Linear models :

Regression, Analysis of Variance, Asymptotic Theory

Jean-Marc Azaïs

Institut de Mathématiques
Université de Toulouse

Chapter 1

Definition of Linear models

In this chapter general linear models are defined . A very short list of fundamental formulas and properties is given

1 Matrix form of basis models

1.1 Simple linear regression.

The word "regression" comes mainly from the work of Sir Francis Galton with the paper *Regression towards mediocrity in hereditary stature.* Galton's first studied the sizes of daughter peas against the sizes of mother peas and then the stature of persons. Galton observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. Rather, the characteristics in the offspring regress towards a mediocre point (a point which has since been identified as the mean). This is the ethymology of the (strange) word "regression" .

Let us consider a more pedagogical example of the regression of the blood pressure on age

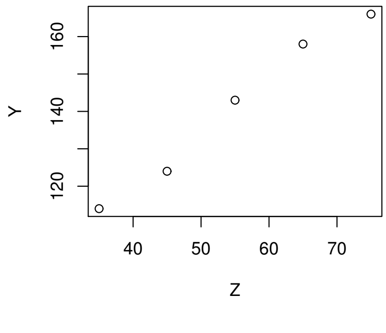


Figure 1.1: Scatter point of (Z_i, Y_i), Z_i mean age of group of women i, Y_i mean blood pressure

The linear relation between the age and the blood pressure leads us to set the following model

$$Y_i = \beta_1 + \beta_2 Z_i + \varepsilon_i \quad i = 1, s, n = 5.$$

Let us consider the vectors $Y = (Y_i)_{1 \leq i \leq 5}$ and $\varepsilon = (\varepsilon_i)_{1 \leq i \leq 5}$.

The model above can be written

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ 1 & Z_3 \\ 1 & Z_4 \\ 1 & Z_5 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}.$$

Matricially :

$$Y = X\beta + \varepsilon \quad \text{with} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ 1 & Z_3 \\ 1 & Z_4 \\ 1 & Z_5 \end{pmatrix}. \tag{1.1}$$

N.B. To keep classical notation, we will denote matrix and vectors with the same kind of symbols. Nevertheless X will be in general a matrix, while Y et Z will be size-n-vectors.

1.2 One way analysis of variance

We consider the example of the measurement of the heights of several trees of three forests using the model

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where Y_{ij} is the height of the j th tree of Forest i and μ_i is the true (unobservable) mean of Forest i. This model can be written

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{25} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \\ Y_{35} \\ Y_{36} \\ Y_{37} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{34} \\ \varepsilon_{35} \\ \varepsilon_{36} \\ \varepsilon_{37} \end{pmatrix}$$

In matricial form

$$Y = X\beta + \varepsilon \text{ with } \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}. \tag{1.2}$$

N. B.: In the example above Y the coordinate of Y are indexed by two indices(i, j) and will still call it “vector” and not “matrix”. On one hand, strictly speaking, a vector is a member of a vectorial space that has to be closed under addition and multiplication, in that sense, matrices or even functions can be viewed as vectors. In that sense Y is indeed a vector. But on the other hand, we will use matrix calculation and its conventions that demand a vector to be a column vector. In this case it is necessary to “unroll” the two-way array Y in lexicographic order to make it a “vector ”in this strict sense, as it is done above.

1.3 Multiple linear regression

The main message of the two examples above is that the analysis of variance model and the simple regression model are very similar. In fact they are almost the same and the class of such model is even larger. Let us consider, for example, the observation of

- Y, a vector of the n yields of a chemical reaction (expressed as percentage);

- $Z^{(1)}$, a vector that consists of the n measurements of the associated temperature of the substratum ;

- $Z^{(2)}$, a vector that consists of the n measurements of the pH of the substratum.

We assume that the variable to be explained, the dependent variable, the yield Y depends linearly on temperature and pH (explanatory variables or independent variables) $Z^{(1)}$ and $Z^{(2)}$. We set the following multiple regression model :

$$Y_i = \beta_1 + \beta_2 Z_i^{(1)} + \beta_3 Z_i^{(2)} + \varepsilon_i, \tag{1.3}$$

for $i = 1, s, n$. In matrix form:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1^{(1)} & Z_1^{(2)} \\ 1 & Z_2^{(1)} & Z_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & Z_n^{(2)} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Or

$$Y = X\beta + \varepsilon \quad \text{with} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & Z_1^{(1)} & Z_1^{(2)} \\ 1 & Z_2^{(1)} & Z_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & Z_n^{(2)} \end{pmatrix}. \tag{1.4}$$

2 Linears models: basic definition and fundamental hypotheses

Fundamental definition: we will say that the variable Y that consists of n observations $Y_i, i = 1, ...n$ obeys to a linear model in the statistical sense if we can write :

$$Y = X\beta + \varepsilon \tag{1.5}$$

where

i. X is a known n, k matrix $k < n$

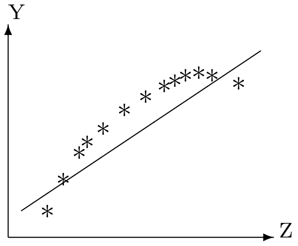
ii. β is an unknown vector of size k

iii. the random vector ε, that represents the error of the model, satisfies the **four fundamental hypotheses** .

- **FH1 : The errors are centered**

$$\mathbb{E}\left(\varepsilon\right) = 0.$$

In other words it means that the assumed model is correct in the sense that no relevant effect has been forgotten. A counter-example is given by the following linear regression example:



In this example it is clear that a curvature has been forgotten and that a better model would be

$$Y_i = \beta + \beta_2 Z_i + \beta_3 (Z_i)^2 + \varepsilon_i.$$

- **FH2 : The variance of the error is constant (homoscedaticity)** :

$$\text{Var}(\varepsilon_i) = \sigma^2, \quad \text{for all } i.$$

In practice this fundamental assumption in one of the most difficult to check. In particular it is not in general automatically implied by a smpling design.(see examples after).

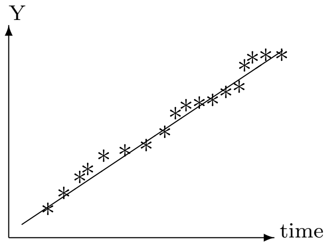
let us consider the following counter-example. The survival of insects to the administration of insecticides A and B. Four repetitons are performed yielding the following data.

	Survival rate	
	product A	product B
rep1	0,01	0,37
rep2	0.02	0.26
rep3	0.02	0.60
rep4	0.04	0.44
⋮	⋮	⋮

At first sight , Insecticide A is much more efficient than insecticide B implying that the survival rate with A is close to 0 but also less variable than with Insecticide B. This is an heteroscedastic situation.

- **FH 3 : The variables ε_i are independent .**

It is generally considered than this assumption is true when each observation (statistical unit) is the resultant of an independent sampling. This is the case in the forest example if each tree has been correctly sampled (which is not that easy and demands spatial method and GPS positioning) . In contrast in temporal problems, as it is the case often in econometrics, some inertia may occur as in the following counter-example.



In this example you can easily check that if the curve is above (for example) the line at some time, it is more likely to be also above at the next time.

- **FH4 : The errors are Gaussian (or normal), i.e. :**

$$\varepsilon_i \sim \mathcal{N}(m_i, \sigma_i^2) \text{ for all } i,$$

Where m_i and σ_i^2 are some parameters. Consequently to **FH1** and **FH2** , in fact,

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ for all } i.$$

It is an easy consequence of **FH1-FH4** that ε is a Gaussian vector, more precisely

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

where I_n is the identity matrix of size n. Consequently Y is also a Gaussian vector

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

This last equality could have been chosen as a definition of a linear model, this is formally correct but in practice it is better to distinguish the four hypotheses. In particular, as we shall see, the hypothesis of Gaussianity **FH4** is of less importance, especially for large data. In several cases we shall consider non Gaussian linear model where **FH4** is simply removed or replaced by a weaker form, for example that the error are i.i.d with finite fourth moment.

To check **FH4** is not easy. For small data, classical normality tests as Kolmogorov-Smirnov or Shapiro-Wilks tests are not directly applicable because we observe not directly the errors but their estimation, the residuals. For large data theses tests are worthless because **FH4** it is not needed, see Chapter 6.

A crude graphical method to check **FH4** is to compute a Quantile-Quantile plot : Q-Q plot on the residuals

3 Fondamental formulas

3.1 Four formulas

We consider the general linear model given by (1.5) that we can call the X linear model, since X is known. X is often called the "design matrix". Of course we assume **FH1-FH4**. In addition, and for simplification, up to Chapter 5 we will assume that X is full ranked. That is equivalent to

- $Rank(X) = k$
- $Ker(X) = 0$
- $X'X$ is invertible, where the prime means the transpose.

The ingredients

- The first remark is that the linear model is a statistical model with $k+1$ parameters $\theta = (\beta, \sigma)$. In fact it is even an exponential model in the sense of theoretical statistics implying some optimality of estimators that will be admitted and described later.

-The method used is least squares method we define the residual sum of squares

$$RSS(\beta) = \| Y - X\beta \|^2 = (Y - X\beta)'(Y - X\beta),$$

and we minimize it. Let $\widehat{\beta}$ be the argument minimum and $RSS := RSS(\widehat{\beta})$ be the minimum.

The formulas

- **F1 :** The minimum of the sum of squares is attained at a single point

$$\widehat{\beta} = (X'X)^{-1}X'Y. \tag{1.6}$$

This formula has several consequences. Firstly, the solution is explicit that means that we can easily compute its distribution. Secondly, it is of low complexity because we have to solve a k, k linear system which is in general easy. So linear models can have large sizes and can adapt very well to the reality . Thirdly it implies that $\widehat{\beta}$ is itself Gaussian as a linear function of the Gaussian vector Y.

In fact numerically, we solve the **normal equations** : $X'X\widehat{\beta} = X'Y$.

- **F2 :**

$$\mathbb{E}\,(\widehat{\beta}) = \beta. \tag{1.7}$$

This is a direct consequence of (1.6). It means that the least square estimator is unbiased.

As announced this implies that the unbiased estimator, function of a sufficient statistics, is optimal (of minimum variance) among all unbiased estimators. This is the

Rao-Blackwell Theorem. This means that if $\widetilde{\beta}$ is another unbiased estimator and if C is any vector in $\mathbb{R}^k$ defining a linear combination $C'\beta$, then

$$\mathrm{Var}\,(C'\widetilde{\beta}) \geq \mathrm{Var}\,(C'\widehat{\beta}).$$

- **F3 :** $\mathrm{Var}\,(\widehat{\beta}) = \sigma^2(X'X)^{-1}$.

  This expression gives the variance-covariance matrix of the vector $\widehat{\beta}$ at the cost of the estimation of σ^2 . It achieves one of the most important goals of Statistics: not only estimate but estimate also the precision of the estimation.

- **F4 :** Set:

$$RSS = (Y - X\widehat{\beta})'(Y - X\widehat{\beta}) = \parallel Y - X\widehat{\beta} \parallel^2 = \parallel Y - \widehat{Y} \parallel^2,$$

  Then RSS is a random variable that is independent of $\widehat{\beta}$ and with distribution $\sigma^2\chi^2(n-k)$. This last distribution is a Chi-square distribution with $(n-k)$ degrees of freedom multiplied by the scalar factor σ^2.

  If you keep in mind that, because of the law of large number, a $\chi(d)$ is close to d , a natural unbiased estimator is

$$\widehat{\sigma}^2 = \mathrm{CMR} = \frac{RSS}{n-k} = \frac{\parallel Y - \widehat{Y} \parallel^2}{n-k}, \tag{1.8}$$

  This implies also, as for $\widehat{\beta}$,that it is optimal by sufficiency techniques.

Proof : We will make a geometrical proof of **F1-4** using orthogonal projections. Let us recall first that if $u \in \mathbb{R}^n$ is a vector and if E is a sub-linear space of $\mathbb{R}^n$ and P_E the orthogonal projector on E then $P_E u$ can be characterized by

-either : $P_E u$ belongs to E and $u - P_E u$ is orthogonal to E.

-or : $P_E u$ is the minimizer in $v \in E$ of the program
$\parallel u - v \parallel^2$ minimum.

The equivalence of the two characterization is due to the Pythagore Theorem.

We will prove the following lemma

Lemma 1.1 *Suppose that $E = [X] := Im(X)$ when X is a full rank matrix and $[X]$ is the space generated by the columns of X. Then*

$$P_E = X(X'X)^{-1}X'$$

Proof of the lemma

We use the first characterization. Since $P_E u$ belongs to E we search it as Xw. We want $u - Xw$ to be orthogonal to E, it suffices that it is orthogonal to the generating system $X_1, ...X_k$ where X_J is the jth column of X. So we have to solve

$$\text{for all } j = 1, ..., k : \langle u, X_J \rangle = \langle Xw, X_J \rangle.$$

This system can be rewritten

$$X'Xw = X'u \;\; ; \;\; w := (X'X)^{-1}X'u \;\; ; \; \mathbb{P}_E u := X(X'X)^{-1}X'u$$

∎

We now turn to the proof of **F1-F4**

F1 : Because of the characterization of the projection $\widehat{Y} := X\widehat{\beta} = P_{[X]}Y \in \mathbb{R}^n$, minimizing $\| Y - Y_X \|^2$. Then it suffices to apply the lemma.

F2 : Since the expectation is a linear operator it commutes with matrices:

$$\mathbb{E}\,(\widehat{\beta}) = \mathrm{E}\left[(X'X)^{-1}X'Y\right] = (X'X)^{-1}X'\mathbb{E}\,(Y) = (X'X)^{-1}X'(X\beta) = \beta.$$

On the other side, at a cost of change of parameterization, see Coursol [6] p. 13-14 or, for example Bickel and Docksum [4], the linear model is a statistical exponential family with sufficient statistics $X\widehat{\beta}$ and RSS. This implies that $\widehat{\beta}$ which is a linear function of $X\widehat{\beta}$ and is an unbiased estimator with minimal variance among the unbiased estimators.

F3 :

$$\mathrm{Var}\,(\widehat{\beta}) = (X'X)^{-1}X'(\mathrm{Var}(Y))X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

F4 : Because projection is linear $P_{[X]}Y = X\beta + P_{[X]}\varepsilon$. Ainsi,

$$\mathrm{RSS}(\widehat{\beta}) = \| Y - P_{[X]}Y \|^2 = \| \varepsilon - P_{[X]}\varepsilon \|^2 .$$

Let $[X]^{\perp}$ be the orthogonal of $[X]$ (the image of X)

$$\varepsilon - P_{[X]}\varepsilon = P_{[X]^{\perp}}\varepsilon.$$

The dimension of $[X]^{\perp}$ is $n-k$. Because of classical results on isotropic Gaussian variables or beacause of the Cochran theorem (see Appendix) the random variable $RSS = RSS(\widehat{\beta})$ has a $\sigma^2\chi^2(n-k)$ distribution.

Note that a $\chi^2(n-k)$ variable can be represented as the sum of $(n-k)$ squares of standard Gaussian variables so it has expectation $(n-k)$. As a consequence

$$\mathbb{E}\,(\widehat{\sigma^2}) = \sigma^2.$$

The same sufficiency arguments as above imply optimality .

Last but not least

$$\widehat{\beta} \text{ depends on the projection of the data on } [X]$$

and

$$\widehat{\sigma^2} \text{ depends on the projection of the data on } [X]^{\perp},$$

so they are independent. ∎

Remark : If we don't assume normality **FH4** and without any other hypothesis, the least squares estimator $\widehat{\beta}$ remains optimal among the **linear** unbiased estimators. This is Gauss-Markov theorem (see Exercises).

3.2 A worked example : explicit equation in case of simple linear regression .

We consider the classical particular case of Model (1.5) corresponding to simple linear regression . We will compute the expression of estimator and their variance using matrix calculations. The unique clever argument is a change of parameterization. We start from the model

$$Y_i = \beta_1 + \beta_2 Z_i + \varepsilon_i,$$

and we define $\overline{Z}$ as the mean of the regressor: $\overline{Z} = Z_1 + \cdots + Z_n$ and we can rewrite the model as

$$Y_i = \beta_1 + \beta_2\overline{Z} + \beta_2(Z_i - \overline{Z}) + \varepsilon_i = \widetilde{\beta_1} + \beta_2\widetilde{Z_i} + \varepsilon_i,$$

Where $\widetilde{\beta_1} := \beta_1 + \beta_2\overline{Z}$ and $\widetilde{Z_i} := Z_i - \overline{Z}$. $\widetilde{Z}$ is a centered variable. In fact doing that, we have made the model orthogonal and this will be presented in details in Chapter 5. This simplify drastically the computation, at the end we can return to the original parameterization.

In matrix form we have **and forgetting the tilde for notational ease**

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

We set

$$X = \begin{pmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_n \end{pmatrix}$$

and, of course,

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

We can use

$$X'X = \begin{pmatrix} n & \sum Z_i \\ \sum Z_i & \sum (Z_i)^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \sum (Z_i)^2 \end{pmatrix} \text{ yielding } (X'X)^{-1} = \begin{pmatrix} n^{-1} & 0 \\ 0 & \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}.$$

In addition,

$$X'Y = \begin{pmatrix} 1 & \dots & 1 \\ Z_1 & \dots & Z_n \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i Z_i \end{pmatrix}$$

thus

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} n^{-1} & 0 \\ 0 & \left(\sum (Z_i)^2\right)^{-1} \end{pmatrix} \begin{pmatrix} \sum Y_i \\ \sum Y_i Z_i \end{pmatrix} = \begin{pmatrix} n^{-1} \sum Y_i \\ \left(\sum Y_i Z_i\right) \left(\sum (Z_i)^2\right)^{-1} \end{pmatrix}.$$

We can compute the variance-covariance matrix

$$\mathrm{Var}\,(\beta) = \sigma^2 (X'X)^{-1} = \begin{pmatrix} \sigma^2 n^{-1} & 0 \\ 0 & \sigma^2 \left(\sum (Z_i)^2\right)^{-1} \end{pmatrix}.$$

Now we reintroduce the tildes with this notation we have

$$\begin{aligned} \widehat{\widetilde{\beta}}_1 &= \bar{Y} \\ \widehat{\beta}_2 = \widehat{\widetilde{\beta}}_2 &= \frac{\sum_{i=1}^n Y_i (Z_i - \bar{Z})}{\sum_{i=1}^n (Z_i - \bar{(Z)})^2} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_{i=1}^n (Z_i - \bar{Z})^2} \\ \widehat{\beta}_1 = \widehat{\widetilde{\beta}}_1 &- \widehat{\beta}_2 \bar{Z}. \end{aligned}$$

Concerning variances and covariance

$$\begin{aligned} \mathrm{Var}\,(\widehat{\beta}_2) &= \frac{\sigma^2}{\sum_{i=1}^n (Z_i - \bar{Z})^2} \\ \mathrm{Cov}\,(\widehat{\beta}_1, \widehat{\beta}_2) &= -\bar{Z}\mathrm{Var}\,(\widehat{\beta}_2) = -\frac{\sigma^2 \bar{Z}}{\sum_{i=1}^n (Z_i - \bar{Z})^2} \\ \mathrm{Var}\,(\widehat{\beta}_1) &= \sigma^2 \left(1/n + \frac{(\bar{Z})^2}{\sum_{i=1}^n (Z_i - \bar{Z})^2} \right) = \frac{\sigma^2}{n \sum_{i=1}^n (Z_i - \bar{Z})^2} \left(\sum_{i=1}^n (Z_i - \bar{Z})^2 + n(\bar{Z})^2 \right) \\ &= \frac{\sigma^2}{n \sum_{i=1}^n (Z_i - \bar{Z})^2} \left(\sum_{i=1}^n Z_i^2 \right) \end{aligned}$$

So we found out all classical formulas about simple linear regression.

4 Fundamental tests and confidence intervals

4.1 Student's test for a linear combination

Let us consider for example the slope β_2 in the simple linear regression model. A natural question is whether $\beta_2 = 0$ and this can be generalized, in the general linear model (1.5), into the question $C'\beta = 0$? where C is the coefficient of a linear combination. Other examples are $\mu_1 - \mu_2 = 0$ or $2\mu_1 - \mu_2 - \mu_3 = 0$ in the forest example.

More precisely we want to test

$$\text{against} \quad \begin{array}{lcl} H_0 & : & C'\beta = 0 \\ H_1 & : & C'\beta \neq 0 \end{array}.$$

Proposition 1.1 *In the case of the general linear model and under the null hypothesis H_0 $(C'\beta = 0)$, Then*

$$\widehat{T} = \frac{C'\widehat{\beta}}{\sqrt{\widehat{\sigma}^2 C'(X'X)^{-1}C}}$$

follows a Student's distribution with parameter $(n-k)$.

Proof : Using **F2** and **F3**,

$$\operatorname{Var}(C'\widehat{\beta}) = \sigma^2 C'(X'X)^{-1}C.$$

A natural estimator of $\operatorname{Var}(C'\widehat{\beta})$ is $\widehat{\sigma^2 C'(X'X)^{-1}C}$. We normalize $C'\widehat{\beta}$ by its estimated standard error which is a classical demarche. In some sense we can say that we have "Studentized" the variable $C'\widehat{\beta}$. We obtain

$$\widehat{T} = \frac{C'\widehat{\beta}}{\sqrt{\sigma^2 C'(X'X)^{-1}C}} \times \frac{\sqrt{\sigma^2}}{\sqrt{\widehat{\sigma}^2}}.$$

Under H_0, since $C'\beta = 0$, $C'\widehat{\beta}$ is a centered Gaussian variable so after making is variance equal to 1

$$\frac{C'\widehat{\beta}}{\sqrt{\sigma^2 C'(X'X)^{-1}C}} \sim \mathcal{N}(0,1).$$

Moreover , $\frac{n-k}{\sigma^2}\widehat{\sigma}^2 = \frac{n-k}{\sigma^2}\|P_{[X]^\perp}\varepsilon\|^2$follows a $\chi^2(n-k)$ distribution and is independent of $C'P_{[X]}\varepsilon$ since the two spaces are orthogonal.

This corresponds strictly to the definition a Student distribution with $(n-k)$ degrees of freedom ∎

This proposition permits to contruct the test of H_0 against H_1 by choosing as rejection (of H_0) region $|\widehat{T}| > T_{n-k,1-\alpha/2}$ where $T_{n-k,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student $T(n-k)$ distribution. The properties of symmetry of the T distribution imply that, under H_0, the rejection probability is α ensuring that the level is at the nominal value.

Under H_1 some calculations proves that the probability of rejection (which is now the good decision) is always greater than α. Furthermore it is close to 1 if $C'\beta$ is far from zero.

Remark: $C'\beta = 0$ defines a sub-model of the general linear model (1.5), in that case a general Fisher test exists as described in the next session. Some calculation show that these two tests are the same.

4.2 Fisher test of a sub-model

The Fisher test is a generalization of the Student test when the co-dimension of H_0 is larger than one. In other words H_0 consists of assuming the nullity of more than one parameter. An obvious example is given by the equality of means in one-way analysis of variance. The general model H_1 is

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, ..., I, j = 1, ...n_{ij} > 0.$$

H_0 is the sub-model corresponding to the equality of means

$$Y_{ij} = \mu + \varepsilon_{ij}.$$

The framework

We consider general model (1.5) with X being a rank k matrix $k < n$. Here we don't need X to be full-ranked. We define
RSS as the residual sum of squares of this model.
$X\widehat{\beta} =\| Y - X\widehat{\beta} \|^2$ is the estimated response.

We consider the sub-linear model

$$Y = X^{(0)}\beta^{(0)} + \varepsilon, \tag{1.9}$$

Since it is a sub-model $[X^{(0)}]$ is strictly included in $[X]$ and $dim[X^{(0)}] = k_0 < k = dim[X]$. In this model $X^{(0)}\widehat{\beta}^{(0)}$ is the estimated response and

$$RSS_0 =\| Y - X^{(0)}\widehat{\beta}^{(0)} \|^2 .$$

is the residual sum of square. To give a general presentation of the two models we set

$$Y = R + \varepsilon$$

and the test problem can be written as

$$\text{versus} \begin{array}{lcl} H_0 & : & R \in [X^{(0)}] \\ H_1 & : & R \in [X] \setminus [X^{(0)}] \end{array} .$$

Proposition 1.2 *With the notation above, we define the test statistics of H_0 against H_1 by*

$$\widehat{F} = \frac{(\ RSS_0 - \ RSS)/(k - k_0)}{RSS/(n-k)}.$$

Then,under H_0, the statistics $\widehat{F}$ follows a Fisher distribution with parameters $(k - k_0, n - k)$. Under H_1 it takes larger values.

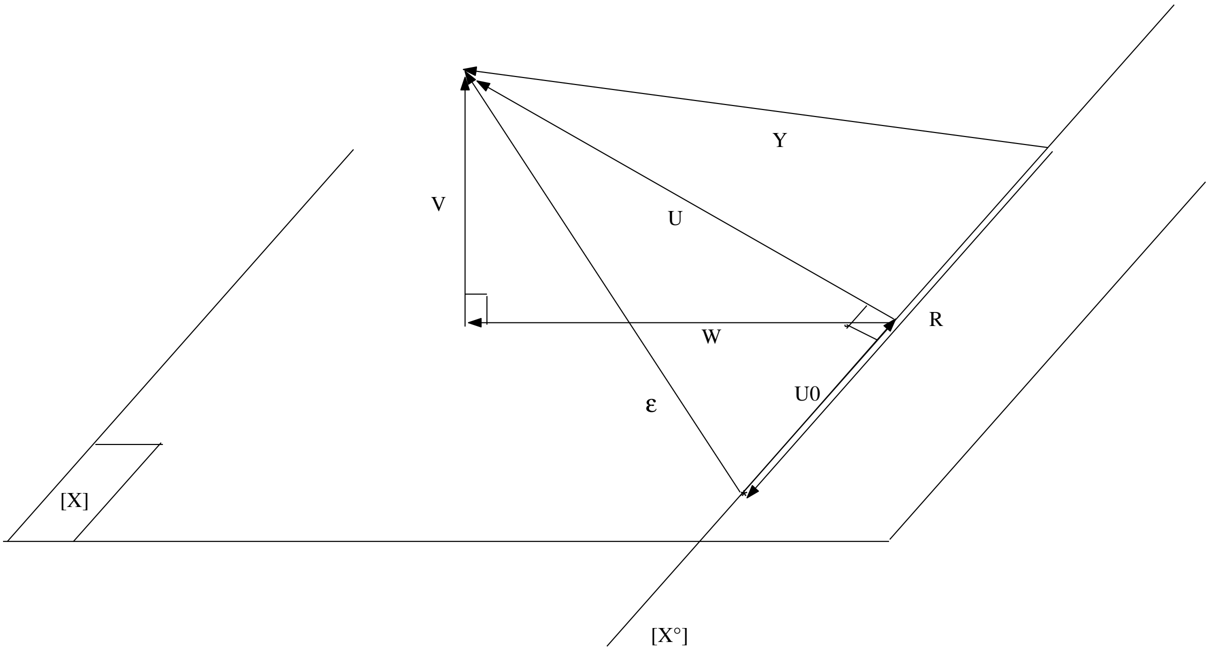


Figure 1.2: A graphical presentation of the Fisher test

The consequence of this proposition is that the rejection region (of H_0) will be defined by:

$$\widehat{F} > F_{(k-k_0,n-k,1-\alpha)},$$

where $F_{(k-k_0,n-k,1-\alpha)}$ is the $1-\alpha$ fractile of the Fisher distribution. This ensure, as for the Student test, that the level is effectively α.

Proof : The geometrical proof refers to Figure 4.2. Under H_0

$$\text{RSS} = \|Y - P_{[X]}Y\|^2 = \|P_{[X]^\perp}Y\|^2 = \|P_{[X]^\perp}\varepsilon\|^2 = \|V\|^2,$$

Where $V = P_{[X]^\perp}\varepsilon$ is indicated in Figure 4.2. We have used the fact that because $R \in [X]$, $P_{[X]^\perp}(R) = 0$.

Similarly,

$$\text{RSS}_0 = \|Y - P_{[X^{(0)}]}Y\|^2 = \|P_{[X^{(0)}]^\perp}Y\|^2 = \|P_{[(X^{(0)})]^\perp}\varepsilon\|^2 = \|U\|^2,$$

with $U := P_{[(X^{(0)})]^\perp}\varepsilon$. Let now A be the orthogonal of $[X^{(0)}]$ in $[X]$:

$$A \overset{\perp}{\oplus} [X^0] = [X].$$

Let $W = P_A\varepsilon$ (see Figure 4.2). By the Pythagore theorem

$$\|U\|^2 = \|V\|^2 + \|W\|^2$$

or

$$\|P_{[X^{(0)}]^\perp}\varepsilon\|^2 = \|P_{[X]^\perp}\varepsilon\|^2 + \|P_A\varepsilon\|^2.$$

Since ε is a isotropic Gaussian vector (zero expectation with a variance which is multiple of the identity). The projections onto two orthogonal spaces are independent with $\sigma^2\chi^2$ distribution. More precisely

$$\begin{aligned}\text{RSS} &= \|P_{[X]^\perp}\varepsilon\|^2 \text{ has a distribution } \sigma^2\chi^2(n-k),\\ \text{RSS}_0 - \text{RSS} &= \|P_{[X^{(0)}]^\perp}\varepsilon\|^2 - \|P_{[X]^\perp}\varepsilon\|^2 = \|P_A\varepsilon\|^2 \text{ has a distribution } \sigma^2\chi^2(k-k_0),\end{aligned}$$

and they are independent. Thus

$$\widehat{F} = \frac{\|P_A\varepsilon\|^2/(k-k_0)}{\|P_{[X]^\perp}\varepsilon\|^2/(n-k)}$$

corresponds strictly to the definition of the Fisher distribution with parameters $(k-k_0, n-k)$: $F_{(k-k_0,n-k)}$. ∎

4.3 Fisher test of the joint nullity of several linear combinations

Suppose that in a medical experiment we want to study the 5 level of a factor "treatment with 5 levels with a one-way analysis of variance model

$$Y_{ij} = \beta_i + \varepsilon_{ij} \quad i = 1, ..., 5$$

and that the relevant null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_3$ and $\beta_4 = \beta_5$, that can be written $C'\beta = 0$ with

$$C' = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

In the general case where C is of dimension p, k by some calculations it can be proved that

$$\widehat{F} = \frac{\widehat{\beta}'C(C'(X'X)^{-1}C)^{-1}C'\widehat{\beta}}{p\widehat{\sigma}^2}.$$

follows, under H_0 $C'\beta = 0$, a distribution $F(p, n-k)$.

4.4 Confidence intervals; confidence region

Let us begin with the simplest case on a linear combination $C'\beta$. We can extend the results of Section 4.1 to the test of

$$\text{versus} \quad \begin{aligned} H_0 &: C'\beta = c_0 \\ H_1 &: C'\beta \neq c_0 \end{aligned},$$

where c_0 is some value that doesn't need to be null. The relevant test statistics is now

$$\widehat{T} = \frac{C'\widehat{\beta} - c_0}{\sqrt{\widehat{\sigma}^2 C'(X'X)^{-1}C}},$$

the rest being identical.

In statistics, there is an equivalence between confidence intervals (or regions) and family of tests. If we have a family of tests of level α of the hypotheses $C'\beta = c_0$, the set of c_0 that are accepted gives a confidence region (which is here an interval) which is of confidence $1-\alpha$. This is direct. In ou case it yields

$$CI = \left[C'\widehat{\beta} - T_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma}^2 C'(X'X)^{-1}C}\, , \, C'\widehat{\beta} + T_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma}^2 C'(X'X)^{-1}C}\right].$$

In exactly the same manner, the resuts of Section 4.3 can be applied to the case $C'\beta$ of dimension $p > 1$. If

c_0 is some value $\mathbb{R}^p$, The statistics of the Fisher test of

$$\text{versus} \begin{array}{lcl} H_0 & : & C'\beta = c_0 \\ H_1 & : & C'\beta \neq c_0 \end{array}.$$

is

$$\widehat{F} = \frac{(\widehat{\beta}'C - c_0')(C'(X'X)^{-1}C)^{-1}(C'\widehat{\beta} - c_0)}{p\widehat{\sigma}^2}$$

that still follows, under H_0 a distribution $F(p, n-k)$. The set of the c_0 accepted at a level α is now the ellipsoid CR defined by

$$CR = \{c \in \mathbb{R}^p : (\widehat{\beta}'C - c_0')(C'(X'X)^{-1}C)^{-1}(C'\widehat{\beta} - c_0) \leq p\widehat{\sigma}^2 F_{(p,n-k),1-\alpha}\}.$$

Remark 1: The Scheffé method that will be presented in Chapter 3, Section 3, is based on the projections of this ellipsoid.

Remark 2 : In a linear model, a classical tool to measure the adequation of a model is the determination coefficient or R-square defined by

$$R^2 = \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2} = \frac{\|\widehat{Y} - \overline{Y}\|^2}{\|Y - \overline{Y}\|^2}.$$

This R-square must be carefully interpreted. Firstly the larger the model, the larger the R-square. Thus the R-square prefers always the largest model and is not a criterion of choice of models, see Chapter 7. Secondly, depending on the randomness of the phenomenon to explain, a R-square of 0.8, for example, can be either very good or very bad. So comparison of R-squares must be conducted between models of approximatively the same size and on the same kind of data.

5 About the fundamental hypotheses.

The hypothesis of Gaussianity is difficult to check in practice. Classical Gaussianity tests (Kolmogorov-Smirnov, Cramer-Von Mises, Anderson-Darling or Shapiro-Wilks) demands the observation of the ε_i that are non-observables. When applied to their estimation : the residuals $\widehat{\varepsilon_i} = (Y - \widehat{Y})_i$, they loose their properties. As already said a visual method like QQ-plots permit to detect huge departure from Gaussianity. But as explained in Chapter 6 and as anyone can check by a small simulation, except for very small data sets and for very non Gaussian data, most of the properties remains approximatively true without **FH4**. We say that the linear model is **robust** to non-Gaussianity.

Concerning all the fundamental hypotheses here is a list without details or proofs of the properties that are conserved.

Properties of the least squares estimator $\widehat{\beta}$

We consider

$$\widehat{\beta} = (X'X)^{-1}X'Y.$$

- $\widehat{\beta}$ is unbiased as soon as **FH1** is true : $\mathbb{E}\,(\widehat{\beta}) = \beta$,
- The variance-covariance matrix of $\widehat{\beta}$ remains equal to $\sigma^2(X'X)^{-1}$ under **FH2** et **FH3**, But this has little interest unless **FH1**is true.
- Under **FH1-FH3** $\widehat{\beta}$ is not optimal among the unbiased estimators but among the linear unbiased estimators only.
- Under **FH3-FH4** $\widehat{\beta}$ is Gaussian. But it converges to a Gaussian distribution under a very large set of hypotheses,see, for example, Chapter 6.

Properties of the estimator $\widehat{\sigma}^2$

Of course we need **FH2** in order σ^2 to be defined. Then

- Without Gaussianity, under **FH1-3**, $\widehat{\sigma^2}$ remains unbiased (see exercise 6).
- Under the same framework $\widehat{\sigma^2}$ converges to σ^2 for large data but the speed of convergence depend on the Kurtosis of the distribution of errors , see Chapter 6.

Properties of the tests F and T

In this section we assume **FH1-3** and not **FH4** . Without entering into the details , It is proved in Chapter 6 that the properties of the T and F test remains true for large data.

Correlated errors

Some correlation can be assumed beween the errors, for example that they form an ARMA process. This is the ARMAX model: Autoregressive?moving-average model with exogenous inputs model. We refer to the literature ,Amemiya [2], Green [8], Guyon [9] ou Jobson [10].

6 Exercises

Exercise 1.1

(*) Let Y obey to a X-linear model and let $T \in \mathbb{R}^n$ be a deterministic vector. Prove that

$$\mathbb{E}\left(\|T - Y\|^2\right) = n\sigma^2 + \|T - X\beta\|^2.$$

Exercise 1.2

(**) [Gauss-Markov Theorem] We assume a X-linear model without **FH4** and we prove the optimality of $\widehat{\beta}$ among linear unbiased estimators.
This means that is $\widetilde{\beta}$ is another unbiased estimator

$$\mathrm{Var}\,(\widetilde{\beta}) - \mathrm{Var}\,(\widehat{\beta}) \text{ is a semi-definite positive matrix },$$

or equivalently for every linear combination $C'\beta$ of the parameters

$$\mathrm{Var}\,(C'\widetilde{\beta}) \geq \mathrm{Var}\,(C'\widehat{\beta}).$$

i. Set $\widetilde{\beta} = MY$ where M est une matrice de taille (k, n). Show that $MX = I_n$.

ii. Write $\widehat{\beta} = TP_{[X]}Y$, and show that $MP_{[X]} = TP_{[X]}$.

iii. Show that $\widetilde{\beta} = \widehat{\beta} + MP_{[X]^\perp}Y$, The sum being orthogonal. Conclude.

Exercise 1.3

(**)Let β_1 et β_2 two real unknown parameters and let :

- Y_1 an unbiased estimator of $\beta_1 + \beta_2$ with variance σ^2;
- Y_2 an unbiased estimator of $2\beta_1 - \beta_2$ with variance $4\sigma^2$;

- Y_3 an unbiased estimator of $6\beta_1 + 3\beta_2$ with variance $9\sigma^2$,

These estimators are assumed to be independent. What estimator of β_1 and β_2 could you propose ? You can use the preceding exercise.

Exercise 1.4

(**) [Estimation of the variance] We consider the non-Gaussain X-linear model (without **FH4**) and we want to compute $\mathbb{E}\,\widehat{\sigma^2}$.

i. Show that $(n-k)\widehat{\sigma^2} = \mathrm{Tr}(\varepsilon' P_{[X]^\perp}\varepsilon)$ where Tr is the trace.

ii. Using the well-known identity $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, show that $(n-k)\mathbb{E}\,(\widehat{\sigma^2}) = \sigma^2 Tr(P_{[X]^\perp}\varepsilon'\varepsilon)$.

iii. Conclude .

Chapter 2

Regression

In this chapter, after some presentation of linear and non-linear regression models, we will try to answer to twoquestions :

- *What kind of phenomenon can be modeled by regression ?*
- *How to improve a regression model to get a better fit to the data ?*

1 Linear and non-linear models

We have seen that multiple regression is a linear model. Polynomial regression is a particular case where the explanatory variables, the regressors, are linked by a nonlinear formula. Periodic regression on sine and cosine functions is an other example. More precisely if Y is a periodic function of t with period 2π:

$$Y_i = \mu + \alpha_1 \cos(t_i) + \beta_1 \sin(t_i) + \alpha_2 \cos(2t_i) + \beta_2 \sin(2t_i) + ... + \varepsilon_i$$

is the general form of the periodic regression model which is again a linear regression model.

A classical **non linear model** can be encountered in pharmacokinetics. If you consider the plasmatic concentration of a drug after a bolus injection or an oral administration, general considerations based on differential equations lead to a compartment model. More precisely we can set

$$Y_i = \beta_1 \exp(-\alpha_1 t_i) + \beta_2 \exp(-\alpha_2 t_i) + \varepsilon_i \quad \text{for} \quad i = 1, \cdots, n.$$

The unknown parameters are $\alpha_1, \alpha_2, \beta_1, \beta_2$ and the dependence in α_1, α_2 is definitively non-linear because on the non linearity of the exponential function.

An other classical example is given by the logistic regression. More precisely

$$Y_i = \frac{\beta_1 + \beta_2 \exp(\beta_3 x_i)}{1 + \beta_4 \exp(\beta_3 x_i)} + \varepsilon_i \quad \text{for} \quad i = 1, \cdots, n.$$

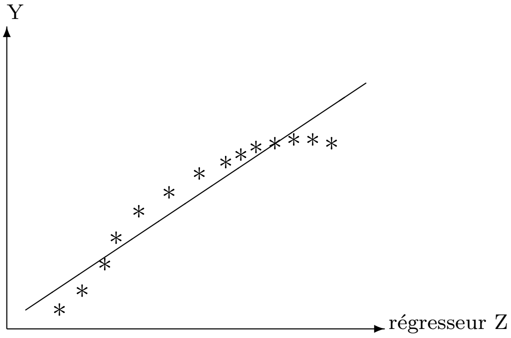
The unknown parameters are $\beta_1, \beta_2, \beta_3, \beta_4$ and the relations again non-linear .

In such a case, the least squares method is also used. But the solution is no longer explicit: in fact the non linear model can be linearized by a Taylor formula around a prior, defining a tangent linear model, the estimator of which are easy to obtain. In second step this estimate is used as a prior and the process is iterated. Finally the complexity is much higher.

2 Graphical control

Once a regression model has used, it is mandatory to check graphically the validity of the fondamental hypotheses.

• In simple linear regression a scatter plot of Z, Y with the regression line gives an almost exhaustive information. For example



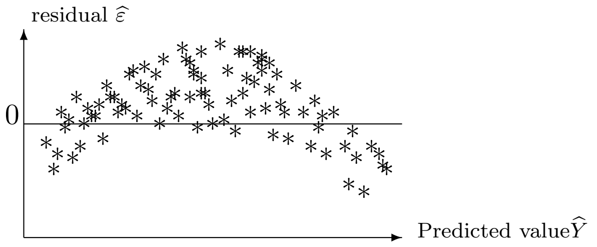
On this plot we see a curvature of the cloud of points and there is a strong evidence that **FH1** is no longer true.

• In case of multiple regression, this kind of graphics is not possible. Remember that the fundamental hypotheses concerns the errors ε_i that are unobservables, so we must use their estimations

$$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i.$$

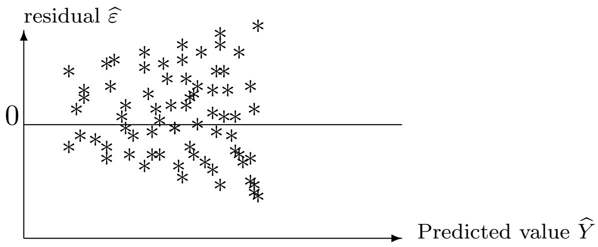
1/To check FH1 et FH2: adequacy of the model and homoscedasticity we use the classical plot of residuals against fitted value $(\widehat{Y}_i)_i$. This graphics must be almost systematically done.

Rougly speaking two main pathological patterns can be detected. The first one is "banana shape " as in the following example



In such an example, it can be considered that the adequacy hypothesis **FH1** is not satisfied. In other words the regression formula must be enriched by proposing other regressors.

The other typical pathological pattern is the "trumpet shape" as in the following example:



In this example there is a strong evidence that the variance is not homogenous. A possible solution is to use a transformation of variable.

Remark: Some authors and some software use Studentized residuals. These last variables are the residuals $\widehat{\varepsilon}_i$ divided by the estimation of their standard error. They follow a Student distribution, i.e. almost a normal distribution. Ordinary residual are in the same unit as the observation. For example if we observe a residual of 0.5 cm on the height of man, we know that the fit is very good. On the other hand, Studentized residuals are a-dimensional. We know that a Studentized residual of 5.1 is very large but we have no practical interpretation of this 5.1.

Can we transform the model ?

• We can freely transform the regressors (dependent variables) $Z^{(1)}, \cdots, Z^{(p)}$ using every possible algebraic transformation : power, square root, exponential circular functions, etc... as soon as the resulting regression formula can be interpreted. This has to be done in case of residual plot of the first kind ("banana"). In second step a selection of model procedure as in Chapter 7 can be performed to remove un-necessary variables.

• On the other hand, the response Y can be transformed only in the case the graphics of residual show some evidence of heteroscedasticity. The linear model assumes that the

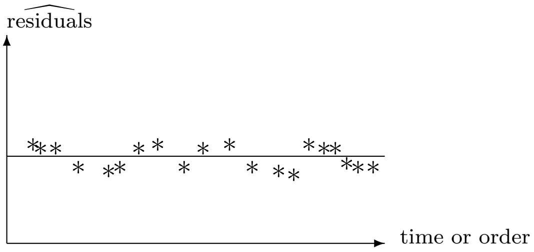
Kind of relation	Domain for Y	Transformation
$\sigma = (\text{const})Y^k,\ k \neq 1$	$\mathbb{R}_+^*$	$Y \mapsto Y^{1-k}$
$\sigma = (\text{const})\sqrt{Y}$	$\mathbb{R}_+^*$	$Y \mapsto \sqrt{Y}$
$\sigma = (\text{const})Y$	$\mathbb{R}_+^*$	$Y \mapsto \log Y$
$\sigma = (\text{const})Y^2$	$\mathbb{R}_+^*$	$Y \mapsto Y^{-1}$
$\sigma = (\text{const})\sqrt{Y(1-Y)}$	$[0,1]$	$Y \mapsto \arcsin\left(\sqrt{Y}\right)$
$\sigma = (\text{const})\sqrt{1-Y}.Y^{-1}$	$[0,1]$	$Y \mapsto (1-Y)^{1/2} - 1/3(1-Y)^{3/2}$
$\sigma = (\text{const})(1-Y^2)^{-2}$	$[-1,1]$	$Y \mapsto \log(1+Y) - \log(1-Y)$

Table 2.1: Table of the changes of variable for the response Y

absolute error is constant, i.e. independent of the amplitude of the response. In many case the error is proportional to the response: the larger the response the larger the error. In such a case, a logarithmic transform of the response will fix the problem. A list of the transformations to be used is given in Table 2.1 depending on the relation between the mean response and the standard error. An alternative which is more rigorous but much more complex is to use a generalized liner model with a all chosen link function see for example McCullagh et Nelder [14].

Note that these transformations are based on Talylor expansion and are valid for rather large data in the other cases the use of a generalized linear model is necessary.

2/ To check independence FP 3, The relevant graph consist of a scatterplot of residuals against the time that can, in most of the times, be found as the order in the file. An example is given by the following graph:



On the graph we can see some "runs" of residual with the same sign. This can be checked by a special test "run test". See Exercise 5

In case of evidence of correlation between residuals and thus between errors, a classical approach is to use an ARMA model. The resulting model of regression with ARMA errors is called ARMAX. See Amemiya [2], Green [8], Guyon [9] ou Jobson [10]).

3/ To check normality FH4

The QQ plot is a scatterplot with in abscissa the order statistics of the residuals (the residual after sorting) $\widehat{\varepsilon}_{(i)}$ where the (i) means the sorting and in ordinate the quantile i/n of the standard normal distribution.

More precisely the two coordinates are

- $\widehat{\varepsilon}_{(i)}$ is the fractile i/n of the empirical distribution. -$Z_{i/n}$ is is the fractile of the $N(0,1)$

and if the residuals are approximatively of distribution $N(0,\sigma^2)$ we know that $\widehat{\varepsilon}_{(i)}$ is closed to $\sigma Z_{i/n}$. So the scatter plot is closed to a line.

3 Random regressors

The presentation we made assumes that the regressors are deterministic variables and that they are perfectly known. This is rarely the case. Or presentation can be generalized using conditional models.

Suppose now that the regressors are random $Z^{(1)}, Z^{(2)}, \cdots, Z^{(p)}$, exactly observed and independent of the errors ε. In that (good) case we can put us conditionally to the n observations of $Z^{(1)}, Z^{(2)}, \cdots, Z^{(p)}$. The distribution of the errors doesn't change because of independence so the conditional model is indeed a linear model in the sense of our definition and the generalization is free.

Alternatively, suppose that the regressor are random because they are observed with error. Consider for example the yield of a chemical reaction that depends on the temperature of the substratum. The following temperature has been design for the experiment : 150:10: 210 but that in fact the actual temperature differs from the nominal in an unknown manner. In that case we obtain an error in variable model whose solution is not given by least square method. The method is called "total least squares", based on principal component analysis, is difficult to implement see Exercise **??**. In practice, as soon as the errors are small, regression is still used though not perfectly rigorous.

4 Choosing among the regressors

This topic will be considered in detail in Chapter 7. Basically if the number of regressor is large, it is very likely to this that some of them are superfluous. In the whole model, classical outputs of softwares give a T test of significance of each variable.

This test, for variable $Z^{(i)}$, means "if I keep all others variables, can I remove variable i"

Suppose that they are 10 regressors and that the T tests of the variables 1,5,8 are non significative. Can we remove all three variables ? No because each test is performed keeping all other variables ! A very crude but recommended method is **backward selection**:

We start with the whole model

At each step the least significative variable is identified.

- if it is not significative (at a given level) the variable is removed and we pass to the next step
- if it is significative, the algorithm stops, the last model is the chosen one .

This algorithm has a variant which is forward regression which is just the contrary : starting from the void model and adding stepwise the most significant variable until the variable added is non-significative.

A third variant "stepwise" mixes forward and backward steps.

Note that the Rsquare defined by :

$$R^2 = \frac{\|\widehat{Y} - \overline{Y}\|^2}{\|Y - \overline{Y}\|^2}$$

always prefers the whole model.

4.1 Measure of colinearity

The colinearity between regressors is an important issue. It implies correlation between the estimates of the coefficients β_i and also an inflation of variance. This last inflation with respect to an ideal situation where the model is orthogonal is measured by the Variance inflation factor VIF which is defined as follows

$$VIF_i = \frac{1}{1 - R_i^2}$$

whereR_i^2 is defined as the Rsquare of the regression of $Z^{(i)}$on all the others regressor.

By definition this coefficient is always larger than 1. It take the value 1 in case of orthogonality. A value larger than 10 is generally considered as an indication of large colinearity.

Some authors define the Tolerance TOL defined as 1/ VIF.

5 Exercises

(English version coming soon)

Exercise 2.1

(Transformation de variables) Soit des observations Y_{ij} qui suivent le modèle suivant :

$$Y_{ij} = \mu_i + \varepsilon_{ij} \cdot \sqrt{\mu_i(1-\mu_i)}, \quad \text{pour} \ \ i = 1, \cdots, I, \ j = 1, \cdots, J, \tag{2.1}$$

où les erreurs ε_{ij} vérifient les postulats habituels **P1-4**. Ce modèle correspond au 5ème cas du tableau 2.1. On pose

$$Z_{ij} = \arcsin(\sqrt{Y_{ij}}).$$

i. Écrire un développement limité à l'ordre 1 de la fonction $x \mapsto \arcsin(\sqrt{x})$ au point x_0.

ii. On admet que l'on peut négliger le reste : soit c'est une hypothèse que l'on assume, soit on suppose que σ^2 tend vers zéro. Dans ce dernier cas on utilise ce que l'on appelle la δ méthode ou Théorème de Slutsky (voir Van der Vaart [21] ou Dacunha-Castelle et Duflo [7] p. 91). Montrer que Z_{ij} suit un modèle d'analyse de la variance à un facteur.

iii. Montrer que si les Y_{ij} sont des données de comptage sur de grands effectifs, avec une probabilité de succès qui dépend de l'indice i on est approximativement dans la situation du modèle (2.1).

iv. Traiter de même tous les cas du tableau 2.1.

Exercise 2.2

(Test de runs) Ce test est utilisé pour tester la présence ou non de corrélations dans les ε_i. On commence d'abord par le décrire dans le cas où l'on observe des variables aléatoires $Y_1, \ldots, Y_n$ dont on veut tester l'indépendance. On les suppose de médiane zéro. On compte en fait le nombre de "paquets" ou "runs" R de même signe que $Y_1, \cdots, Y_n$.
Par exemple, si $Y_1, \ldots, Y_9 = (1.1, 1.3, -2, -1, 4.5, 1.6, -2.7, -1.3, 4)$, il y a 5 runs pour $n = 9$ données.

i. Montrer que si on suppose qu 'aucun des Y_i n'est nul, alors :

$$R = 1 + \sum_{i=1}^{n-1} \mathbb{I}_{\mathrm{Y_i Y_{i+1}<0}} := 1 + \sum_{\mathrm{i=1}}^{\mathrm{n-1}} \mathrm{Z_i}$$

ii. On suppose que les Y_i sont indépendantes et de loi diffuse (c'est-à-dire absolument continue par rapport à la mesure de Lebesgue). Montrer que $\mathbb{E}(R) = \dfrac{n+1}{2}$,

iii. Montrer que si $|i-j| > 1$, Z_i et Z_j sont indépendants. Montrer que Z_i et Z_{i+1} sont également indépendants. En déduire $\operatorname{Var}(R)$.

iv. En utilisant le théorème de la limite centrale, construire pour des grands échantillons une statistique libre qui suit une loi normale centrée réduite sous l'hypothèse H_0 d'indépendance et qui tend vers $\pm\infty$ sous les alternatives H_1 d'intrication et de répulsion. Nous laissons au lecteur le soin de deviner le sens de ces deux derniers mots.

Remarque : Pour ce qui est du test de l'indépendance des erreurs dans un modèle linéaire, on appliquera ce test aux estimateurs $\widehat{\varepsilon_i}$ en négligeant leurs liaisons (toujours présentes, même sous l'hypothèse d'indépendance des ε_i) et en négligeant le fait que leur médiane n'est qu'approximativement nulle. Il existe d'autres versions de ce test sous des hypothèses d'échangeabilité (voir par exemple Lecoutre et Tassi, [11]).

6 Software examples

We will use a data set from Tomassone *et al* [20] that studies Pine Processionary (Thaumetopoea pityocampa), one of the most destructive species to pines. We want to study the influence of several variables on the density, $X11$ (or $X12 = \log(X11)$) of the population.

A list of explanatory variables is used. More precisely we consider

- the altitude of the plot : $X1$;
- its step in degrees : $X2$;
- the number of pine of the plot : $X3$;
- the height of the tree at the center of the plot: $X4$;
- the diameter of this tree : $X5$;
- a note on the density of population of trees : $X6$;
- The orientation of the plot (from 1=south, 2=others) : $X7$;
- the mean heigh of main trees : $X8$;
- number of strata of vegetation : $X9$;
- A measure of mixture of populations (de 1=mixed , to 2=non mixed) : $X10$.

The observations are really quantitative even for $X7$ or $X9$ because a mean is taken over many sub-plots. This example will be considered in other chapters and in particular in chapter 7. In the present chapter we consider the regression of $X11$ or $X12$ on $X1$, $X2$, $X4$ et $X5$.

Sofware : R :

The commands are

```
library(car)
lm.proc=lm(X11~X1+X2+X4+X5,proc)
Anova(lm.proc,type="III")
par(mfrow=c(2,2))
plot(lm.proc,las=1)
```

<u>**Software : SAS**</u> :

We assume that the data `sasuser.proc` has been created :

```
proc reg data=sasuser.proc all;
model X11=X1 X2 X4 X5;
plot r.*p.;run;quit;
```

Here is an extract of the output (which is very long because of the command `all`) :

```
            Correlation

Variable     X1            X2           X4            X5           X11

X1         1.0000       0.0861        0.3211        0.2876       -0.5337
X2         0.0861       1.0000        0.1346        0.1175       -0.4647
X4         0.3211       0.1346        1.0000        0.9050       -0.3576
X5         0.2876       0.1175        0.9050        1.0000       -0.1578
X11       -0.5337      -0.4647       -0.3576       -0.1578        1.0000

        Analysis of Variance

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

Model                     4       13.10487        3.27622      11.51    <.0001
Error                    27        7.68670        0.28469
Corrected Total          31       20.79157

Root MSE              0.53357    R-Square     0.6303
Dependent Mean        0.81406    Adj R-Sq     0.5755
Coeff Var            65.54360

        Parameter Estimates
                            Parameter     Standard
Variable   Label      DF     Estimate        Error  t Value   Pr > |t|
Intercept  Intercept   1      6.60309      1.02423     6.45     <.0001
X1         X1          1     -0.00281   0.00078216    -3.60     0.0013
X2         X2          1     -0.04565      0.01346    -3.39     0.0022
X4         X4          1     -0.75510      0.21591    -3.50     0.0016
X5         X5          1      0.16847      0.05154     3.27     0.0029
```

```
        Parameter Estimates
                                                                   Variance
Variable   Label      DF      Tolerance      Inflation       95% Confidence Limits

Intercept  Intercept   1              .              0        4.50153        8.70464
X1         X1          1        0.89499        1.11733       -0.00442       -0.00121
X2         X2          1        0.97975        1.02067       -0.07326       -0.01803
X4         X4          1        0.17631        5.67193       -1.19810       -0.31209
X5         X5          1        0.18093        5.52696        0.06273        0.27422

        Output Statistics

Dep    Var  Predicted Std Error
Obs    X11 Value Mean   Predict   95% CL Mean     95% CL Predict     -2-1 0 1 2

1   2.3700   1.6964   0.1625   1.3631   2.0298   0.5520   2.8409   |      |**     |
2   1.4700   1.2602   0.1799   0.8912   1.6293   0.1049   2.4155   |      |       |
3   1.1300   1.3642   0.2090   0.9353   1.7931   0.1884   2.5400   |      |       |
4   0.8500   1.0823   0.1651   0.7436   1.4210  -0.0637   2.2283   |      |       |
5   0.2400   0.3341   0.1658  -0.0060   0.6743  -0.8123   1.4806   |      |       |
6   1.4900   1.0255   0.1084   0.8031   1.2479  -0.0916   2.1427   |      |*      |
7   0.3000   0.0136   0.2569  -0.5135   0.5407  -1.2015   1.2287   |      |*      |
8   0.0700  -0.1807   0.2677  -0.7299   0.3686  -1.4055   1.0442   |      |*      |
9   3.0000   1.8174   0.1836   1.4406   2.1942   0.6596   2.9752   |      |****   |
10  1.2100   0.8020   0.2604   0.2677   1.3364  -0.4162   2.0203   |      |*      |
:      :       :          :       :        :        :        :        :      :
```

Discussion : The global Fisher test is significative, which is a minimal property. The Q-Q plot is coherent with normality. But the two residual plots : raw and studentized show a strange behavior and in particular a triangle left and below with no observations. This is due to the constraint of positivity of the observation. This is why in other studies $X12 = \log(X11)$ will be used. A very fast interpretation of the signs of the coefficients is that the processionary is less present on plots with difficult access.

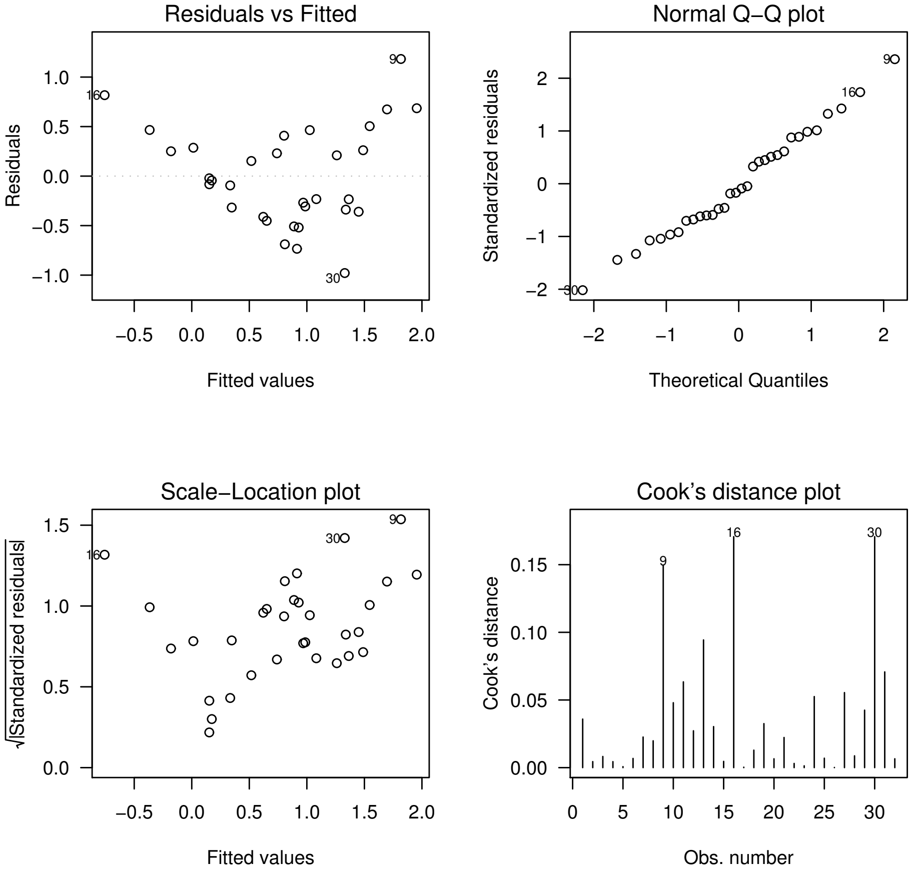


Figure 2.1: Graphics from the regression (R)

Chapter 3

Analysis of variance

This chapter presents the notion of interaction between to qualitative variables or factors; a generalization to several factors; a definition of nested and crossed situations; and multiple comparison problem.

1 The general framework

Analysis of variance (Abbreviated as ANOVA) consists of explaining a quantitative variable by several qualitative variables of factors. Let us consider two examples :

a. Varieties comparison : the dependent variable: the variable to explain is the yield. It can be explained by:

   - 2 factors : variety × location,
   - 3 factors : variety × location × year
   - 4 factors : genetic family × individual× location × year.

b. Annual income of an executive. The dependent variable is the annual income of an executive as a function of seven factors: domain of activity× age group × size of the firm × region × diploma× position × sex.

As we can see, a high number of factor may permit to better modell complex situations. We don't have to hesitate to put a large number of factors in the model. Limitations are : A) we must learn how to build models with many factors. This has to do with Aa) defining order-two and larger interaction, Ab) defined relations between factors : crossed factors, nested factors, factors included one into the other. B) the second limitation concerns the dimension of the model that grows exponentially with the number of factors and must remain smaller than the number n of observations. We will present now the more classical model : the model with two crossed factors. The reason why the factors are called " crossed" will be explained only in Section 3.2

2 Two crossed factors

2.1 Presentation

Let us consider the historical example of varieties comparisons. The first problem Ronald Fisher had to consider when he began his career at Rothamsted experimental station. To compare, for example I cereal varieties on a quantitative criterion as the yield per hectare, we have at our disposition J locations that are mainly J different regions of culture. In Location i Variety j is experimented $n_{i,j}$ times with $n_{i,j} > 0$. In most experiments this number of replication is designed to be constant $n_{i,j} = r > 1$ (balanced case) but, because of missing data, this number is eventually unbalanced .

Let Y_{ijk} be the observation on the kth observation of Variety i in Location j. In a first step we assume that the response depends on the couple (i, j) and we set the following **one-way** analysis of variance model.

$$Y_{ijk} = \beta_{ij} + \varepsilon_{ijk}, \quad \text{where} \tag{3.1}$$

- i is the variety index , $i = 1, ..., I$;
- j is the location index , $i = 1, ..., J$;
- k is the repetition index $k = 1, ..., n_{ij} > 0$.

We assume in addition that at least one combination (i, j) is observed at least two times: $n_{ij} \geq 2$. This ensures that the number of observations

$$n = \sum_{i,j} n_{ij},$$

is greater than the dimension of the model which is IJ as we will see.

Suppose for example that $I = 2$, $J = 3$ and $n_{ij} = 2$ for all i, j. The model can be written in matrix form as

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{21} \\ \beta_{22} \\ \beta_{23} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{131} \\ \varepsilon_{132} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{231} \\ \varepsilon_{232} \end{pmatrix}$$

For the moment, the model we have set is a a one-way (or one factor) analysis of variance model associated to the qualitative variable variety× location that takes IJ values. This proves that the dimension is IJ. As it is easy to check, the model is regular and since the least squares estimator of d reals $X_1, ..., X_d$ is their mean $\overline{X}$ we have

$$\widehat{\beta}_{ij} = Y_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \ .$$

Our goal is now to introduce the two original factors, but carefully avoiding to write non regular models. We define

- The general mean $\mu = \beta_{..}$, where the dots means the a mean over the indices replaced by the dots. it is estimated by $\widehat{\beta}_{..} = \frac{1}{IJ} \sum_{i=1,...,I;j=1,...,J} \widehat{\beta}_{ij}$;

- The differential effect of the modality i of the first factor. This is defined with respect to the preceding mean:

  $\alpha_i = \beta_{i.} - \beta_{..}$, estimated by par $\widehat{\beta}_{i.} - \widehat{\beta}_{..}$;

- The differential effect of the modality j of the second factor. This is defined with respect to the preceding mean facteur $\gamma_j = \beta_{.j} - \beta_{..}$, estimated by $\widehat{\beta}_{.j} - \widehat{\beta}_{..}$;

- The quantity we need to arrive to β_{ij} is called the *interaction*. As a matter of fact, in general we don't have $\beta_{ij} = \beta_{i.} + \beta_{.j} + \beta_{..}$ and the quantity missing is :

$$\delta_{ij} = \beta_{ij} - \beta_{i.} - \beta_{.j} + \beta_{..} = (\beta_{ij} - \beta_{..}) - (\beta_{i.} - \beta_{..}) - (\beta_{.j} - \beta_{..}).$$

The main message of this section is the necessity of this term to achieve the decomposition. Finally the initial model (3.1) : $Y_{ijk} = \beta_{ij} + \varepsilon_{ijk}$ can be rewritten in the form

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + \delta_{ij} + \varepsilon_{ijk}, \tag{3.2}$$

with $k \in \{1, \cdots, n_{ij}\}$ pour $(i, j) \in \{1, \cdots, I\} \times \{1, \cdots, J\}$, and the following implicitly defined constraints.

- $\sum_i^I \alpha_i = 0$ and$\sum_j^J \gamma_j = 0$;
- for all $j = 1, \cdots, J$: $\sum_i \delta_{ij} = 0$;
- for all $i = 1, \cdots, I$: $\sum_j \delta_{ij} = 0$.

But we insist on the fact that this model must not be regarded as a model on its own (that would be irregular) but as a rewriting of Model 3.1. In other words, when a calculation has to be performed it is in general simpler to make it with Model 3.1.

The matrix form of Model 3.2 in the example $I = 2$, $J = 3$ and $n_{ij} = 2$ for all i, j, is:

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \delta_{11} \\ \delta_{12} \\ \delta_{13} \\ \delta_{21} \\ \delta_{22} \\ \delta_{23} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{131} \\ \varepsilon_{132} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{231} \\ \varepsilon_{232} \end{pmatrix}.$$

Remind that the design matrix X is not full-ranked and that special techniques that will be presented in Chapter 5 are needed to study it directly.

Testing strategies

The model (3.2) is very different depending on whether the δ part is present or not. If not we introduce the following definition

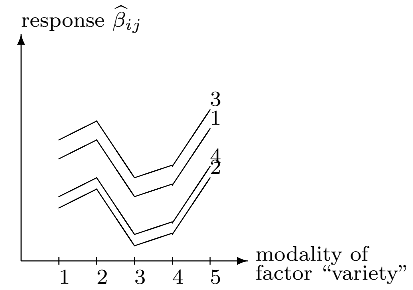
Definition 3.1 *When the interaction part (δ) is absent in Model (3.2), the model is called "additive". In the other case the model is called general or "interactive". Finally the quantities $\alpha_i, i = 1, ..., I$ and $\gamma_j, j = 1, ..., j$ define the main effect of the two factors.*

There is a huge difference between the additive model and the interactive model.

Firstly the size of the general model: IJ is much more important than the size of the additive model $I + J - 1$. For example if $I = 20,\ J = 10$ the respective sizes are 200 and 29.

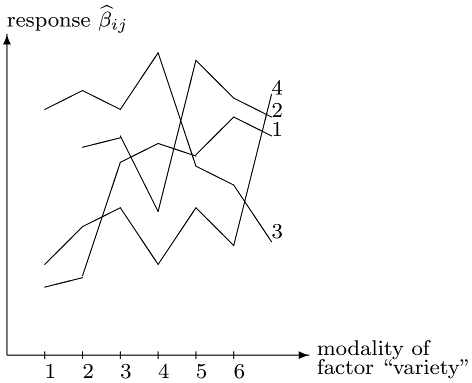
Secondly the behaviors are different. If we represent β_{ij} or its estimation $\widehat{\beta}_{ij}$ as a function of each of the factors, the typical behaviors are as in its example with $I = 6$ and $J = 4$.

Additive model



The curves are strictly parallels meaning that if we compare for example two varieties, their difference of yield are constant in every location.

General interactive model



There is no typical behavior in this case. The response is as general as possible.

2.2 General model in case in case of equi-replication

To simplify we assume that the number of replications is constant. In other words, $n_{ij} = (const) = K \geq 2$ The model is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}\ ,\ i = 1, \cdots, I\ ,\ j = 1, \cdots, J\ ,\ k = 1, \cdots, K,$$

with the above constraints. We define the hypotheses :

- $H_0^{(1)}$: "all the α_i's are zero ": test of the principal effect of the first factor ;

- $H_0^{(2)}$: "all the γ_j's are zero": test of the principal effect of the second factor;

- $H_0^{(3)}$: "all the γ_{ij}'sare zero": test of the interaction.

Note that all these hypotheses can be expressed in Model (3.1) and this is the good point of view. Let us introduce some extra notation.

- $E := \mathbb{R}^n$ is the space of the observations ($n = I \cdot J \cdot K$) equipped with the classical Euclidean norm. An element of E is denoted $(Y_{ijk})_{ijk}$.
- $E_0 = [\mathbb{1}] = \{(\mathrm{Y_{ijk}})_{\mathrm{ijk}} \in \mathrm{E}\ :\ \mathrm{Y_{ijk}} = (\mathrm{const}) = \mathrm{m}\}$ is the space of constants generated by the principal diagonal.
- $E_1 = \{(Y_{ijk})_{ijk} \in E : Y_{ijk} = a_i$ for some $a_i's$ such that $\sum_i a_i = 0$.
  E_1 consist of the vectors whose coordinates depend on i only and that are centered.
- $E_2 = \{(Y_{ijk})_{ijk} \in E : Y_{ijk} = b_j$for some $b_j's$ such that $\sum_j b_j = 0\}$.
  E_2 consist of the vectors whose coordinates depend on j only and that are centered.
- $E_3 = \{(Y_{ijk})_{ijk} \in E : Y_{ijk} = c_{ij}$ for some $c_{ij}'s$ such that $, \forall i,\ \sum_j c_{ij} = 0; \forall j,\ \sum_i c_{ij} = 0\}$.
  E_3 is the space of the interaction.

We have the following easy relations :

- E_0, E_1 E_2 et E_3 are orthogonal;

- $E_0 + E_1$ corresponds to the model with the first factor ;

- $E_0 + E_2$ corresponds to the model with the second factor;

- $E_0 + E_1 + E_2$ corresponds to the additive model ;

- $E_0 + E_1 + E_2 + E_3$ correspond to the whole model.

- $P_{E_0}(Y) = \big(Y_{...}\big)_{ijk}$,

- $P_{E_0+E_1}(Y) = P_{E_0} + P_{E_1}(Y) = \big(Y_{i..}\big)_{ijk}$,
- $P_{E_0+E_2}(Y) = P_{E_0} + P_{E_2}(Y) = \big(Y_{.j.}\big)_{ijk}$,
- $P_{E_0+E_1+E_2+E_3}(Y) = \big(Y_{ij.}\big)_{ijk}$,

By combination it is easy to deduce the expression of each projector. Let us consider, for example, the case of the interaction, i.e. the test of $H_0^{(3)}$. The sum of squares (SS) associated to this effect is defined as the difference of RSS between the models with and without interaction. It is a consequence of orthogonality that this quantity is $\|P_{E_3}Y\|^2$. The numerator of the Fisher test statistics is just the **mean square**: the SS divided by $(I-1)(J-1)$. As for the denominator, it is the estimator of the variance. Finally

$$\widehat{F} = \frac{\big(\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2\big)/(I-1)(J-1)}{\big(\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2\big)/(n - I.J)}.$$

Considering in details all the other cases we get the following analysis of variance table that gives the exact expression of every test.

Source	Sum of squares	Degrees of freedom	$\widehat{F}$
Factor 1	$\sum_{i,j,k}(Y_{i..} - Y_{...})^2$	$I-1$	$\frac{(n-I.J)}{(I-1)}\frac{\sum_{i,j,k}(Y_{i..} - Y_{...})^2}{\sum_{i,j,k}(Y_{i,j,k} - Y_{ij.})^2}$
Factor 2	$\sum_{i,j,k}(Y_{.j.} - Y_{...})^2$	$J-1$	$\frac{(n-I.J)}{(J-1)}\frac{\sum_{i,j,k}(Y_{.j.} - Y_{...})^2}{\sum_{i,j,k}(Y_{i,j,k} - Y_{ij.})^2}$
Fac 1 × Fac 2	$\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2$	$(I-1)(J-1)$	$\frac{(n-I.J)\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2}{(I-1)(J-1)\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2}$
Residual	$\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2$	$n - I \cdot J$	

To avoid making the presentation more cumbersome, we have not indicated the mean squares as it is classical in the software outputs.

Remark 1 : Note that for example

$$SC_1 = \sum_{i,j,k}(Y_{i..} - Y_{...})^2$$

can be written also

$$SC_1 = J \cdot K \cdot \sum_i (Y_{i..} - Y_{...})^2.$$

The chosen form is more simple and more coherent with Euclidean norms.

Remark 2: Is the analysis of variance table relevent ? A careful exploration of the table shows that most of information is redundant: basically the $\hat{F}$ are sufficient.

2.3 Additive model : equi-replicated case

Removing E_3 and $H_0^{(3)}$, we can perform the same kind of computations, obtaining the following table. The main difference is that, for dimension reasons, K can take now the value 1.

We get

Source	Sum of squares	Degres of freedom	$\widehat{F}$
Factor 1	$\sum_{i,j,k}(Y_{i..} - Y_{...})^2$	$I-1$	$\dfrac{(n-I-J+1)\sum_{i,j,k}(Y_{i..} - Y_{...})^2}{(I-1)\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2}$
Factor 2	$\sum_{i,j,k}(Y_{.j.} - Y_{...})^2$	$J-1$	$\dfrac{(n-I-J+1)\sum_{i,j,k}(Y_{.j.} - Y_{...})^2}{(J-1)\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2}$
Residual	$\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2$	$n-I-J+1$	

2.4 Choosing a model

Remark first that in the equireplicated case with $K = 1$, the interactive model can't be assumed because it is too large.

In the other cases, we have to set the complete model and the first question is to test the interaction. If it is significative, the two factors are relevant, the model is the good one and the test of the mains effect can have only an interest for the description of the magnitude of the effects.

If it is not, two strategies are possible.

- The **pooling strategy**. In that case we will assume that non significative means null : the non significative interaction term is removed from the model and its sum of squares is **pooled** with the residual sum of square.

In a second step, mains effects are tested, and removed if non significative, with the problem that they can be both non significative and a kind of backward procedure can be used as in regression. See for example proc glmselect in SAS.

-The **non pooling strategy**. As a precaution we don't consider non-significative as null and we keep the interaction in the model. We must then define the hypothesis of nullity of main effect in the complete model. In case of equi-repetition this is again easy and we can define the absence of the first factor in the regular model 3.1 as

$$\beta_{i.} = (const) \tag{3.3}$$

In the non equi-replicated case, there are several definition and thus several analysis of variance table . (3.3) is a possible choice that corresponds in the classical terminology to the type III analysis of variance and it is recommended as a first choice. An other possible choice is type I analysis of variance that is sequential and depend on the order of writing. For example the type I analysis of variance of the interactive model A,B,A*B defines

- the sum of squares associated to the first factor A, by difference between the void model (with the constant only) and the model with the first factor A.
- the sum of squares associated to B, by difference between the model with A and the additive model with A,B
- the sum of squares associated to the interaction by its only possible definition: difference between A,B and A,B,A*B.

We see that the Type I analysis of A,B,A*B differs from that of B,A,A*B.

Except from special models, as polynomial regression where there is natural order between the terms of the model, this make little sense.

2.5 Difference between balanced experiments and others

First remark that in real life the unbalanced case (non equireplicated) is the most common because mainly of missing data: even if a balanced experiment has been designed, in most of the cases some data disappear and the final data set is unbalanced.

The balanced case has the property that analysis of variance table is unique and explicit as explained is Section 2.2. More precisely the model is orthogonal (under some constraint system) in the sense of Chapter 5.

In the unbalanced case, several analysis of variance table can be constructed. We have briefly described Type I and Type III that are the most classical. Note that the different estimators are means of means and are, in general, not ordinary means. The sum of squares have in general no explicit expression because mainly the least squares estimators are solutions of a linear system and are not explicit. The computation cost is more important That's for example the difference between `proc anova` (for balanced experiments) and `proc glm` in SAS.

As a curiosity we mention that the sum of square associated to one factor is given in the book [18] p. 87-93 :

$$SC = \sum_{i=1}^{I} w_i \left(\left[\frac{\sum w_i \widehat{\beta}_{i.}}{\sum w_i} \right] \widehat{\beta}_{i.} \right)^2$$

where for every $i = 1, \cdots, I$, the weight w_i is defined by :

$$\operatorname{Var}\left(\widehat{\beta}_{i.}\right) = \frac{\sigma^2}{w_i} \quad \text{et donc} \quad \frac{\sigma^2}{w_i} = \frac{\sigma^2}{J^2} \sum_{j=1}^{I} \frac{1}{n_{ij}}.$$

3 Extensions

3.1 Multiple comparisons

Let us consider, for example, a comparative experiment with two factors. One is the factor of interest, for example drug, and the second is a control factor, for example subject. Traditionally, in design of experiment technology, the first one is called "treatment" and the second "bloc".

The main purpose of the experiment is the test on the "treatment" factor. If it non significative, the experiment is negative and the story stops. But in the other case a natural question arises: If among the I levels of the treatment, some difference exists what form do they take ? For example, can we sort the I treatments or can we perform comparison with the control which is, for example, Treatment 1 ?

If we have decided, a priori, to compare Treatments 1 and 2, this can be done easily by the T test on H_0 $\alpha_1 = \alpha_2$ or equivalently $\beta_{1.} = \beta_{2.}$

$$\widehat{T}_{12} = \frac{\widehat{\beta_{1.}} - \widehat{\beta_{2.}}}{\sqrt{\widehat{\mathrm{Var}}\,(\widehat{\beta_{1.}} - \widehat{\beta_{2.}})}}.$$

This test has a level α (note that α has nothing to do with α_i). But suppose, for example, that $I = 12$, if we test each one of $12 \times 11/2 = 66$ pairwise comparisons by a α T test, the probability of making at least an error : the FWER (FamilyWise Error Rate) will be much larger that α. Below we present briefly some method to control this FWER.

Case of Pairwise comparisons

i. Tukey method : it is adapted to balanced casee or to one way analysis of variance. It gives simultaneous confidence intervals (All of them are correct with probability $1-\alpha$) for all the difference between the means $\alpha_i - \alpha_j; 1 \leq i < j \leq I$. In the balanced case it is the most precise .

ii. Bonferroni's method : This a very crude method based on a simple union bound : if we want to control a global risk of α on $I(I-1)/2$ comparisons, a solution is to perform each of the T tests at the level

$$\alpha' = \frac{\alpha}{I(I-1)/2}.$$

It is particularly adapted to the case where I is small and the design is unbalanced.

iii. Scheffé method : It is a very robust method which consist of constructing a confidence ellipsoid for the vector $\alpha_1, ..., \alpha_I$, using a general Fisher test. In a second step, this ellipsoid is projected on the axes that correspond to the coordinates $\alpha_i - \alpha_{i'}$.

Comparison to a control. :

In some situations, the aim of the experiment is the comparison to a particular treatment (say I): the control. This can be the placebo in medical experiments. In that case, all the $I(I-1)/2$ comparisons do not have to be considered, but only the $I-1$ comparisons to the control. The Bonferroni method is easily adapted by choosing

$$\alpha' = \frac{\alpha}{(I-1)}.$$

The equivalent of the Tukey method is now the Dunnet method that constructs simultaneous confidence intervals for the

$$\alpha_i - \alpha_I \ ; \ 1 \leq i \leq (I-1).$$

The reader can find a detailed presentation in Miller [16].

3.2 Several factors, crossed factors and nested factors

When more than two factors are present, we can define: the mean, the main effects, the interactions of order two, of order three etc... A particular case has to be detailed: the nested case.

Definition 3.2 *Two factor are crossed if their levels make sense independently one to another.*
Factor B is nested to factor A is for example $B = 2$ makes sense only if we know the value of A

Here are some example of factors that are usually crossed.

- variety *location$\Longrightarrow$ to predict a yield;
- Type of car $*$ type of traject $\Longrightarrow$ to predict fuel consomption ;
- Type of supermarket $*$ region $\Longrightarrow$ To predict annual sales of a supermarket

Here are some examples of nested factors :

- Burger / sample in burger $\Longrightarrow$ for a bacteriological test ;
- Plant / worker $\Longrightarrow$ for the yield of a worker ;
- doe/ number of the litter / number of rabbit. $\Longrightarrow$ in animal genetics.

If the first example, suppose that 3 samples are taken at random in a burger. There are no relation between all the samples sharing the same number.

The case of nested factor must be declared with you favorite software. Indeed the main effect of the nested factor must not be introduced. More precisely, in case of two factors, the second nested to the first the decomposition is

$$Y_{ijk} = \mu + \alpha_i + \gamma'_{ij} + \varepsilon_{ijk}, \tag{3.4}$$

with then constraints $\sum_i \alpha_i = 0$ and for all i, $\sum_j \gamma'_{ij} = 0$.

Remark: consider the burger example and suppose we have 4 burgers and that we take 3 samples by burger numbered 1, 2 or 3. Then we are in the nested situation. Suppose that alternatively we decide to number the samples globally from 1 to 12. This can definitively be done and gives another relation between the two factors . The factor burger with 4 levels is now included in the sample factor with 12 levels. We give no details.

3.3 Testing homogeneity of variances

The graphs of residuals can show a larger dispersion in some region of the experience. This is the case in particular if we suspect that the variance depend on the level of some factor. A natural test is the Bartlett test which is the likelihood ratio test between the homosedastic and the heteroscedastic models. Nevertheless this test if very affected by non normality of residuals, even for large data sets. A safer choice can be the Levene ([12]) test or its modification based on squares. Let us present them in the case of one-way analysis of variance.

Let the Y_{ij} be the observations. From the analysis of variance analysis we compute the residuals: $\widehat{\varepsilon}_{ij}$.

In a second step we perform an new analysis of variance and the corresponding Fisher test on

$$|\widehat{\varepsilon}_{ij}|.$$

or

$$(\widehat{\varepsilon}_{ij})^2.$$

Both give a test of homogeny of the variance.
As for regression the two solutions in case of heterogeny are

- transformation of the response variable Y with the same rules ;

- using a generalized linear model.

4 Computer example

4.1 Balanced two ways ANOVA

Data are taken from Calas *et al.* (1998) [5]. In this experiment two solution for disinfecting the roots of teeths are compared of two strains of germs of Prevotella nigrescens, a wild sampled strain and a reference one (NCTC 9336).

The response in the mean number of germs remaining after the treatment.

SAS software :

```
proc glm data=sasuser.dents;
class trait germe;
model lnbac=trait germe trait*germe;
output out=sortie predicted=p student=r;
lsmeans trait germe trait*germe/out=graph;
run; quit;
proc gplot data=sortie;
plot r*p;run; quit;
proc gplot data=graph;
plot lsmean*germe=trait;run; quit;
```

Comments :

- The second line declares `trait` and `germe` as qualitative variables or factors
- The third line declares the standard two-ways ANOVA model.
- The forth line writes the outputs on a new data.
- The fifth ask for ajusted means (commande`lsmeans`) and write them in the data "graph".
- Le last lines make the graphs

Note that the Levene test is available only in case of one-way ANOVA by `means .../hovtest=levene;`. In the other cases you have to save the residuals and to reanalyze them.

```
Dependent Variable: LNBAC
                                   Sum of          Mean
Source                  DF        Squares        Square    F Value     Pr > F
Model                    3     26.8258464     8.9419488      22.43     0.0001
Error                   60     23.9183125     0.3986385
Corrected Total         63     50.7441589

                  R-Square           C.V.      Root MSE            LNBAC Mean
                  0.528649       98.91739       0.63138               0.63829
```

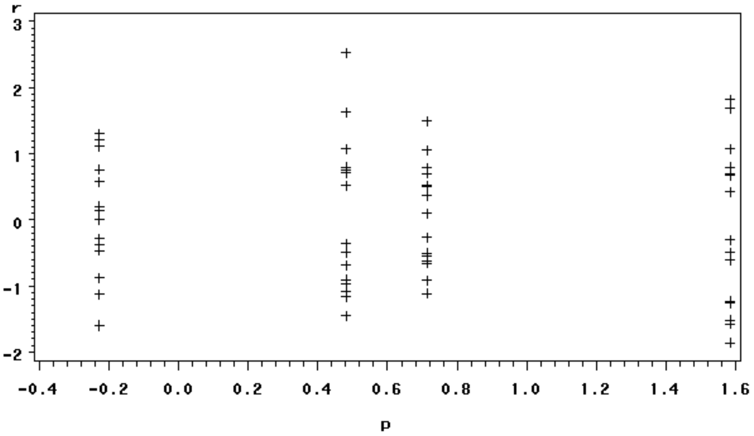


Figure 3.1: Residual Plot

```
Source                  DF     Type III SS     Mean Square    F Value    Pr > F
TRAIT                    1      0.10378062      0.10378062       0.26    0.6118
GERME                    1     16.70226119     16.70226119      41.90    <.0001
TRAIT*GERME              1     10.01980464     10.01980464      25.14    <.0001

  Least Squares Means

  TRAIT        LNBAC         GERME        LNBAC         TRAIT    GERME          LNBAC
              LSMEAN                     LSMEAN                                LSMEAN
  1       0.67855719         1       1.14914344         1        1         1.58508813
  2       0.59801969         2       0.12743344         1        2        -0.22797375
                                                        2        1         0.71319875
                                                        2        2         0.48284063
```

R software :

The same analysis can be performed by

```
dents=read.table("C:/Donnees/dents.txt",header=TRUE)
attach(dents)
germe=as.factor(germe)
trait=as.factor(trait)
library(car)
```

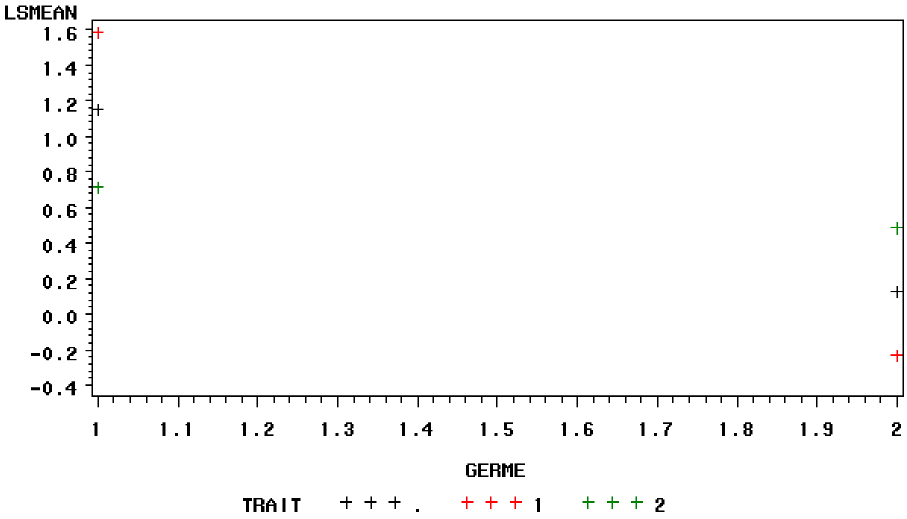


Figure 3.2: Plot of interactions

```
dents.lm=lm(LNBAC~germe:trait-1,contrasts=list(germe=contr.sum,trait=contr.sum))
summary(dents.lm)
anova(dents.lm)
Anova(dents.lm,type="II")
Anova(dents.lm,type="III")
plot(dents.lm$fit,dents.lm$res)
interaction.plot(trait,germe,LNBAC,fixed=TRUE,col = 2:3,leg.bty = "o")
interaction.plot(germe,trait,LNBAC,fixed=TRUE,col = 2:3,leg.bty = "o")
plotMeans(LNBAC,germe,trait)
library(Rcmdr)
levene.test(LNBAC,trait:germe)
```

Comments on the commands :

- Before making an ANOVA it must be checked that the factors have been declared as qualitative variables. This is the object of commands 3 and 4.

- Once the model stated the types I, II and III analysis of variance (commands 8, 9 and 10) are performed once you have called the command `car` (command 5). **You must use the option `lcontrasts` in `lm` as shown, unless type III analysis will be false.**

- Graphs are constructed in the last 4 commands (`plotMeans` makes the same as `interaction.plot` but with another presentation).

- last command performs the Levene test.

here is a partial output + :

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
germe1:trait1   1.5851     0.1578  10.042 1.82e-14 ***
germe2:trait1  -0.2280     0.1578  -1.444  0.15386
germe1:trait2   0.7132     0.1578   4.518 2.98e-05 ***
germe2:trait2   0.4828     0.1578   3.059  0.00332 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6314 on 60 degrees of freedom
Multiple R-Squared: 0.6886,     Adjusted R-squared: 0.6679
F-statistic: 33.18 on 4 and 60 DF,  p-value: 1.36e-14

>Anova Table (Type III tests)

Response: LNBAC
             Sum Sq Df F value    Pr(>F)
(Intercept) 26.0744  1 65.4086 3.475e-11 ***
germe       16.7023  1 41.8983 1.968e-08 ***
trait        0.1038  1  0.2603    0.6118
germe:trait 10.0198  1 25.1351 5.033e-06 ***
Residuals   23.9183 60

>Levene's Test for Homogeneity of Variance

      Df F value  Pr(>F)
group  3  2.3318 0.08315 .
      60
```

General comments : This data set is perfectly balanced implying that type I II and III coincide. Note that the listing propose a Student test of individual terms (including the intercept) and that this test makes no sense in our case. An examination of the Fisher tests shows that the interaction is significative but that mains effect are not. Looking at the graphics permits to understand this paradox: one treatment is efficient on one strain and the other treatment on the other strain. Eventually the Leven test is coherent with homoscedasticity.

4.2 Unbalanced two-ways ANOVA

We consider the time for germination of carrots seeds (variable `jg`), measured in days as a function of the variety and of the type of soil. This data is extracted from Searle [18]. The data is unbalanced: the number of replication varies from 1 to 3. The commands are

R software :

```
attach(carotte)
sol=as.factor(sol)
var=as.factor(var)
library(car)
carotte.lm=lm(jg~var*sol,contrasts=list(var=contr.sum,sol=contr.sum))
anova(carotte.lm)
Anova(carotte.lm,type="II")
Anova(carotte.lm,type="III")
```

Giving the results :

```
> anova(carotte.lm)
Analysis of Variance Table

Response: jg
          Df  Sum Sq Mean Sq F value   Pr(>F)
var        2  93.333  46.667  3.5000 0.075085 .
sol        1  83.901  83.901  6.2926 0.033393 *
var:sol    2 222.766 111.383  8.3537 0.008888 **
Residuals  9 120.000  13.333

> Anova(carotte.lm,type="II")
Anova Table (Type II tests)

Response: jg
           Sum Sq Df F value   Pr(>F)
var       124.734  2  4.6775 0.040475 *
sol        83.901  1  6.2926 0.033393 *
var:sol   222.766  2  8.3537 0.008888 **
Residuals 120.000  9
```

```
> Anova(carotte.lm,type="III")
Anova Table (Type III tests)

Response: jg
            Sum Sq Df  F value    Pr(>F)
(Intercept) 3497.5  1 262.3114 5.784e-08 ***
var          192.1  2   7.2048  0.013546 *
sol          123.8  1   9.2829  0.013865 *
var:sol      222.8  2   8.3537  0.008888 **
Residuals    120.0  9
```

SAS software :

```
proc glm data=carotte;
class var sol;
model jg=var sol var*sol;
lsmeans var*sol;
run; quit;
```

Giving the results :

```
Dependent Variable: JG
                                   Sum of          Mean
Source                  DF        Squares        Square  F Value    Pr > F
Model                    5     400.000000     80.000000     6.00    0.0103
Error                    9     120.000000     13.333333
Corrected Total         14     520.000000

                  R-Square           C.V.      Root MSE            JG Mean
                  0.769231       24.34322       3.65148            15.0000

 Source                 DF        Type I SS     Mean Square  F Value     Pr > F
VAR                      2       93.3333333      46.6666667     3.50    0.0751
SOL                      1       83.9007092      83.9007092     6.29    0.0334
VAR*SOL                  2      222.7659574     111.3829787     8.35    0.0089

Source                  DF      Type III SS   Mean Square  F Value     Pr > F
VAR                      2       192.127660     96.063830     7.20    0.0135
SOL                      1       123.771429    123.771429     9.28    0.0139
```

```
VAR*SOL                     2       222.765957      111.382979      8.35     0.0089

                                       Least Squares Means

                                    VAR    SOL               JG
                                                         LSMEAN
                                    1      1          9.0000000
                                    1      2         16.0000000
                                    2      1         14.0000000
                                    2      2         31.0000000
                                    3      1         18.0000000
                                    3      2         13.0000000
```

General comments : The unbalanced character of the design is easily checked by comparing type I and Type III analysis. For the rest of the discussion we focus on type III analysis. The experiment shows clearly (everything is significative) that the time to germination depend in a complex manner on the soil and the variety. To have a reliable prediction, it is necessary to use the lsmeans at the crossed level `sol*var`.

4.3 Nested ANOVA

The data set is from the book of Milliken et Johnson [17]). We want to compare 8 insecticides (variable `produit`) originated from 4 firms . Each firm produces exactly two products numbered 1 or 2. Three replications have been made and the final observation is the number of alive mosquitos out of a box of 400.

SAS software :

```
proc glm data=sasuser.insect;
class firme produit;
model nb=firme produit(firme);
means firme/tukey;
run;quit;
```

thart gives:

```
                                           Sum of
  Source                      DF          Squares     Mean Square    F Value    Pr > F
  Model                        7      20605.33333      2943.61905      49.33    <.0001
  Error                       16        954.66667        59.66667
  Corrected Total             23      21560.00000
```

```
                  R-Square     Coeff Var      Root MSE       nb Mean
                  0.955720      7.022200      7.724420      110.0000

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
firme                        3     19971.66667      6657.22222     111.57    <.0001
produit(firme)               4       633.66667       158.41667       2.66    0.0714

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
firme                        3     19971.66667      6657.22222     111.57    <.0001
produit(firme)               4       633.66667       158.41667       2.66    0.0714


                     Tukey's Studentized Range (HSD) Test for nb

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

                     Alpha                                   0.05
                     Error Degrees of Freedom                  16
                     Error Mean Square                   59.66667
                     Critical Value of Studentized Range  4.04609
                     Minimum Significant Difference        12.759

             Means with the same letter are not significantly different.

              Tukey Grouping          Mean      N    firme
                           A       141.167      6    b
                           A
                           A       134.667      6    a
                           B        91.833      6    c
                           C        72.333      6    d
```

R software :

```
attach(insect)
firme=as.factor(firme)
produit=as.factor(produit)
insect.lm=lm(nb~produit:firme+firme,contrasts=list(produit=contr.sum,firme=contr.sum))
anova(insect.lm)
```

```
Anova(insect.lm,type="III")
insect.aov=aov(nb~firme)
TukeyHSD(insect.aov,ordered=TRUE)
```

The results are :

```
Response: nb
              Df  Sum Sq Mean Sq F value    Pr(>F)
firme          3 19971.7  6657.2 111.574 6.135e-11 ***
produit:firme  4   633.7   158.4   2.655    0.0714 .
Residuals     16   954.7    59.7

Anova Table (Type III tests)

Response: nb
              Sum Sq Df  F value    Pr(>F)
(Intercept)   290400  1 4867.039 < 2.2e-16 ***
firme          19972  3  111.574 6.135e-11 ***
produit:firme    634  4    2.655    0.0714 .
Residuals        955 16

  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = nb ~ firme)

$firme
        diff       lwr      upr
c-d 19.50000  5.099148 33.90085
a-d 62.33333 47.932481 76.73419
b-d 68.83333 54.432481 83.23419
a-c 42.83333 28.432481 57.23419
b-c 49.33333 34.932481 63.73419
b-a  6.50000 -7.900852 20.90085
```

Discussion : Since Type I and Type III are equals the design is balanced. The residual plots have nothing particular and have been omitted. The nested effect is rather non-significative but the result is not clear-cut. The firm clearly differ and the Tukey multiple comparison procedure show that firm a and b don't differ .

5 Exercises

Exercise 3.1

(*) Soit le jeu de données suivant pour deux facteurs à deux niveaux (données totalement inventées pour que les calculs tombent juste) :

facteur 1	1	1	1	1	1	2	2	2
facteur 2	1	1	1	2	2	1	1	2
réponse	19	15	14	10	6	9	11	6

Calculez les estimateurs dans le cas des paramétrisations (3.1) et (3.2). On calculera plus rapidement si on représente les données dans un tableau 2×2.

Exercise 3.2

(*) [Hétéroscédasticité] Soit un modèle d'analyse de la variance à un facteur, vérifiant les postulats **P1, P3** et **P4** et tel que pour les I modalités du facteur on ait $\mathrm{Var}\,(\varepsilon_{ij}) = \sigma_i^2$ pour tout $j = 1, \ldots, n_i$. La variance des erreurs dépend donc de la modalité considérée. Déterminer alors les estimateurs $\widehat{\mu}_i$ et $\widehat{\sigma}_i^2$ par maximum de vraisemblance des différents paramètres du modèle (soit $2I$ paramètres). Si on suppose maintenant que $\mu_i = \mu$ pour $i = 1, \ldots, I$, que valent alors les $\widehat{\sigma}_i^2$ et $\widehat{\mu}$?

Exercise 3.3

(**) Montrer que $H_0^{(1)}$ est équivalente à $\beta_{i.} = (cte)$ dans le modèle (3.1). Montrer que $H_0^{(3)}$ est équivalente à $\forall (i,j) \neq (i',j') \in \{1, \ldots, I\} \times \{1, \ldots, J\} : \beta_{ij} - \beta_{i'j} - \beta_{ij'} + \beta_{i'j'} = 0$.

Exercise 3.4

(**) Pour introduire des effets différentiels dans un modèle d'analyse de la variance à **un** facteur,

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik},$$

quel type de contrainte doit-on utiliser ?

$$\sum_{i=1}^{I} \alpha_i = 0 \text{ ou bien } \sum_{i=1}^{I} n_i \alpha_i = 0 \text{ ?}$$

Eléments de solution : Par cohérence avec la décomposition marginale utilisée dans le cas où il y a deux facteurs, on serait tenté de poser la première contrainte. Ce n'est pas la bonne réponse. En effet :

- *L'estimation et donc la définition précise du paramètre μ importe peu, puisque ce paramètre de moyenne générale est souvent sans intérêt. N'oublions pas que le but d'une expérience est de comparer; toute l'attention est donc focalisée sur les effets différentiels α_i. Le premier type de contraintes a peu d'intérêt.*

- *Le second type n'a d'autre intérêt que calculatoire. Si on l'utilise, l'estimateur $\widehat{\mu}$ de μ est la moyenne générale : $Y_{..}$ (estimateur sous l'hypothèse nulle), nécessaire pour la construction du test de Fisher. En particulier, la formule $SC = \sum_{ik}(Y_{i.} - Y_{..})^2 = \sum_{ik}(\widehat{\alpha}_i)^2$ ne serait pas vraie sinon.*

Cette réponse est en contradiction avec celle que l'on fait pour deux facteurs. Nous voyons donc une fois de plus sur cet exemple, l'intérêt de travailler avec des modèles réguliers, comme l'est le modèle de notre première présentation.

Exercise 3.5

(**) [Décomposition de type I] Soit un modèle linéaire et supposons que l'on ait scindé le paramètre β en différents sous-ensembles $\beta_1, \cdots, \beta_m$. Une telle décomposition est appelée une *partition*. Une partition naturelle est la décomposition en effet principaux et interaction dans le modèle à deux facteurs croisés (3.2). En toute rigueur, ce modèle est non-régulier et le lecteur pourra consulter le chapitre 5. La partition fait que l'on peut écrire

$$Y = X_1 \cdot \beta_1 + \cdots + X_m \cdot \beta_m + \varepsilon.$$

Notez bien que l'ordre importe dans l'écriture du mode et que nous supposerons toujours que les effets principaux sont avant les interactions, que les interactions doubles sont avant les triples, etc... De manière générale, on définit les espaces V_i $i = 1, \cdots, m$ de la façon suivante :

$$V_i \text{ est l'orthogonal de } [X_1, \cdots, X_{i-1}] \text{ dans } [X_1, \cdots, X_i].$$

On définit le test associé au i-ème élément du modèle comme le test de l'hypothèse nulle :

$$P_{[V_i]} X \cdot \beta = 0.$$

On considère désormais un modèle d'analyse de la variance à deux facteurs et on désire pondérer la décomposition par les effectifs :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

avec maintenant les contraintes suivantes : $\forall j,\ \sum_i n_{ij}\gamma_{ij} = 0,\quad \forall i,\ \sum_j n_{ij}\gamma_{ij} = 0$, et également $\sum_i n_{i+}\alpha_i = 0,\quad \sum_j n_{+j}\beta_j = 0$ (le + veut dire la somme sur l'indice qu'il remplace). Montrer que l'hypothèse $\forall i,\ \alpha_i = 0$ s'écrit dans le modèle (3.1) sous la forme : "$h_i := \sum_j n_{ij}\beta_{ij}$, ne dépend pas de i". Montrer que le test de cette hypothèse est obtenu par la décomposition de type I.

Exercise 3.6

(**) Nous allons illustrer une nouvelle fois l'abominable complexité de l'option `solution` de SAS en analyse de la variance à deux facteurs. Voici un exemple volontairement simple et dont les données ont été inventées. L'utilisation de `solution` donne la valeur -12 à la fin du tableau pour $a * b = (1, 1)$. Comment s'interprète t-elle ?

Les données sont

a	1	1	1	1	1	2	2	2
b	1	1	2	2	2	1	2	2
Y	5	3	25	27	32	12	12	21

```
proc glm ;
class a b;
model y= a b a*b/solution;
lsmeans a b a*b;run; quit;
```

Voici un extrait des résultats numériques obtenus :

```
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        3     792.0000000     264.0000000      22.96    0.0055
Error                        4      46.0000000      11.5000000
Corrected Total              7     838.0000000

            R-Square     Coeff Var      Root MSE        y Mean
            0.945107      17.84824      3.391165      19.00000

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
a                            1       6.8571429       6.8571429       0.60    0.4831
b                            1     555.4285714     555.4285714      48.30    0.0023
a*b                          1      61.7142857      61.7142857       5.37    0.0814

                                        Standard
Parameter             Estimate             Error    t Value    Pr > |t|
Intercept          24.00000000 B      2.39791576      10.01      0.0006
a         1         4.00000000 B      3.09569594       1.29      0.2659
a         2         0.00000000 B       .                .         .
b         1       -12.00000000 B      4.15331193      -2.89      0.0446
b         2         0.00000000 B       .                .         .
a*b       1 1     -12.00000000 B      5.18009009      -2.32      0.0814
a*b       1 2       0.00000000 B       .                .         .
a*b       2 1       0.00000000 B       .                .         .
a*b       2 2       0.00000000 B       .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
was used to solve the normal equations. Terms whose estimates are followed by
the letter 'B' are not uniquely estimable.
```

```
a      y LSMEAN       b    y LSMEAN      a   b     y LSMEAN
1      16.0000        1      8.0000      1   1       4.0000
2      18.0000        2     26.0000      1   2      28.0000
                                         2   1      12.0000
                                         2   2      24.0000
```

Chapter 4

Analysis of covariance

In this chapter the way of mixing quantitative and qualitative variable in a linear model is presented. We define "heterogeneity of the slopes" which is in some sense a kind of interaction between a factor and a variable.

1 The model

In many situations, the set of explanatory variable contains both quantitative and qualitative variables. The Analysis of Covariance model (ANACOVA) is the linear model that mixes these variables. We begin with the simplest case of aone variable and one factor.

1.1 An exemple

We consider the stature, the height, of young girls at different ages from 6 to 10. For a fixed girl and in this particular period, a linear growth gives a very good fit. But every one knows that some individuals are taller and they may grow more rapidly. To take this into account the intercept and the slope of the regression must be specific to the individual.

Let Y_{ij} be the height of individual i at age number j, a possible model is

$$Y_{ij} = \mu_i + \beta_i \cdot age_j + \varepsilon_{ij}.$$

It is easy to check that this model is linear.

Two main questions are to be considered :

i. Are the slopes β_i different ? If yes, the model will be called "with heterogeneity of slopes".

ii. Are the intercepts μ_i different ? If yes, the model will be called "with heterogeneity of intercepts " .

As for analysis of variance, we can introduce differential effects $i = 1, \cdots, I$:

$$\mu_i = \mu + \alpha_i \quad \text{et} \quad \beta_i = \beta + \gamma_i$$

with the usual conditions

$$\sum_i \gamma_i = \sum_i \alpha_i = 0.$$

Giving the model :

$$Y_{ij} = \mu + \alpha_i + \beta \cdot age_j + \gamma_i \cdot age_j + \varepsilon_{ij}, \tag{4.1}$$

This model, which is not regular, must be viewed as a rewriting of the first one. The last term corresponding to the " heterogeneity of slopes" that depends both on the quantitative and the qualitative variable can be viewed as an interaction. It has to be declared in this manner for example in SAS

```
heigth= individual + age + age * individual.
```

We test, first, the heterogeneity of slopes

- for all $i = 1, \cdots, I,\ \gamma_i = 0$.

If this test is non-significative, it is worth testing the heterogeneity of intercepts.

- for all $i = 1, \cdots, I,\ \alpha_i = 0$.

If again this test is non-significative, the factor can removed from the model.

1.2 The general model

In case of several variables and factors we can mix all the methods of regression (transformation of regressors) and analysis of variance (constructing interactions) to obtain rather complicated models. It difficult to describe all possible model and large size example are tedious.

Finally the analysis of covariance model are all linear models and the general formula apply. Only the interpretation is a little particular as you can see in the numerical example.

2 Numerical example

Data are from Tanner [19]. the, height (cm) of young girls has been measured at every age between 6 and 10 years.

ind	6 ans	7 ans	8 ans	9 ans	10 ans
1	116	122	126.6	132.6	137.6
2	117.6	123.2	129.3	134.5	138.9
3	121	127.3	134.5	139.9	145.4
4	114.5	119	124	130	135.1
5	117.4	123.2	129.5	134.5	140
6	113.7	119.7	125.3	130.1	135.9
7	113.6	119.1	124.8	130.8	136.3

In this period that excludes early childhood and puberty the growth of an human being is almost linear with some differences between males an females.

To make the interpretation easier we introduce the variable `agec` which is `age` after centering. **SAS software** :

To have tests and estimators in a convenient form we need 3 call of `proc glm`

```
proc glm data=...;
class ind;
model taille=agec ind agec*ind;run;quit;
proc glm data=...;
class ind;
model taille=age ind age*ind;run;quit;
proc glm data=...;
class ind;
model taille=agec*ind/solution noint;run;quit;
```

```
--------------------------------------------------------------------------------------
                                      Sum of
  Source                  DF         Squares     Mean Square    F Value    Pr > F
  Model                   13     2491.474429      191.651879     981.39    <.0001
  Error                   21        4.101000        0.195286
  Corrected Total         34     2495.575429

                  R-Square     Coeff Var      Root MSE    taille Mean
                  0.998357      0.346566      0.441911       127.5114

--------------------------------------------------------------------------------------
  Source                  DF       Type I SS     Mean Square    F Value    Pr > F
  age                      1     2169.515571     2169.515571    11109.4    <.0001
  ind                      6      316.459429       52.743238     270.08    <.0001 (A)
  age*ind                 6        5.499429         0.916571        4.69    0.0035

  Source                 DF      Type III SS     Mean Square     F Value    Pr > F
  age                     1      2169.515571     2169.515571     11109.4    <.0001
  ind                     6        4.253299         0.708883        3.63    0.0125 (B)
  age*ind                 6        5.499429         0.916571        4.69    0.0035

-------------------------------------------------------------------------------------
  Source                 DF        Type I SS     Mean Square     F Value    Pr > F
  agec                    1      2169.515571     2169.515571     11109.4    <.0001
  ind                     6       316.459429       52.743238      270.08    <.0001 (C)
  agec*ind                6        5.499429         0.916571        4.69    0.0035

  Source                 DF      Type III SS     Mean Square     F Value    Pr > F
  agec                    1      2169.515571     2169.515571     11109.4    <.0001
  ind                     6       316.459429       52.743238      270.08    <.0001 (C)
  agec*ind                6        5.499429         0.916571        4.69    0.0035

-------------------------------------------------------------------------------------
                                                Standard
          Parameter             Estimate           Error     t Value     Pr > |t|
          ind        1       126.9600000      0.19762880      642.42       <.0001
          ind        2       128.7000000      0.19762880      651.22       <.0001
            :                     :               :            :           :
          agec*ind   1         5.3800000      0.13974467       38.50       <.0001
          agec*ind   2         5.3900000      0.13974467       38.57       <.0001
            :                     :               :            :           :
```

The last command uses a regular model giving estimator with the `/solution` option in a nice form.

R software :

```
ind=as.factor(ind)
library(car)
fille.new=groupedData(taille~age|ind,data=fille)
plot(fille.new)
anova(lm(taille~age*ind))
Anova(lm(taille~age*ind),type="III",contrasts=list(age=contr.sum,ind=contr.sum))
anova(lm(taille~agec*ind))
summary(lm(taille~(agec:ind-1)+ind))
```

Giving the results (extract) :

```
Response: taille
          Df  Sum Sq Mean Sq    F value    Pr(>F)
age        1 2169.52 2169.52 11109.4433 < 2.2e-16 ***
ind        6  316.46   52.74   270.0824 < 2.2e-16 *** (A)
age:ind    6    5.50    0.92     4.6935  0.003547 **
Residuals 21    4.10    0.20

Anova Table (Type III tests)

Response: taille
             Sum Sq Df   F value    Pr(>F)
(Intercept) 1067.06  1 5464.0736 < 2.2e-16 ***
age          289.44  1 1482.1565 < 2.2e-16 ***
ind            4.25  6    3.6300  0.012520 *     (B)
age:ind        5.50  6    4.6935  0.003547 **
Residuals      4.10 21

Response: taille
          Df  Sum Sq Mean Sq    F value    Pr(>F)
agec       1 2169.52 2169.52 11109.4433 < 2.2e-16 ***
ind        6  316.46   52.74   270.0824 < 2.2e-16 *** (C)
agec:ind   6    5.50    0.92     4.6935  0.003547 **
Residuals 21    4.10    0.20

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
ind1      126.9600     0.1976  642.42   <2e-16 ***
ind2      128.7000     0.1976  651.22   <2e-16 ***
  :           :           :       :         :
agec:ind1   5.3800     0.1397   38.50   <2e-16 ***
agec:ind2   5.3900     0.1397   38.57   <2e-16 ***
  :           :           :       :         :
```

Comments : The fit is excellent . First $R^2 \simeq 0.998$ but moreover the standard error is of about 4 mn which is very good and imply the use of data of high quality. The analysis shows that the growth is linear (because of the excellent fit) and that the speed of growing (the slope) depends on the individual. As for the test of `ind` three tests are proposed

- Type I with `age` (A);
- Type III with `age` (B);

- Type I or III with `agec` (C)(they are identical, since the model is orthogonal)

It is easy to check that (A) and (C) are identical and test the hypothesis " the mean height are identical"

This is the one that makes sense and the hypothesis is rejected.

(B) tests " the height extrapolated at age 0 are equal'. This mean little since in the early ages, the growth of human being is not linear. For example this extrapolated is about 70 cm which is not the size of a newborn.

We see one of the few interests of The TypeI decomposition : we don't need to center the variable to get the good test.

Discussion: The experiment shows that in the period considered, the growth is linear. Some individuals are taller and some grow faster than the others.

Chapter 5

Non regular models and orthogonality

In this chapter we present the main tools to study non regular linear models and to define orthogonality that simplify the processing of models.

1 Non regular models

Some models cannot be parametrized in a regular manner : they are over-paramterized. The most common model is the additive model in two-ways analysis of variance. Consider, for example, the very simple case where $I = J = 2$ and each combination is observed once. With the notation of Chapter 3 :

$$\begin{aligned} Y_{11} &= \mu + a_1 + b_1 + \varepsilon_{11} \\ Y_{12} &= \mu + a_1 + b_2 + \varepsilon_{12} \\ Y_{21} &= \mu + a_2 + b_1 + \varepsilon_{21} \\ Y_{22} &= \mu + a_2 + b_2 + \varepsilon_{22}. \end{aligned}$$

The design matrix X is defined by

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

We can see that every vector of the form $(\alpha+\beta, -\alpha, -\alpha, -\beta, -\beta)$ returns the value 0 when multiplied by X.

The values $\mu, a_i, b_i,\ (i = 1, 2)$ are not uniquely determined : they are not identifiable.

Definition 5.1 *A linear model is said non regular when the design matrix X is not full ranked, i.e. when the kernel of X,* Ker(X) is not restricted to $\{0\}$

let K :=Ker(X) := $\{z \in \mathbb{R}^n,\ X \cdot z = 0\}$ the kernel of X. let us make two remarks :

- $X\widehat{\beta}$ remains unique because it is the projection of X on Y sur $[X]$,
- $\widehat{\beta}$ cannot be unique because if $\widehat{\beta}$ is a solution and if $z \in K$, $\widehat{\beta}+z$ is another solution. The set of solution is indeed $\{\widehat{\beta}+z,\ z \in K\}$.

We use pseudo-inverses

Definition 5.2 *Let M a matrix, M^- is the pseudo-inverse of M if $MM^-M = M$.*

Proposition 5.1 *If $(X'X)^-$ is a pseudo-inverse of $X'X$, then $\widehat{\beta} = (X'X)^-X'Y$ is a solution of the normal equations.*

$$(X'X)\widehat{\beta} = X'Y.$$

Proof : We know that $P_{[X]}Y$ is uniquely defined, as a consequence, there exists $u \in \mathbb{R}^k$ such that $X'Y = X'P_{[X]}Y = X' \cdot X \cdot u$. Define $\widehat{\beta} = (X'X)^-X'Y$. Then, if is easy to check that $\widehat{\beta}$ satisfies the normal equations.

$$\begin{aligned} X'X\widehat{\beta} &= (X'X)(X'X)^-X'Y \\ &= (X'X)(X'X)^-X'Xu \\ &= X'Xu = X'Y \end{aligned}$$

from the definition of pseudo inverse. ■

Among all possible pseudo-inverse of $X'X$ some are more interessant than others .

Identifiability constraints Suppose that rg(X) = dim$[X]$ = $h < k$. There is $(k-h)$ redondant parameters. We define a matrix H with $(k-h)$ rows and k columns and with full rank such that:

$$\mathrm{Ker}(H) \cap \mathrm{Ker}(X) = \{0\}.$$

This means that if we restrict our attention to the β's that satisfy $H\beta = 0$ the kernel of X is now $\{0\}$ and the model is identifiable. In addition we have

Proposition 5.2

- *The matrix $(X'X + H'H)$ is invertible, its inverse is a pseudo-inverse of $X'X$*
- *The vector $\widehat{\beta} = (X'X + H'H)^{-1}X'\,Y$ is the unique solution of the normal equations that satisfies $H\widehat{\beta} = 0$.*

Proof : Dimensions considerations imply that there exists a unique $\widehat{\beta}$ such that

$$X\widehat{\beta} = P_{[X]}Y \quad \text{with} \quad KH\widehat{\beta} = 0.$$

Let us consider the minimization problem in β of $\|Y - X\beta\|^2 + \|H\beta\|^2$.

The value of $\widehat{\beta}$ obtained above is indeed a solution of this minimization problem because it minimize separately each term. The problem can be expressed as

$$\left\| \begin{pmatrix} Y \\ \hline 0 \end{pmatrix} - \begin{pmatrix} X \\ \hline H \end{pmatrix} \beta \right\|^2 \text{ is minimum,}$$

where the bar means the concatenation of matrices. The matrix in the right hand side is full-ranked because.

$$\begin{aligned} \begin{pmatrix} X \\ \hline H \end{pmatrix} \beta = 0 \quad &\Rightarrow \quad X\beta = 0 = H\beta \\ &\Rightarrow \quad \beta \in \operatorname{Ker}([H]) \cap \operatorname{Ker}([X]) \Rightarrow \beta = 0. \end{aligned}$$

We know that the least squares solution of this problem is given by :

$$\widehat{\beta} = \left(\begin{pmatrix} X \\ \hline H \end{pmatrix}' \begin{pmatrix} X \\ \hline H \end{pmatrix} \right)^{-1} \begin{pmatrix} X \\ \hline H \end{pmatrix}' \begin{pmatrix} Y \\ \hline 0 \end{pmatrix} = (X'X + H'H)^{-1}X'Y.$$

It remains to show that $(X'X + H'H)^{-1}$ is a pseudo-inverse of $(X'X)$, which is a direct consequence of

$$(X'X)(X'X + H'H)^{-1}(X'X) = X'P_{[X]}X = X'X$$

because, by definition, $P_{[X]}X = X$. ■

Example 5.1 *Consider the one-way analysis of variance model*

$$Y_{i,j} = \mu_i + \varepsilon_{i,j} \;\; \} \;\; i = 1, \cdots, 4 \;\; et \;\; j = 1, 2.$$

This is a regular model of dimension 4, but if we introduce differential effects

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \qquad i = 1, \cdots, 4 \;\; and \;\; j = 1, 2,$$

we obtain a model of rank 4 but with 5 parameters. The constraint $\sum_{i=1}^{4} \alpha_i = 0$ will restore the identifiability. But the other constraint $\alpha_4 = 0$ will do the same but with another parameterization

The sweep operator

Most of the statical softwares use the ”sweep” operator to handle colinearity between the columns of the design matrix X. In some situation this makes sense, in other not. The program examine sequentially the columns of the matrix X. If the column j is numerically considered (this depends arbitrary on a threshold on the singular values) as collinear to the preceding columns, the column j is removed or, in others words, the constraint $\beta_j = 0$ is chosen. This the default choice and it is not always the better choice. In the introducing example of two-ways analysis of variance the sweep operator will use the constraints $a_2 = b_2 = 0$ which violate symmetry between the levels of the factors.

Estimable functions and contrasts

Some linear function of β : $C'\beta$ have the nice property that they don't depend on the particular solution of the over-parameterized model. They don't depend on the type of constraint chosen. These functions are called "estimable functions". It can be checked that they satisfy $C'\beta = D'X\beta$, where D is a full ranked matrix.

A classical example of functions that are, in general, estimable in analysis of variance is given by the **contrasts**: A linear combination $C'\beta$ is a contrast if the sum of the weights C_i vanishes. In other word if $C'\mathbb{1} = 0$. In the introducing example, $a_1 - a_2$ is a contrast.

2 Orthogonality for regular models

Orthogonality is a notion that permits to simplify the computation but also the interpretation in a linear model. Orthogonality is associated to a partition of the parameter β and consequently of the design matrix X and of information matrix $X'X$. let us give first some exemple of such a partition.

Example 5.2 *Consider the multiple regression model on three variables $Z^{(1)}, Z^{(2)}$ and $Z^{(3)}$:*

$$Y_i = \mu + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \beta_3 Z_i^{(3)} + \varepsilon_i \quad , \quad i = 1, \cdots , \quad n > 4.$$

The vector β contains 4 terms : $\mu = \beta_0, \beta_1, \beta_2, \beta_3$ and the matrix X four columns. In this case, as it is common in regression, we can consider the finest partition : $\{\beta_0\}, \{\beta_1\}, \{\beta_2\}, \{\beta_3\}$. The orthogonality will correspond to the orthogonality, for the ordinary Euclidean metric, of the four lines : $[\mathbb{1}]$, $[\mathrm{Z}^{(1)}]$, $[\mathrm{Z}^{(2)}]$ et $[Z^{(3)}]$. This is equivalent to the fact that the information matrix is diagonal.

Example 5.3 *Let us consider the quadratic regression model depending on two variables $Z^{(1)}$ et $Z^{(2)}$*

$$Y_i = \beta_0 + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \gamma_1 \big(Z_i^{(1)}\big)^2 + \gamma_2 \big(Z_i^{(2)}\big)^2 + \delta Z_i^{(1)}.Z_i^{(2)} + \varepsilon_i \quad , \quad i = 1, \cdots , n > 6.$$

Here we can consider the partition

- *the constant β_0;*

- *linear effects β_1, β_2;*

- *squares γ_1, γ_2;*

- *cross product δ.*

Orthogonality is this the orthogonality of $[\mathbb{1}]$, $[(\mathrm{Z}^{(1)}, \mathrm{Z}^{(2)})]$, $[\left((\mathrm{Z}^{(1)})^2, (\mathrm{Z}^{(2)})^2\right)]$ *and* $[Z^{(1)}Z^{(2)}]$. *And in that case the information matrix is block diagonal.*

We state now our main definition.

Definition 5.3 (Orthogonality for regular models) *Consider a partition in a regular linear model*

$$Y = X\beta + \varepsilon = X_1\beta_1 + \cdots + X_m\beta_m + \varepsilon,$$

where X_i is a matrix of size (n, k_i) ; $\beta_i \in \mathbb{R}^{k_i}$ and $\sum k_i = k < n$). This partition is said orthogonal if the following sub-spaces of $\mathbb{R}^n$

$$[X_1], \cdots, [X_m]$$

are orthogonal.

Equivalently, the model is orthogonal if the the information matrix has a block diagonal structure corresponding to the partition.

To make sense the partition must be natural.

- In regression the most common is the finest partition that separates every variables.

- In analysis of variance, the partition corresponds to the mean; the mains effects and the interactions.

Orthogonality gives two nice properties:

Proposition 5.3 *Consider a regular linear model equipped with and orthogonal partition.*

$$Y = X_1\beta_1 + \cdots + X_m\beta_m + \varepsilon.$$

Then :

- *The estimators of the different components $(\widehat{\beta_i})_{1 \leq i \leq k}$ are independent (non correlated under non Gaussian model).*

- *For $\ell = 1, \cdots, m$, the expression of $\widehat{\beta}_\ell$ does not depend on the presence or absence of the others $\beta_{j'}$ in the model.*

Proof : orthogonality implies that :

$$P_{[X]}Y = P_{[X_1]}Y + P_{[X_2]}Y + \cdots + P_{[X_m]}Y.$$

For $i = 1, \cdots, m$, the estimator $\widehat{\beta}_i$ satisfies

$$X_i\widehat{\beta}_i = P_{[X_i]}Y, \tag{5.1}$$

which implies independence by the properties of the isotrope normal distribution. Note that since X is full-ranked , X_i must also be full-ranked. From (5.1) we deduce

$$\widehat{\beta}_i = (X_i'X_i)^{-1}X_i'Y, \tag{5.2}$$

that gives the second assertion. ■

In addition we get an approximative independence of the Fisher tests on the components of the partition. They are only linked by the estimation of σ^2. When the number of residual degrees of freedom is large, this estimation is almost exact and the link is very small.

An example of application of the second property is the following : consider an orthogonal (for the finest partition) multiple regression model, then

$$\widehat{\beta}_j = \sum_{i=1}^{n} \frac{Z_i^{(j)}Y_i}{\left(Z_i^{(j)}\right)^2}.$$

It is the same as in the simple regression model with the sole variable $Z^{(j)}$.

Example orthogonal polynomial regression:

Let us consider the quadratic regression model:

$$Y = \mu + \beta_1 Z + \beta_2 Z^2 + \varepsilon \tag{5.3}$$

Where the variable Z takes equally spaced values from 1 to n, $\overline{Z} = (n+1)/2$. let $< .,. >$ be the scalar product of $\mathbb{R}^n$ and consider the variables :

$$T^{(0)} := \mathbb{1} \quad ; \quad \mathrm{T}^{(1)} := \mathrm{Z} - \overline{\mathrm{Z}} \quad ; \quad \mathrm{T}^{(2)} := (\mathrm{Z} - \overline{\mathrm{Z}})^2 - \frac{1}{\mathrm{n}} < (\mathrm{Z} - \overline{\mathrm{Z}})^2, \mathbb{1} > .$$

Because it is one-to one, this define a good change of variable, Model (5.3) is equivalent to

$$Y = \gamma_0 T^{(0)} + \gamma_1 T^{(1)} + \gamma_2 T^{(2)} + \varepsilon. \tag{5.4}$$

The information matrix takes the value

$$X'X = \begin{pmatrix} \|T^{(0)}\|^2 & < T^{(0)}, T^{(1)} > & < T^{(0)}, T^{(2)} > \\ < T^{(1)}, T^{(0)} > & \|T^{(1)}\|^2 & < T^{(1)}, T^{(2)} > \\ < T^{(2)}, T^{(0)} > & < T^{(2)}, T^{(1)} > & \|T^{(2)}\|^2 \end{pmatrix}.$$

This is the Gram matrix (matrix of scalar products) of $T^{(0)}$; $T^{(1)}$; $T^{(2)}$. By symmetry $< T^{(0)}, T^{(1)} >$ and $< T^{(1)}, T^{(2)} >$ are zero. On the other hand $T^{(2)}$ has been chosen so that $< T^{(2)}, \mathbb{1} >= 0$, giving the fact that the information matrix is diagonal and thus the model orthogonal.

We have in Model (5.4):

$$\widehat{\gamma}_i = \frac{< T^{(i)}, Y >}{\|T^{(i)}\|^2}.$$

3 Orthogonality for non-regular models.

Definition 5.4 *Let us consider a partition is a non-regular linear model:*

$$Y = X_1\beta_1 + \cdots + X_m\beta_m + \varepsilon.$$

Consider a system of constraints $C_1\theta_1 = 0, \cdots, C_m\theta_m = 0$ that make the model identifiable. We say that these constraints make the partition orthogonal if the vectorial sub-spaces

$$V_i = \{X_i\beta_i : \beta_i \in Ker(C_i)\} \quad , \quad i = 1, \cdots, m$$

are orthogonal.

This definition makes perfect sense with the exemple of two-ways analysis of variance.

Proposition 5.4 *Consider the two-ways analysis of variance model:*

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{i,j,k}; \quad i = 1, \cdots, I, \ j = 1, \cdots, J, \ k = 1, \cdots, n_{ij},$$

with $n_{ij} \geq 1$, $\sum_{ij} n_{ij} = n > IJ$. There exists a system of constraints making the partition orthogonal if and only if

$$n_{ij} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{5.5}$$

where for example, $n_{i+} = \sum_i n_{ij}$ and $n_{++} = \sum_j \sum_i n_{ij}$.

In such a case the constraints are

$$(i) \sum_i^I \alpha_i n_{i+} = 0 \quad \text{and} \quad (ii) \sum_j^J \beta_j n_{+j} = 0;$$

$$(iii)\ \forall i = 1, \cdots, I, \ \sum_j^J n_{ij}\gamma_{ij} = 0 \quad \textit{and} \quad (iv)\ \forall j = 1, \cdots, J, \ \sum_i^I n_{ij}\gamma_{ij} = 0.$$

Remarks : 1/ A result very similar holds true for the additive model.
2/ Note that if we use the type III decomposition, we implicitly use the "non weighted system of constraints"

$$\sum_i \alpha_i = 0; \ \sum_j \beta_j = 0; \ \forall i, \ \sum_j \gamma_{i,j} = 0 \text{ et } \forall j, \ \sum_i \gamma_{i,j} = 0.$$

With this system, orthogonality demands equi-repetition $n_{ij} = const.$

Proof : a) Clearly the orthogonality of μ and α is equivalent to (i). The one of μ and β is equivalent to (ii) .
The orthogonality of the space generated by μ and α with the space generated by γ is equivalent to (iv).

b) it remains to study the orthogonality of $[\alpha]$, the space generated by α with $[\beta]$. Let us define the two spaces:

$$\begin{aligned} A &:= \{(\alpha_1, \cdots, \alpha_I) \in \mathbb{R}^I : \sum_i \alpha_i n_{i+} = 0\}, \\ \text{and symmetrically} \quad B &:= \{(\beta_1, \cdots, \beta_J) \in \mathbb{R}^J : \sum_i \beta_j n_{+j} = 0\}. \end{aligned}$$

Let $V^{(\alpha)}$ and $V^{(\beta)}$ be two vectors in $[\alpha]$ and $[\beta]$ respectively, we have

$$V^{(\alpha)}_{ijk} = \alpha_i \text{ with } \alpha = (\alpha_i)_i \in A,$$

$$V^{(\beta)}_{ijk} = \beta_j \text{ with } \beta = (\beta_j)_j \in B.$$

As a consequence if (5.5) holds true

$$< V^{(\alpha)}, V^{(\beta)} >= \sum_{ij} n_{ij}\alpha_i\beta_j = \sum_{ij} \frac{n_{i+}\alpha_i n_{+j}\beta_j}{n_{++}} = 0.$$

c) Reciprocal if $[\alpha]$ and $[\beta]$ are orthogonal, for all $\alpha = (\alpha_i)_i$ in A and $\beta = (\beta_j)_j$ in B we have

$$\sum_{ij} n_{ij}\alpha_i\beta_j = 0. \tag{5.6}$$

let us fix α. As (5.6) holds true for all β and that the only relation satisfied by the β_j is $\sum_j \beta_j n_{+j} = 0$, it implies that $(\sum_i n_{ij}\alpha_i)$ is proportional to n_{+j} as a function of j.

Summing in j , we see that the the coefficient of proportionality is necessarily zero. We have proved that for every vector α in A :

$$\sum_{ij} n_{ij}\alpha_i = 0, \quad \text{for all } j = i, \cdots , J.$$

Again it implies that the vector n_{ij} is proportional to n_{i+}. Let C_j be the coefficient of proportionality : $n_{ij} = C_j n_{i+}$, Summing in j

$$C_j n_{++} = n_{+j}.$$

Plugging this formula in the preceding gives the result. ■

4 Exercices

Exercise 5.1

* Show the equivalence :

$$X \text{ full-ranked } \Leftrightarrow X' \cdot X \quad \text{invertible.}$$

Exercise 5.2

* Find the result for additive model corresponding to Proposition 5.4
Same question pour nested model.

Chapter 6

Asymptotic properties

This chapter studies the behavior of the estimators and the test statistics in a linear model with an increasing size. It shows that, under weak conditions, the estimators are constant, asymptotically normal and that the T and F statistics converge to the asymptotic χ^2 test. (Coming soon !)

Chapter 7

Asymptotic selection of models for regression

1 Introduction

In many situations one uses several variables in a regression model as a precaution. Some of the k variables are relevant and some not. Fitting the whole model with say k variables seems to make sense but it often leads to disagreeable phenomenum : the over-fitting in the sense that the estimated coefficients follow the data and thus the errors. They don't follow the model. A good example is given by Figure 1 where a smooth signal is observed with some noise and is estimated by a piecewise constant model with 42 sub-intervals.

To avoid this kind of phenomenon, we must introduce the parsimony principle: to estimate too many parameters leads to an inflation of variance and a poor performance of the estimation of the response. Thus is is often better to set to zero the coefficients of some explanatory variables that seem to have a small or non-significative influence.

More precisely, we consider a regression model with k regressors and n observations.

$$Y_i = \sum_{j=1,..k} \beta_j Z_i^{(j)} + \epsilon_i, i = 1, ..., n$$

We will assume in most of the parts of this paper that $n > k$ and that the model is regular.

The whole model, the true model, the over-models and the false models

The model with all regressors will be called the "whole model". It will be denoted by $\bar{m}$. Among the coefficients $\beta_1, \dots, \beta_k$ of the whole model, some may be zero and the corresponding variables are not needed. They will be called the superfluous variables. The goal of choice of model is to identify these variables and consequently the true model that consists of all variables with non-zero beta. This model will be denoted m^*; to identify m^* instead of $\bar{m}$ permits to avoid the over-fitting.

The identification can be false in two direction

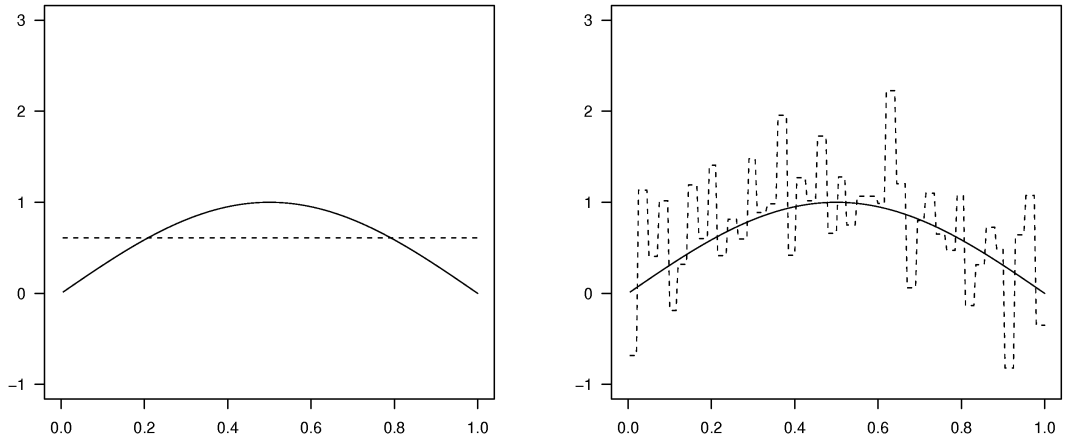


Figure 7.1: Over-fitting with a piecewise constant estimation over 42 sub-intervals.

- we can chose an "over-model": it is a model m that strictly contains m^*. As a consequence in contains some superfluous variables.
- we can chose a "false model" that does not contain some variables of m^*. For us a false model may contain some superfluous variables, it does not matter.

As an example if $(\bar{m}) = \{1, 2, 3, 4, 5\}$ and $m^* = \{1, 4, 5\}$

- $\{1, 3, 4, 5\}$ is an over-model
- $\{1, 5\}$ and $\{1, 3, 5\}$are two false-models. The first one is a sub-model (of the true one) but it does not matter.

We will consider the "all sub-set regression" in the sense that will search the true model among the set $\mathcal{M}$ of the 2^k sub-models of $\bar{m}$. Some exception to that case are

- nested model, for example in polynomial regression : the jth regressor $Z_i^{(j)}, i = 1, ..., n$ is a power of the second regressor $Z_i^{(2)}$ (The first one is the constant) and we want to chose the degree of the polynomial. There are k sub-models only of $\bar{m}$ to consider.
- model with an intercept. In almost all the cases, the first regressor is the "all-one vector" $\mathbb{1}_{\rm n}$ and in many case one does not want to question the presence of this vector in the model. In that case the set $\mathcal{M}$ of models to be considered is of size 2^{k-1}. This case is very similar to all sub-set regression, so we will omit the details.

Elementary methods

Test or thresholding.

Tests: Let two models m_1 and m_2 of the set $\mathcal{M}$ of considered models. Let α a level that may depend on n. When m_1 and m_2 are nested, one method is to perform a classical α F test between them. This method leads to two problems; first the number of tests to perform is very large $(3^k - 2^k)$ and second it may be non-consistent in the sense that m_1 can be chosen preferable to m_2 and m_2 to m_3 while m_3 is chosen preferable to m_1

Thresholding: One very crude method is to adjust the whole model and perform an α T-test of each of the k variables and keep the significative ones. This certainly makes sense if the model (or equivently the regressors) is orthogonal. In the other case it can lead to strange decisions. For example if $Z^{(1)}$ and $Z^{(2)}$ are very collinear and very collinear to Y, the thresholding method will discard both variables because, when $Z^{(1)}$ is present, $Z^{(2)}$ is no longer needed and vice versa. Nevertheless we will be able to prove some properties of this method.

Backward regression: to avoid the problem encountered in the example above, the backward selection method starts with the whole model and then

- at each step, the least significant variable is removed from the model and calculations are made anew.

-this is done while the variable to be removed is non-significant at a α level. If the variable is significant, of course it is kept, the procedure stops and the model is chosen .

Stepwise regression is a variant of the preceding where at every step we may add or remove a variable. We skip the details.

Forward regression is exactly the contrary of backward regression: we start with the empty model or the model with the sole constant and we add at each step the most significative variable. We end when the variable to be add is non significative at α level. The forward regression which is also called "L^2 boosting " can be applied in the case $k > n$.

PRESS or cross-validation

Let $m \subset \bar{m}$ and let us consider the associated regression model

$$Y_i = \sum_{j \in m} \beta_j Z_i^{(j)} + \epsilon_i,$$

We denote by X_m the design matrix (the matrix of the linear model) associated to model $m \in \mathcal{M}$. We want to estimate the quadratic error of model m

$$\mathbb{E}\left(\|X_{m^*}\beta - X_m\widehat{\beta_m}\|^2\right).$$

A way of estimating this is a "leave-one-out cross-validation".

For $i = 1, \dots, n$ we, define Y^{-i} the vector obtained be removing the ith observation. For $m \in \mathcal{M}$ we define μ_m^{-i} as the scalar which is the prediction
based on Y^{-i} and on Model m
and taken at the point i.

We define the PRESS as

$$\text{PRESS}\ _m = \sum_{i=1}^{n} (Y_i - \mu_m^{-i})^2. \tag{7.1}$$

Note that one nice property of the PRESS is that the two random variables Y_i and μ_m^{-i} are independent.
The chosen model is the one with the lowest PRESS.

In general, selection with PRESS is computationally very expensive and has properties equivalent to C_p or AIC (see above) see [13] . In particular, as we will see, it tends to over-estimate the size of the model. For linear model the situation is nicer since we can compute a simplified form: it can be proved that

$$\text{PRESS}_m = \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i^m)^2}{(1-h_i)^2},$$

where $h_i = X_i'(X'X)^{-1}X_i$, X_i being the ith row of X see, for example, [15] p. 252 for a proof.

To remedy to this drawback one can use "leave-p-out cross-validation, but in this case the computational cost is even larger. A less costly alternative in v-fold cross-validation , see [3] for a detailed study .

2 Methods based on L_0 penalties

As a general principle, the likelihood method choses always the largest model and this is true for our regression model. Note that, as we will see later , maximizing the likelihood is equivalent, for linear model to minimize the sum of square. Heuristic considerations (based on Kullback information or on Bayesian models) have lead to use the following penalized likelihood criterions [1]

$$\begin{aligned} AIC &= -2\log(\text{maximized likelihood}) + 2|m| \\ BIC &= -2\log(\text{maximized likelihood}) + \log(n)|m| \end{aligned}$$

where

- the maximized likelihood is the maximum of the likelihood
- the likelihood is computed on n independent observations.
- The penalty $2|m|$ or $\log(n)|m|$ favors small models($|m|$ is the size of m).
- the criterions have to be minimized.

These criterion can be extended as

$$GIC = -2\log(\text{maximized likelihood}) + c(n)|m| \tag{7.2}$$

where $c(n)$ is a function to be fixed later.

We can define also

$$AIC_c = -2\log(\text{maximized likelihood}) + n\frac{n+|m|+1}{n-|m|-3}$$

and

$$C_p(m) = \frac{SS(m)}{\widehat{\sigma_{\overline{m}}}} + 2|m|$$

Lemma 7.1 *In the linear model*

$$Y = X\beta + \epsilon$$

we have

- *the maximum likelihood estimator of σ^2 is $\widehat{\sigma^2} = 1/nSS$ where SS is the sum of squares*
$$SS = \|Y - X\widehat{\beta}\|^2.$$
-
$$-2\log(\text{ maximized likelihood}) = n\log(\widehat{\sigma^2}) + SS/\widehat{\sigma^2}$$
-
$$-2\log(\text{ maximized likelihood}) = n\log(\widehat{\sigma^2}) + n$$
-
$$-2\log(\text{ maximized likelihood}) = n\log(SS/n) + n = n\log(SS) + (const).$$

The proof is omitted, each result being an easy consequence of the preceding one. Note that the constants (*const*) appearing in the formulas above play no role and can be omitted.

All the criterions as PRESS AIC BIC GIC permit a rather easy comparison of models **but** , if we perform a "all subset selection", the number of sub-model to compare is 2^k which is soon very large. Some "leaps and bounds algorithm" exist that permit to avoid to examine all the possibilities but practical limitations are about $k = 30$. In the other cases only a partial exploration is performed by a stepwise algorithm. This works rather well in practice but no theoretical results are known in that case. For these reasons for large size, one often prefers L^1 penalties as LASSO.

The criterion PRESS AIC BIC GIC permit to consider the "large dimension case" : $k > n$ only if we limit us to models of size $|m| < S$ with S much smaller than n . Such models are called sparse models. But in this case again computational problems are heavy.

3 Comparison of models with AIC, GIC

This section is devoted to the study of the relations of AIC (BIC, GIC) with tests.

Assumption 7.1 *Though we will use Gaussian likelihood to estimate and compute the criterion, we will work under one of the two following hypotheses when n tend to infinity and when k may depend on n but must satisfy $k_n = o(n)$*

- *the Gaussian case : the $\epsilon_i, i = 1, \dots, n$ of the linear model errors are independent with distribution $N(0, \sigma^2)$, the variance σ^2 is of course unknown.*
- *the errors are centered independent with the same symmetric distribution and finite variance σ^2 and finite order four moment. We assume in addition the Huber condition : H^n the maximal diagonal element of the "hat matrix" $X(X'X)^{-1}X'$ tends to zero (X is the matrix associated to the whole model).*

3.1 AIC

Lemma 7.2 *Suppose that m_1 and m_2 are two nested models $m_1 \subset m_2$, the model m_1 is preferable to m_2 for AIC iff*

$$\widehat{F}_{m_2/m_1} = \frac{(SS(m_1) - SS(m_2))/(|m_2| - |m_1|)}{SS(m_2)/(n - |m_2|} > \frac{n - |m_2|}{|m_2| - |m_1|} \left(\exp\left(2 \frac{(|m_2| - |m_1|)}{n}\right) - 1 \right) \tag{7.3}$$

Again the proof can be omitted. The reader familiar with linear model has recognized in the left-hand-size of (7.3) the statistics of the Fisher test. In other words, AIC performs a Fisher test, but with a different critical value. We have obviously the same kind of result for GIC replacing the 2 by $c(n)$.

Suppose now that the number n tends to infinity that m_1 is the true or an over-model and that Assumption 1 is satisfied, then using law or large number and Central limit theorem under Lindeberg condition (see for example Th 8.3 of Azaïs and Bardet) (since we are under the null hypothesis) the limit distribution of $\widehat{F}_{m_2/m_1}$ is

$$\widehat{F}_{m_2/m_1} \Rightarrow \chi^2(p)/p$$

where $\Rightarrow$ is the convergence in distribution. Obviously the right-hand side of (7.3) converges to 2.

As a consequence AIC performs asymptotically a F test with critical value $2p$ where p is the difference of degrees of freedom between the two hypotheses.
This corresponds to the following levels.

difference p	level
1	0.104
2	0.068
3	0.049
4	0.037
5	0.028

As a consequence, considering the case where m_1 is the true model, the result above shows that AIC has a probability that tends to a positive limit to prefer every over-model m_2 to the true model. So the probability of choosing m^* cannot tend to 1.

3.2 BIC, GIC

If we consider the GIC as defined by (7.2). The calculation above shows that the criterion will prefer model m_1 to m_2 iff

$$\widehat{F}_{m_2/m_1} = \frac{(SS(m_1) - SS(m_2))/(|m_2| - |m_1|)}{SS(m_2)/(n - |m_2|)} > \frac{n - |m_2|}{|m_2| - |m_1|}\left(\exp(c(n)\frac{(|m_2| - |m_1|)}{n}) - 1 \right)$$

Now we assume that $c(n) \to +\infty$, $c(n) = o(n)$ to get that the right hand side is equivalent to $c(n)$.

As a consequence the probability of preferring a given over-model is, for n suffiently large, smaller that the probability of a $\xi^2(d)$ distribution to be smaller that K for every K so it tends to zero.

Since k is assumed to be fixed, the number of over-models is bounded and we obtain immediately that the probability of "preferring an over-model to m^*" tends to zero.

3.3 Case of a false model

Suppose that m is "false". It is not in general a sub-model of the true model. But it can be compared to the model $m_2 = m \cup m*$. Since $m \subset m_2$ we can appy Lemma 7.2 that shows that asymptotically m_2 is prefered to m if $\hat{F}_{m_2/m} \geq \widetilde{c}(n)$ where $\widetilde{c}(n) \simeq c(n)$. Using the same arguments of chapter 8 of Azaïs and Bardet (2005), we see that under our hypotheses the denominator $D = \widehat{\sigma}^2$ of $\hat{F}_{m_2,m}$ tends in probability to σ^2 while the numerator can be written as

$$N = \big(\|P_V(X\beta) + P_V X\widehat{\beta}\|^2\big)/d$$

where V is the orthogonal of $[X_m]$ in $[X_{m_2}]$. Using the normality of $\widehat{\beta}$ (Th 8.2 of the same book) we see that what ever the non-centrality parameter $\|P_V(X\beta)$ is, the numerator can be written as

$$D = 1/d\|\|P_V(X\beta) + Z\|^2$$

where Z has for variance-covariance matrix

$$P_V X(X'X)^{-1}X'P_V = P_V.$$

Using a rotation argument, this matrix can be transformed, for example, into I_d where I_d is the identity of size d and N can be written as the norm a vector in a space of dimension d as

$$N = \frac{\sigma^2}{d}\|\xi + W_n\|^2$$

where ξ converges to the $N(0, I_d)$ distribution and

$$\|W_n\|^2 = \frac{1}{\sigma^2}\|P_V X\beta\|^2 = \frac{1}{\sigma^2}\|P_{m^\perp} X\beta\|^2.$$

This last parameter will be called the non-centrality parameter and denoted by NC_m:

$$NC_m = \frac{\|P_{m^\perp}X\beta\|^2}{\sigma^2}$$

Note that this parameter depends additionally on n but we omit that in the notation.

We set now our uniform born on χ^2 distributions

Lemma 7.3 - *For all integer $d \geq 1$ and for all real $c(n)$ greater than* 2

$$\mathbb{P}\{\chi^2(d) \geq c(n)d\} \leq \exp(-c(n)/2)$$

- *If $NC > 4c(n)d$ then*

$$\mathbb{P}\{\chi'^2(d, NC) \leq c(n)d\} \leq \exp -(\frac{NC}{8d})$$

Proof: The first part is easy to obtain by an exponential inequality or by exact computation using integration by parts. Let χ'^2 be a variable with distribution $\chi'^2(d, NC)$. This variable has the representation

$$\chi'^2 = \|\sqrt{NC}e_1 + Z\|^2,$$

where e_1 is the first vector of the basis and Z is standard normal in $\mathbb{R}^d$. Let $\chi^2 = \|Z\|^2$. Denoting $c(n)$ by c for short, we have

$$\mathbb{P}\{\chi'^2 < cd\} = \mathbb{P}\{\chi' < \sqrt{cd}\} \leq \mathbb{P}\{\chi > \sqrt{NC} - \sqrt{cd}\} \leq \mathbb{P}\{\chi > 1/2\sqrt{NC}\} = \mathbb{P}\{\chi^2 > 1/4NC\}.$$

It suffices to use the first relation. ∎

We turn now to the main results. Suppose that the parameter $c(n)$ satisfies $1 << c(n) << n$ and that every false model m has a non-centrality parameter that satifies $c(n) << NC_m$, then

(i) Suppose now that m is a false model then using the convergence in probability to σ^2 of the denominator of $\hat{F}_{m_2,m}$ we obtain that

$$\begin{aligned}\mathbb{P}\{m \text{ prefered to } m_2\} &= \mathbb{P}\{\|\xi + W_n\|^2 \leq dC(n)(1 + o_p(1))\} \\ &\leq \mathbb{P}\{\|\xi\| \geq d\big(\sqrt{NC_m} - \sqrt{c(n)(1 + o_p(1)}\big)\} \\ &= \mathbb{P}\{\|\xi\| \geq d\big(\sqrt{NC_m}(1 + o_p(1))\big)\},\end{aligned}$$

where W and ξ are defined as above. The convergence in distribution of ξ implies that this probability tends to zero.

As a consequence mixing this case with the case of over-models, we have proven that GIC chooses the true model with a probability that converges to 1.

(ii) Suppose now in addition that the model is Gaussian,then it is possible to give an exponential bounds to the probability of a false model m to be preferred to $m_2 = m \cup m^*$.

The false model m is prefered to $m_2 = m \cup m^*$ if $\hat{F}_{m_2,m} \leq \widetilde{C}_n$. Firstly We can choose n sufficiently large so that $\widetilde{C}_n \leq 2C(n)$, secondly we have

$$\hat{F}_{m_2,m} \overset{D}{=} \frac{\chi'^2(d, NC_m)/d}{\chi^2(n-|m_2|)/(n-|m_2|)}$$

Using large deviation inequality (in fact just the easier part), except with an exponentially small (as a function of n) probability,

$$\chi^2(n-|m_2|) \leq (2(n-|m_2|))$$

so that it suffiices to give bound to

$$\mathbb{P}\{\chi'^2(d, NC_m)/d \leq 4C(n)\}$$

and this by Lemma 7.3 is smaller than $\exp\big(-\frac{NC_m}{8d}\big)$ so we have proved that

$$\mathbb{P}\{m \text{ prefered to } m_2\} \leq \exp(-((const)n) + \exp\big(-\frac{NC_m}{8d}\big).$$

A first example of application is the very simple case where k is fixed and the matrix X associated to the whole model satisfies

$$1/n\ X'X \to M \tag{7.4}$$

where M is some definite positive matrix. In that case the computation below proves that for every false model m

$$NC_m \simeq \gamma_m n$$

with $\gamma_m > 0$.

Computation of NC

Indeed by the Pythagore Theorem

$$NC_m = \|P_{m^\perp} X\beta\|^2 = \|X\beta\|^2 - \|P_m X\beta\|^2$$

We study the two terms separately. Because of our hypothesis

$$\|X\beta\|^2 \simeq n\beta' M\beta.$$

For the seond term

$$\|P_m X\beta\|^2 = \beta' X' X_m (X'_m X_m)^{-1} X'_m X\beta \simeq n\beta' M_{\overline{m},m} M^{-1}_{m,m} M_{m,\overline{m}}\beta$$

where M_{m_1,m_2} is the extraction of the matrix M choosing m_1 for the lines and m_2 for the columns. So that

$$\|P_{m^\perp} X\beta\|^2 \simeq n\beta'(M - M_{\overline{m},m} M^{-1}_{m,m} M_{m,\overline{m}})\beta$$

M can bee seen as the Gram matrix (the matrix of norms and scalar products) of some set of k vectors in $\mathbb{R}^k$, say $V_1, \dots, V_k$ (the choice is up to a rotation). Since M is non singular these vectors are not collinear. A classical linear algebra calculation shows that

$$M - M_{\overline{m},m} M_{m,m}^{-1} M_{m,\overline{m}}$$

is the matrix of the quadratic form that associate to the vector $b \in \mathbb{R}^k$ the quantity

$$\|\Pi_{m^\perp} \sum_{i=1}^{k} b_i V_i\|^2,$$

where $\Pi_{m^\perp}$ is the projector on the orthogonal of the space S_m generated by the vector that are in m. Since m is a false model β has some coordinates that does not belong to m and because of the linear independence of the vectors, $\sum \beta_i V_i$ does not belong to S_m. and we obtain the result. ∎

Note that

- the condition (7.4) is met for example if the regressors are draw from i.i.d. replicates of some random distribution with a second order moment and non-degenerate variance matrix. This is a direct consequence of the law of large numbers.
- under this condition, it is an exercise, to check that the thresholding method and the backward method find the true model with a probability that tends to 1, as soon as the tests are conducted at a level α_n that tends to zero sufficiently slowly.
- The result can be generalized to the case where some normalization $d(n)$ of the information matrix exists such that

  $$1/d(n)\ X'X \to M$$

  In such a case $c(n)$ must be negligible with respect to $d(n)$.
- When $k = k_n$ tends to infinity, we cannot hope to have a property like (7.4) but our result still prove that if $k(n) = o(c(n))$, GIC will chose and over-model with a probability that tends to zero.

4 Asymptotic oracle inequality

In this section we assume normality. Consider the quadratic risk of estimation and we still assume a) condition (7.4) with M regular b) that k fixed and c) that $1 << c(n) << n$. Let $\widehat{m}$ the model chosen by GIC (with probability 1 it is unique). We define the risk of estimation

$$R_n = \mathbb{E}\left(\|\widehat{Y}_{\widehat{m}} - X\beta\|^2\right).$$

This risk can be partitionned into the risks relative to the choice of a particular model

$$R_n = \sum_{m\in\mathcal{M}} R_n(m) := \sum_{m\in\mathcal{M}} \mathbb{E}\left(\|\widehat{Y}_m - X\beta\|^2 1\!\!\mathrm{I}_{\widehat{\mathrm{m}}=\mathrm{m}}\right)$$

Using the decomposition bias, variance, a short computation shows that
if Z is some random variable that can be written

$$Z = \mathbb{E}\left(Z\right) + \epsilon = \mu + \epsilon$$

with ϵ symmetric and E an event that may depend on ϵ but whose distribution is invariant by change of sign of ϵ, then

$$\mathbb{E}\left(Z 1\!\!\mathrm{I}_{\mathrm{E}}\right)^2 = \mu^2\mathbb{P}\left(\mathrm{E}\right) + \mathrm{Var}\left(\epsilon 1\!\!\mathrm{I}_{\mathrm{E}}\right) + 2\mu\mathbb{E}\left(\epsilon 1\!\!\mathrm{I}_{\mathrm{E}}\right) = \mu^2\mathbb{P}\left(\mathrm{E}\right) + \mathrm{Var}\left(\epsilon 1\!\!\mathrm{I}_{\mathrm{E}}\right).$$

Remarking that a change of sign of the errors does not modify the choice of model and using the assumed symmetry of the errors we get

$$\mathbb{E}\left(\|\widehat{Y}_m - X\beta\|^2 1\!\!\mathrm{I}_{\widehat{\mathrm{m}}=\mathrm{m}}\right) = \|\mathrm{P}_{\mathrm{m}^\perp}(\mathrm{X}\beta)\|^2\mathbb{P}\left(\widehat{\mathrm{m}}=\mathrm{m}\right) + \mathbb{E}\left(\|\mathrm{P}_{\mathrm{m}}\epsilon\|^2 1\!\!\mathrm{I}_{\widehat{\mathrm{m}}=\mathrm{m}}\right) = \mathrm{J}_{1,\mathrm{m}} + \mathrm{J}_{2,\mathrm{m}}$$

Then every false model m satisfies

$$\|P_{m^\perp}(X\beta)\|^2 \simeq \gamma_m n \qquad \text{with } \gamma_m > 0.$$

Thus by Lemma 7.3 , for n sufficiently large$NC_m \simeq \frac{\gamma_m n}{\sigma^2} > 4c(n)$:

$$\mathbb{P}\left(\widehat{m} = m\right) \le \exp -(\frac{\gamma_m n}{8d_m}),$$

Where d_m is the number of missing variables in m: $d_m = |m \cup m^*| - |m|$. For over-models the quantity $\|P_{m\perp}(X\beta)\|^2$ vanishes. This implies that

$$\sum_{m\in\mathcal{M}} J_{1,m} \to 0.$$

For the other terms we use the Schwarz inequality

$$\mathbb{E}\left(\|P_m\epsilon\|^2 1\!\!\mathrm{I}_{\widehat{\mathrm{m}}=\mathrm{m}}\right) \le \left(\mathbb{E}\,\|\mathrm{P}_{\mathrm{m}}\epsilon\|^4\mathbb{P}\left(\mathrm{m}\neq\mathrm{m}^*\right)\right)^{1/2}.$$

Let us compute the quantity $\mathbb{E}\,\|P_m\epsilon\|^4$. Let P_{ij} denote the entry i,j of P_m

$$\begin{aligned}\mathbb{E}\,\|P_m\epsilon\|^4 &= \sum_{iji'j'} \mathbb{E}\left(\epsilon_i P_{ij}\epsilon_j\epsilon_{i'}P_{i'j'}\epsilon_{j'}\right)\\ &= \sum_{iji'j'} P_{ij}P_{i'j'}\mathbb{E}\left(\epsilon_i\epsilon_j\epsilon_{i'}\epsilon_{j'}\right).\end{aligned}$$

Because of independance, the last expectation vanishes except if the four indices are pairwise equals. It remains three cases to consider

- $i = i' = j = j'$ which contribution is $m_4 \sum_i P_{ii}^2$ where m_4 is the order 4 moment of the errors.
- $i = j \neq i' = j'$ which contribution is $\sigma^4 \sum_{i \neq i'} P_{ii} P P_{i'i'}$
- $i = i' \neq j = j'$ or $i = j' \neq j = i'$ which contribution is bounded by $2\sigma^4 \sum_{i \neq j} P_{ij}^2$.

Since

$$\sum_i P_{ii}^2 + \sum_{i \neq j} P_{ij}^2 = tr(P_m^2) = tr(P_m) = |m|$$

$$\sum_{i \neq i'} P_{ii} P P_{i'i'} + \sum_i P_{ii}^2 = (tr(P_m))^2 = |m|^2,$$

it is easy to see that $\mathbb{E}\, \|P_m \epsilon\|^4$ is bounded. Note that in the Gaussian case it

is the expectation of square of a $\chi^2(|m|)$ variable which can be easily computed to be $\sigma^4(|m|^2 + 2|m|)$. Finally

$$\sum_{m \in \mathcal{M}, m \neq m^*} J_{2,m} \to 0`$$

Finally we have proven that

$$R_n = \mathbb{E}\left(\|\widehat{Y}_{\widehat{m}} - X\beta\|^2\right) \to |m^*|$$

The risk we have if we know the true model. The risk with a choice of model by GIC is asymptotically the same than the risk when the oracle tell us which is the true model. Such an inequality is called an Oracle inequality

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[2] T. Amemiya. *Advanced Econometrics.* Harvard University Press, Cambridge, 1985.

[3] S. Arlot. *Rééchantillonnage et sélection de modèles.* PhD thesis, Université Paris Sud-Paris XI, 2007.

[4] P. J. Bickel and K. A. Doksum. Mathematical statistics, volume i, 2001.

[5] P. Calas, T. Rochd, P. Druilhet, and J.-M. Azas. In vitro adhesion of two strains of prevotella nigrescens to the dentin of the root cana:the part played by different irrigations solutions. *Journal of Endodontics*, 24(2):112–115, 1998.

[6] J. Coursol. *Techniques statistiques des modles linaires.* cimpa, Nice, 1981.

[7] D. Dacunha-Castelle and M. Duflo. *Probabilités et statistiques. Tome 2.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Problèmes à temps mobile. [Movable-time problems].

[8] W. Green. *Econometrics analysis.* Prentice Hall International Editions, fourth edition, 2000.

[9] X. Guyon. *Modle linaire et conomtrie.* Ellipse, Paris, 2001.

[10] J. Jobson. *Applied Multivariate Data Analysis.* Springer-Verlag, New York, 1991. Springer Series in Statistics.

[11] J.-M. Lecoutre and P. Tassi. *Statistique non paramtrique et robustesse.* Economica, Paris, 1987.

[12] H. Levene. Robust tests for equality of variances. In *Contributions to probability and statistics*, pages 278–292. Stanford Univ. Press, Stanford, Calif., 1960.

[13] K.-C. Li. Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, pages 958–975, 1987.

[14] P. McCullagh and J. Nelder. *Generalized linear models.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983.

[15] A. McQuarrie and C. Tsai. *Regression and time series model selection.* World Scientific Publishing Co. Inc., River Edge, NJ, 1998.

[16] J. Miller and G. Rupert. *Simultaneous statistical inference.* Springer-Verlag, New York, second edition, 1981. Springer Series in Statistics.

[17] G. Milliken and D. Johnson. *Analysis of Messy Data vol. 1 : Designed Experiments.* Van Nostrand Reinhold, New-York, 1984.

[18] S. Searle. *Linear models for unbalanced data.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987.

[19] J. Tanner. *The Physique of the Olympic Athlete.* George Allen and Unwin, London, 1964.

[20] R. Tomassone, S. Audrain, E. Lesquoy, and C. Miller. *La rgression, nouveaux regards sur une ancienne mthode statistique.* Masson, Paris, 1992.

[21] A. van der Vaart. *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

Contents