Le modèle linéaire
exemples traités avec Splus

Aline Julien
Patrick Gasqui
Claire Chabanet

Biométrie — Jouy en Josas
4/5/95

SOMMAIRE

La plupart des données utilisées sont extraites du module “Modèle Linéaire” de la Formation Permanente à la statistique FPSTAT; d’autres sont le résultat d’expérimentations INRA. Toutes sont traitées avec Splus3.1.
Les chapitres traités sont les suivants. Ils sont indépendants les uns des autres.

LA REGRESSION LINEAIRE SIMPLE

L’ANALYSE DE VARIANCE A UN FACTEUR

L’ANALYSE DE VARIANCE A DEUX FACTEURS

L’ANALYSE DE VARIANCE A DEUX FACTEURS : PLAN DESEQUILIBRE

L’ANALYSE DE COVARIANCE

COMPARAISON DE DROITES DE REGRESSION

LA REGRESSION LINEAIRE MULTIPLE

LA REGRESSION LINEAIRE MULTIPLE : CHOIX DE REGRESSEURS

LA REGRESSION LINEAIRE SIMPLE

Résumé des principales commandes :

```
> reg_lm(tension~age,data=arter)
> summary(reg)
> plot(reg)
```

I Le problème

1) Les données

On étudie la liaison entre l'âge et la tension artérielle. On choisit a priori la population étudiée, une population de femmes donnée, et 5 âges compris entre 35 et 75 ans. On tire aléatoirement une femme dans chaque sous-population correspondant à chaque âge, on observe sa tension artérielle.

Les données recueillies sont les suivantes :

AGE (X)	TENSION (Y)
35	114
45	124
55	143
65	158
75	166

2) Choix de l'analyse

La tension artérielle est une variable mesurée, c'est une variable aléatoire. L'âge est une variable contrôlée, fixée, elle n'est pas aléatoire. On cherche à exprimer la tension artérielle, variable à expliquer, en fonction de l'âge, variable explicative. L'âge est la seule variable explicative, elle est quantitative, on met en œuvre une régression linéaire simple.

Cette méthode consiste à modéliser les données par une droite. Si la modélisation est adéquate, cela signifie que la relation entre la tension et l'âge est linéaire. On

pourra alors prévoir la tension artérielle d'une femme si on connaît son âge. Ces prévisions seront possibles "à l'intérieur du domaine d'observation", c'est-à-dire pour des personnes dont l'âge varie de 35 à 75 ans. En effet, rien ne permet de penser que la liaison est linéaire avant 35 ans et après 75 ans.

II Premiers traitements

1) Saisie

On stocke les données âge et tension sous forme de deux vecteurs :

```
> age_c(35,45,55,65,75)
> tension_c(114,124,143,158,166)
```

Il est souhaitable de stocker les données dans un objet unique, de type "data frame", pour utiliser l'une ou l'autre des fonctions de modélisation de Splus. Un data frame est une liste d'objets. Il peut contenir tous les objets nécessaires à la modélisation : la variable expliquée et les variables explicatives. Ici, le data frame s'appellera "arter".

```
> arter_data.frame(age,tension)
> arter
   age tension
1  35     114
2  45     124
3  55     143
4  65     158
5  75     166
```

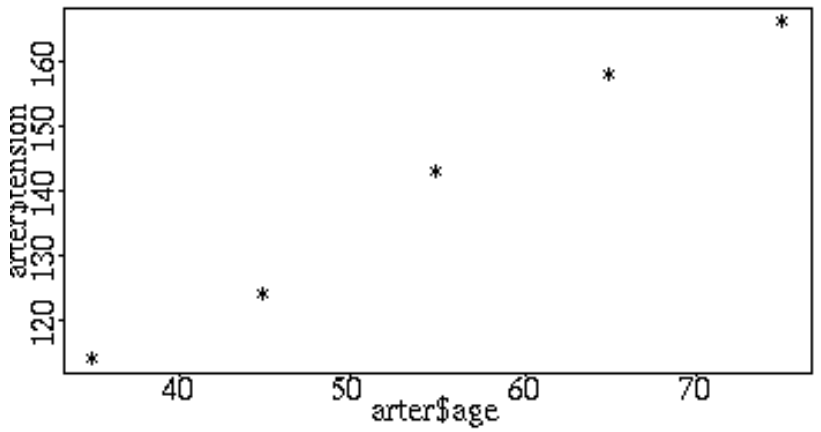
On peut accéder aux variables "age" et "tension" par extraction des objets "age" et "tension" du data frame "arter", on détruit donc les deux vecteurs "age" et "tension".

```
> arter$age
[1] 35 45 55 65 75
> arter$tension
[1] 114 124 143 158 166
> rm(age)
> rm(tension)
```

2) Visualisation graphique

On représente graphiquement les données sous forme d'un nuage de points.

```
> plot(arter$age,arter$tension)
```



La variable expliquée “tension” est portée en ordonnée, alors que “age”, la variable explicative, est placée en abscisse.

III Analyse : Régression

1) Le modèle et sa mise en œuvre

Le modèle étudié est le suivant : $Y = \alpha + \beta X + \varepsilon$

Y : tension artérielle,
X : âge,
α : ordonnée à l'origine de la droite cherchée,
β : pente de la droite cherchée,
ε : erreurs indépendantes, d'espérance nulle, telles que $var(\varepsilon) = \sigma^2$.

C'est la fonction **lm** (**l**inear **m**odel) qui permet de réaliser les calculs.

```
> reg_lm(tension~age,data=arter)
```

La formule “tension ~ age” décrit le modèle. Une formule s'écrit toujours de la manière suivante : var.expliquée ~ modèle. Ici le modèle est réduit à un terme, le terme “age”, car il n'y a qu'un régresseur. Le modèle “age” est équivalent au modèle “1+age”, où “1” décrit le terme constant et “age” le régresseur. Par défaut, le terme constant est inclus dans le modèle. Si l'on voulait écrire un modèle sans terme constant, il faudrait le dire explicitement en écrivant : “-1+age”.
L'équation de la droite ajustée sera donc : $\widehat{Y} = \widehat{\alpha} + \widehat{\beta} X$ et $\widehat{var(\varepsilon)} = \widehat{\sigma^2}$

2) Visualisation des résultats

a) résultats numériques

```
> summary(reg)
```

```
Call: lm(formula = tension ~ age, data = arter)
Residuals:
  1  2  3  4  5
 0.6 -3.2 2 3.2 -2.6

Coefficients:
            Value Std.Error t value Pr(>|t|)
(Intercept) 65.1000 5.8284  11.1695 0.0015
   age      1.3800  0.1026  13.4461 0.0009

Residual standard error: 3.246 on 3 degrees
 of freedom
Multiple R-Squared: 0.9837

F-statistic: 180.8 on 1 and 3 degrees of freedom,
the p-value is 0.0008894

Correlation of Coefficients:
    (Intercept)
age -0.9685
```

Call : le modèle.

Residuals : résidus associés à chaque donnée.

Coefficients :

- dans la colonne "value" on trouve les coefficients estimés de la droite de régression. "Intercept" représente l'ordonnée à l'origine, tandis que "age" représente la pente de la droite. On a donc :

$$\widehat{\beta} = 1,38 \qquad \widehat{\alpha} = 65,1$$

- dans la colonne "Std.Error", on trouve l'estimation de l'écart-type de chaque coefficient.
- la colonne "t value" contient les statistiques de Student du test des hypothèses nulles "α=0" sur la première ligne, et "β=0" sur la deuxième ligne. Ces

statistiques sont égales à $\frac{\hat{\alpha}}{\widehat{ET(\hat{\alpha})}}$ et $\frac{\hat{\beta}}{\widehat{ET(\hat{\beta})}}$ respectivement. Elles sont à comparer à la valeur critique lue dans la table de Student pour un nombre de degrés de liberté égal à 3 (=n-2), et un niveau de confiance donné. Pour un risque d'erreur de 5% (c'est-à-dire un niveau de confiance de 95%), la valeur critique vaut 3,182. Les statistiques de Student sont supérieures à cette valeur critique, ceci entraîne le rejet des hypothèses nulles.

- En fait, il est inutile de lire la valeur critique dans une table de Student : la colonne "Pr(>|t|)" (p-value) indique la probabilité d'obtenir une statistique supérieure en valeur absolue à la statistique calculée, si H_0 est vraie; elle s'interprète comme le risque d'erreur réel lorsque l'on rejette l'hypothèse nulle. Ici, les risques sont si faibles (<5%) que l'on peut rejeter l'hypothèse nulle dans les deux cas.
  On peut toujours tester l'hypothèse nulle "α=0" dans le but de simplifier le modèle. Mais on ne peut interpréter le résultat du test si ce test suppose une extrapolation en dehors du domaine d'observation. Ici, l'interprétation de ce test supposerait l'extrapolation de la droite à des âges compris entre 0 et 35 ans. Cette extrapolation est abusive, et une tension nulle à la naissance n'a pas de sens. On ne peut donc interpréter le test de nullité de l'ordonnée à l'origine, que dans le cas où 0 appartient au domaine d'observation de la variable explicative.

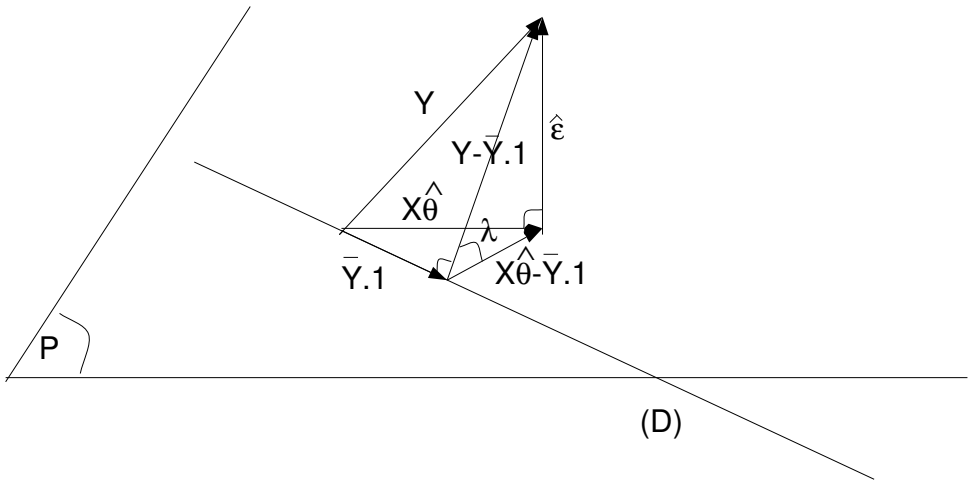
Residual standard error : estimation de l'écart-type résiduel $\widehat{\sigma} = 3,246$.

Multiple R-squared : coefficient de détermination R^2. C'est le résultat de la division de la variablité expliquée par le modèle par la variabilité totale de Y.

Le coefficient de détermination vaut 0,98 dans cet exemple, la régression explique donc 98% de la variabilité totale.
Le coefficient de détermination ne permet pas de valider le modèle. C'est un indicateur global de l'explication qu'apporte le modèle.

Une représentation géométrique permet de visualiser le vecteur des observations, le vecteur des valeurs ajustées et R^2.

On représente les observations par un vecteur Y de R^n. Soit P le plan engendré par les vecteurs colonne de X (le vecteur $\vec{1}$ et le vecteur correspondant aux valeurs prises par la variable explicative), soit D la droite engendrée par le vecteur $\vec{1}$, $\overline{Y}.\vec{1}$ la projection orthogonale de $\vec{Y}$ sur D, $X\vec{\hat{\Theta}}$ la projection orthogonale de Y sur P. On obtient alors :



$$\vec{Y} - \overline{Y}.\vec{1} = \left(X\vec{\hat{\Theta}} - \overline{Y}.\vec{1}\right) + \vec{\hat{\varepsilon}}$$
$$et \quad \vec{Y} = X\vec{\hat{\Theta}} + \vec{\hat{\varepsilon}}$$

Que l'on projette $\vec{Y}$ ou $\overline{Y}.\vec{1}$ sur P, on trouve le même vecteur de résidus. On dit que l'on projette $\vec{Y}$ par soucis de simplicité, en sachant que la table d'analyse de variance présente la projection de $\vec{Y} - \overline{Y}.\vec{1}$. On a alors :

$$R^2 = \frac{\sum\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{\|X\vec{\widehat{\Theta}} - \overline{Y}.\vec{1}\|^2}{\|\vec{Y} - \overline{Y}.\vec{1}\|^2} = corr^2\left(Y, X\hat{\Theta}\right)$$

Sur le graphique, on voit que :

$$\cos\lambda = \frac{\|X\vec{\widehat{\Theta}} - \overline{Y}.\vec{1}\|}{\|\vec{Y} - \overline{Y}.\vec{1}\|}$$

On a alors : $R^2 = \cos^2\lambda$ et donc $0 < R^2 < 1$.

Plus λ est grand, plus la variabilité expliquée par le modèle est petite devant la variabilité résiduelle, et plus R^2 est proche de 0. Le modèle explique bien la variabilité observée, si λ est petit, c'est-à-dire si R^2 est proche de 1.

F-statistic : test de la signification de l'explication apportée par le modèle (test de Ho : "β=0").

$$CMM = \frac{\sum \left(\widehat{Y_i} - \overline{Y}\right)^2}{1}$$

CMM : carré moyen expliqué par le modèle.

$$CMR = \frac{\sum \left(Y_i - \widehat{Y_i}\right)^2}{n-2}$$

CMR : carré moyen résiduel.
n : nombre d'observations.

$$F = \frac{CMM}{CMR}$$

Si les carrés moyens CMM et CMR sont du même ordre de grandeur, le modèle n'apporte aucune explication et il n'est pas intéressant. Pour comparer les ordres de grandeur des carrés moyens, on calcule la statistique F. Le modèle apporte une explication significative, si la statistique calculée est supérieure à la valeur critique lue dans la table de Fisher (pour un niveau de confiance donné, et des degrés de libertés égaux à 1 et 3). Ici F vaut 180,8 et est supérieur à 10.1(valeur lue dans la table de Fisher pour des degrés de libertés égaux à 1 et 3, et un niveau de confiance de 95%), le modèle apporte donc une explication significative.

Correlation of Coefficients : matrice de corrélation des coefficients.
Ici, α et β sont corrélés. Ils doivent donc être interprétés simultanément.

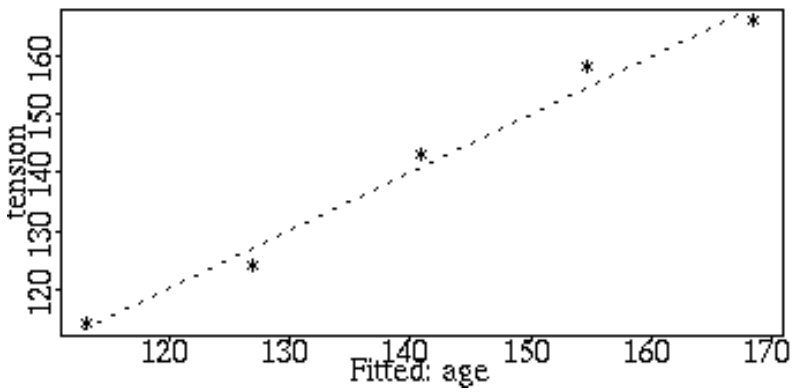
Il peut arriver que le modèle apporte une explication significative et qu'il ne soit pas valide. En particulier, le modèle n'est pas valide si les résidus présentent une structure qui remet en cause le caractère aléatoire de la distribution des erreurs. Il est donc nécessaire d'étudier le modèle avec plus de précautions, et en particulier de faire des représentations graphiques des résidus. En effet, les résidus sont les valeurs estimées des erreurs, et les erreurs doivent vérifier les postulats : espérance nulle, variance constante, indépendance, distribution normale.
Les fonctions **residuals** (ou **resid**), **coefficients** (ou **coef**) et **fitted.values** (ou **fitted**) permettent d'accéder respectivement aux résidus, coefficients et valeurs ajustées. Elles sont très utiles, notamment pour étudier les résidus.

```
> fitted(reg)
     1      2   3      4      5
 113.4 127.2 141 154.8 168.6
```

b) résultats graphiques

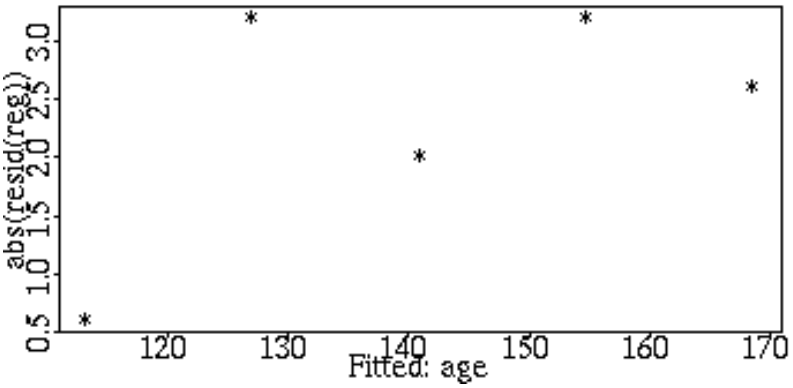
La fonction **plot** est une fonction “générique” qui réalise une représentation graphique. Si les données lui sont fournies sous forme de vecteurs ou de matrices, la représentation graphique est une représentation classique sous forme de nuage de points, car les vecteurs et matrices sont des objets qui n’appartiennent à aucune classe particulière. Si l’argument est un objet de classe “lm” par exemple, c’est la fonction **plot.lm** qui sera réellement exécutée, de façon à fournir une représentation graphique adaptée à la classe de l’objet.

```
> class(reg)
[1] "lm"
> plot(reg)
```



Le premier graphique représente les valeurs observées en fonction des valeurs ajustées. Les pointillés représentent la droite de pente 1, passant par l’origine.

Il permet de visualiser les résidus. En effet, si ceux-ci étaient nuls, tous les points seraient situés sur la droite de pente 1. Les écarts verticaux entre les points et la droite représentent donc les résidus. On pourrait détecter de grands résidus, le cas échéant (ici, les points sont proches de la droite, les résidus sont petits). On peut également vérifier que la distribution des résidus ne présente pas de structure particulière (les points semblent répartis aléatoirement de part et d’autre de la droite).



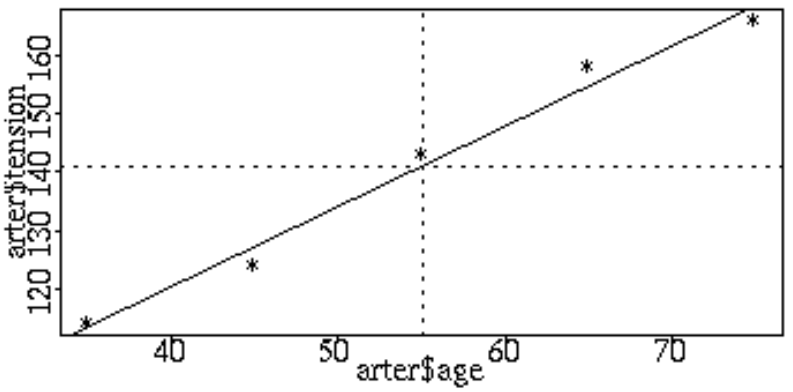
Le deuxième graphique représente la valeur absolue des résidus en fonction des valeurs ajustées. Il permet le contrôle du postulat : "la variance des erreurs est constante". On peut penser qu'ici, elle ne dépend pas des valeurs ajustées.

Le test de Fisher a révélé que le modèle envisagé apportait une explication significative. De plus les résidus sont distribués aléatoirement et leur variance semble constante. Le modèle est donc valide et il s'écrit alors :

$$\widehat{Y} = 65,1 + 1,38X$$

On peut tracer d'autres graphiques utiles à l'analyse du modèle, et y ajouter des éléments supplémentaires : des droites (**abline**), des points (**points**), du texte (**text**) . . .

```
> plot(arter$age,arter$tension)
> abline(reg)
> abline(v=mean(arter$age),lty=2)
> abline(h=mean(arter$tension),lty=2)
```



On représente les observations. On superpose la droite de régression. En effet, si l'argument de la fonction "abline" est un objet de classe lm, S-plus trace la droite correspondant aux coefficients estimés. On trace en pointillée (`lty=2`) une droite verticale au niveau de la moyenne de l'âge (`v=mean(age)`), et une droite horizontale au niveau de la moyenne de la tension (`h=mean(tension)`).
On vérifie ainsi que le point qui a pour coordonnées les moyennes d'âge et de tension, se trouve bien sur la droite de régression.

IV Pour aller plus loin . . .

Dans les résultats affichés par la fonction "summary", on trouve également la corrélation des cofficients. Elle vaut $-0{,}9685$. Les deux coefficients sont corrélés. On peut les "rendre indépendants" en centrant la variable explicative. L'équation du modèle se calcule alors de la façon suivante :

$$
\begin{aligned}
Y_i &= \alpha + \beta X_i + \beta\overline{X} - \beta\overline{X} + \varepsilon_i \\
&\Leftrightarrow \\
Y_i &= \alpha + \beta\left(X_i - \overline{X}\right) + \beta\overline{X} + \varepsilon_i \\
&\Leftrightarrow \\
Y_i &= \underbrace{\alpha + \beta\overline{X}} + \beta\underbrace{\left(X_i - \overline{X}\right)} + \varepsilon_i \\
&\Leftrightarrow \\
Y_i &= \alpha' + \beta Z_i + \varepsilon_i
\end{aligned}
$$

$$
\begin{aligned}
\alpha' &= \alpha + \beta\overline{X} = \overline{Y} \\
Z_i &= X_i - \overline{X}
\end{aligned}
$$

On refait la régression sur la variable age centrée (Z) que l'on inclut dans le data frame. Cette nouvelle régression est une autre paramétrisation du même modèle. Les résultats tels que résidus, carrés moyens résiduels, degrés de liberté et statistiques de Fisher sont donc identiques. Par contre, l'estimation de l'ordonnée à l'origine diffère. Elle est égale à la moyenne des tensions. Cette nouvelle ordonnée à l'origine α' a plus de sens que α, dans la mesure où son estimation correspond à la valeur prédite au "centre" du domaine d'observation ($\overline{X}$). Cette valeur prédite est indépendante de l'estimation de la pente, ce qui explique qu'alors, la corrélation entre les deux coefficients soit nulle.

L'indépendance entre $\overline{Y}$ (la nouvelle ordonnée à l'origine) et la pente est assez intuitive : la droite passe par le point moyen $\left(\overline{X}, \overline{Y}\right)$. Une fois calculée $\overline{Y}$ on en déduit l'ordonnée à l'origine, on peut ensuite estimer la pente en recherchant celle qui minimise le critère des moindres carrés : il suffit de "faire pivoter" la droite autour du point moyen, et de choisir celle qui minimise le critère. On estime ainsi la pente, indépendemment de l'ordonnée à l'origine.

```
> age.centre_arter$age-mean(arter$age)
> arter_data.frame(arter,age.centre)
> rm(age.centre)
> reg1_lm(tension~age.centre,data=arter)
> summary(reg1)
```

```
Call: lm(formula = tension ~ age.centre, data = arter)
Residuals:
   1    2 3   4    5
 0.6 -3.2 2 3.2 -2.6

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) 141.0000   1.4514    97.1452   0.0000
 age.centre   1.3800   0.1026    13.4461   0.0009

Residual standard error: 3.246 on 3 degrees of freedom
Multiple R-Squared: 0.9837
F-statistic: 180.8 on 1 and 3 degrees of freedom,
the p-value is 0.0008894

Correlation of Coefficients:
           (Intercept)
age.centre      0
```

Bibliographie

- La régression nouveaux regards sur une ancienne méthode statistique, R. Tomassone, E. Lesquoy, C. Millier, INRA actualités scientifiques et agronomiques, MASSON.
- Applied Regression Analysis, N.R. Draper, H. Smith, Wiley & Sons.
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- A handbook of Statistical Analysis using Splus, B.S. Everitt (1994), Chapman & Hall.
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.

L'ANALYSE DE VARIANCE A UN FACTEUR

Résumé des principales commandes :

```
> anavar_lm(hauteur~foret,data=arbre)
> summary(anavar)
> anova(anavar)
> plot(anavar)
```

I Le problème

1) Les données

Des forestiers ont réalisé des plantations d'arbres en trois endroits. Plusieurs années plus tard, ils souhaitent savoir si la hauteur des arbres est identique dans les trois forêts. Chacune des forêts constitue une population. Dans chacune des forêts, on tire au sort un échantillon d'arbres, et on mesure la hauteur de chaque arbre.

Forêt 1	Forêt 2	Forêt 3
23,4	18,9	22,5
24,4	21,1	22,9
24,6	21,1	23,7
24,9	22,1	24,0
25,0	22,5	24,0
26,2	23,5	
	24,5	

2) Choix de l'analyse

L'unité expérimentale est l'individu sur lequel chaque observation est effectuée, c'est l'arbre. La hauteur est mesurée pour chaque arbre échantilloné, c'est une variable aléatoire, c'est la variable expliquée. Le seul facteur de variation est le facteur forêt, facteur à trois niveaux (modalités). C'est un facteur qualitatif. On cherche à expliquer la variable hauteur par le facteur forêt. Autrement dit,

on cherche à comparer les hauteurs moyennes des arbres dans les trois forêts. L'analyse de variance est l'outil adapté.

L'analyse de variance consiste à calculer les moyennes empiriques des hauteurs dans chaque forêt et à se demander si la différence existant entre ces moyennes est due au hasard de l'échantillonage, ou bien si elle existe réellement. On teste l'hypothèse nulle "la hauteur des arbres ne dépend pas de la forêt". Si l'hypothèse est rejetée, c'est qu'une des forêts au moins se distingue d'une autre. On peut modéliser la hauteur en l'expliquant par la forêt de provenance.

II Premiers traitements

1) Saisie

Création de deux vecteurs contenant les données : "foret" et "hauteur".
Le vecteur "foret" est créé en répétant les chiffres 1 à 3 respectivement 6, 7 et 5 fois.

```
> foret_rep(1:3,c(6,7,5))
> hauteur_c(23.4,24.4,24.6,24.9,25.0,26.2,
            18.9,21.1,21.1,22.1,22.5,23.5,24.5,
            22.5,22.9,23.7,24.0,24.0)
```

On définit "foret" comme un facteur. On stocke les objets "foret" et "hauteur" dans un data frame qu'on appelle "arbre", et on détruit les deux vecteurs qui ont servi à la création du data frame.

```
> foret_factor(foret)
> arbre_data.frame(foret,hauteur)
> rm(foret)
> rm(hauteur)
> arbre
   foret hauteur
 1     1    23.4
 2     1    24.4
 3     1    24.6
 4     1    24.9
 5     1    25.0
 6     1    26.2
 7     2    18.9
 8     2    21.1
 9     2    21.1
10     2    22.1
```

```
11      2    22.5
12      2    23.5
13      2    24.5
14      3    22.5
15      3    22.9
16      3    23.7
17      3    24.0
18      3    24.0
```

Calcul des moyennes par forêt :

```
> moy_tapply(arbre$hauteur,arbre$foret,mean)
> moy
      1         2     3
 24.75 21.95714 23.42
```

Calcul de la moyenne générale :

```
> moy.g_mean(arbre$hauteur)
> moy.g [1] 23.29444
```

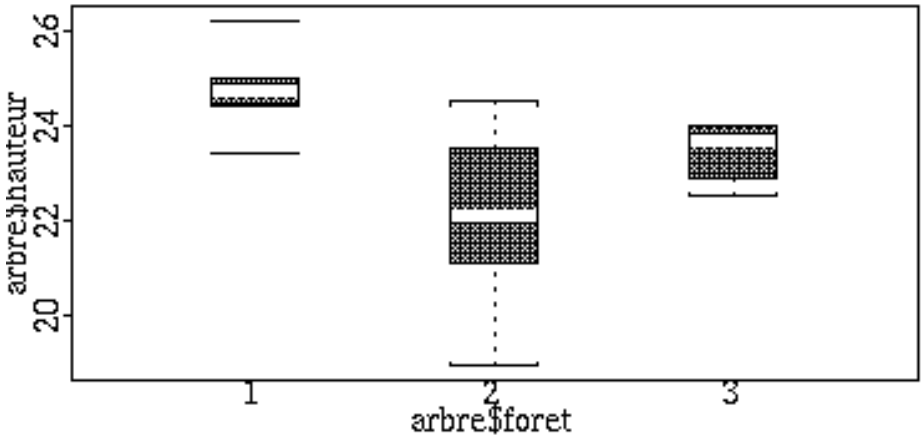
Calcul de la moyenne des moyennes :

```
> moy.moy_mean(moy)
[1] 23.37571
```

La fonction “tapply” applique une fonction donnée (ici, “mean”) à un tableau de valeurs. Ici, elle calcule la moyenne des hauteurs pour chaque forêt.
La moyenne générale n’est pas égale à la moyenne des moyennes, car le plan d’expérience est déséquilibré : le nombre de répétitions dépend de la forêt.

2) Visualisation graphique

```
> plot(arbre$foret,arbre$hauteur)
```



La fonction “plot” reconnait la classe “factor” du premier argument arbre$foret, et appelle la fonction “plot.factor” qui trace une boîte à pattes par forêt.
Cette représentation est un résumé intéressant, mais abusif lorsque le nombre de répétitions est faible, ou lorsque la distribution n’est pas unimodale. Ici, le nombre de répétitions est faible, c’est pourquoi on représente chacune des observations en utilisant l’argument optionnel **character**. On peut ajouter à ce graphique les moyennes locales, et la moyenne globale que l’on représentera par une droite.

```
> plot(arbre$foret,arbre$hauteur,character="o")
> points(1:3,moy,pch="-")
> abline(h=moy.g)
```

III Analyse de variance

1) Le modèle et sa mise en œuvre

On veut expliquer la hauteur en fonction de la forêt. “hauteur” est la variable expliquée, la formule “hauteur~foret” décrit le modèle. Comme en régression linéaire simple, le terme constant (ici μ) est implicitement contenu dans le modèle, c’est-à-dire que la formule “hauteur~foret” est équivalente à “hauteur~1+foret”.

Le modèle étudié est le suivant : $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Y_{ij} : hauteur de l’arbre j de la forêt i,
μ : moyenne générale,
α_i : écart entre la forêt i et la moyenne générale,
ε_{ij} : erreurs indépendantes, d’espérance nulle, telles que $var(\varepsilon_{ij}) = \sigma^2$.

Afin d’obtenir des résultats respectant la contrainte $\sum_{i=1}^{3} \alpha_i = 0$, celle-ci doit être précisée en option (voir **Pour aller plus loin . . .**).

```
> options(contrasts=c("contr.sum","contr.sum"))
> anavar_lm(hauteur~foret,data=arbre)
```

2) Visualisation des résultats

a) résultats numériques

```
> summary(anavar)
```

```
Call: lm(formula = hauteur ~ foret, data = arbre)
Residuals:
   Min     1Q   Median   3Q    Max
 -3.057 -0.7729 0.1464 0.5707 2.543

Coefficients:
             Value  Std. Error  t value Pr(>|t|)
(Intercept)  23.3757   0.3133    74.6211  0.0000
     foret1   1.3743   0.4409     3.1167  0.0071
     foret2  -1.4186   0.4251    -3.3374  0.0045

Residual standard error: 1.317 on 15 degrees
 of freedom
Multiple R-Squared: 0.4933
F-statistic: 7.301 on 2 and 15 degrees of freedom,
the p-value is 0.006107

Correlation of Coefficients:
       (Intercept)  foret1
foret1 -0.0133
foret2 -0.1171     -0.4306
```

La fonction “summary” munie d’un objet de classe lm comme argument, appelle en fait la fonction “summary.lm”, définie pour afficher les résultats d’un objet de classe lm sous une forme adaptée.

Les résidus sont résumés par la médiane, le plus petit, le plus grand, le quartile inférieur et le quartile supérieur. Cela permet de détecter des résidus grands en valeur absolue, ou de soupçonner une dissymétrie de la distribution des résidus. Toutefois, les résidus associés à chaque observation sont accessibles grâce à la fonction “resid”.

```
> resid(anavar)
   1     2     3    4    5    6       7
-1.35 -0.35 -0.15 0.15 0.25 1.4 -3.057143
      8          9         10         11         12
-0.8571429 -0.8571429 0.1428571 0.5428571 1.542857
     13      14   15    16    17   18
  2.542857 -0.92 -0.52 0.28  0.58 0.58
```

Dans la colonne “value” du tableau, on trouve les effets estimés. “Intercept” représente la moyenne $\widehat{\mu}$, “foret1” et “foret2” représentent $\widehat{\alpha_1}$ et $\widehat{\alpha_2}$. On remarque qu'il manque l'effet de la forêt 3. Celui-ci peut se déduire des autres effets, en utilisant la contrainte sur la somme des effets. La fonction “dummy.coef” évite de faire le calcul.

```
> dummy.coef(anavar)
$"(Intercept)":
 (Intercept)
    23.37571
$foret:
        1          2           3
 1.374286 -1.418571 0.04428571
```

La statistique de Fisher :

On cherche à tester l'hypothèse nulle “les moyennes des trois forêts sont égales”. Pour cela, on calcule une statistique, la statistique de Fisher-Snedecor.

$SCE_M = \sum_i (Y_{i.} - Y_{..})^2$ $\qquad$ $SCE_R = \sum_{i,j} (Y_{ij} - Y_{i.})^2$

$CMM = \frac{SCE_M}{I-1}$ $\qquad$ $CMR = \frac{SCE_R}{n-I}$

I : nombre de modalités du facteur,
n : nombre total de mesures.

On a alors : $F = \frac{CMM}{CMR}$.

La valeur calculée dans l'exemple est 7.301. Or, sous l'hypothèse nulle :”les moyennes des 3 forêts sont égales”, F suit une loi de Fisher-Snedecor à (I-1) et (n-I) degrés de libertés, c'est-à-dire 2 et 15. La valeur critique pour un niveau de confiance de 95% est 3,68. La statistique calculée est supérieure à la valeur critique lue dans la table de Fisher, on rejette donc l'hypothèse nulle avec moins de 5% de chances d'émettre une conclusion fausse.

On conclut que les moyennes des 3 forêts ne sont pas égales, c'est-à-dire que la hauteur des arbres varie d'une forêt à l'autre.

Remarque :

La valeur critique peut être lue dans la table de Fisher, on peut aussi l'obtenir avec S-plus. La fonction "qf" donne le **q**uantile de la loi **f**isher-snedecor en précisant en argument, le niveau de confiance donné, ainsi que les deux degrés de liberté définissant la loi.

```
> qf(0.95,2,15)
[1] 3.68232
```

On peut aussi baser notre conclusion sur la lecture de la p-value uniquement. Il y a un risque d'erreur de 0.006107 si l'on rejette l'hypothèse nulle, c'est-à-dire que l'on a 0.6% de chances de se tromper en rejetant Ho. Le risque est très faible, on rejette donc l'hypothèse Ho.

On présente généralement les résultats sous la forme d'une table d'analyse de variance :

```
> anova(anavar)
```

```
Analysis of Variance Table

Response: hauteur

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq F Value       Pr(F)
foret      2  25.30930 12.65465 7.30072 0.006107144
Residuals 15  26.00014  1.73334
```

"Response" : variable expliquée.

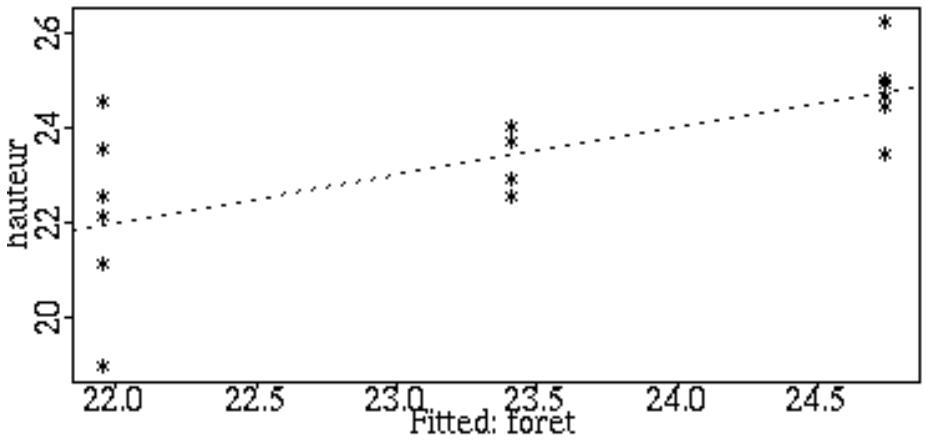
Pour chaque terme du modèle, c'est-à-dire le facteur étudié (foret) et les résidus (residuals), cette table indique :

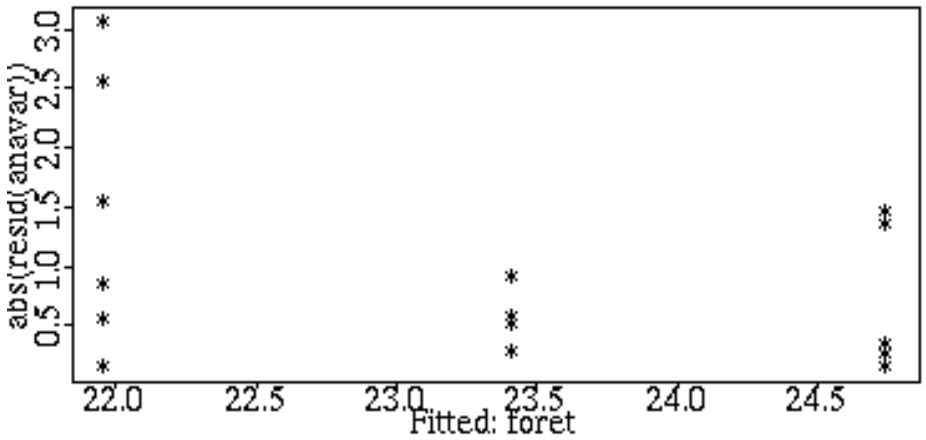
Df : le nombre de degrés de liberté,
Sum of Sq : la somme des carrés des écarts (SCE),
Mean Sq : le carré moyen (CM),
F Value : la statistique calculée,
Pr(F) : la probabilité d'obtenir une statistique de Fisher au moins aussi grande que la statistique calculée, si Ho est vraie. Cette quantité s'interprète comme le risque d'erreur réel lorsque l'on rejette l'hypothèse nulle (p-value).

Le message “Terms added sequentially (first to last)” attire l’attention de l’utilisateur sur le fait que le plan est déséquilibré (le nombre de répétitions par forêt n’est pas constant). Cela est sans conséquence s’il s’agit d’une analyse de variance à un facteur, la table d’analyse de variance affichée peut être interprétée telle que. Par contre, s’il s’agit d’une analyse de variance à plusieurs facteurs, cela signifie que les sommes de carrés affichées dans cette table d’analyse de variance dépendent de l’ordre d’introduction des facteurs dans le modèle. Il est alors conseillé d’afficher les sommes de carrés dites de type III pour interpréter (voir la fonction “drop1.aov” et le chapître “ l’analyse de variance à plusieurs facteurs : plan déséquilibré”).

b) résultats graphiques

```
> plot(anavar)
```

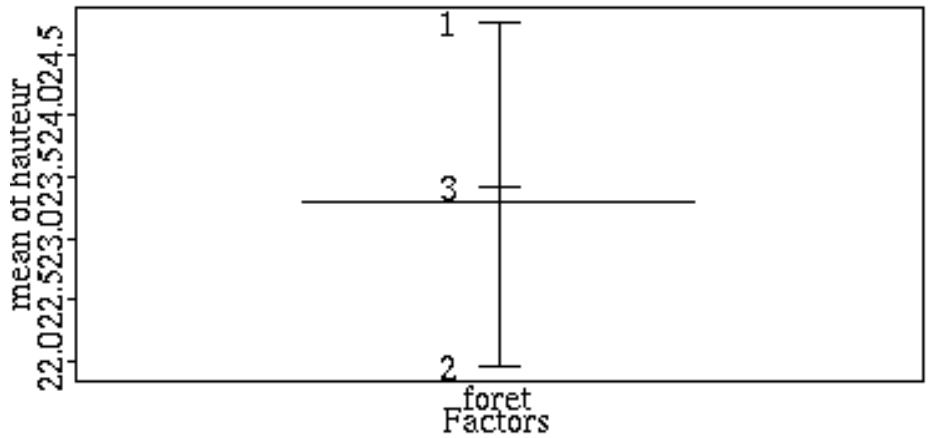




Les deux graphiques obtenus sont similaires à ceux tracés dans le cas de la régression, puisque la fonction appelée est la même : “plot.lm”. Ils permettent une vérification a posteriori des hypothèses d’homogénéité des variances et de distribution aléatoire des résidus autour de zéro.

On peut représenter les résultats de l’analyse de variance grâce à la fonction “plot.design” qui trace chaque moyenne pour chaque niveau de chaque facteur. Les arbres de la forêt 1 sont en moyenne plus hauts que ceux des forêts 2 et 3, alors que ceux de la forêt 2 sont les plus petits.

```
> plot.design(arbre)
```



IV Pour aller plus loin . . .

Les valeurs estimées des effets dépendent du système de contrastes choisi, ou (c’est équivalent) du système de contraintes choisi. Choisir un système de contraste donné revient à choisir une paramétrisation du modèle, tout comme en régression linéaire, on a choisi une nouvelle paramétrisation en centrant la variable explicative.

Par défaut, Splus utilise les contrastes de Helmert, appelés “**contr.helmert**” (pour les facteurs non ordonnés).

Si on choisit de respecter la contrainte : $\sum_{i=1}^{3} \alpha_i = 0$, on utilise les contrastes “**contr.sum**” dits contrastes de somme.

Le calcul des coefficients avec les contrastes d'Helmert se fait de la façon suivante :

- le premier coefficient est la différence entre les moyennes des niveaux 1 et 2, que l'on divise ensuite par deux.

```
> (moy[2]-moy[1])/2
         2
-1.396429
```

- le deuxième coefficient est la différence entre la moyenne du niveau 3 et la moyenne des moyennes des niveaux 1 et 2, le tout divisé par 3.

```
> (moy[3]-mean(moy[1:2]))/3
          3
0.02214286
```

- le nième coefficient sera la différence entre la moyenne du niveau (n+1) et la moyenne des moyennes des niveaux 1,2, . . . ,n, le tout divisé par (n+1).
  Chacun des paramètres mesure l'écart moyen entre le n-ième niveau et les (n-1)-ième niveaux précédents.

Les coefficients estimés avec les contrastes de somme sont calculés par la différence entre la moyenne du niveau étudié et la moyenne des moyennes.

Comparons les coefficients obtenus dans les deux cas.
Par défaut, Splus utilise les contrastes de Helmert si le facteur est non ordonné, et les contrastes polynomiaux si le facteur est ordonné. Cependant, on a choisi d'utiliser les contrastes "contr.sum", en tapant la commande : **`options(contrasts=c("contr.sum","contr.sum"))`**. Le premier système de contrastes est celui des facteurs non ordonnés, et le second, celui des facteurs ordonnés. Les "contraintes de somme" seront donc le système de contraintes (ou contrastes) utilisé dans toute la suite de la session, à moins d'en choisir explicitement un autre.
Vérifions que le système de contrastes "courants" est bien "contr.sum" :

```
> options()$contrasts
[1] "contr.sum"  "contr.sum"
```

Les contrastes "actuels" sont "contr.sum" que le facteur soit ordonné ou pas. On choisit d'utiliser les contrastes de Helmert, et on estime les effets correspondant à cette nouvelle paramétrisation.

```
> options(contrasts=c("contr.helmert","contr.helmert"))
> anavar1_lm(hauteur~foret,data=arbre)
> coef(anavar1)        #avec contr.helmert
 (Intercept)    foret1     foret2
    23.37571 -1.396429 0.02214286
> coef(anavar)        #avec contr.sum
 (Intercept)   foret1    foret2
    23.37571 1.374286 -1.418571
```

Suivant le système de contraintes choisi, les estimations des paramètres sont différentes. Les corrélations entre les paramètres sont différentes également. L'intérêt des contraintes de Helmert, c'est que les paramètres estimés ne sont pas corrélés, si le plan d'expérience est équilibré. Vérifions le.
Pour cela, on crée un nouveau data frame "arbre1" qui contient les nouvelles données pour un plan équilibré (5 mesures par forêt).

```
> foret_rep(1:3,c(5,5,5))
> hauteur_c(arbre$hauteur[1:5],
          arbre$hauteur[7:11],
          arbre$hauteur[14:18])
> arbre1_data.frame(foret=factor(foret),hauteur)
> rm(foret)
> rm(hauteur)
> anavar2_lm(hauteur~foret,arbre1)

Correlation of Coefficients:
       (Intercept) foret1
foret1  0
foret2  0           0

> options(contrasts=c("contr.sum","contr.sum"))
> anavar3_lm(hauteur~foret,arbre1)
Correlation of Coefficients:
       (Intercept) foret1
foret1  0.0
foret2  0.0        -0.5
```

En conclusion, si le plan est équilibré, les paramètres estimés avec les contrastes de Helmert sont indépendants, ce qui n'est pas le cas avec les contrastes de somme. L'indépendance entre les paramètres est très intéressante, elle permet une interprétation des différents paramètres indépendemment les uns des autres (voir la régression linéaire multiple). Les contrastes de Helmert ont cependant l'inconvénient d'être moins naturels à interpréter d'un point de vue biologique.

Pour les deux systèmes de contraintes, les effets du facteur et le terme constant sont indépendants tout de même.

Bibliographie

- Méthodes statistiques, Snedecor & Cochran (1967), ACTA.
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- A handbook of Statistical Analysis using Splus, B.S. Everitt (1994), Chapman & Hall.
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.

L'ANALYSE DE VARIANCE A DEUX FACTEURS

Résumé des principales commandes :

```
> plot.factor(huile,character="o")
> plot.design(huile)
> anavar2_lm(teneur~.+origine:testeur,data=huile)
> summary(anavar2)
> anova(anavar2)
> plot(anavar2)
```

I Le problème

1) Les données

On dispose de variété de tournesols dont l'origine géographique est l'Afrique, la Hongrie ou le Maroc. Elles sont issues de croisements avec différents testeurs (un testeur est une lignée avec laquelle chaque variété a été croisée). On cherche à savoir si leur teneur en huile est identique, quels que soient le pays de provenance et le testeur utilisé. Pour chaque combinaison origine×testeur, on observe la teneur en huile sur deux parcelles. On obtient deux répétitions par combinaison.

	AFRIQUE	HONGRIE	MAROC
T1	43,54 45,30	44,25 42,55	47,28 49,40
T2	47,21 47,73	44,34 46,49	47,75 49,47

2) Choix de l'analyse

La teneur est mesurée sur chaque unité expérimentale (parcelle de tournesol), c'est une variable aléatoire, c'est la variable expliquée. Il y a deux facteurs de variation : le facteur origine à trois niveaux (ou modalités), et le facteur testeur

à deux niveaux. On cherche à expliquer la variable teneur par les deux facteurs origine et testeur, ce qui revient à comparer les teneurs en huile moyennes pour chaque combinaison origine$\times$testeur. L'analyse de variance à deux facteurs est l'outil adéquat.

Le principe de base est le même que pour l'analyse de variance à un facteur, on compare la variabilité due à un facteur de variation à la variabilité résiduelle, celle qui ne peut être expliquée par aucun des facteurs. On cherche d'abord à tester l'existence d'une interaction entre les facteurs origine et testeur, et si l'interaction n'est pas significative, on teste l'existence d'un effet testeur, et l'existence d'un effet origine. Pour cela, on compare la variabilité expliquée par l'effet testeur (respectivement origine) à la variabilité résiduelle.

II Premiers traitements

1) Saisie

Voici une façon de saisir les données :
Les données sont saisies sous forme d'une matrice à 3 colonnes, la première contenant le numéro du testeur, la seconde le numéro d'origine, et la dernière la teneur en huile. On crée ensuite le data frame "huile", dont les composantes sont le facteur "testeur", le facteur "origine" et la variable "teneur".

```
> huile_matrix(c(
    1,1,43.54,
    1,1,45.30,
    1,2,44.25,
    1,2,42.55,
    1,3,47.28,
    1,3,49.40,
    2,1,47.21,
    2,1,47.73,
    2,2,44.34,
    2,2,46.49,
    2,3,47.75,
    2,3,49.47)
    ,ncol=3,byrow=T)
> huile_data.frame(testeur=as.factor(huile[,1]),
                   origine=as.factor(huile[,2]),
                   teneur=huile[,3])
```

On peut accéder à une seule variable, par exemple la teneur en huile, en tapant la commande **huile$teneur**.

Voici les règles de recherche par défaut. Tous les objets (données ou fonctions) nécessaires à l'exécution d'une commande sont trouvés dans l'un des répertoires suivants, les répertoires étant parcourus dans l'ordre où ils sont affichés.

```
> search()
[1] "/home/fidji/aj/.Data"
[2] "/usr/local/splus3.1/library/inra/.Data"
[3] "/usr/local/splus3.1/splus/.Functions"
[4] "/usr/local/splus3.1/stat/.Functions"
[5] "/usr/local/splus3.1/s/.Functions"
[6] "/usr/local/splus3.1/s/.Datasets"
[7] "/usr/local/splus3.1/stat/.Datasets"
[8] "/usr/local/splus3.1/splus/.Datasets"
```

On peut "attacher" un data frame particulier, ce qui permet d'avoir directement accès aux objets qui le composent. Cela ajoute une règle de recherche. On "détache" le data frame lorsque l'attachement n'est plus utile (**detach()**).

```
> attach(huile)
> search()
[1] "/home/fidji/aj/.Data"
[2] "huile"
[3] "/usr/local/splus3.1/library/inra/.Data"
[4] "/usr/local/splus3.1/splus/.Functions"
[5] "/usr/local/splus3.1/stat/.Functions"
[6] "/usr/local/splus3.1/s/.Functions"
[7] "/usr/local/splus3.1/s/.Datasets"
[8] "/usr/local/splus3.1/stat/.Datasets"
[9] "/usr/local/splus3.1/splus/.Datasets"
```

Le fait d'avoir "attaché" le data frame "huile" permet d'accéder directement aux objets qui le composent.

```
> teneur
   1     2    3     4     5     6    7     8
43.54 45.3 44.25 42.55 47.28 49.4 47.21 47.73
   9    10    11    12
44.34 46.49 47.75 49.47
```

moyenne par testeur :

```
> moy.test_tapply(teneur,testeur,mean)
```

moyenne par origine :

```
> moy.orig_tapply(teneur,origine,mean)
```

moyenne générale (la moyenne générale est la moyenne calculée à partir de

l'ensemble des observations, elle est égale à la moyenne des moyennes car le plan est équilibré) :

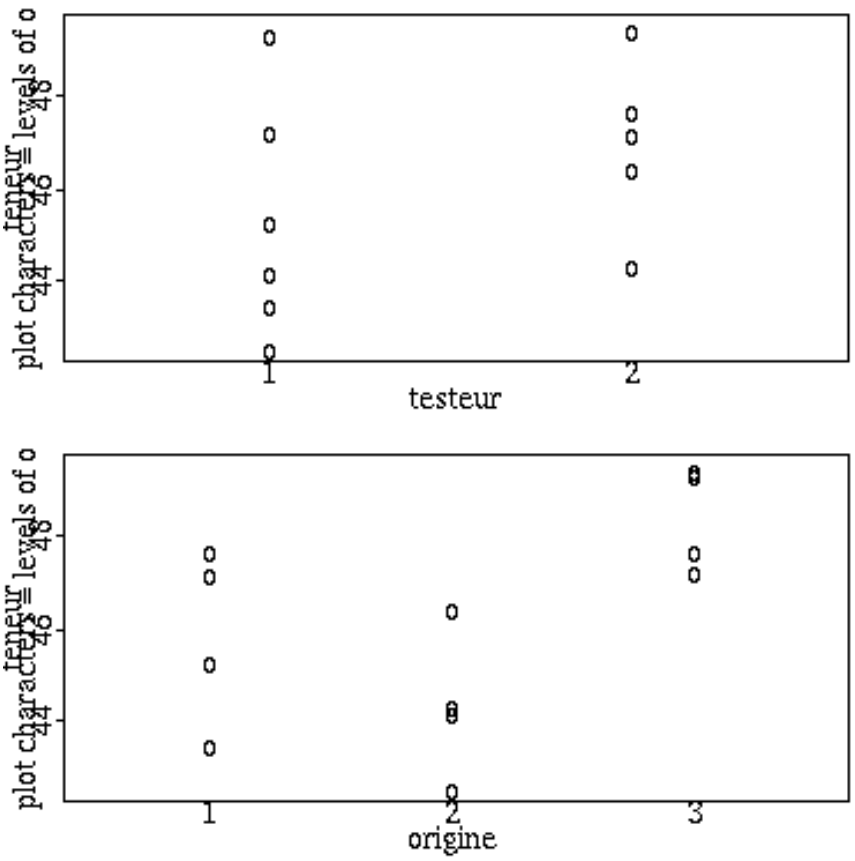
```
> moy.g_mean(teneur)
```

2) Visualisation graphique

Les représentations graphiques qui suivent permettent, le cas échéant, de détecter des "données extérieures", une hétérogénéité des variances ou une dissymétrie de la distribution.

Le nombre de répétitions est faible. Mieux vaut représenter toutes les observations, plutôt que de les résumer (par défaut, Splus trace des boîtes à pattes).

```
> plot.factor(huile,character="o")
```



On peut superposer les moyennes en utilisant les instructions suivantes :

Facteur testeur :

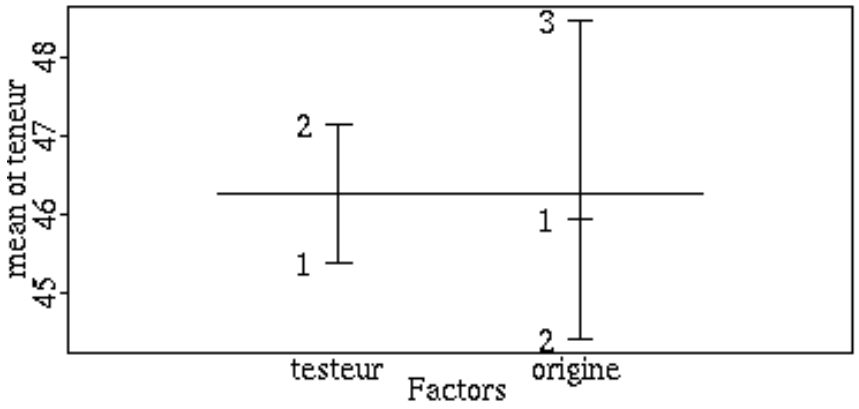
```
> plot(testeur,teneur,character="o")
> points(1:2,moy.test,pch="-")
> abline(h=moy.g,lty=2)
```

Facteur origine :

```
> plot(origine,teneur,character="o")
> points(1:3,moy.orig,pch="-")
> abline(h=moy.g,lty=2)
```

On peut aussi représenter simultanément les moyennes de tous les facteurs et la moyenne générale:

```
> plot.design(huile)
```



III Analyse de la variance

1) Le modèle et sa mise en œuvre

Modèle étudié : $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$.

Y : teneur en huile
μ : moyenne générale
α_i : écart entre l'origine et la moyenne générale
β_j : écart entre le testeur et la moyenne générale
γ_{ij} : interaction
ε_{ijk} : erreurs indépendantes, d'espérence nulle, telles que $var(\varepsilon_{ijk}) = \sigma^2$

On écrit ici le modèle sous une forme abrégée : "teneur ~ . +origine:testeur". Cette écriture est équivalente à "teneur ~ origine+testeur+origine:testeur", ou encore à "teneur ~ 1+origine+testeur+origine:testeur". Le "." est interprété comme "la partie additive du modèle faisant intervenir tous les facteurs et toutes les variables du data frame huile, y compris le terme constant, à l'exception de la variable expliquée". Le terme "origine:testeur" représente l'interaction. Une autre écriture équivalente est "teneur ~ . ^ 2", elle signifie "la partie additive du modèle, et

toutes les interactions d'ordre 2".

On choisit d'utiliser les contraintes "de somme" :
$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$
(ce ne sont pas les contraintes utilisées par défaut)

```
> options(contrasts=c("contr.sum","contr.sum"))
> anavar2_lm(teneur~.+origine:testeur,data=huile)
```

2) Visualisation des résultats

a) résultats numériques

```
> summary(anavar2)
```

```
Call: lm(formula=teneur~.+origine:testeur,data=huile)
Residuals:
    Min      1Q    Median    3Q   Max
 -1.075 -0.865 -2.22e-16 0.865 1.075

Coefficients:
                  Value   Std. Error  t value  Pr(>|t|)
    (Intercept)  46.2758   0.3568   129.6926   0.0000
       testeur  -0.8892   0.3568    -2.4920   0.0470
       origine1  -0.3308   0.5046    -0.6556   0.5364
       origine2  -1.8683   0.5046    -3.7025   0.0101
testeurorigine1  -0.6358   0.5046    -1.2601   0.2544
testeurorigine2  -0.1183   0.5046    -0.2345   0.8224
Residual standard error: 1.236 on 6 degrees of freedom
Multiple R-Squared: 0.8373
F-statistic: 6.176 on 5 and 6 degrees of freedom,
the p-value is 0.02325
```

```
Correlation of Coefficients:

          (Intercept)   testeur origine1 origine2 testeurorigine1
testeur           0.0
origine1          0.0       0.0
origine2          0.0       0.0     -0.5
testeurorigine1   0.0       0.0      0.0      0.0
testeurorigine2   0.0       0.0      0.0      0.0     -0.5
```

Comme dans les exemples précédents, on retrouve une statistique globale (F-statistic) qui permet de tester l'hypothèse nulle "Le modèle n'apporte aucune explication", soit "$\alpha_i = \beta_j = \gamma_{ij} = 0 \qquad \forall i, \forall j$".
Il faut afficher la table d'analyse de variance pour visualiser les statistiques permettant de tester l'interaction, puis l'effet de chacun des facteurs. Elles sont calculées de la manière suivante :

$$SCE_{or} = \sum_{i,j,k} (Y_{i..} - Y_{...})^2 \qquad CM_{or} = \frac{SCE_{or}}{I-1}$$

$$SCE_{te} = \sum_{i,j,k} (Y_{.j.} - Y_{...})^2 \qquad CM_{te} = \frac{SCE_{te}}{J-1}$$

$$SCE_I = \sum_{i,j,k} (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 \qquad CM_I = \frac{SCE_I}{(I-1)(J-1)}$$

$$SCE_R = \sum_{i,j,k} \left(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...}\right)^2 \qquad CM_R = \frac{SCE_R}{(r-1)IJ}$$

I : nombre de modalités pour le facteur origine
J : nombre de modalités pour le facteur testeur
r : nombre de répétitions

$$\text{Facteur origine : } F_{or} = \frac{CM_{or}}{CM_R}$$
$$\text{Facteur testeur : } F_{te} = \frac{CM_{te}}{CM_R}$$
$$\text{Interaction : } F_I = \frac{CM_I}{CM_R}$$

Table d'analyse de variance correspondante :

```
> anova(anavar2)
```

```
Analysis of Variance Table

Response: teneur

Terms added sequentially (first to last)
              Df Sum of Sq Mean Sq  F Value   Pr(F)
testeur        1  9.48741  9.48741  6.20995 0.0470361
origine        2 33.74582 16.87291 11.04411 0.0097472
origine:testeur 2  3.94822  1.97411  1.29215 0.3414604
Residuals      6  9.16665  1.52777
```

Response : variable expliquée,
Df : degrés de liberté,

Sum of Sq : somme des carrés des écarts (SCE),
Mean Sq : carré moyen (CM),
F Value : statistique calculée,
Pr(F) : risque d'erreur réel lorsque l'on rejette l'hypothèse nulle.

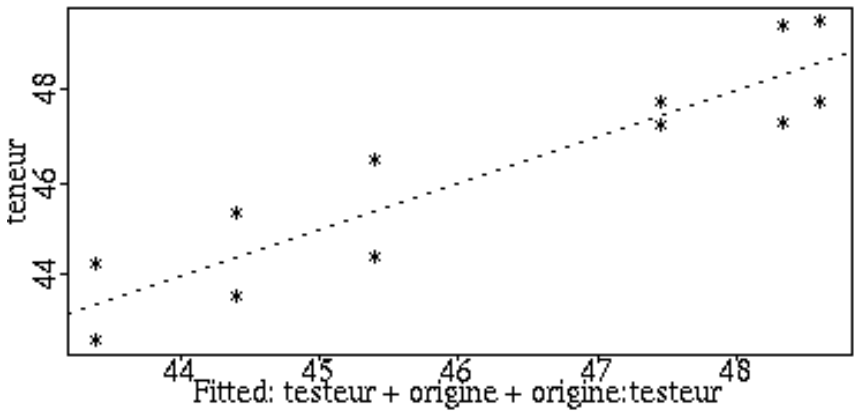
Facteur testeur :
Sous l'hypothèse nulle "les moyennes par modalité du facteur testeur sont égales", le rapport des carrés moyens F_{te} suit une loi de Fisher à 1 et 6 degrés de libertés. Le risque d'erreur lu dans la table d'analyse de variance est faible (< 5%), on rejette donc l'hypothèse nulle. On conclut que les moyennes des testeurs ne sont pas égales, la teneur en huile des tournesols dépend du testeur utilisé (p<5%).

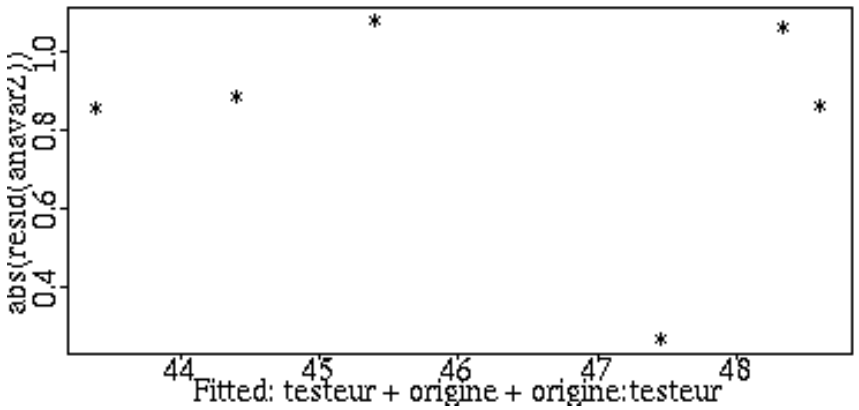
Facteur origine :
Sous l'hypothèse nulle "les moyennes par modalité du facteur origine sont égales", le rapport des carrés moyens F_{or} suit une loi de Fisher à 2 et 6 degrés de liberté. Le risque d'erreur lu dans la table d'analyse de variance est très faible (< 1%), on rejette donc l'hypothèse nulle. On conclut que les moyennes des origines ne sont pas égales, la teneur en huile des tournesols dépend aussi du pays d'origine (p<1%).

Interaction :
Sous l'hypothèse nulle "les termes d'interaction sont égaux", le rapport des carrés moyens F_I suit une loi de Fisher à 2 et 6 degrés de liberté. Le risque d'erreur lu dans la table d'analyse de variance est grand (>34%), on accepte l'hypothèse nulle. On conclut qu'il n'y a pas d'interaction significative entre les deux facteurs.

b) résultats graphiques

```
> plot(anavar2)
```





Ces graphiques sont un outil de validation des postulats. Ils permettent de s'assurer qu'aucun des résidus n'est particulièrement grand, que la variance est homogène (elle ne dépend pas des valeurs ajustées).
Les postulats semblent vérifiés, le modèle est donc adéquat. La table d'analyse de variance permet de conclure que le pays d'origine du tournesol et le testeur utilisé influencent la teneur en huile des tournesols, l'effet de l'origine est "indépendant" de l'effet testeur et réciproquement.

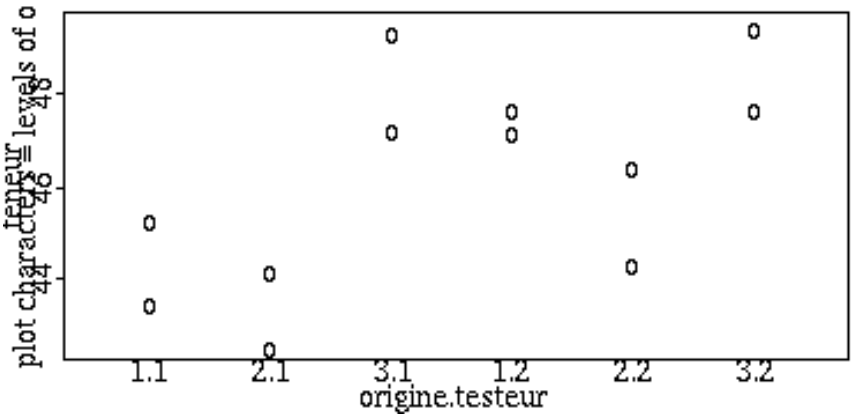
IV Pour aller plus loin . . .

On peut ajouter un facteur d'interaction dans le data frame "huile".

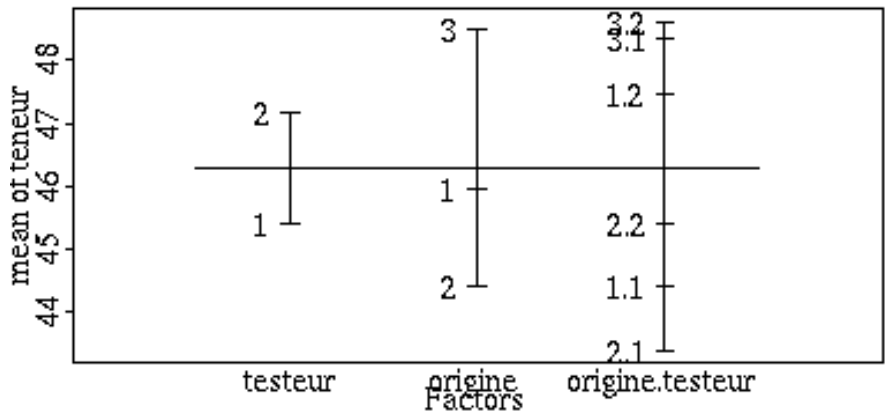
```
> huile$origine.testeur_interaction(origine,testeur)
```

On peut alors effectuer les représentations graphiques avec "plot.factor" et "plot.design". La commande "plot.factor" produit trois graphiques. Les deux premiers sont les mêmes qu'avec "huile", c'est pourquoi seul le troisième graphique est représenté ici.

```
> plot.factor(huile,character="o")
```

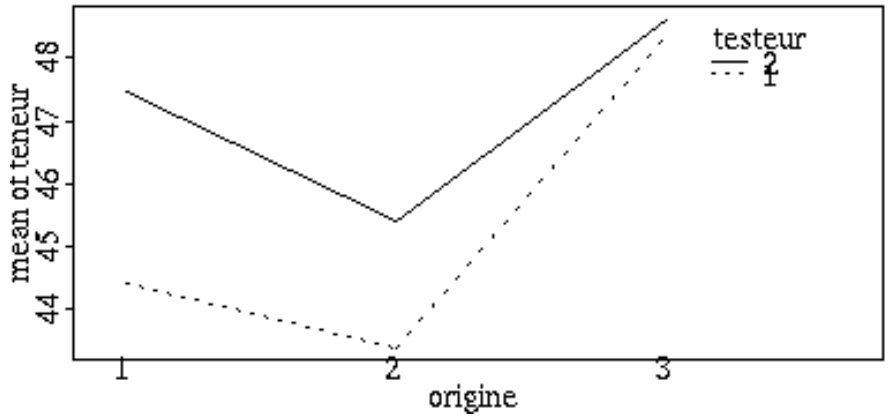



```
> plot.design(huile)
```



Ce graphique permet de visualiser les 6 moyennes correspondant à l'interaction, de repérer les moyennes faibles ou élevées par rapport aux moyennes du même niveau. Cela est particulièrement intéressant dans le cas d'une interaction significative.

```
> interaction.plot(origine,testeur,teneur)
```



On représente ici les moyennes de la variable “teneur” pour chaque origine et pour chaque testeur. Ce graphique est une aide à l’interprétation de l’interaction, **lorsque le test a révélé une interaction significative**.
Lorsqu’il n’y a pas d’interaction, les lignes brisées sont parallèles ou approximativement parallèles : il faut tenir compte du fait que les moyennes empiriques ne sont pas les vraies moyennes inconnues mais des réalisations d’une variable aléatoire. Un écart au parallélisme est une “mesure” de l’interaction. Ici les droites ne sont pas tout à fait parallèles, bien que le test de Fisher ait conclu à l’absence d’interaction entre les deux facteurs. Cela s’explique soit par l’existence d’une interaction légère que le test n’a pas pu déceler faute d’un nombre de répétitions suffisant, soit par le caractère aléatoire des moyennes empiriques : le tirage au sort des unités expérimentales a abouti à ces six moyennes, donc à ce graphique, mais de nouvelles observations aboutiraient à un graphique différent.

```
> detach()
```

Bibliographie

- Méthodes statistiques, Snedecor & Cochran (1967), ACTA.
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- A handbook of Statistical Analysis using Splus, B.S. Everitt (1994), Chapman & Hall.
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.

L'ANALYSE DE VARIANCE A DEUX FACTEURS : PLAN DESEQUILIBRE

Résumé des principales commandes :

```
> plot.design(carotte)
> anavar_lm(jgerm~.+variete:sol,data=carotte)
> drop1.aov(anavar,scope=anavar$call)
> plot(anavar)
```

I Le problème

1) Les données

On désire étudier la germination de trois variétés de graines de carottes dont les graines sont semées dans deux types de sol différents. On cherche à savoir si le nombre de jours avant germination de ces graines est identique, quel que soit la variété et le type de sol. Pour chaque combinaison variété$\times$sol, on mesure le nombre de jours qui s'écoule avant la germination des graines. Ces mesures sont faites sur quinze parcelles.

Y_{ijk}	Variété 1	Variété 2	Variété 3
Sol 1	6 10 11	13 15	14 22
Sol 2	12 19 15 18	31	18 9 12

2) Choix de l'analyse

Le nombre de jours avant germination des graines est compté, donc mesuré, sur chaque unité expérimentale (parcelle de graines de carottes), c'est une variable aléatoire, c'est la variable expliquée. La variété de carotte utilisée et le type de sol dans lequel les graines sont semées sont des facteurs de variations, respectivement à trois et deux niveaux. On cherche à expliquer le nombre de jours avant germination par les facteurs variété et sol, c'est-à-dire que l'on compare les moyennes observées pour chaque combinaison des deux facteurs. On réalise donc une analyse de variance à deux facteurs.

Le nombre de mesures effectuées n'est pas le même pour chaque combinaison : le plan est déséquilibré. La décomposition de la somme de carrés expliquée par le modèle, en une somme de carrés expliquée par le facteur variété et une somme de carrés expliquée par le facteur sol, n'est pas unique. Autrement dit, il existe différentes façons de calculer les sommes de carrés. Les différents modes de calcul sont associés à différentes formulations de l'hypothèse nulle.

II Premiers traitements

1) Saisie

On saisit les données sous forme d'une matrice à trois colonnes. Le data frame "carotte" est composé du facteur "variete", du facteur "sol", et de la variable "jgerm".

```
> carotte_matrix(c(
   1,1,6,
   1,1,10,
   1,1,11,
   1,2,12,
   1,2,15,
   1,2,19,
   1,2,18,
   2,1,13,
   2,1,15,
   2,2,31,
   3,1,14,
   3,1,22,
   3,2,18,
   3,2,9,
   3,2,12,),ncol=3,byrow=T)
> carotte_data.frame(variete=as.factor(carotte[,1]),
                     sol=as.factor(carotte[,2]),
                     jgerm=carotte[,3])
```

Plan équilibré et non équilibré :

On peut vérifier avec Splus, que le plan est équilibré ou qu'il ne l'est pas. Pour cela, on utilise la fonction "replications" qui affiche les nombres de répétitions.

L'argument fourni à la fonction "replications" est la formule qui décrit le modèle. Il s'agit ici d'un modèle avec interaction. Le nombre de répétitions sera donc calculé pour chaque combinaison variété×sol (si le modèle avait été additif, seuls les nombres de répétitions par variété d'une part, par type de sol d'autre part, auraient été calculés).

```
> repl_replications(jgerm~variete*sol,carotte)
> repl
$variete:
 1 2 3
 7 3 5

$sol:
 1 2
 7 8

$"variete:sol":
  1 2
1 3 4
2 2 1
3 2 3
> is.numeric(repl)
[1] F
```

La dernière commande permet de savoir si le plan est équilibré ou non. Si la réponse est F, le plan est déséquilibré, si la réponse est T, le plan est équilibré (dans le cas d'un plan déséquilibré, "repl" est une liste et n'est pas "numeric"; au contraire, si le plan est équilibré, "repl" est un scalaire et il est "numeric").

Dans le cas d'un plan déséquilibré, il est conseillé de baser l'interprétation sur les "moyennes ajustées" (LSMEANS de SAS), et non sur les "moyennes brutes" obtenues directement à partir des observations (ex : la "moyenne brute" pour la variété 1 est calculée en faisant la moyenne de toutes les observations obtenues avec la variété 1). Les moyennes ajustées sont des "moyennes de moyennes" : on calcule d'abord les moyennes pour chaque combinaison des facteurs, les moyennes ajustées pour chaque niveau de facteurs sont calculées à partir des moyennes par combinaison. Si le plan est équilibré, les "moyennes ajustées" coincident avec les "moyennes brutes". Si le plan est déséquilibré, ce n'est pas le cas.

Calcul des moyennes ajustées :

Dans le cas d'un plan déséquilibré, on calcule d'abord les moyennes pour chaque combinaison des deux facteurs, les moyennes pour chaque facteur sont obtenues à partir des moyennes de chaque combinaison.

Moyennes par combinaison :

```
> attach(carotte)
> moy.var.sol_tapply(jgerm,list(sol,variete),mean)
```

Moyennes ajustées par variété :

```
> moy.var_apply(moy.var.sol,2,mean)
```

Moyennes ajustées par type de sol :

```
> moy.sol_apply(moy.var.sol,1,mean)
```

Moyenne générale (calculée à partir des moyennes par combinaison, ou par variété, ou par type de sol) :

```
> moy.moy_mean(moy.var.sol)
```

Tableau des résultats

sol 1	sol 2		sol 1	sol 2	moyennes ajustées (moy.var)
6 10 11	12 19 15 18	variété 1	9	16	12,5
13 15	31	variété 2	14	31	22,5
14 22	18 9 12	variété 3	18	13	15,5
		moyennes ajustées (moy.sol)	13,67	20	16,83

On peut obtenir plus directement les moyennes ajustées, à l'aide de la fonction "dummy.coef" (cf. "Pour aller plus loin").

Interaction :

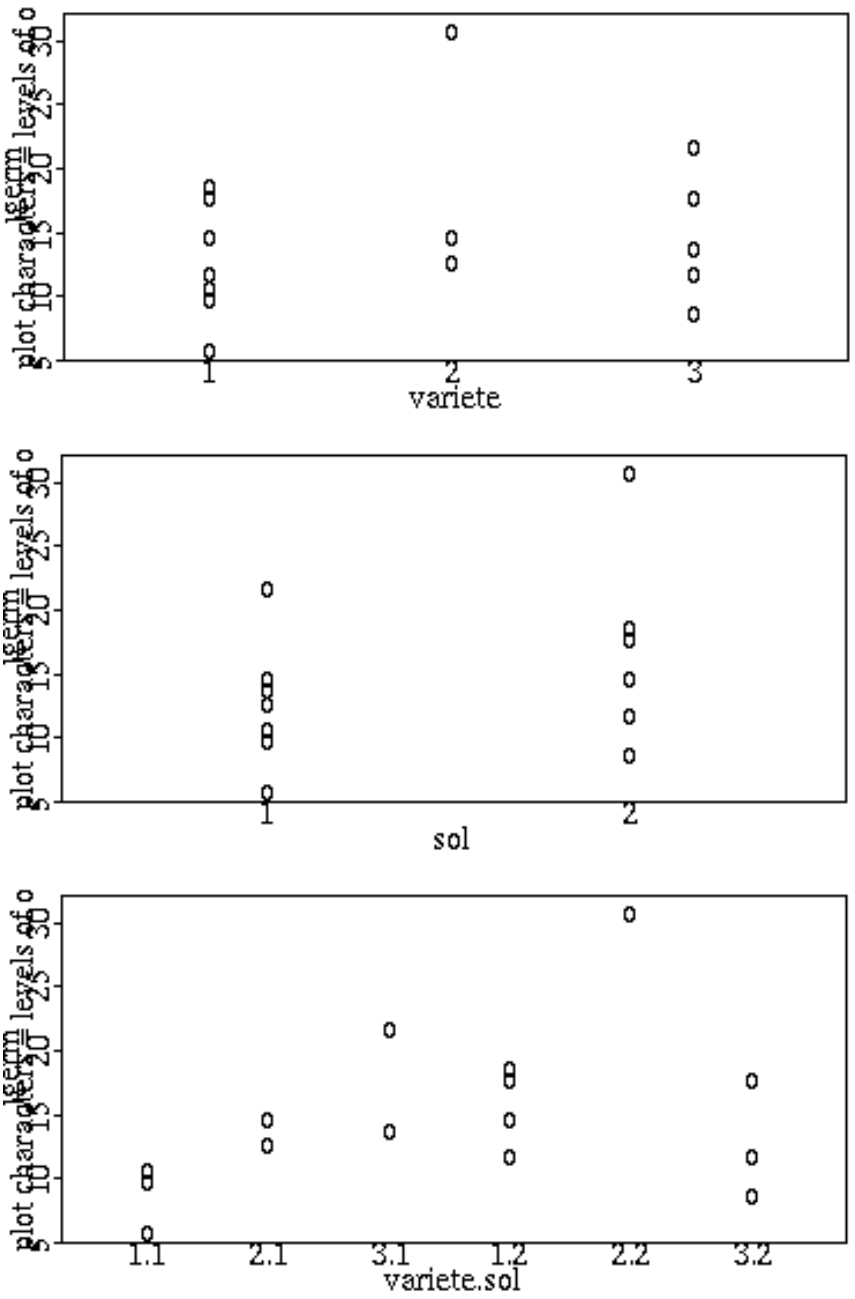
L'interaction peut-être intégrée directement dans le data frame "carotte", sous la forme d'une composante appelée "variete.sol".

Cela n'est pas obligatoire pour étudier le modèle avec interaction, mais cela permet, dans les représentations graphiques, de différencier chacune des combinaisons variété×sol.

```
> carotte$variete.sol_interaction(variete,sol)
> detach()
```

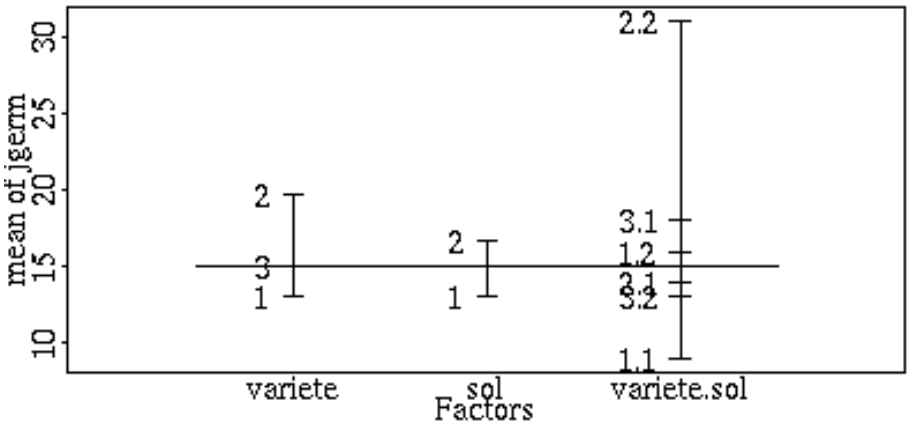
2) Visualisation graphique

```
> plot.factor(carotte,character="o")
```



Le dernier graphique représente le nombre de jours avant germination des graines de carotte en fonction de chaque combinaison des deux facteurs. Cela permet, le cas échéant, de détecter des écarts aux postulats, cela permet également de se faire une idée de l'effet des facteurs, avant de réaliser l'analyse.
On peut aussi représenter les moyennes pour chaque niveau de chaque facteur. Remarque : la fonction "plot.design" représente les moyennes brutes, il aurait été préférable de représenter les moyennes ajustées.

```
> plot.design(carotte)
```



III Analyse de la variance

1) Le modèle et sa mise en œuvre

Modèle étudié : $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$.
Y : nombre de jours avant germination
μ : moyenne générale
α_i : écart entre la variété et la moyenne générale
β_j : écart entre le type de sol et la moyenne générale
γ_{ij} : interaction entre les deux facteurs
ε_{ijk} : erreurs indépendantes, d'espérance nulle, telles que $var(\varepsilon_{ijk}) = \sigma^2$
Cette écriture correspond à un modèle complet à deux facteurs, le terme complet signifie que l'interaction est contenue dans le modèle (on dit aussi "modèle interactif complet").
La mise en œuvre se fait de la même manière que dans le cas équirépété.

```
> options(contrasts=c("contr.sum","contr.sum"))
> anavar_lm(jgerm~variete+sol+variete:sol,data=carotte)
```

2) Visualisation des résultats

a) résultats numériques

Ici, les résultats obtenus avec la fonction “summary” n’apporteront pas beaucoup d’information quant à l’interprétation des résultats, mieux vaut afficher la table d’analyse de variance.
Le plan est déséquilibré, la décomposition de la somme de carrés du modèle n’est pas unique.

La décomposition n’est pas unique car l’analyse de variance consiste à projeter le vecteur des observations sur des sous espaces vectoriels qui ne sont pas orthogonaux. En effet, la matrice X du plan d’expérience (matrice d’incidence) s’écrit :

$$X = \begin{array}{c} \begin{matrix} cte & & sol & & variete & & & interaction & & & & \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$\widetilde{X} = \begin{array}{c} \begin{array}{ccccccccc} cte & sol & variete & & & interaction & & & \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$\widetilde{X}'\widetilde{X} = \begin{array}{c} \begin{array}{ccccccccc} cte & sol & variete & & & interaction & & & \end{array} \\ \begin{pmatrix} 15 & 1 & -4 & -2 & -1 & -1 & 1 & -2 & 0 \\ 1 & 15 & -2 & 0 & 1 & 1 & 7 & 4 & 6 \\ -4 & -2 & 10 & 7 & 5 & 3 & -1 & 4 & 3 \\ -2 & 0 & 7 & 12 & 3 & 5 & -1 & 3 & 6 \\ -1 & 1 & 5 & 3 & 5 & 3 & 3 & 3 & 3 \\ -1 & 1 & 3 & 5 & 3 & 5 & 3 & 3 & 3 \\ 1 & 7 & -1 & -1 & 3 & 3 & 7 & 3 & 3 \\ -2 & 4 & 4 & 3 & 3 & 3 & 3 & 4 & 3 \\ 0 & 6 & 3 & 6 & 3 & 3 & 3 & 3 & 6 \end{pmatrix} \end{array}$$

La matrice X n'est pas de plein rang. On construit la matrice $\widetilde{X}$ à partir de X en supprimant la colonne correspondant au premier niveau de chacun des facteurs, et en la retranchant aux colonnes correspondant aux autres niveaux du même facteur. La matrice $\widetilde{X}$ décrit complètement le plan d'expérience, elle est de plein rang. La matrice $\widetilde{X}'\widetilde{X}$ contient les produits scalaires entre les vecteurs colonnes de $\widetilde{X}$. Si le plan est équilibré, la matrice $\widetilde{X}'\widetilde{X}$ est composée de quatre blocs diagonaux (correspondant à μ, variété, sol, variété:sol)), et de 0 en dehors de ces blocs diagonaux. Les quatre blocs diagonaux sont associés à quatre sous-espaces vectoriels de R^n, ces quatre sous-espaces vectoriels sont orthogonaux, et le plan est dit orthogonal (l'orthogonalité d'un plan dépend à la fois des nombres de répétitions, et du modèle choisi). Ici, la matrice $\widetilde{X}'\widetilde{X}$ n'est pas bloc diagonale,

les sous-espaces vectoriels associés à chacun des termes du modèle ne sont pas orthogonaux. La décomposition de la somme de carrés expliquée par le modèle, sur ces sous-espaces vectoriels n'est pas unique.
Nous présentons deux types de décompositions, les sommes de carrés de type I et les sommes de carrés de type III.

Sommes de carrés de type I :

Ce sont les sommes de carrés affichées par la fonction "anova". Elles sont dites séquentielles : "Terms added sequentially".
On les obtient en projetant successivement sur les sous-espaces vectoriels, dans l'ordre correspondant à l'ordre d'introduction des termes dans le modèle. Ici, on projette la somme de carrés expliquée par le modèle sur le sous-espace vectoriel engendré par le facteur variété. On obtient une part de variabilité expliquée par le facteur variété et un reste. On projette ce reste sur le sous-espace vectoriel engendré par le facteur sol.
On teste d'abord l'effet du facteur variété, puis l'effet du facteur sol qui n'est pas déjà pris en compte par le facteur variété, puis l'interaction. Les sommes de carrés obtenues dépendent de l'ordre d'introduction des facteurs dans le modèle.

```
> anova(anavar)
```

```
Analysis of Variance Table
Response: jgerm
Terms added sequentially (first to last)
            Df Sum of Sq  Mean Sq  F Value      Pr(F)
variete      2   93.3333  46.6667 3.500000 0.07508469
sol          1   83.9007  83.9007 6.292553 0.03339305
variete:sol  2  222.7660 111.3830 8.353723 0.00888845
Residuals    9  120.0000  13.3333
```

Sommes de carrés de type III :

C'est, en général, la méthode conseillée. Les résultats ne dépendent pas de l'ordre d'introduction des facteurs dans le modèle.
On projette la somme de carrés expliquée par le modèle sur le sous-espace vectoriel engendré par le facteur sol et l'interaction. On décompose ainsi la somme de carrés expliquée en une somme de carrés expliquée par le facteur sol et l'interaction d'une part, et un reste d'autre part. Ce reste est la somme de carrés (de type III) expliquée par le facteur variété : tout ce qui n'est pas expliqué par le facteur sol et l'interaction , est supposé expliqué par le facteur variété. On projette ensuite la somme de carrés expliquée sur le sous-espace vectoriel engendré par le facteur variété et l'interaction. Le reste est la somme de carrés expliquée par le facteur sol. Les sommes de carrés obtenues sont donc indépendantes de l'ordre

de projection.
On teste l'hypothèse "les moyennes du facteur sont égales". Ces sommes de carrés tiennent compte des nombres de répétitions n_{ij} par combinaison.
Elles peuvent être obtenues à l'aide de la fonction "drop1.aov", sans omettre l'option "scope=anavar$call", où anavar est le nom de l'objet rendu par lm.

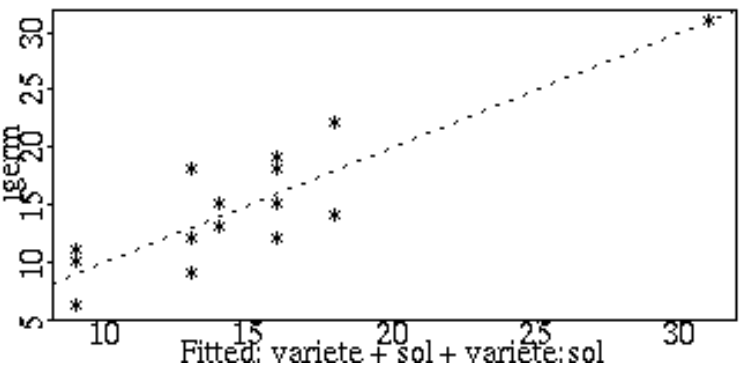
```
> drop1.aov(anavar,scope=anavar$call)
```

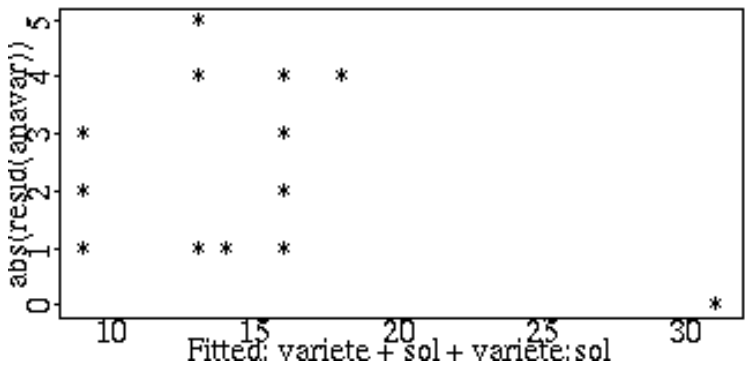
```
Single term deletions
Model:
jgerm ~ variete + sol + variete:sol
            Df Sum of Sq      RSS  F Value      Pr(F)
     <none>              120.0000
    variete  2  192.1277 312.1277 7.204787 0.01354629
        sol  1  123.7714 243.7714 9.282857 0.01386499
variete:sol  2  222.7660 342.7660 8.353723 0.00888845
```

On conclut que l'interaction est significative (avec un risque d'erreur inférieur à 1%). Il existe donc un effet significatif de chacun de ces deux facteurs, puisque l'effet variété dépend du sol, et réciproquement, l'effet sol dépend de la variété.
Le graphique obtenu par "plot.design" indique que le type de sol 1 associé à la variété 1 favorise une germination rapide, alors que le type de sol 2 associé à la variété 2 ralentit la germination.

b) résultats graphiques

```
> plot(anavar)
```



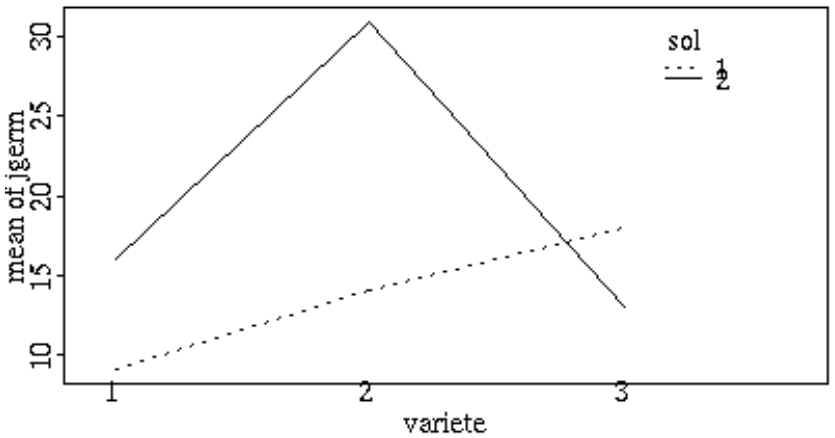


A partir de ces graphiques, on peut valider les postulats. On ne remarque aucune structure des résidus, leur variance semble homogène (elle ne dépend pas des valeurs ajustées), aucun résidu n'est particulièrement grand.
Les postulats sont vérifiés, le modèle est adéquat. Le nombre de jours avant germination des graines de carotte dépend de la variété et du type de sol, ces deux effets étant dépendants l'un de l'autre.

IV Pour aller plus loin . . .

Interaction :

```
> attach(carotte)
> interaction.plot(variete,sol,jgerm)
> detach()
```



L'interaction est significative, il est intéressant de faire cette représentation graphique. Elle est une aide pour l'interprétation de l'interaction entre les deux facteurs.
Ici, on peut vérifier que le type de sol 1 associé à la variété 1 donne un nombre de jours avant germination faible, et que le type de sol 2 associé à la variété 2 donne, au contraire, un nombre de jours avant germination plutôt élevé. Les autres combinaisons des deux facteurs variété et sol, donnent sensiblement le même nombre de jours avant germination.

Moyennes brutes et moyennes ajustées :

On base en général l'interprétation sur les moyennes ajustées.

Dans les premiers traitements, la calcul des moyennes ajustées a été fait "manuellement", à l'aide des fonctions "apply", "tapply" et "mean". Il existe une façon plus directe d'obtenir ces **moyennes ajustées**, il suffit d'utiliser la fonction "dummy.coef".

```
> effets_dummy.coef(anavar)
$"(Intercept)":
 (Intercept)
    16.83333

$variete:
         1        2         3
 -4.333333 5.666667 -1.333333

$sol:
         1        2
 -3.166667 3.166667

$"variete:sol":
         11         21        31         12        22         32
 -0.3333333 -5.333333 5.666667 0.3333333 5.333333 -5.666667

> effets[[1]]+effets[[2]]
    1    2    3
 12.5 22.5 15.5
> effets[[1]]+effets[[3]]
        1  2
 13.66667 20
```

```
> effets[[1]]+outer(effets[[2]],effets[[3]],"+")+effets[[4]]
   1  2
1  9 16
2 14 31
3 18 13
```

Remarquons qu'en sommant les effets obtenus avec "dummy.coef", on obtient bien les moyennes ajustées, autrement dit les moyennes de moyennes, parce que les effets sont calculés avec le modèle interactif complet. Ceci est toujours vrai pour plus de deux facteurs, à condition d'évaluer les effets à partir du modèle interactif complet.

Les **moyennes brutes**, elles, peuvent être obtenues à l'aide de la fonction "model.tables" qui admet comme argument un objet de classe "aov". Il est donc nécessaire d'estimer à nouveau le modèle avec la fonction aov (aov comme Analysis Of Variance).

```
> anavar_aov(jgerm~variete*sol,data=carotte)
> model.tables(anavar,type="means")
Refitting model to allow projection
Tables of means
Grand mean
 15

 variete
     1     2  3
    13 19.67 15
rep  7  3.00  5

 sol
        1     2
    12.52 17.17
rep  7.00  8.00
 variete:sol
Dim 1 : variete
Dim 2 : sol
     1  2
1    9 16
rep  3  4
2   14 31
rep  2  1
3   18 13
rep  2  3
```

Les moyennes brutes sont aussi des moyennes de moyennes pondérées par les effectifs.

La moyenne brute calculée pour la variété 3 vaut 15, elle peut être obtenue de deux façons différentes :

$$15 = (14 + 22 + 18 + 9 + 12)/5 = \frac{2.\left(\frac{14+22}{2}\right) + 3.\left(\frac{18+9+12}{3}\right)}{5}$$

La moyenne ajustée pour la variété 3 vaut 15.5, elle est obtenue de la façon suivante :

$$15.5 = \frac{\left(\frac{14+22}{2}\right) + \left(\frac{18+9+12}{3}\right)}{2}$$

Lorsque le déséquilibre des nombres de répétitions est du au hasard, ou lorsque qu'il ne correspond pas à un choix délibéré de pondération des traitements, la pondération des moyennes par les effectifs ne se justifie pas. C'est la raison pour laquelle on conseille en général de baser les interprétations sur les moyennes ajustée. Une interprétation basée sur les moyennes brutes ne se justifie que dans le cas où on souhaite pondérer les traitements en fonction du nombre de répétitions.

Bibliographie

- Analysis of messy data, vol 1,Milleken & Johnson. *(théorique avec des exemples).*
- Linear Models for Unbalanced Data, S.R. Searle, (1987), New-York : John Wiley & Sons. *.(théorique)*
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.

L'ANALYSE DE COVARIANCE

Résumé des principales commandes :

```
> plot.factor(rats,character="o")
> plot.design(rats)
> anacov_lm(thyroide~regime+
                    I(corporel-mean(colporel)),rats)
> drop1.aov(anacov,scope=anacov$call)
> plot(anacov)
```

I Le problème

1) Les données

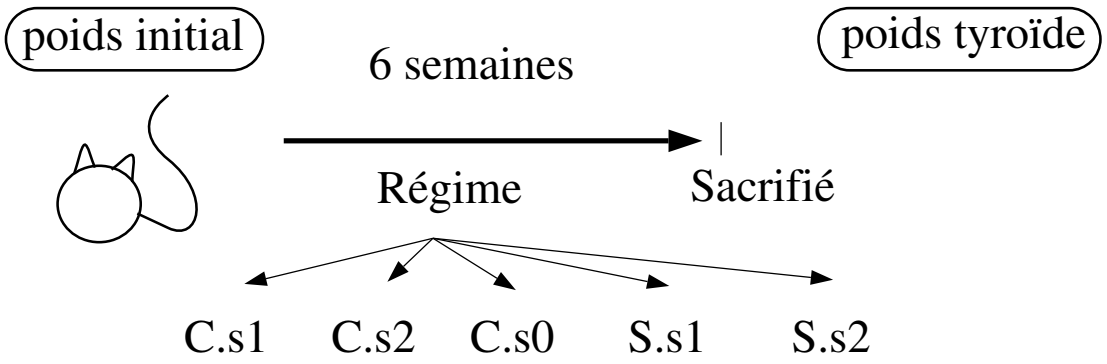
Les données ont été recueillies à l'INRA, par Sylvie Rabot (UEPSD).

On étudie l'effet de deux régimes alimentaires, l'un à base de colza, l'autre à base de soja, sur le développement de la thyroïde du rat. Les régimes alimentaires sont associés ou pas à une souche bactérienne (2 types de souches bactériennes S1 et S2) inoculée par voie orale. On mesure le poids de la thyroïde de 24 rats, abattus après avoir été soumis à l'un des régimes, pendant 6 semaines.

On appellera régime alimentaire chacune des 5 combinaisons suivantes :

- colza associé à la souche 1 (C.s1),
- colza associé à la souche 2 (C.s2),
- colza sans souche (C.s0),
- soja associé à la souche 1 (S.s1),
- soja associé à la souche 2 (S.s2).

L'échantillon de rats a été constitué par tirage au sort, dans une population de rats aussi homogène que possible (même âge, même sexe). Le poids initial des rats varie peu. Cependant, des expériences précédentes ont montré que le poids de la thyroïde est lié au poids total du rat. Le poids de la thyroïde après le régime est donc dépendant du poids corporel initial du rat. Il faudra tenir compte du poids initial pour étudier l'effet du régime.



Les données recueillies sont les suivantes :

poids initial corporel	poids de la thyroide	régime alimentaire
292,1	11,3	C.s1
293,0	14,0	C.s1
274,9	11,7	C.s1
286,9	13,7	C.s1
302,6	14,3	C.s1
277,0	13,5	C.s1
304,1	15,6	C.s1
275,6	12,1	C.s2
289,9	13,0	C.s2
289,3	13,1	C.s2
293,3	13,2	C.s2
296,7	14,2	C.s2
299,2	11,9	C.s2
289,7	13,6	C.s2
308,7	11,2	C.s0
268,0	10,6	C.s0
294,3	10,7	C.s0
285,5	10,8	S.s1
265,0	7,2	S.s1
269,6	9,8	S.s1
293,5	8,8	S.s1
312,8	9,1	S.s2
274,4	7,5	S.s2
298,7	8,9	S.s2

2) Choix de l'analyse

On cherche à expliquer le poids de la thyroïde par le régime alimentaire, et à introduire une correction qui tienne compte de la diversité des poids initiaux. Le poids de la thyroïde est mesuré, c'est la variable expliquée. Le régime alimentaire est fixé, c'est un facteur de variation à 5 niveaux. Le poids initial corporel est mesuré. Il ne dépend pas du régime alimentaire étudié : d'une part il est mesuré avant de soumettre les rats aux régimes, d'autre part le tirage aléatoire des rats pour constituer l'échantillon et pour affecter un des régimes à chaque rat assure que le domaine de variation des poids initiaux ne dépende pas du régime. Il a vraisemblablement une incidence sur le poids de la thyroïde, cette incidence peut être supposée identique quel que soit le régime (effet linéaire et égalité des pentes). C'est une covariable. On cherche à expliquer une variable à l'aide d'un facteur et d'une covariable. Nous allons mettre en œuvre une analyse de covariance.

On peut réaliser une analyse de covariance lorsque :

- la relation qui lie la variable expliquée à la covariable est linéaire,
- le domaine d'observation de la covariable est le même quel que soit le niveau du facteur,
- on a égalité des pentes des droites qui lient covariable et variable expliquée, pour les différents niveaux du facteur.

II Premiers traitements

1) Saisie

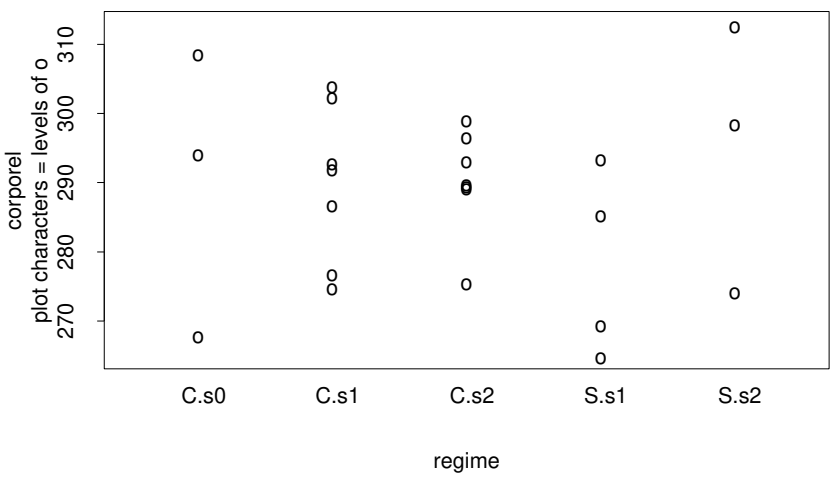
On saisit les données sous forme d'une matrice à trois colonnes. Le data-frame "rats" est composé de la variable "corporel", de la variable "thyroïde", et du facteur "regime".

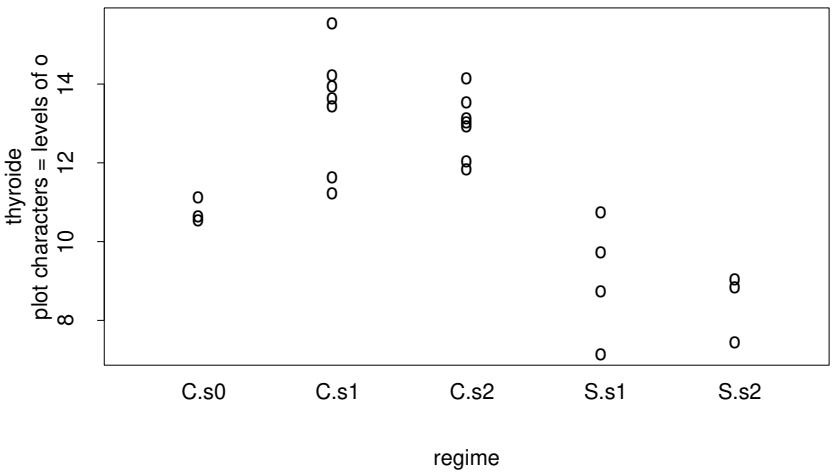
```
> rats_matrix(c(
        292.1,11.3,"C.s1",
        293.0,14.0,"C.s1",
        274.9,11.7,"C.s1",
        286.9,13.7,"C.s1",
        302.6,14.3,"C.s1",
        277.0,13.5,"C.s1",
        304.1,15.6,"C.s1",
        275.6,12.1,"C.s2",
        289.9,13.0,"C.s2",
        289.3,13.1,"C.s2",
```

```
          293.3,13.2,"C.s2",
          296.7,14.2,"C.s2",
          299.2,11.9,"C.s2",
          289.7,13.6,"C.s2",
          308.7,11.2,"C.s0",
          268.0,10.6,"C.s0",
          294.3,10.7,"C.s0",
          285.5,10.8,"S.s1",
          265.0,7.20,"S.s1",
          269.6,9.80,"S.s1",
          293.5,8.80,"S.s1",
          312.8,9.10,"S.s2",
          274.4,7.50,"S.s2",
          298.7,8.90,"S.s2"),ncol=3,byrow=T)
> rats_data.frame(corporel=as.numeric(rats[,1]),
                  thyroide=as.numeric(rats[,2]),
                    regime=as.factor(rats[,3]))
```

2) Visualisation graphique

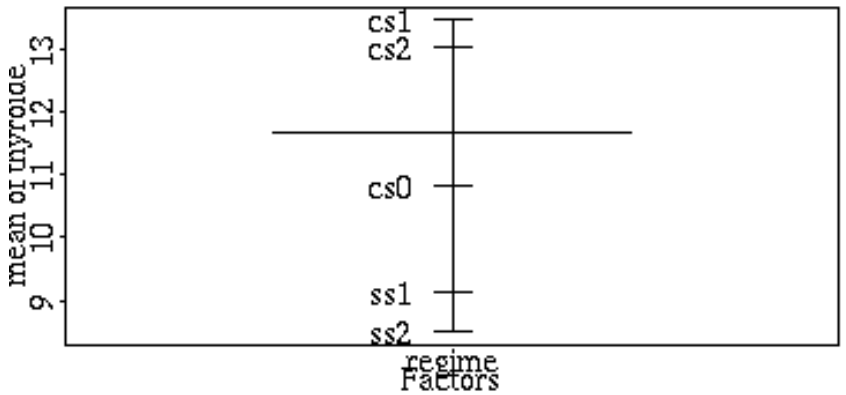
```
> plot.factor(rats,character="o")
```





On vérifie que les poids corporels initiaux ne dépendent pas du régime. Par contre, le poids de la thyroïde semble être affecté par le régime alimentaire.

```
> plot.design(rats)
```



La commande “plot.design(rats)” affiche deux graphiques, les poids corporels moyens par régime d’une part, les poids thyroïdiens moyens par régime d’autre part. Seul le second graphique est intéressant. On peut n’obtenir que le second graphique en tapant “attach(rats); plot.design(thyroide˜regime); detach()”.

III Analyse de la covariance

1) Le modèle et sa mise en œuvre

Le modèle étudié est le suivant : $Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - X_{..}) + \varepsilon_{ij}$

i : indice de régime,
j : indice de rat,
Y_{ij} : poids de la thyroïde du rat j soumis au régime i,
μ : moyenne générale,
α_i : effet du régime alimentaire,
$\beta(X_{ij} - X_{..})$: effet du poids corporel initial avec :

- β : pente de la régression entre Y et X,
- X_{ij} : poids corporel initial du rat,
- $X_{..}$: poids corporel initial moyen,

ε_{ij} : erreurs indépendantes, d'espérance nulle, telles que $var(\varepsilon_{ij}) = \sigma^2$.

On centre la covariable X_{ij} de manière à estimer des paramètres β et μ indépendants l'un de l'autre.

```
> options(contrasts=c("contr.sum","contr.sum"))
> anacov_lm(thyroide~regime+I(corporel-mean(corporel)),rats)
```

Le centrage de la variable "corporel" est indiqué lors de la description du modèle (une alternative aurait été la création d'une nouvelle variable "corporel.centree" dans le data-frame "rats"). La fonction I (comme identity) permet à Splus d'interpréter le "-" comme une opération de soustraction brute, et non pas comme il est interprété habituellement dans une formule, c'est à dire pour supprimer un terme du modèle.

Ce modèle suppose que les pentes des droites de régression entre la variable et la covariable sont égales. Autrement dit, il n'y a pas d'interaction entre le régime et le poids corporel. C'est le modèle d'analyse de covariance.

On peut aussi décrire un modèle avec des pentes qui diffèrent suivant le régime; il suffit d'ajouter un terme d'interaction entre le facteur régime et la covariable. Mais il ne s'agit plus d'un modèle d'analyse de covariance, il s'agit d'un problème de comparaison de droites de régression.

2) Visualisation des résultats

a) résultats numériques

Le plan est non orthogonal, on utilise les sommes de carrés de type III.

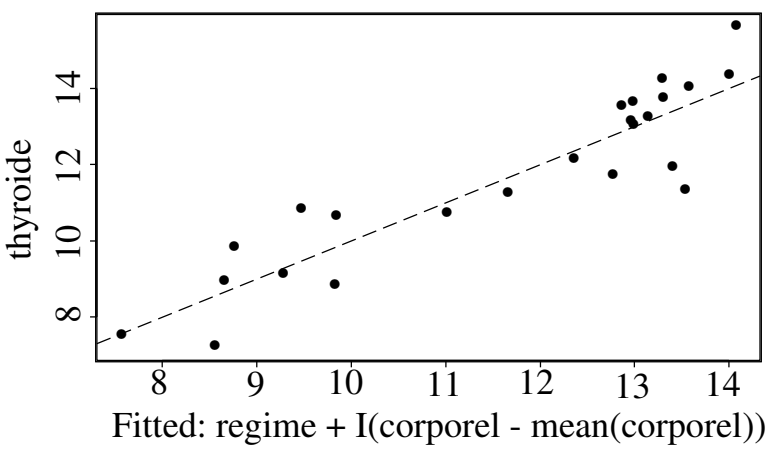
> drop1.aov(anacov,scope=anacov$call)

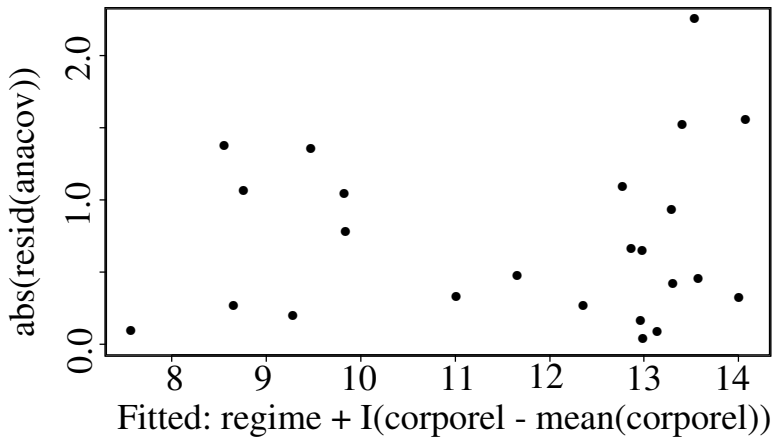
```
Single term deletions

Model:
thyroide ~ regime + I(corporel - mean(corporel))
                              Df Sum of Sq      RSS
                      <none>               19.5874
                      regime   4  86.60197 106.1893
I(corporel - mean(corporel))   1   6.47501  26.0624
                               F Value      Pr(F)
                      <none>
                      regime 19.89593 0.00000206
I(corporel - mean(corporel))  5.95028 0.02529513
```

On conclut qu'il existe un effet régime significatif, et que la correction apportée par la prise en compte du poids corporel initial est utile.

b) résultats graphiques

> plot(anacov)





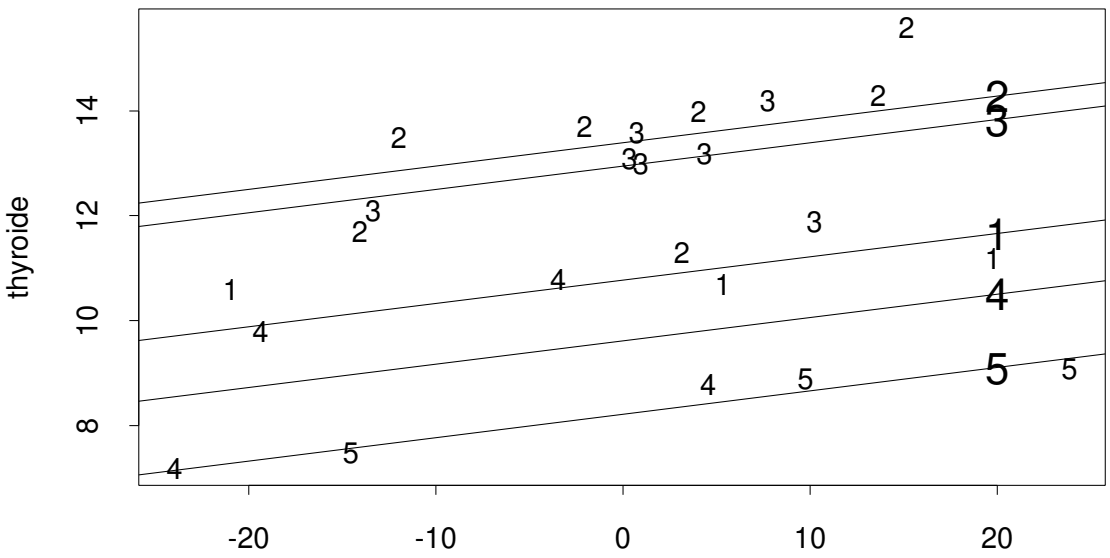
On ne détecte aucun résidu particulièrement grand, aucune structure particulière des résidus, et la variance semble homogène. Les postulats sont vérifiés, le modèle est adéquat. Le régime auquel les rats sont soumis a donc un effet sur le poids de la thyroïde qui varie également en fonction du poids corporel initial.

On représente graphiquement le modèle estimé (droites parallèles), et les observations.

```
attach(rats)
corporel0_corporel-mean(corporel)
coefs_dummy.coef(anacov)
```

La fonction “dummy.coef” permet d’obtenir tous les paramètres estimés, y compris ceux que l’on peut déduire des autres paramètres à l’aide des contraintes.

```
plot(corporel0,thyroide,type="n")
text(corporel0,thyroide,regime)
lapply(as.list(coefs[[1]]+coefs[[2]]),abline,coefs[[3]])
text(20,coefs[[1]]+coefs[[2]]+20*coefs[[3]],lwd=3,cex=1.5)
```



Cette représentation a l'avantage de présenter les données, elle permet donc au lecteur de s'assurer que les conditions d'utilisation de l'analyse de covariance semblent respectées; elle présente aussi le modèle ce qui permet d'interpréter.

IV Pour aller plus loin . . .

On a exprimé le poids de la thyroïde à l'aide d'un modèle d'analyse de covariance, ce modèle impose l'égalité des pentes. Graphiquement, il semble que l'hypothèse d'égalité des pentes soit acceptable. Cependant, cette hypothèse peut être testée, en comparant le modèle d'analyse de covariance à un modèle plus général où les pentes varient suivant le régime.

Soient M le modèle général et M_0 le sous-modèle (analyse de covariance) :

$$M : \quad Y_{ij} = \mu + \alpha_i + \beta_i(X_{ij} - X_{..}) + \varepsilon_{ij} \quad (droites)$$
$$M_0 : \quad Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - X_{..}) + \varepsilon_{ij} \quad (anacov)$$

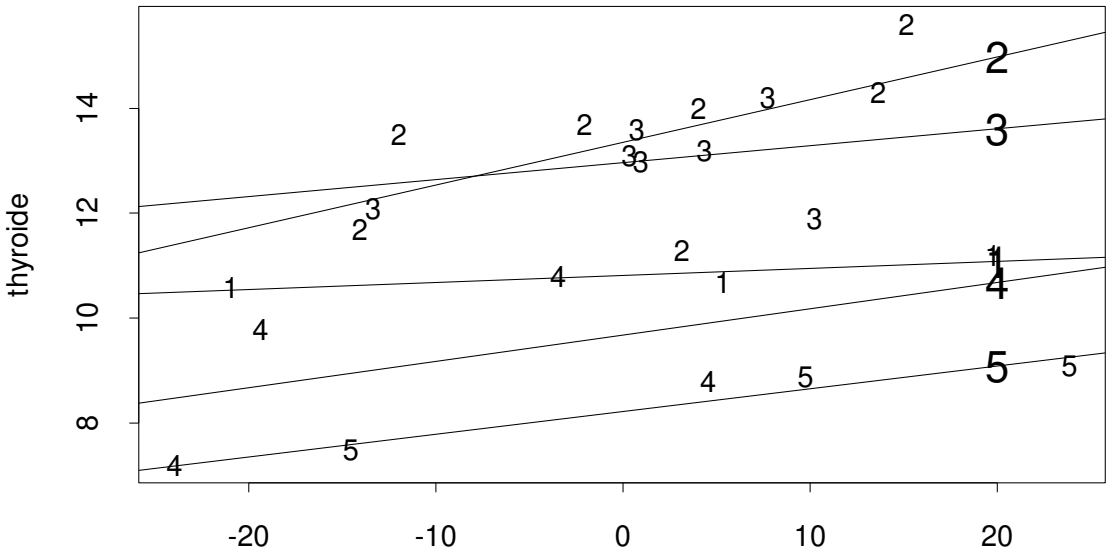
La comparaison du modèle M_0 au modèle M par un test de sous-modèle revient à tester l'hypothèse nulle d'égalité des pentes :

$$H_0 : \qquad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$$

```
> droites_lm(thyroide~regime+
                regime:I(corporel-mean(colporel)),rats)
```

On représente graphiquement le modèle estimé.

```
coefs_dummy.coef(droites)
coefs_cbind(coefs[[1]]+coefs[[2]],coefs[[3]])

plot(corporel0,thyroide,type="n")
text(corporel0,thyroide,regime)
apply(coefs,1,abline)
text(20,coefs[,1]+20*coefs[,2],lwd=3,cex=1.5)

detach()
```



On réalise le test de sous-modèle à l'aide de la fonction "anova".

```
> anova(anacov,droites)
```

```
Analysis of Variance Table

Response: thyroide

                                                Terms Resid.Df
1 regime + regime:I(corporel - mean(corporel))      14
2        regime + I(corporel - mean(corporel))      18
       RSS    Test Df Sum of Sq   F Value     Pr(F)
1 17.64604
2 19.58737 1 vs. 2 -4 -1.941326 0.3850519 0.8157461
```

RSS : residual sum of square : somme de carrés résiduelle.

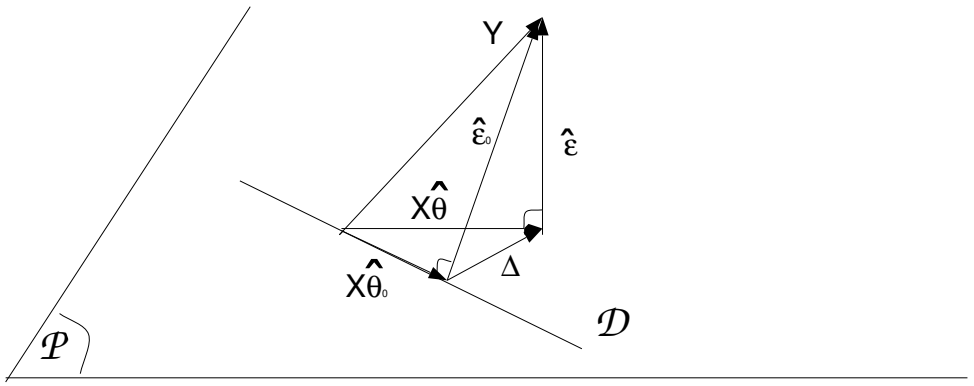
La différence des nombres de degrés de liberté entre M et M_0 vaut 4. La différence des sommes de carrés résiduelles vaut 1.9, et la statistique de Fisher calculée vaut 0.39. Le risque d'erreur si on rejette Ho, est supérieur à 80%. On accepte donc l'hypothèse nulle. Le choix du modèle d'analyse de covariance est justifié.

Représentation géométrique :

Soit P l'espace vectoriel de dimension p représentant le modèle général M.

Soit D le sous-espace vectoriel de dimension q représentant le sous-modèle M_0.

Le projeté de $\vec{Y}$ sur P est le vecteur $X\vec{\hat{\Theta}}$ des valeurs estimées obtenues avec le modèle général; en projetant $\vec{Y}$ sur D, on obtient le vecteur $X\vec{\hat{\Theta}}_0$ des valeurs estimées obtenues avec le sous-modèle M_0. On appelle Δ la différence entre les deux vecteurs des estimés obtenus (c'est aussi la différence entre les deux vecteurs de résidus). Δ mesure l'écart entre les modèles M et M_0.



$$\Delta = SCR_{M_0} - SCR_M$$

Le test consiste à comparer Δ à la variance estimée par le carré moyen résiduel du modèle général. On calcule la statistique suivante :

$$F = \frac{\frac{SCR_{M_0} - SCR_M}{p-q}}{\frac{SCR_M}{n-p-1}}$$

Sous Ho, F suit une loi de Fisher-Snedecor à (p-q) et (n-p-1) degrés de liberté.

Bibliographie

- Méthodes statistiques, Snedecor & Cochran (1967), ACTA.
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.

COMPARAISON DE DROITES DE REGRESSION

Résumé des principales commandes :

```
> lm0_lm(temps~-1+systeme+systeme:debit,pression)
> anova(lm0)
> plot(lm0)
> lm4_lm(temps~-1+systeme1+systeme1:debit,pression)
> anova(lm0,lm4)
> anova(lm4)
> plot(lm4)
```

I Le problème

1) Les données

Lors d'une perfusion faite à l'aide d'un pousse seringue, l'occlusion de la ligne de perfusion interrompt l'administration du produit. La pression dans le système de perfusion augmente alors, jusqu'à atteindre un niveau qui met en alarme le perfuseur et le stoppe. On étudie le délai écoulé entre l'occlusion et le déclenchement de l'alarme. Il dépend du système de perfusion et du débit.

Les systèmes de perfusion expérimentés sont les suivants :

- une seringue seule équipée d'un robinet à trois voies (S1),
- le système précédent avec en plus, une tubulure souple en PVC de 2.3 ml de volume (S2),
- le système précédent avec en plus, la présence d'une bulle d'air de 1 ml dans la seringue (S3),
- une seringue équipée d'un robinet à trois voies et d'une tubulure rigide en polyéthylène de 1.1 ml de volume (S4).

Les débits auxquels sont soumis les systèmes sont : 2.5, 5.0, 7.5, 10.0, 12.5, 15.0, 20.0 et 25.0 ml/h.
On étudie le temps qui s'écoule avant le déclenchement de l'alarme (délai en secondes). Pour chaque couple (système, débit) trois observations sont réalisées.

délai		systèmes			
		S1	S2	S3	S4
débit en ml/h	2.5	1970 1990 1995	2775 2790 2785	3460 3440 3455	1965 1985 1990
	5.0	995 985 1000	1350 1365 1410	1685 1665 1665	970 945 980
	7.5	640 606 615	900 905 915	1125 1110 1130	615 600 635
	10.0	440 448 443	685 690 675	820 835 845	435 430 445
	12.5	358 355 354	560 580 575	665 680 685	355 365 345
	15.0	297 296 299	490 530 500	585 595 600	290 300 290
	20.0	240 240 240	345 335 340	465 460 465	225 255 260
	25.0	190 190 190	275 270 275	390 375 375	195 200 185

2) Choix de l'analyse

Le délai de déclenchement de l'alarme est mesuré, c'est la variable expliquée. Le système de perfusion est un facteur à quatre niveaux. Le débit peut être considéré comme un facteur à huit niveaux, ou comme un régresseur : c'est une variable quantitative.

Pour un système de perfusion donné, le délai est lié linéairement à l'inverse du débit. On peut donc modéliser l'ensemble des données par un ensemble de droites de régression : on estime une ordonnée à l'origine et une pente pour chaque système de perfusion.

II Premiers traitements

1) Saisie

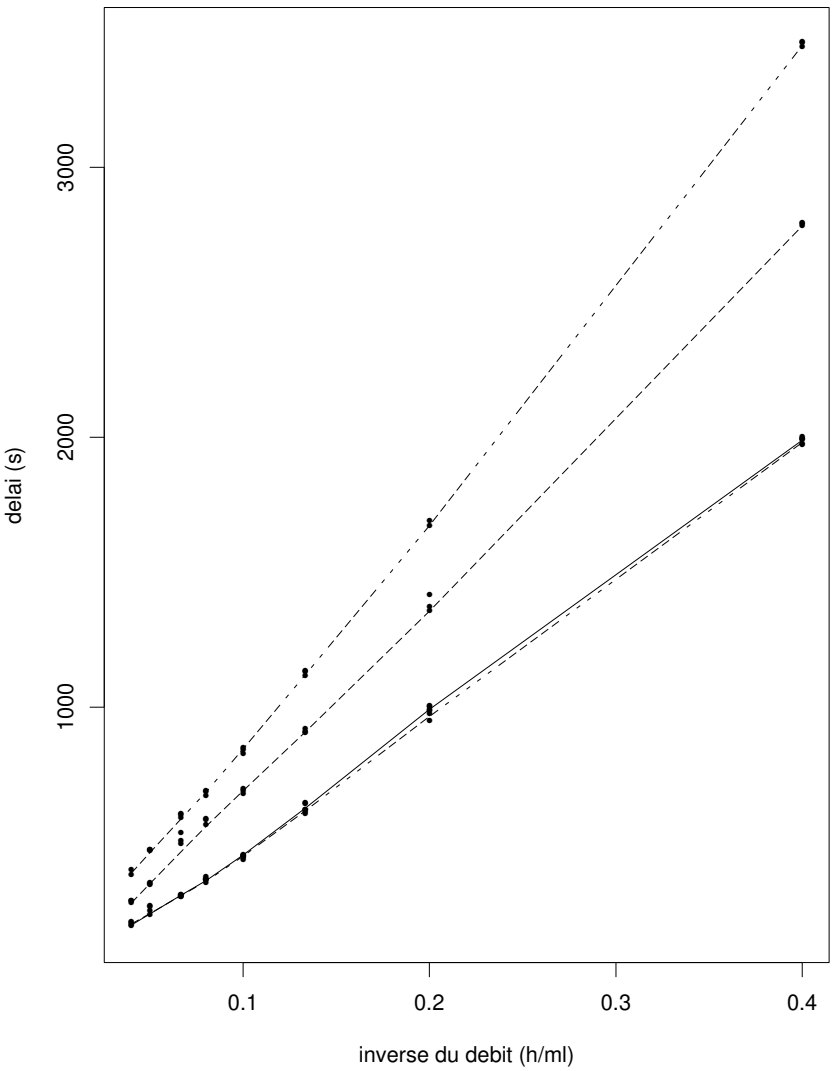
Le data-frame “pression” est créé à partir du fichier unix “data1”. La variable “debit” contient en fait les inverses des débits.

```
> data1_matrix(c(scan("data1")),ncol=3,byrow=T)
> pression_data.frame(debit=1/data1[,2],
                      delai=data1[,1],
                      systeme=as.factor(data1[,3]))
```

2) Visualisation graphique

Pour chaque système de perfusion utilisé, on représente le délai observé en fonction de l’inverse du débit, et on superpose un lissage obtenu à l’aide de la fonction lowess.

```
plot(debit,delai,xlab="inverse du debit (h/ml)",
                 ylab="delai (s)")
for (i in unique(systeme))
{syst_systeme==i
 lines(lowess(debit[syst],delai[syst]),lty=i)}
```



Le délai de déclenchement de l'alarme semble être une fonction linéaire de l'inverse du débit. Les relations linéaires semblent être différentes d'un système de perfusion à l'autre, sauf pour les systèmes S1 et S4. La comparaison des droites de régression permet de tester la similitude entre les systèmes.

III Analyse : comparaison de droites de régression

1) Le modèle général

$$Y_{ij} = \alpha_i + \beta_i \left(\frac{1}{debit} \right) + \varepsilon_{ij}$$

i : indice de système,
j : indice de répétition,
Y_{ij} : délai de déclenchement de l'alarme,
α_i : ordonnées à l'origine des droites,
β_i : pentes des droites,
$\frac{1}{debit}$: inverse du débit,
ε_{ij} : erreurs indépendantes, d'espérance nulle, telles que $var(\varepsilon_{ij}) = \sigma^2$.

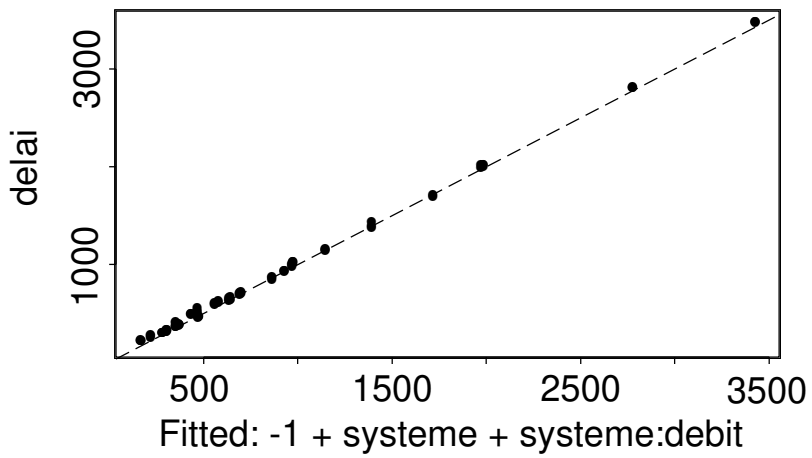
Le modèle ne comprend pas de terme constant.

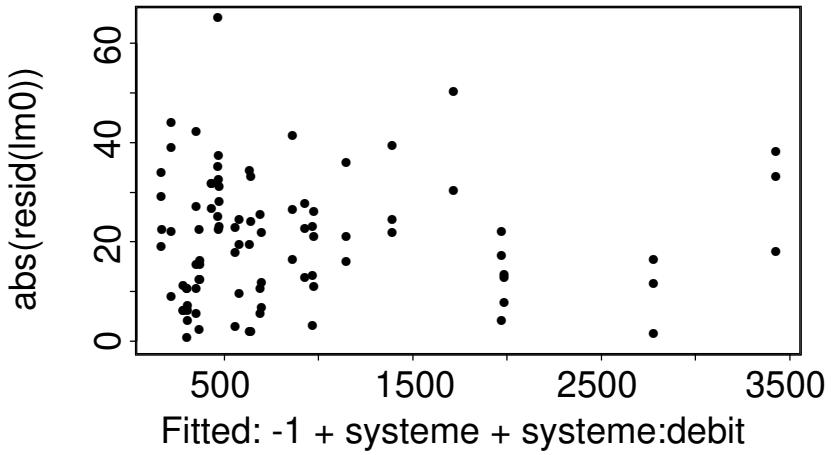
```
> options(contrasts=c("contr.sum","contr.sum"))
> lm0_lm(delai~-1+systeme+systeme:debit,pression)
> anova(lm0)
```

```
Analysis of Variance Table
Response: delai
Terms added sequentially (first to last)
               Df Sum of Sq  Mean Sq  F Value Pr(F)
systeme         4  71993833 17998458 29373.83     0
systeme:debit   4  51202691 12800673 20890.94     0
Residuals      88     53921      613
```

Le modèle apporte une explication significative. On rejette l'hypothèse d'égalité des droites.

```
> plot(lm0)
```





On ne détecte pas de structure particulière des résidus, le modèle est adéquat. Voici les estimations des paramètres :

```
> coef(lm0)
  systeme1 systeme2 systeme3  systeme4 systeme1debit
-33.53225 3.643386 6.878349 -33.61601     5040.74
  systeme2debit   systeme3debit   systeme4debit
    6926.031        8538.916        5005.229
```

Les paramètres des systèmes S1 et S4 sont très proches. On teste le sous-modèle correspondant à l'hypothèse d'égalité des droites pour les systèmes S1 et S4, et d'autres sous-modèles correspondant à d'autres hypothèses.

2) Les sous-modèles

a) Principe du test de sous-modèle

On peut être amené à tester un sous-modèle par rapport à un modèle plus général, lorsque se pose un problème de comparaison de droites de régression, ou pour tester l'effet de facteurs et/ou de régresseurs simultanément, ou encore, pour trouver un modèle plus simple, débarrassé des régresseurs difficiles (ou couteux) à mesurer. On doit d'abord s'assurer de la validité du modèle général (pas de structure dans les résidus).

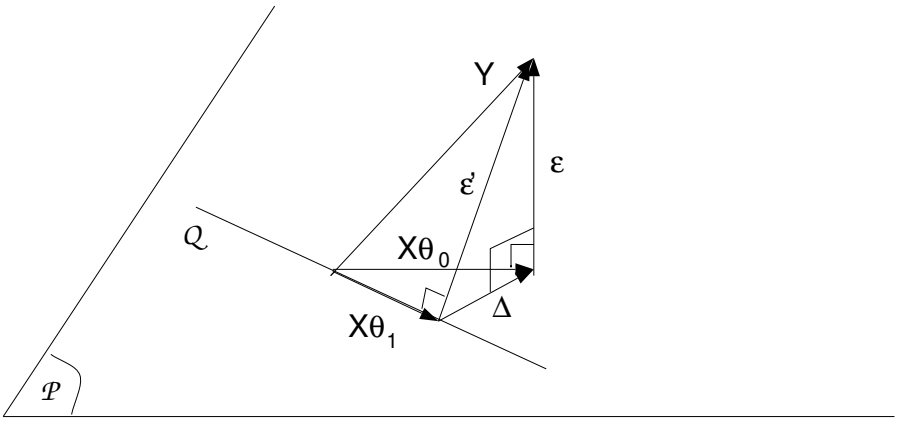
Soit le modèle général suivant, correspondant à des droites de régression qui auraient une pente et une ordonnée à l'origine toutes différentes. On suppose que ce modèle a été validé :

$$M_0 : \qquad Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij}$$

Pour tester l'hypothèse d'égalité des ordonnées à l'origine, on pose le sous-modèle suivant :

$$M_1 : \qquad Y_{ij} = \alpha + \beta_i X_{ij} + \varepsilon_{ij}^{'}$$

ε' contient ε plus la variablité expliquée par des ordonnées à l'origine différentes.
On pose Ho : "le modèle général et le sous-modèle sont équivalents".
Soit P, l'espace vectoriel de dimension p associé au modèle général, et Q, le sous-espace vectoriel de dimension q associé au sous-modèle.



Δ représente la différence entre les variabilités expliquées par les deux modèles. Si Δ est du même ordre de grandeur que la résiduelle du modèle général (aux degrés de liberté près), le sous-modèle est adéquat.

On calcule la statistique suivante :

$$F = \frac{\frac{SCR_{SM} - SCR_{MG}}{p-q}}{\frac{SCR_{MG}}{n-p-1}}$$

Sous l'hypothèse nulle d'équivalence entre les deux modèles, F suit une loi de Fisher-Snedecor à (p-q) et (n-p-1) degrés de liberté. On lit la valeur critique correspondante dans la table de Fisher, pour un risque α donné. Si la statistique calculée est supérieure à la valeur critique lue, on rejette Ho. Les deux modèles ne sont pas équivalents, le sous-modèle ne convient pas. Si la statistique calculée est inférieure à la valeur critique lue, on accepte Ho, les deux modèles sont équivalents.

b) Application

On teste la validité de chacun de ces quatre sous-modèles :

- M_1 : égalité des droites,
- M_2 : même ordonnée à l'origine, des pentes différentes,
- M_3 : même pente, des ordonnées à l'origine différentes,
- M_4 : droite S_1 identique à S_4, les autres différentes.

sous-modèle 1 : égalité des droites.

$$Y_{ij} = \alpha + \beta \frac{1}{debit} + \varepsilon_{ij}$$

```
> lm1_lm(delai~debit,pression)
> anova(lm0,lm1)
```

```
Analysis of Variance Table

Response: delai

                          Terms Resid. Df     RSS
1 -1 + systeme + systeme:debit        88   53921
2                         debit        94 7075830
     Test Df Sum of Sq  F Value Pr(F)
1
2 1 vs. 2 -6  -7021909 1909.981     0
```

Le sous-modèle n'est pas équivalent au modèle général, on rejette l'hypothèse nulle d'égalité des droites.

sous-modèle 2 : Toutes les droites ont même ordonnée à l'origine, mais des pentes différentes.

$$Y_{ij} = \alpha + \beta_i \left(\frac{1}{debit} \right) + \varepsilon_{ij}$$

```
> lm2_lm(delai~systeme:debit,pression)
> anova(lm0,lm2)
```

```
Analysis of Variance Table
Response: delai
             Terms Resid.          Df     RSS
1 -1 + systeme + systeme:debit    88   53920.94
2                 systeme:debit    91   68826.74
    Test    Df   Sum of Sq    F Value       Pr(F)
1
2 -systeme  -3    -14905.8   8.108847   7.910158e-05
```

La p-value est très faible, on rejette l'hypothèse nulle. Le sous-modèle n'est pas équivalent au modèle général, il n'est donc pas valide.

sous-modèle 3 : Toutes les droites ont la même pente, mais des ordonnées à l'origine différentes.

$$Y_{ij} = \alpha_i + \beta \frac{1}{debit} + \varepsilon_{ij}$$

```
> lm3_lm(delai~-1+systeme+debit,pression)
> anova(lm0,lm3)
```

```
Analysis of Variance Table
Response: delai
            Terms Resid.            Df     RSS
1 -1 + systeme + systeme:debit     88    53921
2         -1 + systeme + debit     91   2636600
     Test    Df   Sum of Sq   F Value    Pr(F)
1
2   1 vs.2   -3   -2582679    1404.994     0
```

Ici, la p-value est très faible, on rejette l'hypothèse d'équivalence des deux modèles, le sous-modèle ne peut pas être utilisé.

sous-modèle 4 : Les droites correspondant aux systèmes S1 et S4 sont équivalentes, mais différentes de celles correspondantes aux autres systèmes, qui sont elles-même différentes les unes des autres.

$$Y_{ij} = \alpha_i + \beta_i \left(\frac{1}{debit} \right) + \varepsilon_{ij}$$

avec $\alpha_1 = \alpha_4 \quad et \quad \beta_1 = \beta_4$.

On ajoute au data frame “pression” une variable “systeme1”, qui regroupe les systèmes S1 et S4.

```
> pression$systeme1_pression$systeme
> levels(pression$systeme1)_c("1","2","3","1")
> lm4_lm(delai~systeme1+systeme1:debit,pression)
> anova(lm0,lm4)
```

```
Analysis of Variance Table

Response: delai

                          Terms Resid. Df      RSS
1 -1 + systeme + systeme:debit        88 53920.94
2    systeme1 + systeme1:debit        90 54389.69
     Test Df Sum of Sq   F Value     Pr(F)
1
2 1 vs. 2 -2 -468.7478 0.3825027 0.6832806
```

La p-value est élevée, on accepte l’hypothèse nulle. Le sous-modèle équivaut donc au modèle général.

On peut essayer de simplifier encore le modèle en diminuant le nombre de paramètres.

```
> dummy.coef(lm4)
$"(Intercept)":
 (Intercept)
   -7.684131

$systeme1:
      1        2        3
 -25.89 11.32752 14.56248

$"systeme1:debit":
   1debit   2debit   3debit
 5022.984 6926.031 8538.916
```

On regroupe les ordonnées à l'origine, pour S2 et S3.

```
pression$systeme2_pression$systeme
levels(pression$systeme2)_c("1","2","2","1")
lm5_lm(delai~systeme2+systeme1:debit,pression)
```

```
anova(lm0,lm5)

Response: delai

                          Terms Resid. Df      RSS
1 -1 + systeme + systeme:debit        88 53920.94
2    systeme2 + systeme1:debit        91 54441.22
     Test Df Sum of Sq   F Value     Pr(F)
1
2 1 vs. 2 -3  -520.284 0.2830378 0.8375178
```

Le sous-modèle est accepté. Toute autre tentative de regroupement des ordonnées à l'origine, ou des pentes conduit à rejeter le sous-modèle correspondant. lm5 est donc le "plus petit" modèle que l'on puisse trouver. C'est le modèle lm5 qui est interprété.

```
drop1.aov(lm5,scope=lm5$call)
Single term deletions

Model:
delai ~ systeme2 + systeme1:debit
               Df Sum of Sq      RSS  F Value
        <none>                 54441
      systeme2  1     14854    69295    24.83
systeme1:debit  3  51777764 51832205 28849.32
                      Pr(F)
        <none>
      systeme2 2.968879e-06
systeme1:debit 0.000000e+00
```

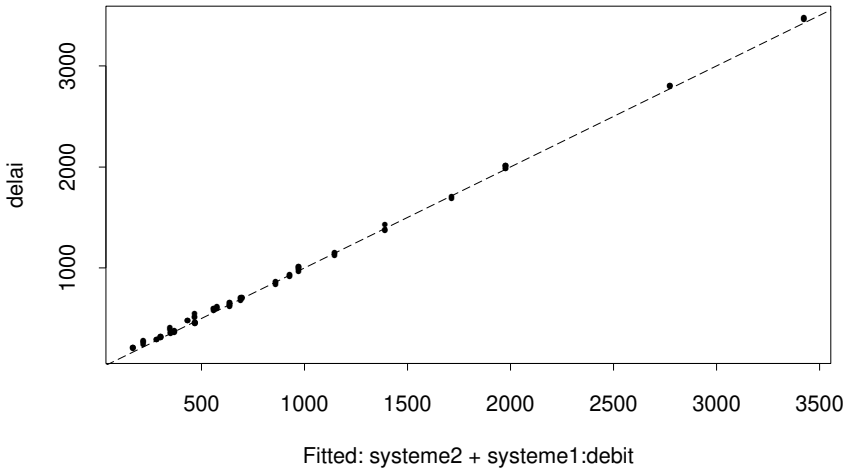
Chacun des termes du modèle est significatif.

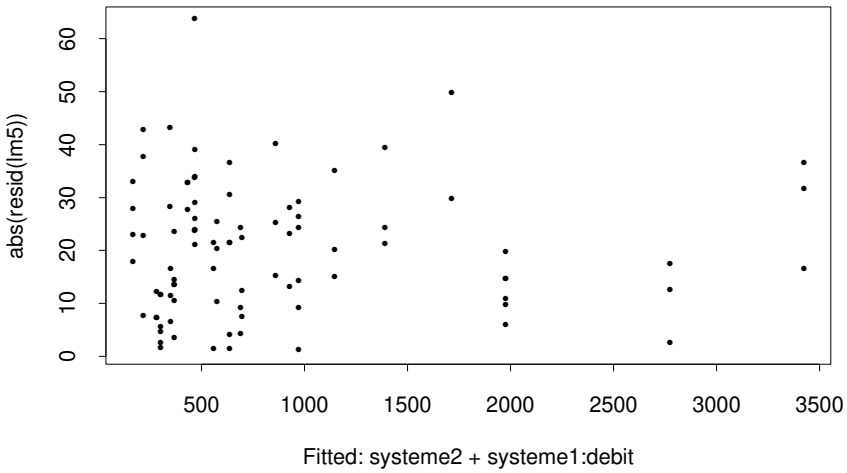
```
dummy.coef(lm5)
$"(Intercept)":
 (Intercept)
   -14.15663

$systeme2:
        1       2
 -19.4175 19.4175

$"systeme1:debit":
 systeme11debit systeme12debit systeme13debit
       5022.984         6918.9       8546.046
```

```
> summary(lm5)

Call: lm(formula = delai ~ systeme2 + systeme1:debit,
 data = pression)
Residuals:
    Min     1Q Median    3Q   Max
 -49.47 -19.98 -3.605 21.53 63.48

Coefficients:
                   Value Std. Error    t value   Pr(>|t|)
   (Intercept)   -14.1566     3.8968    -3.6329     0.0005
      systeme2    19.4175     3.8968     4.9829     0.0000
systeme11debit  5022.9841    31.6386   158.7614     0.0000
systeme12debit  6918.9004    37.5739   184.1414     0.0000
systeme13debit  8546.0462    37.5739   227.4466     0.0000

Residual standard error: 24.46 on 91 degrees of freedom
Multiple R-Squared: 0.999
F-statistic: 23250 on 4 and 91 degrees of freedom,
the p-value is 0

Correlation of Coefficients:
               (Intercept) systeme2 systeme11debit systeme12debit
      systeme2  0.0000
systeme11debit -0.5430      0.5430
systeme12debit -0.4572     -0.4572   0.0000
systeme13debit -0.4572     -0.4572   0.0000         0.4181
```

Les ordonnées à l'origine sont identiques pour S1 et S4 d'une part, S2 et S3 d'autre part. Les pentes sont identiques pour S1 et S4, différentes pour les autres sytèmes.

```
> plot(lm5)
```





Bibliographie

- La régression nouveaux regards sur une ancienne méthode statistique, R. Tomassone, E. Lesquoy, C. Millier, INRA actualités scientifiques et agronomiques, MASSON.
- Applied Regression Analysis, N.R. Draper, H. Smith, Wiley & Sons.
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.

LA REGRESSION LINEAIRE MULTIPLE

Résumé des principales commandes :

```
> pairs(moutarde)
> brush(as.matrix(moutarde),hist=T)
> reg_lm(RC~.,moutarde)
> summary(reg)
> plot(reg)
> qqnorm(resid(reg))
```

I Le problème

1) Les données

Les pieds de moutarde s'adaptent à la sécheresse en développant des racines courtes tubérisées.

On soumet 31 pieds de moutarde à un stress hydrique de 34 jours. On mesure ensuite, pour chacun des pieds :

- le nombre de racines courtes tubérisées,
- la longueur de la tige,
- le potentiel hydrique foliaire,
- le poids de matière sèche des racines,
- le poids de matière sèche des parties aériennes,
- le nombre de feuilles.

On cherche à savoir dans quelle mesure des variables telles que la longueur de la tige ... ont une influence sur le nombre de racines courtes, c'est à dire dans quelle mesure ces variables permettent d'expliquer une plus ou moins bonne adaptation à la sècheresse.

Les données recueillies sont les suivantes :

racines courtes	longueur de la tige	potentiel hydrique foliaire	poids des racines	poids des parties aériennes	nombre de feuilles
0.00	29	65	87	43	2
0.00	35	65	163	122	2
1.10	40	65	175	117	3
0.69	25	60	38	49	2
0.00	30	30	57	23	1
0.00	45	70	270	124	5
0.69	40	65	202	78	4
1.39	50	70	226	74	3
1.61	50	85	525	222	5
1.10	55	80	230	92	3
3.47	60	155	1109	897	4
2.40	80	95	869	628	5
1.10	60	60	553	189	8
3.00	90	100	903	3022	6
3.43	80	145	1216	3049	6
1.61	75	85	912	3273	6
2.83	60	75	689	443	6
1.61	85	85	443	251	5
2.20	65	80	643	424	5
4.09	60	240	1089	843	6
3.09	60	80	825	757	7
1.39	70	80	385	1350	5
4.22	90	180	1335	728	7
4.17	90	175	953	668	3
4.57	95	205	1145	696	6
3.43	75	305	1129	678	6
3.04	70	120	978	529	6
3.26	75	70	795	329	6
5.40	70	300	1618	1075	7
4.16	70	250	1020	881	7
3.91	60	280	1020	624	8

Le nombre de racines courtes est exprimé en log, pour homogénéiser la variance.

2) Choix de l'analyse

Le nombre de racines courtes est la variable à expliquer. Les cinq autres variables sont des variables explicatives continues, ce sont des régresseurs. On met en

œuvre une régression linéaire multiple.

II Premiers traitements

1) Saisie

Les variables sont notées :

RC : nombre de racines courtes tubérisées (en log),
LT : longueur de la tige,
HF : potentiel hydrique foliaire,
PR : poids des racines,
PA : poids des parties aériennes,
FE : nombre de feuilles.

```
> moutarde_matrix(c(scan("moutarde")),
                      ncol=6,byrow=T)
> moutarde_data.frame(RC=moutarde[,1],
                          LT=moutarde[,2],
                          HF=moutarde[,3],
                          PR=moutarde[,4],
                          PA=moutarde[,5],
                          FE=moutarde[,6])
```

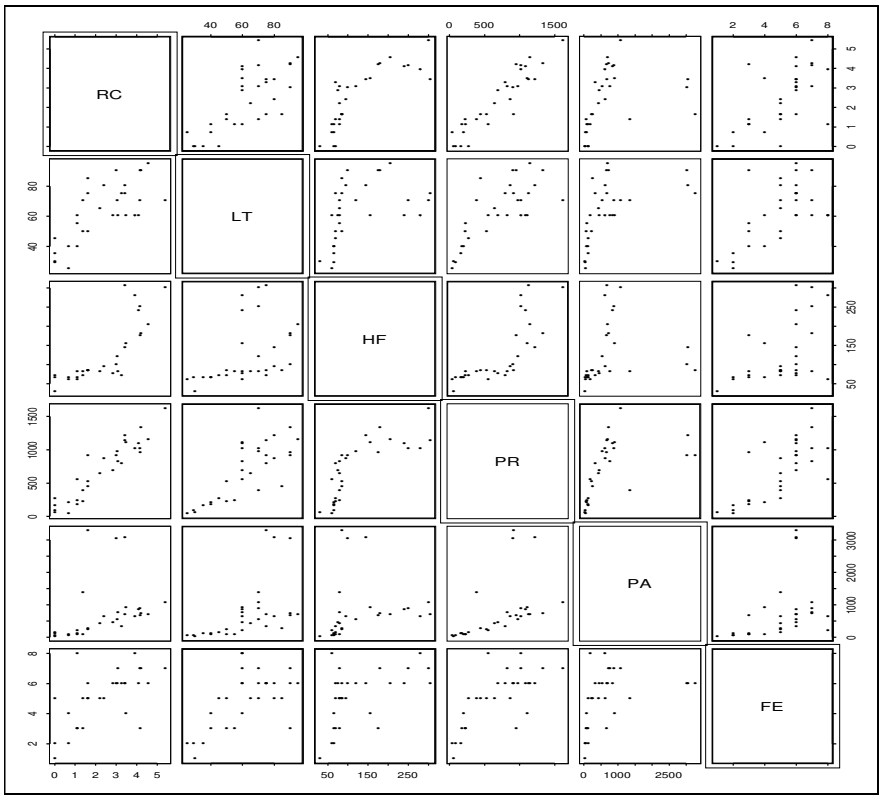
2) Visualisation graphique

Cette étape est très importante. On représente sur un seul graphique toute l'information contenue dans le tableau de données. Tous les nuages de points représentant une variable en fonction d'une autre sont dessinés. Ce type de graphique est un outil indispensable pour "apprendre" les données. Il complète parfaitement la régression multiple. Il permet :

- d'identifier une ou plusieurs observations particulières,
- d'étudier les liaisons entre la variable à expliquer et chacun des régresseurs (présence ou absence de liaison apparente, type de liaison).
- d'étudier les corrélations entre régresseurs.

Ici, la variable RC semble liée à chacun des régresseurs. La liaison entre RC et HF semble quadratique, plutôt que linéaire; on remarque des observations particulières dans le graphique représentant RC en fonction de PA (ces mêmes points se singularisent dans le graphe de PR en fonction de PA : ce sont des plantes qui ont des racines peu développées par rapport aux parties aériennes); on remarque aussi des corrélations entre les régresseurs.

```
> pairs(moutarde)
```

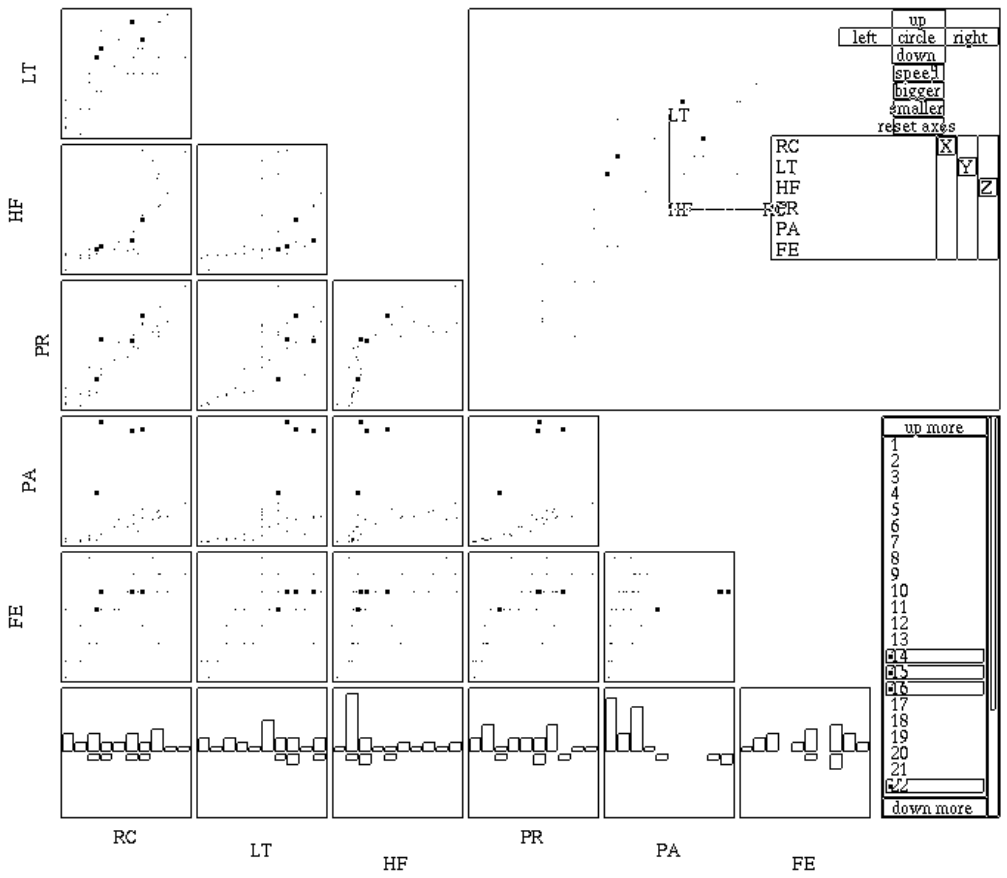


Par rapport à la fonction “pairs”, la fonction “brush” a l’avantage de permettre à l’utilisateur d’interagir avec le graphique : on peut marquer un ensemble de points intéressants, par exemple les plantes dont les parties aériennes sont très développées par rapport aux racines. Ces plantes sont marquées dans chacun des “petits graphiques”. C’est en général un outil puissant pour “comprendre” les données. Ici, les plantes dont les parties aériennes sont très développées sont tout à fait “dans la moyenne” pour les autres variables, et pour les liaisons entre variables, hormis des racines peu développées par rapport aux parties aériennes.

Un histogramme de chacune des variables peut être affiché. Il permet de marquer facilement les points ayant des valeurs particulièrement élevées ou faibles de l’une ou l’autre des variables.

Une représentation d’un triplet de variables au choix, avec rotation interactive du nuage de points est également fournie.

```
> brush(as.matrix(moutarde),hist=T)
```



III Analyse : régression linéaire multiple

1) Le modèle et sa mise en œuvre

$RC = \alpha + \beta_1 LT + \beta_2 HF + \beta_3 PR + \beta_4 PA + \beta_5 FE + \varepsilon$

α : terme constant,
$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$: coefficients associés à chaque variable,
ε : erreurs indépendantes, d'espérance nulle, telles que $var(\varepsilon) = \sigma^2$.

Ce modèle de régression multiple est un modèle additif, comprenant un terme constant et faisant intervenir toutes les variables du data-frame "moutarde". On peut donc résumer l'écriture du modèle à l'aide d'un "point".

```
> options(contrasts=c("contr.sum","contr.sum"))
> reg_lm(RC~.,moutarde)
```

2) Visualisation des résultats

a) résultats numériques

> summary(reg)

```
Call: lm(formula = RC ~ ., data = moutarde)
Residuals:
     Min      1Q  Median    3Q    Max
 -0.8858 -0.3149 0.02299 0.429 0.6658
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -0.4289  0.4089    -1.0490  0.3042
         LT  0.0136  0.0079     1.7106  0.0995
         HF  0.0032  0.0021     1.5385  0.1365
         PR  0.0029  0.0006     4.9673  0.0000
         PA -0.0003  0.0001    -2.1074  0.0453
         FE -0.0547  0.0739    -0.7408  0.4657
Residual standard error: 0.508 on 25 degrees of freedom
Multiple R-Squared: 0.91
F-statistic: 50.54 on 5 and 25 degrees of freedom,
the p-value is 2.861e-12
```

```
Correlation of Coefficients:
   (Intercept)      LT      HF      PR      PA
LT -0.7311
HF -0.3617      0.2611
PR  0.5476     -0.5136 -0.7428
PA  0.0573     -0.1572  0.3099 -0.3460
FE -0.4310     -0.1135  0.0813 -0.4007  0.0467
```

Comme en régression simple, on affiche le “résumé” des résidus, l’estimation des coefficients ainsi qu’un test de nullité pour chacun d’eux, l’estimation de l’écart-type résiduel, le coefficient de détermination, la statistique du test d’explication du modèle et enfin les corrélations entre les paramètres.

Attention à l’interprétation du test de nullité de chacun des paramètres, dans le cas de fortes corrélations : il est préférable de tester le sous-modèle où on a supprimé simultanément les variables corrélées.
La statistique de Fisher est calculée à l’aide de la somme de carrés expliquée par la régression, de la somme de carrés résiduelle, et des degrés de liberté associés.

$$SCreg = \sum \left(\widehat{Y_i} - \overline{Y}\right)^2 \qquad CMreg = \frac{SCreg}{p}$$
$$SCr = \sum \left(Y_i - \widehat{Y_i}\right)^2 \qquad CMr = \frac{SCr}{n-p-1}$$
$$F = \frac{CMreg}{CMr}$$

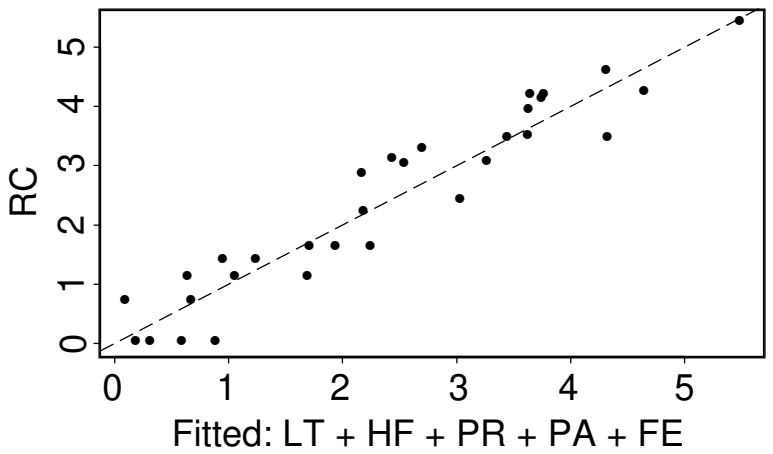
où p est le nombre de régresseurs, et n le nombre d'observations c'est à dire le nombre d'unités expérimentales observées, ou encore le nombre de (p+1) uplets d'observations.

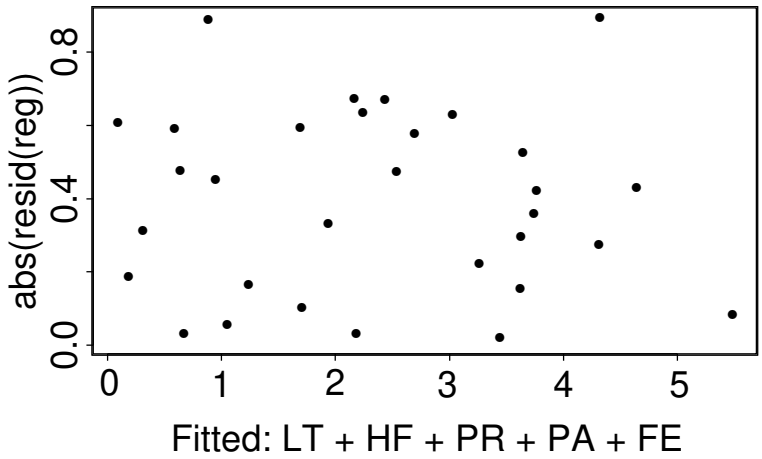
Sous Ho : "$\beta_i = 0, \quad \forall i$", F suit une loi de Fisher à p et (n-p-1) degrés de liberté. Si la statistique calculée est supérieure à la valeur critique lue dans la table de Fisher, on rejette Ho, et on conclut que le modèle apporte une explication significative. Ici, la p-value est très petite, on a très peu de chances de se tromper en rejetant Ho. On conclut que la part d'explication apportée par le modèle est significative. Le modèle obtenu s'écrit :

$$RC = -0.43 + 0.013LT + 0.003HF + 0.0029PR - 0.0003PA - 0.054FE$$

b) résultats graphiques

```
> plot(reg)
```

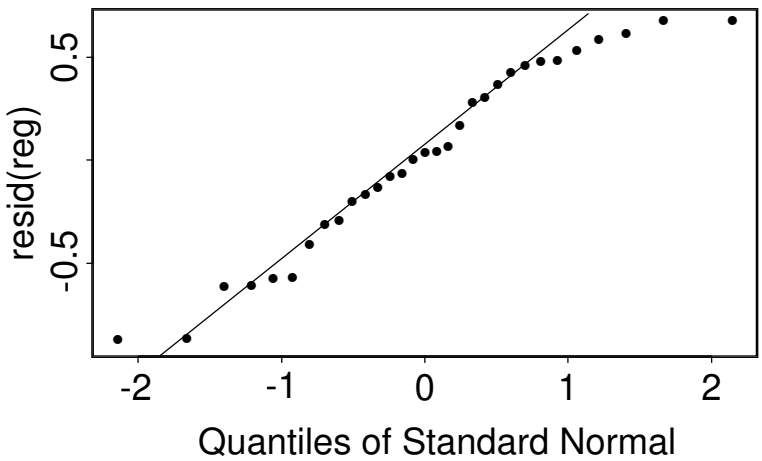




On vérifie les postulats visuellement. Il ne semble pas y avoir de structure particulière des résidus, et leur variance semble homogène.

Il existe un outil graphique permettant de vérifier partiellement la normalité des erreurs : c'est la droite de Henry. On représente les quantiles empiriques des résidus, en fonction des quantiles théoriques de la loi normale centrée réduite. Si les résidus sont distribués suivant une loi normale, les points oscillent autour d'une droite. On superpose la droite passant par les 1er et 3ième quartiles empiriques et théoriques; cette droite sert de point de repère.

```
> qqnorm(resid(reg))
> qqline(resid(reg))
```



Les résidus sont distribués normalement et ne présentent pas de structure particulière. Le modèle est adéquat.

IV Pour aller plus loin . . .

L'examen des données suggère une relation quadratique entre les variables HF et RC. On ajoute donc un terme de degré 2 en HF, et on teste le sous-modèle "reg" contre le grand modèle "reg2".

```
> reg_lm(RC~.,moutarde)
> reg2_lm(RC~.-HF+poly(HF,2),moutarde)
> anova(reg,reg2)
Analysis of Variance Table

Response: RC

                                          Terms Resid. Df
1                    LT + HF + PR + PA + FE        25
2 LT + HF + PR + PA + FE - HF + poly(HF, 2)        24
       RSS    Test Df Sum of Sq  F Value     Pr(F)
1 6.452083
2 6.154164 1 vs. 2  1 0.2979193 1.161825 0.2918024
> summary(reg2,corr=F)

Call: lm(formula = RC ~ . - HF + poly(HF, 2), data = moutarde)
Residuals:
     Min       1Q  Median      3Q     Max
 -0.9342 -0.3282 0.05395 0.3756 0.7055

Coefficients:
               Value Std. Error t value Pr(>|t|)
 (Intercept)  0.2165  0.4573     0.4735  0.6401
          LT  0.0100  0.0086     1.1668  0.2548
          PR  0.0027  0.0006     4.4000  0.0002
          PA -0.0003  0.0001    -2.0887  0.0475
          FE -0.0314  0.0767    -0.4089  0.6863
poly(HF, 2)1  1.8182  0.9862     1.8438  0.0776
poly(HF, 2)2 -0.7363  0.6831    -1.0779  0.2918

Residual standard error: 0.5064 on 24 degrees of freedom
Multiple R-Squared: 0.9141
F-statistic: 42.58 on 6 and 24 degrees of freedom,
the p-value is 1.24e-11
```

L'ajout du polynôme de degré 2 n'améliore pas la qualité du modèle de façon significative. Plusieurs termes semblent non significatifs.

Dans un but de prédiction, on recherche un sous-modèle dont les variables explicatives sont faciles à mesurer et le moins destructives possible. Dans ce cadre, on peut envisager le modèle qui a pour variables explicatives FE et LT. Malheuresement, ce sous-modèle est rejeté.

La tentative de recherche d'un sous-modèle dont toutes les variables sont faciles à observer ayant échoué, on n'a pas plus de raisons de conserver une variable explicative plutôt qu'une autre. On supprime donc un à un les termes les moins explicatifs (t value la plus faible), et on teste le sous-modèle obtenu par rapport au "grand" modèle "reg2". On choisira le dernier modèle accepté selon cette procédure.

```
> reg3_update(reg2,~.-FE)
```

```
> anova(reg2,reg3)
Analysis of Variance Table

Response: RC

                                          Terms Resid. Df
1 LT + HF + PR + PA + FE - HF + poly(HF, 2)        24
2                LT + PR + PA + poly(HF, 2)        25
       RSS Test Df   Sum of Sq   F Value     Pr(F)
1 6.154164
2 6.197030  -FE -1 -0.04286614 0.1671693 0.6862654
```

```
> summary(reg3,corr=F)

Call: lm(formula = RC ~ LT + PR + PA + poly(HF, 2),
  data = moutarde)
Residuals:
     Min      1Q Median      3Q     Max
 -0.9801 -0.323 0.0604 0.3925 0.7091

Coefficients:
               Value Std. Error t value Pr(>|t|)
 (Intercept)  0.1823  0.4420     0.4124  0.6835
          LT  0.0093  0.0082     1.1246  0.2715
          PR  0.0026  0.0005     4.8088  0.0001
          PA -0.0003  0.0001    -2.1058  0.0454
poly(HF, 2)1  1.8921  0.9532     1.9850  0.0582
poly(HF, 2)2 -0.8151  0.6444    -1.2650  0.2175

Residual standard error: 0.4979 on 25 degrees of freedom
Multiple R-Squared: 0.9135
F-statistic: 52.83 on 5 and 25 degrees of freedom,
the p-value is 1.738e-12
```

```
> reg4_update(reg3,~.-LT)
> anova(reg2,reg4)
Analysis of Variance Table

Response: RC

                                          Terms Resid. Df
1 LT + HF + PR + PA + FE - HF + poly(HF, 2)        24
2                     PR + PA + poly(HF, 2)        26
       RSS   Test Df  Sum of Sq   F Value     Pr(F)
1 6.154164
2 6.510518 -LT-FE -2 -0.3563536 0.6948537 0.5089118
```

```
> summary(reg4,corr=F)

Call: lm(formula = RC ~ PR + PA + poly(HF, 2),
 data = moutarde)
Residuals:
    Min      1Q  Median     3Q   Max
 -1.005 -0.4477 0.08819 0.3885 0.804

Coefficients:
               Value Std. Error t value Pr(>|t|)
 (Intercept)  0.5426  0.3060     1.7731  0.0879
          PR  0.0029  0.0005     6.1040  0.0000
          PA -0.0003  0.0001    -1.9522  0.0618
poly(HF, 2)1  1.7727  0.9521     1.8620  0.0739
poly(HF, 2)2 -1.0678  0.6070    -1.7591  0.0903

Residual standard error: 0.5004 on 26 degrees of freedom
Multiple R-Squared: 0.9092
F-statistic: 65.05 on 4 and 26 degrees of freedom,
the p-value is 3.678e-13
```

```
> reg5_update(reg4,~.-poly(HF,2)+HF)
```

```
> anova(reg2,reg5)
Analysis of Variance Table

Response: RC

                                        Terms Resid. Df
1 LT + HF + PR + PA + FE - HF + poly(HF, 2)        24
2                             PR + PA + HF        27
       RSS    Test Df Sum of Sq  F Value    Pr(F)
1 6.154164
2 7.285390 1 vs. 2 -3 -1.131226 1.470518 0.247604
```

```
> summary(reg5,corr=F)

Call: lm(formula = RC ~ PR + PA + HF, data = moutarde)
Residuals:
     Min      1Q  Median      3Q     Max
 -0.9832 -0.3169 0.05363 0.4516 0.8454

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -0.0339  0.1859    -0.1825  0.8565
         PR  0.0033  0.0004     7.5632  0.0000
         PA -0.0002  0.0001    -1.8056  0.0821
         HF  0.0024  0.0021     1.1764  0.2497

Residual standard error: 0.5195 on 27 degrees of freedom
Multiple R-Squared: 0.8983
F-statistic: 79.54 on 3 and 27 degrees of freedom,
the p-value is 1.6e-13
```

```
> reg6_update(reg5,~.-HF)
```

```
> anova(reg2,reg6)
Analysis of Variance Table

Response: RC

                                        Terms Resid. Df
1 LT + HF + PR + PA + FE - HF + poly(HF, 2)        24
2                                  PR + PA        28
       RSS               Test Df Sum of Sq  F Value
1 6.154164
2 7.658835 -LT-FE-poly(HF, 2) -4 -1.504671 1.466979
      Pr(F)
1
2 0.2432753
```

```
> summary(reg6,corr=F)

Call: lm(formula = RC ~ PR + PA, data = moutarde)
Residuals:
     Min      1Q   Median     3Q    Max
 -0.9739 -0.3685 -0.04669 0.4245 0.8638

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)   0.0240   0.1805     0.1331   0.8951
         PR   0.0037   0.0003    14.1984   0.0000
         PA  -0.0003   0.0001    -2.3897   0.0238

Residual standard error: 0.523 on 28 degrees of freedom
Multiple R-Squared: 0.8931
F-statistic: 117 on 2 and 28 degrees of freedom,
the p-value is 2.531e-14
```

```
> reg7_lm(RC~PR,moutarde)
```

```
> anova(reg2,reg7)
Analysis of Variance Table

Response: RC

                                          Terms Resid. Df
1 LT + HF + PR + PA + FE - HF + poly(HF, 2)        24
2                                          PR        29
       RSS                  Test Df Sum of Sq  F Value
1 6.154164
2 9.220858 -LT-PA-FE-poly(HF, 2) -5 -3.066694 2.391898
       Pr(F)
1
2 0.06777223
```

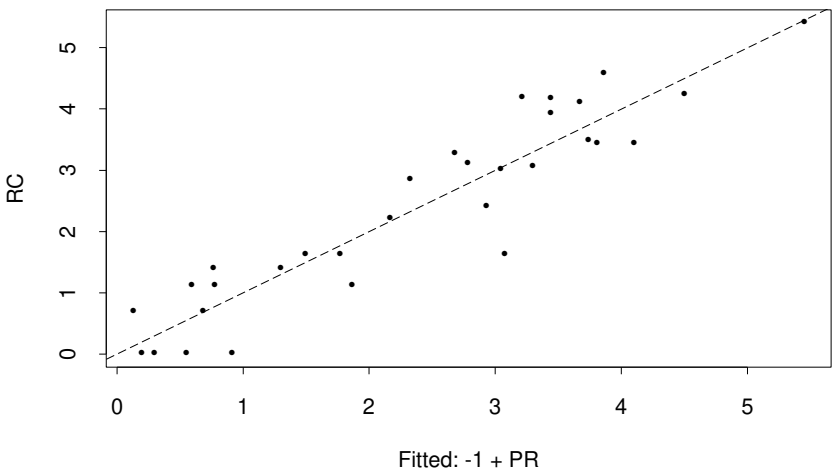
```
> summary(reg7,corr=F)

Call: lm(formula = RC ~ PR, data = moutarde)
Residuals:
    Min      1Q    Median      3Q    Max
 -1.463 -0.2909 -0.009448 0.4835 0.9602

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  0.0242  0.1946     0.1244  0.9018
         PR  0.0033  0.0002    14.0143  0.0000

Residual standard error: 0.5639 on 29 degrees of freedom
Multiple R-Squared: 0.8713
F-statistic: 196.4 on 1 and 29 degrees of freedom,
the p-value is 1.91e-14
```

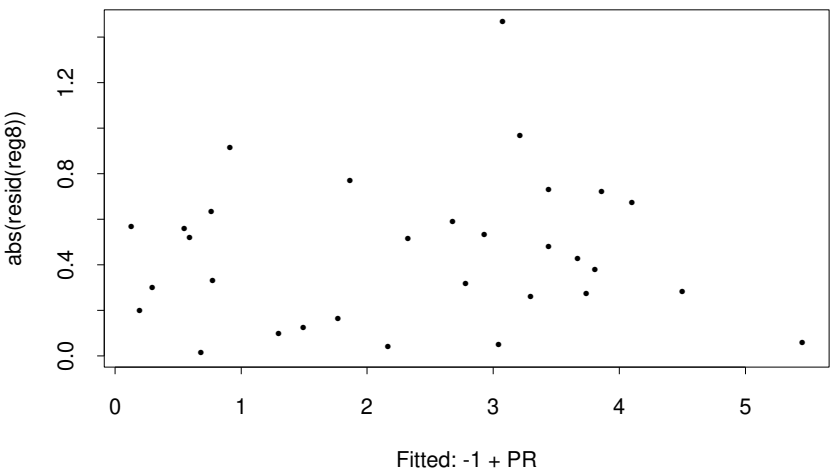
```
> reg8_lm(RC~-1+PR,moutarde)
```

```
> anova(reg2,reg8)

Analysis of Variance Table

Response: RC

                                        Terms Resid. Df
1 LT + HF + PR + PA + FE - HF + poly(HF, 2)        24
2                                    -1 + PR        30
       RSS                  Test Df Sum of Sq  F Value
1 6.154164
2 9.225782 -LT-PA-FE-poly(HF, 2) -6 -3.071618 1.996449
      Pr(F)
1
2 0.1058887
```

```
> summary(reg8)
Call: lm(formula = RC ~ -1 + PR, data = moutarde)
Residuals:
    Min      1Q   Median    3Q    Max
 -1.462 -0.2847 0.009653 0.492 0.9602

Coefficients:
     Value Std. Error t value Pr(>|t|)
PR  0.0034  0.0001    27.5939  0.0000

Residual standard error: 0.5546 on 30 degrees of freedom
Multiple R-Squared: 0.9621
F-statistic: 761.4 on 1 and 30 degrees of freedom,
the p-value is 0
```

```
> plot(reg8)
```





En conclusion, le poids des racines suffit à lui seul à expliquer la résistance au stress hydrique.

En simplifiant le modèle pas à pas, on a mis en œuvre une méthode de choix de régresseurs : la méthode descendante. D'autres méthodes existent. C'est l'objet du chapitre suivant.

Bibliographie

- La régression nouveaux regards sur une ancienne méthode statistique, R. Tomassone, E. Lesquoy, C. Millier, INRA actualités scientifiques et agronomiques, MASSON.
- Applied Regression Analysis, N.R. Draper, H. Smith, Wiley & Sons.
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- A handbook of Statistical Analysis using Splus, B.S. Everitt (1994), Chapman & Hall.
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.

LA REGRESSION LINEAIRE MULTIPLE : CHOIX DE REGRESSEURS

Résumé des principales commandes :

```
> brush(pins,hist=T)
> reg_lm(NIDS~.,data=pins)
> back_stepwise(pins[,1:5],pins[,"NIDS"],method="backward")
> leap_leaps(as.matrix(pins[,1:5]),pins[,"NIDS"])
```

I Le problème

1) Les données

Un expérimentateur étudie l'influence des caractéristiques du peuplement forestier sur le développement de la processionnaire du pin (chenille). Il étudie des parcelles forestières de 10 hectares. Sur chacune de ces parcelles, des placettes de 5 ares sont échantillonnées. Sur chacune des placettes, il observe

- l'altitude (en m),
- la pente (en °),
- le nombre de pins,
- la hauteur de l'arbre échantillonné au centre de la placette,
- le diamètre de cet arbre,
- le nombre de nids de processionnaires par arbre.

altitude	pente	nb de pins	hauteur	diamètre	log(nb nids)
1200	22	1	4.0	14.8	2.37
1342	28	8	4.4	18.0	1.47
1231	28	5	2.4	7.8	1.13
1254	28	18	3.0	9.2	0.85
1357	32	7	3.7	10.7	0.24
1250	27	1	4.4	14.8	1.49
1422	37	22	3.0	8.1	0.30
1309	46	7	5.7	19.6	0.07
1127	24	2	3.5	12.6	3.00
1075	34	9	4.3	12.0	1.21
1166	24	17	5.5	16.7	0.38
1182	41	32	5.4	21.6	0.70
1179	15	0	3.2	10.5	2.64
1256	21	0	5.1	19.5	2.05
1251	26	2	4.2	16.4	1.75
1536	38	31	5.7	17.8	0.06
1554	27	20	5.6	20.2	0.13
1305	30	6	3.8	15.7	1.00
1316	34	8	3.1	11.4	0.41
1427	39	19	4.6	15.2	0.72
1575	20	32	5.2	18.9	0.67
1397	26	16	4.2	14.8	0.12
1377	29	4	5.3	19.8	0.97
1574	24	23	5.2	17.8	0.07
1396	45	13	4.7	15.2	0.10
1393	27	5	4.7	18.3	0.68
1433	23	18	6.5	21.0	0.13
1349	24	1	2.7	5.8	0.20
1208	23	2	3.5	11.5	1.09
1198	28	15	3.9	11.3	0.18
1228	31	6	5.4	21.8	0.35
1229	21	11	5.8	16.7	0.21
1310	36	17	5.2	17.8	0.03

2) Choix de l'analyse

L'unité expérimentale est la placette. Le nombre de nids par arbre est la variable expliquée. Les autres variables correspondent à des caractéristiques forestières, ce sont les variables explicatives ou régresseurs. Pour des raisons partiellement statistiques, c'est le logarithme du nombre de nids par arbre qui est analysé.

On réalise une régression linéaire multiple avec toutes les variables explicatives. On examine graphiquement le modèle obtenu, de façon à le valider.

Une fois le modèle validé on tente de le simplifier. On a deux raisons de simplifier le modèle : dans un but de prédiction, il est souhaitable que les variables explicatives soient de préférence des variables faciles à mesurer, et mesurable avec le plus de précision possible; il est plus facile d'interpréter un modèle qui soit le plus simple possible. Si l'objectif poursuivi n'est pas la prédiction mais plutôt la compréhension du phénomène, ou si les variables explicatives sont aussi faciles (ou difficiles) à mesurer les unes que les autres, on peut mettre en œuvre une procédure automatique de sélection de variables.

II Premiers traitements

1) Saisie

Les variables sont notées :

ALT	:	altitude,
P	:	pente,
NB	:	nombre de pins,
H	:	hauteur de l'arbre échantillonné au centre de la placette,
D	:	diamètre de l'arbre échantillonné au centre de la placette,
NIDS	:	logarithme du nombre de nids de processionnaires.

```
> pins_matrix(c(scan("pins")),ncol=6,byrow=T)
> pins[,6]_log(pins[,6])
> pins_data.frame(ALT=pins[,1],
                  P=pins[,2],
                  NB=pins[,3],
                  H=pins[,4],
                  D=pins[,5],
                  NIDS=pins[,6])
```

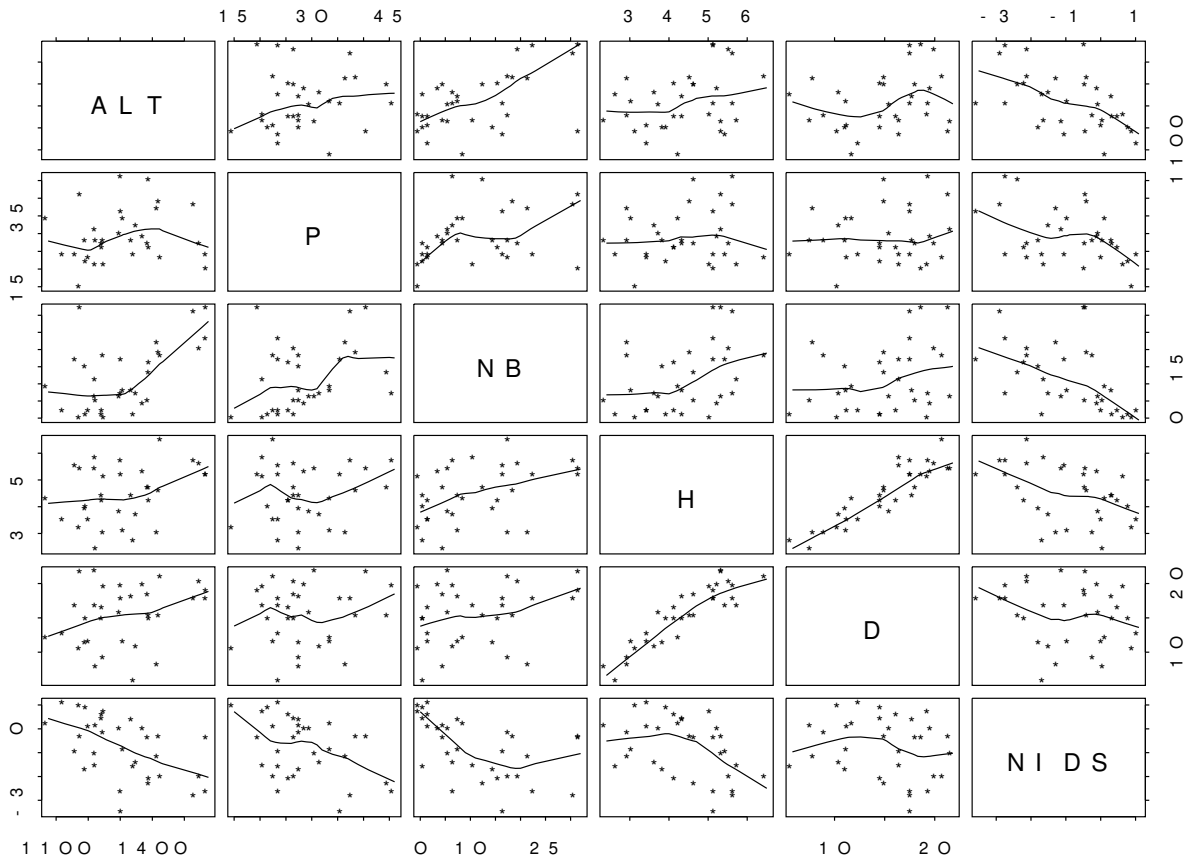
2) Visualisation graphique

La fonction pairs permet, grâce à son argument “panel” de superposer des informations graphiques sur chacun des nuages de points, elle permet notamment de superposer un lissage.

Le nombre de nids semble lié à la plupart des variables explicatives. Une liaison linéaire est plausible. La relation qui lie le nombre de nids aux variables explicatives semble très “lâche”.

Remarquons que les variables H et D sont fortement corrélées. C’est un résultat auquel on pouvait s’attendre a priori.

```
> pairs(pins, panel=function(x,y)
                    { points(x,y); lines(lowess(x,y))})
```



III Analyse

1) Modèle général

Le modèle général est celui qui tient compte de toutes les variables explicatives :

$$NIDS = \alpha + \beta_1 ALT + \beta_2 P + \beta_3 NB + \beta_4 H + \beta_5 D + \varepsilon$$

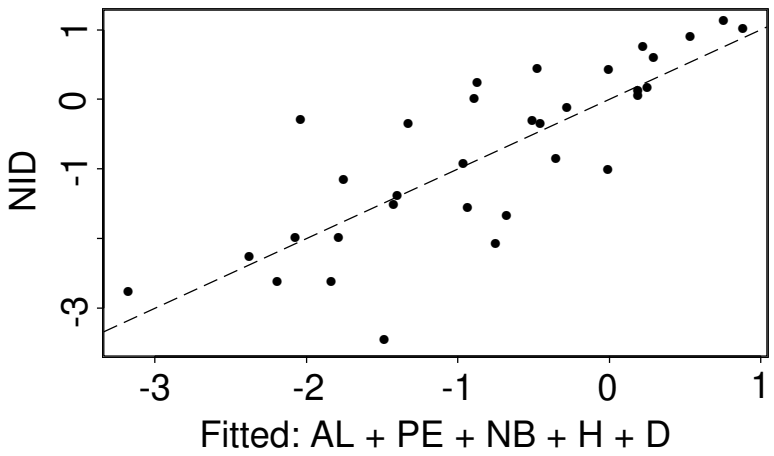
α : terme constant,
$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$: coefficients associés à chaque variable,
ε : erreurs indépendantes, d'espérance nulle, telles que $var(\varepsilon) = \sigma^2$.

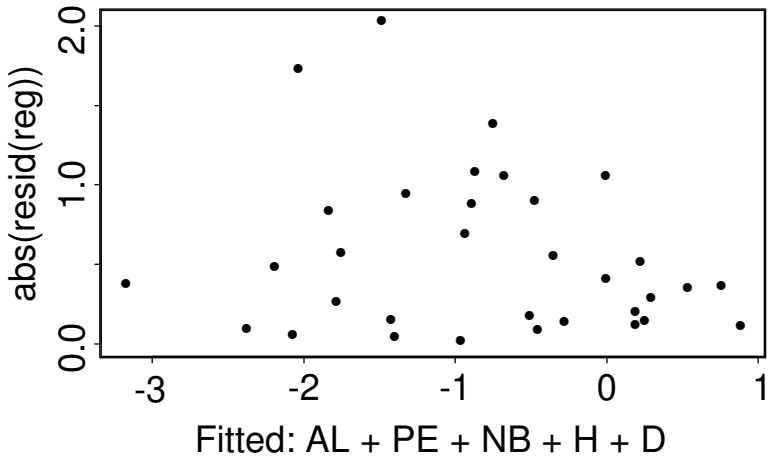
```
> options(contrasts=c("contr.sum","contr.sum"))
> reg_lm(NIDS~.,data=pins)
> summary(reg)
```

```
Call: lm(formula = NIDS ~ ., data = pins)
Residuals:
    Min      1Q  Median      3Q   Max
 -2.017 -0.2509 0.07468 0.3631 1.711

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  7.6208  1.8166     4.1950  0.0003
        ALT -0.0039  0.0013    -2.9023  0.0073
          P -0.0567  0.0206    -2.7447  0.0106
         NB -0.0022  0.0199    -0.1112  0.9123
          H -1.3436  0.3446    -3.8988  0.0006
          D  0.2810  0.0797     3.5242  0.0015

Residual standard error: 0.805 on 27 degrees of freedom
Multiple R-Squared: 0.6472
F-statistic: 9.908 on 5 and 27 degrees of freedom,
the p-value is 1.825e-05
```

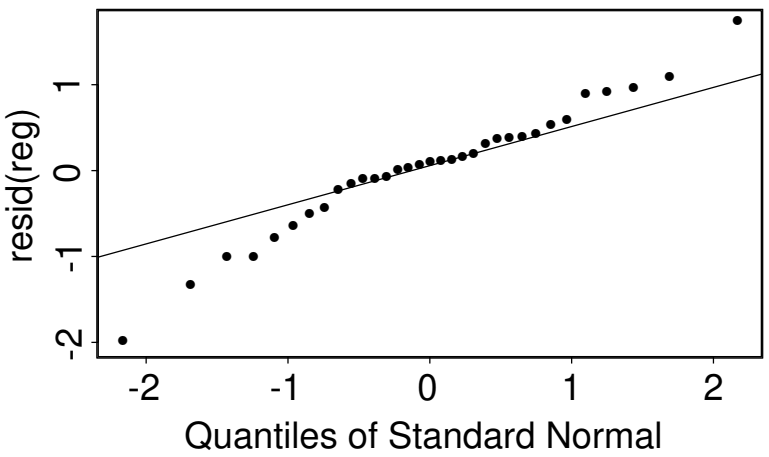
```
Correlation of Coefficients:
   (Intercept)      ALT       P      NB       H
ALT -0.8861
  P -0.3607      0.0708
 NB  0.5514     -0.4749 -0.2932
  H -0.2423      0.0298  0.0354 -0.3273
  D  0.1674     -0.0958 -0.0455  0.2281 -0.8982
```

```
> plot(reg)
```


NID
1
0
-1
-3
-3
-2
-1
0
1
Fitted: AL + PE + NB + H + D


abs(resid(reg))
2.0
1.0
0.0
-3
-2
-1
0
1
Fitted: AL + PE + NB + H + D

```
> qqnorm(resid(reg))
```



Les postulats semblent vérifiés : pas de structure apparente dans les résidus, homogénéité des variances, normalité des résidus. Le modèle est valide. On rejette l'hypothèse de nullité de l'ensemble des coefficients, c'est à dire que les régresseurs dans leur ensemble, apportent une explication significative au nombre de nids de processionnaires. Le développement de la processionnaire dépend des caractéristiques forestières.

2) Choix de régresseurs

Globalement, les régresseurs apportent une explication significative. Cependant les régresseurs ne sont sans doute pas tous nécessaires (ni utiles) pour expliquer le nombre de nids : la variable NB par exemple semble superflue. Le problème se pose en termes de choix d'un ensemble de régresseurs. Il existe plusieurs méthodes : des méthodes pas à pas (ascendantes, descendantes ou mixtes), ou des méthodes directes.

Parmi les méthodes pas à pas, la méthode descendante est préférable à la méthode ascendante : en effet, le modèle complet est estimé à la première étape, il fournit une estimation de la variance résiduelle la moins biaisée possible, et une référence pour les tests des sous-modèles envisagés par la suite. Cependant, les méthodes pas à pas n'assurent pas la sélection du meilleur sous-modèle possible, puisque l'ajout ou le retrait d'une variable explicative n'est pas remis en cause.

Ces méthodes restent cependant des outils de description des données.

a) la méthode de sélection descendante

La méthode de sélection descendante consiste à ajuster le modèle général, et à supprimer les régresseurs les moins explicatifs, un à un. A chaque étape, une

statistique de test est calculée, c'est elle qui détermine l'acceptation ou le rejet du sous-modèle obtenu. Le processus s'arrête lorsque le sous-modèle obtenu est rejeté.

La statistique utilisée à chaque étape peut être l'une ou l'autre des 3 statistiques suivantes, suivant le logiciel ou la façon de procéder :

$$F_1 = \frac{\left(SCR_{M_{i+1}} - SCR_{M_i}\right)/1}{SCR_{M_i}/ddl_{res_{M_i}}}$$

$$F_2 = \frac{\left(SCR_{M_{i+1}} - SCR_{M_i}\right)/1}{SCR_{M_G}/ddl_{res_{M_G}}}$$

$$F_3 = \frac{\left(SCR_{M_{i+1}} - SCR_{M_G}\right)/\left(ddl_{res_{M_{i+1}}} - ddl_{res_{M_G}}\right)}{SCR_{M_G}/ddl_{res_{M_G}}}$$

M_G : sous-modèle général,
M_i : sous-modèle retenu à l'étape précédente,
M_{i+1} : sous-modèle testé.

De façon rigoureuse, on conseillerait plutôt l'emploi de la statistique F_3 qui consiste à tester un sous-modèle en le comparant au modèle complet, supposé sans biais.

On met en œuvre la méthode de sélection descendante à l'aide de la fonction "stepwise". La statistique de test calculée à chaque étape est F_1.

```
> selection_stepwise(pin[,1:5],pin[,"NIDS"],
                              method="backward")
> selection                                              # F1
$rss:
[1] 17.50338 22.97170 31.39837 40.62531 49.59603

$size:
[1] 4 3 2 1 0

$which:
      ALT  P NB H D
4(-3)   T  T  F T T
3(-2)   T  F  F T T
2(-1)   F  F  F T T
1(-5)   F  F  F T F
0(-4)   F  F  F F F

$f.stat:
[1]  0.01511265 10.30969960 12.10533768  9.69760290
[5]  7.28692864

$method:
[1] "backward"
```

Avec :
rss : somme de carrés résiduelle à chaque étape,
size : nombre de régresseurs à chaque étape,
which : régresseurs supprimés à chaque étape,
f.stat : statistique de test,
method : méthode utilisée.

0.015 est à comparer à un $F(1,27)$.

10.3 est à comparer à un $F(1,26)$.

Le premier sous-modèle est accepté alors que le second est rejeté. Les régresseurs choisis sont donc : ALT, P, H, D.

On peut choisir F_2 pour statistique de test :

```
> anova(reg,update(reg,.~.-NB),
            update(reg,.~.-NB-PE),
            update(reg,.~H+D),
            update(reg,.~H),
            update(reg,.~1))
```

```
Response: NIDS                                          # F2
  Terms Resid.          Df   RSS    Test Df Sum of Sq
1 ALT + P + NB + H + D  27 17.49537
2     ALT + P  + H + D  28 17.50338 -NB   -1 -0.008012
3          ALT + H + D  29 22.97170 -P    -1 -5.468320
4                H + D  30 31.39837 -ALT  -1 -8.426672
5                    H  31 40.62531  -D   -1 -9.226937
6                    1  32 49.59603       -1 -8.970719
     F Value      Pr(F)
1
2    0.01236   0.9122822
3    8.43907   0.0072427
4   13.00459   0.0012419
5   14.23962   0.0008032
6   13.84420   0.0009219
```

On peut aussi choisir F_3 comme statistique de test.

```
> anova(reg,update(reg,~.-NB))                          # F3
Analysis of Variance Table

Response: NIDS

                  Terms Resid. Df      RSS Test Df
1 ALT + P + NB + H + D        27 17.49537
2     ALT + P  + H + D        28 17.50338  -NB -1
     Sum of Sq    F Value     Pr(F)
1
2 -0.008012161 0.01236489 0.9122822
```

```
> anova(reg,update(reg,~.-NB-PE))
Analysis of Variance Table

Response: NIDS

                  Terms Resid. Df      RSS   Test Df
1 ALT + P + NB + H + D        27 17.49537
2          ALT + H + D        29 22.97170 -PE-NB -2
  Sum of Sq  F Value     Pr(F)
1
2 -5.476332 4.225718 0.0253136
```

Que l'on choisisse l'une ou l'autre des statistiques F_1,F_2 ou F_3 ne change rien à l'ensemble des régresseurs sélectionnés, mais ça n'est pas toujours le cas.

Le modèle sélectionné est :

$$NIDS = \alpha + \beta_1 ALT + \beta_2 P + \beta_4 H + \beta_5 D + \varepsilon$$

b) La méthode "leaps and bounds"

Cette méthode consiste à envisager tous les sous-modèles possible et à calculer un critère pour chaque sous-modèle. C'est ensuite à l'utilisateur de choisir le sous-modèle qui lui convient, en tenant compte du critère, et éventuellement de considérations non statistiques (ex : variable facile à mesurer ...). Le critère statistique utilisé peut être la statistique de Mallows ou le coefficient de détermination par exemple.

Comme le coefficient de détermination, la statistique de Mallows est un critère global d'évaluation de la qualité d'une régression.

$$C_q = \frac{SCE_q}{S^2} + (2q - n)$$

q : nombre de régresseurs du sous-modèle (le terme constant est ici compté comme un régresseur; q est en fait le nombre de paramètres du sous-modèle),
SCE_q : somme de carrés résiduelle du sous-modèle à q paramètres,
S^2 : variance résiduelle du modèle général,
n : nombre total d'observations.

Si le sous-modèle considéré est sans biais (équivalent au modèle général), l'espérance de C_q est approximativement égale à q. L'écart entre C_q et q est donc une mesure du biais, c'est une mesure de la qualité du sous-modèle.

```
> leap_leaps(as.matrix(pins[,1:5]),pin[,"NIDS"])
> leap
$Cp:
 [1] 23.328392 24.545704 30.668221 30.927410
 [5] 35.771919 37.980179 41.206406 15.431000
 [9] 18.288826 19.840874 20.066921 22.648823
[13] 23.123829 25.133335 10.330601 13.523884
[17] 15.340080 15.984615 17.398430  2.819653
[21]  3.289909  3.594711  4.119999  5.096062
[25] 10.678608 16.549629 17.337616  4.040302
[29]  4.464081  4.807858  5.152027  5.180683
[33]  5.182297  5.558399  5.927650  6.765202
[37] 10.100483  6.006123  6.037823  6.436340
[41]  7.087036  7.088764 13.109085 13.263390
[45]  8.000004

$size:
 [1] 2 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 5 5 5 5
[24] 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 8

$label:
 [1] "ALT"               "NB"
 [3] "P"                 "H"
 [5] "V"                 "FO"
 [7] "D"                 "ALT,P"
 [9] "ALT,FO"            "ALT,NB"
[11] "ALT,H"             "P,H"
[13] "ALT,V"             "ALT,D"
[15] "ALT,P,FO"          "ALT,P,H"
```

```
[17] "ALT,P,NB"           "ALT,P,V"
[19] "ALT,P,D"            "ALT,P,H,D"
[21] "ALT,P,H,FO"         "ALT,P,D,V"
[23] "ALT,P,V,FO"         "ALT,P,D,FO"
[25] "ALT,P,NB,FO"        "ALT,P,NB,V"
[27] "ALT,P,NB,D"         "ALT,P,H,D,V"
[29] "ALT,P,H,D,FO"       "ALT,P,NB,H,D"
[31] "ALT,P,D,V,FO"       "ALT,P,H,V,FO"
[33] "ALT,P,NB,H,FO"      "ALT,P,NB,D,V"
[35] "ALT,P,NB,V,FO"      "ALT,P,NB,D,FO"
[37] "ALT,P,NB,H,V"       "ALT,P,H,D,V,FO"
[39] "ALT,P,NB,H,D,V"     "ALT,P,NB,H,D,FO"
[41] "ALT,P,NB,H,V,FO"    "ALT,P,NB,D,V,FO"
[43] "P,NB,H,D,V,FO"      "ALT,NB,H,D,V,FO"
[45] "ALT,P,NB,H,D,V,FO"

$which:
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
 [1,]    T    F    F    F    F    F    F
 [2,]    F    F    T    F    F    F    F
 [3,]    F    T    F    F    F    F    F
 [4,]    F    F    F    T    F    F    F
 [5,]    F    F    F    F    F    T    F
 [6,]    F    F    F    F    F    F    T
 [7,]    F    F    F    F    T    F    F
 [8,]    T    T    F    F    F    F    F
 [9,]    T    F    F    F    F    F    T
[10,]    T    F    T    F    F    F    F
[11,]    T    F    F    T    F    F    F
[12,]    F    T    F    T    F    F    F
[13,]    T    F    F    F    F    T    F
[14,]    T    F    F    F    T    F    F
[15,]    T    T    F    F    F    F    T
[16,]    T    T    F    T    F    F    F
[17,]    T    T    T    F    F    F    F
[18,]    T    T    F    F    F    T    F
[19,]    T    T    F    F    T    F    F
```

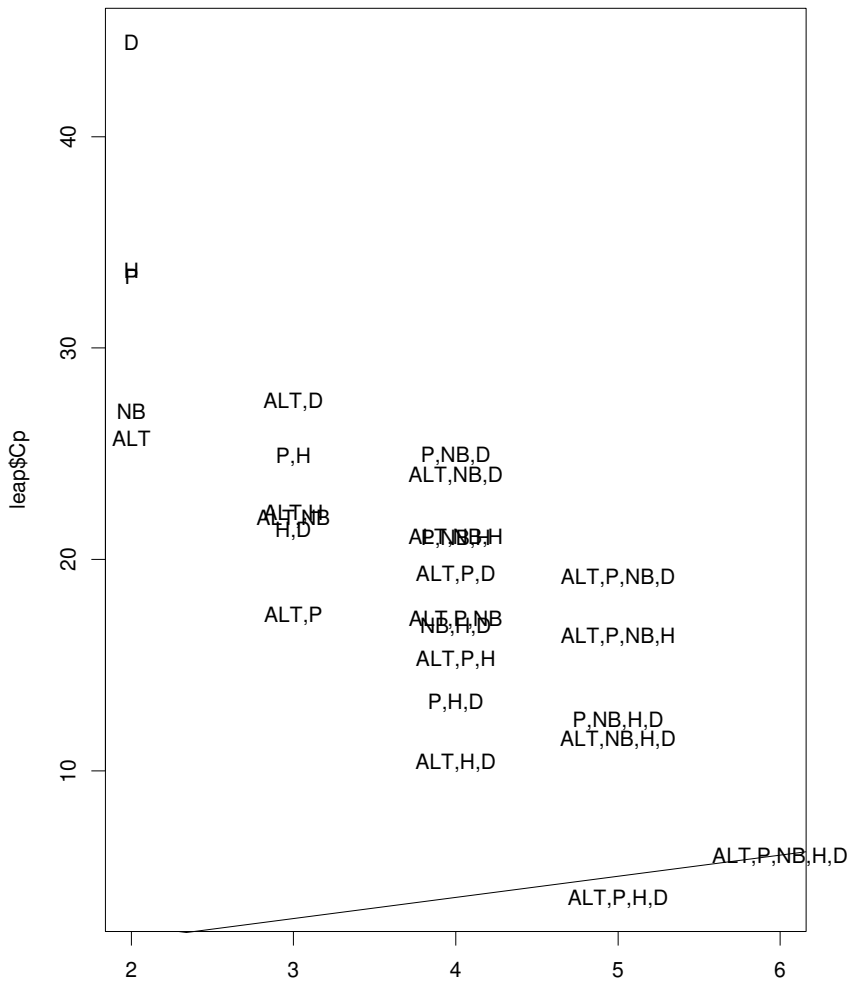
```
[20,]    T    T    F    T    T    F    F
[21,]    T    T    F    T    F    F    T
[22,]    T    T    F    F    T    T    F
[23,]    T    T    F    F    F    T    T
[24,]    T    T    F    F    T    F    T
[25,]    T    T    T    F    F    F    T
[26,]    T    T    T    F    F    T    F
[27,]    T    T    T    F    T    F    F
[28,]    T    T    F    T    T    T    F
[29,]    T    T    F    T    T    F    T
[30,]    T    T    T    T    T    F    F
[31,]    T    T    F    F    T    T    T
[32,]    T    T    F    T    F    T    T
[33,]    T    T    T    T    F    F    T
[34,]    T    T    T    F    T    T    F
[35,]    T    T    T    F    F    T    T
[36,]    T    T    T    F    T    F    T
[37,]    T    T    T    T    F    T    F
[38,]    T    T    F    T    T    T    T
[39,]    T    T    T    T    T    T    F
[40,]    T    T    T    T    T    F    T
[41,]    T    T    T    T    F    T    T
[42,]    T    T    T    F    T    T    T
[43,]    F    T    T    T    T    T    T
[44,]    T    F    T    T    T    T    T
[45,]    T    T    T    T    T    T    T

$int:
[1] T
```

Où :
Cp est la statistique de Mallows obtenue pour chaque sous-modèle,
size , le nombre de coefficients du sous-modèle y compris le terme constant (q),
label , les variables retenues pour chaque sous-modèle,
which , la matrice booléenne des variables retenues ou non,
int , précise si le terme constant a été pris en compte.

On représente graphiquement la statistique C_q en fonction du nombre de paramètres q pour chacun des sous-modèles. On superpose la droite de pente 1 passant par l'origine. Les modèles les meilleurs sont les moins biaisés, ce sont ceux qui ont le C_q le plus petit, le plus proche de la droite. Un modèle à quatre régresseurs convient. On choisit les régresseurs ALT, P, H et D.

Si l'expérimentateur préférait un modèle à trois régresseurs, il choisirait les régresseurs ALT, H et D, à moins que pour une raison ou pour une autre, il ne préfère voir intervenir dans le modèle la variable P à la place de la variable ALT, il choisirait alors P, H et D.

```
> plot(leap$size,leap$Cp,type="n")
> text(leap$size,leap$Cp,leap$label)
> abline(0,1)
```



Pour ce jeu de données, la méthode "leaps and bounds" amène à choisir le même ensemble de régresseurs que la méthode pas à pas descendante, mais ça n'est pas toujours le cas.

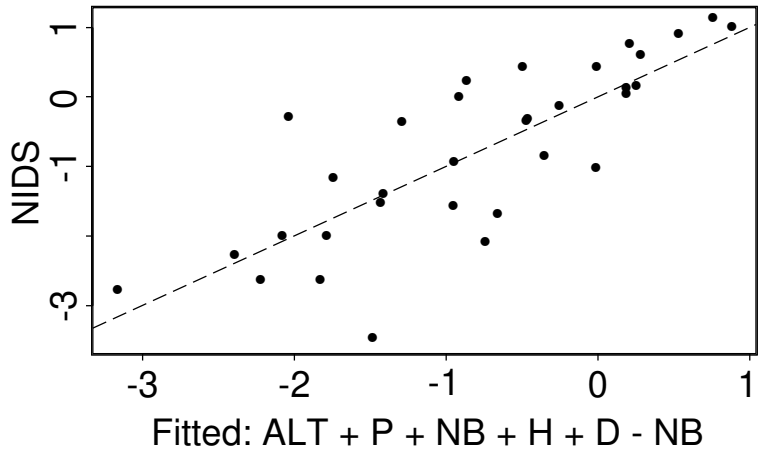
On utilise les outils habituels (test de sous-modèle, examen des résidus) pour valider le modèle choisi.

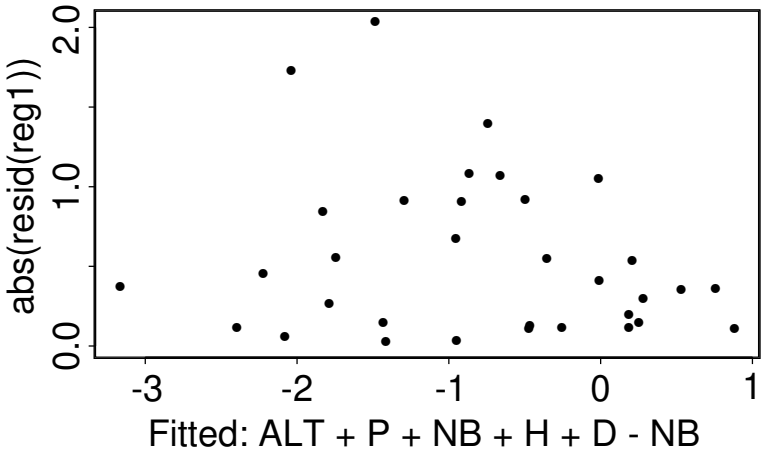
```
> reg1_lm(NIDS~.-NB,pins)
> anova(reg,reg1)
Analysis of Variance Table

Response: NIDS

                       Terms Resid. Df      RSS Test Df
1      ALT + P + NB + H + D        27 17.49537
2 ALT + P + NB + H + D - NB        28 17.50338  -NB -1
     Sum of Sq    F Value     Pr(F)
1
2 -0.008012161 0.01236489 0.9122822
```

> summary(reg1)

```
Call: lm(formula = NIDS ~ . - NB, data = pins)
Residuals:
    Min      1Q  Median      3Q   Max
 -2.021 -0.2501 0.09002 0.3518 1.711

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  7.7321  1.4886     5.1943  0.0000
        ALT -0.0039  0.0011    -3.4188  0.0019
          P -0.0573  0.0194    -2.9576  0.0062
          H -1.3561  0.3198    -4.2401  0.0002
          D  0.2831  0.0763     3.7117  0.0009

Residual standard error: 0.7906 on 28 degrees of freedom
Multiple R-Squared: 0.6471
F-statistic: 12.83 on 4 and 28 degrees of freedom,
the p-value is 4.677e-06

Correlation of Coefficients:
   (Intercept)      ALT       P       H
ALT -0.8502
  P -0.2495     -0.0813
  H -0.0784     -0.1511 -0.0670
  D  0.0513      0.0146  0.0230 -0.8951
```

```
> plot(reg1)
```





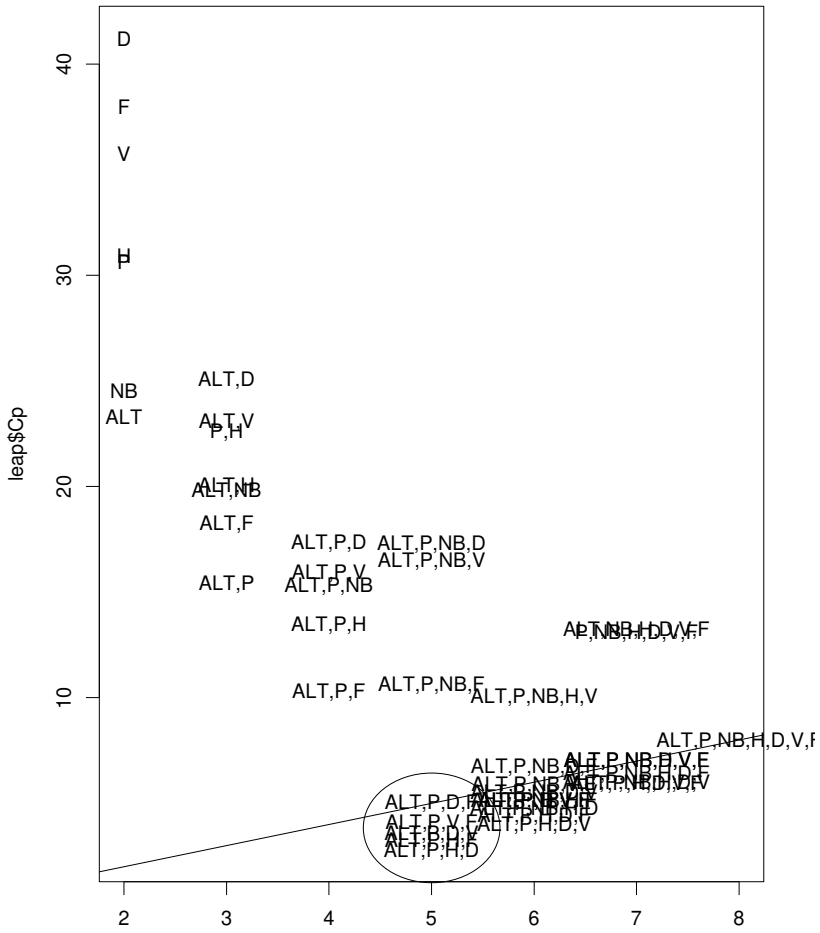
IV Pour aller plus loin . . .

Les variables H et D sont très corrélées. Cela peut causer des problèmes d'interprétation. Nous en verrons une illustration à l'occasion des modèles estimés ci-dessous.

Ce ne sont peut-être pas les variables H et D elles-même qui sont les plus aptes à expliquer le développement de la processionnaire. On peut penser à d'autre variables comme le volume du tronc , ou sa forme plus ou moins élancée, et introduire V=H*D et F=H/D comme variables explicatives potentielles.

La variable H/D est une variable qui représente la forme plus ou moins élancée du tronc. On ne peut pas la nommer F dans les commandes S, car F a un sens particulier : False. On la nomme FOR et on affiche F dans les graphiques.

```
pins$V_pins$H*pins$D
pins$FOR_pins$H/pins$D
leap_leaps(as.matrix(pins[,c(1:5,7,8)]),pins[,"NIDS"])
```



On retient 5 modèles candidats :
“ALT,P,D,F”,”ALT,P,V,F”,”ALT,P,D,V”,”ALT,P,H,F”,”ALT,P,H,D”. Le critère de Mallows est équivalent pour ces 5 modèles. Voici les estimations obtenues pour chacun de ces 5 modèles.

```
> summary(lm(NIDS~ALT+P+D+FOR,pins))      # 1

Call: lm(formula = NIDS ~ ALT + P + D + FOR, data = pins)
Residuals:
    Min      1Q   Median     3Q   Max
 -2.114 -0.3331 -0.06541 0.6423 1.511

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  12.6823   2.1158     5.9940   0.0000
        ALT  -0.0037   0.0012    -3.0993   0.0044
          P  -0.0574   0.0202    -2.8378   0.0084
          D  -0.1292   0.0481    -2.6851   0.0120
        FOR -16.4433   4.3554    -3.7754   0.0008

Residual standard error: 0.8248 on 28 degrees of freedom
Multiple R-Squared: 0.616
F-statistic: 11.23 on 4 and 28 degrees of freedom,
the p-value is 1.461e-05

Correlation of Coefficients:
     (Intercept)     ALT       P       D
 ALT -0.4902
   P -0.1383     -0.0763
   D -0.4838     -0.3333 -0.1093
 FOR -0.6817     -0.2007 -0.0713  0.6761
```

```
> summary(lm(NIDS~ALT+P+V+FOR,pins))      # 2

Call: lm(formula = NIDS ~ ALT + P + V + FOR, data = pins)
Residuals:
  Min      1Q   Median     3Q  Max
 -2.1 -0.2777 -0.01631 0.5909 1.47

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  10.8552   1.8466     5.8784   0.0000
        ALT  -0.0036   0.0012    -3.0617   0.0048
          P  -0.0574   0.0199    -2.8914   0.0073
          V  -0.0154   0.0053    -2.9115   0.0070
        FOR -13.6633   3.6113    -3.7835   0.0007

Residual standard error: 0.8103 on 28 degrees of freedom
Multiple R-Squared: 0.6293
F-statistic: 11.88 on 4 and 28 degrees of freedom,
the p-value is 9.076e-06

Correlation of Coefficients:
    (Intercept)     ALT       P       V
ALT -0.6729
  P -0.1979     -0.0776
  V -0.1714     -0.3408 -0.1024
FOR -0.5566     -0.1370 -0.0468  0.4875
```

```
> summary(lm(NIDS~ALT+P+D+V,pins))      # 3

Call: lm(formula = NIDS ~ ALT + P + D + V, data = pins)
Residuals:
    Min      1Q Median     3Q   Max
 -2.122 -0.2785 0.1505 0.5981 1.372

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  2.9593  1.8353     1.6125  0.1181
        ALT -0.0036  0.0012    -3.0780  0.0046
          P -0.0595  0.0197    -3.0261  0.0053
          D  0.5192  0.1334     3.8923  0.0006
          V -0.0723  0.0177    -4.0791  0.0003

Residual standard error: 0.8024 on 28 degrees of freedom
Multiple R-Squared: 0.6365
F-statistic: 12.26 on 4 and 28 degrees of freedom,
the p-value is 6.972e-06

Correlation of Coefficients:
    (Intercept)     ALT       P       D
ALT -0.8237
  P -0.2342     -0.0814
  D -0.5610      0.1379  0.0193
  V  0.5714     -0.2144 -0.0423 -0.9660
```

```
> summary(lm(NIDS~ALT+P+H+FOR,pins))      # 4

Call: lm(formula = NIDS ~ ALT + P + H + FOR, data = pins)
Residuals:
    Min      1Q   Median     3Q   Max
 -2.058 -0.3119 -0.04984 0.4986 1.563

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  11.3445   1.8480     6.1387   0.0000
        ALT  -0.0037   0.0012    -3.1788   0.0036
          P  -0.0567   0.0196    -2.8993   0.0072
          H  -0.4721   0.1521    -3.1030   0.0043
        FOR -11.8502   3.2829    -3.6097   0.0012

Residual standard error: 0.7978 on 28 degrees of freedom
Multiple R-Squared: 0.6407
F-statistic: 12.48 on 4 and 28 degrees of freedom,
the p-value is 5.968e-06

Correlation of Coefficients:
    (Intercept)     ALT       P       H
ALT -0.6488
  P -0.1850     -0.0792
  H -0.2460     -0.3151 -0.1089
FOR -0.5840     -0.0706 -0.0320  0.3252
```

```
> summary(lm(NIDS~ALT+P+H+D,pins))      # 5

Call: lm(formula = NIDS ~ ALT + P + H + D, data = pins)
Residuals:
    Min      1Q  Median      3Q   Max
 -2.021 -0.2501 0.09002 0.3518 1.711

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  7.7321  1.4886     5.1943  0.0000
        ALT -0.0039  0.0011    -3.4188  0.0019
          P -0.0573  0.0194    -2.9576  0.0062
          H -1.3561  0.3198    -4.2401  0.0002
          D  0.2831  0.0763     3.7117  0.0009

Residual standard error: 0.7906 on 28 degrees of freedom
Multiple R-Squared: 0.6471
F-statistic: 12.83 on 4 and 28 degrees of freedom,
the p-value is 4.677e-06

Correlation of Coefficients:
    (Intercept)     ALT       P       H
ALT -0.8502
  P -0.2495     -0.0813
  H -0.0784     -0.1511 -0.0670
  D  0.0513      0.0146  0.0230 -0.8951
```

Il est très délicat d'interpréter les estimations obtenues lorsque l'une au moins des corrélations entre les paramètres est élevée. Dans les modèles 1 et 3 par exemple, les coefficients de D sont significatifs avec des signes opposés. Que doit on conclure concernant l'effet du diamètre sur le développement de la processionnaire ? Ce phénomène s'explique par la corrélation très élevée entre D et V. L'interprétation du modèle 3 risque d'aboutir à des conclusions erronées.

On recherche donc un modèle dont les coefficients sont les moins corrélés possible, on choisit le modèle 4. On conclut que les conditions favorables au développement de la processionnaire sont :

- une altitude faible,
- peu de pente,
- des arbres petits,
- de forme trapue : diamètre grand par rapport à la hauteur.

Bibliographie

- La régression nouveaux regards sur une ancienne méthode statistique, R. Tomassone, E. Lesquoy, C. Millier, INRA actualités scientifiques et agronomiques, MASSON.
- Applied Regression Analysis, N.R. Draper, H. Smith, Wiley & Sons.
- Statistical Models in S, J.M. Chambers, T.J. Hastie (1992).
- Modern Applied Statistics with Splus, W.N. Venables, B.D. Ripley (1994), Springer-Verlag.