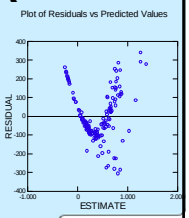
Utilisation approfondie des logiciels statistiques, module 7

Introduction à R et à MYSTAT

Guy Mélard, U.L.B.
Mars 2010
(gmelard@ulb.ac.be)



Objectif de la leçon

Elle a pour but de fournir une introduction aux autres logiciels statistique de manière à faciliter leur utilisation en complément de SPSS, avec l'illustration sur des exemples concrets de statistique appliquée

⇨R

⇨MYSTAT

D'autres logiciels pourraient être envisagés (Minitab, …) pour lesquels il existe des versions d'évaluation.

Mais, ne pas considérer des logiciels spécialisés (économétrie, séries chronologiques, …)

Exemple : étude clinique en rhumatologie

22 polyarthritiques
- ✓âge du patient (AGE)
- ✓sévérité de l'affection (0 à 4: SEVER)
- ✓type de traitement
  - ◆anti-inflammatoires (TANTINFL) et/ou
  - ◆stéroïdes (TSTEROI)
- ✓dosages de 3 enzymes: 5'NU (FNU), ADA, PNP
- ✓pourcentages de lymphographie: LYMPHOST, OKT4, OKT8

+ 3 groupes de contrôle: dosage d'un seul enzyme pour chacun d'eux

Questions:

1. le dosage des enzymes est-il plus bas chez les arthritiques?
2. si oui, y-a-t-il une dépendance entre ce dosage et ...
   - 2.1 l'âge?
   - 2.2 le traitement?
   - 2.3 la sévérité de l'affection?
   - 2.4 les trois pourcentages?

Introduction à R

⇨R est un logiciel statistique dérivé de S

⇨Le langage "S" a été développé en 1976 chez Lucent Technologies (auparavant AT&T Bell Labs) par une équipe dirigée par John Chambers

⇨Cela a été le premier langage de programmation créé spécifiquement pour la visualisation et l'exploration des données, la modélisation statistique et la programmation sur des données

⇨En 1988, S est devenu un produit commercial appelé SPlus, maintenant Spotfire S+, dans sa 8e version et commercialisé par Tibco (http://www.tibco.com/)

Introduction à R

⇨Commencé en 1997, R (http://www.r-project.org/) est essentiellement un langage de programmation interprété libre (« open source ») développé de manière collaborative avec de multiples fonctions statistiques et, de plus, la possibilité pour des contributeurs (vous ou moi) d'ajouter des « packages »

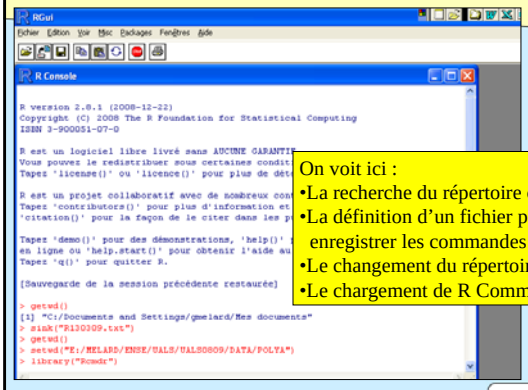
⇨Avant de faire cela, regardez d'abord ce qui existe : dans un site miroir de CRAN (Comprehensive R Archive Network), par exemple parmi les « task views »

⇨Pour plus d'informations sur R voyez par exemple http://en.wikipedia.org/wiki/R_programming_language.

Introduction à R

- R est entièrement libre de droits et gratuit
- De plus il peut être installé sous différents systèmes (PC, sous Windows ou Linux, Mac, multiple systèmes Unix)
- Le code source est disponible
- Une fois R installé, vous pouvez ajouter des « packages » et les mettre à jour
- R de base consiste en une interface (Rgui.exe sur PC) et un programme d'exécution (R.exe)
- L'interface sert presque uniquement à gérer R, pas à traiter des applications statistiques

Introduction à R

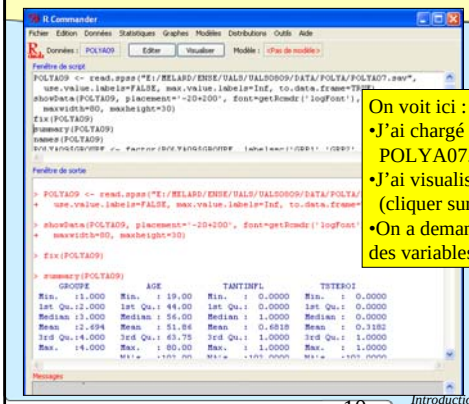


On voit ici :
- La recherche du répertoire courant
- La définition d'un fichier pour enregistrer les commandes
- Le changement du répertoire courant
- Le chargement de R Commander

Introduction à R Commander

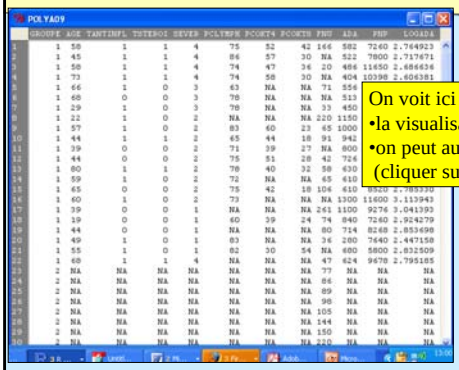
- Parmi ces packages (voir les « Tasks Views »), il y a R Commander, une application statistique graphique, sous le nom de package Rcmdr
- Comme les autres packages R, vous devez l'installer depuis un site miroir CRAN.
- Patience parce qu'il nécessite un grand nombre d'autres packages qui doivent être chargés
- R Commander est évidemment gratuit
- Une fois installé, chargez le package en tapant dans Rgui : Library("Rcmdr").
- R Commander n'est pas R mais emploie R
- R Commander existe en français

Introduction à R Commander



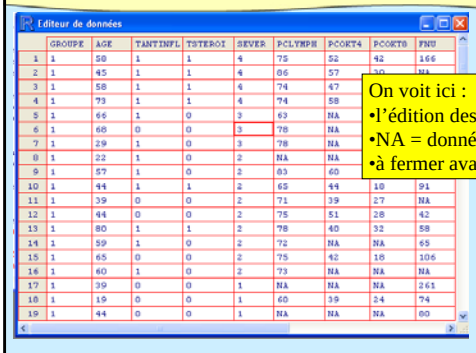
On voit ici :
- J'ai chargé le fichier SPSS POLYA07.SAV
- J'ai visualisé les données (cliquer sur Visualiser)
- On a demandé une synthèse des variables

Introduction à R Commander



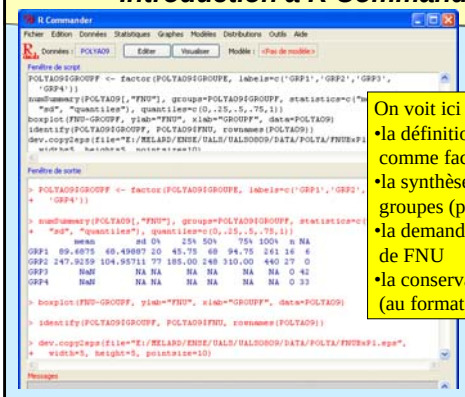
On voit ici :
- la visualisation des données
- on peut aussi les éditer (cliquer sur Editer)

Introduction à R Commander



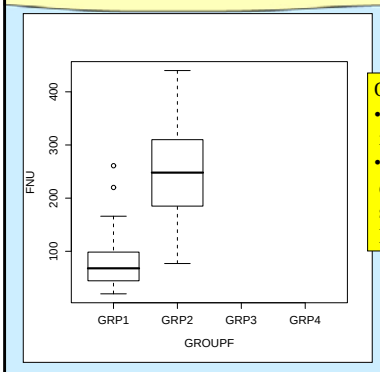
On voit ici :
- l'édition des données
- NA = donnée manquante
- à fermer avant de poursuivre

Introduction à R Commander



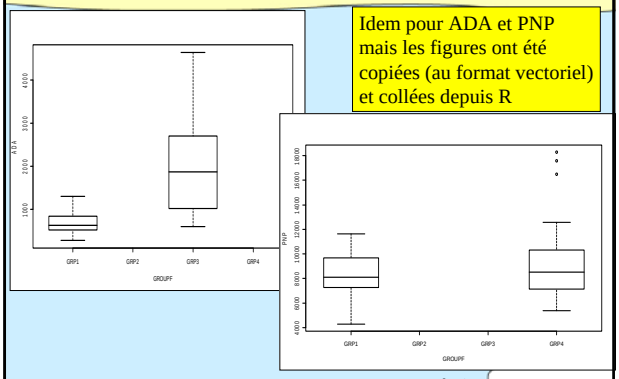
On voit ici :
- la définition de GROUPE comme facteur
- la synthèse de FNU par groupes (pas très jolie !)
- la demande d'un box plot de FNU
- la conservation de la figure (au format EPS)

Introduction à R Commander



On voit ici :
- la figure (au format EPS) insérée dans la présentation
- les sorties graphiques doivent être sauvegardées séparément et individuellement

Introduction à R Commander



Idem pour ADA et PNP mais les figures ont été copiées (au format vectoriel) et collées depuis R

Introduction à R Commander

- ⇨ R Commander n'est pas SPSS ni SAS mais il peut lire les fichiers SPSS (.SAV) (certainement ceux de la version 13 ou 14) et les fichiers Excel (version 2003)
- ⇨ Il peut afficher les données en tableau et même les éditer
- ⇨ Attention un des concepts les plus dérangeants est celui de "facteur" à utiliser pour grouper les données (sous-échantillons), dans les tableaux de contingence, ou comme facteur dans une analyse ANOVA.
- ⇨ Même si des variables ordinaires sont clairement des facteurs, elle doivent être explicitement déclarées comme facteurs, possiblement avec un changement de nom

N.B. Valeurs manquantes (missing values)

- ⇨ L'effet est le suivant : les procédures statistiques ne porteront pas sur les cas pour lesquels une des variables utilisées prend comme valeur une des valeurs manquantes
- ⇨ Données manquantes notées NA (« not available »)
- ⇨ Il y a aussi des valeurs impossibles NaN (« not a number »)

Fenêtre de syntaxe

- ⇨ Les instructions de R sont automatiquement collées en fonction des commandes choisies dans les menus
- ⇨ On peut modifier la syntaxe
  - ✓ soit pour réaliser des fonctions qui ne sont pas disponibles par les menus
  - ✓ soit pour appliquer la commande à d'autres variables (par copier/coller multiples suivi d'une édition)
- ⇨ On peut sauver le contenu de la fenêtre de syntaxe (script)
- ⇨ On peut exécuter le contenu d'un fichier de script

Fenêtre de sortie

- La sortie sous forme de texte apparaissent automatiquement
- C'est du texte pur (pas très esthétique …)
- … mais pas la sortie sous forme graphique qu'il faut sauvegarder depuis R (aux formats ps/eps/pdf) ou copier/coller individuellement
- On peut éditer la sortie
- On peut sauvegarder la sortie

Exemple (1)

- Comparaison avec le code SPSS (ci-dessous)
  - ✓ Toutes ces opérations ont déjà été faites dans POLYA07.SAV
  - ✓ … et les variables existent déjà

```
TITLE           'ENZYMES DANS LA POLYARTHRITE' .
DATA LIST       FIXED/
                GROUPE 1, AGE 2-3, TANTINFL 4,
                TSTEROI 5, SEVER 6, PCLYMPH 7-8,
                PCOKT4 9-10, PCOKT8 11-12, FNU 13-15,
                ADA 16-19, PNP 20-24 .
VAR LABELS      TANTINFL,TRAITEMENT ANTI INFLAMMATOIRE/
                TSTEROI ,TRAITEMENT STEROIDES/ .
MISSING VALUES  PCLYMPH,PCOKT4,PCOKT8(0)/
                FNU(999)/ .
N OF CASES      124 .
IF              (GROUPE = 1 OR GROUPE = 3) LOGADA = LG10(ADA) .
IF              (GROUPE = 1 OR GROUPE = 4) INVPNP = -1000000/PNP .
BEGIN DATA
```


Exemple (2)

- Suite de la comparaison avec le code SPSS (ci-dessous)
  - ✓ Ces commandes sont toutes réalisables dans R Commander
  - ✓ … mais à condition d'avoir défini GROUPF, TANTINFL, TSTEROI et SEVER comme facteurs

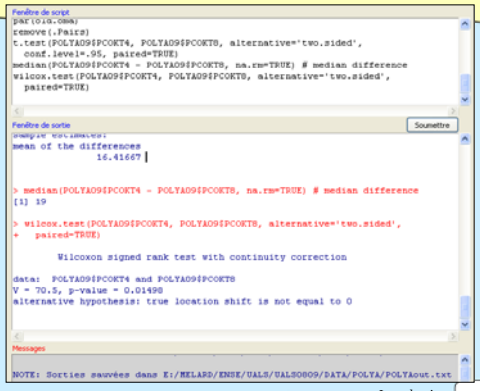
```
BEGIN DATA
158114755242166 582 7260
...
4                18300
END DATA .
DESCRIPTIVES    PCLYMPH,PCOKT4,PCOKT8 / STATISTICS ALL .
FREQUENCIES     SEVER / STATISTICS ALL .
CROSSTABS       TABLES = TANTINFL BY TSTEROI /
                         SEVER BY TANTINFL, TSTEROI / STATISTICS = CHISQ.
MEANS           TABLES = FNU, ADA, PNP, LOGADA,INVPNP BY GROUPE .
CORRELATIONS    AGE, ADA, PNP WITH AGE .
NONPAR CORR     AGE, ADA, PNP WITH AGE .
```


Exemple (3)

- Suite de la comparaison avec le code SPSS (ci-dessous)
  - ✓ Les tests de Student de comparaison des moyennes et le test de Mann-Whitney (ou Wilcoxon, échantillons indépendants) ne sont pas possible dans R Commander parce que GROUPF possède 4 niveaux
  - ✓ Le test de Kolmogorov-Smirnov n'est pas disponible
  - ✓ En revanche les deux tests pour échantillons appariés existent

```
T-TEST          GROUPS = GROUPE(1,2) / VARIABLES = FNU .
T-TEST          GROUPS = GROUPE(1,3) / VARIABLES = ADA, LOGADA .
T-TEST          GROUPS = GROUPE(1,4) / VARIABLES = PNP, INVPNP .
NPAR TESTS      K-S = FNU BY GROUPE(1,2)/
                K-S = ADA BY GROUPE(1,3)/
                K-S = PNP BY GROUPE(1,4) .
NPAR TESTS      M-W=FNU BY GROUPE(1,2)/
                M-W=ADA, LOGADA BY GROUPE(1,3)/
                M-W=PNP, INVPNP BY GROUPE(1,4)/ .
T-TEST          PAIRS = PCOKT4 WITH PCOKT8 .
NPAR TESTS      WILCOXON = PCOKT4 WITH PCOKT8 .
FINISH .
```


Exemple (4)



Les menus de base

File	Fichier
Edit	Edition
Data	Données
Statistics	Statistiques
Graphs	Graphes
Models	Modèles
Distributions	Distributions
Tools	Outils
Help	Aide

Menu File

- Change directory
- Open script file
- Save script
- Save script as
- Save output
- Save output as
- Save R workspace
- Save R workspace as
- Exit - from Commander
  - from Commander and R

Menu Edit

- Cut
- Copy
- Paste
- Delete
- Find
- Select all
- Undo
- Redo
- Clear Window

Menu Data (1)

- New data set
- Open data set
- Import data - from text file or clipboard
  - |- from SPSS data set
  - |- from Minitab data set
  - |- from STATA data set
  - |- from Excel, Access, or dBase data set
- Data in packages - List data sets in packages
  - |- Read data set from attached package
- Active data set . . .
- Manage variables in active data set . . .

Menu Data (2)

- Active data set - Select active data set
  - |- Refresh active data set
  - |- Help on active data set (if available)
  - |- Variables in active data set
  - |- Set case names
  - |- Subset active data set
  - |- Remove cases from active data set
  - |- Stack variables in active data set
  - |- Remove cases with missing data
  - |- Save active data set
  - |- Export active data set
- Manage variables in active data set . . .

Menu Data (3)

- Manage variables in active data set - Recode variable
  - |- Compute new variable
  - |- Add observation numbers to data set
  - |- Standardize variables
  - |- Convert numeric variables to factors
  - |- Bin numeric variable
  - |- Reorder factor levels
  - |- Define contrasts for a factor
  - |- Rename variables
  - |- Delete variables from data set

Menu Statistics (1)

- Summaries . . .
- Contingency Tables . . .
- Means . . .
- Proportions . . .
- Variances . . .
- Nonparametric tests . . .
- Dimensional analysis . . .
- Fit models . . .

Menu Statistics (2)

- Summaries - Active data set
  - |- Numerical summaries
  - |- Frequency distributions
  - |- Count missing data
  - |- Table of statistics
  - |- Correlation matrix
  - |- Correlation test
  - |- Shapiro-Wilk Normality test
- Contingency Tables - Two-way table
  - |- Multi-way table
  - |- Enter and analyze two-way table

Menu Statistics (3)

- Means - Single-sample t-test
  - |- Independent-samples t-test
  - |- Paired t-test
  - |- One-way ANOVA
  - |- Multi-way ANOVA
- Proportions - Single-sample proportion test
  - |- Two-sample proportions test
- Variances - Two-variances F-test
  - |- Bartlett's test
  - |- Levene's test
- Nonparametric tests - Two-sample Wilcoxon test
  - |- Paired-samples Wilcoxon test
  - |- Kruskal-Wallis test
  - |- Friedman rank sum test

Menu Statistics (4)

- Dimensional analysis - Scale reliability
  - |- Principal-components analysis
  - |- Factor analysis
  - |- Cluster analysis
    - |- k-means cluster analysis
    - |- Hierarchical cluster analysis
    - |- Summarize hierarchical clustering
    - |- Add hierarchical clustering to data set
- Fit models - Linear regression
  - |- Linear model
  - |- Generalized linear model
  - |- Multinomial logit model
  - |- Proportional-odds logit model

Menu Graphs (1)

- Color palette
- Index plot
- Histogram
- Stem-and-leaf display
- Boxplot
- Quantile-comparison plot
- Scatterplot
- Scatterplot matrix
- Line graph
- XY conditioning plot
- Plot of means
- Band graph
- Bar graph
- Pie chart . . .

Menu Graphs (2)

- 3D graph - 3D scatterplot
  - |- Identify observations with mouse
  - |- Save graph to file
- Save graph to file - as bitmap
  - |- as PDF/Postscript/EPS
  - |- 3D RGL graph

Menu Models (1)

- Select active model
- Summarize model
- Add observation statistics to data
- Confidence intervals
- Akaike information criterion (AIC)
- Bayasian information criterion (BIC)
- Hypothesis tests - ANOVA table
  - |- Compare two models
  - |- Linear hypothesis
- Numerical diagnostics - Variance-inflation factors
  - |- Breusch-Pagan test for heteroscedasticity
  - |- Durbin-Watson test for autocorrelation
  - |- RESET test for nonlinearity
  - |- Bonferroni outlier test

Menu Models (2)

- Graphs - Basic diagnostic plots
  - |- Residual quantile-comparison plot
  - |- Component+residual plots
  - |- Added-variable plots
  - |- Influence plot
  - |- Effect plots

Menu Distributions (1)

- Continuous distributions . . .
- Discrete distributions . . .

Menu Distributions (2)

- Continuous distributions
  - |- Normal distribution . . .
  - |- t distribution . . .
  - |- Chi-squared distribution . . .
  - |- F distribution . . .
  - |- Exponential distribution . . .
  - |- Uniform distribution . . .
  - |- Beta distribution . . .
  - |- Cauchy distribution . . .
  - |- Logistic distribution . . .
  - |- Lognormal distribution . . .
  - |- Gamma distribution . . .

Menu Distributions (3)

- Continuous distributions
  - | - Normal distribution - Normal quantiles
  - | |- Normal probabilities
  - | |- Plot normal distribution
  - | |- Sample from normal distribution
  - | - t distribution - t quantiles
  - | |- t probabilities
  - | |- Plot t distribution
  - | |- Sample from t distribution
  - |- Chi-squared distribution . . .

Menu Distributions (4)

- Continuous distributions
  - |- . . .
  - |- Weibull distribution - Weibull quantiles
  - | |- Weibull probabilities
  - | |- Sample from Weibull distribution
  - |- Gumbel distribution - Gumbel quantiles
  - | |- Gumbel probabilities
  - | |- Plot Gumbel distribution
  - | |- Sample from Gumbel distribution
- Discrete distributions . . .

Menu Distributions (5)

- Discrete distributions
  - |- Binomial distribution
  - | - Binomial quantiles
  - | |- Binomial tail probabilities
  - | |- Binomial probabilities
  - | |- Plot binomial distribution
  - | |- Sample from binomial distribution
  - |- Poisson distribution . . .
  - |- Geometric distribution . . .
  - |- Hypergeometric distribution . . .
  - |- Negative binomial distribution . . .

Menu Tools

- Load package(s)
- Load Rcmdr plug-in package(s)
- Options

Menu Help & boutons

- Commander help
- Introduction to the R Commander
- Help on active data set (if available)
- About Rcmdr

Buttons: Edit data set; View data set; Submit (lines from the script window)

Information field buttons: active data set; active model

Right-click "context" menus for script and output windows.

Regression linéaire et non linéaire

⇨ Comparaison avec SPSS

Exemple : *regression*/ANOVA *variables= y, x1, x2 / dependent = y / by = f1, f2 / with x1, x2*

⇨ R (et R Commander) : (notation de Wilkinson & Rogers)

$y \sim x1 + x2$	régression linéaire avec constante
$y \sim x1 + x2 + 0$	régression linéaire sans constante
$y \sim \text{poly}(x1,2)$	régression polynomiale de degré 2
$\log(y) \sim x1 + x2$	régression linéaire de log(*y*) avec constante
$y \sim f1$	modèle ANOVA à un facteur
$y \sim f1*f2$	modèle ANOVA à deux facteurs avec interaction
$y \sim f1+f2$	modèle ANOVA à deux facteurs sans interaction
$y \sim f1*f2 - f1:f2$	modèle ANOVA à deux facteurs sans interaction
$y \sim f1 + x1$	modèle d'analyse de covariance avec constante
$y \sim f1*x1$	régression linéaires pour chaque niveau de *f*1

Introduction à MYSTAT (1)

⇨ MyStat est une version étudiante (actuellement) gratuite de Systat

⇨ Mystat est pour Windows seulement

⇨ Il possède beaucoup des fonctionnalités de Systat mais pas toutes bien entendu

⇨ Systat est apparu vers la fin 1970 et est devenu un produit commercial vers 1983.

⇨ Il a été vendu à SPSS Inc. en 1995 et appartient maintenant à Cranes Software, Bengalore (India) depuis 2002

⇨ Version la plus récente : version 13, octobre 2009

⇨ Pour plus d'information : http://www.systat.com/SystatProducts.aspx

Introduction à MYSTAT (2)

⇨ Un peu comme SPSS, les utilisateurs novices peuvent employer SYSTAT par l'interface par menus pour réaliser des analyses simples et produire de beaux graphiques 2D et 3D pour des présentations et des rapports

⇨ Les utilisateurs plus avancés peuvent accélérer considérablement leur recherche en employant le langage de commande intuitif de SYSTAT avec la possibilité de sauvegarder des macro-commandes complexes

⇨ On peut trouver Mystat à l'adresse : http://www.systat.com/MystatProducts.aspx.

⇨ Mystat vient avec la documentation complète de Systat (more than 2000 pp.)

Exemple : poissons d'un lac finlandais

fishcatch.txt

NAME: Fish Catch

TYPE: Sample

SIZE: 159 observations, 8 variables

SUBMITTED BY: Juha Puranen […]

Description of fishcatch.dat

This data set includes measurements on 159 fish caught from the same lake (Laengelmavesi) near Tampere, Finland. The data include 7 species of fish: bream, whitefish, roach, parkki (no English translation), smelt, pike, and perch. For each fish, two measurements were taken: weight in grams and length (in centimeters) from the nose to the end of the tail.

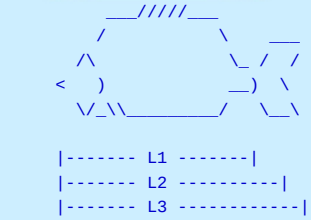
Exemple : poissons d'un lac finlandais

VARIABLE DESCRIPTIONS:
1 Obs Observation number ranges from 1 to 159
2 Species (Numeric)

Code	Finnish	Swedish	English	Latin
1	Lahna	Braxen	Bream	Abramis brama
2	Siika	Iiden	Whitewish	Leusiscus idus
3	Saerki	Moerten	Roach	Leuciscus rutilus
4	Parkki	Bjoerknan	?	Abramis bjrkna
5	Norssi	Norssen	Smelt	Osmerus eperlanus
6	Hauki	Jaedda	Pike	Esox lucius
7	Ahven	Abborre	Perch	Perca fluviatilis

3 Weight Weight of the fish (in grams)
4 Length1 Length from the nose to the beginning of the tail (in cm)
5 Length2 Length from the nose to the notch of the tail (in cm)
6 Length3 Length from the nose to the end of the tail (in cm)
7 Height% Maximal height as % of Length3
8 Width% Maximal width as % of Length3
9 Sex 1 = male 0 = female

Exemple : poissons d'un lac finlandais



|------- L1 -------|
|------- L2 ----------|
|------- L3 ------------|

Values are aligned and delimited by blanks.
Missing values are denoted with NA.
There is one data line for each case.
SPECIAL NOTES:
I have usually calculated
Height = Height%*Length3/100
Width = Width%*Length3/100

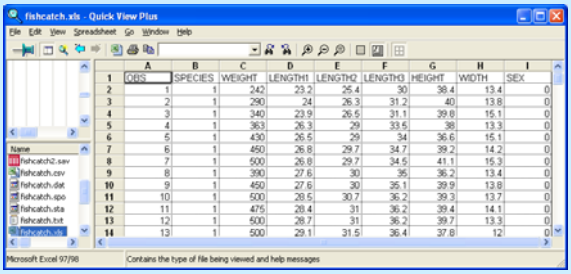
Ne faites pas comme l'étudiant d'il y a 3 ans qui a régressé le poids sur les variables LENGTH1, LENGTH2, LENGTH3, HEIGHT, WIDTH, SEX et SPECIES !

Exemple : poissons d'un lac finlandais

fishcatch.dat

```
   2      1    290.0    24.0    26.3    31.2    40.0   13.8   NA
   3      1    340.0    23.9    26.5    31.1    39.8   15.1   NA
   4      1    363.0    26.3    29.0    33.5    38.0   13.3   NA
   5      1    430.0    26.5    29.0    34.0    36.6   15.1   NA
   6      1    450.0    26.8    29.7    34.7    39.2   14.2   NA
   7      1    500.0    26.8    29.7    34.5    41.1   15.3   NA
   8      1    390.0    27.6    30.0    35.0    36.2   13.4   NA
   9      1    450.0    27.6    30.0    35.1    39.9   13.8   NA
  10      1    500.0    28.5    30.7    36.2    39.3   13.7   NA
  11      1    475.0    28.4    31.0    36.2    39.4   14.1   NA
  12      1    500.0    28.7    31.0    36.2    39.7   13.3   NA
  13      1    500.0    29.1    31.5    36.4    37.8   12.0   NA
  14      1       NA    29.5    32.0    37.3    37.3   13.6   1
. . .
 104      7      5.9     7.5     8.4     8.8    24.0   16.0   NA
. . .
 OBS   SPECIES WEIGHT   LENGTH1 LENGTH2 LENGTH3 HEIGHT   WIDTH SEX
```

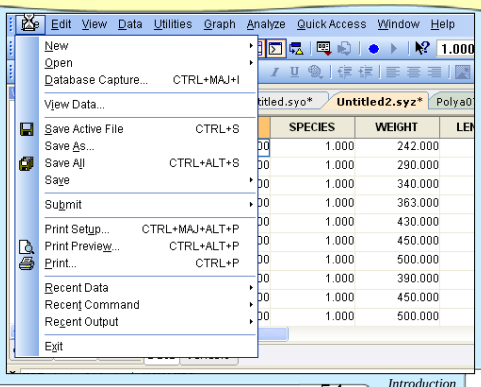

Les données



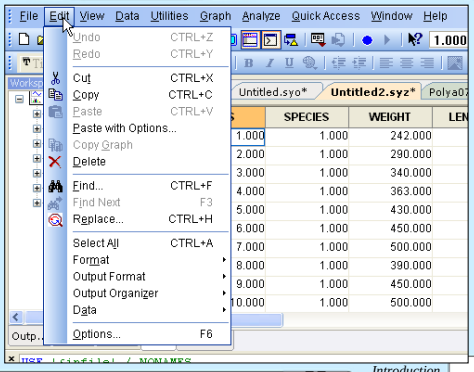
Les menus de base

- File — Fichier
- Edit — Edition
- View — Visualisation
- Data — Données
- Utilities — Outils
- Graph — Graphes
- Analyze — Analyses
- Quick Access — Accès rapide
- Window — Fenêtres
- Help — Aide

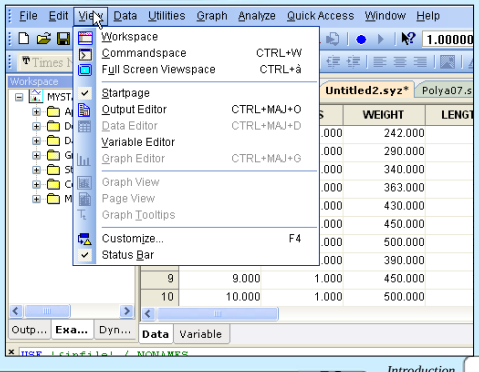
Menu File



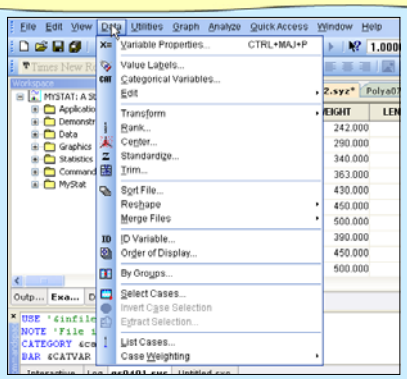
Menu Edit



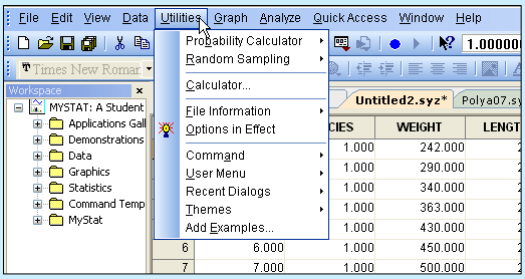
Menu View



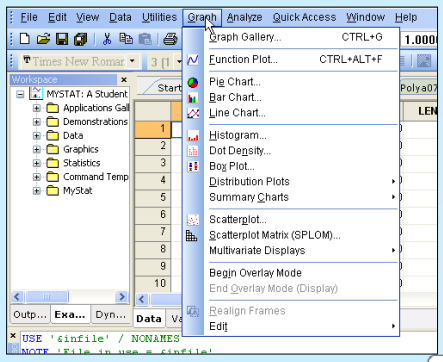
Menu Data



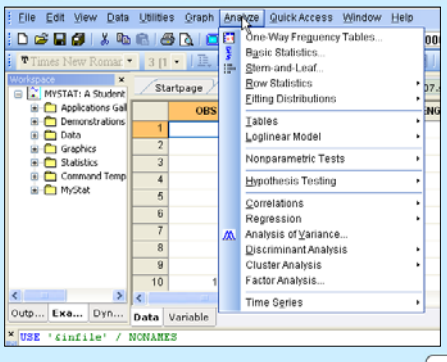
Menu Utilities



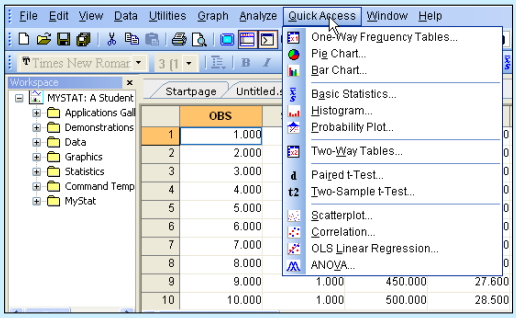
Menu Graph



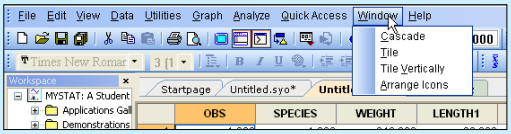
Menu Analyze



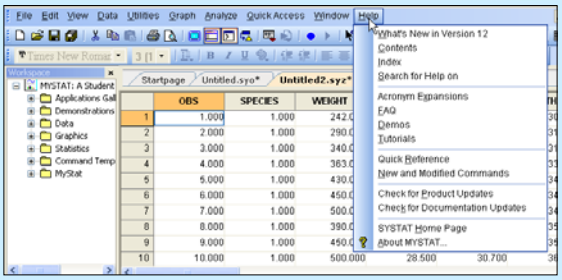
Menu Quick Access



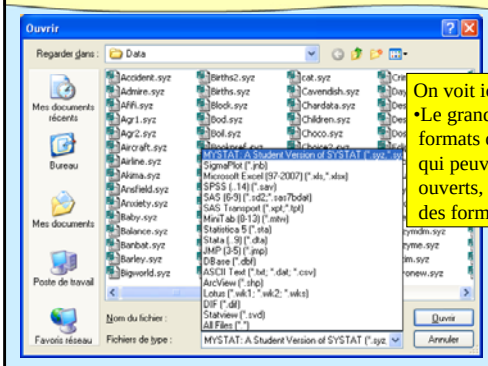
Menu Window



Menu Help



Introduction à MYSTAT



On voit ici :
- Le grand nombre de formats de fichiers qui peuvent être ouverts, y compris des formats très récents

Résultats de la régression 1

Dependent Variable	WEIGHT
N	158
Multiple R	0.931
Squared Multiple R	0.867
Adjusted Squared Multiple R	0.863
Standard Error of Estimate	132.861

Regression Coefficients B = (X'X)⁻¹X'Y

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t	p-value
CONSTANT	-724.539	77.133	0.000	.	-9.393	0.000
LENGTH1	32.389	45.134	0.904	0.001	0.718	0.474
LENGTH2	-9.184	48.367	-0.275	0.000	-0.190	0.850
LENGTH3	8.747	16.283	0.283	0.003	0.537	0.592
HEIGHT	4.947	2.768	0.114	0.213	1.787	0.076
WIDTH	8.636	6.972	0.055	0.444	1.239	0.217

Durbin-Watson D Statistic	0.355
First Order Autocorrelation	0.817

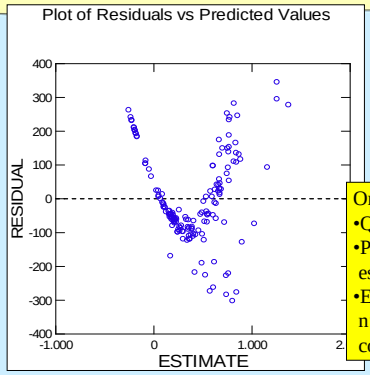
Information Criteria

AIC	2.001.289
AIC (Corrected)	2.002.035
Schwarz's BIC	2.022.727

Analysis of Variance

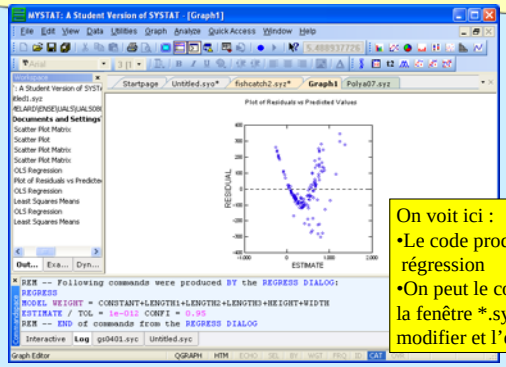
Source	SS	df	Mean Squares	F-ratio	p-value
Regression	17.560.908.093	5	3.512.181.619	198.966	0.000
Residual	2.683.127.473	152	17.652.154		

Résultats de la régression 1



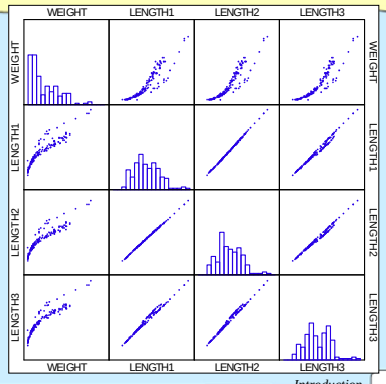
On voit ici :
- Que ce n'est pas bon
- Parce que la relation est plutôt multiplicative
- Et parce que l'espèce n'est pas prise en compte

Résultats de la régression 1

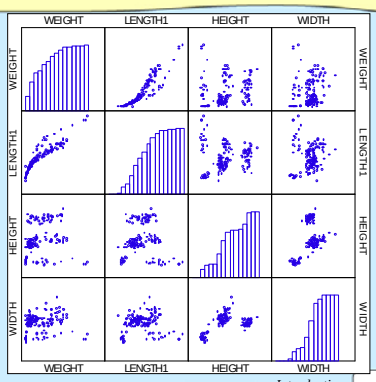


On voit ici :
- Le code produit par la régression
- On peut le copier dans la fenêtre *.syc, le modifier et l'exécuter

Poids en fonction des longueurs



Graphique (avec hauteur et largeur incorrectes)

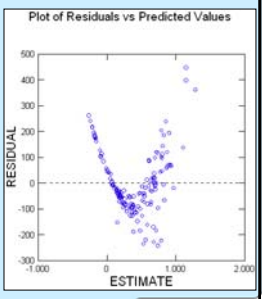


Régression 2 avec variable espèce

Dependent Variable	WEIGHT
N	158
Multiple R	0.942
Squared Multiple R	0.886
Adjusted Squared Multiple R	0.884
Standard Error of Estimate	122.560

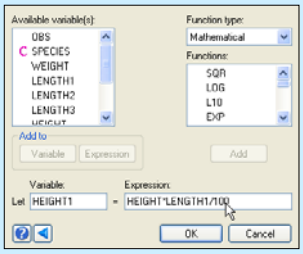
Regression Coefficients B = (X'X)-1X'Y

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t	p-value
CONSTANT	-508.963	47.054	0.000	.	-10.816	0.000
LENGTH1	22.394	2.034	0.625	0.230	11.011	0.000
HEIGHT1	23.416	8.081	0.221	0.128	2.898	0.004
WIDTH1	40.044	21.285	0.162	0.100	1.881	0.062
SPECIES1	-0.539	7.180	-0.004	0.326	-0.075	0.940

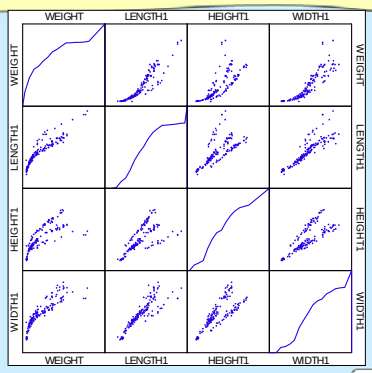


Définition de nouvelles variables

- ⇨ Puis on définit HEIGHT1 et WIDTH1 comme les vraies hauteurs et épaisseurs
- ⇨ On définit les logarithmes des variables LENGTH1, etc.

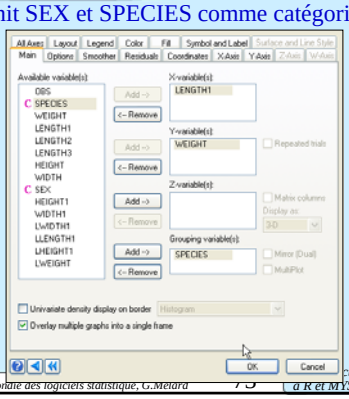


Graphique (avec hauteur et largeur correctes)

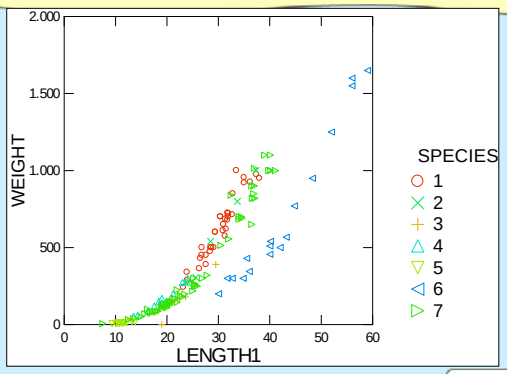


Poids en fonction d'une longueur et de l'espèce

⇨On définit SEX et SPECIES comme catégorielles (C)



Poids en fonction d'une longueur et de l'espèce



Résultats de la régression 3 avec variable qualitative

Dependent Variable	WEIGHT
N	158
Multiple R	0.969
Squared Multiple R	0.939

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
LENGTH1	27.963.151	1	27.963.151	3.331	0.070
LENGTH2	16.681.673	1	16.681.673	1.987	0.161
LENGTH3	13.644.372	1	13.644.372	1.625	0.204
HEIGHT	5.166.905	1	5.166.905	0.615	0.434
WIDTH	5.919.017	1	5.919.017	0.705	0.402
SPECIES	1.457.457.148	6	242.909.525	28.935	0.000
Error	1.225.670.325	146	8.395.002		

Régression 4, en logarithmes

⇨Si la masse des poissons était proportionnelle à leur volume, on devrait avoir une relation multiplicative, et s'ils étaient en forme de parallépipèdes rectangles

WEIGHT = densité*LENGTH1*HEIGHT1*WIDTH1

(en termes des variables d'origine :

WEIGHT = densité*(LENGTH1)³*HEIGHT*WIDTH

⇨On définit les logarithmes des variables LWEIGHT, LENGTH1, etc. , d'où le modèle

$LWEIGHT = b_0 + b_1\ LLENGTH1 + b_2 LHEIGHT1 + b_3 LWIDTH1 + e$

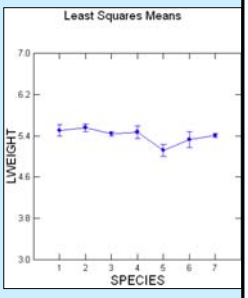
⇨Comme la densité dépend de l'espèce, il est naturel d'avoir des coefficients différents pour chaque espèce

Régression 4, en logarithmes

Dependent Variable	LWEIGHT
N	157
Multiple R	0.998
Squared Multiple R	0.996

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
LLENGTH1	0.899	1	0.899	122.413	0.000
LHEIGHT1	0.146	1	0.146	19.907	0.000
LWIDTH1	0.223	1	0.223	30.318	0.000
SPECIES	0.393	6	0.065	8.909	0.000
Error	1.080	147	0.007		



Conclusion

⇨R, surtout avec R Commander peut déjà servir de logiciel statistique de base

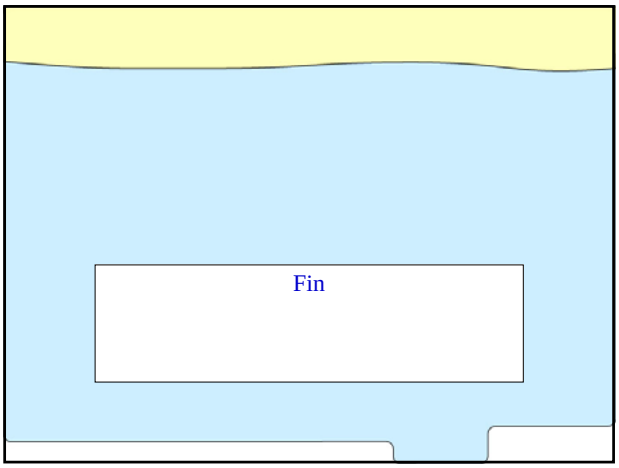
⇨R Commander ne possède que des procédures statistiques de base

⇨Surtout si les données sont lisibles dans un format classique

⇨Complété par R, moyennant de la programmation, de nombreuses méthodes sont disponibles

⇨MyStat est une version gratuite de Systat qui est déjà très puissante, même si pas aussi puissante que Systat ou SPSS

⇨On peut le conseiller pour une analyse exploratoire


Fin