CHAPITRE 3-1
LE MODÈLE LINÉAIRE GÉNÉRAL ET SON ESTIMATION PAR LA MÉTHODE DES MOINDRES CARRÉS

Plan

CHAPITRE 3-1
LE MODÈLE LINÉAIRE GÉNÉRAL ET SON ESTIMATION PAR LA MÉTHODE DES MOINDRES CARRÉS

3-1.1 Le modèle linéaire

3-1.1.1 UN EXEMPLE DE MODÈLE LINÉAIRE

On veut étudier la relation entre la taille de la plus grande ville d'un pays (variable dépendante, dénotée *PLAR*) et la population totale et le PIB per capita du pays (variables indépendantes, dénotées *PTOT* et *GNPC* respectivement). L'un des modèles que l'on pourrait envisager serait

$$PLAR_i = \beta_1 + \beta_2\, PTOT_i + \beta_3\, GNPC_i$$

Si l'on fixe la valeur des paramètres β_1, β_2 et β_3, et si l'on connaît la population totale et le PIB per capita d'un pays, on peut calculer ce que prédit ce modèle quant à la population de la plus grande ville. Par exemple, supposons que l'on fixe [1]

$$\beta_1 = 3500$$

$$\beta_2 = 0{,}01$$

$$\beta_3 = 0{,}1$$

Voici les données relatives au Brésil et au Costa Rica pour 1990, extraites du tableau 1 dans Lemelin et Polèse (1995) :

		PLAR ('000)	*PTOT* ('000)	*PURB* ('000)	*GNPC* ($ US)
7 Brazil	Sao Paulo	17395	150368	112643	2680
13 Costa Rica	San Jose CR	1016	3015	1420	1900

À partir de ces données, on peut calculer que le modèle « prédit » qu'en 1990, la population de Sao Paulo, en milliers, était de

$$3500 + (0{,}01 \times 150368) + (0{,}1 \times 2680) = 5272$$

et celle de San José, CR,

$$3500 + (0{,}01 \times 3015) + (0{,}1 \times 1900) = 3720$$

Comme on peut le voir, ce sont de très mauvaises prédictions : l'écart avec la valeur observée est de 12123 dans le premier cas et de −2704 dans le second. Il serait prématuré de conclure que le modèle est mauvais à partir de deux observations. On peut cependant soupçonner que la relation linéaire est peu appropriée au phénomène étudié.

Combien de paramètres y a-t-il dans le modèle ? Qu'est-ce que la constante du modèle ?

Dans l'exemple qui précède, le modèle comporte trois paramètres (β_1, β_2 et β_3). À chaque paramètre est associée une variable indépendante. Le paramètre β_1 est la constante du modèle : la variable indépendante qui lui est associée est une constante. Noter ici le double emploi du mot « constante » : il désigne à la fois (1) une variable dont la valeur est la même pour toutes les observations et (2) le paramètre β_1 associé à cette variable. En effet, si l'on voulait être complètement explicite, il faudrait présenter les données sous la forme suivante :

		Constante	*PLAR* ('000)	*PTOT* ('000)	*PURB* ('000)	*GNPC* ($ US)
7 Brazil	Sao Paulo	1	17395	150368	112643	2680
13 Costa Rica	San Jose CR	1	1016	3015	1420	1900

Le calcul des « prédictions » du modèle s'écrit alors

$$(3500 \times 1) + (0{,}01 \times 150368) + (0{,}1 \times 2680) = 5272$$

$$(3500 \times 1) + (0{,}01 \times 3015) + (0{,}1 \times 1900) = 3720$$

Le paramètre β_1 est alors multiplié comme les autres par la valeur de la variable correspondante. Il est important de garder à l'esprit que la constante est l'une des variables du modèle, notamment lorsqu'il s'agit de compter le nombre de variables indépendantes (ici, trois). Cela est également important lorsque le modèle comporte des variables indépendantes dichotomiques (appelées variables muettes), pour ne pas introduire de redondance dans le modèle (voir le chapitre 4-2).

Lorsque le modèle ne comprend que deux variables indépendantes dont l'une est une constante, il s'agit de la régression linéaire *simple* :

$$y_i = \alpha + \beta x_i$$

---

1 Ces valeurs se rapprochent des valeurs estimées par la méthode des moindres carrés ordinaires, appliquée aux données pour 1990 présentées au tableau 1 de Lemelin et Polèse (1995). Les valeurs estimées exactes sont : $\beta_1 = 3431$, $\beta_2 = 0{,}01324$ et $\beta_3 = 0{,}09375$. Le coefficient de détermination multiple de la régression est de 0,26.

Il ne sera question ici que du cas général de la régression linéaire *multiple*, dont la régression simple est un cas particulier.

3-1.1.2 LE MODÈLE LINÉAIRE SOUS SA FORME GÉNÉRALE

Généralisons à partir de l'exemple que nous venons d'examiner.

Pour un modèle théorique déterministe, la forme générale du modèle linéaire [2] est donnée par

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

où l'indice souscrit i désigne un individu dans la population ou une observation dans l'échantillon. La variable dépendante est y_i et les variables indépendantes sont x_{i1}, x_{i2}, ..., x_{ik}. Les coefficients β_j sont les paramètres inconnus du modèle, à estimer.

En général, l'une des variables indépendantes, la plupart du temps la première, est une constante :

$x_{i1} = 1$ pour tout i

On peut alors écrire le modèle de la façon suivante :

$$y_i = \sum_{j=1}^{k} \beta_j x_{ij} = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

On utilise l'expression « constante du modèle » pour désigner à la fois la variable indépendante dont la valeur est constante (x_{i1}) et le paramètre qui lui est associé (β_1).

NOTE :

Certains auteurs écrivent

$$y_i = \sum_{j=0}^{h} \beta_j x_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_h x_{ih}$$

C'est pourquoi il faut toujours préciser, quand on donne le nombre de variables indépendantes, si ce nombre inclut la constante ou non (ici, il y a $h + 1 = k$ variables indépendantes en comptant la constante). Ce détail a son importance quand il s'agit de compter les degrés de liberté

[2] Pour être exact, il faudrait parler ici du modèle linéaire général à variable dépendante unique, puisque le terme « modèle linéaire général » désigne plus largement un modèle qui peut comporter plusieurs variables dépendantes.

associés à certaines variables-tests. Dans cet ouvrage, le nombre de variables indépendantes (dénoté *k* la plupart du temps) inclut toujours la constante.

3-1.1.3 LA REPRÉSENTATION DES RELATIONS NON LINÉAIRES DANS LE MODÈLE LINÉAIRE

Le modèle linéaire général permet de représenter des relations non linéaires, pourvu qu'elles soient linéaires par rapport aux paramètres ou linéarisables. Quelques exemples illustrent ce que cela signifie.

Exemple 1 : la transformation logarithmique

La relation exponentielle

$$PLAR_i = K\, PURB_i^h$$

devient linéaire lorsqu'on prend les logarithmes :

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

Mais alors, les variables du modèle ne sont plus *PLAR* et *PURB*, mais ln *PLAR* et ln *PURB*.

Lemelin et Polèse (1995) ont estimé les paramètres de cette relation ; les calculs ont été faits avec les logarithmes népériens [3]. Leurs résultats sont donnés au tableau 2 de l'article : ln K = 2,067 et h = 0,636. Voici les valeurs des variables, arrondies, pour le Brésil et le Costa Rica en 1990, calculées à partir du tableau 1 de Lemelin et Polèse (1995) :

		ln *PLAR* ('000)	ln *PURB* ('000)
7 Brazil	Sao Paulo	9,76	11,63
13 Costa Rica	San Jose CR	6,92	7,26

À partir de ces données, on peut calculer que le modèle « prédit » qu'en 1990, la population de Sao Paulo, en milliers, était de

$$\text{EXP}[2{,}067 + (0{,}636 \times 11{,}63)] = \text{EXP}(9{,}46) = 12883$$

et celle de San José, CR,

$$\text{EXP}[2{,}067 + (0{,}636 \times 7{,}26)] = \text{EXP}(6{,}68) = 800$$

---

[3] Comme on peut le voir en appliquant les logarithmes népériens à la relation exponentielle, la valeur estimée du paramètre *h* n'est pas influencée par le choix de la base des logarithmes (le nombre transcendantal *e* pour les logarithmes népériens ou 10 pour les logarithmes communs). La constante estimée, log *K* ou ln *K*, dépend cependant du choix de la base.

La linéarisation d'un modèle par sa transformation logarithmique est un procédé très fréquent. Nous l'avons déjà vu à propos de l'ajustement d'une courbe de tendance (voir 1-2.3) :

$$y_t = y_0 (1+r)^t$$

devient

$$\log y_t = \log y_0 + t \log(1+r)$$

Dans ce cas, cependant, l'exposant t est une des deux variables indépendantes du modèle (l'autre étant la constante), et $\log y_t$ est la variable dépendante, alors que y_0 et $\log(1+r)$ en sont les paramètres à estimer.

En économie, on applique la transformation logarithmique à la fonction de production Cobb-Douglas, qui se définit comme

$$Y_i = A\, K_i^B\, T_i^C$$

où

Y est la quantité produite

K est la quantité de capital utilisée

T est la quantité de main-d'oeuvre utilisée

A, B et C sont des paramètres.

Quand on applique la transformation logarithmique, le modèle devient linéaire :

$$\log Y_i = \log A + B \log K_i + C \log T_i$$

Exemple 2 : l'ajout de variables indépendantes

La relation

$$\ln PURB_i = a + b \ln PTOT_i + c \ln GNPC_i + d (\ln GNPC_i)^2$$

comporte trois variables indépendantes (en comptant la constante), mais l'une des variables indépendantes apparaît à la fois sous forme linéaire et sous forme quadratique. Cette relation peut néanmoins être traitée comme une relation linéaire. Il suffit pour cela de considérer que $\ln GNPC$ et $(\ln GNPC)^2$ sont deux variables indépendantes différentes. Avec la constante, le modèle compte alors quatre variables indépendantes [4].

Ce procédé peut évidemment se généraliser. Ainsi, la relation cubique

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3$$

devient linéaire quand on définit

$z_{i1} = 1$ (constante), $z_{i2} = x_i$, $z_{i3} = x_i^2$ et $z_{i4} = x_i^3$

Le modèle peut alors s'écrire sous la forme d'une relation linéaire :

$$y_i = \beta_1 + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} = \sum_{j=1}^{4} \beta_j z_{ij}$$

Ce procédé permet aussi de linéariser des polynômes de degré quelconque, notamment pour estimer des *surfaces de tendance*. L'estimation de surfaces de tendance peut servir à décrire les variations dans l'espace des valeurs d'une variable comme, par exemple, le prix des maisons. Il s'agit bien d'un modèle descriptif [5], puisque la relation ne s'appuie sur aucune théorie. Les données requises sont les prix de vente des immeubles et leur localisation, en coordonnées XY [6]. Un polynôme de second degré est de la forme

$$Z_i = \beta_0 + X_i \beta_1 + Y_i \beta_2 + X_i^2 \beta_3 + Y_i^2 \beta_4 + X_i Y_i \beta_5$$

Ce modèle compte six variables indépendantes : la constante, X_i, Y_i, X_i^2, Y_i^2, et X_iY_i. Dans un polynôme de troisième degré, on aurait les quatre variables supplémentaires suivantes : X_i^3, Y_i^3, $X_i^2Y_i$ et $X_iY_i^2$. Plus le degré du polynôme est élevé, plus la surface qu'il décrit peut être complexe, mais plus le nombre de variables indépendantes est élevé : nous verrons que le nombre de paramètres que l'on peut estimer est limité par le nombre d'observations.

3-1.2 Où intervient l'aléatoire ?

Dans l'analyse de régression, on cherche à connaître les paramètres de la relation entre y_i et les x_{ij}. Or, si le modèle théorique déterministe était vrai, chaque observation se conformerait avec exactitude au modèle : pour connaître les paramètres β_j, il suffirait alors de recueillir sur y_i

---

4 On peut comparer ce procédé à l'utilisation, au théâtre ou au cinéma, de plus d'un acteur pour représenter un même personnage à des âges différents.

5 De la même façon que l'ajustement d'une courbe de tendance temporelle est un modèle descriptif. Souvent, un modèle descriptif est utilisé comme complément d'un modèle théorique (voir notamment les travaux de François Desrosiers et Marius Thériault, de l'Université Laval, sur les prix immobiliers dans la région de Québec).

6 Dans un système d'information géographique (SIG) la situation des objets dans l'espace est enregistrée sous la forme de coordonnées, comme la position d'un point dans le plan cartésien ; ces coordonnées sont parfois la latitude et la longitude géographiques de la position, mais pas nécessairement.

et les x_{ij} autant d'observations qu'il y a de paramètres à estimer et de résoudre un système de k équations (une pour chaque observation i) à k inconnues (les β_j) [7].

Ainsi, en physique, la vitesse v d'un objet en chute libre est égale au temps écoulé t, multiplié par la constante d'accélération a [8] :

$$v = a\,t$$

De cette relation, il découle que la distance parcourue d est proportionnelle au carré du temps de chute :

$$d = \int_0^t v\ dt = \int_0^t at\ dt = \frac{1}{2} a t^2$$

Puisque la Loi de l'accélération de la pesanteur est déterministe, il suffit, pour connaître la valeur de la constante a, d'une seule observation précise d'un corps en chute libre : ayant mesuré d et t, il est facile de trouver la valeur de a.

De même, si le modèle

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

était exact, les valeurs observées pour le Brésil et le Costa Rica permettraient de définir un système de deux équations linéaires à deux inconnues :

$$9{,}76 = \ln K + 11{,}63\ h \text{ (Brésil)}$$

$$6{,}92 = \ln K + 7{,}26\ h \text{ (Costa Rica)}$$

La solution de ce système est donnée par

$$\ln K = 2{,}20$$

$$h = 0{,}65$$

Ce serait trop beau ! Nous savons bien que nos modèles, surtout en sciences sociales, sont beaucoup trop simples, *a fortiori* lorsque ils sont linéaires, pour représenter toute la complexité du réel. Nos modèles théoriques ne sont que des approximations et, même quand ils sont bons, ce n'est qu'approximativement que les observations s'y conforment. Il s'ensuit que, si l'on estimait les k paramètres à l'aide de k observations, il y a de fortes chances qu'une nouvelle

---

7 Des méthodes analogues sont utilisées dans certaines circonstances. On parle alors de « calibrage » d'un modèle, plutôt que d'« estimation ».

8 La constante d'accélération de la pesanteur est égale à 980,621 cm/s^2, c'est-à-dire 32,1725 pi/s^2 au niveau de la mer.

observation (k+1) soit incompatible avec le modèle (d'un point de vue déterministe) : la $(k+1)^{\text{ème}}$ équation serait contradictoire avec les autres [9].

Par exemple, aux observations sur Sao Paulo et San José, ajoutons les données relatives à Toronto (calculées à partir du tableau 1 de Lemelin et Polèse, 1995) :

		ln PLAR ('000)	ln PURB ('000)
7 Brazil	Sao Paulo	9,76	11,63
9 Canada	Toronto	8,15	9,93
13 Costa Rica	San Jose CR	6,92	7,26

Si nous appliquons à Toronto les coefficients calculés précédemment, nous obtenons :

$$\ln PLAR = 2{,}20 + 0{,}65 \ln PURB = 2{,}20 + (0{,}65 \times 9{,}93) = 8{,}65 \neq 8{,}15$$

L'équation n'est pas vérifiée dans le cas de Toronto. En fait, nous savons qu'il n'y a pas de solution au système suivant, de trois équations à deux inconnues :

$9{,}76 = \ln K + 11{,}63\ h$ (Brésil)

$8{,}15 = \ln K + 9{,}93\ h$ (Can.)

$6{,}92 = \ln K + 7{,}26\ h$ (Costa Rica)

Plus généralement, avec un échantillon de n observations et k paramètres à estimer (un pour chaque variable indépendante), on peut construire un système de n équations à k inconnues. Si le nombre d'observations n est supérieur au nombre de paramètres à estimer, le nombre d'équations est supérieur au nombre d'inconnues. Or, le plus souvent, un tel système n'a pas de solution parce que les équations sont incompatibles entre elles.

Donc, même quand un modèle est une bonne approximation de la réalité, il subsiste néanmoins un écart entre les prédictions du modèle et les observations. Cet écart s'explique notamment par l'absence parmi les variables indépendantes de nombreux facteurs secondaires dont l'influence individuelle est faible (modèle incomplet, trop simple), mais aussi par des erreurs de mesure sur les variables. Cela se traduit par une « erreur », qui ne paraît pas systématique mais semble plutôt engendrée par le hasard. Pour représenter cette erreur dans le modèle théorique, on lui ajoute une variable aléatoire :

---

9 Il se pose un problème tout à fait similaire dans les sciences dites exactes, comme la physique. Les erreurs de mesure introduisent un élément d'inexactitude dans les observations. Il s'ensuit que, même quand les modèles sont des « lois » déterministes, il subsiste un certain degré d'imprécision dans la connaissance des valeurs des paramètres (comme la constante de l'accélération de la pesanteur).

$$y_i = \sum_{j=1}^{k} \beta_j \, x_{ij} + u_i$$

Le *terme aléatoire*, u_i, est aussi appelé *terme d'erreur*, *erreur stochastique*, ou *erreur* tout court, ou encore perturbation (*disturbance term*). Il est à noter que les valeurs prises par le terme aléatoire sont tout aussi inobservables que les paramètres de la relation : tout ce que l'on observe, ce sont les valeurs de la variable dépendante et des variables indépendantes.

Par exemple, le modèle

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

est un modèle théorique déterministe. Mais le modèle sur lequel s'appuient l'estimation des paramètres et les tests d'hypothèse rapportés dans Lemelin et Polèse (1995) est en réalité

$$\ln PLAR_i = \ln K + h \ln PURB_i + u_i$$

où u_i est un terme aléatoire.

Ainsi, dans l'analyse de régression, l'aléatoire s'introduit par la troisième « porte », que nous décrivions comme suit :

> « [...] il y a trois "portes" par lesquelles l'aléatoire s'introduit dans les modèles :
>
> 1. Il y a d'abord la nature aléatoire, déjà mentionnée, du lien entre un échantillon et la population dont il est tiré.
> 2. Les variables opérationnelles sont des mesures imparfaites des concepts et on peut considérer que l'erreur de mesure est aléatoire (c'est-à-dire déterminée au hasard). On peut donc représenter par un modèle aléatoire l'influence des erreurs de mesure qui interviennent lors de la traduction des hypothèses théoriques en hypothèses opérationnelles (les modèles de la « théorie des erreurs » en sciences physiques ont d'ailleurs été parmi les premiers modèles aléatoires).
> 3. Enfin, certains phénomènes nous apparaissent comme aléatoires en soi et ils ne peuvent pas être représentés adéquatement par des modèles théoriques non aléatoires. Le hasard dans ces modèles est un concept qui recouvre tantôt une indétermination fondamentale (comme en physique des particules), tantôt une multitude de facteurs inobservables (comme c'est plus souvent le cas en sciences sociales [10]), dont les manifestations apparaissent comme régies par des lois de probabilité ».
>
> (Chapitre 2-2) [11].

---

[10] Pensons en particulier aux modèles d'utilité aléatoire (*random utility*) sous-jacents aux modèles de choix discrets (*discrete choice*) logit, probit, etc. Ces modèles sont abordés au chapitre 4-3.

[11] Ce passage est inspiré de Malinvaud, qui écrit : « On sait que l'emploi du calcul des probabilités pour l'analyse des données statistiques est justifié par l'une ou l'autre des deux considérations suivantes. Ou bien le phénomène étudié est assimilé à un processus comportant une détermination aléatoire de certaines grandeurs ;

Selon cette conception, même si les données englobaient la totalité de la population étudiée, l'élément aléatoire ne disparaîtrait pas : car ce qui est aléatoire ici, ce n'est plus tant le lien entre la population et l'échantillon que le lien entre le modèle déterministe (la loi mathématique), dont les paramètres sont inconnus, et les observations, qui s'écartent du modèle de façon aléatoire [12] : ainsi, les observations cessent d'être incompatibles avec le modèle ; elles sont simplement, du point de vue probabiliste, plus ou moins compatibles avec le modèle. Ajoutons cependant que dans ce contexte, la distinction population-échantillon (voir chapitres 2-1 et 2-2) subsiste néanmoins ; mais elle se trouve d'abord dans le fait que les valeurs que prennent les termes aléatoires inobservables sont *tirées de la population infinie des valeurs que pourrait générer le processus aléatoire* sous-jacent à chacun des termes aléatoires. Comme le laisse entendre la formulation qui précède, il n'est pas exclu *a priori* que les valeurs des termes aléatoires associés à différentes observations soient engendrées par des processus aléatoires différents. C'est pour signifier cela que, dans certains contextes, on parle des termes aléatoires au pluriel.

À un premier niveau, donc, la combinaison d'un terme aléatoire et d'un modèle déterministe permet de s'accommoder du caractère approximatif de l'accord entre le modèle et les observations. Pour aller plus loin, il faut caractériser les distributions de probabilité des termes aléatoires u_i. On aura alors un modèle aléatoire et l'on sera en mesure d'appliquer les méthodes de l'induction statistique, en vue notamment

- d'estimer les paramètres des distributions de probabilité des termes aléatoires ;
- d'estimer les paramètres des distributions d'échantillonnage des estimateurs ;
- de faire des tests d'hypothèses.

3-1.3 L'estimateur des moindres carrés ordinaires

Pour compléter la notation utilisée, convenons ce qui suit :

---

ces grandeurs sont alors considérées comme aléatoires dans l'univers [c'est-à-dire dans la population] comme dans l'échantillon observé. Ou bien la sélection des unités observées résulte d'un tirage aléatoire ; la composition de l'échantillon est alors aléatoire, donc aussi les données obtenues, même si elles portent sur des grandeurs non aléatoires » (Malinvaud, 1969, p.62). Malinvaud poursuit en disant que le premier type de justification lui semble plus approprié au contexte de l'économétrie.

12 Il y a quelque chose de la caverne de Platon dans cette conception (voir en annexe le texte de l'allégorie de la caverne de Platon). On traite la réalité observable comme si elle était le reflet imparfait (l'ombre projetée) du modèle théorique déterministe (l'idéal). Le terme aléatoire du modèle représente les imperfections de la réalité observable. L'induction statistique cherche à discerner l'« idéal » (au sens platonicien de ce mot) à travers son reflet.

b_i : valeur estimée du paramètre β_i.

$\hat{y}_i$: valeur de y_i « prédite » ou calculée par le modèle tel qu'estimé.

On a, par définition

$$\hat{y}_i \equiv \sum_j b_j x_{ij}$$

e_i : résidu calculé (ou « erreur ») de la régression pour la $i^{\text{ème}}$ observation.

On a, par définition

$$e_i \equiv y_i - \hat{y}_i \equiv y_i - \sum_j b_j x_{ij}$$

NB. Il ne faut pas confondre e_i, le résidu calculé (observable) avec le terme aléatoire correspondant u_i, inobservable.

3-1.3.1 DÉFINITION

On peut appliquer la méthode des moindres carrés sans compléter la spécification du modèle aléatoire (Voir l'énoncé de ce principe en 2-2.3). Il suffit de reconnaître que le modèle n'est qu'une approximation et que les observations ne s'y conforment qu'approximativement.

Le principe des moindres carrés consiste à choisir les valeurs estimées b_j qui minimisent la somme des carrés des résidus (ou « erreurs »). Cela veut dire minimiser la somme des carrés des différences entre les valeurs observées de y_i et les valeurs « prédites » $\hat{y}_i$:

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \sum_j b_j \, x_{ij} \right)^2 = \sum_i e_i^2$$

L'expression $\sum_i (y_i - \hat{y}_i)^2$ est donc le carré de la distance euclidienne généralisée entre les valeurs observées de la variable dépendante et les valeurs prédites. La solution de ce problème de minimisation est donnée par l'estimateur des moindres carrés ordinaires.

Les résultats de l'estimation sont souvent rapportés sous forme de tableau : voir les tableaux 2 et 4 de Lemelin et Polèse (1995), le tableau 1 de Heikkila *et al.* (1989) ou le tableau 1 de Richardson *et al.* (1990).

3-1.3.2 QUELQUES PROPRIÉTÉS DE L'ESTIMATEUR DES MOINDRES CARRÉS ORDINAIRES

(1) Estimateur linéaire

Cet estimateur est *linéaire*, c'est-à-dire que chaque b_j est calculé comme fonction linéaire des y_i, plus exactement comme une somme pondérée des y_i :

$$b_j = \sum_i w_{ji}\, y_i$$

où chacun des coefficients w_{ji} dépend de l'ensemble des x_{gh} [13].

(2) Somme des résidus nulle

Lorsque le modèle de régression comporte une constante, comme c'est généralement le cas, la somme des résidus de la régression est nulle :

$$\sum_i e_i = 0$$

La démonstration n'est pas donnée ici, parce qu'elle fait appel à l'écriture matricielle.

(3) Relation entre les moyennes

Lorsque le modèle comporte une constante, la moyenne des valeurs prédites est égale à la valeur prédite à partir des valeurs moyennes des variables indépendantes, et toutes deux sont égales à la valeur moyenne observée de la variable dépendante :

$$m_y = m_{\hat{y}} = \sum_j b_j\, m_{x_j}$$

Cette propriété découle de la précédente.

Démonstration :

Nous savons que

$$\sum_i e_i = 0$$

Or

---

[13] Plus précisément, w_{ji} est l'élément j,i de la matrice $(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$. Noter que le fait que l'*estimateur* soit linéaire n'est pas une conséquence de ce que le *modèle* soit linéaire.

$$e_i \equiv y_i - \hat{y}_i \equiv y_i - \sum_j b_j x_{ij}$$

Cela implique notamment que la somme des valeurs « prédites » est égale à la somme des valeurs observées :

$$\sum_i y_i - \sum_i \hat{y}_i \equiv \sum_i \left(y_i - \hat{y}_i\right) \equiv \sum_i e_i \equiv 0$$

$$\sum_i y_i = \sum_i \hat{y}_i$$

Mais, puisque

$$\sum_i \hat{y}_i \equiv \sum_i \left(\sum_j b_j x_{ij} \right) \equiv \sum_j b_j \left(\sum_i x_{ij} \right)$$

alors

$$\sum_i y_i = \sum_i \hat{y}_i$$

implique

$$\sum_i y_i \equiv \sum_i \hat{y}_i \equiv \sum_j b_j \left(\sum_i x_{ij} \right)$$

$$\left(\frac{1}{n}\right)\sum_i y_i \equiv \left(\frac{1}{n}\right)\sum_i \hat{y}_i \equiv \left(\frac{1}{n}\right)\sum_j b_j \left(\sum_i x_{ij} \right) \equiv \sum_j b_j \left[\left(\frac{1}{n}\right)\sum_i x_{ij} \right]$$

Or

$$m_y = \left(\frac{1}{n}\right)\sum_i y_i$$

$$m_{\hat{y}} = \left(\frac{1}{n}\right)\sum_i \hat{y}_i$$

$$m_{x_j} = \left(\frac{1}{n}\right)\sum_i x_{ij}$$

On a donc

$$m_y = m_{\hat{y}} = \sum_j b_j \, m_{x_j}$$

3-1.4 Le coefficient de détermination multiple et l'analyse de la variance

3-1.4.1 CONSTRUCTION DU COEFFICIENT DE DÉTERMINATION MULTIPLE

Le coefficient de détermination multiple est une mesure d'association qui se rattache à la famille des mesures de similarité ; plus exactement, c'est une mesure du degré d'accord entre le modèle et les observations. En statistique, une telle mesure s'appelle une « mesure d'ajustement » (*goodness of fit measure*).

Le coefficient de détermination multiple est basé sur une analyse de décomposition [14] de la variabilité de la variable dépendante, où cette variabilité est mesurée par la somme des carrés des déviations par rapport à la moyenne :

$$\sum_i (y_i - m_y)^2 = (n-1)\, s_y^2$$

Ce type d'analyse de décomposition est appelé en statistique une « analyse de la variance » [15].

Première étape : décomposition de la variablilité

Lorsque le modèle comporte une constante, on peut décomposer la variabilité en deux composantes [16] :

$$\sum_i (y_i - m_y)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - m_y)^2$$

Esquisse de démonstration :

$$\sum_i (y_i - m_y)^2 = (n-1)\, s_y^2$$

Si l'on développe le membre de gauche de cette expression, on obtient

$$\sum_i (y_i - m_y)^2 = \sum_i \left[(y_i - \hat{y}_i) + (\hat{y}_i - m_y)\right]^2$$

$$\sum_i (y_i - m_y)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - m_y)^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - m_y)$$

---

14 Sur l'analyse de décomposition, voir 1-2.2 .

15 Il y a une forme plus spécialisée d'analyse de la variance qui permet d'examiner la relation entre une variable dépendante et plusieurs variables indépendantes catégoriques, en décomposant la variance de la variable dépendante entre la variance *à l'intérieur* des groupes (définis par des combinaisons de catégories des variables indépendantes) et la variance *entre* les groupes. Nous en reparlerons en 4-2.

On peut montrer que, **si le modèle comporte une constante**, le dernier terme est nul [17], de sorte que

$$\sum_i (y_i - m_y)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - m_y)^2$$

Deuxième étape : interprétation des éléments de la décomposition

Dans l'expression $\sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - m_y)^2$, le second terme est une mesure de la variabilité des valeurs prédites par le modèle : c'est la partie « expliquée » de la variabilité. Le premier terme est une mesure de la variabilité des résidus : c'est la partie de la variabilité qui échappe au modèle. On a donc

Variabilité totale = variabilité résiduelle + variabilité « expliquée »

Cette interprétation conduit à la notation suivante, fréquemment utilisée dans les sorties des logiciels d'applications statistiques [18] :

SST (*Sum of Squares, Total*) $= \sum_i (y_i - m_y)^2 = (n-1)\, s_y^2$

SSM (*Sum of Squares, Model*) $= \sum_i (\hat{y}_i - m_{\hat{y}})^2 = \sum_i (\hat{y}_i - m_y)^2 = (n-1)\, s_{\hat{y}}^2$

(parce que, lorsqu'il y a une constante, $m_y = m_{\hat{y}}$)

SSR (*Sum of Squares, Residuals*) $= \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$

En somme,

---

16 S'il n'y a pas de constante, la décomposition n'est plus valide. Il peut même arriver alors que R^2 soit négatif.

17 Cette démonstration fait appel à l'écriture matricielle.

18 Il faut faire attention, cependant, car on trouve aussi la notation suivante : *SSR*, pour *Sum of Squares, Regression*, à la place de *SSM*, et *SSE*, pour *Sum of Squares, Errors*, à la place de *SSR*.

Variabilité totale		Variabilité résiduelle		Variabilité « expliquée »
SST		*SSR*		*SSM*
Sum of Squares, Total	=	*Sum of Squares, Residuals*	+	*Sum of Squares, Model*
$\sum_i (y_i - m_y)^2 = (n-1)s_y^2$		$\sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$		$\sum_i (\hat{y}_i - m_{\hat{y}})^2 = \sum_i (\hat{y}_i - m_y)^2 = (n-1)s_{\hat{y}}^2$

Troisième étape : construction d'une mesure d'ajustement (« goodness of fit »)

Le *coefficient de détermination multiple* est la part de la variabilité « expliquée » dans la variabilité totale :

$$R^2 = \frac{\text{Variabilité «expliquée»}}{\text{Variabilité totale}} = \frac{SSM}{SST}$$

Ou encore, puisque $SST = SSR + SSM$, on a $SSM = SST - SSR$ et

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i e_i^2}{(n-1)s_y^2}$$

Rappelons que la méthode des moindres carrés consiste à prendre comme estimés des paramètres les valeurs des coefficients qui minimisent $\sum_i e_i^2$. À la lumière de la formule énoncée ci-dessus, on voit que la méthode des moindres carrés revient donc à choisir les valeurs des coefficients qui maximisent R^2 *étant donné la spécification*, c'est-à-dire *dans le cadre du modèle retenu* [19].

La valeur du coefficient de détermination multiple est généralement rapportée dans les tableaux de résultats : voir les tableaux 2 et 4 de Lemelin et Polèse (1995).

---

[19] Cela n'est pas la même chose que de comparer, comme nous le verrons plus loin, les R^2 de différents modèles après avoir estimé les paramètres de chacun d'eux de façon à obtenir pour chaque modèle le R^2 le plus élevé possible par la méthode des moindres carrés.

3-1.4.2 DOMAINE DE VARIATION DU COEFFICIENT DE DÉTERMINATION MULTIPLE (VALEURS EXTRÊMES)

Le coefficient de détermination varie entre zéro et un. Mathématiquement en effet, *SST*, *SSM* et *SSR* sont des sommes de carrés : en conséquence, leur valeur ne peut pas être négative. De plus, $SST = SSR + SSM$ *:* il s'ensuit que ni *SSR*, ni *SSM* ne peuvent excéder la valeur de *SST*. Enfin, puisque $R^2 = \frac{SSM}{SST}$, il découle de ce qui précède que le coefficient de détermination R^2 ne peut pas être inférieur à zéro ou supérieur à un. Examinons maintenant dans quelles circonstances R^2 pourrait atteindre à ces valeurs limites.

Le coefficient de détermination R^2 est égal à un lorsque $SSR = 0$, c'est-à-dire lorsque le modèle reproduit parfaitement les observations qui ont servi à en estimer les paramètres. Il est égal à zéro lorsque $SSR = SST$, c'est-à-dire lorsque $SSM = 0$. Mais dans quelles circonstances pourrions-nous obtenir $SSM = 0$? Eh bien, *SSM* est une somme de carrés :

$$SSM = \sum_i \left(\hat{y}_i - m_y\right)^2$$

Par conséquent, on ne peut avoir $SSM = 0$ que si tous les termes de la somme sont nuls, c'est-à-dire si $\hat{y}_i = m_y$ pour chaque observation *i*. Comment peut-on en arriver là ? On peut montrer que cette situation se produit lorsque, à partir du modèle général

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

les coefficients estimés par la méthode des moindres carrés sont tous nuls, à l'exception de la constante. Dans ces conditions,

$$b_2 = b_3 = ... = b_k = 0$$

et on a

$$\hat{y}_i = b_1 = m_y$$

où m_y est la valeur de b_1 telle qu'estimée par la méthode des moindres carrés ordinaires dans ce cas.

En somme, quand le coefficient de détermination est nul, c'est qu'il n'y a pas de relation détectable entre la variable dépendante et les variables indépendantes : quand on estime les

paramètres du modèle, toutes les variables indépendantes disparaissent, à l'exception de la constante, parce qu'elles sont multipliées par un coefficient dont la valeur est estimée à zéro.

Le coefficient de détermination multiple est-il réellement une mesure de similarité, comme nous l'avons affirmé au début ? Il suffit, pour s'en convaincre, de voir que *SSR* est le carré de la distance euclidienne généralisée entre l'ensemble des valeurs observées et l'ensemble des valeurs prédites par le modèle. C'est donc une mesure de *dissimilarité*. Le rapport $\frac{SSR}{SST}$ est donc une mesure de dissimilarité *normée*, dont le domaine de variation s'étend de zéro à un. Par conséquent, $R^2 = 1 - \frac{SSR}{SST} = \frac{SSM}{SST}$ est une mesure de similarité, dont le domaine de variation s'étend aussi de zéro à un.

3-1.4.3 Relation entre R^2 et le coefficient de corrélation simple

On peut montrer que le coefficient de détermination, R^2, est égal au carré du coefficient de corrélation simple entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$:

$$r_{\hat{y}y}^2 = \left(\frac{s_{\hat{y}y}}{s_{\hat{y}} s_y} \right)^2 = R^2$$

3-1.4.4 Coefficient de détermination ajusté

Lorsque les hypothèses classiques sont respectées (ces hypothèses sont définies ci-après), le *coefficient de détermination ajusté*

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}\left(1 - R^2\right) = 1 - \frac{SSR/(n-k)}{SST/(n-1)}$$

est un estimateur non biaisé du « vrai » coefficient de détermination.

Le coefficient de détermination ajusté peut s'interpréter comme une façon de tenir compte du nombre de variables indépendantes dans l'évaluation de la performance d'un modèle. En effet, on peut généralement augmenter le coefficient de détermination R^2 en ajoutant des variables indépendantes au modèle, même si la présence des variables supplémentaires ne s'appuie pas sur une hypothèse théorique.

Or on voit dans la formule du coefficient de détermination corrigé $\overline{R}^2$ que, lorsqu'on ajoute des variables, $\overline{R}^2$ peut diminuer, si R^2 n'augmente pas assez pour compenser l'accroissement de k. Il y a cependant d'autres procédures, plus fiables, pour décider de l'opportunité d'ajouter ou de retrancher telle ou telle de variable : ce sont les tests d'hypothèse.

La valeur du coefficient de détermination ajusté est généralement rapportée dans les tableaux de résultats : voir les tableaux 2 et 4 de Lemelin et Polèse (1995), le tableau 1 de Heikkila *et al.* (1989) ou le tableau 1 de Richardson *et al.* (1990).