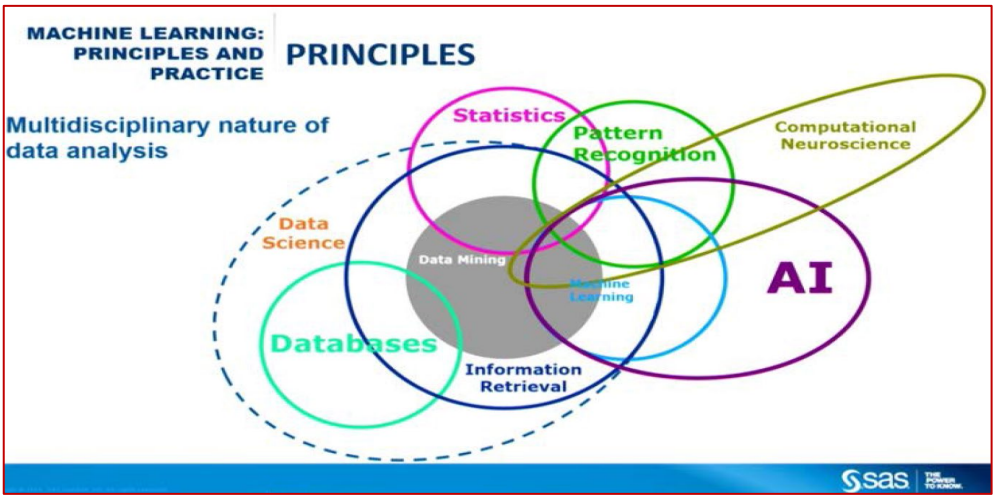
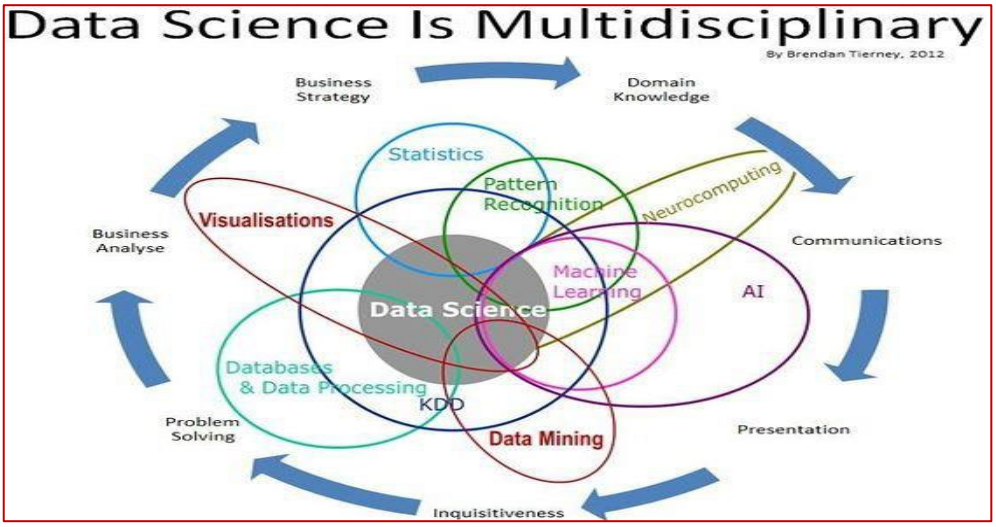
MTH8302 Modèles de régression et d'analyse de la variance

E X E R C I C E S





Les données sont disponibles en fichier Statistica (STW) et en fichier Excel (XLSX)

https://cours.polymtl.ca/mth6301/WEB-mth8302/MTH8301-Data-Devoirs.stw

https://cours.polymtl.ca/mth6301/WEB-mth8302/MTH8301-Data-Devoirs.xlsx

Mth8302-exer-01 Analyse diagnostique / graphique dans les modèles statistiques

Données = Anscombe

Le fichier contient 4 couples de variables, (X1, Y1), (X2, Y2), (X3, Y3) et (X4, Y4).
On considère un modèle de régression linéaire simple pour prédire Y en fonction de X pour chacun des 4 couples (X, Y):

$$\text{modèle 1 :} \quad Y1 = \beta_0 + \beta_1 X1 + \varepsilon$$
$$\text{modèle 2 :} \quad Y2 = \beta_0 + \beta_1 X2 + \varepsilon$$
$$\text{modèle 3 :} \quad Y3 = \beta_0 + \beta_1 X3 + \varepsilon$$
$$\text{modèle 4 :} \quad Y4 = \beta_0 + \beta_1 X4 + \varepsilon$$

QUESTIONS

1a) Compléter les valeurs manquantes du tableau ci bas.

modèle couple	β_0	β_1	R^2	SSreg	SSresid	SStot
1 (X1, Y1)						
2 (X2, Y2)						
3 (X3, Y3)						
4 (X4, Y4)						

R^2 : coefficient de détermination = fraction de la variation de Y expliqué par X

SSreg : somme de carrés de régression (expliquée) par le modèle

SSresid : somme de carrés résiduelle (erreur)

SStot : somme des carrés totale

Commentez le tableau.

1b) Tracez, pour chacun des 4 couples de variables, un nuage de points (« 2D scatterplots ») illustrant la variation de Y en fonction de X.
Commentez les graphiques. Faites un lien avec le commentaire en 1a).

1c) Pour chacun des 4 couples de variables, tracer un graphique des résidus en fonction de la variable explicative X. Commentez. Faites un lien avec le commentaire en 1b).

1d) Pour chacun des 4 couples de variables, tracer un graphique des résidus sur échelle de probabilité gaussienne. Commentez.

1e) Considérons le couple (X3, Y3). La droite de régression est-elle affectée s'il s'avère que l'observation (X3 =13 Y3 = 12,74) est le résultat d'une erreur et peut être éliminée.
Refaire les calculs sans cette observation.
Le modèle est-il adéquat avec cette observation?
Que devient alors la valeur de R^2 sans cette observation?

1f) Proposer une conclusion générale pour ce numéro.

Mth8302-exer02 Régression logistique

Données = Vaccins

Étude sur l'efficacité d'un programme de sensibilisation vaccin contre la grippe.
Un Centre Local de Santé Communautaire (CLSC) a envoyé un dépliant publicitaire encourageant les personnes âgées à recevoir un vaccin contre la grippe. Une étude subséquente, un échantillon de 159 personnes furent choisies au hasard et on leur demanda s'il avait reçu le vaccin.
La variable de réponse Y est

reçu le vaccin Y_reçuVaccin = 1 = oui

pas reçu le vaccin Y_reçuVaccin = 0 = non

Le fichier contient trois autres variables potentiellement explicatives.

X1_âge : âge de la personne

X1_catAge : âge représenté par 5 valeurs typiques 52 – 57 – 62 – 67 - 72

52 = 54 et moins 57 = 55 à 59 62 = 60 à 64 67 = 65 à 69 72 = 70 et plus

X2_indSanté : indice de sensibilisation à sa santé - échelle 0 à 100

0 = aucune sensibilité 100 = sensibilité très élevée

X3_genre : sexe de la personne : F = Femme H = homme

colonnes ajoutées 9-10-11-12 tableau croisé = colonne 7 X colonne 3
Y0 = nombre pas vaccinés Y1 = nombre vaccinés
selon les catégories d'âge (colonne 3) pour l'ensemble de tous les répondants

QUESTIONS

2a) **Ajusté un modèle de régression logistique entre Y et X1_âge.**

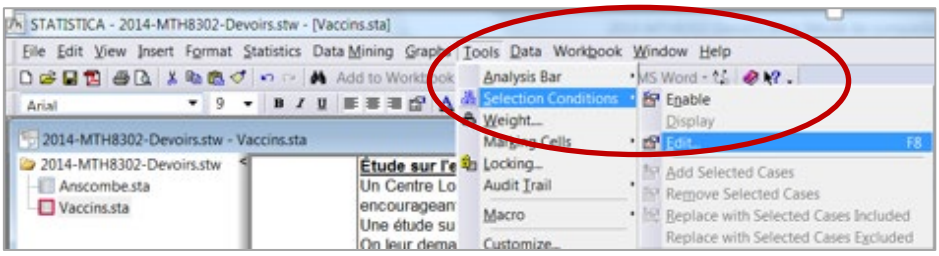
Les variables X2 et X3 ne sont pas tenues en compte dans le modèle.

Module de Statistica *Nonlinear Estimation ….Quick logit regression*

2b) **Refaire 2a) pour les hommes seulement.**

Recommandation : imposer un filtre sur les données avec

Tools…Selection conditions…edit



2c) **Refaire 2b) pour les femmes seulement.**

2d) **Comparer les 3 modèles de 2a) 2b) 2c) : effet de l'âge et du sexe sur Y**

2e) **Refaire l'analyse 2a) avec le tableau agrégé des colonnes 10 / 11 / 12 / 13**

Comparer les modèles 2a) et 2e).

Mth8302-exer03 Régression non linéaire

Données = croissance

L'observation des phénomènes de croissance de matière biologique, animale, végétale,.. donne souvent lieu à une courbe en forme de S (sigmoide). La fonction de croissance mesurée est représentée par une fonction non linéaire car le phénomène est souvent caractérisé par une évolution lente au début suivi par une croissance rapide et se terminant par une stabilisation progressive. Plusieurs fonctions ont été proposées pour modéliser le phénomène. Ces fonctions sont paramétrées par des paramètres a , b , c, d, …Les fonctions suivantes sont parmi les plus employées pour modéliser les phénomènes de croissance.

Voir ModelesNonLineaire.pdf sur le site https://cours.polymtl.ca/mth6301/MTH8302.htm

Gompertz	$Y = a*\exp[-\exp(b - cx)]$
Logistique 3P	$Y = a / [1 + \exp(b - cx)]$
	remarque : ne pas confondre avec le modèle logistique obtenu avec a = 1
Morgan-Mercer	$Y = (b*c + a\,x^d) / (c + x^d)$
Weibull	$Y = a - b \exp(-c\,x^d)$
Probit	$Y = a\,\Phi(b + c\,x)$

Φ (u) : fonction de répartition de la distribution normale centrée-réduite

---

Le fichier de données proposé pour ce numéro est un exemple typique de données de croissance. Employer le module *Nonlinear Estimation* de *STATISTICA* pour réaliser cet exercice.

QUESTIONS

3a) **Tracer le graphique des données.**
3b) **Ajuster la fonction de Gompertz.**
3c) **Ajuster la fonction Logistique.**
3d) **Ajuster la fonction Weibull.**
3e) **Proposer la meilleure fonction choix de fonction pour modéliser les données. Préciser le critère employé pour faire le choix.**

Information sur la fonction Weibull

La distribution de Weibull a pour fonction de densité f avec des paramètres b, c et q positifs :

$$f(x ; b, c, q) = c / b*[(x - q) / b]^{(c-1)}\, e\,\{-[(x - q) / b]^{**c}\} \quad \text{pour } 0 < q <= x < \infty$$

b > 0	**est le paramètre d'échelle de la distribution**
c > 0	**est le paramètre de forme de la distribution**
Q	**est le paramètre de position de la distribution**
Ee	**= 2,71828 … constante d'Euler**

La fonction de répartition F de la distribution de Weibull est:

$$(1)\quad F(x) = e\,\{-[(x - q) / b]^{**c}\}$$

On peut modifier cette fonction en ajoutant un facteur de décalage d et un facteur multiplicatif a:

$$(2)\quad G(x) = d + a*e\,\{-[(x - q) / b]^{**c}\}$$

G(x) n'est plus une fonction de répartition sauf si d = 0 et a = 1
En particulier, si on pose q = 0, on obtient:

$$(3)\quad H = d + a*e\{-[(x) / b]^{**c}\}$$

Avec un changement de notation, on obtient l'équation (4).

$$(4)\quad Y = a - b \exp(-c\,x^{**d})$$

Statistica, avec l'estimation de Quasi-Newton, présente une meilleure convergence avec la forme (3) qu'avec la forme (4) si on spécifie les contraintes (b > 0, c > 0). Il est possible de le faire avec la forme (4) mais il faut ajuster les contraintes.

Complément d'information No3

Ajustement de modèles non linéaires

Tenir en compte la remarque et ajuster le modèle Weibull avec la forme (3). il est possible de tenir en compte des contraintes sur les paramètres lors de l'ajustement d'un modèle non linéaire. La documentation (help) de Statistica donne de l'information et des exemples. Globalement, on impose des conditions logiques sur les paramètres que l'introduit dans la fonction comme par exemple, (a>0) qui vaut 1 si a est positif et vaut 0 si a est négatif.

Ajustement de modèles non linéaires de type Weibull

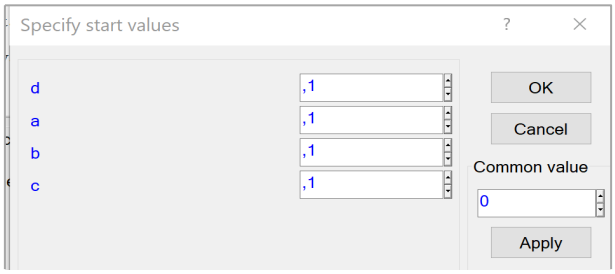
Il est possible d'ajuster le modèle Weibull avec chacune des 2 fonctions proposées

Forme (3) du modèle Weibull **v3 = d + a*exp(-(v2/b)**c)**

ou **Y = d + a*exp(-(x/b)**c)**

Forme (4) du modèle Weibull **v3 = a - b*exp(-(c*(v2**d)))**

Par défaut, le module Nonlinear de Statistica choisit de petites valeurs initiales pour les constantes du modèle proposé. Par exemple, pour le modèle (3) il choisit



Ce choix peut s'avérer critique. Le processus itératif des solutions linéarisées peut ne pas converger car on est trop loin de la solution optimale ou encore parce que l'on est en un point de l'espace des paramètres ou que le système d'équations ne possède pas de solution. Statistica renvoie un message de mise en garde. Par exemple,

Model is: v3 = d + a*exp(-(v2/b)**c) (Croissance.sta in 2020-MTH8302-Devc
Dep. Var. : Y(poids)
Caution: Degenerate solution, results may not be correct !

	Estimate	Standard error	t-value df = 11	p-value	Lo. Conf Limit	Up. Conf Limit
d	423,2953	80,77134	0,00	0,00	245,5188	601,0719
a	22,0291	0,00000	0,00	0,00	22,0291	22,0291
b	155,0690	0,00000	0,00	0,00	155,0690	155,0690
c	-40,8780	0,00000	0,00	0,00	-40,8780	-40,8780

Pour résoudre le mauvais de valeurs initiales, il faut explorer plusieurs ensembles de valeurs
On explore la forme de la fonction proposée avec plusieurs choix des paramètres, on trace la fonction ainsi que les points observés pour identifier un choix raisonnable des paramètres.
Il n'est pas nécessaire d'être près des valeurs optimales. Souvent il s'agit d'avoir le bon signe des paramètres pour assurer la convergence. À titre d'exemple, pour les données du No 3 avec la forme (4), j'ai choisi a = 100 b = 100 c= 0,01 d = 1 et cela a permis d'assurer la convergence de l'estimation.

Mth8302-exer04 Classification de modèles

Il n'y a pas d'ensemble de données associées à cet exercice.

Soient les modèles suivants où Y est une variable continue à prédire,

X1, X2, sont des variables explicatives continues et les β sont des paramètres inconnus.

4a) **M1 : $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2}$**

4b) **M2 : $Y = 1 - \exp(\beta_0 X^{\beta_1})$**

4c) **M3 : $Y = \beta_0 / (1 + \beta_1 X_1 + \beta_2 X_2)$**

4d) **M4 : $Y = \beta_0 + \beta_1 \exp(\beta_2 X)$**

4e) **M5 : $Y = \beta_0 X (1 + \beta_1 X)^2$**

Question 4

Pour chaque modèle, classer le dans une des 3 catégories et justifier votre réponse.

- **LINPAR : linéaire dans les paramètres β**
- **LINARIZ : linéaire après transformations (linéarisable); définir les transformations**
- **INTRNLIN : intrinsèquement non linéaire dans les paramètres β**

Voir la page suivante pour un complément d'information sur le concept de linéarité.

Compléments d'information sur le concept de linéarité d'un modèle dans ses paramètres

définition

Modèle général $Y = \varphi(X_1, X_2,\ldots, X_k; \beta_0, \beta_1, \beta_2 ,\ldots, \beta_p) + \varepsilon \quad \varepsilon \sim N(0,\sigma^2)$ (1)

Modèle LINÉAIRE dans les β si

$\varphi(X_1, X_2,\ldots, X_k; \beta_0, \beta_1, \beta_2 ,\ldots, \beta_p) = \Sigma\, \beta_j\, f_j(X_1, X_2,\ldots,X_k)$ (2)

$f_j(X_1, X_2,\ldots, X_k) = U_j$ ne dépend pas de paramètre inconnu (3)

alors $Y = \Sigma\, \beta_j\, U_j + \varepsilon$ (4)

Il n'existe pas de méthode générale pour décider si le modèle est linéaire ou ne l'est pas. Il faut essayer de trouver des transformations f_j qui permettent de passer de la forme (2) à la forme (4) en satisfaisant la contrainte (3). La recherche de la linéarité est de permettre de détermier exactement les car on a affaire à la résoutiond'un sytème d'équations linéaires.

La variable Y peut elle-même être transformée en une nouvelle variable Y' = H(Y) afin d'examiner la linéarité dans les β. Si Y est transformée alors φ sera transformée en H(φ(β))
L'analyse de la linéarité dans les β se fera avec H(φ(β)).
La recherche de transformations peut s'avérer infructueuse sans possibilité de faire un classement de linéarité ou non. Mais il n'existe pas de méthode générale qui permet de démontrer mathématiquement l'existence ou non de la linéarité d'un modèle.

Supposons que les paramètres β sont transformés par des fonctions dans de nouveaux paramètres γ alors le modèle peut être déclaré linéaire dans les paramètres γ s'il satisfont l'équation (2) en remplaçant β avec les γ. Supposons que les transformations entre les nouveaux paramètres γ et les anciens paramètres β , soient définies par les équations,

$$\gamma_j = g_j(\beta_1, \ldots, \beta_k) \quad \text{pour} \quad j = 1, 2, .., l \quad \text{et la valeur de } l \text{ peut être différente de } k.$$

Une déclaration de linéarité ou non d'un modèle ne s'applique pas obligatoirement aux coefficients d'origine β mais peut s'appliquer à la série des nouveaux coefficients γ obtenus par des transformations. Si l'expression du modèle dans les γ permet de le déclarer linéaire, alors les γ pourront être calculés directement en résolvant un système d'équations linéaires. Cela ne veut pas dire que l'on peut déterminer les coefficients d'origine β. Pour cela il faut inverser les équations

$$\gamma_j = g_j(\beta_1, \ldots, \beta_k)$$

Ce n'est pas assurer d'avance. Mais cela n'a pas d'importance. Mais si les fonctions inverses n'existent pas, on est en présence d'un système non linéaire dans les β. La définition de linéarité du modèle dans les β est simple quand on peut appliquer directement l'équation (2) dans les γ.

Mais si on transforme les β dans de nouveaux paramètres en essayant de transformer l'équation de départ, la réponse de la linéarité ou non du modèle est plus nuancée. La déclaration de linéarité peut s'appliquer aux

1. paramètres d'origine β
2. paramètres transformés γ avec l'existence d'une transformation inverse aux β (cas l = k)
3. paramètres transformés γ sans transformation inverse aux β

Si on n'est pas en présence de ces trois cas, alors le modèle n'est pas linéarisable et il sera déclaré intrinsèquement non-linéaire.

Mth8302-exer05 Étude de modélisation avec plusieurs méthodes

Données = BostonHousing

Description des données

Harrison, D, Rubenfeld, D. (1978)
Hedonic (House) Prices and the Demand for Clean Air
J. of Environmental Economic and Management, v.5, pp. 81-102
Combined information from 10 separate governmental and education sources 506 census tracts (CT) in city of Boston of the year 1970
But : étude de la relation entre 11 indicateurs de la qualité de vie et la valeur d'une résidence.
Le fichier contient 506 observations est divisé en 2 groupes

GROUP = M pour le développement des Modèles (405 observations : 80% des observations)
les données M sont surlignées (vert) et constituent un filtre;
la modélisation statistique est basée sur ces données (405 obs.)
voir *Tools...sélections conditions...edit*

GROUP = T pour Tester le modèle (101 observations : 20% des observations)
pour inclure ce groupe dans une analyse, éditer le filtre

VARIABLES explicatives

- **X1 CRIM : CRIMe Rate Per Capita by town**
- **X2 NOX: : Nitric OXide concentration (parts per 10 million)**
- **X3 AGE : Proportion of owner-occupied units built prior to 1940.**
- **X4DIS : Weighted DIStances to five Boston employment centers**
- **X5 RM : Average number of RooMs per dwelling**
- **X6 LSTAT : % of the Lower STATus of the population**
- **X7 RAD : Index of accessibility to RADical highways**
- **X8 CHAS : CHASrles river dummy variable (1 if census tract bounds the river; 0 otherwise)**
- **X9 INDUS : Proportion of non-retail INDUStrial business acres per town**
- **X10 TAX : Full value property TAX rate per $10,000**
- **X11 PT : Pupil-Teacher ratio by town**
- **X12 RLZ : Proportion of Residential Land Zoned for lots over 25,000 sq.ft.**
disponible dans le fichier mais elle ne sera pas employée dans les analyses

RÉPONSE

Y MV : Median Value of owner occupied-homes (in $1000's)

Les modèles seront développés uniquement avec le groupe M des 405 observations

QUESTIONS

5a) Ajustez un *Modèle de Régression Ordinaire* (MRO) de Y basé sur les 11 variables X1,..., X11
remarque : consultez l'annexe pour de l'aide sur cette question.

5b) Les données présentent-elles un problème de multi colinéarité?

5c) Développez un *Modèle de Régression avec Sélection pas à pas Avant* (*Forward Stepwise*) (MRF)

5d) Développez un *Modèle de Régression avec Sélection pas à pas Arrière* (*Backward Stepwise*) (MRB)
Complétez le tableau 5d de la page 10 qui résume les modèles.

5e) Comparez les prédictions des 3 modèles sur l'ensemble test T constitué des 101 observations. Choisir le meilleur modèle selon des critères ; préciser la nature de ces critères. **voir les pages 11 à 18 pour de l'aide**

Tableau 5d : synthèse des modèles

Var	Nom	coefficient	MRO ordinaire	MRF sélection avant	MRB sélection arrière
X0	intercepte	b0			
X1	CRIM	b1			
X2	NOX	b2			
X3	AGE	b3			
X4	DIS	b4			
X5	RM	b5			
X6	LSTAT	b6			
X7	RAD	b7			
X8	CHAS	b8			
X9	INDUS	b9			
X10	TAX	b10			
X11	PT	b11			
		SS resid résiduelle			
		MSE = $sigma^2$ (ANOVA)			
		R^2			
		R^2 ajusté			

remarque : **laisser la cellule vide si la variable n'est pas retenue dans le modèle**

AIDE pour la question 5e)

Comment calculer les prédictions des observations qui n'ont pas participé au calcul du modèle si on utilise STATISTICA ?

Il n'y a au moins deux méthodes pour réaliser cette opération.

Cette opération peut se faire directement dans le module *General Regression Models.*

MÉTHODE 1

Nous allons illustrer en utilisant un sous ensemble de 100 observations appelé BostonHousing100.sta choisies au hasard provenant du fichier d'origine *BostonHousing.sta* qui en contenait 506 dont 405 étaient employées (groupe M) pour calculer le modèle. Dans le nouveau fichier, le groupe M est formé de 80 observations et nous avons retenu 7 des 11 variables disponibles.

Voici une copie d'une partie partielle du nouveau fichier Statistica *BostonHousing100.sta.*

Harrison, D, Rubenfeld, D. (1978) Hedonic Prices and the Demand For Clean Air
J. of Environmental Economic and Management, v.5, 81-102
Combined information from 10 separate governemental and education sources
100 census tracts (CT) (choisi au hasard) parmi les 506 qui étaient disponibles dans le fichier d'origine : BostonHousing.sta
Le fichier contient un filtre pour identifier le variable (v10) Group = M ou T
voir *Tools ... Selection conditions ... Edit ...*
Seulement les 80 observations du groupe M seront retenus pour développer l'équation de prédiction de MV
basée sur les 7 variables (indicateurs). On peut employer le méthode de régression ordinaire (MRO) ou une autre méthode.
Le fichier contient 100 observations est divisé en 2 groupes
GROUP = M pour le développement du (des) Modèles (80 observations)
les données M sont surlignées (vert) et constituent un filtre; toute analyse statistique
est basée sur ces données voir *Tools ... Selections conditions ... edit*
GROUP = T pour Tester le modèle (20 observations)
pour inclure ce groupe dans une analyse, il faut désactiver le filtre
INDICATEURS 7 variables furent retenues sur les 11 qui étaitent disponibles dans le fichier d'origine.

X3	AGE	= Proportion of owner occupied units built prior to 1940
X4	DIS	= Weighted DIStances to five Boston employment centers
X5	RM	= Average number of RooMs per dwelling
X6	LSTAT	= % of the Lower STATus of the population
X9	INDUS	= Proportion of non-retail INDUStrial business acres per town
X10	TAX	= Full value property TAX rate per $10,000
X11	PT	= Pupil-Teacher ratio by town
	RLZ	= Proportion of Residential Land Zoned for lots over 25,000 sq.ft.

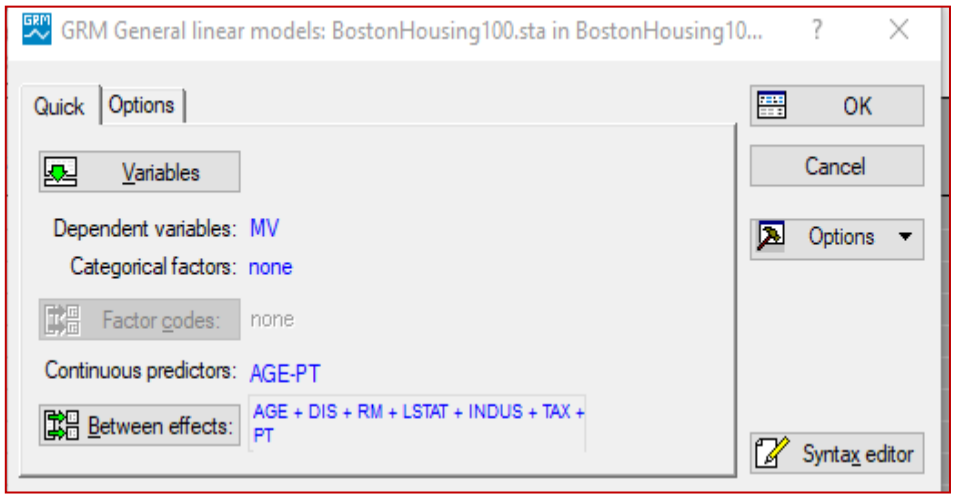
RÉPONSE
Y MV = Median Value of owner occupied-homes in $1000's

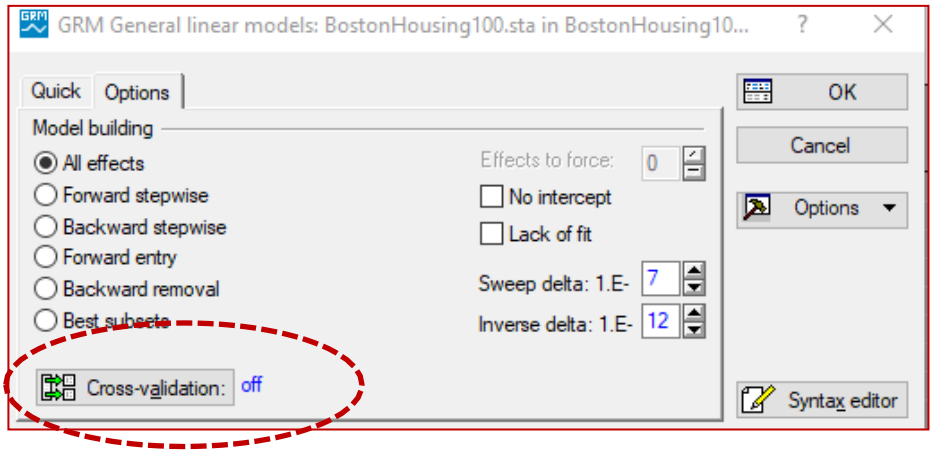
But de ce fichier But de ce fichier
Montrer comment avec Statistica développer un modèle sur une partie des données (ici le groupe M) et
appliquer le modèle pour obtenir des prédictions sur une autre partie (T).
On peut faire cette manipulation selon avec 2 méthodes. Elle seront illustrées dans le document qui accompagne ce workbook.
Cela permet de prédire la valeur de MV et de les comparer avec les vraies valeurs disponibles.
On peut répéter l'approche avec plusieurs types de méthodes de modélisation.
On peut ainsi choisir le meilleur modèle parmi plusieurs selon des critères appropriés.

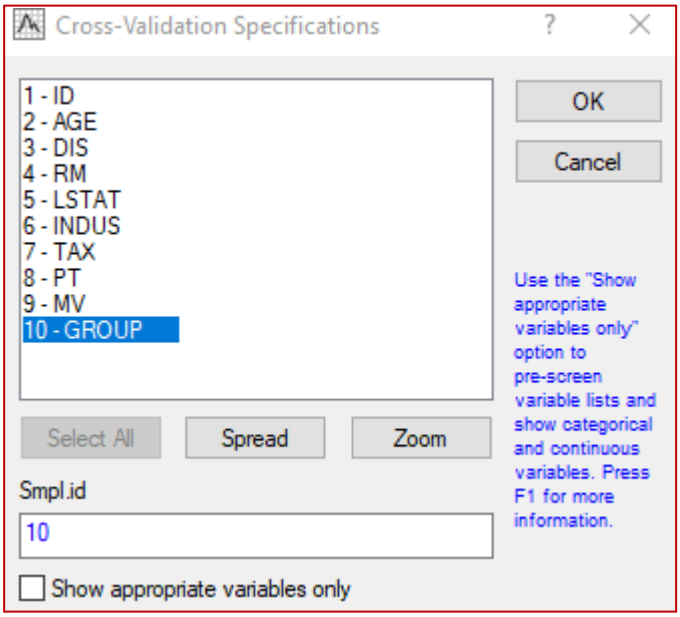
1 ID	2 AGE	3 DIS	4 RM	5 LSTAT	6 INDUS	7 TAX	8 PT	9 MV	10 GROUP
1	65,2	4,0900	6,575	4,98	2,31	296	15,3	24,0	M
2	78,9	4,9671	6,421	9,14	7,07	242	17,8	21,6	T
3	45,8	6,0622	6,998	2,94	2,18	222	18,7	33,4	M
4	84,5	4,4619	6,096	10,26	8,14	307	21,0	18,2	T
5	81,7	4,2579	5,990	14,67	8,14	307	21,0	17,5	M
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
97	97	92,7	1,8209	5,454	18,06	27,74	711	20,1	15,2	M
98	98	98,0	1,8226	5,093	29,68	27,74	711	20,1	8,1	M
99	99	72,9	2,7986	5,390	21,14	9,69	391	19,2	19,7	M
100	100	80,8	2,5050	6,030	7,88	11,93	273	21,0	11,9	T

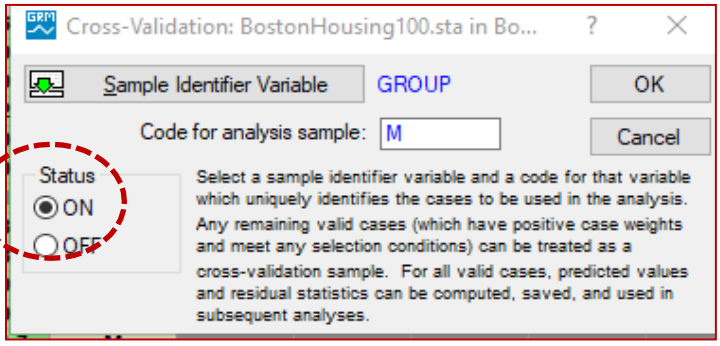
Avec le module General Regression Model

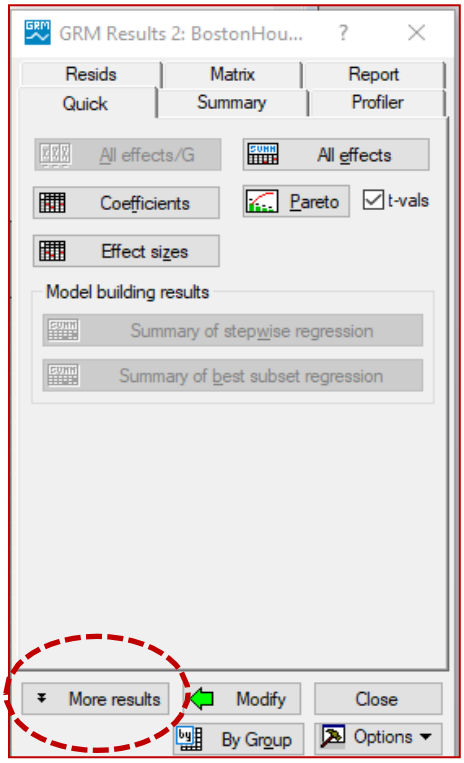
Nous allons monter avec une série de captures d'écran qui sont suffisamment explicites pour que vous soyez en mesure de répéter l'opération vous-même.

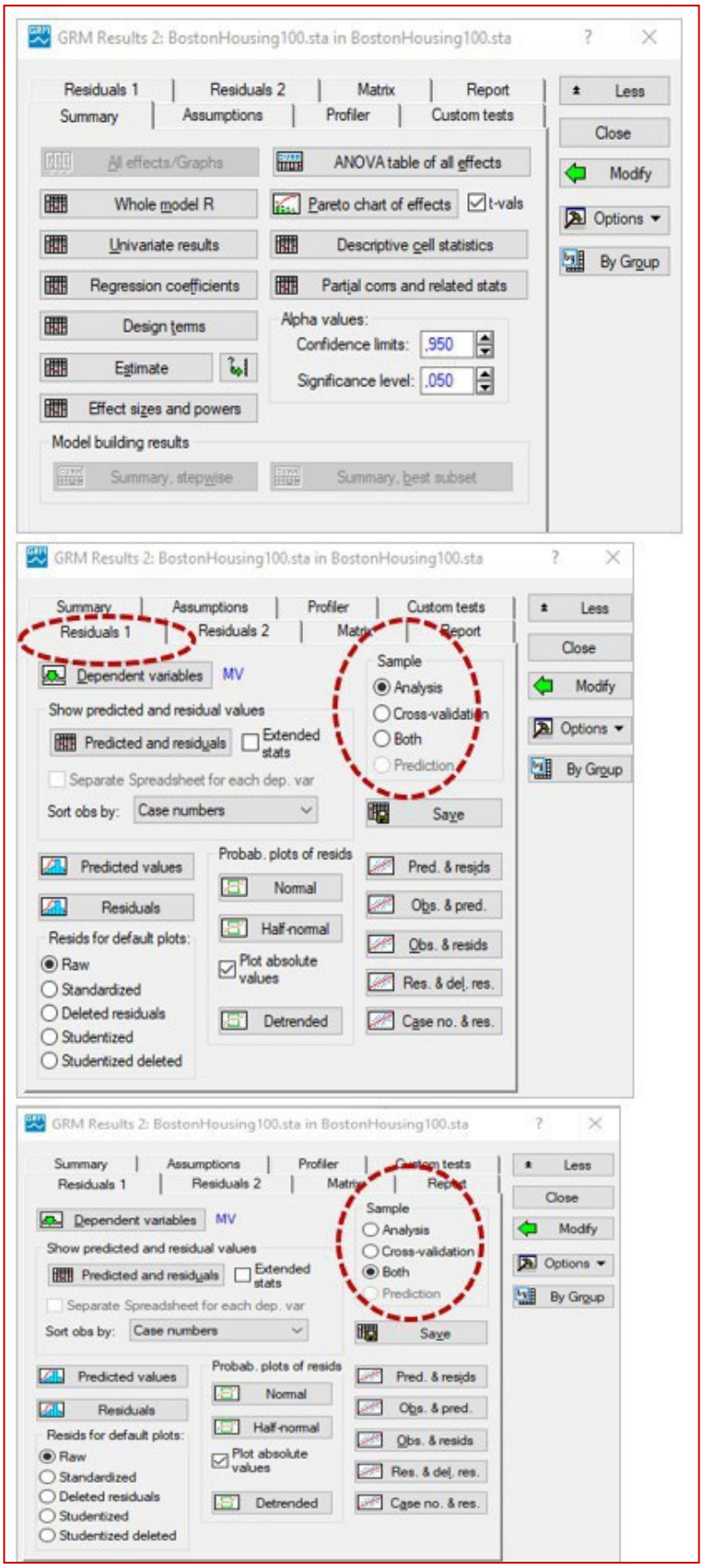


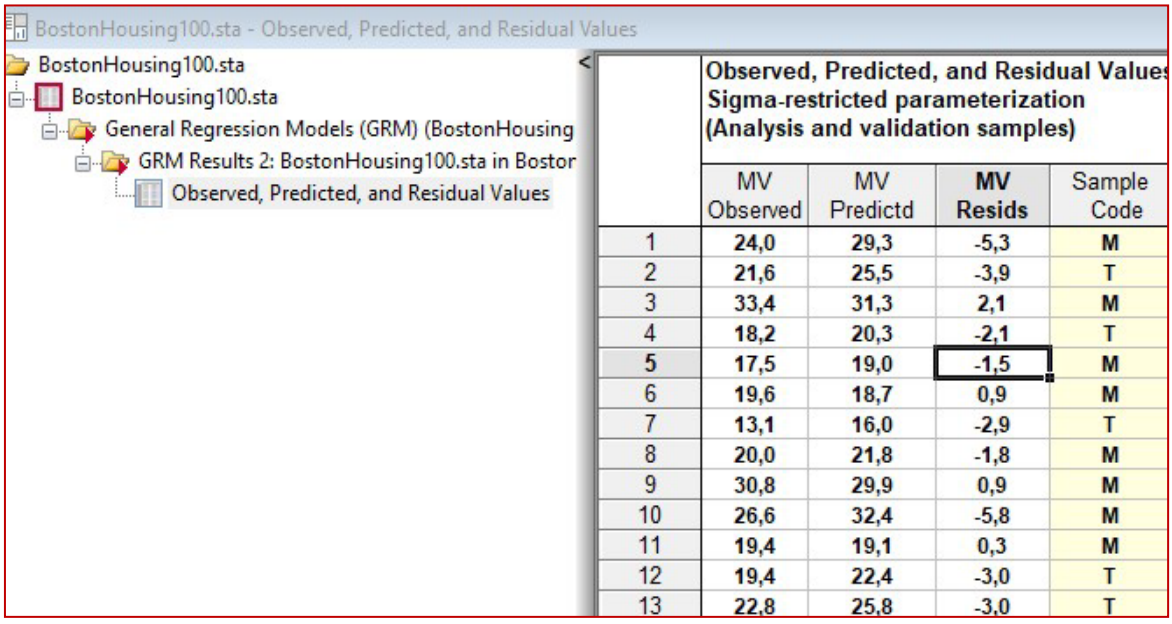



Cross-Validation Specifications
1 - ID
2 - AGE
3 - DIS
4 - RM
5 - LSTAT
6 - INDUS
7 - TAX
8 - PT
9 - MV
10 - GROUP
OK
Cancel
Select All
Spread
Zoom
Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.
Smpl.id
10
Show appropriate variables only


Cross-Validation: BostonHousing100.sta in Bo...
Sample Identifier Variable
GROUP
OK
Code for analysis sample:
M
Cancel
Status
ON
OFF
Select a sample identifier variable and a code for that variable which uniquely identifies the cases to be used in the analysis. Any remaining valid cases (which have positive case weights and meet any selection conditions) can be treated as a cross-validation sample. For all valid cases, predicted values and residual statistics can be computed, saved, and used in subsequent analyses.


GRM Results 2: BostonHou...
Resids
Matrix
Report
Quick
Summary
Profiler
All effects/G
All effects
Coefficients
Pareto
t-vals
Effect sizes
Model building results
Summary of stepwise regression
Summary of best subset regression
More results
Modify
Close
By Group
Options

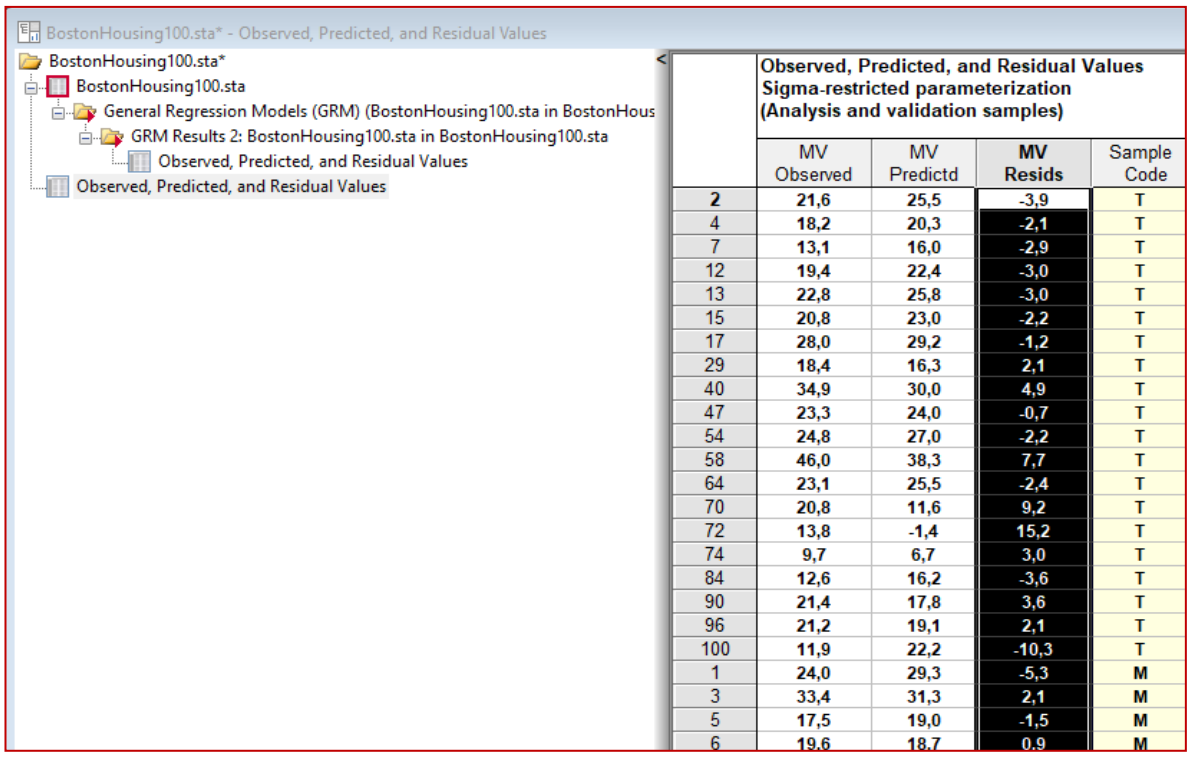

GRM Results 2: BostonHousing100.sta in BostonHousing100.sta
Residuals 1
Residuals 2
Matrix
Report
Summary
Assumptions
Profiler
Custom tests
Less
Close
Modify
Options
By Group
All effects/Graphs
ANOVA table of all effects
Whole model R
Pareto chart of effects
t-vals
Univariate results
Descriptive cell statistics
Regression coefficients
Partial corrs and related stats
Design terms
Alpha values:
Confidence limits: .950
Estimate
Significance level: .050
Effect sizes and powers
Model building results
Summary, stepwise
Summary, best subset
Dependent variables
MV
Show predicted and residual values
Predicted and residuals
Extended stats
Separate Spreadsheet for each dep. var
Sort obs by: Case numbers
Sample
Analysis
Cross-validation
Both
Prediction
Save
Predicted values
Residuals
Resids for default plots:
Raw
Standardized
Deleted residuals
Studentized
Studentized deleted
Probab. plots of resids
Normal
Half-normal
Plot absolute values
Detrended
Pred. & resids
Obs. & pred.
Obs. & resids
Res. & del. res.
Case no. & res.



Observed, Predicted, and Residual Values
Sigma-restricted parameterization
(Analysis and validation samples)

	MV Observed	MV Predictd	MV Resids	Sample Code
1	24,0	29,3	-5,3	M
2	21,6	25,5	-3,9	T
3	33,4	31,3	2,1	M
4	18,2	20,3	-2,1	T
5	17,5	19,0	-1,5	M
6	19,6	18,7	0,9	M
7	13,1	16,0	-2,9	T
8	20,0	21,8	-1,8	M
9	30,8	29,9	0,9	M
10	26,6	32,4	-5,8	M
11	19,4	19,1	0,3	M
12	19,4	22,4	-3,0	T
13	22,8	25,8	-3,0	T

Les résultats en ordre de *Sample Code* correspondant à la variable GROUP qui fut crée.



Observed, Predicted, and Residual Values
Sigma-restricted parameterization
(Analysis and validation samples)

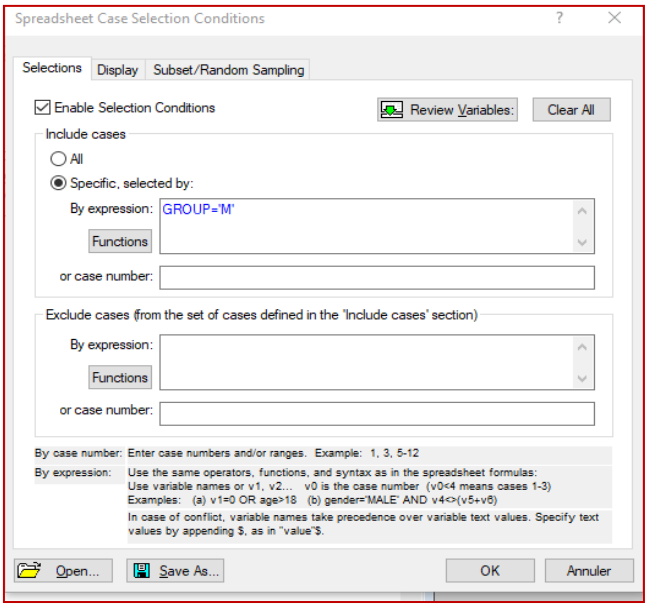
	MV Observed	MV Predictd	MV Resids	Sample Code
2	21,6	25,5	-3,9	T
4	18,2	20,3	-2,1	T
7	13,1	16,0	-2,9	T
12	19,4	22,4	-3,0	T
13	22,8	25,8	-3,0	T
15	20,8	23,0	-2,2	T
17	28,0	29,2	-1,2	T
29	18,4	16,3	2,1	T
40	34,9	30,0	4,9	T
47	23,3	24,0	-0,7	T
54	24,8	27,0	-2,2	T
58	46,0	38,3	7,7	T
64	23,1	25,5	-2,4	T
70	20,8	11,6	9,2	T
72	13,8	-1,4	15,2	T
74	9,7	6,7	3,0	T
84	12,6	16,2	-3,6	T
90	21,4	17,8	3,6	T
96	21,2	19,1	2,1	T
100	11,9	22,2	-10,3	T
1	24,0	29,3	-5,3	M
3	33,4	31,3	2,1	M
5	17,5	19,0	-1,5	M
6	19,6	18,7	0,9	M

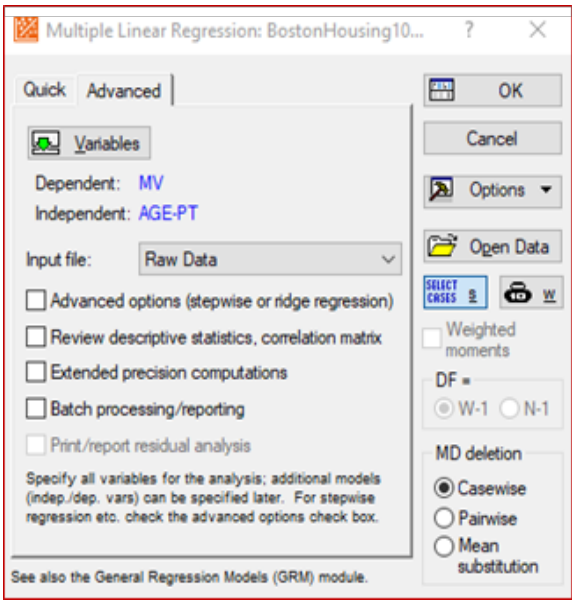
<u>Remarque</u>

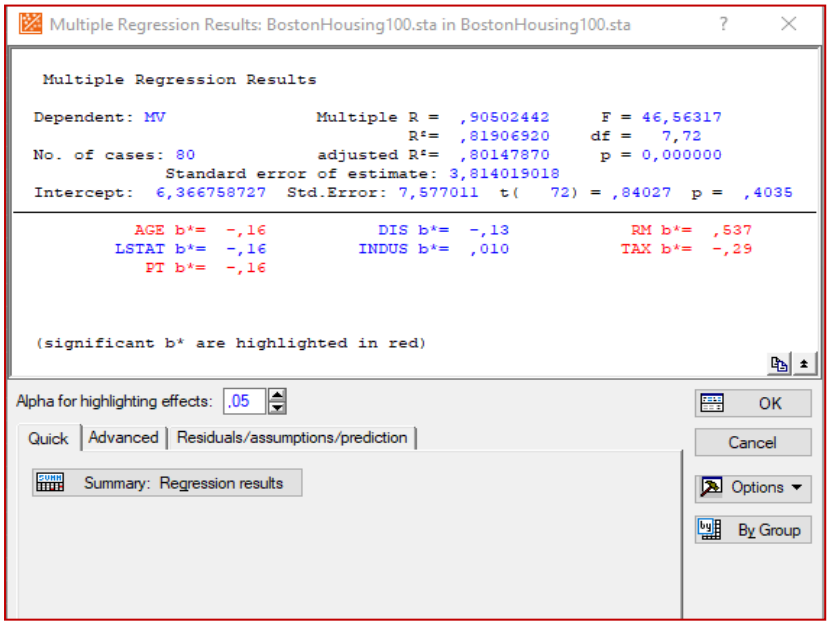
Le fichier initial *BostonHousing100.sta* contenait une variable qui séparait les données en 2 groupes. C'est la variable GROUP. Il n'est pas nécessaire d'imposer un filtre sur les données avec la commande *Tool Selection Condition Edit*

Méthode 2

On fait la création d'une variable pour séparer les données en 2 groupes.
Ici c'est la variable GROUP dans le fichier. La cariable GROUP agit comme filtre.
On introduit un filtre dans les données avec la commande *Tool Selection Condition Edit*
en imposant la condition GROUP = M pour identifier les données avec lesquelles on fera les Calculs.
Ensuite, on lance l'analyse de son choix, par exemple *Multiple Regression*.
Voici l'exemple en question.





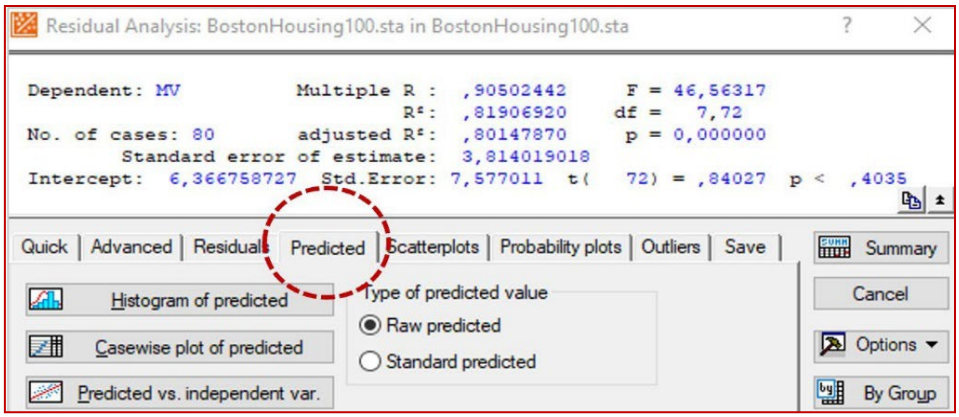


Regression Summary for Dependent Variable: MV (BostonHousing100.sta in B
R= ,90502442 R²= ,81906920 Adjusted R²= ,80147870
F(7,72)=46,563 p<0,0000 Std.Error of estimate: 3,8140
Include condition: GROUP='M'

N=80	b*	Std.Err. of b*	b	Std.Err. of b	t(72)	p-value
Intercept			6,366759	7,577011	0,84027	0,403537
AGE	-0,161775	0,078355	-0,054695	0,026491	-2,06463	0,042561
DIS	-0,134560	0,074894	-0,598831	0,333301	-1,79667	0,076583
RM	0,536922	0,075389	6,606792	0,927663	7,12197	0,000000
LSTAT	-0,162086	0,095678	-0,190542	0,112475	-1,69407	0,094573
INDUS	0,010040	0,093314	0,012153	0,112950	0,10759	0,914618
TAX	-0,288263	0,078184	-0,014442	0,003917	-3,68698	0,000437
PT	-0,158309	0,058348	-0,608556	0,224297	-2,71317	0,008332

Ces coefficients sont identiques à ceux du module General Regression Models.
Ils ne furent pas montrés précédemment.

On peut obtenir les prédictions des 80 observations avec lesquelles le modèle fut construit.
Avec l'onglet *Predicted*

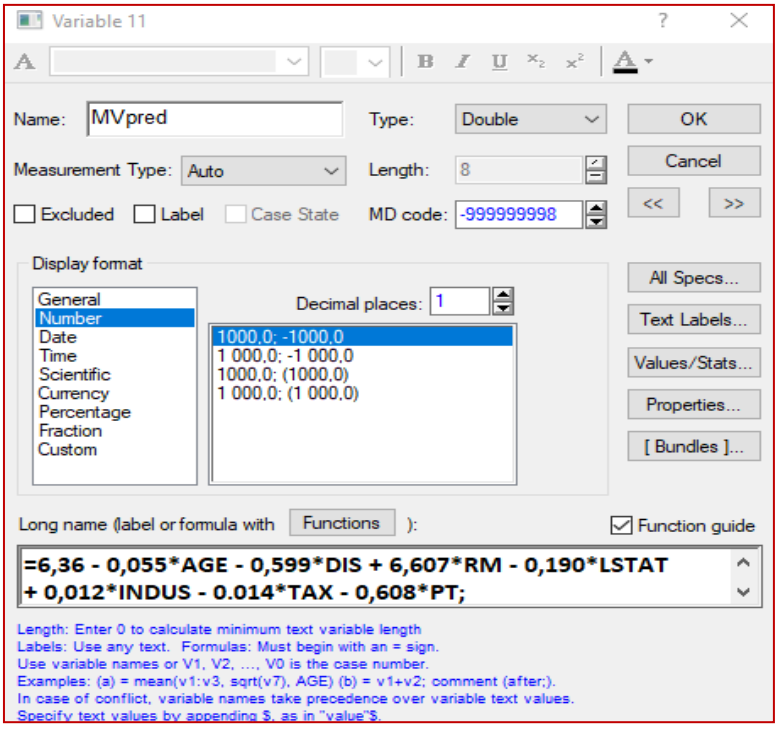


Pour obtenir les prédictions de l'ensemble T qui ne fut pas utilisé dans les calculs, il faut faire 2 opérations.

a) **Enlever la condition du filtre sur les données.**
b) **Créer un colonne additionnelle dans le fichier *BostonHouse100.sta*.**

Par exemple MVpred et introduire l'équation de sa définition avec le résultat de l'analyse

MVpred = 6,36 – 0,055*AGE – 0,599*DIS + 6,607*RM – 0,190*LSTAT
+ 0,012*INDUS -0.014*TAX – 0,608*PT



9 MV	10 GROUP	11 MVpred
24,0	M	29,4
21,6	T	25,6
33,4	M	31,4
18,2	T	20,4
17,5	M	19,1
19,6	M	18,9
13,1	T	16,1
20,0	M	21,9
30,8	M	30,1
26,6	M	32,5
19,4	M	19,2

Cette méthode est plus longue que la précédente.
C'est la seule si on a plusieurs modèles provenant de plusieurs analyses / modules à insérer dans le fichier de données dans le but de faire des comparaisons.

Mth8302-exer06- Étude d'un modèle de régression multiple

Données = BodyFat-F

Données physiologiques de 20 femmes en santé, âgées entre 25 et 35 ans

La mesure de l'indice de gras (BodyFat) est compliquée, longue et couteuse : on doit faire l'immersion de la personne dans l'eau.
Peut-on développer un modèle fiable qui permettrait de prédire plus simplement et plus rapidement Y_BodyFat avec les variables faciles à mesurer :

- **X1_epTricep : épaisseur peau triceps**
- **X2_circHanches : circonférence hanches**
- **X3_circBras : circonférence du milieu du bras**

QUESTIONS

On pense qu'il est naturel que toutes les variables soient positivement corrélées entre elles et, en particulier, avec l'indice de gras.

6a) Calculez la matrice de corrélation.
Produire un scattergramme (dispersion de 2 variables) global entre toutes les variables.
La corrélation positive entre les variables est-elle valide?

6b) Développez le Modèle de Régression Multiple Ordinaire (MRMO) entre Y et X1, X2, X3.

Examinez le signe des coefficients dans le modèle MRO.
Le modèle semble-t-il satisfaisant ?
Quel est la cause ?
Que peut-on faire pour obtenir un modèle plus satisfaisant ?
Proposez 2 autres modèles, Mod1, Mod2 incorporant toutes les variables X.
pour l'obtenir un modèle plus satisfaisant.

6c) Développez un modèle alternatif Mod1 ; précicez la méthode employée.

6d) Développez un deuxième modèle alternatif Mod2 ; précisez la méthode employée.

6e) Comparez et faites un choix entre Mod1 et Mod2 à l'aide de critères appropriés.

Mth8302-exer07 Étude de prédiction d'activité biologique : modélisation PLS

Données = Penta

INTRODUCTION

Les nouveaux médicaments sont développés avec des produits chimiques qui sont biologiquement actifs (génie du vivant). Tester des molécules pour déceler l'activité biologique est un processus couteux et il serait utile de prédire l'activité biologique avec des mesures dont le coût serait plus faible. Il est même possible, sans même faire le composé, de calculer certaines caractéristiques comme la taille, la lidrophobicité (habileté à se dissoudre), et la polarité de groupes chimiques clés sur différents site de la molécule ainsi que l'activité du composé. Ce domaine de recherche est appelé chimie computationnelle.

Le fichier de données, Penta, contient 31 observations et les variables

- **NOM : nom du composé**
- **15 mesures X : S1, L1,…, P5**
- **Réponse Y_logRAI : logarithme de l'activité bradykinine (enzyme de conversion)**
- **CLASSE ; classement des données : entrainement, test**

Le fichier est divisé en 2 parties; les 15 premières observations forment l'ensemble d'entrainement du modèle PLS (étude 1978 de Ufkes); les autres constituent l'ensemble test et proviennent de l'étude 1982. Les peptides utilisés dans la deuxième étude étaient différents de ceux de la première étude et la bradykinine employée dans les deux études provenait de sources différentes.

Références

Ufkes, J. G. R. et al (1978) Structure-Activity Relationships of Bradykinin-Potentiating Peptides *European Journal of Pharmacology, vol 50, p. 119*
Ufkes, J. G. R. et al (1982) Further Studies on Structure-Activity Relationships of Bradykinin-Potentiating Peptides, *European Journal of Pharmacology, vol 79, p. 155*

Objectif

Développer un modèle PLS basé sur la première étude et examiner sa performance à prédire les données de la deuxième étude.

QUESTIONS

7a) Développez un premier modèle PLS (noté M1) sur les seules données de test (15 premières observations) pour l'activité bradikinine. Considérez un modèle avec toutes les composantes.

7b) Développez un deuxième modèle PLS (noté M2) basé sur les 2 premières composantes seulement. Justifiez l'abandon des composantes au-delà des 2 premières.

7c) Développez un troisième modèle PLS (noté M3) basé sur les 2 premières composantes basés seulement sur les régresseurs S1 P1 S3 P3 L3 S4 L4 P4. Justifiez l'abandon des autres variables L1 S2 L2 P2 S5 L5 P5.

7d) Employez le modèle M3 pour prédire l'activité brakinine pour les données de la deuxième étude. Commentez le résultat, proposez une conclusion et, possiblement, une explication.

Mth8302-exer08-Étude de modélisation avec MARS et réseaux de neurones

Données = BostonHousing ce numéro constitue une suite de Mth8302-exer05

Description des données

Harrison, D, Rubenfeld, D. (1978)
Hedonic (House) Prices and the Demand for Clean Air
J. of Environmental Economic and Management, v.5, pp. 81-102
Combined information from 10 separate governmental and education sources
506 census tracts (CT) in city of Boston of the year 1970

But étude de la relation entre 11 indicateurs de la qualité de vie et la valeur d'une résidence

Le fichier de 506 observations est divisé en 2 groupes

GROUP = M pour le développement des Modèles (405 observations: 80% des observations)
les données M sont surlignées (vert) et constituent un filtre;
la modélisation statistique est basée sur ces données (405 obs.)
voir *Tools...selections conditions...edit*
GROUP = T pour Tester le modèle (101 observations: 20% des observations)
pour inclure ce groupe dans une analyse, éditer le filtre

INDICATEURS

X1	CRIM	: CRIMe Rate Per Capita by town
X2	NOX	: Nitric OXide concentration (parts per 10 million)
X3	AGE	: Proportion of owner-occupied units built prior to 1940.
X4	DIS	: Weighted DIStances to five Boston employment centers.
X5	RM	: Average number of RooMs per dwelling
X6	LSTAT	: % of the Lower STATus of the population
X7	RAD	: Index of accessibility to RADical highways
X8	CHAS	: CHASrles river dummy variable (1 if census tract bounds the river; 0 otherwise)
X9	INDUS	: Proportion of non-retail INDUStrial business acres per town
X10	TAX	: Full value property TAX rate per $10,000
X11	PT	: Pupil-Teacher ratio by town
	RLZ	: Proportion of Residential Land Zoned for lots over 25,000 sq.ft.

disponible dans le fichier mais elle ne sera pas employée

RÉPONSE

Y_MV : Median Value of owner-occupied homes (in $1000's)

Les modèles seront développés avec le groupe M des 406 observations.

QUESTIONS

8a) Ajuster un Modèle de Régression MARS (MARS) de Y basé sur les 11 variables X1,..., X11.

Inclure des termes d'interaction dans le modèle.

Inclure des graphiques qui identifient les nœuds employés dans le modèle.

8b) Développer 20 réseaux de neurones pour Y. Retenir les 2 meilleurs.

8c) Comparer la performance des modèles développés en 8a) et 8b) sur l'ensemble T des 101 observations. Préciser vos critères.

8d) Compléter la comparaison en 8c) en incluant le meilleur modèle retenu lors du No5e)
Présenter vos résultats dans un tableau.

8e) Identifier les forces et faiblesses des différentes méthodes de modélisation employées dans la question 5e) et les questions 8a) et 8b).
Présenter les forces et faiblesses dans un tableau.

8f) Proposer un conclusion générale sur le processus de modélisation statistique à l'aide de modèles de régression includant les réseaux de neuronnes.

Mth8302-exer09 Modèle d'analyse de la variance à un facteur

Données = Corn

remarque : des définitions utiles pour cet exercice sont présenter suite à l'énoncé (page 24)

Le fichier donne le rendement de maïs cultivé, YldA, YldB, YldC avec 3 fertilisants A, B, C.
16 Grandes Étendues de Sol (GES) furent divisées en 3 parcelles de terrain (PT).
Sur chaque PT on répandit un des 3 fertilisants A, B, C. L'attribution du fertilisant fut réalisée en attribuant au hasard un des 3 fertilisants à chaque PT dans chaque GES.
C'est la méthode 1.

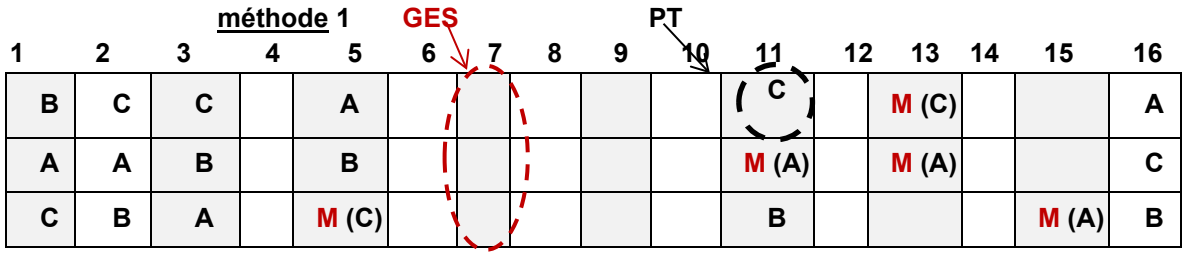


Figure 1 : exemple d'une répartition hypothétique des fertilisants A, B, C M (= manquante)

On mesura le rendement (Yield) obtenu lors de la récolte. Pour des raisons inconnues, 5 mesures ne sont pas disponibles (manquantes = M) comme on peut le constater dans la figure 1 et dans le tableau des données.

Tableau des données

GES	YldA	YldB	YldC	GES	YldA	YldB	YldC	GES	YldA	YldB	YldC	GES	YldA	YldB	YldC
1	452	546	785	5	356	459	M	9	664	589	772	13	M	654	M
2	874	547	458	6	754	665	669	10	682	534	732	14	435	665	597
3	554	774	886	7	558	467	857	11	M	456	689	15	M	546	830
4	447	465	536	8	574	365	821	12	547	651	654	16	245	537	827

QUESTIONS

9a) En vue de l'analyse statistique, ildentifier les variables, leur type, leur rôle.
Réorganiser les données dans un autre tableau en fonction de leur rôle.
remarque : l'analyse des données en 9f) sera exécutée sur ce nouveau fichier

9b) Quelle est l'unité statistique?

Méthode alternative (méthode 2) pour réaliser la comparaison des fertilisants.
On répand les fertilisants A, B, C sur 3 très grandes étendues (TGE) dans un premier temps. Ensuite, on subdivise chacune de ces très grandes étendues en 16 parcelles (GES) de terrain comme dans la méthode 1. Cette méthode 2 est illustrée dans la figure 2.

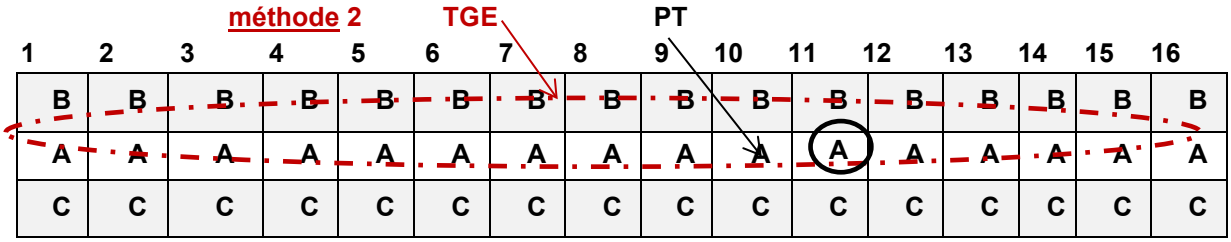


Figure 2 : exemple d'une répartition hypothétique des fertilisants A, B, C

9c) Est-ce qu'il y une différence dans la définition de l'unité statistique par rapport à la méthode 1?
9d) Comparer les 2 méthodes : avantages, inconvénients
9e) Est-ce qu'il y aurait eu une différence pour faire l'analyse statistique?
Quelle serait votre recommandation : méthode 1 ou la méthode 2?

9f) Effectuez l'analyse statistique des données. Inclure les éléments :

- **Représentation graphique de la réponse en fonction du facteur.**
- **Tableau de l'analyse de la variance.**
- **Décision concernant l'influence du facteur sur la réponse.**
- **Analyse des résidus.**
- **Égalité des variances.**
- **Comparaisons multiples a posteriori.**
- **Analyse non paramétrique**

remarque : le but visé de l'analyse non paramétrique est d'expérimenter avec sa mise en œuvre et aussi de comparer le résultat avec ceux de l'analyse paramétrique.

9g) Interpréter les résultats obtenus. Formuler une conclusion.

Dans une nouvelle étude on considère de varier un deuxième facteur variété de mais (V), dont les modalités sont V1 et V2.
Le premier facteur, type de fertilisant (F), de modalités A, B, C sera aussi présent.
On envisage de maintenir le nombre total d'observations à 48 comme dans le premier plan de collecte des données. Le plan est équilibré :
le facteur V est présent dans 24 (= 48 / 2) observations,
le facteur F est présent dans 16 (= 48 / 3) observations,
les 6 combinaisons de traitement (VF) sont présentes dans 8 (= 48 / 6) observations.
Les 2 méthodes d'assignation (design expérimental) des facteurs aux unités expérimentales sont envisagées :

Design Expérimental 1 (DE1): assignation en mode complètement aléatoire (CRD).
Design Expérimental 2 (DE2) : assignation en mode parcelles divisées (SplitPlot)
dans laquelle un premier facteur est attribué à de grandes parcelles (TGE)
et, ensuite un deuxième facteur est attribué à des parcelles divisées (PT).

9h) Veuillez compléter le tableau suivant

critères	DE1 : CRD	DE2 : SplitPlot
Avantages		
Inconvénients		
Unités expérimentales		
Recommendation		
Analyse statistique et méthode d'assignation des traitements		

page suivante : définitions

Définitions

Unité expérimentale = Unité statistique = Sujet

= élément inanimé ou vivant sur lequel on fixe les facteurs X et l'on collecte les données de la réponse Y

plus petite quantité de matériel (ou sujet) sur laquelle on peut appliquer un traitement.
Le matériel peut être inanimé (matériau) ou vivant comme des animaux ou des personnes.
On vise a avoir des unités statistiques les plus semblables (homogènes).
Cette exigence est beaucoup plus facile à réaliser lorsque que le matériel est inanimé.
Elle plus difficile avec des unités vivantes.
Les unités peuvent avoir plusieurs tailles comme dans la méthode 2 ou un facteur est appliqué sur de grandes unités et ensuite, un deuxième facteur est appliqué sur des unités statistiques plus petites.
Cela peut être souhaitable ou commode de procéder ainsi.

Traitement

combinaison des modalités d'un ou plusieurs facteurs contrôlés qui sera appliqué sur une unité statistique.

Design (protocole) expérimental

méthode (protocole) d'attribution des traitements aux unités statistiques.
Complétement aléatoire (« Completely Randomized Design ou CRD ») :
la figure (page 22) est un exemple de l'assignation au hasard des traitements aux unités statistiques ou vice versa. Dans ce cas, il n'y a aucune contrainte à la randomisation. Cette méthode est généralement souhaitable mais il y a des situations ou ce n'est ni possible ni souhaitable de procéder ainsi. Par exemple, avec deux facteurs ou plus, lorsque qu'un des facteurs est plus difficile à changer que les autres, on peut employer la méthode des unités expérimentales divisées aussi appelée méthode des parcelles divisées (« SplitPlot »). La figure 2 illustre la méthode.

Mth8302-exer10 Étude produit alimentaire

Données = saucisses

Contexte

Une compagnie de produits alimentaires ayant développé une nouvelle saucisse à base de fèves Soya décida de conduire une expérience afin de tester l'effet de deux facteurs H et T du congélateur H : humidité (%) avec les niveaux : 20 - 30 – 40
T : température (degrés C) avec les niveaux : (- 5) / (-8) / (- 11) / (-14)
sur le changement de la couleur de la saucisse. Pour chacune des 12 combinaisons des facteurs H et T, 500 saucisses furent entreposées pour une période de 90 jours. À la fin de la période, on nota le pourcentage de saucisses Y ayant subi un changement appréciable de couleur.

QUESTIONS

10a) Il n'y a de répétition des mesures de Y pour les différents traitements.
On peut analyser les données avec un modèle additif seulement.
Les facteurs sont quantitatifs mais ils seront considérés comme catégoriques pour l'analyse.
Faire l'analyse de Y avec un modèle de variance.
Employer le module ANOVA.

10b) Tracez le graphique de la réponse en fonction de l'humidité et de la température en plaçant la température sur l'axe horizontal.
Une interaction semble-t-elle présente?

10c) Effectuez le test de Tukey pour vérifier la présence d'une interaction.
Remarque : le test de Tukey permet de détecter un effet d'interaction potentiel lorsqu'il n'y a pas de répétition.
Remarque : ne pas confondre ce test avec la méthode de comparaisons multiples de Tukey (HSD).

10d) Les deux facteurs étant quantitatifs on peut ajuster un modèle de régression polynomial du deuxième degré : $Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_{11} X1^2 + \beta_{22} X2^2 + \beta_{12} X1X2$
X1 et X2 sont les variables de codage de H et T.

Remarques

- **les variables X1 et X2 représente les variables de codage pour les variables humidité (H) et température (T). Elles sont définies par**

  $X1 = (H - m1) / d1 \qquad X2 = (T - m2) / d2$

  avec $m1 = (\max H + \min H)/2 \qquad d1 = (\max H - \min H)/2$

  $m2 = (\max T + \min T)/2 \qquad d2 = (\max T - \min T)/2$

  Les variables X1 et X2 varient sur l'intervalle [-1 ,1] et elles sont adimensionnelles.

- **la variable de réponse Y est une fraction entre 0 et 1. Il est généralement recommandé de transformer ce type de réponse afin de stabiliser la variance.**
  La transformation suggérée dans le cas d'une telle réponse est définie par

  $U = \arcsin(Y^{0,5})$

remarque :
cette transformation est vraiment utile si les valeurs observées de Y occupent une grande partie de l'intervalle [0, 1]. Dans le cas contraire, la transformation ne corrige pas le problème de la variance car celle-ci est pratiquement constante sur un petit intervalle.
Dans cet exercice, on analysera Y et non pas U.

Mth6302-exer11 Étude de marketing

données = MarketShare

Source Kutner & all 5th ed. p. 1350 data set C.3 Market Share

But de l'étude impact de facteurs pouvant affecter la part de marché d'un produit détenue par une entreprise.

Période entre septembre 1999 et août 2002 (36 mois)

Variables explicatives catégoriques X

X1 = escompte (oui, non)

X2 = promotion emballage (oui, non)

groupe = variable (1, 2, 3, 4) créée par le croisement de X1 et X2

1 = (non, non) 2 = (non, oui) 3 = (oui, non) 4 = (oui, oui)

Variables explicatives continues X

X3 = indice Nielsen = indice d'exposition publicité produit

X4 = prix mensuel

Variable à expliquer Y = part marché du produit (%)

QUESTIONS

11a) Ajuster et interpréter sommairement chacun des 4 modèles suivants.

MODÈLE	VARIABLES X	EFFETS PRINCIPAUX	EFFETS INTERACTION
M1	groupe, X3	groupe, X3	----------
M2	groupe, X3	groupe, X3	groupe*X3
M3	X1, X2, X3, X4	X1, X2, X3, X4	X1*X2
M4	X1, X2, X3, X4	X1, X2, X3, X4	X1*X2, X1*X3, X1*X4, X2*X3, X2*X4, X3*X4

11b) Comparer les modèles en présentant vos résultats avec le tableau de la page suivante.

11c) Proposer une conclusion sur l'influence des facteurs X1, X2, X3, X4 sur Y.

11d) Quel autre modèle M5 pourrait-on proposé?

Aide

explorer les procédures disponibles avec l'onglet *Advanced Linear / Nonlinear Model*

GLM : General Linear Model

GRM : General Regression Model

TABLEAU - QUESTION 11b)

MODÈLE	EFFETS SIGNIFI-CATIFS	ERREUR EXPÉRIMEN-TALE (SIGMA)	ANALYSE DES RÉSIDUS	CONCLUSION SOMMAIRE	COMMENTAIRE
M1					
M2					
M3					
M4					

Mth8302-exer12 Expérience comportement rats de laboratoire – 4 facteurs

Données = drug

Source Kutner & all 5th ed. p. 1356 C12. Drug Effect Experiment
Reference T.G. Heffner, R. B. Drawbaugh, M.J. Zigmond
Amphetamine and Operant Behavior in Rats: Relationship between Drug Effects and Control Response Rate. *Journal of Comparative and Physiological Psychology* (1974) pp. 10031-10043

Objectif Analyser Influence du niveau de dosage de l'amphétamine sur le comportement de rats de laboratoire en tenant en compte d'un classement initial et d'une cédule de renforcement.

Variables

id-animal 24 rats (unité expérimentale, sujet): rat1,…, rat24
design les rats sont classés en 6 sous-groupes selon les facteurs A et B
chaque rat reçoit une dose d'amphétamine selon 4 niveaux (facteur C)
et il est mesuré (Y) à 2 périodes de temps (facteur D)
facteur A cédule de renforcement
FR-2 - reçoit eau après 2 coups sur le levier
FR-5 : reçoit eau après 5 coups sur le levier
facteur B chaque rat est classé selon sa rapidité
lente / moyenne / vite
facteur C dose d'amphétamine (mg/kg)x10 administré dans un ordre au hasard
0 (=solution saline) / 5 / 10 / 20
facteur D période de temps t1 / t2
Y-réponse nombre de coups de levier par seconde pour que le rat reçoive de l'eau

QUESTIONS

12a) Effectuer une analyse de la variance de Y en tenant en compte le plan de collecte des données.
Utiliser un modèle faisant intervenir les effets principaux et les interactions d'ordre 2.
Ne pas faire l'analyse des résidus.

12b) Interpréter sommairement la signification des résultats de l'analyse en 12a).

12c) Refaire une deuxième analyse avec un nouveau modèle tenant compte des résultats de l'analyse 12a). Effectuez une analyse des résidus.

12d) Tracer des graphiques de la réponse Y en fonction de chacun des effets principaux et des effets d'interaction avec le modèle employé en 9c).

12e) Définir une nouvelle variable appelée *groupe* selon les 6 combinaisons de A et B suivantes

groupe	A	B	groupe	A	B
1	FR-2	lente	4	FR5	lente
2	FR-2	moyenne	5	FR5	moyenne
3	FR-2	vite	6	FR5	vite

Calculer une nouvelle variable de réponse appelée Ym en faisant la moyenne de Y sur les 2 valeurs de D.
Tracer le graphique de Ym en fonction du facteur C (axe horizontal) selon les 6 groupes.
Il s'agit d'un seul graphique où sont superposées (overlay) les données de Ym pour les 6 groupes.
Ajouter la droite de Ym en fonction de C pour chaque groupe.
Aide :
employer *Graph…Categorized Graphs….Scatterplots…. Layout = Overlay…. Fit type = linear*

12f) À l'examen du graphique en 12e), décrire l'influence des facteurs A, B et C sur la réponse Ym en tenant en compte la variable groupe.

Mth8302-exer 13 Exercice théorique

On dispose de n observations (x_{i1}, x_{i2}, y_i) **i = 1, 2, …, n**
On considère l'ajustement d'un modèle de régression de la variable Y continue sur les deux variables explicatives continues X_1, X_2 :

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

Toutes les variables furent centrées-réduites :

moyennes: $\overline{x_1} = \overline{x_2} = \overline{y} = 0$ **écart types:** $S_{X1} = S_{X2} = S_y = 1$

Notation

r_{y1} **le coefficient de corrélation entre Y et X_1**
r_{y2} **le coefficient de corrélation entre Y et X_2**
r_{12} **le coefficient de corrélation entre X_1 et X_2**

Puisque les variables sont centrées réduites, on a les résultats suivants :

$$\Sigma x_{i1} = \Sigma x_{i2} = \Sigma y_i = 0 \qquad \Sigma x_{i1}^2 = \Sigma x_{i2}^2 = \Sigma y_i^2 = n - 1$$

$$r_{y1} = \Sigma y_i\, x_{i1} \qquad r_{y2} = \Sigma y_i\, x_{i2} \qquad r_{12} = \Sigma x_{i1}\, x_i$$

QUESTIONS

13a) Écrire les équations de moindres carrés pour l'estimation des paramètres.

13b) Montrer que $\hat{\beta}_0 = 0$

13c) Écrire le système des 2 équations à résoudre pour β_1, β_2 **en fonction des coefficients de corrélation** r_{y1}, r_{y2}, r_{12}

13d) Résoudre le système.

Mth8302-exer14 Exercice théorique

Soit le modèle de régression multiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

QUESTIONS

14a) Montrer la relation suivante entre la statistique F pour tester

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$$

et le coefficient de détermination R^2

$$F = [R^2 (N - m - 1)] / [m (1 - R^2)]$$

14b) Monter que le $R^2_{ajusté}$ peut s'écrire

$$R^2_{ajusté} = 1 - [SSresid / (N - m - 1) / SStot / (N-1)]$$

Notation

m : nombre de variables explicatives

N : nombre d'observations

SSresid : somme des carrés résiduels

SStot : somme totale des carrés de Y

R^2: carré du coefficient de corrélation multiple

Mth8302-exer15 Étude de trafic sur les voies rapides

Données = traffic

Description des variables

X1_pop: population de la région ou la section de la route est localisée
X2_voies: nombre de voies de la section de route
X3_largeur: largeur section en pieds
X4_accès: 1 = oui 2 = non
X5_classe: 1 = rural noninterstate 2 = rural interstate
3 = urban interstate 4 = urban noninterstate
X6_camions: disponibilité en 5 catégories selon tonnage et restrictions
X7_région: 1 = rurale 2 = urbaine (moins de 50 000 h)
3 = urbaine (plus de 50 000 h)
Y_ADDT : moyenne (sur une année) du nombre véhicules passant dans une section route chaque jour

Il y a 3 variables continues X1, X2, X3 et 4 variables catégoriques X4, X5, X6, X7.

QUESTIONS

15a) Identifier le meilleur modèle M1 à effets principaux de 5 variables selon le critère du $R^2_{\text{ajusté}}$.
remarque : ne pas faire l'analyse des résidus et l'identification des points influents.

15b) Développer un modèle M2 basé sur : X1, X2, X2*X2, X4, X1*X4.
Présenter les principaux résultats : coefficients, ANOVA, analyse des résidus.

15c) Comparer M1 et M2.

Mth8302-exer16 Étude entre le taux de mortalité et la pollution atmosphérique

Données **AirPollutionMortality-59villes**

Description des données

Les données proviennent d'une étude (en 1960) dont le but était d'examiner l'effet de 3 indicateurs de la pollution atmosphérique sur le taux de mortalité. Les 3 indicateurs de pollution sont : HC, NOX, SO2x. La variable à expliquer est l'indice de mortalité (Mortality). Cet indice fut normalisé pour enlever l'effet de l'âge des individus. Cette opération d'ajustement n'est pas documentée mais elle est probablement le résultat d'une régression avec l'age moyen dans chaque aglomération. On voulait aussi tenir compte (enlever) l'effet de 4 variables d'environnement et de 7 variables démographiques sur le taux de mortalité. Les données représentent 59 agglomérations urbaines (SMA) pour lesquelles on dispose de 14 variables explicatives dont 4 sont reliées à l'environnement, 7 à la démographie et 3 à la pollution atmosphérique. Les variables reliées à la pollution furent transformées sur l'échelle logarithmique afin de rendre leurs distributions plus symétriques en forme de cloche (pas nécessairement normales).

ENVIRONNEMENT

1. **JanTemp: Mean January temperature (degrees Fahrenheit)**
2. **JulyTemp: Mean July temperature (degrees Fahrenheit)**
3. **RelHum: Relative Humidity**
4. **Rain: Annual rainfall (inches)**

DÉMOGRAPHIQUES

5. **Education: Median education**
6. **PopDens: Population density**
7. **%NonWhite: Percentage of non-whites**
8. **%WC: Percentage of white-collar workers**
9. **Pop: Population**
10. **PopHouse: Population per household**
11. **Income: Median income**

POLLUTION

12. **logHCPot: log (HC pollution potential)**
13. **logNOxPot: log (Nitrous Oxide pollution potential)**
14. **logSO2xPot: log (Sulfur Dioxide pollution potential)**

QUESTIONS

16a) **Ajuster un modèle de régression multiple de Mortality avec les 14 variables explicatives.**

16b) **Completer les tableaux 16b1, 16b2, 16b3 en identifiant les meilleurs modèles à 4 / 5 / 6 / 7 / 8 varables variables selon chacun des critères : R^2 $R^2_{ajusté}$ C_p**

16c) **Quelles variables sont retenues à la suite de l'étude 16b)?**

16d) **Employer la méthode de *sélection pas à pas* avant (*forward stepwise*) pour retenir des modèles à considérer et compléter le tableau 16d).**

16e) **À la suite des étapes 16c) et 16d), proposer un modèle final de régression pour la mortalité. Donner l'équation.**
La pollution est-elle un facteur de mortalité?

16f) **Effectuer une analyse pour identifier, s'il y a lieu, des observations influentes avec le modèle final en 16e). Refaire un nouveau modèle en excluant ces observations.**

Mettre un X dans les cases appropriées

Tableau 16b1 critère R^2

variables	retenues 4	retenues 5	retenues 6	retenues 7	retenues 8
1. JanTemp					
2. JulyTemp:					
3. RelHum					
4. Rain					
5. Education					
6. PopDens					
7. %NonWhite					
8. %WC					
9. Pop					
10. PopHouse					
11. Income					
12. logHCPot					
13. logNOxPot					
14. logSO2xPot					

Tableau 16b2 critère $R^2_{ajusté}$

variables	retenues 4	retenues 5	retenues 6	retenues 7	retenues 8
1. JanTemp					
2. JulyTemp:					
3. RelHum					
4. Rain					
5. Education					
6. PopDens					
7. %NonWhite					
8. %WC					
9. Pop					
10. PopHouse					
11. Income					
12. logHCPot					
13. logNOxPot					
14. logSO2xPot					

Tableau 16b3 critère Cp

variables	retenues 4	retenues 5	retenues 6	retenues 7	retenues 8
1. JanTemp					
2. JulyTemp:					
3. RelHum					
4. Rain					
5. Education					
6. PopDens					
7. %NonWhite					
8. %WC					
9. Pop					
10. PopHouse					
11. Income					
12. logHCPot					
13. logNOxPot					
14. logSO2xPot					

Tableau 16d : variables retenues avec la méthode de sélection avant pas à pas

variables	retenues
1. JanTemp	
2. JulyTemp:	
3. RelHum	
4. Rain	
5. Education	
6. PopDens	
7. %NonWhite	
8. %WC	
9. Pop	
10. PopHouse	
11. Income	
12. logHCPot	
13. logNOxPot	
14. logSO2xPot	

16e) Équation du modèle FINAL

16f) Équation revisée du modèle FINAL (s'il y a lieu)

Mth8302-exer17 Modèles d'ANOVA

Données : Amphétamine

TITRE : effet du médicament de l'amphétamine sur le comportement

24 souris de laboratoire mâles, type albino, de poids approximativement égaux et de même souche furent utilisées dans une expérience concernant l'effet de l'amphétamine sur leur comportement lorsque privées d'eau. C'est l'objectif principal de cette expérience.
L'expérience fut réalisée en 2 parties : étude 1 et étude 2.

Étude 1

12 souris (s01, ..., s12) furent entrainées à activer un levier pour obtenir de l'eau jusqu'à ce qu'elles obtiennent un taux relativement stable d'activation. Sur la base de ce résultat, les souris furent classées en 3 catégories (lente, moyenne, vite) de vitesse initiale. Ce facteur est représenté par la variable XB_vitesse dans le fichier.

Les souris reçurent de l'amphétamique selon 4 niveaux de dosage (mg/kg). Ce facteur est représenté par la variable XC_dose dans le fichier et les modalités fixées dans cette expérience sont fixées: 0 (solution saline) / 0,5 / 1,0 / 1,8 (mg/kg). Les 4 niveaux furent administrés au hasard pour chaque souris. Une heure après réception, une séance expérimentale commence. La souris reçoit de l'eau après 2 coups (appui) sur le levier. C'est le facteur XA_levier dont la modalité est 2 dans l'étude 1. Chaque dose fut administrée 2 fois représenté par la variable rep_dose (1 et 2). Cette variable est un facteur de répétition. La réponse mesurée Y est définie par

Y = nombre de coups sur le levier / temps écoulé durant la session
temps est mesuré en seconde.

Étude 2

12 souris additionnelles (s13, ..., s24) furent utilisées pour la deuxième expérience.
L'étude2 est semblable à l'étude1 sauf que les souris reçoivent de l'eau après 5 coups.
Dans ce cas la variable XA_levier = 5

ANALYSES

étude1 seulement : Y en fonction de XB_vitesse, XC_dose
étude1 et étude2 combinées : Y en fonction de XA_levier, XB_vitesse, XC_dose
remarque : les modèles proposés pour l'analyse seront composés d'effets principaux et des effets d'interactions doubles

QUESTIONS

17a) Pour chacune des 2 études, décrire la nature et le rôle des facteurs dans le modèle d'analyse de variance/covariance qui sera employé pour faire l'analyse.

17b) Pourquoi les souris furent-elles initialement classées en catégories de vitesse?

17c) Pourquoi les souris recoivent-elles la dose d'amphétamine dans un ordre dicté par le hasard?

17d) Proposer le modèle qui sera employé pour faire l'analyse de l'étude 1.
Exécuter cette l'analyse et présenter les principaux résultats sous forme de tableaux et de graphiques. Proposer une conclusion principale de cette étude.

17e) Répondez aux mêmes questions que 17d) pour les 2 études combinées.

Mth8302-exer18

Mth8302-exer19

Mth8302-exer20

Mth8302-exer21

Mth8302-exer22

Mth8302-exer30 Étude exploratoire sur la classification des vins du Portugal

Données WineQuality

Source P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis
Modeling Wine Preferences by Data Mining from Physicochemical Properties.
Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Description des données

Les données proviennent d'une étude pour établir les relations entre 11 caractéristiques physico-chimiques (pH, …) du vin et une évaluation sensorielle de sa qualité par un groupe d'experts. Les vins proviennent du Portugal de la famille *Vinho Verde*. L'évaluation sensorielle est un indice sur une échelle de 0 (très mauvais) à 10 (excellent). On peut considérer cette variable (QUALITY dans le fichier) comme étant quasi continue, même si elle ne l'est pas strictement et, on peut s'en servir pour faire de la modélisation par régression. Cette variable est aussi recodée en trois catégories, nommée QUALITY2. Le fichier contient **6 497 observations sur 11 variables physico-chimiques.** Il y a **1 599 échantillons de vin rouge** et **4 898 échantillons de vin blanc**. L'information sur les variables ainsi que le recodage est décrite dans l'entête du fichier *Statistica.* Celui-ci contient aussi une variable SET qui divise le fichier en trois sous-ensembles : ***train*** (3 916 obs.) ***validate*** (1 317 obs.) ***test*** (1 264 obs,).

L'ensemble ***train*** est employé pour développer les modèles,
l'ensemble ***validate*** permet de valider les modèles
l'ensemble ***test*** permet de tester et comparer les modèles qui seront développés selon plusieurs méthodes.

QUESTIONS

30a) Comparer, au moyen de tableaux croisés et de tests statistiques appropriés, les 3 sous-ensembles (SET) pour décider s'ils sont relativement semblables (homogènes) pour la variable de réponse Y (QUALITY).

30b) Refaire l'analyse sur les 11 autres variables explicatives X. Identifier les variables explicatives X pour lesquelles on observe des différences significatives entre les 3 sous-ensembles.

30c) Avec la méthode de régression *stepwise* (pas à pas), développer un modèle de régression de Y basé sur les 11 variables physico-chimiques pour l'ensemble de tous les vins rouges et vins blancs. Identifier les variables critiques (importantes) de la qualité du vin.

30d) La conclusion est-elle la même si on fait les analyses 19b) et 19c) en séparant les vins par couleur ?

remarque : développer les modèles sur l'ensemble ***train*** seulement ;
les différents modèles de prédiction seront comparés sur l'ensemble ***test.***

30e) Employer le module ***Classification Trees*** du module Mult/Exploratory du menu *Statistics* pour développer un arbre de classification de la variable QUALITY2.
Appliquer et comparer les résultats des méthodes: *discriminant* et *C&RT_style*

30f) Refaire l'analyse c) avec le module General and Classification Trees (G&RT) du menu Data Mining.

30g) Comparer le résultat de 19e) avec ceux de 19d).

31h) Développer 5 réseaux de neurones avec le module SANN du menu Data Mining.
Utiliser les réseaux de type *MLP* seulement.

30i) Comparer les réseaux avec les courbes ROC (Receiver Operation Curve).
Remarque : pour interpréter les résultats consulter *CourbesROC& Courbes LIFT.pdf*

30j) Identifier, en ordre d'importance, les caractéristiques chimiques du vin qui lui donne sa qualité.

Mth8302-exer31 Travail de recherche sur une expérience scientifique

Effectuez un recherche dans une revue scientifique (ou sur l'internet) afin de trouver un article qui utilise une ou l'autre des méthodes d'analyse statistiques suivantes :
régression, modèle d'analyse de la variance, réseau de neurones, arbres de classification.

L'article doit contenir :

- **le fichier des données**
- **une analyse statistique**
- **l'interprétation des résultats**

QUESTIONS

On demande de

31a) présenter sommairement la problématique étudiée;

31b) faire la liste des variables, leur rôle, l'espace de variation;

31c) discuter sommairement les résultats présentés;

si possible : critiquer le plan de collecte de données;

31d) refaire l'analyse des données en utilisant le logiciel *STATISTICA;*

31e) comparer le résultat de votre analyse statistique avec l'analyse statistique des auteurs;

31f) écrire un rapport sommaire des étapes 18a) à 18e).

- **Identifier clairement la référence de l'article ou le site Internet.**
- **Inclure un fichier workbook nommé «*nomfamille*.stw » de *STATISTICA***

contenant les données ainsi que les procédures d'analyse employées

à l'étape 31e).

Mth8302-exer32 Planification d'une étude statistique

Vous avez suivi des formations en analyse statistique des données d'expériences et des données observationnelles. On vous consulte pour obtenir une proposition pour la planification et l'analyse statistique d'un projet de recherche.

Documents de référence

- **Planification d'une étude statistique (pdf)**
- **Étapes pour entreprendre un projet d'expérimental (pdf)**
- **Information pour entreprendre un projet expérimental (pdf)**
- **Études statistiques: expérimentales VS observationnelles (pdf)**
- **Rôle de la statistique dans l'industrie et les affaires (pdf)**
- **Processus modélisation statistique (pdf)**
- **Modèles statistiques (pdf)**
- **Taguchi et la conception robuste en ingénierie (pdf)**
- **Introduction au contrôle statistique des processus - SPC (pdf)**
- **Étude Répétabilité & Reproductibilité (R&R) : Évaluation processus de mesurage (pdf)**

Problématique