De la régression simple et l'analyse de la variance aux modèles linéaires généralisés : synthèse et chronologie

Pierre Dagnelie

Faculté universitaire des Sciences agronomiques
B-5030 Gembloux (Belgique)

pierre@dagnelie.be

Résumé

Ce document présente la filiation qui conduit de la régression simple et de l'analyse de la variance la plus élémentaire aux différentes formes de modèles linéaires : modèle linéaire général, modèle linéaire mixte et modèles linéaires généralisés. Il peut servir d'introduction, pour les personnes peu compétentes en la matière, et de rappel ou de synthèse, pour les personnes plus averties.

À ces dernières personnes, ce document peut suggérer une présentation éventuellement utile dans une optique de vulgarisation, en particulier dans certains enseignements de base et lors de contacts de type « consultation statistique ». Il nous semble en effet que, dans de nombreuses circonstances, les modèles les plus élaborés ont intérêt à être présentés en partant des modèles les plus simples, et non pas de façon trop abrupte.

Summary

This document sketches the relationship between the simplest concepts of regression and analysis of variance and the more elaborate linear models : general linear model, linear mixed model and generalized linear models. It can serve as an introduction for people with little expertise in this field, and as a recall or a synthesis for those more informed.

To these persons, this document may suggest a possibly useful presentation in terms of popularization, especially in some basic courses and during contacts in statistical consulting. It seems to us that in many circumstances, more sophisticated models should be introduced starting with the simplest models, and not in a too sharp way.

Mots-clés

Régression, analyse de la variance, modèle linéaire, modèle mixte, modèle linéaire généralisé.

Keywords

Regression, analysis of variance, linear model, mixed model, generalized linear model, GLM.

1. Introduction

Les modèles linéaires mixtes et généralisés sont largement utilisés à l'heure actuelle, souvent sans que leur position exacte par rapport aux modèles antérieurs, plus simples, soit bien connue. Nous croyons utile dans ces conditions de rappeler la filiation entre, au départ, la régression simple et l'analyse de la variance, et en fin de parcours, les modèles généralisés. Nous traitons ce sujet en considérant successivement :

– la régression simple (paragraphe 2),
– la régression multiple et le modèle linéaire (paragraphe 3),
– l'analyse de la variance (paragraphe 4),
– le modèle linéaire mixte et les modèles linéaires généralisés (paragraphe 5).

Nous renvoyons fréquemment le lecteur qui désire disposer d'informations complémentaires à nos deux tomes de *Statistique théorique et appliquée* [Dagnelie, 2006, 2007], à l'aide de mentions du type « STAT1, § . . . » et « STAT2, § . . . ».

Nous illustrons aussi les différents problèmes envisagés par des exemples simples, mais correspondant à des situations réelles. Il s'agit le plus souvent de sous-ensembles de données, arrondies ou simplifiées, provenant d'exemples présentés dans les mêmes ouvrages et auxquels nous renvoyons par des mentions « STAT1, ex. . . . » et « STAT2, ex. . . . ». Dans chaque cas, nous exposons autant que possible les calculs numériques de façon détaillée, de manière à permettre au lecteur qui le souhaite de suivre pas à pas la « mécanique » sous-jacente.

Bien sûr, d'autres documents peuvent tout autant être consultés. On peut citer notamment les livres de Draper et Smith [1998], Hocking [2003], McCulloch et Searle [2001], et Rencher et Schaalje [2008].

Nous mettons en outre l'accent sur certains problèmes de terminologie et sur quelques éléments historiques, en nous inspirant notamment de l'article de David [1995], des compléments publiés ultérieurement et du document PDF correspondant [David, 2006-2007], ainsi que des sites web de Miller [2008] et Verduin [2008].

Enfin, nous terminons cette présentation par une synthèse chronologique et un récapitulatif de quelques formules (paragraphe 6).

2. La régression simple

2.1. La régression linéaire simple

1° Considérons à titre d'exemple les données suivantes, relatives à une variable indépendante ou explicative x et une variable dépendante ou à expliquer y :

x_i	y_i
1	95
3	83
7	58
14	28

Pour ces données, le but de la régression linéaire simple est de déterminer l'équation de la droite qui, dans le plan $(x\,,\,y)$, s'ajuste le mieux à l'ensemble des quatre points observés. L'équation de cette droite, dite *droite de régression*, est généralement présentée sous la forme :

$$y = a + b\,x\,,$$

les paramètres a et b étant respectivement l'*ordonnée à l'origine* ou *terme indépendant* et le *coefficient de régression*.

La figure 1 présente les points observés et la droite de régression, le détail des calculs étant donné au paragraphe 2.1.5°.

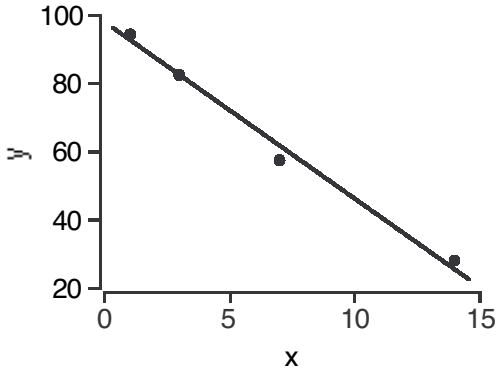


Figure 1

2° Pour un ensemble de valeurs observées $(x_i\,,\,y_i)$, relatives à n individus $(i = 1\,,\,\ldots\,,n)$, on détermine généralement les valeurs de a et b de la manière suivante :

$$a = \bar{y} - b\,\bar{x} \quad \text{et} \quad b = \mathrm{SPE}/\mathrm{SCE}_x\,,$$

$\bar{x}$ et $\bar{y}$ étant les moyennes des deux séries d'observations, tandis que SCE_x et SPE sont respectivement la somme des carrés des écarts par rapport à la moyenne, pour la variable x, et la somme des produits des écarts par rapport aux moyennes, pour les deux variables considérées simultanément.

Cette procédure de détermination des paramètres a et b est connue sous le nom de *méthode des moindres carrés*, car elle revient à rendre minimum la somme des carrés des écarts entre les valeurs observées y_i de la variable dépendante et les valeurs correspondantes $y(x_i)$ de la droite de régression, c'est-à-dire la somme :

$$\sum_{i=1}^{n}[y_i - y(x_i)]^2 \quad \text{ou} \quad \sum_{i=1}^{n}(y_i - a - b\,x_i)^2\,.$$

Dans le cas de la figure 1, il s'agit des écarts, mesurés verticalement, entre chacun des quatre points observés et la droite de régression.

Cette méthode de calcul date de la fin du 18e et du début du 19e siècle. Elle est essentiellement due à Adrien Marie Legendre (1752-1833) et Carl Friedrich Gauss (1777-1855) [Gauss, 1809, 1855 ; Legendre, 1805]. Le concept de régression date, lui, de la fin du 19e siècle, et est dû à Francis Galton (1822-1911) [Galton, 1885, 1886] [1].

[1] Galton a introduit le mot *régression* pour désigner, dans des problèmes d'hérédité, la diminution progressive (« *regression* ») des écarts par rapport à la moyenne, d'une génération à la suivante. Le terme *régression* a ensuite été utilisé d'une manière tout à fait générale, sans aucune relation avec sa signification initiale.

Informations complémentaires : STAT1, § 4.7.2 et 4.9.1.

3° La somme des carrés des écarts qui est minimisée est la *somme des carrés des écarts résiduelle*. On peut montrer qu'elle est aussi :

$$\mathrm{SCE}_{y.x} = \mathrm{SCE}_y - \mathrm{SPE}^2/\mathrm{SCE}_x \quad \text{ou} \quad \mathrm{SCE}_y - b\,\mathrm{SPE}\,,$$

SCE_y étant la somme des carrés des écarts par rapport à la moyenne, pour la variable y.

Informations complémentaires : STAT1, § 4.7.4 et 4.9.1.

4° À l'équation de régression, qui constitue le modèle observé, on peut associer un modèle théorique :

$$y = \alpha + \beta\,x + \epsilon \quad \text{ou} \quad y_i = \alpha + \beta\,x_i + \epsilon_i\,.$$

Les paramètres α et β sont les valeurs théoriques qui correspondent aux valeurs observées a et b, tandis que les symboles ϵ ou ϵ_i désignent les écarts aléatoires entre les valeurs de la variable explicative et les valeurs correspondantes de la droite de régression.

Pour pouvoir utiliser pleinement ce modèle, on doit supposer que les valeurs de la variable explicative sont connues sans erreur, et que les écarts ϵ_i sont des variables aléatoires de moyenne nulle, de même variance, indépendantes les unes des autres, et de distribution normale (distribution de Gauss ou de Laplace-Gauss). Dans ces conditions, on peut résoudre une série de problèmes tels que la détermination de limites de confiance pour les paramètres α et β, la réalisation de tests de signification de ces paramètres, l'estimation de valeurs de la variable dépendante à partir de valeurs de la variable explicative, etc. [2]

Informations complémentaires : STAT1, § 7.3.4 ; STAT2, § 14.1, 14.3, etc.

5° Pour l'exemple considéré ci-dessus, on a successivement :

$$\bar{x} = 6{,}25\,, \quad \bar{y} = 66\,, \quad \mathrm{SCE}_x = 98{,}75 \quad \text{et} \quad \mathrm{SPE} = -508\,,$$

$$a = 98{,}2 \quad \text{et} \quad b = -5{,}14\,,$$

$$y = 98{,}2 - 5{,}14\,x\,,$$

ainsi que :

$$\mathrm{SCE}_y = 2.638 \quad \text{et} \quad \mathrm{SCE}_{y.x} = 24{,}69\,.$$

Informations complémentaires : STAT1, ex. 4.10.1 ; STAT2, ex. 3.6.1.

2.2. La régression par l'origine

1° Dans certains cas, et notamment dans de nombreux problèmes de calibrage, on doit imposer à la droite de régression de passer par l'origine, c'est-à-dire par le point de coordonnées $(0\,,\,0)$. L'équation de la droite est alors :

$$y = b\,x\,.$$

Les données suivantes illustrent une telle situation, les quatre points observés et la droite de régression étant présentés dans la figure 2 :

[2] On peut noter que toutes les conditions citées ne sont pas nécessaires pour tous les problèmes. La condition de normalité par exemple ne s'impose qu'en ce qui concerne les déterminations de limites de confiance et les tests d'hypothèses.

x_i	y_i
4	25
8	50
12	72
16	92

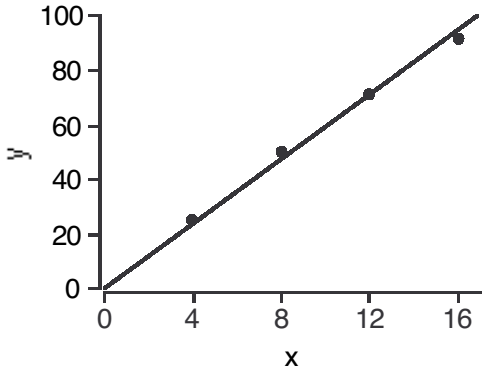


Figure 2

2° Toujours par la méthode des moindres carrés, on peut montrer que, dans ces conditions, le coefficient de régression est :

$$b = \mathrm{SP}/\mathrm{SC}_x\,,$$

si on désigne par SC_x et SP respectivement la somme des carrés des valeurs observées de la variable x et la somme des produits des valeurs observées des deux variables.

Les sommes des carrés et des produits des écarts par rapport aux moyennes (SCE_x et SPE) sont donc remplacées ici par les simples sommes des carrés et des produits (SC_x et SP).

Le même principe s'applique aussi à la somme des carrés des écarts résiduelle, qui est telle que :

$$\mathrm{SCE}_{y.x} = \mathrm{SC}_y - \mathrm{SP}^2/\mathrm{SC}_x \;\text{ ou }\; \mathrm{SC}_y - b\,\mathrm{SP}\,,$$

SC_y étant la somme des carrés des valeurs observées de la variable y.

Informations complémentaires : STAT1, § 4.7.6.3°.

3° Pour les données présentées ci-dessus, on obtient :

$$\mathrm{SC}_x = 480\,, \quad \mathrm{SP} = 2.836 \quad \text{et} \quad b = 5{,}91\,,$$

$$y = 5{,}91\,x\,,$$

et aussi :

$$\mathrm{SC}_y = 16.773 \quad \text{et} \quad \mathrm{SCE}_{y.x} = 16{,}97\,.$$

Informations complémentaires : STAT2, ex. 14.3.4.

2.3. La régression curvilinéaire

1° On appelle *courbes de régression*, et on qualifie de *curvilinéaires*, ou parfois de *non linéaires*, les régressions dont la forme n'est pas à une droite.

Certains types de régression curvilinéaire peuvent être facilement convertis en régression linéaire, par des *changements* ou *transformations de variables*. Tel est le cas par exemple pour la régression exponentielle d'équation :

$$y = c\, e^{bx},$$

à laquelle correspond une courbe, croissante ou décroissante, dont l'asymptote horizontale se confond avec l'axe des abscisses. En remplaçant la variable y par son logarithme, on obtient en effet :

$$\log_e y = a + b\, x \quad \text{ou} \quad y' = a + b\, x\,,$$

si on pose :

$$\log_e c = a\,.$$

2° D'autres types de régression curvilinéaire, telle que la *régression quadratique* ou, d'une manière plus générale, la *régression polynomiale* :

$$y = a + b_1\, x + b_2\, x^2 \quad \text{ou} \quad y = a + b_1\, x + b_2\, x^2 + \ldots + b_p\, x^p,$$

nécessitent le recours à la régression multiple, dont il sera question au paragraphe 3.

Les variables x, x^2, ... sont alors considérées comme étant deux ou plusieurs variables explicatives, qui pourraient tout aussi bien être désignées par x_1, x_2, ... (x_1 au lieu de x, x_2 au lieu de x^2, ...).

3° Dans ces deux premiers types de problèmes, la méthode des moindres carrés donne naissance à des sytèmes d'équations, dites *équations normales*, dans lesquelles les paramètres recherchés a, b, ... apparaissent sous une forme linéaire. Ces sytèmes d'équations sont donc très simples à résoudre.

Dans d'autres situations par contre, les paramètres à déterminer n'apparaissent pas sous une forme linéaire, mais par exemple comme des exposants, dans les équations normales. La résolution du système d'équations implique alors l'utilisation de méthodes de calcul particulières, à caractère itératif.

Il est bon de savoir que l'expression « non linéaire » est souvent utilisée pour désigner cette seule catégorie de régression. Dans cette optique, la qualification « non linéaire » ne concerne pas la forme de la ligne de régression (une courbe, et non pas une droite), mais bien la nature des équations obtenues par la méthode des moindres carrés (équations « non linéaires »). En vue de lever toute ambiguïté, on précise parfois qu'il s'agit d'une régression *non linéaire en les paramètres*.

Informations complémentaires : STAT1, § 4.10 ; STAT2, § 15.2.

3. La régression multiple et le modèle linéaire

3.1. Le cas de deux variables explicatives

1° La régression multiple permet de traiter de nombreux problèmes dans lesquels une variable dépendante doit être exprimée en fonction, non pas d'une seule variable explicative, mais bien de deux ou plusieurs variables explicatives.

Dans le cas particulier de deux variables explicatives x_1 et x_2, l'équation de régression s'écrit :

$$y = a + b_1\, x_1 + b_2\, x_2 \quad \text{ou} \quad y = b_0 + b_1\, x_1 + b_2\, x_2\,.$$

Les données pourraient se présenter par exemple comme suit :

x_{i1}	x_{i2}	y_i
9	52	38
13	70	57
22	78	43
11	83	84
18	68	17

2° En appliquant comme ci-dessus le principe des moindres carrés (paragraphe 2.1.2°), on obtient, pour un ensemble d'observations x_{i1} , x_{i2} et y_i $(i = 1, \ldots, n)$, relatives aux deux variables explicatives et à la variable dépendante :

$$b_0 = \bar{y} - b_1\,\bar{x}_1 - b_2\,\bar{x}_2\,,$$

$$b_1 = \frac{\mathrm{SCE}_2\,\mathrm{SPE}_{1y} - \mathrm{SPE}_{12}\,\mathrm{SPE}_{2y}}{\mathrm{SCE}_1\,\mathrm{SCE}_2 - \mathrm{SPE}_{12}^2} \quad \text{et} \quad b_2 = \frac{\mathrm{SCE}_1\,\mathrm{SPE}_{2y} - \mathrm{SPE}_{12}\,\mathrm{SPE}_{1y}}{\mathrm{SCE}_1\,\mathrm{SCE}_2 - \mathrm{SPE}_{12}^2}\,.$$

Dans ces expressions, $\bar{x}_1$, $\bar{x}_2$ et $\bar{y}$ sont les moyennes des trois variables, SCE_1 et SCE_2 sont les sommes des carrés des écarts relatives aux deux variables explicatives, et les différents symboles SPE désignent les sommes des produits des écarts, en ce qui concerne d'une part les deux variables explicatives considérées simultanément (SPE_{12}), et d'autre part les deux variables explicatives associées chacune à la variable dépendante (SPE_{1y} et SPE_{2y}).

Pour que ces relations puissent être appliquées, il faut toutefois que le dénominateur $\mathrm{SCE}_1\,\mathrm{SCE}_2 - \mathrm{SPE}_{12}^2$ soit différent de zéro, ce qui implique que le coefficient de corrélation r_{12} des variables explicatives doit être différent de -1 et de $+1$.

Le concept de régression double, puis multiple, a été développé par Karl Pearson (1857-1936) à la fin du 19e et au début du 20e siècle [Pearson, 1896 ; Pearson *et al.*, 1903].

Informations complémentaires : STAT2, § 16.2.2.

3° Un modèle théorique, des conditions d'application, et divers problèmes semblables à ceux que nous avons évoqués en matière de régression simple (paragraphe 2.1.4°) pourraient être présentés ici également. Le modèle théorique correspondant à l'équation de régression introduite ci-dessus est par exemple :

$$y = \beta_0 + \beta_1\,x_1 + \beta_2\,x_2 + \epsilon \quad \text{ou} \quad y_i = \beta_0 + \beta_1\,x_{i1} + \beta_2\,x_{i2} + \epsilon_i\,.$$

Informations complémentaires : STAT2, § 16.1.

4° En ce qui concerne l'exemple considéré, on a :

$$\bar{x}_1 = 14{,}6\,, \quad \bar{x}_2 = 70{,}2 \quad \text{et} \quad \bar{y} = 47{,}8\,,$$
$$\mathrm{SCE}_1 = 113{,}2 \quad \text{et} \quad \mathrm{SCE}_2 = 560{,}8\,,$$
$$\mathrm{SPE}_{12} = 106{,}4\,, \quad \mathrm{SPE}_{1y} = -230{,}4 \quad \text{et} \quad \mathrm{SPE}_{2y} = 670{,}2\,,$$

et l'équation de régression :

$$y = -31{,}2 - 3{,}84\,x_1 + 1{,}92\,x_2\,.$$

Dans l'espace à trois dimensions $(x_1\,,\, x_2\,,\, y)$, cette équation est celle d'un plan appelé *plan de régression*.

Quant à la somme des carrés des écarts résiduelle, que nous ne définirons explicitement qu'au paragraphe 3.2.6°, on obtient :

$$\mathrm{SCE}_{y.x} = 287{,}4\,.$$

Informations complémentaires : Dagnelie [1986], ex. 2.4.1.

3.2. Le cas général

1° En vue d'aborder la régression multiple dans le cas général d'un nombre quelconque p de variables explicatives, il s'impose d'avoir recours à des notations matricielles.

Dans cette optique, l'ensemble des variables explicatives $x_1\,,\ x_2\,,\ \ldots,\ x_p$ est tout d'abord regroupé en un vecteur-ligne $\boldsymbol{x}$:

$$\boldsymbol{x} = \left[\, x_1 \;\; x_2 \;\; \ldots \;\; x_p \,\right].$$

Et de même, l'ensemble des coefficients de régression $b_1\,,\ b_2\,,\ \ldots,\ b_p$ est regroupé en un vecteur-colonne $\boldsymbol{b}$:

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}.$$

L'équation de régression linéaire à p variable explicatives :

$$y = b_0 + b_1\, x_1 + b_2\, x_2 + \ldots + b_p\, x_p\,,$$

devient alors :

$$y = b_0 + \boldsymbol{x}\,\boldsymbol{b}\,,$$

soit une forme tout à fait semblable à celle de la régression simple (paragraphe 2.1.1°).

On notera que des notions générales de calcul matriciel peuvent être trouvées dans de nombreux ouvrages, tels que ceux de Graybill [2002], Healy [2000], et Searle [2006], ou encore Dagnelie [1986], Draper et Smith [1998], et Rencher et Schaalje [2008] [3].

2° Les valeurs observées relatives à un ensemble de n individus peuvent être désignées par x_{ij} $(i = 1\,,\ \ldots\,,n$ et $j = 1\,,\ \ldots\,,p)$ pour les p variables explicatives, et y_i $(i = 1\,,\ \ldots\,,n)$ pour la variable dépendante.

Ces données peuvent être réunies en une matrice $\boldsymbol{X}$ d'une part, et un vecteur-colonne $\boldsymbol{y}$ d'autre part :

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \quad \text{et} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Dans la matrice $\boldsymbol{X}$, les différentes lignes sont donc relatives aux différents individus observés, et les différentes colonnes aux différentes variables considérées.

3° Les moyennes des p variables explicatives constituent alors un vecteur-ligne $\bar{\boldsymbol{x}}$:

$$\bar{\boldsymbol{x}} = \left[\, \bar{x}_1 \;\; \bar{x}_2 \;\; \ldots \;\; \bar{x}_p \,\right],$$

semblable au vecteur $\boldsymbol{x}$ des variables elles-mêmes.

De plus, les sommes des carrés des écarts des variables explicatives considérées individuellement SCE_{x_j} $(j = 1\,,\ \ldots\,,p)$, et les sommes des produits des écarts de ces mêmes variables

[3] Pour rappel, en ce qui concerne le problème particulier rencontré ici, le produit d'un vecteur-ligne par un vecteur-colonne s'obtient tout simplement en effectuant la somme des produits des différents éléments du premier vecteur $(x_1\,,\ x_2\,,\ \ldots)$ par les éléments correspondants du deuxième vecteur $(b_1\,,\ b_2\,,\ \ldots)$. Ce principe s'étend facilement au cas du produit d'un vecteur et d'une matrice, ou de deux matrices.

considérées deux à deux $\mathrm{SPE}_{x_j x_{j'}}$ (j et $j' = 1\,,\,\ldots\,,p$ et $j \neq j'$), peuvent être présentées sous la forme d'une matrice carrée symétriques $\boldsymbol{A}_{xx}$:

$$\boldsymbol{A}_{xx} = \begin{bmatrix} \mathrm{SCE}_{x_1} & \mathrm{SPE}_{x_1x_2} & \ldots & \mathrm{SPE}_{x_1x_p} \\ \mathrm{SPE}_{x_1x_2} & \mathrm{SCE}_{x_2} & \ldots & \mathrm{SPE}_{x_2x_p} \\ \vdots & \vdots & & \vdots \\ \mathrm{SPE}_{x_1x_p} & \mathrm{SPE}_{x_2x_p} & \ldots & \mathrm{SCE}_{x_p} \end{bmatrix}.$$

Et de même, les sommes des produits des écarts des différentes variables explicatives et de la variable dépendante $\mathrm{SPE}_{x_j y}$ ($j = 1\,,\,\ldots\,,p$) permettent de constituer un vecteur-colonne $\boldsymbol{a}_{xy}$:

$$\boldsymbol{a}_{xy} = \begin{bmatrix} \mathrm{SPE}_{x_1y} \\ \mathrm{SPE}_{x_2y} \\ \vdots \\ \mathrm{SPE}_{x_py} \end{bmatrix}.$$

4° À titre d'illustration, pour les données à deux variables explicatives considérées aux paragraphes 3.1.1° et 3.1.4°, on a :

$$\boldsymbol{X} = \begin{bmatrix} 9 & 52 \\ 13 & 70 \\ 22 & 78 \\ 11 & 83 \\ 18 & 68 \end{bmatrix} \quad \text{et} \quad \boldsymbol{y} = \begin{bmatrix} 38 \\ 57 \\ 43 \\ 84 \\ 17 \end{bmatrix},$$

et aussi :

$$\bar{\boldsymbol{x}} = \begin{bmatrix} 14{,}6 & 70{,}2 \end{bmatrix} \quad \text{et} \quad \bar{y} = 47{,}8\,,$$

$$\boldsymbol{A}_{xx} = \begin{bmatrix} 113{,}2 & 106{,}4 \\ 106{,}4 & 560{,}8 \end{bmatrix} \quad \text{et} \quad \boldsymbol{a}_{xy} = \begin{bmatrix} -230{,}4 \\ 670{,}2 \end{bmatrix}.$$

5° Sur la base de ces notations, on peut montrer que la méthode des moindres carrés, appliquée au problème de la régression multiple, conduit aux résultats suivants :

$$b_0 = \bar{y} - \bar{\boldsymbol{x}}\,\boldsymbol{b} \quad \text{et} \quad \boldsymbol{b} = \boldsymbol{A}_{xx}^{-1}\,\boldsymbol{a}_{xy}\,,$$

$\boldsymbol{A}_{xx}^{-1}$ étant la matrice inverse de $\boldsymbol{A}_{xx}$. Ces relations sont tout à fait semblables aux résultats relatifs à la régression simple (paragraphe 2.1.2°), qui auraient pu être présentés sous la forme :

$$a = \bar{y} - \bar{x}\,b \quad \text{et} \quad b = \mathrm{SCE}_x^{-1}\,\mathrm{SPE}\,.$$

Au même titre qu'en régression simple et en régression à deux variables explicatives, les valeurs SCE_x d'une part et $\mathrm{SCE}_1\,\mathrm{SCE}_2 - \mathrm{SPE}_{12}^2$ d'autre part doivent être différentes de zéro, il faut ici que le déterminant de $\boldsymbol{A}_{xx}$ soit non nul, pour pouvoir inverser cette matrice. Ceci implique qu'aucune des variables explicatives ne peut être exactement une fonction linéaire d'une ou plusieurs autres variables explicatives.

6° Dans les mêmes conditions, la somme des carrés des écarts résiduelle est :

$$\mathrm{SCE}_{y.x} = a_{yy} - \boldsymbol{a}'_{xy}\,\boldsymbol{A}_{xx}^{-1}\,\boldsymbol{a}_{xy} \;\text{ ou }\; a_{yy} - \boldsymbol{a}'_{xy}\,\boldsymbol{b}\,,$$

a_{yy} étant la somme des carrés des écarts de la variable dépendante, et le vecteur-ligne $\boldsymbol{a}'_{xy}$ étant la transposée du vecteur-colonne $\boldsymbol{a}_{xy}$. Ces expressions sont tout à fait comparables aux expressions correspondantes relatives à la régression simple (paragraphe 2.1.3°) :

$$\mathrm{SCE}_{y.x} = \mathrm{SCE}_y - \mathrm{SPE}\,\mathrm{SCE}_x^{-1}\,\mathrm{SPE} \quad \text{ou} \quad \mathrm{SCE}_y - \mathrm{SPE}\,b\,.$$

7° Ici également, tout ce qui a été dit antérieurement en matière de modèle théorique, de conditions d'application, etc. (paragraphes 2.1.4° et 3.1.3°) peut s'appliquer. En particulier, le modèle théorique est :

$$y = \beta_0 + \boldsymbol{x}\,\boldsymbol{\beta} + \epsilon \quad \text{ou} \quad y_i = \beta_0 + \boldsymbol{x}_i\,\boldsymbol{\beta} + \epsilon_i\,.$$

Informations complémentaires : STAT2, § 16.3.2.

8° Les différentes relations matricielles que nous avons présentées permettent de retrouver les résultats numériques du paragraphe 3.1.4°. On obtient successivement :

$$\boldsymbol{A}_{xx}^{-1} = \begin{bmatrix} 0{,}010751 & -0{,}002040 \\ -0{,}002040 & 0{,}002170 \end{bmatrix},$$

$$\boldsymbol{b} = \begin{bmatrix} 0{,}010751 & -0{,}002040 \\ -0{,}002040 & 0{,}002170 \end{bmatrix} \begin{bmatrix} -230{,}4 \\ 670{,}2 \end{bmatrix} = \begin{bmatrix} -3{,}844 \\ 1{,}924 \end{bmatrix},$$

$$b_0 = 47{,}8 - \begin{bmatrix} 14{,}6 & 70{,}2 \end{bmatrix} \begin{bmatrix} -3{,}844 \\ 1{,}924 \end{bmatrix} = -31{,}2\,,$$

$$y = -31{,}2 + \begin{bmatrix} -3{,}844 \\ 1{,}924 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} = -31{,}2 - 3{,}84\,x_1 + 1{,}92\,x_2\,,$$

ainsi que :

$$\mathrm{SCE}_{y.x} = 2.462{,}8 - \begin{bmatrix} -230{,}4 & 670{,}2 \end{bmatrix} \begin{bmatrix} -3{,}844 \\ 1{,}924 \end{bmatrix} = 287{,}4\,.$$

3.3. La régression sans terme indépendant : le modèle linéaire

1° Comme en régression simple (paragraphe 2.2), on peut envisager le cas de la régression sans terme indépendant, c'est-à-dire le cas d'un plan (pour deux variables explicatives) ou d'un hyperplan (pour plus de deux variables explicatives) passant par l'origine.

Ici aussi, les sommes des carrés et des produits des écarts par rapport aux moyennes doivent être remplacées par les simples sommes des carrés et des produits des valeurs observées, qu'on peut désigner par SC_{x_j}, $\mathrm{SP}_{x_j x_{j'}}$ et $\mathrm{SP}_{x_j y}$, au lieu de SCE_{x_j}, $\mathrm{SPE}_{x_j x_{j'}}$ et $\mathrm{SPE}_{x_j y}$. Et ces différentes sommes des carrés et des produits peuvent être exprimées très simplement en fonction des données initiales.

On peut en effet montrer facilement que, pour les variables explicatives, le produit $\boldsymbol{X'X}$ de la transposée de la matrice des données par cette matrice elle-même réunit l'ensemble des sommes des carrés SC_{x_j} et des produits $\mathrm{SP}_{x_j x_{j'}}$. Et de même, quand les variables explicatives sont associées à la variable dépendante, le produit matriciel $\boldsymbol{X'y}$ est constitué des sommes des produits $\mathrm{SP}_{x_j y}$.

Les produits matriciels $\boldsymbol{X'X}$ et $\boldsymbol{X'y}$ remplacent donc respectivement la matrice $\boldsymbol{A}_{xx}$ et le vecteur $\boldsymbol{a}_{xy}$, de telle sorte que la deuxième relation du paragraphe 3.2.5° devient :

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}\,.$$

Il faut évidemment, ici aussi, que le déterminant de $\boldsymbol{X'X}$ ne soit pas nul.

2° Une propriété essentielle de cette formulation est le fait que, par la régression multiple sans terme indépendant, on peut traiter aussi tous les problèmes de régression linéaire, simple ou multiple, avec terme indépendant. Il suffit, pour ce faire, d'associer au terme indépendant b_0 une variable particulière x_0, parfois qualifiée de variable instrumentale ou pseudo-variable, qui est égale à 1 pour tous les individus considérés.

Le modèle théorique du paragraphe 3.2.7° devient alors :

$$y = \boldsymbol{x}\,\boldsymbol{\beta} + \epsilon \quad \text{ou} \quad y_i = \boldsymbol{x}_i\,\boldsymbol{\beta} + \epsilon_i\,,$$

ou encore :

$$\boxed{\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}}\,,$$

$\boldsymbol{y}$ étant le vecteur des valeurs de la variable dépendante, $\boldsymbol{X}$ la matrice des valeurs des variables explicatives (y compris les valeurs 1 de la pseudo-variable x_0), $\boldsymbol{\beta}$ le vecteur des coefficients de régression (y compris le terme indépendant β_0), et $\boldsymbol{\epsilon}$ le vecteur des résidus aléatoires.

En outre, en fonction de ce qui a été vu au paragraphe 3.2.6°, la somme des carrés des écarts résiduelle est :

$$\text{SCE}_{y.x} = \boldsymbol{y'\,y} - \boldsymbol{y'X}\,(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \;\text{ ou }\; \boldsymbol{y'\,y} - \boldsymbol{y'Xb} \;\text{ ou }\; \boldsymbol{y'}\,(\boldsymbol{y} - \boldsymbol{Xb})\,,$$

$\boldsymbol{y'}$ étant la transposée de $\boldsymbol{y}$, et le produit $\boldsymbol{y'\,y}$ étant la somme des carrés des valeurs observées de la variable dépendante.

3° L'approche présentée ici a l'avantage d'aboutir à un traitement unique de tous les cas de régression linéaire. Elle est souvent présentée *ex abrupto*, sous le nom de *modèle linéaire* ou *modèle linéaire général*, sans passer par les rappels que nous avons développés au cours des paragraphes précédents.

La terminologie est cependant assez floue, l'expression « modèle linéaire général » étant parfois associée à l'utilisation de la régression multiple en analyse de la variance, ou associée à la méthode des moindres carrés généralisés, dont il est question au paragraphe suivant, ou limitée au cas où on étudie simultanément, non pas seulement deux ou plusieurs variables explicatives, mais aussi deux ou plusieurs variables dépendantes. Nous reviendrons sur ce point dans la synthèse finale (paragraphe 6.1.2°).

Informations complémentaires : STAT2, § 16.3.3.

4° Nous reprenons une fois encore les données du paragraphe 3.1.1°, en y ajoutant une pseudo-variable x_0 uniformément égale à 1.

On a alors :

$$\boldsymbol{X} = \begin{bmatrix} 1 & 9 & 52 \\ 1 & 13 & 70 \\ 1 & 22 & 78 \\ 1 & 11 & 83 \\ 1 & 18 & 68 \end{bmatrix} \quad \text{et} \quad \boldsymbol{y} = \begin{bmatrix} 38 \\ 57 \\ 43 \\ 84 \\ 17 \end{bmatrix},$$

et aussi :

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 5 & 73 & 351 \\ 73 & 1.179 & 5.231 \\ 351 & 5.231 & 25.201 \end{bmatrix} \quad \text{et} \quad \boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 239 \\ 3.259 \\ 17.448 \end{bmatrix}.$$

On peut vérifier que les éléments de cette matrice et de ce vecteur sont bien les sommes des carrés et des produits des valeurs des différentes variables. En particulier, le premier élément de la matrice $\boldsymbol{X}'\boldsymbol{X}$ est le nombre d'individus considérés, tandis que les autres éléments de la première ligne et de la première colonne de cette matrice sont les sommes des valeurs observées des variables explicatives. De même, le premier élément du vecteur $\boldsymbol{X}'\boldsymbol{y}$ est la somme des valeurs observées de la variable dépendante.

On en déduit ensuite :

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 9{,}005172 & -0{,}013773 & -0{,}122565 \\ -0{,}013773 & 0{,}010751 & -0{,}002040 \\ -0{,}122565 & -0{,}002040 & 0{,}002170 \end{bmatrix},$$

$$\boldsymbol{b} = \begin{bmatrix} 9{,}005172 & -0{,}013773 & -0{,}122565 \\ -0{,}013773 & 0{,}010751 & -0{,}002040 \\ -0{,}122565 & -0{,}002040 & 0{,}002170 \end{bmatrix} \begin{bmatrix} 239 \\ 3.259 \\ 17.448 \end{bmatrix} = \begin{bmatrix} -31{,}16 \\ -3{,}844 \\ 1{,}924 \end{bmatrix},$$

$$y = -31{,}2 - 3{,}84\, x_1 + 1{,}92\, x_2\,,$$

ainsi que :

$$\mathrm{SCE}_{y.x} = \begin{bmatrix} 38 & 57 & 43 & 84 & 17 \end{bmatrix} \left(\begin{bmatrix} 38 \\ 57 \\ 43 \\ 84 \\ 17 \end{bmatrix} - \begin{bmatrix} 1 & 9 & 52 \\ 1 & 13 & 70 \\ 1 & 22 & 78 \\ 1 & 11 & 83 \\ 1 & 18 & 68 \end{bmatrix} \begin{bmatrix} -31{,}16 \\ -3{,}844 \\ 1{,}924 \end{bmatrix} \right) = 287{,}4\,.$$

Ces résultats sont bien identiques à ceux qui ont été obtenus aux paragraphes 3.1.4° (régression à deux variables explicatives) et 3.2.8° (cas général).

3.4. Les moindres carrés généralisés

1° Nous avons toujours supposé jusqu'à présent que les résidus ϵ_i de la régression sont de moyenne nulle, de même variance, indépendants les uns des autres, et de distribution normale, ces conditions étant celles de la méthode initiale des moindres carrés, dite aussi des *moindres carrés ordinaires* (paragraphes 2.1.2° et 2.1.4°).

Les conditions d'égalité des variances et d'indépendance ou de non-corrélation des résidus peuvent être exprimées également en considérant que les variances et les covariances des résidus constituent une matrice qui se présente de la manière suivante :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix},$$

σ^2 étant la variance commune des résidus relatifs aux différents individus observés.

On peut s'écarter de ces conditions en admettant que la matrice $\boldsymbol{\Sigma}$ est quelconque, mais symétrique par rapport à sa diagonale descendante, ce qui conduit à la notion de *moindres carrés généralisés* [4].

2° Dans le cas des moindres carrés ordinaires, la matrice $\boldsymbol{\Sigma}$ est souvent présentée aussi sous la forme :

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad \text{ou} \quad \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I},$$

le symbole $\boldsymbol{I}$ désignant d'une manière générale une matrice identité, constituée de valeurs 1 dans la diagonale descendante et 0 en dehors de cette diagonale.

Par analogie, en ce qui concerne les moindres carrés généralisés, la matrice $\boldsymbol{\Sigma}$ peut être considérée comme étant telle que :

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{M}.$$

La relation du paragraphe 3.3.1° relative au calcul des coefficients de régression, ou des paramètres du modèle linéaire, devient alors :

$$\boldsymbol{b} = (\boldsymbol{X}' \boldsymbol{M}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{M}^{-1} \boldsymbol{y}\,,$$

pour autant que la matrice $\boldsymbol{M}$ soit connue.

Cette notion de moindres carrés généralisés est essentiellement due à Aitken [1935].

3° Différentes formes particulières des matrices $\boldsymbol{\Sigma}$ et $\boldsymbol{M}$ ont été considérées. Nous en présentons deux.

On peut tout d'abord supposer que les résidus sont toujours indépendants, comme dans la régression tout à fait classique, mais que des poids différents doivent être attribués aux différentes observations, car celles-ci ne sont pas toutes connues avec la même précision. On considère alors, en général, des poids w_i inversément proportionnels aux variances σ_i^2 des différentes observations, ce qui conduit à la forme suivante de la matrice $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1/w_1 & 0 & \dots & 0 \\ 0 & 1/w_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1/w_n \end{bmatrix}.$$

Il s'agit là de la *régression pondérée*.

Une deuxième situation particulière relativement importante correspond à la matrice :

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}.$$

[4] La matrice $\boldsymbol{\Sigma}$, de même que la matrice $\boldsymbol{M}$ dont il est question quelques lignes plus loin, doit aussi être telle que le carré de tout élément extérieur à la diagonale descendante est inférieur ou égal au produit des éléments diagonaux correspondants (matrice semi-définie positive). Cette restriction est comparable à celle qui concerne toute covariance, dont le carré est toujours inférieur ou égal au produit des variances correspondantes [STAT1, § 4.5.2.2°].

Ce type de structure, dite de *symétrie composée*, suppose que les variances des différents résidus sont égales, et que les différents résidus sont éventuellement corrélés entre eux, mais avec une même valeur commune ρ du coefficient de corrélation.

Informations complémentaires : STAT2, § 16.5.2.

4° Nous reprenons à titre d'illustration les données du paragraphe 2.2.1°, relatives au cas d'une droite de régression passant par l'origine. Dans de telles situations, les variances des observations sont souvent d'autant plus élevées qu'on s'éloigne de l'origine.

On peut considérer par exemple que les variances sont proportionnelles aux valeurs de la variable explicative, les poids w_i devant alors être inversément proportionnels à ces valeurs. Si on suppose en outre que les résidus sont indépendants les uns des autres, la matrice $\mathbf{\Sigma}$ est, pour les données considérées :

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 16 \end{bmatrix} \quad \text{ou} \quad \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

Sans entrer dans le détail des calculs, qui sont extrêmement simples, en raison du fait que la matrice $\boldsymbol{X}$ ne comprend qu'une seule colonne, on obtient dans ces conditions l'équation de régression pondérée :

$$y = 5{,}97\, x\,.$$

Comme on pouvait s'y attendre, ce résultat est quelque peu différent de celui du paragraphe 2.2.3°, à savoir [5] :

$$y = 5{,}91\, x\,.$$

4. L'analyse de la variance

4.1. L'analyse de la variance à un critère de classification

1° L'*analyse de la variance à un critère de classification* (ou un facteur) a essentiellement pour but de comparer les moyennes d'un certain nombre de populations.

Pour trois populations et, respectivement, trois, trois et deux observations par population, les données de départ peuvent être par exemple :

Populations		
1	2	3
44	37	12
63	60	46
76	74	–

[5] On peut noter que, dans ce cas très particulier, la droite de régression pondérée est simplement la droite qui passe, non seulement par l'origine, mais aussi par le point moyen $(\bar{x}\,,\,\bar{y})$, c'est-à-dire par le point de coordonnées (10 , 59,75) [STAT1, § 4.7.6.4°].

Le principe de l'analyse est de subdiviser la somme des carrés des écarts totale, relative à l'ensemble de toutes les observations, en une composante factorielle, liée aux différences existant entre les populations, et une composante résiduelle, relative à la variabilité existant au sein des différentes populations, et de comparer ensuite ces deux composantes.

La somme des carrés des écarts totale est calculée par rapport à la moyenne générale de toutes les observations, tandis que la somme des carrés des écarts résiduelle est déterminée par rapport aux moyennes des différentes séries d'observations. Et la somme des carrés des écarts factorielle peut être obtenue par simple différence.

Les résultats de l'analyse sont généralement présentés sous la forme d'un *tableau d'analyse de la variance*. Il s'agit du tableau suivant pour les données qui figurent ci-dessus :

Sources de variation	DL	SCE	CM	F	P
Variation factorielle	2	1.374	687	1,91	0,24
Variation résiduelle	5	1.794	359		
Variation totale	7	3.168			

Dans ce tableau, les sigles DL, SCE, CM, F et P désignent respectivement les nombres de degrés de liberté, les sommes des carrés des écarts, les carrés moyens (analogues à des variances), la valeur observée d'une variable de Fisher-Snedecor, et la probabilité correspondant à cette valeur.

Dans sa version la plus simple, l'analyse de la variance date des années 1920, et est due à Ronald Aylmer Fisher (1890-1962) [Fisher, 1918 ; Fisher et Mackenzie, 1923] [6].

Informations complémentaires : STAT2, § 9.2 et 9.3, et ex. 9.2.1 et 9.3.1.

2° À la décomposition de la somme des carrés des écarts totale en deux parties, on peut associer le modèle théorique suivant :

$$y_{gh} - \mu = \alpha_g + \epsilon_{gh} \quad \text{ou} \quad y_{gh} = \mu + \alpha_g + \epsilon_{gh}\,.$$

Dans ce modèle, le symbole y_{gh} désigne les observations (le premier indice étant relatif aux populations et le deuxième indice aux observations de chacune des populations), les différences $y_{gh} - \mu$ sont les écarts par rapport à la moyenne générale μ, les α_g sont les écarts liés aux différences existant entre les populations (composante factorielle), et les ϵ_{gh} sont les écarts résiduels, à l'intérieur des différentes populations (composante résiduelle).

En vue d'effectuer en toute rigueur le test d'égalité des moyennes, éventuellement accompagné de procédures complémentaires (détermination de limites de confiance pour les moyennes et les différences de moyennes, comparaisons multiples des moyennes, etc.), il faut supposer que les écarts résiduels ϵ_{gh} possèdent les mêmes propriétés de nullité de la moyenne, d'égalité des variances, d'indépendance et de normalité qu'en matière de régression (paragraphe 2.1.4°).

3° Dans le modèle théorique :

$$y_{gh} = \mu + \alpha_g + \epsilon_{gh}\,,$$

l'idée initiale est de considérer les écarts factoriels α_g comme des valeurs fixes, non entachées de fluctuations aléatoires.

[6] Dans les premiers travaux de Fisher, la comparaison des carrés moyens était basée sur la méthode de l'erreur-standard. Ultérieurement seulement, cette comparaison a fait intervenir des distributions particulières, dites distributions z, puis les distributions F de Fisher-Snedecor [Fisher, 1924 ; Snedecor, 1934].

Mais en réalité, dans certaines situations, les termes α_g doivent au contraire être considérés comme aléatoires. Tel est le cas quand les différentes populations qui sont comparées dans l'analyse de la variance ont été choisies au hasard dans un ensemble plus vaste, dont elles sont représentatives.

En vue par exemple de chiffrer la variabilité des productions en fourrage d'un vaste ensemble de prairies, dans une région donnée, on peut être amené à limiter l'observation à un petit nombre de prairies, au sein de chacune desquelles un petit nombre d'échantillons de fourrage seront prélevés et analysés. Les deux niveaux de variation sont, dans ces conditions, les différences entre prairies, et les différences entre échantillons, à l'intérieur des diverses prairies considérées.

Dans l'optique d'un choix aléatoire des prairies dans lesquelles des observations seront réalisées, parmi l'ensemble de toutes les prairies envisagées au départ, les composantes α_g doivent évidemment être considérées comme aléatoires.

La réalisation pratique de l'analyse de la variance ne diffère pas selon qu'on considère les termes α_g comme fixes ou comme aléatoires, mais l'interprétation des résultats est par contre sensiblement différente d'un cas à l'autre. Dans le premier cas, le problème est réellement un problème de comparaison des moyennes, tandis que dans le deuxième cas, il s'agit en fait d'un problème d'estimation de variances, ou de composantes de la variance (estimation de la variance entre prairies et de la variance entre échantillons, dans l'exemple qui vient d'être cité).

En outre, des hypothèses supplémentaires, de nullité de la moyenne et de normalité des α_g , et d'indépendance des α_g et des ϵ_{gh} , doivent être introduites dans le deuxième cas.

La distinction entre ces deux types d'analyse de la variance à un critère de classification a été mise en évidence par EISENHART [1947], qui a introduit les expressions *modèle I* et *modèle II*, devenues ultérieurement *modèle fixe* (ou à effets fixes) et *modèle aléatoire* (ou à effets aléatoires).

Informations complémentaires : STAT2, § 9.3.

4° Quel que soit le modèle envisagé, l'analyse de la variance à un critère de classification peut facilement être traitée comme un cas particulier de régression mutiple sans terme indépendant (paragraphe 3.3).

En vue d'être concret, nous considérons uniquement les données présentées ci-dessus. Pour les huit observations en question, le modèle théorique peut s'écrire de manière détaillée de la façon suivante :

$$\begin{cases} y_{11} = \mu + \alpha_1 + \epsilon_{11} \\ y_{12} = \mu + \alpha_1 + \epsilon_{12} \\ y_{13} = \mu + \alpha_1 + \epsilon_{13} \\ y_{21} = \mu + \alpha_2 + \epsilon_{21} \\ y_{22} = \mu + \alpha_2 + \epsilon_{22} \\ y_{23} = \mu + \alpha_2 + \epsilon_{23} \\ y_{31} = \mu + \alpha_3 + \epsilon_{31} \\ y_{32} = \mu + \alpha_3 + \epsilon_{32} \, , \end{cases}$$

la valeur α_1 intervenant pour les trois observations relatives à la première population, la valeur α_2 pour les trois observations relatives à la deuxième population, et la valeur α_3 pour les deux observations relatives à la troisième population.

Le tout peut être présenté aussi comme suit, sous forme matricielle :

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}.$$

Cette écriture est strictement équivalente à celle du paragraphe 3.3.2° :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,,$$

moyennant les conventions :

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{et} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}.$$

Le vecteur $\boldsymbol{\beta}$ ainsi défini réunit les différents paramètres du modèle, à savoir la moyenne générale et les quantités qui mesurent l'importance des différences existant entre les populations, tandis que la matrice $\boldsymbol{X}$ contient les valeurs d'autant de variables qu'il y a de paramètres, les valeurs 1 indiquant les observations pour lesquelles interviennent les différents paramètres. Les variables qui constituent cette matrice sont des *variables indicatrices*, et la matrice elle-même est souvent appelée *matrice d'incidence* ou *matrice de l'expérience*.

Un tel traitement de l'analyse de la variance à un critère de classification comme un problème de régression multiple s'inscrit dans la ligne du *modèle linéaire* ou *modèle linéaire général*, dont il a été question aux paragraphes 3.3.2° et 3.3.3°. Cette approche, basée sur la méthode des moindres carrés ordinaires, apparaît initialement dans le livre d'Anderson et Bancroft [1952], qui est d'ailleurs partiellement sous-titré « *Analysis of experimental models by least squares* ».

Informations complémentaires : STAT2, § 16.4.1 et 16.4.2.

5° En vue de compléter l'information, nous poursuivons l'examen de l'exemple envisagé ci-dessus, en en esquissant le traitement numérique.

Avant toute chose, il faut noter que la matrice d'incidence $\boldsymbol{X}$ présentée ci-dessus est telle que chacune de ses colonnes peut être exprimée comme une fonction linéaire des trois autres colonnes. Il en résulte que le produit $\boldsymbol{X}'\boldsymbol{X}$ ne peut pas être inversé.

Pour remédier à cette situation, il suffit d'éliminer par exemple la dernière colonne de la

matrice $\boldsymbol{X}$ et le dernier élément du vecteur $\boldsymbol{\beta}$. Les données du problème sont alors :

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 44 \\ 63 \\ 76 \\ 37 \\ 60 \\ 74 \\ 12 \\ 46 \end{bmatrix} \quad \text{et} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}.$$

En appliquant exactement la même procédure qu'au paragraphe 3.3.4°, on obtient, pour la régression multiple, une somme des carrés des écarts résiduelle $\mathrm{SCE}_{y.x}$ strictement égale à la somme des carrés des écarts résiduelle de l'analyse de la variance, à savoir 1.794 (paragraphe 4.1.1°). La régression multiple permet ainsi de reconstituer entièrement le tableau d'analyse de la variance, avec une parfaite concordance entre les deux approches.

4.2. L'analyse de la variance à deux critères de classification

1° L'*analyse de la variance à deux critères de classification* peut être considérée comme une généralisation de l'analyse à un critère, permettant de comparer les moyennes d'un ensemble de populations, en tenant compte de deux facteurs susceptibles d'influencer ces moyennes, et non plus d'un seul facteur.

Une telle situation se présente par exemple quand un certain nombre de méthodes d'analyse chimique sont comparées entre elles en fonction des résultats obtenus pour un certain nombre d'échantillons, chacun des échantillons considérés étant soumis à chacune des méthodes d'analyse. Des résultats extrêmement simples pourraient être, pour trois méthodes d'analyse et deux échantillons seulement :

	Méthodes		
Éch.	1	2	3
1	33	29	38
2	23	24	24

Le principe de l'analyse de la variance est ici de subdiviser la somme des carrés des écarts totale en trois composantes : une somme des carrés des écarts factorielle pour chacun des deux facteurs (méthodes d'analyse et échantillons) et une somme des carrés des écarts résiduelle. On obtient ainsi le tableau :

Sources de variation	DL	SCE	CM	F	P
Méthodes d'analyse	2	21,0	10,5	1,03	0,49
Échantillons	1	140,2	140,2		
Variation résiduelle	2	20,3	10,2		
Variation totale	5	181,5			

Informations complémentaires : STAT2, § 10.2 et 10.3, et ex. 10.3.6.

2° Le modèle théorique suivant peut être associé à l'analyse de la variance qui vient d'être esquissée :

$$y_{gh} = \mu + \alpha_g + \beta_h + \epsilon_{gh}\,,$$

μ étant, comme précédemment (paragraphe 4.1.2°), une moyenne générale, tandis que les termes α_g et β_h sont relatifs aux effets de l'un et l'autre des deux facteurs étudiés, et les ϵ_{gh} sont toujours des écarts résiduels. D'une manière plus générale, le modèle peut également contenir en outre un terme d'interaction, quand deux ou plusieurs observations sont disponibles pour chacune des combinaisons des deux facteurs.

Qu'il y ait ou qu'il n'y ait pas de terme d'interaction, trois possibilités doivent être envisagées. D'une part, dans certains problèmes, les α_g et les β_h doivent être considérés, les uns et les autres, comme fixes. D'autre part, les α_g et les β_h doivent parfois être considérés, les uns et les autres, comme aléatoires. Enfin, dans certaines situations, ces termes sont les uns fixes et les autres aléatoires.

Selon les cas, on parle respectivement d'un modèle fixe (ou à effets fixes), d'un modèle aléatoire (ou à effets aléatoires), et d'un *modèle mixte* (ou à effets mixtes). Le problème que nous avons envisagé ci-dessus est de type mixte, les trois méthodes d'analyse étant bien définies, tandis que les deux échantillons analysés sont représentatifs de tous les échantillons qui auraient pu ou qui pourraient être soumis à l'analyse.

La réalisation de l'analyse de la variance, à l'exclusion toutefois du calcul des valeurs F et des probabilités correspondantes, ne diffère pas d'un cas à l'autre. Le calcul de ces valeurs, par contre, ainsi que les conditions d'application et l'interprétation des résultats, dépendent du modèle considéré.

Pour être complet, nous devons mentionner en outre l'existence de modèles hiérarchisés, dans lesquels les deux critères de classification n'interviennent pas sur pied d'égalité, mais sont subordonnés l'un à l'autre. Ces modèles sont le plus souvent aléatoires ou mixtes.

La distinction entre modèle fixe, modèle aléatoire et modèle mixte apparaît en particulier dans le livre de Mood [1950].

Informations complémentaires : STAT2, § 10.3 et 10.5.

3° Dans les différents cas envisagés, les problèmes d'analyse de la variance peuvent être traités par la régression multiple ou le modèle linéaire, selon les mêmes principes que ceux qui ont été présentés aux paragraphes 4.1.4° et 4.1.5°.

Par analogie avec ce que nous avons fait pour un seul critère de classification, le modèle théorique relatif au problème particulier considéré ci-dessus peut s'écrire de la manière suivante :

$$\begin{cases} y_{11} = \mu + \alpha_1 + \beta_1 + \epsilon_{11} \\ y_{12} = \mu + \alpha_1 + \beta_2 + \epsilon_{12} \\ y_{21} = \mu + \alpha_2 + \beta_1 + \epsilon_{21} \\ y_{22} = \mu + \alpha_2 + \beta_2 + \epsilon_{12} \\ y_{31} = \mu + \alpha_3 + \beta_1 + \epsilon_{31} \\ y_{32} = \mu + \alpha_3 + \beta_2 + \epsilon_{12}\,. \end{cases}$$

La présentation matricielle correspondante est :

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}.$$

Ici également, les différents paramètres du modèle sont réunis en un seul vecteur, et dans la matrice d'incidence, la première colonne est relative à la moyenne générale μ, les trois colonnes

suivantes aux trois termes α_g qui correspondent aux effets du facteur méthodes d'analyse, et les deux dernières colonnes aux deux termes β_h d'effets du facteur échantillons.

Informations complémentaires : STAT2, § 16.4.3.

4° Dans l'optique de la résolution numérique du problème, comme au paragraphe 4.1.5°, la matrice $\boldsymbol{X}$ doit être simplifiée, pour permettre l'inversion du produit $\boldsymbol{X}'\boldsymbol{X}$. On peut supprimer par exemple la quatrième et la sixième colonne de cette matrice, de manière à ne conserver que deux paramètres (α_1 et α_2) pour les trois variantes du facteur méthodes d'analyse, et un paramètre (β_1) pour les deux variantes du facteur échantillons. Les données sont alors :

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 33 \\ 23 \\ 29 \\ 24 \\ 38 \\ 24 \end{bmatrix} \quad \text{et} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \end{bmatrix}.$$

Selon les mêmes principes que précédemment, on obtient, pour la somme des carrés des écarts résiduelle relative à la régression multiple, une valeur identique à celle de l'analyse de la variance, soit 20,3 (paragraphe 4.2.1°).

En outre, on peut éliminer de la matrice $\boldsymbol{X}$, dans un premier temps tout ce qui concerne le facteur méthodes d'analyse, et dans un deuxième temps ce qui concerne le facteur échantillons, en considérant ainsi successivement les matrices :

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{et} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

On obtient alors, pour ces deux situations particulières des sommes des carrés des écarts résiduelles respectivement égales à :

$$41{,}3 \quad \text{et} \quad 160{,}5\,.$$

On peut en déduire les sommes des carrés des écarts relatives aux deux facteurs, en soustrayant de chacune de ces valeurs la composante résiduelle :

$$41{,}3 - 20{,}3 = 21{,}0 \quad \text{et} \quad 160{,}5 - 20{,}3 = 140{,}2\,.$$

Les différentes régressions multiples permettent donc, ici également, de reconstituer l'ensemble du tableau d'analyse de la variance.

4.3. L'analyse de la variance à trois et plus de trois critères de classification

De nombreux modèles d'analyse de la variance existent également pour trois ou plus de trois facteurs ou critères de classification. Ces modèles peuvent être facilement convertis en problèmes de régression multiple, du moins quand les nombres d'observations relatifs aux différentes populations sont égaux.

D'une manière générale aussi, les conditions d'application classiques restent inchangées. Il s'agit toujours de conditions de nullité des moyennes, d'égalité des variances, d'indépendance et de normalité, en ce qui concerne les écarts résiduels et, éventuellement aussi, les facteurs aléatoires.

Des problèmes particuliers se posent toutefois, pour deux et pour plus de deux critères de classification, d'une part dans le cas d'effectifs inégaux, et d'autre part en présence d'écarts résiduels non indépendants. Des effectifs inégaux peuvent résulter notamment de l'existence de données manquantes, et des écarts résiduels non indépendants peuvent provenir, entre autres, du fait que des observations sont effectuées successivement, à différents moments, sur les mêmes individus (observations successives ou observations répétées dans le temps, dites aussi « données longitudinales »).

Informations complémentaires : STAT2, § 11.2 à 11.4 et 16.4.5.

5. Diverses extensions du modèle linéaire général

5.1. Le modèle linéaire mixte

1° Le *modèle linéaire mixte* permet de séparer, dans le modèle linéaire général, les éléments fixes et les éléments aléatoires. La présentation la plus courante est du type :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\,\boldsymbol{\gamma} + \boldsymbol{\epsilon}\,.$$

Dans cette relation, la matrice $\boldsymbol{X}$ et le vecteur $\boldsymbol{\beta}$ jouent le même rôle que précédemment (paragraphes 3.3.2° et 4.1.4°), mais uniquement pour les composantes fixes du modèle, tandis que la matrice $\boldsymbol{Z}$ et le vecteur $\boldsymbol{\gamma}$ jouent un rôle analogue pour les composantes aléatoires du modèle, le vecteur $\boldsymbol{\epsilon}$ étant toujours un vecteur de résidus aléatoires. Tout comme le vecteur $\boldsymbol{\epsilon}$, le vecteur $\boldsymbol{\gamma}$ est donc constitué de variables aléatoires.

Dans l'optique des moindres carrés généralisés (paragraphe 3.4), on suppose que les différents éléments du vecteur $\boldsymbol{\epsilon}$ sont caractérisés par une matrice de variances et covariances $\boldsymbol{\Sigma}$, le plus souvent de l'un ou l'autre type particulier. Mais on suppose aussi que les différents éléments de $\boldsymbol{\gamma}$ sont caractérisés par une autre matrice de variances et covariances $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$, et que les éléments de $\boldsymbol{\gamma}$ sont indépendants des éléments de $\boldsymbol{\epsilon}$.

2° Le modèle linéaire mixte a été mis en œuvre dès les années 1950, essentiellement dans le domaine de la génétique animale [Henderson, 1953 ; Henderson *et al.*, 1959]. Il n'a toutefois connu une utilisation plus générale qu'au cours des années 1990, en relation avec le développement de nouvelles procédures de calcul dans le cadre des logiciels statistiques les plus importants. La procédure MIXED du logiciel SAS en est un exemple [SAS, 1992].

Le modèle linéaire mixte est souvent qualifié simplement de « modèle mixte », ce qui peut prêter à confusion, par comparaison avec le concept plus ancien que nous avons présenté au paragraphe 4.2.2°, en matière d'analyse de la variance.

L'utilisation du modèle linéaire mixte soulève, par rapport aux modèles classiques d'analyse de la variance, un certain nombre de difficultés particulières, tant en ce qui concerne l'estimation des différents paramètres que la réalisation des tests d'hypothèses. Mais ce modèle permet de faire face à des problèmes tels que ceux qui ont été évoqués au paragraphe 4.3 (données manquantes et observations successives). Des informations peuvent être trouvées à ce sujet dans les articles de Littell [2002], McLean *et al.* [1991], et Piepho *et al.* [2003], et dans les livres de Demidenko [2004], McCulloch et Searle [2001], et Verbeke et Molenberghs [1997].

3° L'exemple considéré au cours du paragraphe 4.2 permet d'illustrer de façon extrêmement simple les notations introduites ci-dessus. Pour cet exemple, en séparant le facteur fixe (méthodes d'analyse) du facteur aléatoire (échantillons), on a en effet :

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{et} \quad \boldsymbol{\gamma} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

les matrices $\boldsymbol{X}$ et $\boldsymbol{Z}$ provenant de la scission en deux parties de la matrice d'incidence du paragraphe 4.2.3°, tandis que les vecteurs $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$ proviennent de la division en deux parties du vecteur des paramètres du même paragraphe. Le modèle complet est donc :

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}.$$

4° Quant aux données numériques de cet exemple (paragraphe 4.2.1°), l'application du modèle linéaire mixte conduit à des résultats identiques à ceux de l'analyse de la variance classique, en ce qui concerne le facteur fixe (méthodes d'analyse), auquel on s'intéresse principalement. Cette stricte équivalence des résultats est liée au caractère équilibré des données (même nombre d'observations, égal à 1, pour les différentes combinaisons des deux facteurs).

Par contre, des résultats différents sont obtenus par les deux méthodes quand les nombres d'observations sont inégaux. À titre d'illustration, on peut supposer par exemple que la dernière observation du paragraphe 4.2.1° est manquante.

Dans ces conditions, l'analyse de la variance classique conduit, après estimation de la donnée manquante [STAT2, § 10.4.3], à une valeur F égale à 5,41 et une probabilité correspondante égale à 0,29 . Au contraire, les *packages* LME4 et NLME du logiciel R (<www.r-project.org>), de même que la procédure MIXED du logiciel SAS (<www.sas.com>), fournissent, pour le modèle linéaire mixte, une valeur F égale à 3,17 et une probabilité égale à 0,37 . Mais aucune conclusion valable ne peut être déduite de la différence existant entre ces deux résultats, notamment en raison du très petit nombre d'observations à partir desquelles les calculs sont réalisés.

5.2. Le modèle linéaire généralisé

1° Le modèle linéaire classique :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}\,,$$

dont il a été question notamment aux paragraphes 3.3.2° et 4.1.4°, exprime le fait que les différentes composantes du vecteur $\boldsymbol{y}$ sont égales au produit matriciel $\boldsymbol{X\beta}$, moyennant l'adjonction de valeurs aléatoires $\boldsymbol{\epsilon}$, qui sont de moyennes nulles. On peut donc affirmer aussi qu'en moyenne, les composantes du vecteur $\boldsymbol{y}$ sont égales à $\boldsymbol{X\beta}$:

$$\boldsymbol{\mu_y} = \boldsymbol{X\beta}\,.$$

Le *modèle linéaire généralisé* a pour principe de considérer que ce ne sont pas les moyennes des différentes composantes de $\boldsymbol{y}$ qui sont égales à $\boldsymbol{X\beta}$, mais bien une certaine fonction de ces moyennes :

$$\boldsymbol{f}(\boldsymbol{\mu_y}) = \boldsymbol{X\beta}\,.$$

La fonction considérée f, qui doit être monotone (croissante ou décroissante), est communément appelée *fonction de lien*.

Le modèle linéaire généralisé permet d'analyser ainsi des données qui sont éventuellement caractérisées par d'autres distributions que les distributions normales (distributions binomiales, distributions de Poisson, etc.).

Les changements de variables qui apparaissent du fait de l'utilisation de la fonction f peuvent faire penser aux transformations de variables dont il a été question au paragraphe 2.3.1°, en

matière de régression curvilinéaire. On notera cependant que les deux approches sont sensiblement différentes, dans la mesure où les écarts résiduels $\boldsymbol{\epsilon}$ n'interviennent pas de la même manière dans les deux cas. D'une façon générale, les deux approches ne conduisent donc pas aux mêmes résultats.

2° Le modèle linéaire généralisé est dû à Nelder et Wedderburn [1972]. Comme pour le modèle linéaire mixte, le recours au modèle généralisé n'a cependant connu une expansion importante qu'au cours des années 1990, en relation avec le développement de nouvelles procédures de calcul, telle que la procédure GENMOD de SAS [Hilbe, 1994].

Une présentation simple du modèle linéaire généralisé est donnée par Myers et Montgomery [1997], et d'autres informations peuvent être trouvées dans les livres de Dobson [2002], Lindsey [1997], et McCulloch et Searle [2001].

3° Parmi les problèmes qui peuvent être traités à l'aide du modèle linéaire généralisé, figure notamment l'étude de la mortalité d'un certain nombre d'individus (des insectes par exemple) en fonction de doses croissantes de l'une ou l'autre substance. Classiquement, ces problèmes sont abordés par la méthode des probits ou des logits (régression logistique), en vue notamment de la détermination de doses effectives ou létales (doses létales 50 par exemple) [STAT2, § 15.5].

Dans l'optique du modèle linéaire généralisé, les fonctions de lien correspondantes sont, respectivement pour les probits et les logits :

$$f(\mu_y) = \Phi^{-1}(\mu_y) \quad \text{et} \quad f(\mu_y) = \log[\mu_y/(1-\mu_y)]\,,$$

Φ^{-1} étant la fonction inverse de la fonction de répartition (fonction cumulative) de la distribution normale réduite.

4° Les données suivantes, qui expriment des pourcentages de mortalité en fonction de doses d'insecticide, illustrent ce type de situation :

x_i	y_i
2	6
3	23
4	68
5	88
6	98

La figure 3 présente les points observés correspondants, ainsi que la courbe qu'on peut y ajuster par le modèle linéaire généralisé, et son asymptote supérieure.

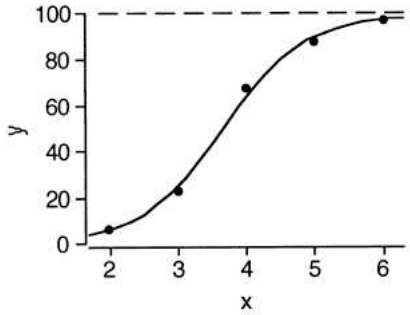


Figure 3

Si on pose :

$$y' = y/100\,,$$

l'équation de cette courbe est :

$$\log_e[y'/(1-y')] = -6{,}06 + 1{,}66\,x \quad \text{ou} \quad y' = 1/(1+e^{6{,}06-1{,}66x})\,.$$

Cette équation a été obtenue à l'aide de la fonction *glm* du logiciel R [X, 2008] [7].

Les résultats qu'on peut déduire de cette équation sont très semblables à ceux que donnent aussi les méthodes classiques des probits et des logits.

Informations complémentaires : STAT2, ex. 15.5.1 et 15.5.2.

5.3. Le modèle linéaire mixte généralisé

Le même principe qu'au paragraphe précédent peut être appliqué au modèle linéaire mixte (paragraphe 5.1), qui peut s'écrire :

$$\boldsymbol{\mu_y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}\,.$$

On obtient alors le *modèle linéaire mixte généralisé* :

$$\boldsymbol{f}(\boldsymbol{\mu_y}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}\,.$$

Ce modèle est donc à la fois une extension du modèle mixte et du modèle généralisé. Il apparaît dans la publication de Gilmour *et al.* [1985], et il est présenté notamment par Engel et Keen [1994], et Molenberghs *et al.* [2002].

6. Synthèse

6.1. Synthèse chronologique

1° Le schéma récapitulatif de la page suivante constitue une sythèse qui positionne :

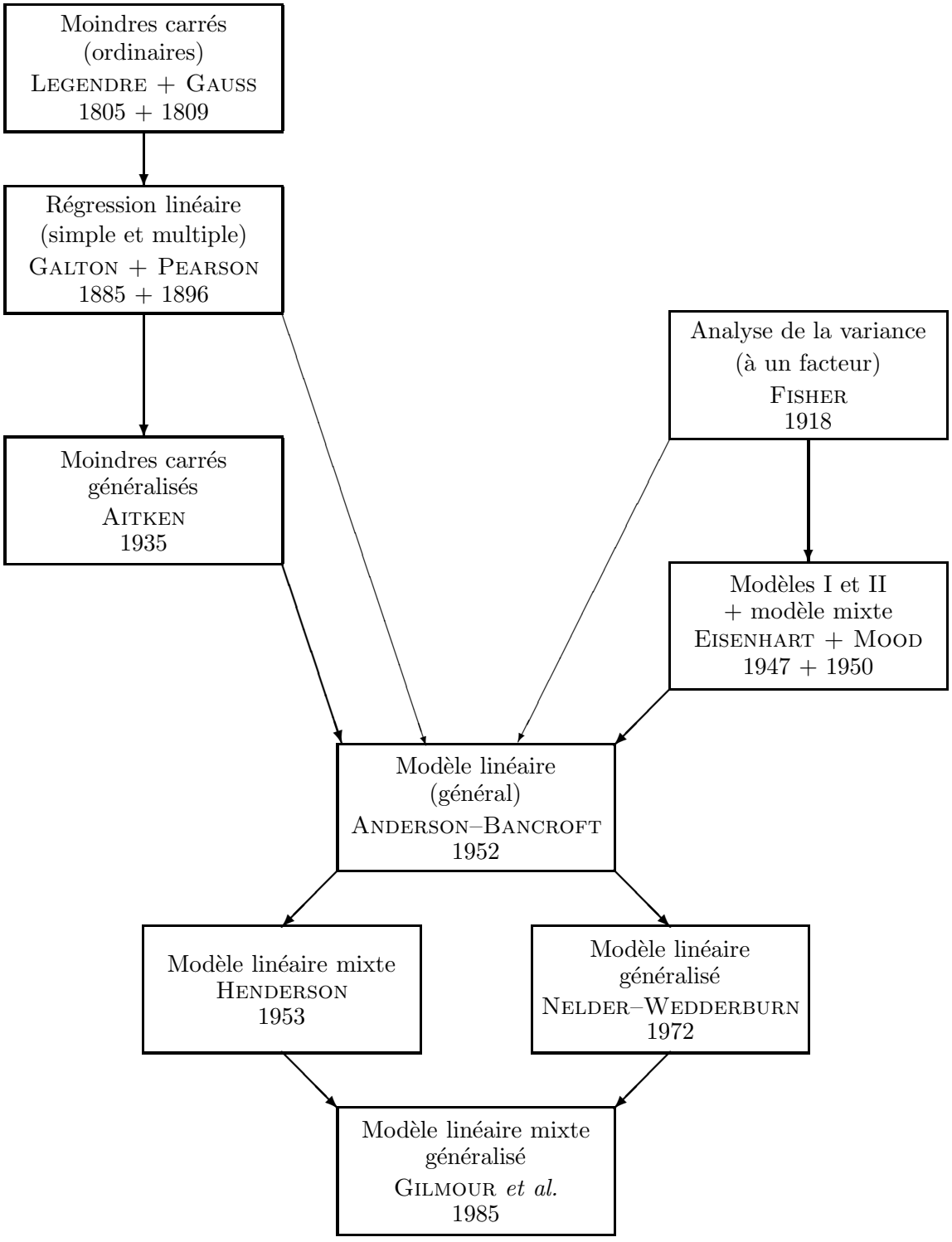
– dans la branche « régression » :
  - les moindres carrés classiques (moindres carrés ordinaires),
  - la régression classique (simple et multiple),
  - les moindres carrés généralisés ;

– dans la branche « analyse de la variance » :
  - l'analyse de la variance la plus simple,
  - les modèles I et II (fixe et aléatoire), rapidement suivis du modèle mixte ;

– dans la « descendance » de ces deux branches :
  - le modèle linéaire ou linéaire général,
  - le modèle linéaire mixte,
  - le modèle linéaire généralisé,
  - le modèle linéaire mixte généralisé.

[7] En vue d'éviter toute confusion, on notera que le sigle « glm » (ou « GLM ») est relatif, selon les auteurs, soit au modèle linéaire général, soit au modèle linéaire généralisé. Pour le logiciel R, ce sigle concerne le modèle linéaire généralisé, alors que pour les logiciels Minitab et SAS, il concerne le modèle linéaire général.

Schéma récapitulatif



Les noms et les dates qui apparaissent dans ce schéma concernent dans chaque cas les premières publications que nous avons mentionnées. Ces informations sont données à titre indicatif, en fonction notamment et dans l'optique des listes de *First (?) occurrence of common terms* de David [1995, 2006-2007], dont l'auteur dit bien « *Establishing a first occurrence is hazardous ; hence the question mark in our title* ».

2° La position que nous attribuons au cadre « modèle linéaire » mérite quelques commentaires supplémentaires.

On peut en effet considérer que le modèle linéaire fait partie intégrante de la branche « régression », que ce soit en association avec les moindres carrés ordinaires, auquel cas il n'y a guère de différences entre « régression multiple » et « modèle linéaire », ou en association avec les moindres carrés généralisés. Nous préférons considérer que le modèle linéaire constitue la jonction entre les deux branches, régression et analyse de la variance.

D'autre part, nous faisons converger quatre flèches vers le cadre « modèle linéaire », car on peut considérer soit, dans une optique élémentaire, que la simple application de la régression multiple en analyse de la variance à un critère de classification constitue déjà un modèle linéaire (flèches plus fines), soit au contraire, dans une optique plus large, que le concept de modèle linéaire englobe nécessairement les moindres carrés généralisés et les différents modèles d'analyse de la variance (flèches plus grosses).

3° On remarquera en particulier l'importance des développements qui ont entouré l'année 1950, depuis les concepts de base de modèles I et II (modèles fixe et aléatoire) d'analyse de la variance, en 1947, jusqu'au modèle linéaire mixte en 1953, et cela avec application du principe des moindres carrés ordinaires dans le premier cas et des moindres carrés généralisés dans le deuxième cas.

Comme nous l'avons signalé, on notera aussi que l'utilisation pratique des derniers modèles envisagés a été largement dépendante, jusqu'au début des années 1990, du développement de nouvelles procédures de calcul dans le cadre des principaux logiciels statistiques.

6.2. Récapitulation de quelques formules

1° L'énumération des formules relatives aux différents cas considérés constitue une autre synthèse utile. Ces formules sont successivement, en ce qui concerne les modèles proprement dits :

– pour la régression simple classique :

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

– pour la régression simple par l'origine :

$$y_i = \beta\, x_i + \epsilon_i\,,$$

– pour la régression multiple classique :

$$y_i = \beta_0 + \boldsymbol{x}_i\, \boldsymbol{\beta} + \epsilon_i\,,$$

– pour la régression multiple sans terme indépendant et le modèle linéaire (général) :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \;\; (\text{ou } \boldsymbol{\mu_y} = \boldsymbol{X}\boldsymbol{\beta}),$$

– pour le modèle linéaire mixte :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\,\gamma} + \boldsymbol{\epsilon} \text{ (ou } \boldsymbol{\mu_y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}),$$

– pour le modèle linéaire généralisé :

$$\boldsymbol{f}(\boldsymbol{\mu_y}) = \boldsymbol{X\beta}\,,$$

– et pour le modèle linéaire mixte généralisé :

$$\boldsymbol{f}(\boldsymbol{\mu_y}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}\,.$$

2° Dans les cas les plus simples, les formules de détermination des coefficients de régression b ou $\boldsymbol{b}$, qui sont aussi les formules d'estimation des valeurs théoriques correspondantes β et $\boldsymbol{\beta}$, sont :

– pour la régression simple classique :

$$b = \mathrm{SPE}/\mathrm{SCE}_x \;\text{ ou }\; \mathrm{SCE}_x^{-1}\,\mathrm{SPE}\,,$$

– pour la régression simple par l'origine :

$$b = \mathrm{SP}/\mathrm{SC}_x \;\text{ ou }\; \mathrm{SC}_x^{-1}\,\mathrm{SP}\,,$$

– pour la régression multiple classique :

$$\boldsymbol{b} = \boldsymbol{A}_{xx}^{-1}\,\boldsymbol{a}_{xy}\,,$$

– pour la régression multiple sans terme indépendant et le modèle linéaire (général) :

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}\,,$$

– et dans le cas des moindres carrés généralisés :

$$\boldsymbol{b} = (\boldsymbol{X'}\,\boldsymbol{M}^{-1}\boldsymbol{X})^{-1}\,\boldsymbol{X'}\,\boldsymbol{M}^{-1}\,\boldsymbol{y}\,.$$

3° Enfin, comme les sommes des carrés des écarts résiduelles constituent le principal point de passage entre la régression et l'analyse de la variance, il peut être intéressant d'en rappeler également les différentes formulations. On a successivement :

– pour la régression simple classique :

$$\mathrm{SCE}_{y.x} = \mathrm{SCE}_y - \mathrm{SPE}^2/\mathrm{SCE}_x \;\text{ ou }\; \mathrm{SCE}_y - b\,\mathrm{SPE}\,,$$

– pour la régression simple par l'origine :

$$\mathrm{SCE}_{y.x} = \mathrm{SC}_y - \mathrm{SP}^2/\mathrm{SC}_x \;\text{ ou }\; \mathrm{SC}_y - b\,\mathrm{SP}\,,$$

– pour la régression multiple classique :

$$\mathrm{SCE}_{y.x} = a_{yy} - \boldsymbol{a}'_{xy}\,\boldsymbol{A}_{xx}^{-1}\,\boldsymbol{a}_{xy} \;\text{ ou }\; a_{yy} - \boldsymbol{a}'_{xy}\,\boldsymbol{b}\,,$$

– et pour la régression multiple sans terme indépendant et le modèle linéaire (général) :

$$\mathrm{SCE}_{y.x} = \boldsymbol{y'\,y} - \boldsymbol{y'X}\,(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \;\text{ ou }\; \boldsymbol{y'\,y} - \boldsymbol{y'Xb}\,.$$

Références bibliographiques

Les différents sites web qui sont mentionnés ont été consultés
en dernier lieu le 23.09.2008.

Aitken A.C. [1935]. On least squares and linear combinations of observations. *Proc. R. Soc. Edinb.* **55**, 42-48.

Anderson R.L., Bancroft T.A. [1952]. *Statistical theory in research.* New York, McGraw-Hill, 399 p.

Dagnelie P. [1986]. *Analyse statistique à plusieurs variables.* Gembloux, Presses agronomiques, 362 p.

Dagnelie P. [2006]. *Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions.* Bruxelles, De Boeck et Larcier, 734 p.

Dagnelie P. [2007]. *Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique.* Bruxelles, De Boeck et Larcier, 511 p.

David H.A. [1995]. First(?) occurrence of common terms in mathematical statistics. *Amer. Stat.* **49** (2), 121-133.

David H.A. [2006-2007]. *First (?) occurrence of common terms in statistics and probability.* Document PDF, <www.stat.iastate.edu/preprint/articles/2006-07_original.pdf>, 34 p.

Demidenko E. [2004]. *Mixed models : theory and applications.* New York, Wiley, 704 p.

Dobson A.J. [2002]. *An introduction to generalized linear models.* Boca Raton, Chapman and Hall/ CRC, 225 p.

Draper N.R., Smith H. [1998]. *Applied regression analysis.* New York, Wiley, 706 p.

Eisenhart C. [1947]. The assumptions underlying the analysis of variance. *Biometrics* **3** (1), 1-21.

Engel B., Keen A. [1994]. A simple approach for the analysis of generalized linear mixed models. *Stat. Neerl.* **48** (1), 1-22.

Fisher R.A. [1918]. The causes of human variability. *Eugen. Rev.* **10**, 213-220.

Fisher R.A. [1924]. On a distribution yielding the error functions of several well known statistics. *Proceedings of the International Congress of Mathematics* (Toronto) 2, 805-813 ; <digital.library.adelaide.edu.au/dspace/handle/2440/15183>.

Fisher R.A., Mackenzie W.A. [1923]. Studies in crop variation. II. The manurial response of different potato varieties. *J. Agric. Sci.* **13**, 311-320 ; <digital.library.adelaide.edu.au/dspace/handle/2440/15179>.

Galton F. [1885]. Opening address. Section H. Anthropology. *Nature* **32**, 507-510 [8].

Galton F. [1886]. Regression towards mediocrity in hereditary stature. *J. Anthrop. Inst.* **15**, 246-263 [8].

Gauss C.F. [1809]. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Hamburg, Perthes et Besser ; <gallica.bnf.fr/ark:/12148/bpt6k3357v.notice>.

Gauss C.F. [1855]. *Méthode des moindres carrés : mémoires sur la comparaison des observations* (trad. J. Bertrand) [9]. Paris, Mallet-Bachelier ; <gallica.bnf.fr/ark:/12148/bpt6k996041.notice>.

Gilmour A.R., Anderson R.D., Rae A.L. [1985]. The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72** (3), 593-599.

Graybill F.A. [2002]. *Matrices with applications in statistics.* Belmont, Brooks/Cole, 461 p.

Healy M.J.R. [2000]. *Matrices for statistics.* Oxford, University Press, 147 p.

Henderson C.R. [1953]. Estimation of variance and covariance components. *Biometrics* **9** (2), 226-252.

Henderson C.R., Kempthorne O., Searle S.R., von Krosigk C.M. [1959]. The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15** (2), 192-218.

[8] Les deux publications de Galton étaient disponibles sur internet à l'adresse <www.galton.org> en décembre 2007, mais ce site n'était pas accessible au moment de mettre la dernière main à ce travail (septembre 2008).

[9] Y compris la traduction de la partie de Gauss [1809] consacrée à la méthode des moindres carrés.

Hilbe J.M. [1994]. Generalized linear models. *Amer. Stat.* **48** (3), 255-265.

Hocking R.R. [2003]. *Methods and applications of linear models.* New York, Wiley, 741 p.

Legendre A.M. [1805]. *Nouvelles méthodes pour la détermination des orbites des comètes* [10]. Paris, Didot ; <imgbase-scd-ulp.u-strasbg.fr/displayimage.php?album=417&pos=0>.

Lindsey J.K. [1997]. *Applying generalized linear models.* New York, Springer, 256 p.

Littell R.C. [2002]. Analysis of unbalanced mixed model data : a case study comparison of ANOVA versus REML/GLS. *J. Agric. Biol. Environ. Stat.* **7** (4), 472-490.

McCulloch C.E., Searle R. [2001]. *Generalized, linear, and mixed models.* New York, Wiley, 325 p.

McLean R.A., Sanders W.L., Stroup W.W. [1991]. A unified approach to mixed linear models. *Amer. Stat.* **45** (1), 54-64.

Miller J. [2008]. *Earliest known uses of some of the words of mathematics.* <members.aol.com/jeff570/>.

Molenberghs G., Renard D., Verbeke G. [2002]. A review of generalized linear mixed models. *J. Soc. Franç. Stat.* **143** (1-2), 53-78.

Mood A.M. [1950]. *Introduction to the theory of statistics.* New York, McGraw-Hill, 433 p.

Myers R.H., Montgomery D.C. [1997]. A tutorial on generalized linear models. *J. Qual. Technol.* **29** (3), 274-291.

Nelder J.A., Wedderburn R.W.M. [1972]. Generalized linear models. *J. R. Stat. Soc., Ser. A*, **135** (3), 370-384.

Pearson K. [1896]. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Phil. Trans. R. Soc., Ser. A*, **187**, 253-318.

Pearson K., Yule G.U., Blanchard N., Lee A. [1903]. The law of ancestral heredity. *Biometrika* **2** (2), 211-236.

Piepho H.P., Büchse A., Emrich K. [2003]. A hitchhiker's guide to mixed models for randomized experiments. *J. Agron. Crop Sci.* **189** (5), 310-322.

Rencher A.C., Schaalje G.B. [2008]. *Linear models in statistics.* New York, Wiley, 672 p.

SAS [1992]. *SAS technical report P-229, SAS/STAT software : changes and enhancements, release 6.07.* Cary, SAS Institute Inc., 620 p.

Searle S.R. [2006]. *Matrix algebra useful for statistics.* New York, Wiley, 472 p.

Snedecor G.W. [1934]. *Calculation and interpretation of analysis of variance and covariance.* Ames, Collegiate Press, 96 p.

Verbeke G., Molenberghs G. [1997]. *Linear mixed models in practice : a SAS-oriented approach.* New York, Springer, 306 p.

Verduin K. [2008]. *A short history of probability and statistics.* <www.leidenuniv.nl/fsw/verduin/stathist/stathist.htm>.

X [2008]. *R : a language and environment for statistical computing – reference index (version 2.7.2).* Vienna, R Foundation for Statistical Computing, document PDF, <www.r-project.org>, 2727 p.

Remerciements

Nous tenons à remercier tout particulièrement MM. Florent Duyme et François Piraux (Arvalis – Institut du Végétal, Boigneville, France) pour les remarques et les suggestions qu'ils ont formulées à propos d'une version provisoire de ce document.

[10] Y compris un « *Appendice* » de 9 pages « *Sur la méthode des moindres quarrés* ».