95.982	[8.13233, 8.47137]

Etudions les résidus.

On peut rechercher les points aberrants parmi les données en examinant ce que l'on appelle la représentation linéaire des résidus (fig. 1.3).

Cette représentation montre que, sur les 101 résidus observés, seuls 4 sortent de l'intervalle ± 2 fois l'écart-type représenté par les deux demi-droites supérieure et inférieure: il s'agit des résidus e_{18} = -0.6581, e_{59} = -0.8847, e_{71} = -0.7044 et e_{100}= 0.6356.

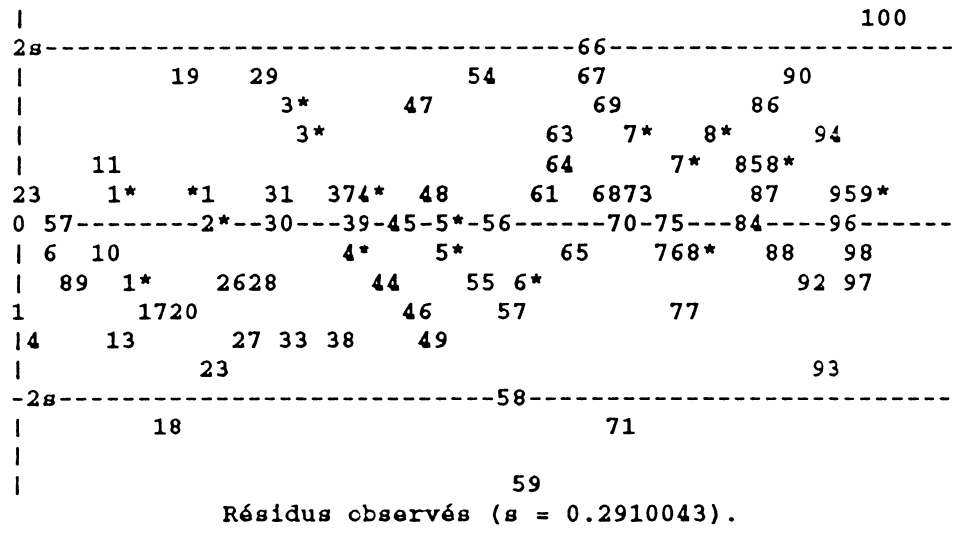

Fig. 1.3: Représentation linéaire des résidus
(Régression de Répul par Incar)

Dans le cas de la loi normale, il y a en moyenne 5% de valeurs à l'extérieur de l'intervalle ± 2 fois l'écart-type; la série des résidus observés ne présente donc rien d'extraordinaire. Il est toutefois intéressant d'examiner les unités statistiques correspondantes. On peut noter une grande différence entre les valeurs observées de certaines variables sur les unités statistiques en question et les moyennes calculées sur la totalité des observations (pour juger de ces différences, on la compare à l'écart-type; c'est pourquoi nous avons fait figurer simultanément la moyenne et l'écart-type des variables en-dessous des valeurs observées).

*	frcar	incar	insys	prdia	papul	pvent	répul	prono
18*	86	1.7	19.8	10	14	10.5	659	2
59*	75	2.32	30.9	8	10	6	345	2
71*	100	2.31	23.1	8	12	1	416	2
100*	116	0.60	5.2	33	38	10	5067	1
moy.	91.9	1.86	20.97	19.1	25.9	9.5	1286.6	
e-t	16.2	0.65	8.67	5.64	7.22	4.34	638.8	

C'est particulièrement vraie pour la pression artérielle pulmonaire (papul) et surtout la pression diastolique (prdia), pour laquelle les valeurs observées sortent de l'intervalle moyenne ± 2 x écart-type suivant le signe des résidus. On peut donc imaginer que la variable explicative Incar devrait être complétée par l'une de ces deux variables pour mieux reconstruire la résistance pulmonaire.

L'histogramme est donné en figure 1.4; les résidus paraissent un peu dissymétriques mais le test d'ajustement du χ^2 permet d'accepter l'hypothèse de normalité (l'écart-type étant le seul paramètre estimé de la loi normale ajustée, le degré de liberté est égal à 5):

Classes	Effectifs	Probabilité	condition (np_i)
1	12	0.0948	9.57
2	8	0.1393	14.07
3	24	0.2106	21.27
4	26	0.2280	23.02
5	14	0.1767	17.85
6	12	0.0980	9.90
7	5	0.0526	5.32

Résidus observés Test du Chi-2: 5.2587 Ddl: 5 Prob. cr.: 0.3851

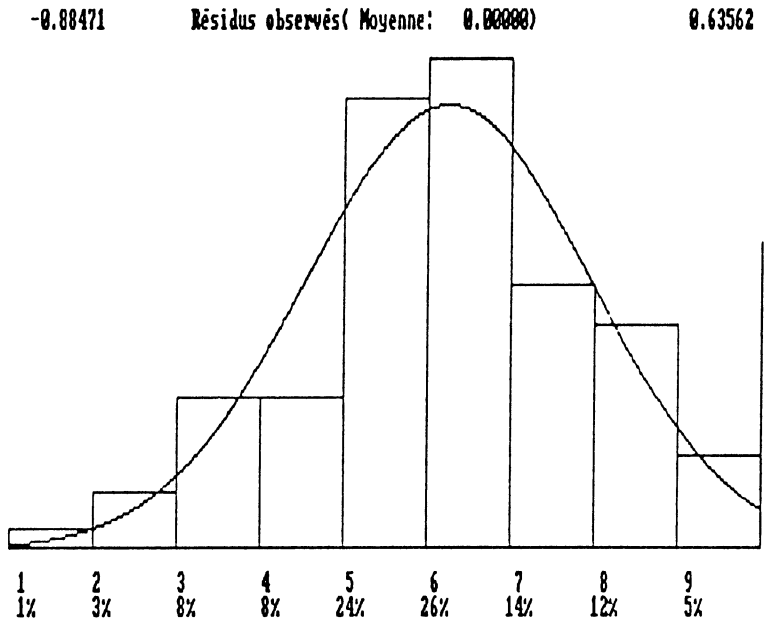

Fig. 1.4: Histogramme des résidus observés

On peut donc accepter l'hypothèse de normalité des résidus, et donc de la variable résiduelle (le nombre d'observations est suffisamment élevé pour que cette approximation soit justifiée). Les tests de Student et de Fisher peuvent donc être utilisés et montrent évidemment que l'on ne peut pas accepter l'hypothèse d'un coefficient de régression α nul ou d'un coefficient de corrélation ρ nul. Les intervalles de confiance sur α et β (p. 13) sont calculés pour un risque de première espèce égal à 0.05.

Chapitre 2

INTRODUCTION
A LA REGRESSION MULTILINEAIRE

1. ESTIMATION DES COEFFICIENTS DE REGRESSION.

1.1 Modèle multilinéaire.

Nous avons étudié dans le chapitre précédent le modèle linéaire simple: une variable expliquée, notée Y, et une variable explicative notée X. Le modèle multilinéaire consiste à introduire plusieurs variables explicatives que nous noterons X_1, X_2, ..., X_p. La distinction entre modèle linéaire et régression linéaire que nous avons précisée reste valable dans le cas multilinéaire: en régression, les variables X_j, j=1,...,p sont des variables aléatoires, dans le modèle multilinéaire, ce sont des variables contrôlées.

Le modèle de régression s'écrit donc:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

où ε est la variable résiduelle, indépendante des v.a. explicatives X_1, X_2, ..., X_p d'espérance nulle et de variance σ^2. Nous disposons d'un échantillon de (X_1, X_2,..., X_p, Y) de taille n.

En raisonnant conditionnellement aux valeurs observées x_i, on peut considérer l'espérance conditionnelle de Y:

$$E(Y / X_j=x_j \ \forall j=1,p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Le modèle multilinéaire est un système de n équations dont chacune correspond à une suite de valeurs des variables contrôlées $X_1,\ldots, X_p$:

$$Y(1) = \beta_0 + \beta_1 x_1(1) + \beta_2 x_2(1) +\ldots + \beta_p x_p(1) + \varepsilon(1)$$
$$Y(2) = \beta_0 + \beta_1 x_1(2) + \beta_2 x_2(2) +\ldots + \beta_p x_p(2) + \varepsilon(2)$$
$$\ldots\ldots\ldots$$
$$Y(i) = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) +\ldots + \beta_p x_p(i) + \varepsilon(i)$$
$$\ldots\ldots\ldots$$
$$Y(n) = \beta_0 + \beta_1 x_1(n) + \beta_2 x_2(n) +\ldots + \beta_p x_p(n) + \varepsilon(n)$$

Dans ce modèle nous faisons l'hypothèse que les v.a. $\varepsilon(i)$ constituent un échantillon indépendant d'une v.a. ε d'espérance nulle et de variance σ^2.

Dès l'instant que l'on raisonne en régression conditionnellement aux observations $x_j(i)$ des v.a. X_j, c'est-à-dire en les supposant fixées, le modèle de régression est équivalent au modèle linéaire: c'est ainsi que nous allons raisonner jusqu'à nouvel ordre. L'expression X=x signifiera "pour $X_j = x_j$, $\forall\, j = 1,\ldots,p$ ".

Il est fréquent d'utiliser la notation matricielle pour exprimer le modèle linéaire. Pour cela, on note:

— X la matrice à n lignes numérotées de 1 à n et p+1 colonnes numérotées j de 0 à p dont le terme général $x_j(i)$ est l'observation de la j^e variable explicative sur l'unité statistique de rang i. La ligne x(i) est définie par la suite $(x_j(i))$ j=0,p et la colonne x_j par la suite $(x_j(i),i=1,n$. La variable X_0 prend par définition la valeur 1 pour toute unité statistique i: le coefficient β_0 est alors le coefficient de régression de cette variable.

— Y le vecteur colonne à n lignes dont le terme y(i) est la variable expliquée sur l'unité statistique de rang i.

— ε le vecteur colonne à n lignes dont le terme $\varepsilon(i)$ est la variable résiduelle de rang i.

— β le vecteur colonne à p+1 lignes dont le terme général est le coefficient de régression β_j.

Le terme aléatoire est ε; la matrice X est connue et le vecteur β est certain mais inconnu. Le modèle linéaire s'écrit alors de la façon suivante:

$$Y = X\beta + \varepsilon$$

Il n'existe mathématiquement qu'une contrainte dans le choix des variables explicatives: elles ne doivent pas être linéairement liées, c'est-à-dire qu'il ne faut pas que l'une des variables se déduise des autres par combinaison linéaire. Nous verrons pourquoi dans le paragraphe suivant.

Rien n'empêche donc, a priori, de choisir comme variables explicatives des fonctions d'autres variables explicatives, pourvu que ce ne soit pas des fonctions linéaires. Par exemple, on peut introduire en X_1 une variable X, en X_2 son carré X^2, en X_3 son cube X^3 etc... Ce modèle est dit polynomial en X; il est souvent utilisé lorsque la variable X est le temps (Cf chapitre précédent). C'est encore un modèle linéaire par rapport aux coefficients β_j.

Nous discuterons ultérieurement du choix raisonné des variables explicatives, qui est un problème ardu de la régression.

1.2 Estimation des coefficients de régression.

Pour estimer les coefficients de régression, on procède comme dans le chapitre précédent; le critère des moindres carrés s'écrit:

$$S(b_0, b_1, b_2, \ldots, b_p) = \sum_{i=1}^{n} [y(i) - \sum_{j=0}^{p} b_j x_j(i)]^2$$

en posant $x_0(i)=1$ $\forall i=1,\ldots,n$.

Cette fonction admet un minimum au point où toutes les dérivées partielles sont nulles:

$$\forall j=0,\ldots,p \quad \partial S/\partial b_j = 0$$

On a:

$$\forall j=0,\ldots,p \qquad \partial S/\partial b_j = -2 \sum_{i=1}^{n} [(y(i) - \sum_{j=0}^{p} b_j\, x_j(i)]\, x_j(i) = 0$$

On retrouve le système des "équations normales", à p+1 inconnues et p+1 équations:

$$\begin{aligned}
&b_0 \sum_{i=1}^{n} x_0(i)^2 + b_1 \sum_{i=1}^{n} x_0(i)x_1(i) + \ldots + b_p \sum_{i=1}^{n} x_0(i)x_p(i) = \sum_{i=1}^{n} x_0(i)y(i)\\
&\qquad\qquad\qquad\qquad \ldots\ldots\ldots\\
&b_0 \sum_{i=1}^{n} x_0(i)x_1(i) + b_1 \sum_{i=1}^{n} x_1(i)^2 + \ldots + b_p \sum_{i=1}^{n} x_1(i)x_p(i) = \sum_{i=1}^{n} x_1(i)y(i)\\
&\qquad\qquad\qquad\qquad \ldots\ldots\ldots\\
&b_0 \sum_{i=1}^{n} x_0(i)x_p(i) + b_1 \sum_{i=1}^{n} x_0(i)x_p(i) + \ldots + b_p \sum_{i=1}^{n} x_p(i)^2 = \sum_{i=1}^{n} x_p(i)y(i)
\end{aligned}$$

La matrice M du système linéaire ci-dessus possède p+1 lignes et p+1 colonnes et a pour terme général $\Sigma\, x_k(i)\, x_l(i)$; elle est égale au produit matriciel $X^t X$, X^t étant la matrice transposée de X. Ce système se met sous la forme matricielle très simple suivante:

$$M B = X^t Y$$

Ces notations sont analogues aux notations utilisées dans le paragraphe précédent.

Pour calculer les coefficients de régression $b_0, b_1, \ldots, b_p$, il suffit donc de calculer la matrice inverse de M, si elle existe: on retrouve exactement la même démarche qu'en régression simple. Si la matrice M n'est pas inversible, cela signifie que les variables explicatives sont liées: l'une au moins peut être reconstruite exactement à l'aide des autres par une combinaison linéaire. Il est indispensable d'éliminer ces variables de l'ensemble des variables explicatives.

Notons en outre que les programmes ne détectent pas toujours les matrices non inversibles.

A partir de maintenant, nous supposons que la matrice M est inversible. En notant M^{-1} la matrice inverse, on a:

$$B = M^{-1} X^t Y$$

Les termes du vecteur B ainsi calculé sont les estimations des coefficients de régression β_j pour j=0,...,p. On en déduit:

— la valeur de la variable estimée en chaque point:

$$y_e(i) = \sum_{j=0}^{p} b_j \, x_j(i)$$

On note Y_e le vecteur colonne défini par la suite $(y_e(i))i=1,...n$.

— les résidus observés:

$$e(i) = y(i) - y_e(i).$$

Ces résidus ne sont pas les observations de la variable ε puisqu'ils sont calculés à l'aide des coefficients de régression estimés b_j. On note E le vecteur colonne défini par la suite (e(i))i=1,n.

Les notations précédentes permettent d'écrire les relations:

$$Y = X\beta + \varepsilon = X\,B + E$$

2. PROJECTION ORTHOGONALE ET VARIABLES REDUITES.

2.1 Opérateur de projection orthogonale.

Nous allons maintenant étudier l'application de $\mathbf{R}^n$ dans $\mathbf{R}^n$ qui à Y fait correspondre $Y_e = X\,B$ de façon à en proposer une interprétation géométrique.

Le vecteur Y, constitué des n observations y(i), appartient en effet à $\mathbf{R}^n$ et la variable estimée $Y_e = X\,B$, qui appartient aussi à $\mathbf{R}^n$, s'exprime de la façon suivante:

$$Y_e = X\,M^{-1}\,X^t\,Y$$

$$= X\,[X^t\,X\,]^{-1}\,X^t\,Y$$

Nous définissons ainsi l'application P qui à Y fait correspondre Y_e définie par la formule ci-dessus. Cherchons l'image Y_{ee} de Y_e par cette application:

$$Y_{ee} = X\,[X^t\,X\,]^{-1}\,X^t\,Y_e$$

$$= X\,[X^t\,X\,]^{-1}\,X^t\,X\,[X^t\,X\,]^{-1}\,X^t\,Y$$

$$= X\,[X^t\,X\,]^{-1}\,X^t\,Y$$

$$= Y_e$$

Nous venons de montrer une propriété caractéristique d'un projecteur, qui est l'idempotence:

$$P\,[\,P(Y)\,] = P(Y) = Y_e$$

Calculons le produit scalaire $[Y - Y_e]\,.\,Y_e$ défini dans $\mathbf{R^n}$ comme la somme des produits des termes de même rang; ce produit scalaire s'exprime à l'aide des produits matriciels ci-dessous:

$$[Y - Y_e]^t\,Y_e = Y^t\,Y_e - Y_e{}^t\,Y_e$$

$$= Y^t\,Y_e - [X\,M^{-1}\,X^t\,Y]^t\,[\,X\,M^{-1}\,X^t\,Y]$$

$$= Y^t\,Y_e - Y^t\,X\,[M^{-1}]^t\,X^t\,X\,M^{-1}\,X^t\,Y$$

$$= Y^t\,Y_e - Y^t\,X\,[M^{-1}]^t\,X^t\,Y$$

$$= Y^t\,Y_e - Y^t\,Y_e \qquad ([M^{-1}]^t = M^{-1})$$

$$= 0$$

Cette propriété caractérise les opérateurs de projection orthogonale définis sur l'espace euclidien $\mathbf{R^n}$.

La représentation géométrique (fig. 2.1) montre l'orthogonalité du sous-espace F engendré par les variables X_j et de $Y\text{-}Y_e$; cette dernière n'est autre que la variable E définie par les résidus e(i). On généralise ainsi la propriété que nous avons donnée sans démonstration dans le chapitre 1 à l'ensemble des variables explicatives:

Théorème: la série des résidus (e(i))i=1,n est centrée et non-corrélée aux variables explicatives X_j pour j=1 à p.

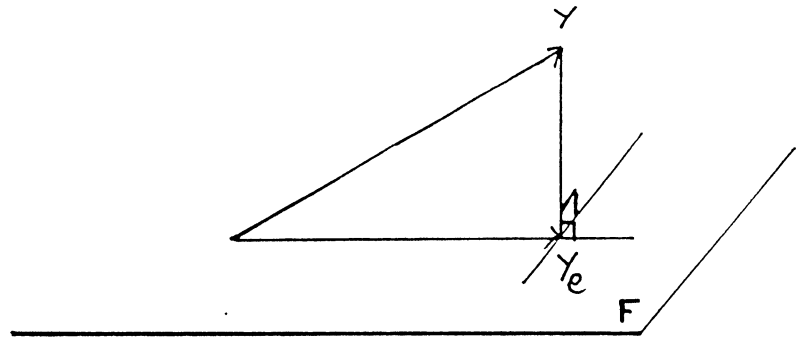

Fig. 2.1: Interprétation géométrique
de la régression multilinéaire

2.2 Variables centrées réduites (régression multilinéaire).

Pour simplifier les notations nous avons introduit la variable x_0 constante et égale à 1. On peut réécrire les équations normales en tenant compte de cette propriété:

La première équation s'écrit :

$$n\, b_0 + b_1 \sum_{i=1}^{n} x_1(i) + \ldots + b_p \sum_{i=1}^{n} x_p(i) = \sum_{i=1}^{n} y(i)$$

Soit, en divisant par n:

$$b_0 + b_1 \bar{x}_1 + \ldots + b_p \bar{x}_p = \bar{y}$$

On peut diviser les autres équations normales par n puis leur soustraire l'équation ci-dessus: on obtient alors:

$$\forall j=1,\ldots,p \quad b_1\, s(x_j)^2 + \ldots + b_p\, cov\,(x_j, x_p) = cov\,(x_j, y)$$

où $s(x_j)^2$ et $cov(x_j, x_k)$ sont les variances et covariances des séries statistiques $x_j(i)$ et $x_k(i)$, pour i=1,n.

Le système des équations normales peut donc être exprimé à l'aide de la matrice de covariances de $(X_1, X_2, ..., X_p)$, le terme constant b_0 étant déduit des autres coefficients a_j par la relation:

$$b_0 = \bar{y} - (b_1\bar{x}_1 + ... + b_p \bar{x}_p)$$

Cette propriété du coefficient constant permet de supposer que les variables mises en jeu dans le modèle sont centrées. La matrice $X^t X$, de dimension p puisqu'il n'y a plus de terme constant, est égale alors à n fois la matrice des covariances entre les variables explicatives.

En outre, il est souvent commode, dans la pratique, d'étudier les coefficients de régression calculés sur les variables centrées réduites. La relation entre ces derniers et les coefficients de régression sur les variables initiales est immédiate; le modèle de régression est, sur les variables initiales:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \varepsilon$$

En centrant les variables, on fait disparaître la constante. Notons Y' et X_j' les variables centrées réduites, σ_j^2 les variances des variables explicatives X_j et σ_y^2 la variance de la variable expliquée; il vient:

$$Y' = \beta_1 \sigma_1 X_1' / \sigma_Y + \beta_2 X_2' \sigma_2 / \sigma_Y + ... + \beta_p X_p' \sigma_p / \sigma_Y + \varepsilon / \sigma_Y$$

Soit:

$$Y' = \beta_1' X_1' + \beta_2' X_2' + \beta_3' X_3' + ... + \beta_p' X_p' + \varepsilon'$$

La variance de la variable résiduelle est divisée par σ_Y^2 et les coefficients de régression β_j' se déduisent des coefficients de régression β_j par la formule:

$$\beta_j' = \beta_j \sigma_j / \sigma_Y$$

Les coefficients de régression β'_j sont relatifs à des variables centrées réduites: on peut donc interpréter directement leur taille et les comparer entre eux.

Toutes ces propriétés sont évidemment vraies pour les estimations.

Dans les représentations graphiques, nous serons amenés à supposer que les variables sont centrées et réduites. La figure 2.2 donne ainsi l'interprétation géométrique dans $\mathbf{R^n}$ de la régression d'une variable centrée Y par deux variables centrées X_1 et X_2; la variable constante X_0 égale à 1 ne figure pas sur ce schéma puisqu'elle appartient à l'orthogonal du sous-espace qui est représenté et qui est engendré par les variables X_1 et X_2 considérées.

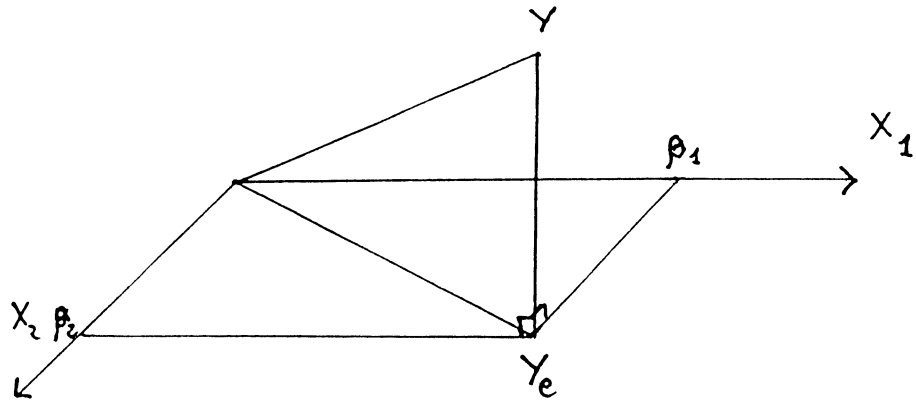

Fig. 2.2: Régression de Y par X_1 et X_2
(variables centrées)

3. PROPRIETES DES ESTIMATEURS ET DES RESIDUS

3.1 Propriétés des estimateurs des coefficients de régression.

Nous avons calculé dans le paragraphe précédent un vecteur $b = (b_0, b_1, \ldots, b_j, \ldots, b_p)$ qui minimise la somme des carrés des erreurs. Ce vecteur b est une estimation du vecteur de régression $\beta = (\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_p)$ dont l'existence est supposée par le modèle linéaire.

Théorème: le vecteur B est l'estimateur linéaire efficace du vecteur de régression β et sa matrice de covariance est donnée par:

$$V_B = \sigma^2 M^{-1}$$

Un estimateur efficace est un estimateur de variance minimale dans la classe des estimateurs sans biais.

Nous admettrons ce théorème dont la démonstration est donnée dans tous les ouvrages classiques (Saporta, 1990 par exemple).

L'estimateur de V_B est obtenu en remplaçant la variance résiduelle σ^2 par son estimateur sans biais s^2 (Cf paragr. suivant).

Conséquence: la v.a. B_j est un estimateur sans biais du coefficient de régression β_j et sa variance σ_j^2 est égale au terme diagonal de la matrice V_B.

Remarques:

— les estimateurs des coefficients de régression ne sont donc pas indépendants; nous verrons ultérieurement qu'ils peuvent être fortement corrélés, comme nous l'avons vu dans le chapitre précédent à propos du coefficient directeur de la droite de régression et de son terme constant;

— on montre en outre que, si la variable résiduelle ε est gaussienne, B est l'estimateur du maximum de vraisemblance et est lui-même gaussien; les estimateurs B_j de chaque coefficient de régression β_j sont alors gaussiens et on peut en donner des intervalles de confiance ou effectuer des tests de Student: pour tester l'égalité d'un coefficient de régression β_j à une valeur spécifiée β_j^0, on utilise la statistique T_j définie par:

$$T_j = (B_j - \beta_j^0) / S_j$$

où S_j est l'estimateur sans biais de l'écart-type de B_j, qui suit la loi de Student à de degré de liberté n-2 si la variable résiduelle est gaussienne (paragr. 1.2);

— les estimateurs B_j et B_k n'étant pas indépendants, tester l'égalité du coefficient β_j à β_j^0 puis de β_k à β_k^0 n'est pas équivalent à tester l'égalité du couple (β_j, β_k) à (β_j^0, β_k^0).

3.2 Etude des résidus. Coefficient de corrélation multiple. Prévision.

L'étude de la variable résiduelle rencontre les mêmes difficultés en régression multilinéaire qu'en régression linéaire: les coefficients de régression β_j n'étant pas connus, la variable ε n'est pas observable.

On ne dispose que des résidus e(i) en chaque point; nous avons vu que ces résidus sont centrés et non corrélés aux variables explicatives.

Soit $s^2(e)$ la variance observée des résidus:

$$s^2(e) = 1/n \sum_{i=1}^{n} [y(i) - \sum_{j=0}^{p} b_j x_j(i)]^2 = 1/n \sum_{i=1}^{n} e^2(i)$$

Cette variance nous donne une estimation sans biais $s^2 = n\, s^2(e)/ (n-p-1)$ de la variance résiduelle. C'est la réalisation de la variable $S^2 = SCR/(n - p -1)$, où SCR est la somme des carrés des résidus:

$$SCR = \sum_{i=1}^{n} [\, Y(i) - Y_e(i) \,]^2$$

$Y_e(i)$ est ici l'estimateur ci-dessous:

$$Y_e(i) = \sum_{j=0}^{p} B_j x_j(i)$$

Théorème: la statistique SCR/σ^2 suit la loi du χ^2 de degré de liberté n-p-1. Elle est non corrélée aux v.a. B_j et aux v.a. $Y_e(i)$ (indépendante si la variable résiduelle est gaussienne).

Le degré de liberté s'explique par l'orthogonalité aux p+1 variables explicatives.

Ce théorème nous permet d'estimer la variance résiduelle par intervalle de confiance.

Définition: on appelle coefficient de corrélation multiple le coefficient de corrélation entre la variable expliquée Y et la variable estimée Y_e. Le coefficient de détermination est le carré du coefficient de corrélation multiple. Ces coefficients sont notés R et R^2.

Théorème: le coefficient de détermination vérifie l'égalité ci-dessous:

$$s^2(e) = (1 - R^2)\ s^2(y)$$

Théorème: si tous les coefficients de régression β_j, j=1,...,p sont nuls, la statistique:

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

suit la loi de Snedecor de degrés de liberté p et n - p - 1.

On rejettera donc l'hypothèse que tous les coefficients de régression (sauf le terme constant) sont nuls, ou, ce qui revient au même, que la valeur théorique de R^2 est égale à 0, pour les valeurs de F appartenant à la région critique $[F_\alpha, +\infty[$, où F_α dépend du risque de première espèce α et des degrés de liberté.

Nous admettons bien sûr toutes les propriétés précédentes dont des exemples sont donnés dans le paragraphe 2.

La figure 2.3.1 donne la représentation des variables non centrées, la figure 2.3.2 celle des variables centrées. La seconde peut être considérée comme la projection orthogonale de la première sur le sous-espace de $\mathbf{R}^n$ orthogonal au vecteur constant égal à 1.

La figure 2.3.2 permet d'interpréter géométriquement la variance des résidus et le coefficient de corrélation multiple:

— le coefficient de corrélation multiple est le cosinus de l'angle θ formé par les vecteurs Y et Y_e;

— la variance des résidus est le carré de la norme du vecteur Y - Y_e divisé par n;

— les coefficients de régression b_1 et b_2 sont les coordonnées de Y_e sur les axes représentant les variables explicatives.

— le terme constant, qui est le coefficient de régression sur la variable X_0, est invisible sur la figure 2.3.2. Par contre on peut le voir sur la figure 2.3.1.

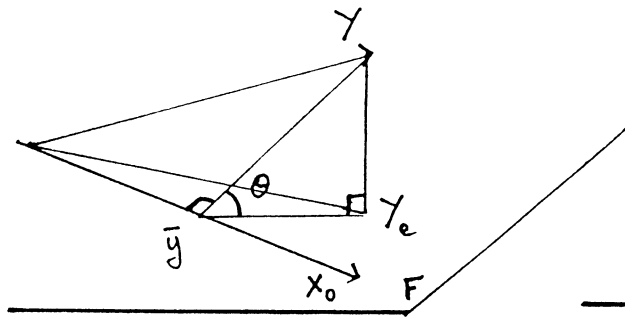

Fig. 2.3.1: Variance des résidus et corrélation multiple (variables non centrées)

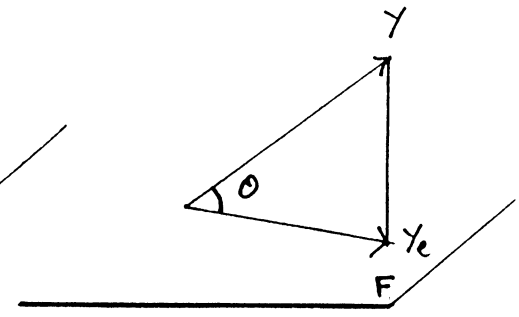

Fig. 2.3.2: Variance des résidus et corrélation multiple (variables centrées)

Pour prédire la variable expliquée en fonction des variables explicatives, on procède comme en régression simple: on peut prédire soit l'espérance conditionnelle, soit la valeur d'un point, en utilisant les observations des variables explicatives. La prédiction ponctuelle est la même, seule change la variance de l'estimateur. Pour estimer E(Y/X=x), on utilise les estimateurs des coefficients de régression:

Théorème: la statistique Y' définie par:

$$Y' = \sum_{j=1}^{p} B_j \, x_j$$

est un estimateur efficace de l'espérance conditionnelle E(Y/X=x):

$$E(Y/X=x) = \sum_{j=1}^{p} \beta_j \, x_j$$

et sa variance est égale à:

$$V(Y') = \sigma^2 \, [x]^t \, M^{-1} \, [x]$$

où $[x]^t$ est le vecteur ligne $[x_0,....,x_j....x_p]$.

La variance de la prévision d'une observation est obtenue en ajoutant simplement σ^2 à la variance précédente.

Pour obtenir des estimateurs, il suffit de remplacer la variance résiduelle σ^2 par son estimateur sans biais s^2.

Pour obtenir une estimation par intervalle de confiance, on utilise la statistique $(Y' - y)/ S_{Y'}$, où $S_{Y'}^2$ est l'estimateur sans biais de la variance V(Y'). Cette statistique suit en effet le loi de Student de degré de liberté n - p-1.

4. EXEMPLE NUMERIQUE.

Nous avons suggéré dans le chapitre 2 de compléter l'index cardiaque par la pression artérielle pulmonaire ou la pression diastolique pour mieux expliquer la résistance pulmonaire des malades étudiés. Nous donnons ci-dessous les résultats numériques obtenus en introduisant la pression diastolique.

Pourquoi avoir choisi la pression diastolique ? Pourquoi pas les deux variables envisagées? En examinant leur coefficient de corrélation (0.928), on peut se dire qu'introduire les deux variables est inutile car les informations qu'elles apportent sur les données sont presque les mêmes. Nous verrons ulté-

rieurement comment en juger de manière satisfaisante (par l'étude des coefficients de corrélation partielle).

Le modèle de régression que nous considérons est donc le suivant:

$$\text{Répul} = \beta_0 + \beta_1 \text{ Incar} + \beta_2 \text{ Prdia} + \varepsilon$$

Les paramètres statistiques des variables explicatives considérées sont les suivants:

VARIABLES	MOYENNES	ECARTS-TYPES	VARIANCES
Incar	1.846	.65575	.43000
Prdia	19.259	5.78051	33.41429
Répul	7.049929	.529936	.28083

Les variances des variables explicatives sont très différentes l'une de l'autre; pour interpréter les coefficients de régression, il sera utile d'examiner leurs estimations sur les variables réduites.

MATRICE DES CORRELATIONS ENTRE LES VARIABLES

	Incar	Prdia	Répul
Incar	1.000		
Prdia	-0.361	1.000	
Répul	-0.839	0.761	1.000

La matrice ci-dessus fait apparaître de fortes corrélations en valeur absolue entre la variable expliquée et les variables explicatives. Le coefficient de détermination (supérieur aux carrés des coefficients de corrélation), sera élevé; le coefficient de corrélation entre les variables explicatives est relativement faible; il semble donc qu'il y ait complémentarité entre ces deux variables explicatives (ce raisonnement n'est valable que pour deux variables).

Dans le tableau ci-dessous, on peut lire que le coefficient de détermination est égal à 0.946; la valeur du F que l'on en déduit est égale à 864.502, sa

probabilité critique P(F > 864.502) est numériquement nulle: cela montre qu'elle appartient à la région critique quel que soit le risque de première espèce choisi: l'hypothèse de nullité des coefficients de régression est rejetée.

La variance résiduelle estimée est ici égale à 0.0155, nettement plus faible que dans la régression à l'aide de la seule variable Incar (0.0846); l'écart-type résiduel estimé est égal à 0.125 contre 0.291 précédemment. L'introduction de la pression diastolique améliore donc considérablement la reconstruction de la résistance pulmonaire.

ANALYSE DE VARIANCE

	ddl	Somme des Carrés	Variance Estimée	Pourcentage de var.tot
Tot	100	283.6404D-01	283.6404D-03	1
Exp	2	268.4260D-01	268.1155D-03	946.3603D-03
Res	98	152.1440D-02	155.2490D-04	0.0536

Corrélation multiple 0.9728 Détermination 0.9464
F(2 , 98) 864.502 Probabilité critique 0.0000

Comme nous l'avons suggéré précédemment, nous examinons tout d'abord les coefficients de régression calculés sur les variables centrées réduites: ils sont de taille relativement proche en valeur absolue l'un de l'autre, ce qui signifie qu'aucune des variables n'a une importance prédominante dans la régression. Ils sont du signe du coefficient de corrélation correspondant, ce qui paraît naturel (cette propriété n'est pas toujours vérifiée).

COEFFICIENTS DE REGRESSION

N°	Estimation	Ecart-type	Estimation (var. réd.)	T de Student
Incar	-0.52458	0.020271	-0.6491	-25.878
Prdia	0.04835	0.002300	0.5274	21.024
Cst	7.08705	0.068631	0	103.264

Le t de Student est très élevé, supérieur à 20 en valeur absolue et a donc une probabilité critique numériquement nulle (P(|T|>21.024)=0). Nous avons une propriété supplémentaire ici par rapport au test du F sur le coefficient de

détermination: on rejette l'hypothèse de nullité de chaque coefficient.

Les intervalles de confiance des coefficients de régression sont définis comme l'ensemble des valeurs β_j tels que la valeur observée t de la statistique $T = (B_j - \beta_j)/s_j$ soit comprise entre -1.96 et 1.96 pour un risque de première espèce $\alpha = 0.05$ (la loi de Student de degré de liberté 98 est confondue avec la loi normale centrée réduite). On en déduit:

$$b_0 \in [\ 6.9525\ ,\ 7.2216\ ]$$
$$b_1 \in [-0.5643\ ,-0.4849\ ]$$
$$b_2 \in [\ 0.04384,\ 0.05286\ ]$$

Examinons maintenant les résidus obtenus: souvenons-nous que c'est par l'étude des résidus que nous avons été amenés à introduire la variable pression diastolique en seconde variable explicative: la résistance pulmonaire des unités statistiques 18, 59, 71 et 100 était particulièrement mal reconstruite par l'index cardiaque.

Les résidus concernant ces malades sont égaux, dans l'ordre précédent, à -0.188, -0.413 et -0.231 et 0.163 au lieu de -0.658, -0.885 et -0.704 et 0.636. On retrouve ici l'amélioration constatée sur un plan général quand on compare les variances résiduelles estimées. Seul le second (n° 59) reste anormal par rapport à l'écart-type résiduel estimé (0.125) puisqu'il est supérieur à 3 x 0.125 en valeur absolue: en examinant les données (p. 39) on remarque une forte valeur de l'index systolique sur l'unité statistique 59, contrairement aux autres unités 18 et 71: c'est peut-être l'explication de la particularité de cette observation.

La figure 2.4 et le test d'ajustement du χ^2 confirment la normalité des résidus.

La figure 2.5 donne la représentation linéaire des résidus. On y retrouve le résidu n° 59, mais trois résidus supérieurs à 2 fois l'écart-type apparaissent: il s'agit des n° 19, 63, 67. ils restent dans des limites acceptables (inférieurs à 3 fois l'écart-type).

Le nombre de valeurs à l'extérieur de l'intervalle $\pm$ 2 x s n'est pas contradictoire avec l'hypothèse que la variable résiduelle suive la loi normale; cette

hypothèse n'est donc pas contredite par les observations effectuées; les tests de Student, de Fisher et les intervalles de confiance sont justifiés.

Régression de la résistance pulmonaire
par l'index cardiaque et la pression diastolique

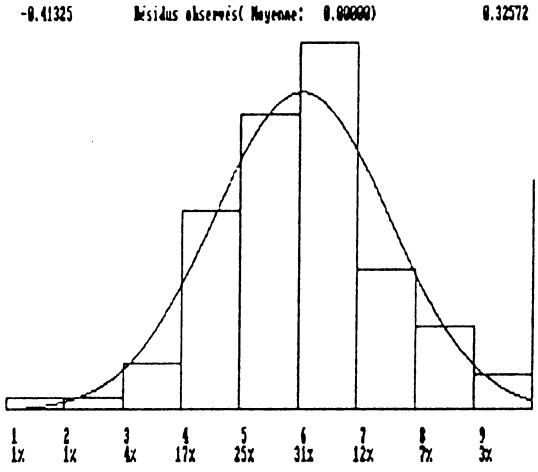

Fig 2.4 Histogramme des résidus observés.
Ajustement par la loi normale.

TEST DU χ^2

Cl.	Eff.	Probab.	np_i
1	6	0.0869	8.78
2	17	0.1579	15.94
3	25	0.2464	24.89
4	31	0.2500	25.25
5	12	0.1647	16.64
6	7	0.0705	7.12
7	3	0.0236	2.38

Résidus observés
Test du Chi-2: 3.7168 Ddl: 5
Prob. cr.: 0.5909

```
|                                    6367
|          19
2s-------------------------------------------------------
|                                                  89
|     11    2*          41             6*          88     100
|      12         29          47              72   81   91  97
|       15         30     42        56   65 70       84       99
| 6   1*1*              39 45      5* 64        **82 86
0 58----17-22--------**---4*5*---------69-7*-------9*93-------
|3  9           27  3*    *3  5155  61                    98
2           20 26 3*35               62       7579       9*
|             23 28                58              8387   95 101
14 7                          49             73      85
|         18       32            54
-2s----------25-----------------------71-------------------
|
|
|
|                                    59
```

Fig. 2.5: Représentation linéaire des résidus observés (s=0.124599)

Chapitre 3

CORRELATIONS PARTIELLES
REGRESSIONS PAS A PAS

1. NOTION DE CORRELATION PARTIELLE.

Dans l'exemple numérique étudié, nous avons constaté que l'introduction de la pression diastolique pour compléter l'index cardiaque améliore considérablement la régression: la variance résiduelle estimée passe de 0.0812 à 0.0154 et le coefficient de détermination augmente de 0.7044 à 0.9464. La variable X = Prdia apporte donc une information sur la variable expliquée Y = Répul étrangère à celle qui est donnée par la première variable explicative X_1 = Incar. Cette information supplémentaire se mesure par le coefficient de corrélation partielle.

1.1 Coefficient de corrélation partielle.

Définition: on appelle coefficient de corrélation partielle de X et Y conditionnellement à X_1 le coefficient de corrélation $R(X,Y/X_1)$ entre $Z_1 = Y - (\beta_1 X_1 + \beta_0)$, et $Z_2 = X - (\alpha_1 X_1 + \alpha_0)$, où β_1, β_0 et α_1, α_0 sont les coefficients des régressions de Y et de X par X_1.

Les variables Z_1 et Z_2 sont toutes deux non corrélées à X_1 puisqu'elles sont définies par les variables résiduelles des régressions de Y et X par X_1. L'estimateur de ce coefficient $R(X,Y/X_1)$ est l'estimateur empirique, calculé sur

les valeurs observées, que nous appellerons aussi coefficient de corrélation partielle et que nous noterons de la même façon.

L'interprétation géométrique est donnée par la figure 3.1, dans laquelle nous avons supposé que les variables sont centrées et réduites.

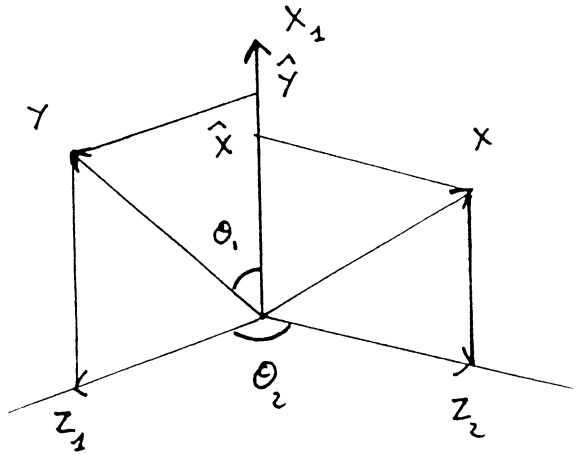

Fig. 3.1: Coefficient de corrélation partielle
de Y et X conditionnellement à X_1

Ce coefficient de corrélation partielle se généralise facilement à des groupes de variables: si X_1, X_2, ..., X_q sont les variables explicatives considérées, le coefficient de corrélation partielle de X et de Y conditionnellement à X_1, X_2, ..., X_q est le coefficient de corrélation des résidus obtenus par la régression de Y par les variables X_j et des résidus obtenus par la régression de X par les variables X_j. On le note $R(Y,X/X_1, X_2,..., X_q) = R(Y,X/X_{.q})$

Théorème: si la variable résiduelle ε est gaussienne et si le coefficient de corrélation partielle théorique est nul, la statistique $F_{1,n-q-2}$ définie par:

$$F_{1,n-q-2} = (n - q - 2) \frac{R(Y,X/X_{.q})^2}{1 - R(Y,X/X_{.q})^2}$$

suit la loi de Snedecor de degrés de liberté 1 , n - q - 2.

Dans notre exemple, ce coefficient de corrélation partielle est estimé à R(Prdia, Répul/Incar) = 0.905 et $F_{1,98}$ = 418.41. Une table nous donne comme région critique du test sur $F_{1,98}$ [6.90, + ∝ [pour un risque de première espèce égal à 0.01 et des degrés de liberté égaux à 1 et 98: il est clair que l'on rejette l'hypothèse de nullité du coefficient de corrélation partielle théorique et que la diminution de la variance résiduelle due à l'introduction de la variable Prdia comme variable explicative ne peut être un effet du hasard.

L'étude des autres coefficients de corrélation partielle montre que l'on aurait pu introduire à la place de la pression diastolique la variable pression artérielle pulmonaire (Papul) dont le coefficient de corrélation partielle est 0.936, supérieur au précédent (0.905):

```
Corrélations partielles avec Répul

Variable explicative considérée: Incar
R2 = 0.70443        F(1,99) = 235.9418        Prob. crit. = 0.0000

         Frcar  Insys  Prdia  Papul  Pvent
Répul    0.357 -0.354  0.905  0.936  0.156

Coef. de corr. 0.3540:
           Valeur du F  14.04 Probabilité critique  0.0004

Coef. de corr. 0.1560:
           Valeur du F   2.44 Probabilité critique  0.1170
```

Seul le coefficient de corrélation partielle de la résistance pulmonaire (Répul) et de la pression ventriculaire (Pvent) n'est pas significatif (r = 0.156, F = 2.44 P(F > 2.44) = 0.117) et ne témoigne pas d'une liaison non aléatoire entre les deux variables compte tenu de l'information apportée par l'index cardiaque.

Nous retrouvons ici ce que nous avions pressenti et expliqué dans le chapitre 2: l'information apportée sur la résistance pulmonaire par l'index cardiaque peut être complétée par la pression diastolique ou la pression articulaire pulmonaire.

A l'aide de la figure 3.1, on montre que le carré du coefficient de corrélation partielle mesure la diminution relative de la variance résiduelle observée. Plus précisément, on montre que:

$$R^2(Y,X/X_1) = \frac{s_q(e)^2 - s_{q+1}(e)^2}{s_q(e)^2}$$

où $s_q(e)^2$ et $s_{q-1}(e)^2$ sont les variances empiriques des résidus dans les régressions de Y par $X_1, ..., X_q$ et par $X_1, ..., X_q$, X respectivement.

L'introduction d'une variable parmi l'ensemble des variables explicatives a donc pour effet de diminuer la variance des résidus, ou, ce qui revient au même, d'augmenter le coefficient de détermination. Cette propriété n'est pas nécessairement vérifiée par l'estimateur sans biais de la variance résiduelle qui dépend du nombre de variables explicatives.

En conclusion, l'examen des coefficients de corrélation partielle nous permet donc de mieux choisir les variables explicatives supplémentaires en évitant les redondances d'information et en mettant en évidence la complémentarité des variables: l'augmentation du coefficient de détermination est d'autant plus importante que le coefficient de corrélation partielle de la variable introduite dans l'ensemble des variables explicatives est élevé.

Ainsi, suivant que l'on ajoute la pression articulaire pulmonaire ou la pression diastolique, le coefficient de détermination atteint 0.94636 ou 0.96356.

1.2 Coefficients de détermination partielle.

La formule précédente permet de généraliser la notion de coefficient de corrélation partielle à un ensemble de variables $Z_1, Z_2, ..., Z_r$, en mesurant la diminution de la variance des résidus quand on introduit les Z_j comme variables explicatives de la régression: la statistique $R^2(Y,Z_1,Z_2,...,Z_r/X_1,...,X_q)^2$, que l'on note $R^2(Y,Z_{.r}/X_{.q})^2$, analogue à un coefficient de détermination est appelée coefficient de détermination partielle:

$$R^2(Y,Z_{.r}/X_{.q}) = \frac{s_q(e)^2 - s_{q+r}(e)^2}{s_q(e)^2}$$

où $s_{q+r}(e)^2$ est la variance des résidus dans la régression de Y par les variables X_j, j = 1,q et Z_k, k = 1,r.

Théorème: si la variable résiduelle ε est gaussienne et si le coefficient de détermination partielle théorique est nul, la statistique $F_{r,n-q-r-1}$ définie par:

$$F_{r,n-q-r-1} = \frac{(n - q - r - 1)}{r} \frac{R(Y,Z_{.r}/X_{.q})^2}{1 - R(Y,Z_{.r}/X_{.q})^2}$$

suit la loi de Snedecor de degrés de liberté r , n - q - r - 1.

Par exemple, en ajoutant simultanément la pression diastolique et la pression artérielle pulmonaire à l'index cardiaque, les sommes des carrés des résidus passent de 8.3837 à 0.94642, les statistiques $R^2(Y,Z_{.r}/X_{.q})$ et $F_{r,q}$ prennent les valeurs 0.88711 et 179.16. La probabilité critique $P(F_{r,q} > 179.16)$ est numériquement nulle: l'information apportée par les deux variables est hautement significative. Cette conclusion n'est d'ailleurs pas étonnante puisque l'on a déjà refusé l'hypothèse de nullité du coefficient de régression de la pression diastolique.

2. ANALYSE DES CORRELATIONS PARTIELLES

2.1 Choix raisonné des prédicteurs.

Le choix raisonné des prédicteurs est effectué par l'utilisateur, en fonction des données qu'il étudie et de ses objectifs.

L'avantage, par rapport aux procédures automatiques que nous présentons dans le paragraphe suivant, est de toujours contrôler le système en cours de constitution et de pouvoir tenir compte simultanément de plusieurs critères, en particulier de la colinéarité entre les variables explicatives (il est préférable en effet que le coefficient de détermination de chaque variable explicative dans la régression par les autres soit faible) et de la variance résiduelle estimée.

La démarche consiste à étudier, après chaque introduction d'un prédicteur, les coefficients de corrélation partielle des variables restantes avec la variable expliquée (pour augmenter le coefficient de détermination de la

régression), la variance résiduelle (pour diminuer la variance des estimations), les coefficients de détermination entre les variables explicatives et le système des prédicteurs déjà introduits (pour éviter les colinéarités).

Nous avons appliqué cette démarche aux données sur l'infarctus:

— on calcule les coefficients de corrélation des variables explicatives avec la variable expliquée: le premier prédicteur est celui pour lequel ce coefficient est, en valeur absolue, le plus élevé en étant significativement non nul bien que cette restriction ne soit pas toujours justifiée comme nous l'indiquons en 2.2 .

CORRELATIONS ENTRE LES VARIABLES

	Frcar	Incar	Insys	Prdia	Papul	Pvent
Répul	0.287	-0.839	-0.833	0.761	0.716	0.318

Nous introduisons ainsi pour expliquer la résistance pulmonaire l'index cardiaque Incar dont le coefficient de corrélation (-0.839) est très largement significatif. Cette décision est uniquement basée sur les résultats numériques: en tenant compte de considérations médicales ou autres, on aurait pu choisir l'index systolique Insys dont le coefficient de corrélation avec la résistance pulmonaire est presque égal au précédent (-0.833).

— on calcule les coefficients de corrélation partielle des variables explicatives restantes et de la variable expliquée conditionnellement au prédicteur précédemment introduit, et leur coefficient de détermination avec le système de prédicteurs.

Les résultats numériques ci-dessous font apparaître deux coefficients de corrélation partielle hautement significatifs (leur probabilité critique est numériquement nulle): les variables concernées sont la pression diastolique et la pression artérielle pulmonaire (Prdia et Papul).

Les coefficients de détermination avec le système de prédicteurs considéré (limité ici à l'index cardiaque) sont tous deux voisins de 0: il est à peu près équi-

valent, au plan statistique, de choisir l'une ou l'autre de ces variables: on peut choisir par exemple celle qui est la plus facile ou la moins chère à observer.

```
                Variables explicatives considérées:
Incar
R2 = 0.70443          F(1,99) = 235.9418          Prob. crit. = 0.0000
Variance résiduelle estimée = 8.4683D-02
-------------------------------------------------------------------
Coefficients de détermination de chaque var. par rapport aux var.
explicatives:
Frcar:1 (R2 = 0.013)|Insys:3 (R2 = 0.787)|Prdia:1 (R2 = 0.013)|
Papul:5 (R2 = 0.073)|Pvent:6 (R2 = 0.080)|
-------------------------------------------------------------------
                      Corrélations partielles
          Frcar  Insys  Prdia  Papul  Pvent
Répul     0.357 -0.354  0.905  0.936  0.156

Coef. de corr. 0.9050 Valeur du F 443.51 Prob. critique  0.0000
```

Nous avons choisi la pression diastolique Prdia dont le coefficient de détermination avec l'index cardiaque est le plus faible.

— on recalcule les coefficients de corrélation partielle et les coefficients de détermination.

```
                Variables explicatives considérées:
Incar  Prdia
R2 = 0.94626          F(2,98) = 864.5018          Prob. crit. = 0.0000
Variance résiduelle estimée = 1.5525D-02
-------------------------------------------------------------------
Coefficients de détermination de chaque var. par rapport aux var.
explicatives:
Frcar:1 (R2 = 0.160)|Insys:3 (R2 = 0.817)|Papul:5 (R2= 0.866)|
Pvent:6 (R2 = 0.118)|
-------------------------------------------------------------------
                      Corrélations partielles
          Frcar  Insys  Papul  Pvent
Répul     0.019 -0.029  0.615 -0.069

Coef. de corr. 0.6150 Valeur du F  59.00 Prob. critique  0.0000
```

Le coefficient de détermination de la pression artérielle avec les prédicteurs déjà introduits est élevé (0.866); on montre que cela peut augmenter

la variance des estimateurs des coefficients de régression, ce qui serait gênant, mais on diminuera (peut-être) la variance résiduelle estimée. Le coefficient de corrélation partielle, largement significatif, ne laisse guère le choix.

— les coefficients de corrélation partielle après introduction de la pression artérielle pulmonaire sont les suivants:

```
                Variables explicatives considérées:
Incar  Prdia  Papul
R2 = 0.96663        F(3,97) = 936.69          Prob. crit. = 0.0000
Variance résiduelle estimée = 9.7569D-03
------------------------------------------------------------------
Coefficients de détermination de chaque var. par rapport aux var.
explicatives:
Frcar:1 (R2 = 0.160)|Insys:3 (R2 = 0.818)|Pvent1:6(R2 = 0.118)|
------------------------------------------------------------------
                     Corrélations partielles
         Frcar  Insys  Pvent
Répul    0.029  0.018 -0.073
```

Aucun de ces coefficients n'est significatif: on peut considérer que le système est complet.

Cette démarche donne finalement des résultats satisfaisants. Mais le système obtenu ne possède pas de propriété d'optimalité; en particulier, il dépend de l'ordre d'introduction des prédicteurs.

2.2 Algorithmes de sélection d'un système de prédicteurs.

Il peut être intéressant d'automatiser la démarche précédente: on parle alors d'algorithme de construction d'un ensemble de prédicteurs. Il existe trois algorithmes principaux:

— Le premier est le même que dans le paragraphe précédent, mais le choix de la variable explicative à introduire éventuellement est uniquement basé sur des considérations numériques; en général, l'algorithme ne tient pas compte du coefficient de détermination des variables explicatives avec le système de pré-

dicteurs et se limite à introduire la variable dont le coefficient de corrélation partielle est le plus élevé, tant que ce coefficient est significatif pour un risque de première espèce fixé.

On aurait ainsi introduit en deuxième variable explicative la pression artérielle pulmonaire au lieu de la pression diastolique suivant leur coefficient de corrélation partielle avec la résistance pulmonaire (0.936 contre 0.905). Le système final aurait été le même que précédemment, la pression diastolique étant introduite en troisième variable explicative (coefficient de corrélation partielle égale à 0.291, $P(F > 8.94)=0.0036$).

Certains programmes vérifient toutefois que la variable à introduire n'est pas colinéaire aux prédicteurs déjà considérés pour éviter des difficultés d'ordre numérique.

Cet algorithme définit ce que l'on appelle régression ascendante.

```
/// Introduction de la variable Incar
          F  235.931 Probabilité critique  0.00000
/// Introduction de la variable Papul
          F  696.582 Probabilité critique  0.00000
/// Introduction de la variable Prdia
          F    8.944 Probabilité critique  0.00363
```

— Le deuxième algorithme consiste à procéder de façon inverse: au départ, toutes les variables explicatives disponibles sont introduites, et l'on élimine au fur et à mesure celle dont le coefficient de corrélation partielle avec la variable expliquée conditionnellement aux autres prédicteurs est le plus faible, tout en étant non significatif.

Cette régression, dite descendante, est parfois efficace et donne en général un autre système de prédicteurs:

```
/// Elimination de la variable Pvent
          F    0.471 Probabilité critique  0.50145
/// Elimination de la variable Insys
          F    0.670 Probabilité critique  0.42015
/// Elimination de la variable Prcar
          F    0.079 Probabilité critique  0.77594
```

Sur ces données, on obtient le même système de prédicteurs, mais il s'agit d'un cas particulier

— Dans le troisième algorithme, chaque introduction d'une variable explicative dans le système de prédicteur est suivie d'une procédure d'élimination analogue à la précédente. Il faut donc définir deux risques: le premier concerne l'introduction et le second, l'élimination. Le second risque doit être supérieur au premier pour assurer la convergence de l'algorithme. Cet algorithme permet l'introduction forcée de variables explicatives au départ: la régression est effectuée suivant ces variables, on complète le système de prédicteurs par une procédure d'introduction, et, en cas d'introduction, on effectue une procédure d'élimination. On recommence ensuite une procédure d'introduction et ainsi de suite. C'est la régression Stepwise.

Nous avons forcé l'introduction de la fréquence cardiaque et de la pression ventriculaire; les résultats sont les suivants, les risques pour l'introduction et l'élimination étant fixés à 0.05:

```
/// Introduction de la variable Incar
          F  218.352 Probabilité critique  0.00000
/// Introduction de la variable Papul
          F  555.559 Probabilité critique  0.00000
/// Elimination de la variable Pvent
          F    0.068 Probabilité critique  0.79093
/// Elimination de la variable Frcar
          F    0.542 Probabilité critique  0.46974
/// Introduction de la variable Prdia
          F    8.944 Probabilité critique  0.00363
```

L'introduction forcée de variables explicatives, possible aussi en régression ascendante, est parfois indispensable pour faire démarrer les algorithmes ascendants: il peut se produire en effet que tous les coefficients de corrélation soient non significatifs, alors qu'un modèle comportant plus d'une variable donne un coefficient de détermination élevé.

A. Bensaber et B. Bleuse-Trillon (1989) en donnent un exemple sur les données d'Anderson (1971) relatives à la consommation de viande aux Etats-

Unis: l'ajustement d'un polynôme en fonction du temps t donne de bons résultats, alors qu'aucun des coefficients de corrélation de t, t^2, t^3, t^4 et t^5 avec la variable expliquée n'est significatif (matrice ci-dessous). Un algorithme ascendant ne donne donc pas de modèle.

	t	t^2	t^3	t^4	t^5	Cons.
t	1.000					
t^2	0.971	1.000				
t^3	0.921	0.986	1.000			
t^4	0.872	0.959	0.992	1.000		
t^5	0.827	0.929	0.975	0.995	1.000	
Cons	-0.327	-0.228	-0.117	-0.020	0.061	1.000

Consommation de viande aux Etats Unis
Corrélations avec la variable t = temps, t^2, t^3, t^4, t^5

Il est préférable de procéder à une régression descendante: on cherche ici à ajuster un polynôme en t de degré minimum: à partir de l'ajustement du polynôme de degré 5, on teste l'égalité à 0 des coefficients de t^k, pour k variant de 5 à 1, et l'on s'arrête au premier rejet; on obtient ainsi un modèle polynomial de degré 3 dont le coefficient de détermination R^2= 0.6708 est hautement significatif (F=12.906, P[F>12.906]=0.0001).

Les algorithmes pas à pas présentent d'autres inconvénients que le précédent. Ils sont tout d'abord construits sur des critères uniquement numériques, l'utilisateur n'ayant pas d'autre choix que d'en attendre les résultats. La différence entre deux coefficients de corrélation partielle n'est pas prise en compte dans le choix de la variable à introduire ou à éliminer, le rang est à lui seul déterminant.

Le critère numérique est en outre critiquable: le risque de première espèce que l'on choisit est calculé sur la loi de chaque coefficient de corrélation partielle (loi de Snedecor), et non sur la loi du plus grand ou du plus petit d'entre eux; la différence est importante, et est discutée dans Draper et Smith (1981, p.311).

BIBLIOGRAPHIE

Anderson T.W. (1971): *The Statistical Analysis of Time series*, Wiley, New York. (application du modèle linéaire pour l'étude des séries chronologiques; introduction aux polynômes orthogonaux pour le modèle polynomial).

Anderson T.W. (1958): *An Introduction to Multivariate Statistical Analysis*, J. Wiley & Sons, New York (la régression dans le modèle gaussien).

Bensaber A., Bleuse-Trillon B. (1989): *Pratique des chroniques et de la prévision à court terme.* Masson, Paris (plus accessible que l'Anderson, avec des exemples numériques)

Cailliez F. et Pagès J.P. (1976): *Introduction à l'analyse des données*, SMASH, 9 rue Duban, 75016 Paris (présentation de la régression et de l'analyse des données à base de l'algèbre linéaire et de la dualité entre un espace euclidien et son dual).

Draper N.R. et Smith H. (1981): *Applied Regression Analysis*, J. Wiley & Sons, New York. (ouvrage fondamental en régression).

Foucart T. et Lafaye J.Y. (1983): *Régression linéaire sur Micro-ordinateur*, Masson, Paris. (présentation géométrique de la régression accompagnée de programmes).

Foucart T. (1985): *Analyse Factorielle. Programmation sur Micro-ordinateur.* Masson, Paris. (présentation simple de l'analyse des données).

Foucart T. (1991): *Introduction aux tests statistiques. Enseignement assisté par ordinateur.* Technip, Paris (tests du F, de Student avec des exemples et des simulations).

Bibliographie

46

Malinvaud E.: *Méthodes Statistiques de l'Econométrie*, Dunod, Paris, 1964. (ouvrage de base en économétrie.)

Mardia K.V., Kent J.T., Bibby J.M.: *Multivariate Analysis*, Academic Press, Londres 1979. (des remarques intéressantes sur le choix des variables explicatives).

Saporta G. (1990): *Probabilités, analyse des données et statistique*, Technip, Paris (ouvrage général sur la statistique; l'exposé sur la régression est complet et bien fait).

Tomassone R., Lesquoy E. et Millier C. (1983): *La régression. Nouveaux regards sur une ancienne méthode statistique*, Masson, Paris. (destiné aux utilisateurs de la régression; bourré de commentaires judicieux, mais des erreurs dans les exemples, surtout à la fin du livre, peut-être corrigés dans la nouvelle édition).

Weisberg S. (1980): *Applied Linear Regression*, Wiley, New-York. (seul défaut: en anglais).

ANNEXE: DONNEES ETUDIEES
(Saporta, 1991)

	frcar	incar	insys	prdia	papul	pvent	répul	prono
1*	90	1.71	19	16	19.5	16	912	2
2*	90	1.68	18.7	24	31	14	1476	1
3*	120	1.4	11.7	23	29	8	1657	1
4*	82	1.79	21.8	14	17.5	10	782	2
5*	80	1.58	19.7	21	28	18.5	1418	1
6*	80	1.13	14.1	18	23.5	9	1664	1
7*	94	2.04	21.7	23	27	10	1059	2
8*	80	1.19	14.9	16	21	16.5	1412	2
9*	78	2.16	27.7	15	20.5	11.5	759	2
10*	100	2.28	22.8	16	23	4	807	2
11*	90	2.79	31	16	25	8	717	2
12*	86	2.7	31.4	15	23	9.5	681	2
13*	80	2.61	32.6	8	15	1	460	2
14*	61	2.84	47.3	11	17	12	479	2
15*	99	3.12	31.8	15	20	11	513	2
16*	92	2.47	26.8	12	19	11	615	2
17*	96	1.88	19.6	12	19	3	809	2
18*	86	1.7	19.8	10	14	10.5	659	2
19*	125	3.37	26.9	18	28	6	665	2
20*	80	2.01	25	15	20	6	796	2
21*	82	3.15	38.4	13	20	6	508	2
22*	110	1.66	15.1	23	31	6.5	1494	1
23*	80	1.5	18.7	13	17	12	907	1
24*	118	1.03	8.7	19	27	10	2097	1
25*	95	1.89	19.9	25	27	20	1143	1
26*	80	1.45	18.1	19	23	15	1269	1
27*	85	1.3	15.1	13	18	10	1108	1
28*	105	1.84	17.5	18	22	10	957	1
29*	122	2.79	22.9	25	36	10	1032	2
30*	81	1.77	21.9	18	27	11	1220	2
31*	118	2.31	19.6	22	27	10	935	2
32*	87	1.2	13.8	34	41	20	2733	1
33*	65	1.19	18.3	15	18	13	1210	1
34*	84	2.15	25.6	27	37	10	1377	2
35*	103	0.91	8.8	30	33.5	10	2945	1
36*	75	2.54	33.9	24	31	16	976	2
37*	90	2.08	23.1	20	28	6	1077	2
38*	90	1.93	21.4	11	18	10	746	2
39*	90	0.95	10.6	20	24	6	2021	1
40*	65	2.38	36.6	16	22	12	739	2
41*	95	0.99	10.4	20	27.5	8	2222	1
42*	95	0.85	8.9	19	22	15.5	2071	1
43*	86	2.05	23.8	21	28	10	1093	2
44*	82	2.02	24.6	16	22	14	871	2
45*	70	1.44	20.6	19	26.5	11	1472	1
46*	92	3.06	33.3	10	15	6	392	2
47*	94	1.31	13.9	26	40	15	2443	1
48*	79	1.29	16.3	24	31	10	1922	1
49*	67	1.47	21.9	15	18	16	980	2

Données Infarctus du myocarde (début)

	frcar	incar	insys	prdia	papul	pvent	répul	prono
50*	75	1.21	16.1	19	24	4	1587	1
51*	80	2.41	30.9	19	24	7	797	2
52*	61	3.28	54	12	16	7	390	2
53*	110	1.24	11.3	22	27.5	11	1774	1
54*	116	1.85	15.9	33	42	13	1816	1
55*	75	2	26.7	16	22	5	880	2
56*	92	1.97	21.4	18	27	3	1096	1
57*	110	0.96	8.80	15	19	16	1583	2
58*	95	2.56	26.9	8	13	3	406	2
59*	75	2.32	30.9	8	10	6	345	2
60*	80	2.65	33.1	13	19	9	574	2
61*	102	1.60	15.7	24	31	16	1550	1
62*	86	1.67	19.4	18	23	8.5	1102	2
63*	60	0.82	13.7	22	32	13	3122	1
64*	100	1.76	17.6	23	33	2	1500	2
65*	80	3.28	41	12	17	2	415	2
66*	108	2.96	27.4	24	35	6.5	946	2
67*	92	1.37	14.8	25	46	11	2686	1
68*	100	1.38	13.8	20	31	11	1797	1
69*	80	2.85	35.6	25	32	7	898	2
70*	87	2.51	28.8	16	24	20	765	1
71*	100	2.31	23.1	8	12	1	416	2
72*	120	1.18	9.9	25	36	8	2441	1
73*	115	1.83	15.9	25	30	8	1311	1
74*	101	2.55	25.2	23.2	30.5	9	957	2
75*	92	2.17	23.5	19	24	3	885	2
76*	87	1.42	16.1	20	26	10	1465	1
77*	80	1.59	19.9	13	20.5	4	1031	2
78*	88	1.47	16.7	23	32.5	10	1769	1
79*	104	1.23	11.8	27	33	11	2146	1
80*	90	1.45	16.1	17	24	8.5	1324	2
81*	67	0.85	12.7	26	33	11	3106	1
82*	87	2.37	27.2	15	22	10	743	2
83*	108	2.40	22.2	26	31	4	1033	2
84*	120	1.91	15.9	18	27	15	1131	1
85*	108	1.50	13.9	28	43	16	1813	1
86*	86	2.36	27.4	24	34	8	1153	2
87*	112	1.56	13.9	24	29	4	1487	1
88*	80	1.34	17	16	25	16	1493	1
89*	95	1.65	17.4	20	33	7	1600	1
90*	90	2.04	22.7	28	41	10	1608	1
91*	90	3.03	33.6	17	23.5	7	620	2
92*	94	1.21	12.9	17	22	3	1455	1
93*	51	1.34	26.3	11	17	6	1015	1
94*	110	1.17	10.6	29	35	10.5	2393	1
95*	96	1.74	18.1	24	29	6	1333	1
96*	132	1.31	9.9	23	28	12	1710	1
97*	135	0.95	7	15	20	7	1684	1
98*	105	1.92	18.3	18	24	3	1000	1
99*	99	0.83	8.4	23	27	8	2602	1
100*	116	0.60	5.2	33	38	10	5067	1
101*	112	1.54	13.8	25	31	8	1610	1

Données Infarctus du myocarde (fin)