1.
Régression linéaire et analyse de variance

L'objet d'une régression linéaire est d'explorer les relations entre une variable quantitative Y (traditionnellement dénommée « variable à expliquer » ou encore « variable dépendante ») et une série de variables X_1, X_2, ..., X_p (dénommées, elles, « variables explicatives », ou encore « variables indépendantes »). La principale vertu de cette méthode est de permettre, par l'intermédiaire d'un *modèle mathématique*, d'évaluer la force de l'association entre Y et chacun des X_i, les autres variables explicatives étant maintenues à un niveau constant. Dans une telle situation, on dit couramment que l'on *ajuste* sur les X_j ($j \neq i$).

Le concept d'ajustement est central en modélisation statistique. S'il trouve ses origines dans l'analyse de variance — un cas particulier de régression linéaire — on le retrouve dans la régression logistique, l'analyse de variance multivariée (MANOVA) ou encore le modèle de Cox.

Dans ce chapitre, nous aborderons successivement :

- en introduction, une discussion des liens reliant régression, prédiction, ajustement, et analyse de variance ;
- une série d'exemples concrets ;
- quelques observations, notamment sur la généralisation du coefficient de corrélation et celle du test t ;
- enfin, nous verrons qu'il est utile de savoir vérifier les conditions de validité d'une régression linéaire au moyen d'un « diagnostic de régression ».

Introduction

Régression, prédiction et ajustement

Si l'on considère trois variables (Y, X_1, X_2) et leurs mesures correspondantes $(y_i, x_{1i}, x_{2i})_{1 \leq i \leq n}$, effectuer une « régression » linéaire de Y à partir de X_1 et X_2, c'est rechercher a_0, a_1 et a_2 tels que $y_i \approx a_0 + a_1 x_{1i} + a_2 x_{2i}$ [1]. Géométriquement, cela équivaut à chercher le plan passant « au mieux » par les points (y_i, x_{1i}, x_{2i}), comme indiqué sur la figure 1.1.

---

[1] Où « ≈ » signifie « approximativement égal à ». Le choix du terme « régression » est purement historique. Il proviendrait d'un des premiers usages de cette technique par Galton, en 1889, à propos d'une étude sur l'hérédité dont le titre était : « *Law of universal regression* ».

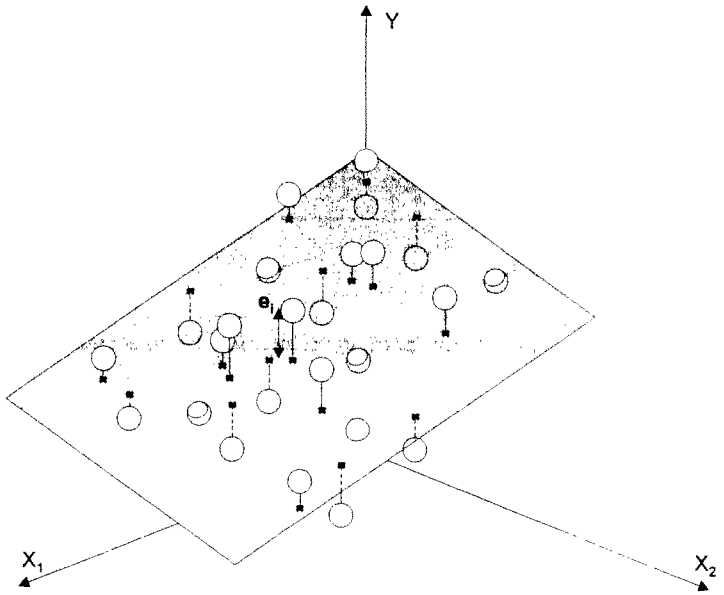

Fig. 1.1 — *Recherche du plan passant « au mieux » par les points (y_i, x_{1i}, x_{2i}) représentés par des motifs « O » sur la figure ; ce plan minimise la somme des carrés des distances d'un point à sa projection parallèlement à Y (motif « x » sur la figure) ; cette distance est notée e_i.*

Formellement, ce plan de régression est celui qui minimise la somme des carrés des distances e_i séparant les points (y_i, x_{1i}, x_{1i}) de leur projection parallèlement à l'axe Y (2). Cela revient à minimiser la somme des e_i^2 où :

$$y_i = a_0 + a_1x_{1i} + a_2x_{2i} + e_i \quad (3).$$

Au premier abord, une régression linéaire apparaît donc avant tout comme un outil utile pour prédire une variable (Y) à partir d'autres variables (ici X_1 et X_2). Pourtant, en biométrie, il est bien rare que cette technique soit utilisée à cette fin. Le plus souvent, elle sert en effet à rechercher une liaison entre Y et X_1 alors que le niveau de X_2 est constant, en d'autres termes à réaliser un ajustement sur X_2 (on peut, de façon symétrique, rechercher une liaison entre Y et X_2 à X_1 constant).

Comment la relation de régression $y_i = a_0 + a_1x_{1i} + a_2x_{2i} + e_i$ introduite plus haut, permet-elle de réaliser un tel ajustement ?

2 Il aurait été sûrement plus intuitif de minimiser la somme des distances entre les points et leur projection, plutôt que la somme des carrés des distances. Encore une fois, ce sont en partie des impératifs calculatoires qui ont imposé cette dernière approche.

3 Le point de coordonnées $(a_0 + a_1x_{1i} + a_2x_{2i}, x_{1i}, x_{2i})$ correspond à la projection parallèlement à Y de (y_i, x_{1i}, x_{2i}) sur le plan P d'équation $y = a_0 + a_1x_1 + a_2x_2$; $e_i = [y_i - (a_0 + a_1x_1 + a_2x_2)]$ est donc bien la distance séparant un point (y_i, x_{1i}, x_{1i}) de sa projection sur P.

Une première réponse à cette question nécessite de faire des hypothèses à propos des e_i ou plus précisément à propos de la variable ε, dont les e_i sont des réalisations. Si ε est indépendante de X_1 et X_2 et si ε a une espérance égale à 0, alors a_1 traduira bel et bien la force de la relation entre Y et X_1 à X_2 constant. En effet, si l'on a :

$$Y = a_0 + a_1X_1 + a_2X_2 + \varepsilon,$$

alors pour X_2 constant égal à x, si l'on recherche la variation (Y-Y') de Y correspondant à une augmentation de X_1 de 1, on a :

$$Y = a_0 + a_1X_1 + a_2x + \varepsilon \qquad (a),$$
$$Y' = a_0 + a_1(X_1 + 1) + a_2x + \varepsilon' \qquad (b),$$

et en effectuant (b) – (a) on obtient :

$$Y' - Y = a_1 + \varepsilon' - \varepsilon.$$

Si ε est indépendante de X_1 et X_2 alors l'espérance $\varepsilon' - \varepsilon$ de $E(\varepsilon' - \varepsilon)$ vaut 0 et donc finalement :

$$E(Y' - Y) = a_1.$$

Le coefficient a_1 est donc égal à la variation moyenne de Y correspondant à une augmentation de X_1 de 1 quand X_2 est maintenu constant. Ce coefficient traduit donc bel et bien la force de la relation entre Y et X_1 à X_2 constant.

Venons-en maintenant à quelques considérations de géométrie. Sur la figure 1.1 nous constatons que la relation « naturelle » entre Y et X_1, que l'on observe visuellement en projetant les points (y_i, x_{1i}, x_{1i}) sur le plan (Y, X_1) parallèlement à X_2, est sensiblement moins forte que la relation entre Y et X_1 à X_2 constant. Pour essayer d'éliminer l'effet de X_2 sur la liaison de Y avec X_1, cherchons la projection sur le plan (Y, X_1) parallèlement à une droite D appartenant au plan (Y, X_2), qui donne des points projetés aussi alignés que possible (voir fig. 1.2). On peut montrer (4) que D est définie par l'intersection du plan de régression ci-dessus avec le plan (Y, X_2). Ajuster, c'est donc projeter. On tente d'annuler l'effet de X_2 sur la relation entre Y et X_1, en transformant par projection les points (y_i, x_{1i}, x_{2i}) en points de coordonnées $(y'_i = y_i - a_2x_{2i}, x_{1i}, 0)$, où la relation des y'_i avec les x_{1i} est aussi forte que possible (les « points projetés sont aussi alignés que possible »).

4 En effet, une régression plane consiste à chercher a_0, a_1 et a_2 tels que $y_i = a_0 + a_1x_{1i} + a_2x_{2i} + e_i$ avec la somme des e_i^2 minimum (et la moyenne des e_i nulle). Cela revient à chercher a_2 puis (a_1, a_0) tels que $(y_i - a_2x_{2i}) = a_0 + a_1x_{1i} + e_i$ avec toujours somme des e_i^2 minimum ; l'application $(y_i, x_{1i}, x_{2i}) \rightarrow (y_i - a_2x_{2i}, x_1, 0)$ est bien une projection sur (Y, X_1) ; et si $u_i = y_i - a_2x_{2i}$, $u_i = a_0 + a_1x_{1i} + e_i$ est bien la droite qui passe « au mieux » par les points (u_i, x_{1i}).

En pratique, si les variables sont effectivement reliées par une relation linéaire, l'ajustement trouvera bien la liaison de Y avec X_1, à X_2 constant. Si tel n'est pas le cas, la régression n'aboutira pas à un résultat faux, elle sera seulement moins efficace, X_2 n'étant plus véritablement maintenue constante.

En faisant une régression, on ne fixe donc que rarement X_2, on essaie seulement de minimiser son influence dans la relation liant Y à X_1 en s'appuyant sur une équation linéaire.

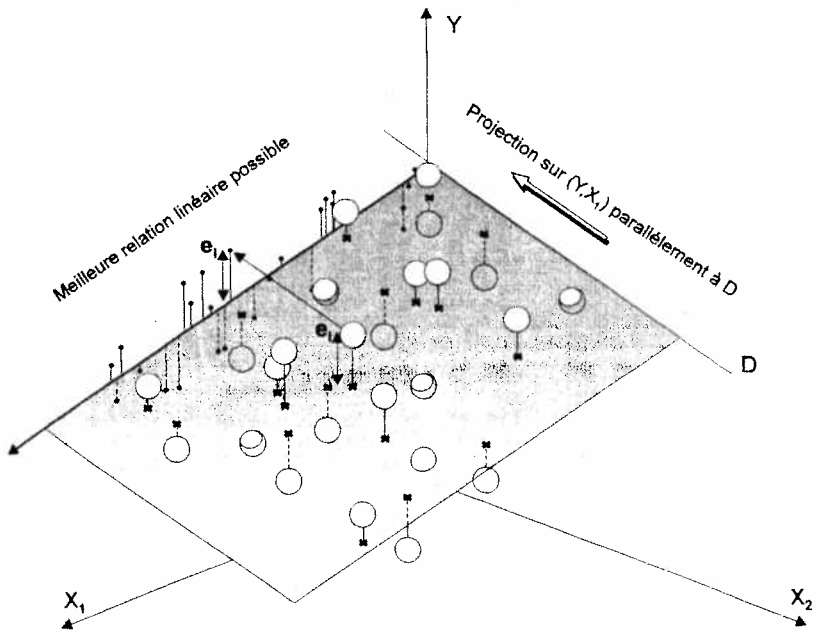

Fig. 1.2 — *Ajuster, c'est obtenir des points projetés le plus proche possible d'une droite.*

Régression linéaire et analyse de variance

Une différence de traditions

Les liens qui associent l'analyse de variance (ANOVA pour « ANalysis Of VAriance ») avec la régression linéaire ne sont pas toujours évidents.

L'analyse de variance regroupe un ensemble de techniques visant à optimiser des protocoles expérimentaux pour individualiser l'influence de différents facteurs sur un paramètre à mesurer. En effet, plutôt que de réaliser une expérience par facteur à explorer, il peut être judicieux de combiner l'action de ces facteurs afin de recueillir une information identique (voire plus riche) avec un coût moindre.

Nous examinerons ainsi plus bas un exemple où l'on tente de déterminer dans quelle mesure l'administration d'un régime normo ou hypercalorique, associé ou non à une supplémentation polyvitaminique, influence la prise de poids de n rats dénutris (un quart des animaux recevant chaque combinaison des deux facteurs après tirage au sort). Si nous essayons de formaliser cette expérience, en notant Y la variable à expliquer (le gain de poids), X_1 le facteur « calorie » ($X_1 = 1$ pour un régime hypercalorique, 2 dans le cas contraire) et X_2 le facteur « vitamine » ($X_2 = 1$ pour une supplémentation et 2 pour un placebo), la régression linéaire :

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + e_i$$

permet de rechercher la liaison de Y avec X_1, à X_2 constant, ainsi que celle de Y avec X_2, à X_1 constant. Voilà qui apporte une réponse satisfaisante à la question ayant motivé l'expérience.

Si une analyse de variance peut ainsi s'écrire sous la forme d'une régression linéaire, la tradition veut cependant que les notations utilisées soient différentes. Si la variable dépendante est toujours notée Y, les facteurs sont généralement notés par les lettres A, B, C, etc. Quand un facteur (A par exemple) a deux modalités, on considère souvent qu'ils correspondent à $A = -1$ et $A = 1$ (plutôt que $A = 1$ ou 2 comme proposé plus haut ; en pratique, il est néanmoins possible de coder librement les facteurs, les logiciels se chargeant de tout modifier en fonction de leurs propres contraintes). La constante de l'équation de régression est généralement notée μ, notons α le coefficient du facteur A, β celui du facteur B, etc. Si, dans notre exemple, les facteurs « calorie » et « vitamine » sont désormais notés A et B, et si y_i, a_i et b_i sont les réalisations de Y, A et B, on devrait ainsi écrire :

$$y_i = \mu + \alpha \cdot a_i + \beta \cdot b_i + e_i .$$

En fait, a_i et b_i valant chacun ± 1, $\mu + \alpha \cdot a_i + \beta \cdot b_i$ ne peut prendre comme valeurs que $\mu \pm \alpha \pm \beta$. On préfère alors changer de notations, l'équation ci-dessus s'écrivant désormais :

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \quad (i = 1, 2 ; j = 1, 2 ; 1 \leq k \leq n / 4)$$

où n est l'effectif total, $\alpha_1 = -\alpha_2$, $\beta_1 = -\beta_2$ et y_{ijk} correspond à la k^e mesure effectuée dans le groupe correspondant à la i^e modalité de A et à la j^e modalité de B.

En résumé, une telle analyse de variance stipule que le gain de poids y_{ijk} peut se décomposer sous la forme :

- d'un gain moyen de poids μ (qui ne tient compte ni d'un apport calorique ni d'une supplémentation vitaminique) ;
- de deux effets « calorie » α_1 et α_2. Plus précisément, α_1 quantifie le gain de poids apporté spécifiquement par le régime hypercalorique, et α_2 l'effet spécifique du régime normocalorique (à supplémentation vitaminique identique). On

remarquera que la somme des effets spécifiques doit être nulle puisque μ représente le gain moyen de poids ;

– de deux « effets vitamine » β_1 et β_2. Ici, de même, β_1 quantifie le gain de poids apporté spécifiquement par la supplémentation vitaminique (à apport calorique constant).

L'interaction de deux facteurs

Pour raffiner les conclusions d'une telle expérience, on pourrait chercher l'existence d'une synergie entre les apports caloriques et vitaminiques, telle que le gain de poids résultant de la prise de ces deux produits soit supérieur à la somme des gains de poids escomptés après la prise de chacun d'entre eux. Une telle synergie est désignée en statistique sous le terme d'« interaction » (ici entre les facteurs « calorie » et « vitamine »). Si l'on désire tester cette interaction, il faut ajouter le produit $\alpha_i\beta_j$ dans l'équation de notre analyse de variance, qui devient alors :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + e_{ijk} \; (i = 1, 2 \; ; j = 1, 2 \; ; 1 \leq k \leq n / 4)$$

Ainsi, $\alpha_1\beta_1$ représente-t-il l'effet propre de la conjonction d'un régime hypercalorique avec une supplémentation vitaminique ; $\alpha_1\beta_1 > 0$ traduisant une potentialisation des deux facteurs pour accroître le poids d'un animal.

L'essentiel

La régression linéaire

En quelques mots : quand on désire tester, en pratique, la liaison entre une variable Y *quantitative* et une variable X_1 (*quantitative ou qualitative binaire*) avec ajustement sur les variables X_2, X_3, ..., X_p (*quantitatives ou qualitatives binaires*), c'est sur le coefficient a_1 de la régression linéaire $y_i = a_0 + a_1x_{1i} + a_2x_{2i} + ... + a_px_{pi} + e_i$ que va porter le test. Une absence de liaison se traduira par l'hypothèse nulle $a_1 = 0$, l'hypothèse alternative $a_1 \neq 0$ se décomposant en $a_1 > 0$ (liaison positive entre X_1 et Y) et $a_1 < 0$ (liaison négative).

> Le test de l'hypothèse nulle $a_1 = 0$ est valide si les résidus e_i suivent une loi normale de même variance (cette variance ne doit donc dépendre ni de y_i ni des x_{ji} $(1 \leq j \leq p)$; et si les e_i ne sont pas corrélés. Ces conditions de validité sont identiques pour l'analyse de variance.

En pratique : chez 101 patients déprimés hospitalisés en psychiatrie, on désire tester, à l'entrée à l'hôpital, l'existence d'une association linéaire entre la monotonie de la voix (mesurée par l'Af0s, paramètre quantitatif évaluant la variabilité de la hauteur de la voix) et l'intensité de la dépression, mesurée par l'échelle de dépression de Hamilton (HDRS). La voix étant sensiblement modifiée par les médicaments psychotropes, il est jugé nécessaire d'ajuster sur les variables

binaires : antidépresseur tricyclique (oui = 1, non = 0), antidépresseur sérotoninergique (1,0), neuroleptique (1,0), benzodiazépine (1,0).

Effectuons dans un premier temps les calculs sans ajustement. Nous allons donc nous pencher sur l'équation de régression :

$$HDRS_i = a_0 + a_1 \cdot Af0s_i + e_i \, .$$

Le logiciel SPSS conduit aux résultats (5) :

```
      * * * *  M U L T I P L E  R E G R E S S I O N  * * * *

Equation Number 1    Dependent Variable..   HDRS ❶

  Descriptive Statistics are printed on Page   16

Block Number  1.  Method:  Enter      AF0S ❷

Variable(s) Entered on Step Number
   1..    AF0S


Multiple R            .18482 ❸
R Square              .03416
Adjusted R Square     .02450
Standard Error       6.61739

Analysis of Variance
                 DF        Sum of Squares       Mean Square
Regression        1             154.86343         154.86343
Residual        100            4378.97971          43.78980

F =       3.53652       Signif F =  .0629

---------- Variables in the Equation ----------

Variable              B        SE B       Beta          T   Sig T

AF0S          -5.151939 ❹  2.739572 ❺  -.184817    -1.881  .0629 ❻
(Constant)    25.658871    2.886383                 8.890  .0000

End Block Number   1   All requested variables entered.
```

La variable à expliquer est introduite en ❶ sous le terme de variable dépendante ; la variable explicative « Af0s » est en ❷. Nous reviendrons plus tard sur le coefficient R ❸.

En ❹ nous trouvons la valeur du coefficient a_1 associé à « Af0s », en ❺ se trouve l'écart type de ce coefficient et enfin en ❻, le « p » correspondant au test de l'hypothèse nulle $a_1 = 0$. Ici, $t = -1{,}881$ et $p = 0{,}0629$.

5 Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

Voyons maintenant le calcul avec ajustement. Après l'introduction des variables reflétant la prise d'un traitement psychotrope, l'équation de régression devient :

$$HDRS_i = a_0 + a_1 \cdot AF0s_i + a_2 \cdot ADtric_i + a_3 \cdot ADserot_i + a_4 \cdot neurol_i + a_5 \cdot benzo_i + e_i ,$$

avec les résultats correspondants (6) :

```
          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Listwise Deletion of Missing Data

Equation Number 1     Dependent Variable..    HDRS  ❶

Block Number  1.  Method:  Enter
   AF0S    BENZO    NEUROL SEROT TRICYC  ❷

Variable(s) Entered on Step Number
   1..     TRICYC
   2..     AF0S
   3..     NEUROL
   4..     BENZO
   5..     SEROT


Multiple R            .30104 ❸
R Square              .09062
Adjusted R Square     .01729
Standard Error       6.45864


Analysis of Variance
                 DF        Sum of Squares        Mean Square
Regression        5             257.73062          51.54612
Residual         62            2586.26938          41.71402

F =       1.23570       Signif F =  .3033 ❹

------------------ Variables in the Equation ------------------

Variable             B         SE B       Beta          T  Sig T

TRICYC          1.015702   2.522127    .050602      .403  .6885
AF0S           -6.768304 ❺ 3.006781 ❻ -.281938    -2.251  .0279 ❼
NEUROL           -.359492   1.761851   -.025682     -.204  .8390
BENZO          -2.061915   1.646296   -.158308    -1.252  .2151
SEROT          -2.071424   2.703120   -.097334     -.766  .4464
(Constant)     29.337322   3.618463                8.108  .0000

End Block Number   1   All requested variables entered.
```

La variable dépendante est toujours en ❶, les variables indépendantes en ❷. Le coefficient de corrélation multiple R est en ❸. Quand une régression s'écrit $y_i = a_0 + a_1x_{1i} + ... + a_px_{pi} + e_i$, R est égal au coefficient de corrélation usuel calculé entre Y et la variable $(a_0 + a_1X_1 + ... + a_pX_p)$; il évalue ainsi la qualité globale de la régression. Plus précisément, comme dans le cas univarié, R^2 est égal au pourcentage de variance que partagent Y et $(a_0 + a_1X_1 + ... + a_pX_p)$. Le test cor-

6 Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

respondant à l'hypothèse nulle R = 0 est en ❹, ce test est d'un intérêt pratique limité.

Si, en biométrie le paramètre R^2 est de peu de valeur, il est des disciplines où il joue un rôle essentiel, notamment quand la régression linéaire est plus envisagée sous l'angle de la modélisation que sous l'angle de l'ajustement. En économie, on peut ainsi souhaiter modéliser les relations entre une variable à expliquer (la consommation des ménages par exemple) et une liste de variables explicatives (le revenu, le nombre d'enfants, la catégorie socio-économique, etc.) par une égalité du type $Y = a_0 + a_1X_1 + ... + a_pX_p + \varepsilon$ (où ε suit une loi normale d'espérance nulle). Ce modèle stipule ainsi que superposé à la relation $Y = a_0 + a_1X_1 + ... + a_pX_p$, il existe un bruit de fond représenté par ε. L'importance relative du bruit de fond ε étant égale à $1 - R^2$, R^2 quantifie donc d'une certaine façon l'adéquation du modèle à la réalité.

En ❺ nous trouvons la valeur du coefficient a_1 associé à la variable « Af0s » dans le modèle ajusté, en ❻ se trouve son écart type et enfin, en ❼, le « p » correspondant au test de l'hypothèse nulle $a_1 = 0$. Ici, $t = -2{,}251$ pour $p = 0{,}0279$, alors que pour le modèle sans ajustement « p » valait 0,0629. L'ajustement renforce ainsi l'association entre les variables « HDRS » et « Af0s ».

Il ne faudrait surtout pas croire qu'un ajustement obtenu par une régression se déroule toujours sans surprise. Bien au contraire, il arrive fréquemment que des associations deviennent significatives après ajustement sur certaines variables, puis perdent cette significativité après l'introduction de variables supplémentaires sans que cela soit facilement interprétable.

Reprenons l'exemple étudiant la monotonie de la voix dans la dépression. Ravis de l'effet positif de l'ajustement sur les variables « traitement médicamenteux », nous décidons de raffiner le résultat. L'anxiété est un facteur susceptible de modifier la voix ; il serait donc utile d'ajuster sur cette variable. Il est possible de mesurer l'anxiété au moyen d'une échelle, ici celle de Tyrer. Notre régression devient maintenant :

$$HDRS_i = a_0 + a_1 \cdot Af0s_i + a_2 \cdot ADtric_i + a_3 \cdot ADserot_i + \\ + a_4 \cdot neurol_i + a_5 \cdot benzo_i + \mathbf{a_6 \cdot Tyrer_i} + e_i .$$

Et le logiciel SPSS aboutit aux résultats :

```
          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Listwise Deletion of Missing Data

Equation Number 1     Dependent Variable..    HDRS

Block Number  1.  Method:  Enter
   AF0S  BENZO    NEUROL SEROT TRICYC TYRER

Variable(s) Entered on Step Number
   1..     TYRER
   2..     TRICYC
   3..     NEUROL
```

```
   4..    AF0S
   5..    BENZO
   6..    SEROT

Multiple R            .56951
R Square              .32434
Adjusted R Square     .23988
Standard Error       5.49927

Analysis of Variance
                 DF        Sum of Squares        Mean Square
Regression        6             696.82294          116.13716
Residual         48            1451.61342           30.24195

F =       3.84027       Signif F =  .0033

------------------ Variables in the Equation ------------------

Variable                B         SE B        Beta           T  Sig T

TYRER            .395651     .097778     .504630       4.046  .0002  ❷
TRICYC          2.518001    2.521001     .125599        .999  .3229
NEUROL          1.122732    1.792981     .078251        .626  .5342
AF0S           -4.502569    3.225064    -.173997      -1.396  .1691  ❶
BENZO            .678385    1.592096     .053829        .426  .6719
SEROT           -.215551    2.941329    -.009915       -.073  .9419
(Constant).    16.972478    4.612254                   3.680  .0006

End Block Number   1   All requested variables entered.
```

En ❶, le « p » correspondant au coefficient de la variable Af0s n'est plus que de 0,169 (contre 0,0279 dans l'ancien modèle) ; l'introduction de la variable explicative « Tyrer » fait donc disparaître la significativité de l'association (HDRS, Af0s).

Un tel résultat est difficile à interpréter. Certes, il peut correspondre à la mise en évidence d'un facteur de confusion : l'anxiété. L'association [dépression / monotonie de la voix] proviendrait alors d'une association [anxiété / monotonie de la voix] et de la présence fréquente d'une composante anxieuse dans la symptomatologie dépressive. Il se peut que ce soit aussi le simple effet du « hasard ». Si, en effet, on introduit dans une régression $Y = a_0 + a_1X_1 + ... + a_pX_p + \varepsilon$ une variable explicative sans relation avec Y ou les X_i, il y a toutes les chances que l'on puisse observer malgré tout de légères fluctuations aléatoires dans l'estimation des différents coefficients a_i. Il peut arriver que ces modifications soient suffisantes pour modifier la significativité d'un effet.

Au total, il faut toujours garder à l'esprit qu'une variable d'ajustement ne doit pas être choisie en fonction des résultats auxquels elle aboutit, mais en fonction de l'intérêt qu'il y a à la maîtriser dans le phénomène étudié.

Il est, de même, toujours préférable de bien spécifier *avant* d'analyser les données quelles variables seront utilisées pour l'ajustement.

L'analyse de variance (ANOVA)

On ne peut prétendre résumer en quelques pages les différentes techniques d'analyse de variance. Ces dernières ne sont en effet que le reflet formel de plans d'expérience que l'on peut diversifier à l'infini, au gré de l'astuce des expérimentateurs. Pour illustrer ce point, nous allons maintenant aborder cinq exemples typiques d'analyse de variance :

- le premier est une analyse de variance à un facteur, en principe la situation la plus simple ;
- le second est relatif à la méthode des « blocs » ;
- le troisième est un « plan factoriel d'ordre 2 », il représente le prototype d'une analyse de variance ;
- le quatrième est un plan expérimental utilisant des « carrés latins » ;
- le cinquième repose sur un modèle mixte, les notions d'effet fixe et d'effet aléatoire y seront donc discutées.

Une ANOVA à un facteur

En quelques mots : pour comparer deux moyennes, nous avons vu dans la partie sur les méthodes univariées qu'il était possible d'utiliser un test t. Il est des situations où ce sont trois moyennes ou davantage que l'on souhaite comparer ; si ces moyennes correspondent à k groupes désignés par la variable « groupe = 1, 2, ..., k », il est possible de procéder à une telle comparaison à partir d'une analyse de variance à un facteur — le facteur « groupe ». Si Y est la variable mesurée et n l'effectif total étudié, nous obtenons ainsi l'équation :

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (1 \le i \le k,\ 1 \le j \le \text{effectif du } i^e \text{ groupe})$$

(où y_{ij} correspond à la mesure de Y effectuée sur le j^e sujet du i^e groupe et α_i à la i^e modalité de la variable « groupe ») (7).

La comparaison de ces k moyennes se fait en deux temps :

- un premier test permet d'accepter ou de rejeter l'hypothèse d'une égalité globale des k moyennes ;
- si cette hypothèse est rejetée, il est ensuite possible de rechercher dans quelle configuration se situent les moyennes les unes par rapport aux autres. C'est ce que l'on appelle une étude de contrastes (8). Il n'existe malheureusement pas d'unanimité sur la conduite à tenir pour une telle étude.

---

7 Dans l'équation $y_{ij} = \mu + \alpha_i + e_{ij}$, le facteur « groupe » associé à α ayant plus de deux classes, il n'est pas aussi simple qu'à l'habitude de déterminer l'expression d'une régression linéaire équivalente. Ce point sera traité pages 134 et 221. On retiendra avant tout que l'interprétation des termes « α_i » est toujours la même : leur somme est (généralement) choisie nulle, et ils correspondent chacun à l'effet spécifique d'une des modalités du facteur qu'ils représentent.

8 Si $m_1, ..., m_k$ sont les k moyennes, un contraste se définit formellement comme une combinaison linéaire des m_i : $c_1m_1 + ... + c_km_k$ vérifiant $c_1 + ... + c_k = 0$. Comparer les moyennes m_i et m_j revient à considérer le contraste : $c_i = 1$, $c_j = -1$ et $c_h = 0$ ($h \neq i, j$).

En pratique : pour optimiser une technique de culture monocouche de fibroblastes humains, on désire comparer la performance de quatre milieux : un milieu minimum (milieu = « mini »), un milieu minimum enrichi en sérum de veau fœtal à 5 % (milieu = « veau5 »), un milieu minimum enrichi en sérum de veau fœtal à 10 % (milieu = « veau10 ») et un milieu minimum enrichi en sérum humain (milieu = « humain »). Après tirage au sort, 36 boîtes sont remplies d'un des quatre milieux, la variable mesurée est ici « logcells », le logarithme du nombre de cellules obtenues après une période de pousse de 3 jours ([9]). Les données récoltées sont ainsi ([10]) :

obs	milieu	logcells	obs	milieu	logcells	obs	milieu	logcells
1	mini	3.72	13	mini	4.43	25	mini	3.95
2	veau5	5.59	14	veau5	5.43	26	veau5	5.41
3	veau10	5.8	15	veau10	5.41	27	veau10	5.76
4	humain	5.55	16	humain	5.61	28	humain	5.67
5	mini	4.33	17	mini	4.13	29	mini	4.35
6	veau5	5.26	18	veau5	5.65	30	veau5	5.41
7	veau10	5.68	19	veau10	6.23	31	veau10	5.69
8	humain	5.65	20	humain	5.53	32	humain	5.29
9	mini	4	21	mini	4.13	33	mini	4.09
10	veau5	5.17	22	veau5	5.27	34	veau5	5.14
11	veau10	5.47	23	veau10	5.79	35	veau10	5.89
12	humain	5.95	24	humain	5.58	36	humain	5.78

Le logiciel SAS, PROC ANOVA aboutit aux résultats suivants ([11]) :

```
                    Analysis of Variance Procedure

                       Class Level Information

                Class        Levels        Values ❶
                MILIEU            4    humain mini veau10 veau5

                  Number of observations in data set = 36

Dependent Variable: LOGCELLS ❷

Source                 DF   Sum of Squares     Mean Square   F Value    Pr > F
Model                   3      14.94280833      4.98093611    116.87    0.0001
Error                  32       1.36382222      0.04261944
Corrected Total        35      16.30663056

          R-Square             C.V.          Root MSE          LOGCELLS Mean
          0.916364         3.957619        0.20644477           5.21638889

Source                 DF         Anova SS     Mean Square   F Value    Pr > F
MILIEU ❸                3      14.94280833      4.98093611    116.87    0.0001 ❹
```

[9] Quand une variable représente un comptage, son logarithme satisfait généralement mieux à la condition de normalité requise, pour la variable à expliquer, dans une analyse de variance.

[10] Données fictives.

[11] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

La variable dépendante est en ❷, la variable indépendante indiquant les groupes à comparer est en ❶. Le test permettant d'évaluer l'égalité globale des quatre moyennes correspond au test de l'effet de la variable « milieu » ❸, le « p » correspondant est en ❹ soit $p < 0{,}0001$. Il y a donc, selon toute vraisemblance, une différence d'efficacité de ces quatre milieux.

Voyons maintenant quels sont les milieux qui se détachent des autres. La procédure PROC ANOVA nous permet de poursuivre l'analyse de variance par une étude de contrastes :

```
                    Analysis of Variance Procedure

                T tests ❶ (LSD) for variable: LOGCELLS

NOTE: This test controls the type I comparisonwise error rate not the
                    experimentwise error rate.

                Alpha= 0.05 ❺ df= 32  MSE= 0.042619
                    Critical Value of T= 2.04
                Least Significant Difference= 0.1982

        Means with the same letter are not significantly different. ❷

              T Grouping                Mean ❹   N  MILIEU

                       A              5.74667     9  veau10
                       A ❸
                       A              5.62333     9  humain
                       B              5.37000     9  veau5
                       C              4.12556     9  mini

             Bonferroni ❻ (Dunn) T tests for variable: LOGCELLS

NOTE: This test controls the type I experimentwise error rate, but
      generally has a higher type II error rate than REGWQ.

                Alpha= 0.05  df= 32  MSE= 0.042619
                    Critical Value of T= 2.81
                Minimum Significant Difference= 0.2737
Means with the same letter are not significantly different.

            Bon Grouping                Mean   N  MILIEU

                       A              5.74667     9  veau10
                       A ❼
                B      A              5.62333     9  humain
                B
                B                     5.37000     9  veau5
                       C              4.12556     9  mini

            Tukey's ❽ Studentized Range (HSD) Test for variable: LOGCELLS

NOTE: This test controls the type I experimentwise error rate, but
      generally has a higher type II error rate than REGWQ.

                Alpha= 0.05  df= 32  MSE= 0.042619
```

```
                Critical Value of Studentized Range= 3.832
                Minimum Significant Difference= 0.2637

Means with the same letter are not significantly different.

      Tukey Grouping              Mean      N  MILIEU

                   A          5.74667      9  veau10
                   A
               B   A          5.62333      9  humain
               B
               B              5.37000      9  veau5

                   C          4.12556      9  mini
```

Plusieurs approches sont possibles pour étudier des contrastes.

– En ❶ nous trouvons l'approche la plus élémentaire : de simples tests t entre chacune des moyennes. Une convention d'écriture (❷) nous permet de regrouper visuellement les moyennes voisines, nous remarquons ainsi en ❸ que le milieu au sérum de veau fœtal à 10 % ne se différencie pas significativement du milieu au sérum humain ; par contre, le milieu au sérum de veau fœtal 5 % semble inférieur, et le milieu minimum encore bien inférieur (les moyennes en ❹ nous aident à évaluer l'importance des différents écarts).

Un reproche est parfois fait à cette approche. Si l'on compare toutes les moyennes entre elles, il est nécessaire de réaliser six tests. Chacun de ces tests étant susceptible de conclure de façon non appropriée à une différence significative, le risque global de trouver une telle différence à tort devient bien supérieur au 5 % que l'on s'octroie habituellement (❺). Plusieurs méthodes ont été proposées pour pallier cet inconvénient.

– La plus connue de ces méthodes est celle de Bonferroni ❻. Elle consiste à diminuer le seuil de significativité utilisé pour la comparaison de chaque paire de moyennes. Puisque ici il y a six tests, le seuil à utiliser sera maintenant de 5 % divisé par 6 soit 0,83 %. Il devient donc plus difficile de mettre en évidence une différence significative, c'est bien le cas dans notre exemple où, en ❼, nous remarquons maintenant que le milieu à base de sérum humain ne peut plus être différencié du milieu au sérum de veau fœtal à 5 %.

– La méthode de Tukey (❽) est une alternative classique. Bien que plus puissante que celle de Bonferroni, elle ne permet pas ici d'amélioration par rapport à cette dernière.

En pratique, la solution ne vient pas de l'utilisation d'une méthode particulière, mais de la valeur à donner aux résultats que l'on trouve. Il faut pour cela bien séparer les calculs que l'on fait dans un but confirmatoire, de ceux qui le sont dans un but exploratoire.

Ainsi, dans le protocole d'une expérience, il est licite (et recommandé) de prévoir, outre le test global d'identité des moyennes, un nombre limité de comparaison de moyennes et autres contrastes (2 ou 3 par exemple). Les tests correspondant à ces comparaisons auront un poids bien supérieur à tous ceux qui pourront être menés par ailleurs, dans le feu de l'action. Ils auront ainsi une authentique

valeur confirmatoire, ce qui veut dire qu'en cas de découverte d'un résultat significatif, on saura exactement à quoi s'en tenir : le risque d'erreur correspond au risque de première espèce. Les autres tests auront, eux, une valeur exploratoire, ce qui signifie qu'ils donnent une indication sur le phénomène observé, sans plus ; une nouvelle expérience devant être réalisée si l'on veut en savoir davantage.

La méthode des blocs

En quelques mots : quand on compare la moyenne d'un paramètre dans plusieurs groupes, il est généralement utile de pouvoir disposer de groupes les plus homogènes possible afin d'obtenir une puissance élevée pour la comparaison.

Il est des situations expérimentales où l'on connaît par avance certains facteurs susceptibles de nuire à une telle homogénéité. Cela sera par exemple le cas si l'on compare le rendement de quatre variétés de maïs en les semant sur un lot de parcelles (six par exemple) ; les différences de fertilité de ces dernières vont introduire une variabilité parasite, nuisible pour la comparaison. L'idéal serait de découper chaque parcelle en quatre, de répartir aléatoirement chaque variété dans chaque quart pour comparer la productivité de chaque espèce de maïs au sein de chaque parcelle, et finalement résumer ces six comparaisons en une seule conclusion. En d'autres termes, il s'agit d'ajuster sur la variable parcelle ([12]).

Si Y est la variable dépendante, A la variable identifiant les p groupes à comparer (ici les quatre variétés de maïs) et B celle identifiant les q blocs (les six parcelles), le modèle d'analyse de variance est :

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \; (1 \leq i \leq p, \; 1 \leq j \leq q),$$

il porte souvent le nom d'analyse de variance à deux facteurs croisés.

En pratique : quatre variétés de maïs (variété = a, b, c ou d) sont cultivées sur six parcelles de terre (parcelle = 1, 2, 3, 4, 5 et 6). On désire comparer la performance de chacune des variétés à qualité de terrain comparable, on utilise donc la méthode des blocs, chaque variété de maïs étant cultivée dans un quart de chaque parcelle choisi aléatoirement. La variable dépendante est ici le nombre de boisseaux récoltés par hectare (variable « récolte »). Les données se présentent sous la forme ([13]) :

obs	variété	parcelle	récolte	obs	variété	parcelle	récolte
1	a	1	232	13	C	1	190
2	a	2	279	14	C	2	208
3	a	3	251	15	C	3	235
4	a	4	278	16	C	4	190

[12] En plus d'un simple ajustement, on remarquera ici qu'à l'intérieur de chaque bloc on retrouve chaque variété de maïs en nombre identique, les blocs sont ainsi dits « complets ». Cela contribue encore davantage à réduire la dispersion des mesures.

[13] Données fictives.

5	a	5	294	17	c	5	224
6	a	6	284	18	c	6	215
7	b	1	193	19	d	1	225
8	b	2	220	20	d	2	201
9	b	3	240	21	d	3	249
10	b	4	249	22	d	4	251
11	b	5	238	23	d	5	234
12	b	6	220	24	d	6	198

La procédure SAS, PROC ANOVA aboutit aux résultats ([14]) :

```
                          The SAS System

                  Analysis of Variance Procedure
                     Class Level Information

Class     Levels    Values ❶
VARIETE        4    a b c d
PARCELLE       6    1 2 3 4 5 6

                Number of observations in data set = 24

                  Analysis of Variance Procedure

Dependent Variable: RECOLTE ❷

Source            DF       Sum of Squares      Mean Square    F Value     Pr > F
Model              8     15597.16666667     1949.64583333       5.75     0.0018
Error             15      5083.33333333      338.88888889 ①
Corrected Tot     23     20680.50000000

       R-Square             C.V.          Root MSE          RECOLTE Mean
       0.754197         7.892362       18.40893503            233.25000000

Source            DF          Anova SS      Mean Square    F Value     Pr > F
VARIETE ❸          3     11655.16666667     3885.05555556      11.46     0.0004 ❹
PARCELLE           5      3942.00000000      788.40000000 ②     2.33     0.0941 ❻

                 Level of          ----------RECOLTE----------
                 VARIETE       N         Mean ❺          SD

                 a             6    269.666667       23.3295235
                 b             6    226.666667       20.1362029
                 c             6    210.333333       18.1622319
                 d             6    226.333333       22.9230597
```

Les variables explicatives sont bien qualitatives en ❶, la variable à expliquer est en ❷. Le test recherchant une différence globale de rendement pour les quatre variétés de maïs doit être recherché en ❸, le « p » correspondant est en ❹ soit p = 0,0004. L'observation des moyennes des récoltes des différentes variétés en

[14] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

❺ nous permet de préciser ce résultat : il semble que ce soit la variété « a » qui soit la plus performante alors que la « c » apparaît en retrait.

On peut se demander quelle a été l'efficacité de l'ajustement sur la variable « parcelle ». Certes, cette dernière ne conduit pas à un effet significatif au seuil de 5 % (❻), mais ce résultat est sans conséquence : une variable ajustement ne doit pas nécessairement conduire à un effet significatif, il suffit qu'elle soit connue pour être susceptible d'introduire une variabilité parasite dans la mesure de la variable à expliquer. Il est possible de se faire une idée sur le gain apporté par l'utilisation de blocs en calculant le nombre d'observations virtuelles que nous a fait gagner cet ajustement. Ce nombre est égal au rapport des variances résiduelles sans et avec ajustement. Désignons par VarRes la variance résiduelle sans ajustement et $\text{VarRes}_{\text{ajust}}$ la variance résiduelle avec ; la variable « variété » comptant quatre classes et la variable d'ajustement « parcelle » en comptant six, nous avons :

$$\text{VarRes}_{\text{ajust}} = ① = 338{,}9$$
$$\text{VarRes} = \left[(6-1) \times ② + 6 \times (4-1) \times ①\right] / (6 \times 4 - 1) = 436{,}6$$

Le rapport des deux vaut donc r = 436,6 / 338,9 ≈ 1,30. La méthode des blocs nous a donc fait gagner environ 30 % d'observations supplémentaires, soit à peu près deux parcelles (chaque parcelle représentant un sixième des observations, soit approximativement 17 %).

Un plan factoriel d'ordre 2

En quelques mots : un plan factoriel d'ordre 2 est généralement utilisé pour étudier l'effet de deux facteurs sur un paramètre quantitatif. Cela sera par exemple le cas si l'on désire étudier l'effet d'un régime normo ou hypercalorique (facteur A) ainsi que l'effet d'une supplémentation polyvitaminique (facteur B) sur le poids de rats dénutris. Deux protocoles sont alors envisageables.

– Le plus simple est de considérer deux expériences distinctes. Une première comparera n rats bénéficiant d'un régime normocalorique à n rats bénéficiant d'un régime hypercalorique. Une seconde comparera n rats recevant un placebo à n rats recevant une supplémentation polyvitaminique. Ce protocole nécessite donc 4n rats.

– Le second (le plan factoriel) nécessite seulement 2n rats ; ces derniers recevront en revanche une combinaison des facteurs A et B. Ainsi, un premier groupe (noté ①) de n / 2 rats recevra un régime normocalorique ainsi qu'un placebo, un second groupe (noté ②) de n / 2 rats recevra un régime hypercalorique ainsi qu'un placebo ; le groupe ③ de n / 2 rats recevra un régime normocalorique ainsi qu'une supplémentation polyvitaminique, enfin le groupe ④ recevra à la fois un régime hypercalorique et une supplémentation polyvitaminique.

Pour étudier le facteur A, on comparera le groupe ① au ② et le groupe ③ au ④ (le niveau du facteur B sera donc identique) ; pour étudier le facteur B, on comparera le groupe ① au ③ et le ② au ④. Les 2n rats sont ainsi disponibles

pour l'étude de chacun des facteurs. Par ce protocole, il est en outre possible d'étudier l'interaction entre les facteurs A et B. Dans notre exemple, cette interaction correspondrait à une éventuelle potentialisation de la supplémentation polyvitaminique avec le régime hypercalorique. Cette interaction était, bien sûr, impossible à appréhender avec le premier protocole.

Si Y désigne la variable aléatoire correspondant au gain de poids d'un rat, l'analyse de variance permettant de formaliser ce plan factoriel s'écrit :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + e_{ijk} \quad (i = 1, 2 ; j = 1, 2 ; 1 \leq k \leq n / 2)$$

où y_{ijk} correspond à la k^e mesure effectuée dans le groupe défini par la i^e modalité du facteur A et la j^e modalité du facteur B ; comme à l'habitude, α_i correspond à l'effet de la i^e modalité du facteur A, β_j à l'effet de la j^e modalité du facteur B et la produit $\alpha_i\beta_j$ à l'interaction de A et B.

En pratique : passons maintenant à une application numérique de l'exemple introduit ci-dessus. Le gain de poids des rats est désigné par la variable « poids » (exprimée en grammes), les deux facteurs sont les variables « calorie » (calorie = 0 pour un régime normocalorique et 1 pour un régime hypercalorique) et « vitamine » (vitamine = 0 pour un placebo et 1 pour une supplémentation polyvitaminique). Les données recueillies prennent ainsi la forme ([15]) :

obs	calorie	vitamine	poids	obs	calorie	vitamine	poids
1	0	0	84	17	0	0	66
2	0	1	61	18	0	1	59
3	1	0	87	19	1	0	89
4	1	1	103	20	1	1	90
5	0	0	56	21	0	0	56
6	0	1	84	22	0	1	74
7	1	0	92	23	1	0	101
8	1	1	107	24	1	1	116
9	0	0	81	25	0	0	79
10	0	1	73	26	0	1	74
11	1	0	77	27	1	0	95
12	1	1	95	28	1	1	112
13	0	0	62	29	0	0	89
14	0	1	.	30	0	1	74
15	1	0	88	31	1	0	91
16	1	1	96	32	1	1	92

La procédure SAS, PROC GLM aboutit aux résultats ([16]) :

[15] Données fictives.

[16] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.


```
                          The SAS System

                  General Linear Models Procedure

                     Class Level Information

                  Class     Levels    Values ❶

                  CALORIE        2    0 1

                  VITAMINE       2    0 1

                  Number of observations in data set = 32
NOTE: Due to missing values, only 31 observations can be used
                        in this analysis. ❸

Dependent Variable: POIDS ❷

Source                  DF    Sum of Squares      Mean Square   F Value    Pr > F

Model                    3    5059.78917051    1686.59639017     17.28    0.0001

Error                   27    2635.17857143      97.59920635

Corrected Total         30    7694.96774194

            R-Square                 C.V.          Root MSE          POIDS Mean
            0.657545             11.76551        9.87923106         83.96774194

Source                  DF       Type I SS ❹     Mean Square   F Value    Pr > F
CALORIE                  1    4541.79690860    4541.79690860     46.54    0.0001
VITAMINE                 1     253.00704023     253.00704023      2.59    0.1190
CALORIE*VITAMINE         1     264.98522167     264.98522167      2.72    0.1110

Source                  DF      Type III SS ❺    Mean Square   F Value    Pr > F
CALORIE                  1    4535.58866995    4535.58866995     46.47    0.0001
VITAMINE                 1     235.17487685     235.17487685      2.41    0.1322
CALORIE*VITAMINE         1     264.98522167     264.98522167      2.72    0.1110
```

Les facteurs « calorie » et « vitamine » sont présentés en ❶, la variable dépendante « poids » est en ❷. La procédure PROC GLM a été utilisée à la place de la procédure PROC ANOVA car il y a une donnée manquante pour le 14e animal (❸) ([17]).

Deux séries de résultats sont proposées en ❹ et ❺. Si les effectifs sont équilibrés, ces résultats sont identiques, ce qui n'est pas le cas ici. Les différences existant entre ces divers résultats seront expliquées à la page 239 ; nous choisirons tout au long de cet ouvrage ceux présentés en ❺ (somme des carrés de type III).

Le « p » correspondant à l'effet du facteur « calorie » est < 0,0001, alors que pour le facteur « vitamine » et pour l'interaction « calorie × vitamine », on a respectivement p = 0,1322 et p = 0,1110.

[17] En effet, si les groupes étudiés ont des effectifs identiques, il est possible de procéder à une analyse de variance sans avoir à faire l'ensemble des calculs d'une régression linéaire. Il existe ainsi deux types de programmes d'analyse de variance, les plus élémentaires, qui ne sont utilisables qu'avec des effectifs équilibrés (comme PROC ANOVA) et ceux qui tolèrent des effectifs déséquilibrés (comme PROC GLM).

Les carrés latins

En quelques mots : dans certaines expériences, il arrive qu'une série de k traitements soit donnée à des sujets à des moments différents (ou à des endroits différents du corps s'il s'agit de crèmes), et que l'ordre (ou le lieu d'application) dans lequel est donnée la séquence soit potentiellement important. Il est alors indispensable de tenir compte dans l'analyse d'un effet « ordre (ou lieu) d'administration », et de faire attention à ce que chaque traitement soit donné de façon équilibrée en 1re, 2e, ..., ke position. L'utilisation de carrés latins répond à cet impératif.

Prenons comme exemple l'étude d'une stimulation de la mélanogenèse chez le cochon d'inde sous l'influence de topiques particuliers, en l'occurrence d'un psoralène (connu pour son efficacité), des dimères de thymine (dont l'effet est discuté), de l'ADN de sperme de hareng (*idem*) ainsi que d'un simple véhicule servant de témoin. On envisage d'appliquer ces quatre produits en des lieux différents d'un même animal : une face externe de cuisse, le ventre, le dos, le torse. La mélanogenèse dépendant de la zone de peau considérée, il est indispensable de maîtriser ce facteur au moyen d'une administration où, pour chaque groupe de quatre animaux, un produit apparaît, après tirage au sort, une seule fois par animal et par localisation : il s'agit par définition d'un carré latin d'ordre quatre. Pour un groupe de quatre animaux, on pourrait ainsi avoir :

	localisation			
animal	torse	ventre	dos	cuisse
1	témoin	psoralène	thymine	ADN
2	psoralène	témoin	ADN	thymine
3	thymine	ADN	psoralène	témoin
4	ADN	thymine	témoin	psoralène

Si l'on décide d'étudier un plus grand nombre de rats (qui, par construction, devra néanmoins rester un multiple du nombre de traitement, ici quatre), il faudra multiplier le nombre de carrés latins (on pourra les prendre distincts les uns des autres).

Dans l'analyse de variance formalisant ce protocole, il faudra tenir compte d'un facteur « traitement » (A), d'un facteur « lieu d'application » (B) et d'un facteur « animal » (C). Nous obtenons ainsi la relation :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk} \quad (1 \le i \le 4, 1 \le j \le 4, 1 \le k \le \text{nombre d'animaux})$$

En pratique : huit rats sont retenus pour l'expérience décrite ci-dessus, un deuxième carré latin est donc construit. La variable à expliquer « mélanine » correspond à la concentration en mélanine des biopsies des différentes zones étudiées, concentration mesurée par spectrométrie. Les variables explicatives sont, comme prévu, la variable « animal » (qualitative à huit classes, animal = 1, 2, ..., 8), la variable lieu de prélèvement « lieuprel » (qualitative à quatre classes, lieuprel = cuisse, ventre, dos et torse) et la variable traitement « trt » (qualitative à quatre classes, trt = témoin, pso (psoralène), thymine et ADN).

Les données collectées sont ([18]) :

obs	animal	lieuprel	trt	mélanine	obs	animal	lieuprel	trt	mélanine
1	1	torse	témoin	25	17	5	torse	adn	25
2	1	ventre	pso	41	18	5	ventre	thymine	21
3	1	dos	thymine	37	19	5	dos	pso	47
4	1	cuisse	adn	28	20	5	cuisse	témoin	54
5	2	torse	pso	64	21	6	torse	thymine	34
6	2	ventre	témoin	3	22	6	ventre	adn	6
7	2	dos	adn	52	23	6	dos	pso	76
8	2	cuisse	thymine	36	24	6	cuisse	témoin	42
9	3	torse	thymine	36	25	7	torse	pso	64
10	3	ventre	adn	8	26	7	ventre	témoin	23
11	3	dos	pso	51	27	7	dos	adn	55
12	3	cuisse	témoin	26	28	7	cuisse	thymine	50
13	4	torse	adn	16	29	8	torse	témoin	9
14	4	ventre	thymine	13	30	8	ventre	pso	43
15	4	dos	témoin	40	31	8	dos	thymine	52
16	4	cuisse	pso	45	32	8	cuisse	adn	26

Le logiciel SAS, PROC GLM nous propose les résultats ([19]) :

```
                          The SAS System

                  General Linear Models Procedure
                      Class Level Information

            Class     Levels    Values ❶
            ANIMAL         8    1 2 3 4 5 6 7 8
            LIEUPREL       4    cuisse dos patte ventre
            TRT            4    adn pso temoin thymine

Dependent Variable: MELANINE ❷

Source             DF    Sum of Squares      Mean Square   F Value    Pr > F

Model              13    8418.14166667     647.54935897      5.74    0.0004

Error              18    2031.35833333     112.85324074

Corrected Total    31    10449.50000000

              R-Square             C.V.        Root MSE      MELANINE Mean

              0.805602         29.61182      10.62324060        35.87500000
```

[18] Données fictives.

[19] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ANIMAL	7	1105.50000000	157.92857143	1.40	0.2652
LIEUPREL	3	4045.75000000	1348.58333333	11.95	0.0002
TRT	3	3266.89166667	1088.96388889	9.65	0.0005

Source	DF	Type III SS ❸	Mean Square	F Value	Pr > F
ANIMAL	7	1105.50000000	157.92857143	1.40	0.2652
LIEUPREL	3	3554.39166667	1184.79722222	10.50	0.0003
TRT	3	3266.89166667	1088.96388889	9.65	0.0005 ❹

T tests (LSD) for variable: MELANINE ❺

NOTE: This test controls the type I comparisonwise error rate not the experimentwise error rate.

Alpha= 0.05 df= 18 MSE= 112.8532
Critical Value of T= 2.10
Least Significant Difference= 11.159
Means with the same letter are not significantly different.

T Grouping	Mean	N	TRT
A ❻	53.875	8	pso
B	34.875	8	thymine
B			
B ❼	27.750	8	temoin
B			
B	27.000	8	adn

Les variables explicatives sont détaillées en ❶, la variable à expliquer en ❷. Nous pouvons maintenant aller directement aux résultats qui nous intéressent en ❸ (somme des carrés de type III). La question posée étant (dans un premier temps) : « Y a-t-il une différence globale dans la stimulation de la mélanogenèse induite par les quatre traitements ? », nous nous tournons vers l'effet de la variable « trt » : ce dernier est significatif puisque p = 0,0005 (❹).

Une étude de contrastes (❺) nous permet d'affiner cette conclusion. Seul le psoralène semble se détacher des autres traitements (❻ et ❼). L'effet de l'ADN ou des dimères de thymine n'apparaît notamment pas différent de celui du témoin.

Le modèle mixte

En quelques mots : dans une analyse de variance, un facteur peut être à effet fixe ou à effet aléatoire.

Un facteur à effet fixe est un facteur dont les modalités ne changent pas quand l'expérience considérée est réalisée à plusieurs reprises. Dans l'exemple introduit plus haut à propos de la méthode des blocs, on compare quatre variétés de maïs cultivées chacune sur six parcelles de terre. Dans une telle situation, le facteur « variété » est à effet fixe, car si l'on désire dupliquer l'expérience, cela n'a pas de sens de modifier les variétés de maïs à comparer. A l'opposé, si les six parcelles

de terre sont tirées au sort parmi un vaste lot de terres, on peut imaginer qu'une nouvelle expérience conduise à un nouveau tirage au sort, les parcelles ne seront alors plus nécessairement identiques... Dans un tel cas, le facteur « parcelle » est, par définition, aléatoire. On pourrait néanmoins tout aussi bien décider de conserver les six parcelles retenues pour la première expérience, le facteur « parcelle » serait alors fixe, au même titre que le facteur « variété ».

Considérer le facteur « parcelle » fixe ou aléatoire a une portée sur la généralité de la conclusion de l'expérience. Un effet fixe restreint les résultats observés aux seules parcelles cultivées, alors qu'un effet aléatoire permet de généraliser ces résultats à l'ensemble des terres parmi lesquelles les parcelles ont été tirées au sort. Le prix à payer pour une telle généralisabilité est, le plus souvent, une perte de puissance pour le test des effets fixes (ici, « variété ») ainsi qu'une sophistication des logiciels à utiliser.

Un modèle combinant des effets fixes à des effets aléatoires est qualifié de modèle mixte.

En pratique : une société commercialisant de nombreuses races de souris mutantes à l'usage des laboratoires, décide de changer le type de nourriture de ses animaux. Une expérience est entreprise afin de comparer les performances de deux aliments (aliment = 1, 2). Cinq souches de souris (souche = 1, 2, 3, 4, 5) sont tirées au sort sur le catalogue de la société. Pour chaque souche, huit animaux sont choisis, quatre recevant un des deux aliments. La variable dépendante est le gain de poids (dénommé « poids », exprimé en grammes).

L'objectif de ce protocole est bien de déterminer l'aliment optimal pour *l'ensemble* des souches élevées, et non pas seulement pour les cinq souches étudiées. Or, puisqu'il est vraisemblable que toutes les souches de souris ne répondent pas de façon identique à une nourriture donnée, il est possible que les animaux tirés au sort fassent justement partie des répondeurs exceptionnels pour *un* aliment particulier ; d'où une difficulté potentielle pour généraliser les résultats de l'expérience à l'ensemble de la population de souris. Un modèle mixte est particulièrement adapté à une telle situation. En effet, en spécifiant que l'effet « souche » est aléatoire, on considère implicitement que les souches étudiées sont tirées d'une population plus vaste sur laquelle doivent porter les conclusions de l'expérience.

Le modèle d'ANOVA est dans ce cas :

$$y_{ijk} = \mu + \alpha_i + b_j + \alpha_i b_j + e_{ijk} \quad (i = 1, 2 ; 1 \le j \le 5 ; 1 \le k \le 4)$$

où Y correspond à la variable « poids », les α_i correspondent au facteur « aliment », b_j au facteur à effet aléatoire « souche » et $\alpha_i b_j$ à l'interaction « aliment × souche » ([20]).

Les données sont (l'observation n° 37 étant manquante) ([21]) :

[20] Pour les différencier, on note souvent les effets aléatoires par des lettres romaines plutôt que grecques. L'interaction $\alpha_i b_j$ est à effet aléatoire car un au moins de ses termes correspond à un effet aléatoire.

obs	aliment	souche	poids	obs	aliment	souche	poids
1	1	1	28.5	21	2	3	26.0
2	1	1	21.2	22	2	3	29.8
3	1	1	21.3	23	2	3	27.6
4	1	1	22.7	24	2	3	26.5
5	2	1	25.8	25	1	4	19.4
6	2	1	24.3	26	1	4	27.5
7	2	1	18.7	27	1	4	28.4
8	2	1	24.4	28	1	4	28.6
9	1	2	21.3	29	2	4	23.0
10	1	2	18.9	30	2	4	25.9
11	1	2	17.1	31	2	4	20.5
12	1	2	22.6	32	2	4	26.6
13	2	2	19.1	33	1	5	20.0
14	2	2	24.1	34	1	5	23.7
15	2	2	20.9	35	1	5	25.5
16	2	2	32.4	36	1	5	26.8
17	1	3	25.1	37	2	5	.
18	1	3	25.1	38	2	5	31.6
19	1	3	24.4	39	2	5	26.9
20	1	3	23.9	40	2	5	36.6

La procédure SAS, PROC MIXED conduit aux résultats ([22]) :

```
                    The MIXED Procedure
                  Class Level Information

                Class      Levels  Values ❶
                SOUCHE          5  1 2 3 4 5
                ALIMENT         2  1 2

             REML Estimation Iteration History

      Iteration  Evaluations      Objective      Criterion ❷
              0            1  145.74686729
              1            2  142.85085284     0.00003935
              2            1  142.84793852     0.00000014
              3            1  142.84792862     0.00000000
                  Convergence criteria met.


                Covariance Parameter Estimates (REML) ❸

Cov Parm            Ratio       Estimate     Std Error       Z  Pr > |Z|

SOUCHE         0.12840296     1.56831587    4.02987595    0.39    0.6971
SOUCHE*ALIMENT 0.26861950     3.28092309    4.73608365    0.69    0.4885
Residual       1.00000000    12.21401673    3.24523380    3.76    0.0002

                Model Fitting Information for POIDS ❹

              Description                          Value

              Observations                       39.0000
              Variance Estimate                  12.2140
```

[21] Données fictives.

[22] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.


```
              Standard Deviation Estimate          3.4949
              REML Log Likelihood               -105.425
              Akaike's Information Criterion    -108.425
              Schwartz's Bayesian Criterion     -110.841
              -2 REML Log Likelihood             210.8494
              Null Model LRT Chi-Square            2.8989
              Null Model LRT DF                    2.0000
              Null Model LRT P-Value               0.2347

                    Tests of Fixed Effects ❺

          Source      NDF    DDF  Type III ❻  F   Pr > F

          ALIMENT       1      4        2.23  0.2098 ❼

                    Least Squares Means

Level                  LSMEAN ❽      Std Error      DDF      T     Pr > |T|
ALIMENT 1       23.64892139     1.25719872       29   18.81     0.0000
ALIMENT 2       26.04243021     1.27173822       29   20.48     0.0000
```

Les variables explicatives sont en ❶, la variable à expliquer en ❹. En ❷ nous remarquons qu'il ne s'agit plus d'un banal calcul de régression linéaire mais d'une estimation utilisant un algorithme itératif dont nous devons vérifier la bonne convergence : elle est ici excellente puisque trois itérations ont suffi.

La première série de résultats apparaît en ❸, il s'agit des tests correspondant aux effets aléatoires — qu'il a fallu déclarer comme tels au logiciel. Puisqu'il n'y a là que de simples variables d'ajustement, ces résultats sont secondaires.

Le résultat clé correspondant au test de l'effet fixe est en ❺. On notera que seuls sont disponibles ici les résultats de type III (❻). Nous obtenons finalement un effet « aliment » non significatif avec $p = 0{,}2098$ (❼). Les gains de poids moyens correspondant à la prise de chacun des aliments sont présentés en ❽.

Voyons par curiosité ce que donnerait une analyse de variance où tous les facteurs seraient à effets fixes. La procédure SAS, PROC GLM nous propose alors (nous ne retenons ici que les principaux résultats) :

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SOUCHE	4	154.73536222	38.68384055	3.20	0.0270
ALIMENT	1	61.77573825	61.77573825	5.12	0.0314 ❶
SOUCHE*ALIMENT	4	105.07418741	26.26854685	2.18	0.0967 ❷

Le « p » correspondant à l'effet du facteur « aliment » est en ❶, soit $p = 0{,}0314$, résultat très différent de celui trouvé avec le modèle mixte. L'origine d'une telle discordance peut être trouvée en ❷. L'interaction « aliment × souche », si elle n'est pas significative au seuil de 5 %, n'en est pas moins appréciable ; or une telle interaction traduit précisément le fait que certaines souches de souris ont été particulièrement sensibles à l'un des deux aliments. Le modèle mixte prend en compte ce résultat et « corrige » le « p » de façon que cela ne soit pas seulement le tirage au sort de cinq souches particulières qui puisse expliquer à lui seul la différence observée entre les deux aliments.

Le rôle du terme d'interaction dans la différence des résultats observés entre le modèle mixte et le modèle à effets fixes peut d'ailleurs se vérifier de la façon suivante. Si l'on considère un modèle d'analyse de variance sans terme d'interaction, on suggère implicitement que cette dernière est sans importance. Le modèle mixte et le modèle à effets fixes devraient alors aboutir à des résultats voisins.

En pratique, pour le modèle mixte :

$$y_{ijk} = \mu + \alpha_i + b_j + e_{ijk} \quad (i = 1, 2 ; 1 \leq j \leq 5 ; 1 \leq k \leq 4)$$

La procédure PROC MIXED donne les résultats :

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
ALIMENT	1	33	3.78	0.0605 ❶

Alors que pour le modèle à effets fixes :

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \quad (i = 1, 2 ; 1 \leq j \leq 5 ; 1 \leq k \leq 4)$$

PROC GLM propose :

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SOUCHE	4	140.39233161	35.09808290	2.54	0.0579
ALIMENT	1	54.79164378	54.79164378	3.97	0.0546 ❷

Les deux valeurs de « p » sont effectivement voisines : la première est égale à 0,0605 (❶) et la seconde à 0,0546 (❷) ([23]).

L'analyse de covariance

En quelques mots : la régression linéaire est une technique très générale permettant à la fois :

- de prédire une variable quantitative normale Y à partir d'une série de p variables (qualitatives ou quantitatives) X_i ($1 \leq i \leq p$) ;
- de rechercher une relation entre Y et un des X_i, avec ajustement sur les variables restantes X_j ($j \neq i$).

[23] Si les effectifs avaient été « équilibrés » (du fait de la donnée manquante, il n'y a pas ici autant de souris de chaque souche ayant consommé un aliment donné), les résultats du modèle mixte et ceux du modèle à effets fixes auraient été identiques.

Un tel ajustement permet d'analyser au plus près les résultats de nombreux protocoles expérimentaux. Les régressions que l'on utilise généralement dans ces circonstances se font à partir de variables X_i qualitatives et l'on parle alors d'analyse de variance. Rien n'empêche cependant de conserver l'« esprit » de l'analyse de variance (c'est-à-dire de conserver le principe d'une formalisation d'un protocole expérimental) et de procéder à un ajustement aussi bien sur des variables qualitatives que quantitatives. Quand une analyse de variance comporte ainsi une ou plusieurs variables explicatives quantitatives, on emploie préférentiellement le terme d'« analyse de covariance ».

En pratique : l'expérience suivante a pour but de comparer, dans le temps, l'effet de quatre traitements susceptibles de retarder l'atrophie musculaire consécutive à la dénervation expérimentale d'un des principaux muscles d'une patte arrière de rat.

Quarante-huit animaux sont ainsi séparés en quatre groupes au moyen d'un tirage au sort, chaque groupe recevant : une forte dose d'atropine (trt = a), une dose modérée de quinidine (trt = b), une dose modérée d'atropine (trt = c) et, enfin, une solution saline servant de témoin (trt = d).

Au quatrième jour suivant la dénervation (jour = 4), quatre rats sont choisis au hasard dans chaque groupe, ils sont sacrifiés, et le poids de leur membre dénervé (variable « p_musc ») est mesuré. La même procédure est réalisée au huitième et au douzième jour (« jour » = 8 et 12).

Si l'objectif était simplement de rechercher l'effet conjoint des facteurs « traitement » et « jour », un plan factoriel d'ordre 2 serait suffisant. On peut cependant supposer que la variabilité des poids initiaux des rats va bruiter la mesure finale, il serait donc opportun de pouvoir ajuster sur la variable « p_init ». Cette dernière étant quantitative, nous nous trouvons ainsi face à une analyse de covariance. Elle s'écrit formellement :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + a \cdot x_{ijk} + e_{ijk}$$
$$(1 \leq i \leq 4, 1 \leq j \leq 3, 1 \leq k \leq 4)$$

où Y désigne la variable dépendante « p_musc », x_{ijk} la variable « p_init », α correspond à l'effet du facteur « traitement » et β à celui du facteur « jour » ; le paramètre « a » est le coefficient de régression de la variable « p_init » ([24]). Les données ainsi collectées prennent finalement la forme ([25]) :

obs	jour	trt	p_init	p_musc	obs	jour	trt	p_init	p_musc
1	4	a	217	0.94	25	8	c	178	0.67
2	4	a	246	1.16	26	8	c	188	0.72
3	4	a	256	1.26	27	8	c	250	1.08

[24] Les conventions d'écriture d'une analyse de covariance apparaissent ainsi intermédiaires entre celles d'une analyse de variance et celles d'une régression linéaire.

[25] DeLury, D.B. (1948) The analysis of covariance, *Biometrics*, 4, 153-170.

4	4	a	200	0.85	28	8	c	195	0.75
5	4	b	198	1.19	29	8	d	194	0.97
6	4	b	248	1.15	30	8	d	274	1.07
7	4	b	180	0.86	31	8	d	222	1.16
8	4	b	218	1.21	32	8	d	274	1.04
9	4	c	264	1.22	33	12	a	198	0.34
10	4	c	200	0.9	34	12	a	175	0.43
11	4	c	210	1	35	12	a	199	0.41
12	4	c	192	1	36	12	a	224	0.48
13	4	d	181	0.99	37	12	b	233	0.41
14	4	d	266	1.51	38	12	b	250	0.87
15	4	d	274	1.55	39	12	b	289	0.91
16	4	d	180	0.98	40	12	b	255	0.87
17	8	a	265	0.91	41	12	c	204	0.57
18	8	a	248	0.73	42	12	c	234	0.8
19	8	a	238	0.52	43	12	c	211	0.69
20	8	a	180	0.65	44	12	c	214	0.84
21	8	b	186	0.87	45	12	d	186	0.81
22	8	b	220	1.04	46	12	d	286	1.01
23	8	b	199	0.88	47	12	d	245	0.97
24	8	b	240	0.96	48	12	d	215	0.87

La procédure SAS, PROC GLM aboutit aux résultats (26) :

```
                General Linear Models Procedure
                   Class Level Information

                Class      Levels    Values ❶
                JOUR            3    4 8 12
                TRT             4    a b c d

                Number of observations in data set = 48

Dependent Variable: P_MUSC ❷

Source                  DF    Sum of Squares      Mean Square   F Value    Pr > F
Model                   12        2.82737825       0.23561485     18.06    0.0001

Error                   35        0.45656967       0.01304485
Corrected Total         47        3.28394792

                R-Square             C.V.          Root MSE          P_MUSC Mean
                0.860969         12.72875        0.11421404           0.89729167

Source                  DF       Type I SS      Mean Square   F Value    Pr > F

JOUR                     2      1.32687917       0.66343958     50.86    0.0001
TRT                      3      0.79308958       0.26436319     20.27    0.0001
JOUR*TRT                 6      0.15290417       0.02548403      1.95    0.0994
P_INIT                   1      0.55450533       0.55450533     42.51    0.0001
```

26 Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.


```
Source                  DF       Type III SS ❸      Mean Square   F Value    Pr > F

JOUR                     2      1.45417776       0.72708888     55.74    0.0001
TRT                      3      0.56025922       0.18675307     14.32    0.0001
JOUR*TRT                 6      0.12156682       0.02026114      1.55    0.0901
P_INIT                   1      0.55450533       0.55450533     42.51    0.0001

                T tests (LSD) for variable: P_MUSC ❹

NOTE: This test controls the type I comparisonwise error rate not the
      experimentwise error rate.

      Means with the same letter are not significantly different.

T Grouping                Mean     N  JOUR ❺

      A                1.11063    16  4

      B                0.87625    16  8

      C                0.70500    16  12

T Grouping                Mean     N  TRT

      A                1.07750    12  d ❻

      B                0.93500    12  b
      B
      B                0.85333    12  c

      C                0.72333    12  a
```

En ❶ sont introduites les variables explicatives *qualitatives* (il manque donc « p_init »). La variable à expliquer « p_musc » est en ❷.

Comme à l'habitude, les résultats sont en ❸, les effets « jour » et « trt » sont significatifs au seuil de 5 % avec $p < 0{,}0001$, la variable d'ajustement « p_init » aboutit elle aussi à $p < 0{,}0001$, mais ce résultat n'est pas intéressant en lui-même. L'interaction « jour × trt » n'est, quant à elle, pas significative.

L'observation des moyennes de la variable à expliquer « p_musc », calculées pour chaque niveau des facteurs « jour » et « trt », nous permet de préciser ces résultats en ❹. Comme prévu, plus le temps s'écoule, plus le muscle s'atrophie ❺ ; en ce qui concerne les traitements ❻, l'atropine à forte dose semble plus efficace qu'à dose modérée, viennent ensuite une dose modérée de quinidine et enfin la solution témoin.

A propos de la régression linéaire

Nous allons maintenant aborder quelques remarques ou précisions supplémentaires sur la régression, en examinant notamment les relations que cette technique entretient avec un test t ou un coefficient de corrélation.

Corrélation et régression linéaire

En quelques mots : corrélation et régression linéaire sont deux notions souvent confondues, cela n'est pas sans raison. Ainsi, quand deux variables X et Y sont parfaitement corrélées (r = 1 ou −1), nous avons vu, dans la première partie traitant des méthodes univariées, que ces deux variables sont linéairement déterminées, c'est-à-dire que $Y = a_0 + a_1X$.

Deux remarques sont susceptibles de préciser cette proximité entre corrélation et régression linéaire.

– Si X et Y sont deux variables aléatoires dont les réalisations sont notées x_i et y_i, il est possible de calculer le coefficient de corrélation r du couple (X, Y) à partir de la régression $y_i = a_0 + a_1x_i + e_i$; r vaut ainsi :

$$r = a_1 \cdot s(X) / s(Y)$$

où s(X) et s(Y) sont les écarts types de X et Y. *On remarque ainsi que si X et Y ont même variance, le coefficient de corrélation de X avec Y est égal à a_1, la pente de la droite de régression $y_i = a_0 + a_1x_i + e_i$.*

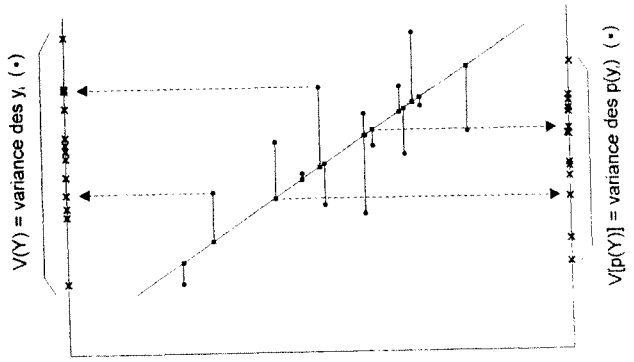

Fig. 1.3 — *Relation entre corrélation et régression : $r^2 = V[p(Y)]/V(Y)$ où les réalisations (x_i, y_i) de (X, Y) sont représentées par des « • » et les projections $p(x_i, y_i)$ sur la droite de régression de Y par rapport à X sont représentées par des « ■ ».*

– De plus, r quantifie la proximité qu'il y a entre la droite de régression $y_i = a_0 + a_1x_i + e_i$ et le nuage de points correspondant aux données. Plus précisément, si V(Y) est la variance des données observées et V[p(Y)] la variance des points projetés sur la droite de régression (voir fig. 1.3.), alors :

$$r^2 = V[p(Y)] / V(Y).$$

C'est pour cela que l'on dit parfois que r^2 représente la part (le pourcentage) de variance expliquée par le modèle de régression.

Ajuster une corrélation : le coefficient de corrélation partielle

En quelques mots : dans le cas multivarié ajusté, le coefficient de corrélation se transforme en coefficient de corrélation partielle ; ce dernier permet de quantifier l'association monotone entre deux variables quantitatives après ajustement sur une ou plusieurs covariables. Numériquement, le coefficient de corrélation partielle provient directement d'équations de régression linéaire ([27]).

En pratique : dans notre exemple recherchant une relation entre dépression (score HDRS) et monotonie de la voix (Af0s), nous avons vu qu'un ajustement sur les variables « traitement médicamenteux » améliorait la pertinence et la significativité des résultats. Si l'on désire maintenant quantifier numériquement cette liaison par un coefficient de corrélation, un ajustement doit, logiquement, aussi être proposé.

Calculons tout d'abord le coefficient de corrélation non ajusté ; le logiciel SPSS nous donne ([28]) :

```
                - -  Correlation Coefficients  - -

              AF0S          HDRS
AF0S          1.0000        -.1848  ❶
             (  116)       (  102)
             P= .          P= .063  ❷
HDRS           -.1848        1.0000
             (  102)       (  109)
             P= .063       P= .

(Coefficient / (Cases) / 2-tailed Significance)
" . " is printed if a coefficient cannot be computed
```

[27] Si l'on désire estimer le coefficient de corrélation partielle entre Y et X_1 après ajustement sur X_2, il faut calculer la première régression $x_{1i} = a_0 + a_2x_{2i} + e_i$, puis la seconde régression $y_i = a'_0 + a'_2x_{2i} + e'_i$, pour finalement estimer le coefficient de corrélation entre les e_i et les e'_i.

[28] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

Le coefficient de corrélation ❶ calculé entre « Af0s » et « HDRS » vaut ainsi : – 0,185 ; avec p = 0,063 (❷).

Calculons maintenant le coefficient de corrélation partielle :

```
- - -  P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S  - - -

Controlling for..    BENZO     NEUROL  TRICYC  SEROT ❸

                AF0S        HDRS

AF0S          1.0000      -.2749❶
              (    0)     (   62) ①
              P= .        P= .028❷

HDRS           -.2749      1.0000
              (   62)     (    0)
              P= .028     P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Le coefficient de corrélation partielle ❶ reliant « Af0s » et « HDRS », après ajustement sur les quatre variables ❸ vaut, lui, – 0,275 avec p = 0,028 ❷.

La régression pratiquée plus haut montrait que l'ajustement sur les variables « traitements médicamenteux » améliorait la significativité de l'association entre « Af0s » et « HDRS » ; le calcul de r nous permet maintenant de quantifier l'augmentation de la force de cette association : r passe de – 0,185 à – 0,275.

Le test t : un cas particulier de régression

En quelques mots : considérons une mesure Y effectuée dans deux groupes de sujets (définis par les relations : groupe = 1 et groupe = 2). Pour comparer la moyenne de Y dans ces deux groupes, on peut utiliser un test t. De façon équivalente on peut considérer la régression de Y sur la variable « groupe » : $y_i = a_0 + a_1 \cdot \text{groupe}_i + e_i$ (pour le sujet numéro i, groupe_i vaut 1 si le sujet appartient au groupe 1 et 2 dans le cas contraire), et tester la nullité de la pente a_1 de cette droite de régression.

En pratique : dans la partie consacrée aux statistiques univariées, nous avons comparé l'âge de patients consultant en médecine générale en fonction de l'existence ou non d'antécédents de cardiopathie ischémique. Il est possible de représenter graphiquement le couple de variables (*cardisch*, âge) ; nous obtenons ainsi :

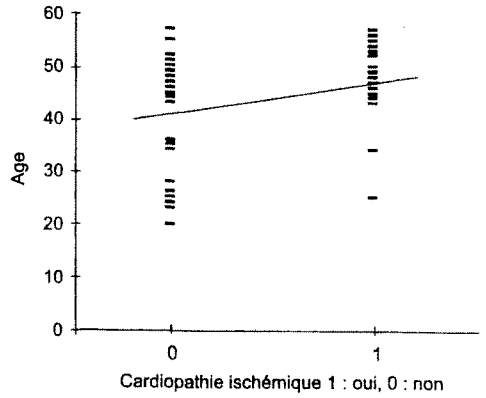

Fig. 1.4 — *Comparer deux moyennes revient à tester la nullité de la pente d'une droite de régression.*

La droite de régression a une pente positive, cela signifie que l'âge moyen des patients « cardiaques » est supérieur à celui des témoins. Une telle différence est-elle significative ? Pour le savoir, il suffit de tester le coefficient a_1 de l'équation : $\hat{\text{age}}_i = a_0 + a_1 \cdot \text{cardisch}_i + e_i$. Le logiciel SPSS nous donne :

```
          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Listwise Deletion of Missing Data
Equation Number 1     Dependent Variable..   AGE
Block Number  1.  Method:  Enter       CARDISCH


Variable(s) Entered on Step Number
   1..     CARDISCH

Multiple R             .42021
R Square               .17657
Adjusted R Square      .16043
Standard Error        9.73623

Analysis of Variance
                 DF        Sum of Squares        Mean Square
Regression        1           1036.70180         1036.70180
Residual         51           4834.50575           94.79423

F =      10.93634       Signif F =  .0017


------------------ Variables in the Equation ------------------
Variable             B         SE B        Beta           T  Sig T
CARDISCH      8.885057     2.686732     .420207  ❶3.307  .0017
(Constant)   40.448276     1.807973              22.372  .0000

End Block Number   1   All requested variables entered.
```

Le résultat est en ❶, soit t = 3,31.

Quel serait le verdict d'un simple test t ? SPSS nous donne cette fois-ci :

t-test for Equality of Means					95%
Variances	t-value	df	2-Tail Sig	SE of Diff	CI for Diff
Equal	3.31❷	51	.002	2.687	(-14.280, -3.490)
Unequal	3.42	49.79	.001	2.600	(-14.109, -3.661)

En ❷ nous retrouvons t = 3,31.

Le codage de variables qualitatives

En quelques mots : quand une analyse de variance a pour but d'expliquer une variable quantitative Y par deux facteurs binaires A et B (A = ± 1 et B = ± 1), nous avons vu page 105 que la traditionnelle relation :

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \ (i = 1, 2 ; j = 1, 2 ; 1 \leq k \leq n / 4)$$

pouvait aussi s'écrire sous la forme d'une régression linéaire :

$$y_i = \mu + \alpha \cdot a_i + \beta \cdot b_i + e_i \ (1 \leq i \leq n)$$

y_i, a_i et b_i désignant les réalisations de Y, A et B ; α et β correspondant, quant à eux, aux coefficients des variables A et B.

Si A n'est plus binaire (par exemple, A = – 1, 0 et 1), comment trouver une correspondance formelle entre analyse de variance et régression linéaire ? L'équation $y_i = \mu + \alpha \cdot a_i + \beta \cdot b_i + e_i$ n'est plus appropriée, elle considère en effet implicitement la variable A comme une variable quantitative ([29]) : A = 1 serait alors plus proche de A = 0 que de A = – 1, ce qui n'est en général pas le cas si A est authentiquement qualitative.

La solution réside dans un « codage » de A au moyen de deux variables qualitatives binaires A_1 et A_2, définies de la façon suivante :

- si A = 1 alors $A_1 = 1$ et $A_2 = 0$;
- si A = 0 alors $A_1 = 0$ et $A_2 = 1$;
- si A = – 1 alors $A_1 = 0$ et $A_2 = 0$ ([30]).

A partir de la régression : $y_i = \mu + \alpha_1 \cdot a_{1i} + \alpha_2 \cdot a_{2i} + \beta \cdot b_i + e_i$, il est à présent possible de tester l'effet du facteur A dans l'analyse de variance ci-dessus en considérant l'hypothèse nulle : ($\alpha_1 = 0$ *et* $\alpha_2 = 0$).

[29] Nous avons vu page 106 qu'une régression linéaire ne pouvait prendre en compte que des variables qualitatives *binaires*.

[30] Il existe en fait un grand nombre de codages équivalents, voir p.221.

Ce type de codage se généralise aisément à une variable qualitative à k classes. Ainsi, dans notre premier exemple d'analyse de variance portant sur une analyse de variance à un facteur (le facteur « milieu ») et comparant les moyennes des rendements de quatre milieux nutritifs sur une culture de fibroblastes (milieu = « humain » pour un milieu à base de sérum humain, milieu = « veau5 » pour un milieu à base de sérum de veau fœtal 5 %, milieu = « veau10 » pour un sérum de veau fœtal 10 % et milieu = « mini » pour un milieu minimum). Si α représente l'effet du facteur « milieu », nous avions considéré le modèle :

$$y_{ij} = \mu + \alpha_i + e_{ij} \ (1 \leq i \leq 4, 1 \leq j \leq n / 4).$$

En définissant maintenant A, A', A" par :

- si milieu = humain, alors A = 1, A' = 0 et A" = 0 ;
- si milieu = veau5, alors A = 0, A' = 1 et A" = 0 ;
- si milieu = veau10, alors A = 0, A' = 0 et A" = 1 ;
- si milieu = mini, alors A = 0, A' = 0 et A" = 0.

Tester la nullité de l'effet « milieu » par une analyse de variance revient à tester la nullité des trois coefficients α *et* α' *et* $\alpha"$ dans la régression linéaire :

$$y_i = \mu + \alpha \cdot a_i + \alpha' \cdot a'_i + \alpha" \cdot a"_i + e_i,$$

et permet de tester par là même l'égalité des quatre milieux de culture en termes de productivité.

En pratique : revenons à l'expérience ci-dessus. La procédure SAS PROC REG nous permet d'estimer la régression : $y_i = \mu + \alpha \cdot a_i + \alpha' \cdot a'_i + \alpha" \cdot a"_i + e_i$. Nous obtenons ainsi les résultats (seules les portions intéressantes ont été retenues) :

Parameter Estimates

Variable	DF	Parameter❶ Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEPT	1	4.125556	0.06881492	59.951	0.0001
A	1	1.497778	0.09731900	15.390	0.0001
A'	1	1.244444	0.09731900	12.787	0.0001
A"	1	1.621111	0.09731900	16.658	0.0001

Dependent Variable: LOGCELLS

Test: GLOBAL❷ Numerator: 4.9809 DF: 3 F value: 116.8700❹

Denominator: 0.042619 DF: 32 Prob>F: 0.0001❸

En ❶ nous trouvons une estimation de μ (« INTERCEPT »), α, α' et $\alpha"$; ces résultats ne sont ici qu'anecdotiques, le test intéressant étant le test global des trois égalités ($\alpha = 0$ *et* $\alpha' = 0$ *et* $\alpha" = 0$) que nous trouvons en ❷. Le « p » est en ❸ ($p < 0{,}0001$), il correspond à un $F(3,32) = 116{,}87$ (❹).

L'analyse de variance testant l'effet du facteur « milieu » nous avait donné les résultats :

Source	DF	Anova SS	Mean Square	F Value	Pr > F
MILIEU	3	14.94280833	4.98093611	116.87❷	0.0001❶

Ici aussi, le « p » vaut p < 0,0001 (❶) et le F, 116,87 (❷).

Le prix à payer

Un simple test t ou un test du chi-2 ont leurs conditions de validité, il en est de même d'une régression linéaire ([31]). Cette technique étant en outre d'un maniement plus délicat, il faudra être encore plus exigeant sur les conditions de son utilisation. Une grande prudence devra ainsi s'exercer en amont et en aval du recueil des données :

- en amont, car il est toujours souhaitable de réfléchir, avant le début de l'étude, aux variables d'ajustement qu'il faudra utiliser ([32]) ;
- en aval, pour vérifier les conditions de validité de la régression linéaire : on qualifie souvent cette vérification de « diagnostic de régression » ([33]).

La vérification des hypothèses sous-jacentes

En quelques mots : ces hypothèses sont de trois ordres :

- la normalité des résidus e_i définis par : $y_i = a_0 + a_1x_{1i} + ... + a_px_{pi} + e_i$. Cette normalité sera généralement évaluée à l'aide d'histogrammes ;
- l'indépendance de var(e_i) avec y_i ainsi qu'avec les x_{ji} (condition portant le nom d'homoscédaticité). Cette condition sera évaluée graphiquement et, en cas de doute, à l'aide d'un test approprié ;
- l'indépendance des résidus e_i. Cette dernière condition est cependant difficile à valider, certains graphiques peuvent être utilisés. Il existe aussi un test historique utile dans un cas particulier fréquent : celui où les données sont temporelles, il s'agit du test de Durby et Watson.

En pratique : dans notre exemple étudiant les relations entre score dépressif et monotonie de la voix, nous avons vu la pertinence d'un ajustement sur quatre variables binaires traduisant la prise de psychotropes.

Regardons maintenant si les différentes conditions nécessaires à l'utilisation d'un tel ajustement sont remplies. La régression correspondante était :

[31] Nous confondrons dans ce chapitre l'analyse de variance et la régression linéaire.

[32] Nous avons vu plus haut que si l'on essaie toutes les combinaisons possibles de variables d'ajustement, il est possible que, fortuitement, l'une d'entre elles modifie les résultats dans un sens qui nous est avantageux. Mais quelle est la validité d'une telle approche ?

[33] Les données et les syntaxes sas et R des exemples ci-dessous sont disponibles sur le site Internet du livre.

$$HDRS_i = a_0 + a_1 \cdot Af0s_i + a_2 \cdot AD\_Tricy_i + \\ + a_3 \cdot AD\_Serot_i + a_4 \cdot Neurol_i + a_5 \cdot Benzo_i + e_i.$$

Vérification de la normalité des résidus e_i

Nous avons vu dans la partie univariée qu'il était discutable de tester statistiquement une normalité, nous nous limiterons donc ici à une approche graphique, utile pour dépister des écarts patents. Le premier outil à envisager est un simple histogramme :

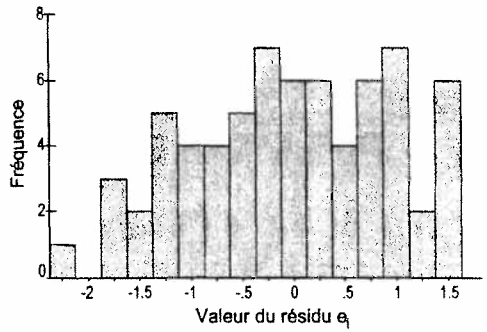

Fig. 1.5 — *Evaluation de la normalité des résidus à l'aide d'un histogramme.*

Le second est un tracé normalisé :

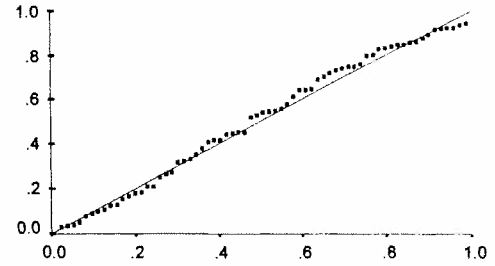

Fig. 1.6 — *Evaluation de la normalité des résidus à l'aide d'un tracé normalisé.*

Tous deux sont en faveur d'une normalité « acceptable ».

Homoscédaticité

La vérification de l'homoscédaticité se fait avant tout graphiquement, par la représentation des couples (y_i, e_i), puis par celle des couples (x_{ji}, e_i) $(1 \leq j \leq p)$.

En pratique, plutôt que d'étudier y_i et e_i, on préfère considérer de façon équivalente y'_i la prédiction de y_i $(y'_i = y_i - e_i)$ [34] ; ainsi que les résidus « standardisés » $e'_i = e_i$ / écart type(e_i). Dans notre exemple, les couples (y'_i, e'_i) ont la représentation graphique suivante :

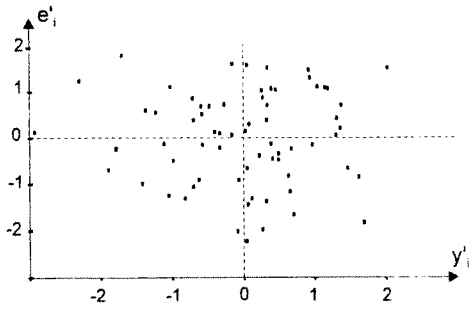

Fig. 1.7 — *Etude des résidus e'_i en fonction de y'_i.*

Plus proche de la solution idéale... :

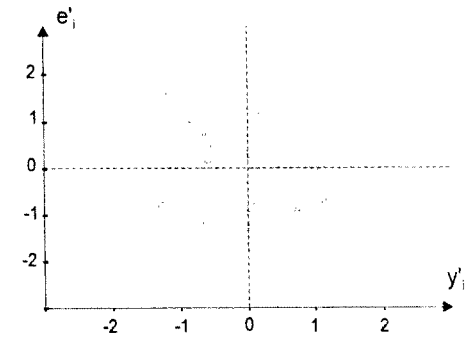

Fig. 1.8 — *Répartition acceptable de résidus.*

[34] Si $y_i = a_0 + a_1x_{1i} + ... + a_px_{pi} + e_i$, la prédiction de y_i est égale à $a_0 + a_1x_{1i} + ... + a_px_{pi}$ (ou de façon équivalente $y_i - e_i$). Plus précisément, si, pour une observation, on ne connaît que la valeur des variables explicatives x_i et pas celle de y_i, la combinaison linéaire des x_i la plus proche de y_i est égale à $a_0 + a_1x_{1i} + ... + a_px_{pi}$.

...que d'une situation anormale du type :

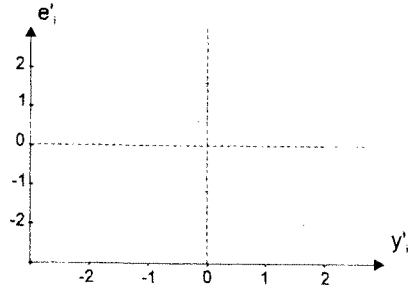

Fig. 1.9 — *Exemple de rejet de l'homoscédaticité, la variance de e'_i croissant avec y'_i.*

Il existe en outre des tests d'homoscédaticité ; le logiciel SAS, PROC REG propose ainsi dans notre exemple :

```
Dependent Variable: HDRS
            Test of First and Second Moment Specification
      DF:   16   Chisq Value:      16.5186   Prob>Chisq:   0.4174 ❶
```

En ❶, on ne peut rejeter la condition d'homoscédaticité, ce que confirmait la représentation graphique précédente.

Indépendance des résidus

Cette étape est en pratique délicate. Si l'on suspecte une corrélation des résidus avec, par exemple, la variable Y, c'est le diagramme (y'_i, e'_i) que l'on observera. Un cas typique étant celui où l'on obtient :

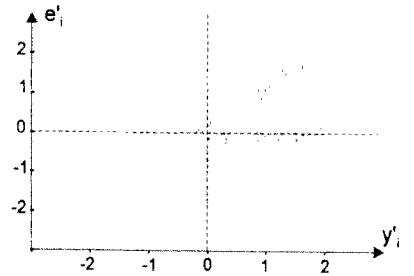

Fig. 1.10 — *Exemple de non-indépendance des résidus e'_i.*

Des résidus ayant des valeurs de y_i voisines sont ici plus proches que ne le voudrait le hasard.

Le modèle est-il robuste ?

En quelques mots : si, en enlevant seulement un ou deux sujets, les coefficients a_i d'une régression $y_i = a_0 + a_1x_{1i} + ... + a_px_{pi} + e_i$ changent du tout au tout, on dit que la régression est peu robuste. Une certaine suspicion entoure alors les résultats qui deviennent moins facilement interprétables.

Pour évaluer la robustesse d'un modèle de régression linéaire, il est possible de procéder de la façon suivante :

- tour à tour, chaque sujet est retiré du jeu de données ;
- on mesure alors les perturbations que cela engendre sur les paramètres à estimer ;
- les individus « sensibles », à l'origine des perturbations les plus fortes, sont finalement observés de près : on recherche notamment d'éventuelles erreurs de mesures ou de recrutement. Si rien de tel n'est décelé, une discussion doit être menée pour établir si l'instabilité observée est rédhibitoire ou si elle est acceptable.

En pratique : reprenons notre exemple recherchant une relation entre la monotonie de la voix (mesurée par l'Af0s) et l'intensité d'une dépression (mesurée par l'échelle de Hamilton, HDRS). Nous avons montré qu'après ajustement sur le traitement médicamenteux, une association significative était trouvée entre ces deux variables. Avant de conclure définitivement, il est important de vérifier la robustesse de notre modèle. La procédure SAS, PROC REG va nous permettre d'évaluer cette stabilité. Nous obtenons ainsi :

Model: MODEL1
Dependent Variable: HDRS ❶

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	257.73062	51.54612	1.236	0.3033
Error	62	2586.26938	41.71402		
C Total	67	2844.00000			

Root MSE	6.45864	R-square	0.0906
Dep Mean	21.00000	Adj R-sq	0.0173
C.V.	30.75543		

Parameter Estimates ❷

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	29.337322	3.61846271	8.108	0.0001
AFOS	1	-676.830367	300.67805053	-2.251	0.0279 ❸
TRICYC	1	1.015702	2.52212671	0.403	0.6885
SEROT	1	-2.071424	2.70312032	-0.766	0.4464
NEUROL	1	-0.359492	1.76185135	-0.204	0.8390
BENZO	1	-2.061915	1.64629642	-1.252	0.2151

❹

❺Obs	Residual	Rstudent ❻	Hat Diag H ❼	Cov Ratio	Dffits	INTERCEP Dfbetas	AFOS Dfbetas ❽
1	4.5407	0.7225	0.0604	1.1149	0.1833	0.0009	0.0729
2	9.0269	1.5109	0.1266	1.0125	0.5752	0.4250	-0.3966■
3	10.3463	1.8089	0.1870	0.9911	0.8675	-0.2847	0.2526
4	9.2041	1.4819	0.0574	0.9460	0.3657	0.2494	-0.2424
5	6.2348	1.0540	0.1596	1.1772	0.4593	-0.0508	0.0708
6	-5.2430	-0.8294	0.0468	1.0813	-0.1837	-0.1092	0.1026
7	0.9598	0.1500	0.0336	1.1383	0.0280	0.0081	-0.0058
8	-11.5138	-2.0150■	0.1787	0.9119	-0.9399	0.1002	0.0377
9	3.9417	0.6732	0.1853	1.2945	0.3210	-0.1019	0.0620
10	2.3841	0.3825	0.0815	1.1832	0.1140	-0.0297	0.0491
11	-1.3388	-0.2091	0.0323	1.1344	-0.0382	-0.0018	-0.0022
12	5.4610	0.8734	0.0665	1.0961	0.2331	0.0422	-0.0083
13	1.2979	0.2092	0.0913	1.2081	0.0663	0.0401	-0.0347
14	6.4963	1.0254	0.0371	1.0333	0.2012	0.0857	-0.0730
15	6.1162	1.0474	0.1813	1.2100	0.4929	-0.0164	0.0997
16	-0.7632	-0.1214	0.0670	1.1799	-0.0325	-0.0076	0.0031
17	-3.8975	-0.6620	0.1766	1.2827	-0.3066	-0.0822	0.0382
18	-1.9654	-0.3450	0.2331■	1.4208	-0.1902	-0.0037	-0.0045
19	-0.8622	-0.1455	0.1718	1.3284	-0.0663	0.0005	0.0042
20	-14.0821	-2.2931■	0.0339	0.6948	-0.4293	-0.1294	0.0954
21	0.4302	0.0688	0.0771	1.1941	0.0199	0.0152	-0.0152
22	4.1728	0.6673	0.0708	1.1359	0.1842	-0.0267	0.0978
23	-4.2865	-0.6892	0.0807	1.1447	-0.2042	0.1319	-0.1583
24	8.1655	1.3081	0.0552	0.9884	0.3161	0.2030	-0.0882
25	-1.4875	-0.2411	0.1012	1.2196	-0.0809	0.0492	-0.0442
26	9.9099	1.5798	0.0339	0.8971	0.2959	0.0900	-0.0666
27	-1.3320	-0.2265	0.1838	1.3439	-0.1075	-0.0016	0.0184
28	-6.6229	-1.0712	0.0814	1.0732	-0.3189	0.0827	-0.1371
29	-2.8019	-0.4462	0.0671	1.1589	-0.1197	-0.0292	0.0126
30	-2.9598	-0.4653	0.0424	1.1271	-0.0980	-0.0527	0.0482
31	-4.0966	-0.6490	0.0538	1.1180	-0.1547	-0.0276	-0.0360
32	7.0134	1.1120	0.0428	1.0211	0.2353	-0.0869	0.1173
33	3.1483	0.4930	0.0341	1.1145	0.0926	-0.0110	0.0219
34	-7.9440	-1.2636	0.0435	0.9870	-0.2693	0.1023	-0.1372
35	9.3001	1.6011	0.1708	1.0385	0.7265	0.0614	0.0216
36	-2.4368	-0.3908	0.0806	1.1813	-0.1157	-0.0225	0.0399
37	0.6510	0.1038	0.0716	1.1862	0.0288	-0.0014	-0.0025
38	2.3547	0.3692	0.0385	1.1314	0.0739	0.0342	-0.0300
39	4.3436	0.6997	0.0838	1.1470	0.2116	0.0505	-0.0827
40	-0.9136	-0.1448	0.0600	1.1705	-0.0366	-0.0255	0.0249
41	6.7389	1.0826	0.0686	1.0559	0.2938	0.2161	-0.2140
42	7.5100	1.2362	0.1078	1.0652	0.4297	-0.3058	0.3599■
43	1.8446	0.2914	0.0534	1.1550	0.0692	0.0135	0.0150
44	6.6540	1.0643	0.0610	1.0514	0.2713	0.2010	-0.1107
45	3.4171	0.5558	0.1041	1.1938	0.1895	-0.0819	0.1140
46	-5.1945	-0.8473	0.1032	1.1460	-0.2874	-0.1913	0.1716
47	-5.7589	-0.9062	0.0347	1.0540	-0.1718	0.0257	-0.0462
48	-7.2931	-1.1746	0.0701	1.0367	-0.3225	-0.1147	0.0743
49	-9.1099	-1.4477	0.0340	0.9319	-0.2717	-0.0843	0.0631
50	-11.3329	-1.8649	0.0793	0.8586	-0.5473	-0.4765	0.3277■
51	-10.3715	-1.6730	0.0520	0.8885	-0.3919	-0.2091	0.0579
52	4.2882	0.6843	0.0667	1.1283	0.1830	0.1460	-0.0892
53	0.7946	0.1241	0.0326	1.1379	0.0228	0.0000	0.0024
54	9.3421	1.5273	0.0838	0.9607	0.4620	0.1103	-0.1806
55	-11.8373	-2.0648	0.1706	0.8863	-0.9365	-0.0988	-0.0064
56	-6.2089	-0.9897	0.0567	1.0623	-0.2427	0.1278	-0.1597
57	-5.8619	-0.9218	0.0328	1.0491	-0.1699	-0.0390	0.0241
58	0.3276	0.0511	0.0323	1.1390	0.0094	0.0015	-0.0007
59	-0.8198	-0.1289	0.0453	1.1530	-0.0281	0.0115	-0.0151
60	3.7015	0.5864	0.0549	1.1278	0.1413	-0.0723	0.0908
61	-3.0592	-0.4809	0.0418	1.1248	-0.1005	0.0353	-0.0483
62	5.4023	0.8500	0.0359	1.0656	0.1641	-0.0329	0.0530
63	0.7430	0.1314	0.2453■	1.4583	0.0749	-0.0528	0.0640
64	-8.2239	-1.3161	0.0529	0.9841	-0.3111	-0.0668	-0.0612
65	6.9588	1.1118	0.0572	1.0368	0.2739	0.1876	-0.0913
66	-7.3975	-1.3288	0.2479■	1.2352	-0.7628	0.0838	0.0409
67	2.4159	0.4271	0.2429■	1.4304	0.2419	0.0547	-0.0490
68	-8.6209	-1.3803	0.0512	0.9662	-0.3206	-0.1054	-0.0254

En ❶ nous retrouvons la variable à expliquer HDRS, en ❷ les résultats du modèle linéaire, et plus particulièrement, en ❸, les caractéristiques du coefficient correspondant à la variable Af0s.

Les résultats qui nous intéressent commencent en ❹. Pour chaque patient ❺, nous disposons entre autres :

- de ❻, le résidu standardisé (des résultats élevés en valeur absolue indiquent que le patient ❺ est mal « compris » par le modèle) ;
- de ❼, qui mesure l'influence qu'exerce le patient ❺ sur l'ensemble du modèle ;
- de ❽, l'influence qu'exerce le patient ❺ sur l'estimation du paramètre qui nous intéresse plus particulièrement : le coefficient de la variable Af0s.

Comment interpréter ces résultats ? Nous avons noté dans le tableau ❹ par un ▪ les patients qui présentaient les valeurs les plus élevées pour ❻, ❼, ou ❽. Regardons de plus près ces patients.

– Les patients 8 et 20 sont les sujets ayant un résidu standardisé ❻ particulièrement élevé. Les valeurs étant tout juste supérieures à 2 en valeur absolue, ces patients ne peuvent pas être considérés comme extrêmement atypiques ([35]), et de plus, les modifications du coefficient de l'Af0s qu'ils entraînent en ❽ sont faibles. Ces deux observations ne sont donc pas inquiétantes.

– Les patients 18, 63, 66 et 67 ont un coefficient ❼ élevé, mais puisqu'ils altèrent peu le coefficient de l'Af0s (❽), ils n'introduisent pas d'instabilité dans la liaison entre score de dépression et monotonie de la voix.

– Il nous reste finalement les patients 2, 42 et 50 dont la caractéristique est justement d'influencer fortement le coefficient de l'Af0s. Il est utile de regarder à quoi correspondent ces sujets :

obs	Af0s	HDRS	AD tricy	AD serot	Neurol	Benzo
2	0.591	34	0	0	1	0
42	1.594	24	0	0	0	1
50	0.739	13	0	0	0	0

Le patient 2 est typique : très déprimé (HDRS = 34), il présente une voix extrêmement monotone (Af0s = 0,591), et consomme en outre des neuroleptiques. Le patient 42 a une voix très modulée (Af0s = 1,594) malgré une dépression notable (HDRS = 24). A l'opposé, le patient 50 a une voix très monotone (Af0s = 0,739) malgré une symptomatologie légère (HDRS = 13) et aucune consommation de médicament.

En résumé, si ces patients correspondent tous les trois à des situations extrêmes, ils n'en sont pas moins authentiques ; il n'est donc absolument pas question de s'en débarrasser. Il est par contre possible de les enlever provisoirement du jeu de données pour évaluer la stabilité des résultats, et notamment les perturbations que cela entraîne sur la significativité du coefficient de l'Af0s. Le programme SAS, PROC REG nous donne alors :

[35] On sait, en effet, que 5 % des mesures d'une variable normale standardisée sont supérieures à 2 en valeur absolue. Un tel degré d'atypie ne peut donc pas être considéré comme « extraordinaire ».

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	291.55400	58.31080	1.490	0.2069
Error	59	2309.46139	39.14341		
C Total	64	2601.01538			

Root MSE	6.25647	R-square	0.1121
Dep Mean	20.87692	Adj R-sq	0.0368
C.V.	29.96835		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	30.767494	3.89799494	7.893	0.0001
AF0S	1	-775.650804 ❶	320.62320344	-2.419	0.0187 ❷
TRICYC	1	1.126266	2.45819376	0.458	0.6485
SEROT	1	-2.346312	2.63672238	-0.890	0.3772
NEUROL	1	-0.945448	1.74944014	-0.540	0.5909
BENZO	1	-2.575330	1.67395088	-1.538	0.1293

Les nouveaux résultats sont proches des anciens ; en particulier, le coefficient de l'Af0s passe simplement de – 676,8 à – 775,7 (❶), le « p » correspondant passant, lui, de 0,0279 à 0,0187(❷).

En conclusion, la stabilité du modèle paraît acceptable.

Les variables explicatives ne sont-elles pas redondantes ?

En quelques mots : quand certaines variables explicatives sont par trop similaires, la qualité numérique du modèle devient médiocre. A l'extrême, si l'une des variables explicatives est une combinaison linéaire des autres ([36]), le modèle est indéterminé. Ce problème correspond à ce que l'on nomme classiquement une situation de multicolinéarité.

Le plus souvent, nous savons *a priori* que certaines variables sont susceptibles d'être proches, nous pouvons ainsi les éliminer avant tout calcul. Parfois, cependant, cela n'est pas prévisible. Il est alors prudent de procéder à une recherche de multicolinéarité, qui se fait en comparant les coefficients a_j de la régression $y_i = a_0 + a_1x_{1i} + ... + a_px_{pi} + e_i$ à leurs écarts types. Ainsi, si pour un ou plusieurs coefficients a_j l'on observe : écart type$(a_j) / a_j > 100$ (par exemple), il sera nécessaire de reconsidérer les variables explicatives à retenir.

[36] Ce qui est très rare en pratique.

En pratique : dans notre modèle recherchant une liaison entre monotonie de la voix (Af0S) et score de dépression (HDRS) avec ajustement sur le traitement psychotrope, les résultats que nous avons obtenus précédemment étaient :

Parameter Estimates

Variable	DF	Parameter Estimate ❷	Standard Error ❶	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	29.337322	3.61846271	8.108	0.0001
AFOS	1	-676.830367	300.67805053	-2.251	0.0279
TRICYC	1	1.015702	2.52212671	0.403	0.6885
SEROT	1	-2.071424	2.70312032	-0.766	0.4464
NEUROL	1	-0.359492	1.76185135	-0.204	0.8390
BENZO	1	-2.061915	1.64629642	-1.252	0.2151

Les écarts types ❶ n'étant pas « grands » devant les estimations des coefficients ❷, cela engage à être confiant.

2.
Modèle linéaire généralisé : régressions logistique et de Poisson

La régression linéaire a été développée dans le but de caractériser les relations associant une variable Y *quantitative*, à une série de variables X_1, X_2, ..., X_p au moyen d'un modèle $Y = a_0 + a_1X_1 + \ldots + a_pX_p + \varepsilon$, où ε est un terme de « bruit » dont les propriétés ont été définies au chapitre précédent. Si Y n'est plus une variable quantitative, où si ε ne vérifie pas les propriétés imposées par le modèle de régression linéaire, alors il est nécessaire de développer une classe de modèles plus généraux, que l'on dénomme précisément modèles linéaires généralisés. Les deux principaux modèles linéaires généralisés utilisés en recherche biomédicale sont la régression logistique et la régression de Poisson.

La régression logistique permet de relier une variable Y qualitative (en général binaire) à une série de variables X_1, X_2, ..., X_p. En épidémiologie, cette modélisation est très utilisée pour relier la survenue d'une maladie (variable binaire) à un groupe de facteurs de risques, en caractérisant notamment le poids spécifique de chaque facteur de risque.

Il n'est pas rare que des variables rencontrées dans des études biomédicales soient des comptages : nombre d'infections urinaires contractées par un patient hospitalisé en gériatrie, nombre d'épisodes dépressifs survenant entre 20 et 40 ans chez des sujets de la population générale, etc. Si l'on souhaite relier ce type de variable à un groupe de variables explicatives en caractérisant le poids spécifique de chacune d'entre elles, c'est une régression de Poisson qu'il faut alors utiliser.

Nous allons voir dans la suite de ce chapitre que la pratique des régressions logistique et de Poisson est proche de la pratique de la régression linéaire. La notion d'ajustement est toujours aussi centrale. La régression linéaire conduit à généraliser la notion de corrélation et de test t ; la régression logistique aboutit, elle, à une généralisation de la notion d'odds-ratio et de test du chi-2. Nous verrons enfin qu'il est important de vérifier la validité des modèles logistique et de Poisson ; la notion de « diagnostic » de régression est aussi importante que dans le cas linéaire.

La régression logistique

Introduction

Pour illustrer cette introduction, considérons une étude fictive dont le but serait d'expliquer l'existence d'une pathologie coronarienne par la présence de