I. LA REGRESSION MULTIPLE (FIN)

I.1. Rappels

Une régression multiple s'accompagne toujours d'une analyse de variance (*ANalyse Of VAriance = ANOVA*). Cette analyse permet de déterminer si l'ensemble des variables explicatives influent de façon significative sur la variable dépendante.

Comme dans le cas de la régression simple, la décomposition de la variance conduit à séparer la part de variance de la variable dépendante expliquée par le modèle de celle due aux résidus.

Ecart total	=	Ecart expliqué par l'équation de régression	+	Ecart résiduel (= inexpliqué)
$(Y_i - \overline{Y})$	=	$(\hat{Y}_i - \overline{Y})$	+	$(Y_i - \hat{Y}_i)$

Où Y_i = observations, $\overline{Y}$ = moyenne des observations et $\hat{Y}_i$ les valeurs prédites par le modèle c.-à-d. les estimations de $E(Y_i)$.

Rappel (voir cours 1). On obtient l'ampleur de chacune de ces dispersions par les sommes suivantes :

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

$$SC_T = SC_E + SC_R$$

Avec :

Source de variation	**Degrés de liberté**	**Carrés moyens**
SC_T = Variance totale (somme des carrés sur le total des données)	n-1	$CM_T = SC_T/(n-1)$
SC_E = Variance résiduelle (variance des résidus)	n-k-1	$CM_E = SC_E/(n-k-1)$
SC_R = Variance de la régression (calculée sur la distance de la régression à la moyenne E(Y))	k	$CM_R = SC_R/k$

On en déduit ensuite les carrés moyens CM_E (erreurs) et CM_R (régression) en divisant SC_E et SC_R par leurs degrés de libertés respectifs (qui sont différents de la régression simple !).

De ce tableau il vient immédiatement :

$$R^2 = \frac{SC_R}{SC_T} \; avec : 0 \le R^2 \le 1 \qquad [1]$$

$(1-R^2)$ *représente la part inexpliquée de la variance totale attribuable soit à l'omission de variables, soit à une formulation incorrecte du modèle, soit enfin à l'erreur instrumentale.*

I.2. La régression est-elle significative dans son ensemble ?

Comme nous sommes dans le cas de la régression multiple, les carrés moyens s'obtiennent à partir d'expressions matricielles :

$$SC_T = \mathbf{Y'Y} - n\overline{Y}^2 \;\; ; \;\; SC_R = \mathbf{b'X'Y} - n\overline{Y}^2 \;\; ; \;\; SC_E = \mathbf{Y'Y} - \mathbf{b'X'Y}$$

Où **Y'** et **X'** sont les transposées des matrices **Y** et **X** respectivement.

On divise par les degrés de libertés respectifs ces expressions pour obtenir les carrés moyens. Notez bien que : CM_E (carré moyen résiduel) = SC_E / (n-k-1) = s^2 et $s = \sqrt{(CM_E)}$ nous donne la dispersion des Yi autour de l'équation de régression (écart –type).

Soit une modèle à k variables :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$$

Hypothèses

H0: $\beta_1 = \beta_2 = ... = \beta_k = 0$ => *Il n'existe aucune contribution d'un quelconque X_k à la réponse*
H1: au moins l'un des $\beta_k \neq 0$ => *Au moins une des variables X_k, apporte une contribution significative à la réponse*

On calcule : $F = \dfrac{CM_R}{CM_E} = \dfrac{SC_R / k}{SC_E / (n-k-1)}$. [2]

Sous l'hypothèse H0 cette quantité est distribuée selon une loi de Fisher avec *k* et (*n*-*k*-1) ddl.

Règle : rejeter H0 si $F > F_{\alpha\,;k,\,n-k-1}$ et favoriser H1. On prend en général $\alpha = 0.05$.

Remarques

1. Très généralement les logiciels calculent la probabilité $P(F_{tabulée} \geq F_{observée})$ sous hypothèse H0. Si cette probabilité est très petite $P(F_{tabulée} \geq F_{observée}) < \alpha$ on rejettera H0. **R** présente le résultat sous la forme : **Prob > F.**
2. Ce test ne permet pas de décider lesquels des coefficients de régression contribuent effectivement de façon efficace (significative) à la réponse.
3. On peut utiliser R^2 pour calculer F en combinant les relations [1] et [2]. Après quelques manipulations algébriques, il vient : $F = \dfrac{R^2 / k}{(1-R^2)/(n-k-1)}$.

I.3. Contribution marginale des variables explicatives

Comme nous venons de le voir, *déclarer que la régression est, dans son ensemble, significative n'implique pas nécessairement que toutes les variables explicatives de l'équation de régression ont une contribution significative*. On devra donc déterminer si la *contribution marginale de chaque variable* est significative. Le test consiste à examiner si l'ajout d'une variable à la suite d'autres variables présentes dans le modèle de régression apporte une contribution significative à la part de variance due à la régression (SC_R). On teste donc de la *pertinence de la dernière variable introduite dans le modèle*. On peut utiliser le test *t* de Student ou encore le test de Fisher-Snedecor.

Test de Student.

On rappelle (cours II) dans le cas du modèle linéaire simple, sous hypothèse de normalité des Yi, la distribution d'échantillonnage du coefficient b_1 est celle d'une loi normale de moyenne $E(b_1) = \beta_1$ et de variance $s^2(b1)$ estimée ainsi : $s(b_1) = \dfrac{s}{\sqrt{\sum (X_i - \bar{X})^2}}$. Les fluctuations de l'écart réduit sont celles de la loi de Student à (n-2) ddl. On a $t = \dfrac{b_1}{s(b_1)} = \dfrac{b_1}{s / \sqrt{\sum (X_i - \bar{X})^2}}$ pour effectuer le test.

Soit l'équation de régression linéaire multiple : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

On souhaite tester la contribution marginale de X_1 : H0: $\beta_1 = 0$, H1 $\neq 0$.

On utilise à nouveau $t = \dfrac{b_1}{s(b_1)}$ avec les valeurs obtenues de l'estimation du modèle à 3 variables.

L'application du test à l'hypothèse H0: $\beta_1 = 0$ permet de tester la contribution de X_1 *comme si elle était la dernière variable introduite dans l'équation* : $Y_i = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \boldsymbol{\beta_1 X_1} + \varepsilon$. On rejette H0 si on obtient : $t > t_{\alpha/2\,;\,(n-k-1)}$ ou bien $t < -t_{\alpha/2\,;\,(n-k-1)}$

Si on ne peut rejeter H0, on devra conclure que l'apport de X_1 est superflu à la suite des autres variables.

Intervalle de confiance

De la même façon que dans le cas de la régression simple, il est possible de calculer pour chaque coefficient l'intervalle de confiance à un niveau souhaité (en général 95%) :

$$b_j - t_{\alpha/2;n-k-1} \cdot s(b_j) \le \beta_j \le b_j + t_{\alpha/2;n-k-1} \cdot s(b_j)$$

Si $\beta_j = 0$ se situe dans l'intervalle, on acceptera l'hypothèse H0, c.-à-d. que la variable X_j est non significative au seuil α, compte tenu de la présence des autres variables.

NB. La variance de chacun des b_j, et donc aussi les $s(b_j)$, s'obtient à partir de la matrice $\mathbf{s^2(b)}$ des variances-covariances (matrice de dispersion) des coefficients de régression :

$$\mathbf{s^2(b)} = s(\mathbf{X'X})^{-1} = CM_E \cdot (\mathbf{X'X})^{-1} = \begin{bmatrix} s^2(b_0) & Cov(b_0,b_1) & \dots & Cov(b_0,b_k) \\ Cov(b_0,b_1) & s^2(b_1) & & Cov(b_1,b_k) \\ & & M & \\ Cov(b_k,b_0) & Cov(b_k,b_1) & \dots & s^2(b_k) \end{bmatrix}$$

Les éléments diagonaux de cette matrice sont les variances cherchées des coefficients de régression. La présence de covariances entre les coefficients indique que l'on ne peut estimer indépendamment chaque paramètre de la régression multiple. Si on retranche une variable quelconque du modèle il convient alors de refaire une nouvelle analyse de régression pour obtenir la nouvelle équation.

II. QUELQUES REMARQUES POUR TERMINER LA REGRESSION MULTIPLE.

L'introduction de variables qualitatives : les variables auxiliaire

Il est possible d'introduire des variables explicatives de nature qualitatives dites variables auxiliaires. Ces données comportent deux, ou plus, modalités. Dans le cas des données ours.txt, la variable sexe ne comporte, évidemment, que deux modalités. On pratique comme suit. Soit à établir le modèle Masse ~ hauteur + sexe. La variable sexe est tout d'abord recodée en deux valeurs (0 = male ; 1= femelle). On a dès lors deux modèles de régression :

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_{i1} \qquad \text{cas femelle}$$
$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2(0) = \beta_0 + \beta_1 X_{i1} \qquad \text{cas male}$$

Il apparaît immédiatement que les deux régressions ne diffèreront que par leurs intercepts…Les deux droites seront parallèles, le décalage représentant l'effet du sexe. Pour tester la contribution marginale sur la variable auxiliaire sexe, on procède à un test t en calculant $t = \dfrac{b_2}{s(b_2)}$ comme habituellement.

Lorsqu'un modèle comporte **plusieurs variables auxiliaires**, on effectue deux régressions, l'une avec le modèle complet et l'autre avec le modèle réduit (celui ne comportant que les variables explicatives). Pour évaluer l'apport des variables auxiliaires on effectue alors un test « **F partiel** ». Par exemple pour un modèle comportant 5 variables dont 2 auxiliaires (X_4 et X_5). Soit à tester H0: $\beta_4 = \beta_5 = 0$, on utilise la quantité :

$$F = \frac{[SC_R(X_1,X_2,X_3,X_4,X_5) - SC_R(X_1,X_2,X_3)]/2}{SC_E(X_1,X_2,X_3,X_4,X_5)/n-5-1}$$

Rejeter H0 si $F > F_{\alpha\,;\,k-g,\,n-k-1}$ où g représente le nombre de paramètres spécifiés en H0 (ici 2).

III. L'ANOVA

III.1 Un facteur

Bien que nous ayons déjà abordé partiellement les principes de l'ANOVA au cours de la régression multiple, nous allons en exposer les principes et méthodes.

L'analyse de la variance à un facteur de classification a pour but la comparaison des moyennes de nA populations, des variantes (ou niveaux n_a), d'un facteur contrôlé (ou facteur A) de variation. On a donc à faire à une variable dépendante continue et une variable explicative discrète (on dit aussi catégorielle). L'analyse consiste à tester si les différences de variation dans chaque groupe (ou échantillon), définies par les modalités de la variable explicative, s'écartent de manière significative de la valeur 0. Son application nécessite quelques conditions :

1. Le paramètre étudié suit une distribution normale (obéit à la loi de Gauss)
2. Les variances des populations sont égales (homoscédaticité)
3. Les échantillons sont prélevés aléatoirement et indépendamment.

Typiquement, les données pouvant faire l'objet d'une ANOVA se présentent sous la forme d'un tableau dans lequel on a consigné les résultats d'une expérience réalisée en plusieurs conditions d'un même facteur. Exemple : on a mesuré l'élongation à 24 heures de l'hypocotyle de germinations de tomates en présence de diverses concentrations d'un extrait d'une plante (*Calluna vulgaris*). Le facteur de contrôle est l'extrait, que l'on a testé avec plusieurs concentrations. Pour chaque concentration on a mesuré l'élongation sur un suffisamment grand nombre de semences. Les semences sont en condition d'indépendance mutuelle pour chacune des expériences.

Aspects théoriques

Les données sont consignées dans un tableau ou $A_1 \ldots A_k$ sont des échantillons:

Facteur	A_1	A_2		A_k	
	x_1^1	x_2^1	···	x_k^1	
	x_1^2	x_2^2	···	x_k^2	
	...	...	...	...	
	x_1^n	x_2^n	···	x_k^n	
Moyennes	$\bar{x}_1$	$\bar{x}_2$	···	$\bar{x}_k$	$\bar{X}$; $\mathrm{Var}(\bar{x}_i)$
Variances	$\mathrm{Var}(x_1)$	$\mathrm{Var}(x_2)$	...	$\mathrm{Var}(x_k)$	$\mathrm{E}(\mathrm{Var}(x_i))$

Conditions expérimentales

Tableau des données expérimentales

Si chacun des *k* échantillon est issu d'une variable aléatoire X_k, le problème est de tester l'hypothèse H0: $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$. $\bar{X}$, la moyenne totale s'écrit : $\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n} x_i^j$. On a alors :

$$\frac{1}{n}\sum_{1}^{k}\sum_{1}^{n}(x_i^j - \bar{X})^2 = \frac{1}{n}\sum_{1}^{k}\sum_{1}^{n}(x_i^j - \bar{x}_i)^2 + \frac{1}{n}\sum_{1}^{k} n_i(\bar{x}_i - \bar{X})^2 ,$$

Ce qui n'est autre que la célèbre décomposition de la **variance totale** en **moyenne des variances** et **variance des moyennes**, que l'on écrit, traditionnellement, en raccourci:

$$\boxed{S^2 = S_R^2 + S_A^2}$$

S_A^2 est dite la *variance due au facteur* (= variance des moyennes)

S_R^2 est dite la *variance résiduelle* (= moyenne des variances)

Modèle sous-jacent : $x_i^j = \mu_i + \alpha_i + e_{ij}$ où les $\{ e_{ij}\}$ sont indépendants, et de distributions identiques $\mathcal{N}(0, \sigma^2)$.

Dans l'hypothèse H0, et seulement dans celle-ci, les x_i sont des variables de même loi. Auquel cas $\frac{nS^2}{\sigma^2}$ suit un χ^2_{n-1} et $\frac{nS_A^2}{\sigma^2}$ suit un χ^2_{k-1}. Si H0 est vraie alors : $\frac{S_A^2/(k-1)}{S_R^2/(n-k)} = F(k-1;n-k)$.
Comme déjà vu précédemment, si ce rapport est supérieur à la valeur critique d'une table de Fisher-Snedecor, on conclut au rejet de H0.

Test Post Hoc (litt. : « après cela »).
Il s'agit donc de connaître, si l'hypothèse H0 est rejetée, quelles sont les niveaux de facteurs qui diffèrent deux à deux. Le test de Tukey, ou test de la différence franchement significative (HSD : honestly significative difference), consiste à calculer l'expression Q pour toutes les différences de moyennes. On calcule les quantités :

$$Q = \frac{\bar{x}_i - \bar{x}_j}{E} \text{ où } \bar{x}_i > \bar{x}_j \text{ et } E = \sqrt{\frac{CM_E}{n}}$$

Cette quantité Q suit une loi particulière (dite des écarts « studentisés ») de paramètres *r* (nombre de groupes) et *ddl* de la ligne des résidus du tableau résultat de l'ANOVA.
Si $Q_{Observé} > Q_{Critique}$, on conclut à une différence significative entre les deux moyennes constituant la paire.

III.2. deux facteurs (avec répétitions).

Voici la description du tableau et de la décomposition de la variance donnée par Jean Vaillant :

On étudie deux facteurs A et B ayant respectivement I et J niveaux.

Données : x_{ijk} est la valeur observée pour la $k$ème répétition du traitement (i,j); $\bar{x}_{i..}$ est la moyenne observée pour le niveau i de A; $\bar{x}_{.j.}$ celle pour le niveau j de B; $\bar{x}_{ij.}$ celle pour le traitement (i,j); n_{ij} est le nombre de répétitions du traitement (i,j); n_{i+} est le nombre de répétitions du niveau i de A; n_{+j} est le nombre de répétitions du niveau j de B.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$n-1$	$CMT = \frac{SCT}{n-1}$	
Facteur A	SCF_A	$I-1$	$CMF_A = \frac{SCF_A}{I-1}$	$F_A = \frac{CMF_A}{CMR}$
Facteur B	SCF_B	$J-1$	$CMF_B = \frac{SCF_B}{J-1}$	$F_B = \frac{CMF_B}{CMR}$
Interaction (A,B)	SCF_{AB}	$(I-1)(J-1)$	$CMF_{AB} = \frac{SCF_{AB}}{(I-1)(J-1)}$	$F_{AB} = \frac{CMF_{AB}}{CMR}$
Résiduelle	SCR	$n-IJ$	$CMR = \frac{SCR}{n-IJ}$	

avec $SCT = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk}-\bar{x})^2$; $SCF_A = \sum_{i=1}^{I} n_{i+}(\bar{x}_{i..}-\bar{x})^2$; $SCF_B = \sum_{j=1}^{J} n_{+j}(\bar{x}_{.j.}-\bar{x})^2$

$SCF_{AB} = \sum_{i=1}^{I}\sum_{i=1}^{J} n_{ij}(\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}+\bar{x})^2$ et $SCR = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk}-\bar{x}_{ij.})^2$

Equation d'ANOVA : $SCT = SCF_A + SCF_B + SCF_{AB} + SCR$

Modèle sous-jacent : $x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ où les e_{ijk} sont i.i.d. selon $\mathcal{N}(0,\sigma^2)$.

Notez que le terme SCF_{AB} permet e mesurer l'interaction entre les deux facteurs. En effet, ce terme est nul quand les variations d'un des deux facteurs sont indépendantes du second car alors :

$$\bar{x}_{ij.} - \bar{x}_{i..} = \bar{x}_{.j.} - \bar{x}$$

Exemple : effet sur la récolte de la durée jour/nuit et de l'ajout d'azote :

Photopériode	Unités d'azote (kg/ha) 10	50	80
8/16	5	15	21
	10	22	29
	8	18	25
14/10	6	25	55
	9	32	60
	12	40	48

⚠ Il existe de nombreux autres plans d'expérimentation (autres que les deux plans simples que nous avons étudiés), conduisant à autant d'équations de décomposition de la variance. Un excellent résumé en a été fait par Jean Vaillant : http://monnano.weebly.com/uploads/1/6/6/3/1663287/plan_exp_anova.pdf

Une autre étude exhaustive par **Doncaster** et **Davey** (Université de Southampton, UK) de tous les plans expérimentaux de l'ANOVA et de l'ANCOVA , ainsi que tous les codes **R** détaillés (et bien plus !) pour chacun de ces plans par **Patrick Doncaster :** https://www.southampton.ac.uk/~cpd/anovas/datasets/

Voici les paramètres de la fonction aov() pour quelques uns des cas les plus courants :

```
# One Way Anova (Completely Randomized Design)
fit <- aov(y ~ A, data = mydataframe)

# Randomized Block Design (B is the blocking factor)
fit <- aov(y ~ A + B, data = mydataframe)

# Two Way Factorial Design
fit <- aov(y ~ A + B + A:B, data = mydataframe)


# Analysis of Covariance
fit <- aov(y ~ A + x, data = mydataframe)
```

For within subjects designs, the dataframe has to be rearranged so that each measurement on a subject is a separate observation. See R and Analysis of Variance.

```
# One Within Factor
fit <- aov(y~A+Error(Subject/A),data=mydataframe)

# Two Within Factors W1 W2, Two Between Factors B1 B2
fit <- aov(y~(W1*W2*B1*B2)+Error(Subject/(W1*W2))+(B1*B2),
   data=mydataframe)
```

Applications

Présentation des données
 # downloader et sauver tomates.txt dans votre répertoire
 # sauvegarder tomates.txt dans un dataframe au moyen de <read.table> (cf. Cours I)
 # Les séparateurs du document sont des tabulations, la première ligne contient les noms des variables.
 # vérifier le contenu de tomates.

 # Cette présentation de données n'est pas compatible avec une ANOVA avec R.
 # Il faut donc réorganiser les données, ce qui a été fait dans « tomates aov.txt » ci-après.

 # downloader et sauver « tomates aov.txt » dans votre répertoire
 # sauvegarder « tomates aov.txt » dans un dataframe au moyen de <read.table>
 # Les séparateurs du document sont des tabulations, la première ligne indique le contenu des colonnes.
 # afficher le contenu de « tomates aov »

EXEMPLE D'ANALYSE A UN FACTEUR (tomates aov.txt)

Une première approche consiste en un affichage graphique des données pour visualiser d'éventuelles différences.

```
> names(dat)=c("Obs","Long","dosage")
> attach(dat)                          # permet de manipuler par names
> dosage = factor(dosage)              # précaution nécessaire !
> is.factor(dosage)                    # vérification
> boxplot(Long~dosage)                 # visualisation des données
> plot(Long~dosage)                    # plot est capable de passer d'un type à l'autre
> fit <-aov(Long~dosage)               # analyse de variance
> summary(fit)                         # = summary(mf <-aov(Long~dosage))
```

Que concluez-vous ?

```
> par(mfrow=c(2,2))                    # 4 graphiques dans une fenêtre
> plot(fit)
```

Ces graphiques permettent de vérifier que les contraintes de base pour l'application de l'ANOVA sont respectées : distribution normale (QQ-plot), résidus vs fitted : erreurs constantes et indépendance des niveaux du facteur.

NB : il peut se trouver qu'un niveau de facteur diffère de tous les autres bien que l'hypothèse H0 soit acceptée… Le boxplot permet de détecter ce genre de situation. On procède au test HSD de Tukey

```
> PostHoc = TukeyHSD(fit, "dosage")  # le test
> PostHoc                            # résultat  Que concluez-vous ?
> plot(PostHoc)                      # voyons cela…
```

#EXEMPLE D'ANALYSE A 2 FACTEURS (tomates 2var.txt)

 # Cette fois on a testé les doses d'extraits en deux conditions différentes
 # de pH pour les cultures.
 # downloader « tomates 2var.txt » dans votre répertoire
 # sauvegarder « tomates 2var.txt » dans un dataframe au moyen de <read.table>
 # afficher le contenu

La question est maintenant : Y-a-t-il une différence d'activité entre les deux conditions de culture et y-a-t-il un effet combiné Dose X pH ?

```
> names(dat) = c("Long", "Dose","pH")
> attach(dat)
> Dose = factor(Dose)                    # deux précautions
> pH = factor(pH)
> interaction.plot(Dose, pH, Long)       #un graphe intéressant. Attention à
                                         # l'ordre des paramètres
> aov2 = aov(Long~ Dose*pH)
> aov2 = aov(Long~ Dose + pH + Dose:pH)         # résultat identique
> boxplot(Long~Dose*Cond)                       # la boite de Box !
> summary (aov2)
> print(model.tables(aov2,"means"),digits=3)   #indique les moyennes
```

Que conclure ?

Rmq : tester les commandes suivantes.

```
> aovlm <- lm(Long~Dose+pH+Dose:pH)
> aovlm
> aovlm <- lm(Long~Dose*pH)
> aovlm
> anova(aovlm)
```

Que concluez-vous par rapport à l'ANOVA ?
Terminons notre travail :

```
> PostHoc <- TukeyHSD(aov(Long~Dose))
> plot(PostHoc)
> PostHoc <- TukeyHSD(aov(Long~Dose*pH))
> plot(PostHoc)
> library(gplots)
> plotmeans(Long~Dose, xlab = "Dosage", ylab = "élongation 24 h", main="MeanPlot\n
avec IC de 95% ")
> plotmeans(Long~pH, xlab = "Dosage", ylab = "élongation 24 h", main="MeanPlot\n
avec IC de 95% ")
```

Pour se distraire… Ma première fonction R !!!
Il est souvent utile de conserver un ensemble de commandes que l'on utilise systématiquement. Pour cela il suffit de créer une fonction et de la conserver dans un fichier texte ayant l'extension R. Par exemple : <MonAnova.R> Il suffira ensuite de l'ouvrir avec R puis d'ajouter à sa suite la ligne d'appel…

```
MyAnova <- function(y,A,B){          # déclaration ; attend 3 paramètres
A = factor(A)                         # deux précautions…
B = factor(B)
interaction.plot(A, B, y)             # un graphe intéressant
x11()                                  # pour conserver le graphe précédent
boxplot(y~A*B)                        # la boite de Box !
aov2 = aov(y~ A*B)
summary (aov2)
print(model.tables(aov2,"means"),digits=3)  #indique les moyennes
}                                      # fin de MonAnova

# Il suffit de l'appeler ainsi
dat <- read.table("tomates 2var.txt", sep = "\t", header = TRUE)
names(dat) = c("Long", "Dose","pH")
MyAnova(Long, Dose, pH)              # that's all !
```

Compléments : exemples de MANOVA et ANCOVA

On considère l'exemple historique "Iris" de Fisher : on dispose d'échantillons de taille 50 de trois espèces d'iris. Les espèces d'iris sont : setosa, versicolor et virginica. Les variables dépendantes sont la longueur des sépales, la largeur des sépales, la longueur des pétales et la largeur des pétales.

```
data(iris)
iris
```

On place les 4 variables dépendantes (VD) dans un data.frame :

```
Y <- cbind(iris$Sepal.Length,iris$Sepal.Width,iris$Petal.Length,iris$Petal.Width)
Y
summary(Y)
```

- **Puis on effectue l'analyse. Attention : test de Wilk car le test F habituel ne convient pas !**
- **Pour que R trouve la variable Species on précise data = iris (sinon, il ne connait pas car on a pas exécuté attach(iris)).**

```
fit <- manova(Y ~ Species, data=iris)
summary(fit, test = "Wilks")
```

D'autres tests sont également disponibles au lieu de Wilks (notez l'écriture compactée des analyses):

```
summary(manova(Y ~ Species, data=iris),test="Hotelling-Lawley")
summary(manova(Y ~ Species, data=iris),test="Roy")
```

Pour obtenir les statistiques univariées :

```
summary.aov(fit)
```

Une ANCOVA (un peu stupide...). Sepal.Length joue le role de covariable.

```
fitancova <- aov(Sepal.Width ~ Species + Sepal.Length, data = iris)
summary(fitancova)
```