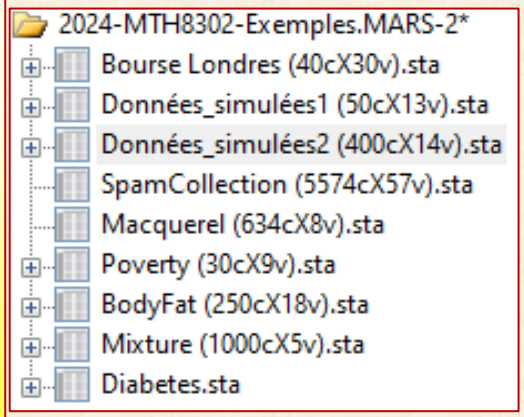
MTH8302 - Analyse de régression et analyse de variance

ch06-ch10 … Data Mining … Machine Learning

Chapitre 6 : Multivariate Adaptative Regression Splines (MARS)
Chapitre 7 : Introduction au Data Mining
Chapitre 8 : Classification and Regression Trees (CRT)
Chapitre 9 : Random Forest (RF)
Chapitre 10 : Artificial Neural Networks (ANN)

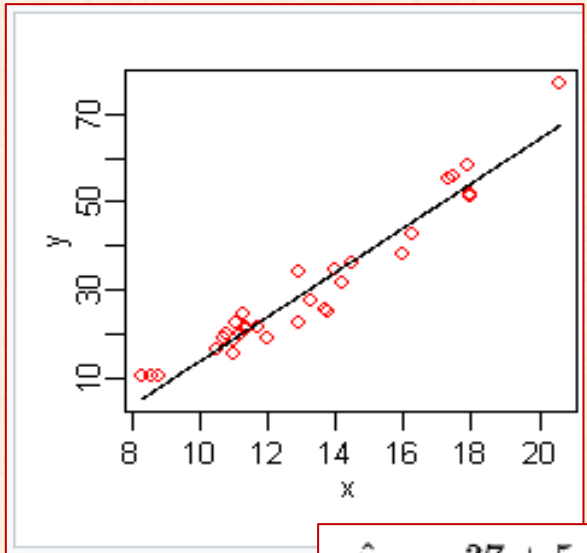
Ch06-Régression MARS — **page**

		n	p
Exemple 1	Bourse Londres..... 31-37	40	10
Exemple 2	Simulées-1 ((X,Y)..... 38-39	50	2
Exemple 3	Simulées-2 (X1 X2… Y).... 40-44	400	25
Exemple 4	Pourriels (spam)..... 45-46	4601	3
Exemple 5	Densité œufs poisson 47-48	634	8
Exemple 6	Indice pauvreté..... 49-54	30	9
Exemple 7	Body Fat..... 55-58	540	18
Exemple 8	Mixture..... 59-62	1000	5
Exemple 9	Diabètes 63-66	442	16



La Régression multivariée par spline adaptative (***MARS*** **pour « *Multivariate Adaptive Regression Splines* ») est une forme de régression non paramétrique pouvant être vue comme une extension des régressions linéaires qui modélisent automatiquement des interactions et des non-linéarités.**
présentée par Jerome H. Friedman et Bernard Silverman en 1991

Multivariate Adaptative Regression Splines (MARS) : idée de base



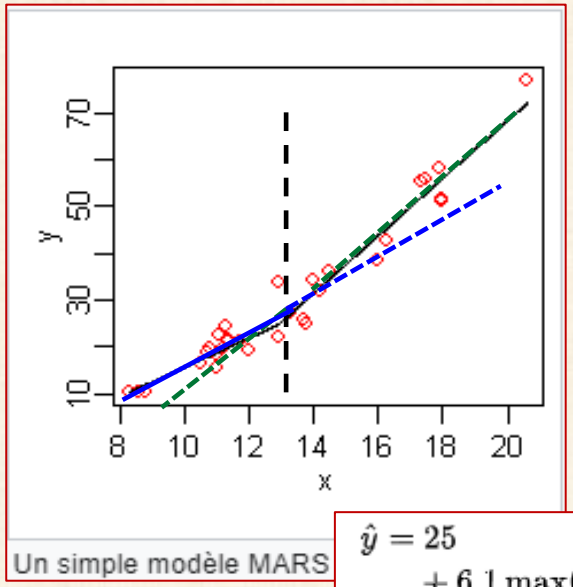
Un modèle linéaire.

$$\hat{y} = -37 + 5.1x$$

comportement Y différent selon les valeurs de X :

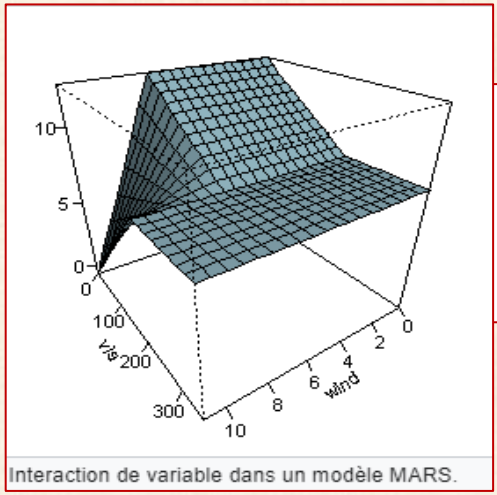
- centre
- grande
- petite

présence nœuds dans les données



Un simple modèle MARS

$$\hat{y} = 25 + 6.1 \max(0, x - 13) - 3.1 \max(0, 13 - x)$$



Interaction de variable dans un modèle MARS.

autre exemple

$$\begin{aligned}\text{ozone} = 5.2 &+ 0.93 \max(0, \text{temp} - 58) \\ &- 0.64 \max(0, \text{temp} - 68) \\ &- 0.046 \max(0, 234 - \text{ibt}) \\ &- 0.016 \max(0, \text{wind} - 7) \max(0, 200 - \text{vis})\end{aligned}$$

https://fr.wikipedia.org/wiki/R%C3%A9gression_multivari%C3%A9e_par_spline_adaptative

Multivariate Adaptative Regression Splines (MARS)

Modèle MARS

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x).$$

Le modèle est une somme pondérée

de fonctions de base

Chaque fonction de base peut prendre l'une des 3 formes suivantes :

- Une constante 1.
  Il n'y a qu'un seul terme de ce type : l'intersection avec l'axe (la valeur moyenne de la variable à expliquer quand la variable explicative prend la valeur zéro). Dans l'exemple ici cette valeur est 5.2

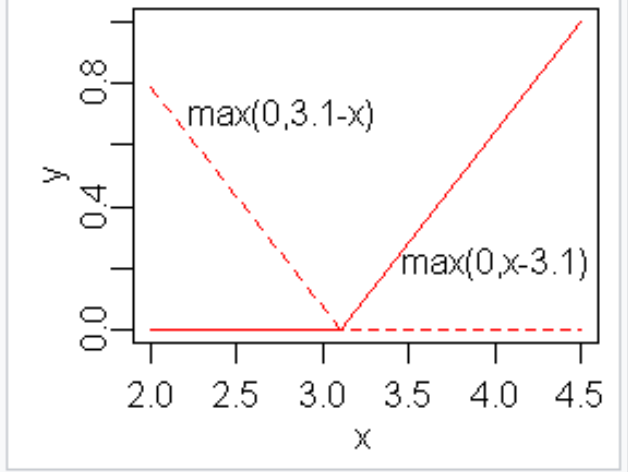
- Une fonction charnière (c'est quoi ?) définition page suivante

- Un produit de fonctions charnières

https://fr.wikipedia.org/wiki/R%C3%A9gression_multivari%C3%A9e_par_spline_adaptative

Multivariate Adaptative Regression Splines (MARS)

fonctions charnières utilisées dans le modèle MARS



Une paire de charnières en miroir fonctionne avec un nœud à x=3,1

Un élément clé des modèles MARS sont *les* fonctions charnière prenant la forme

$$\max(0, x - c) \quad \text{ou} \quad \max(0, c - x)$$

Une fonction charnière est nulle sur une partie de sa plage et peut donc être utilisée pour partitionner les données en régions disjointes, dont chacune peut être traitée indépendamment. Par exemple, une paire de charnières en miroir fonctionne dans l'expression

$$6.1 \max(0, x - 13) - 3.1 \max(0, 13 - x)$$

les fonctions charnières peuvent être multipliées entre elles pour former des fonctions non linéaires.

Les fonctions de charnière sont également appelées fonctions de rampe , de bâton de hockey ou de redresseur .

À la place du max notation utilisée dans cet article, les fonctions de charnière sont souvent représentées par où

$$[\pm(x_i - c)]_+ \text{ où } [\cdot]_+$$

+ signifie prendre la partie positive.

https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_spline

MÉTHODES & OUTILS : Data Mining

CRT Classification Regression Tree

arbres de décision

FA Forêsts Aléatoires

GAM Gen. Additive Models

MARS Multivariate Adaptive Regression Splines

ANN Artificial Neural Networrk

réseau de neurones

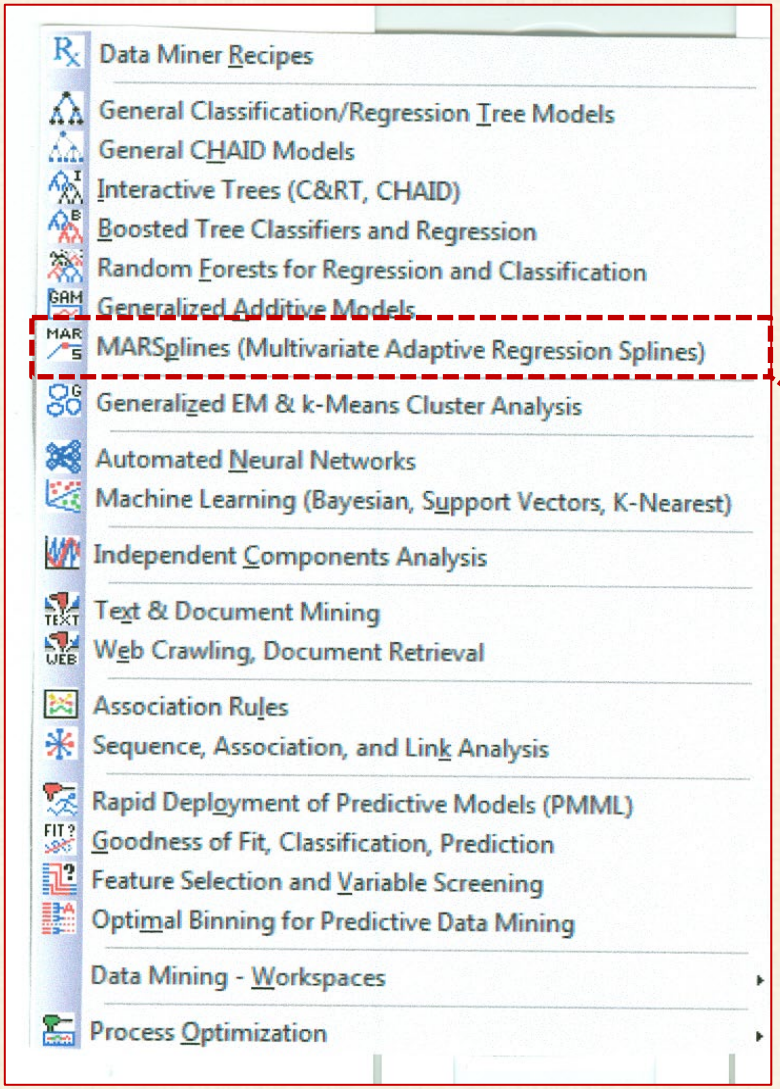
autres procédures

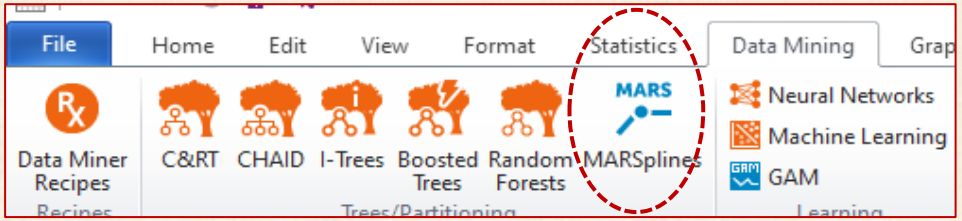
=

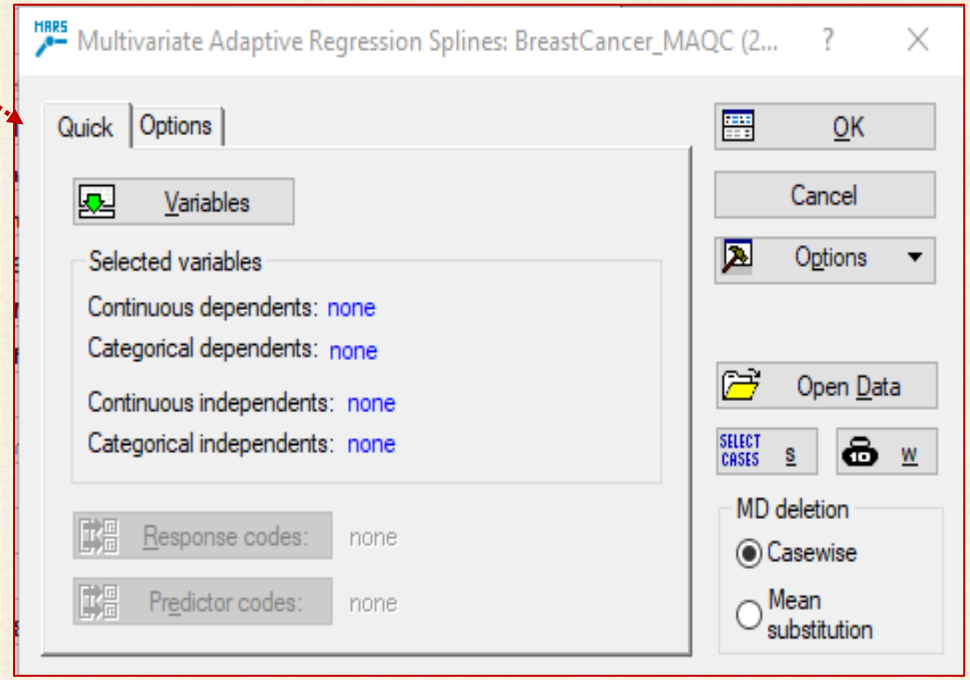
analyse non supervisée

RÉGRESSION MARS : mise en œuvre avec Statistica

MARS pas disponible en JMP Pro
disponible en Python et R







RÉGRESSION MARS

Financial data of 40 UK companies 1983

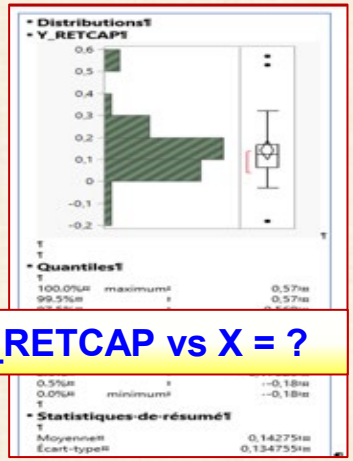
J. Jobson *Applied Multivariate Data Analysis*, vol 1, Regression and Experimental Design 1991, Springer-Verlag

Y_RETCAP	Return on capital employed
X1_WCFTCL	Ratio of working capital flow to total current liabilities
X2_WCFTDT	Ratio of working capital flow to total debt
X3_GEARRAT	Gearing ratio (debt-equity ratio)
X4_LOGSALE	Log to base 10 of total sales
X5_LOGASST	Log to base 10 of total assets
X6_NFATAST	Ratio of net fixed assets to total assets
X7_CAPINT	Capital intensity (ratio of total sales to total assets)
X8_FATTOT	Gross sixed assets to total assets
X9_INVTAST	Ratio of total inventories to total assets
X10_PAYOUT	Payout ratio
X11_QUIKRAT	Quick ratio
X12_CURRAT	Current ratio

Ex1 : Bourse Londres

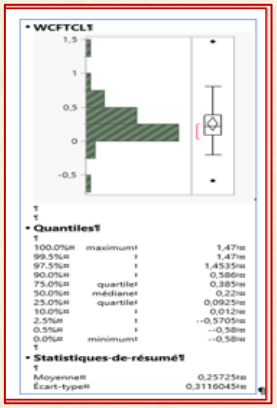
1 ID	2 Y_RETCAP	3 WCFTCL	4 WCFTDT	5 GEARRAT	6 LOGSALE	7 LOGASST	8 NFATAST	9 CAPINT	10 FATTOT	11 INVTAST	12 PAYOUT	13 QUIKRAT	14 CURRAT
1	0,26	0,25	0,25	0,46	4,11	4,30	0,10	0,64	0,12	0,74	0,07	0,18	1,53
2	0,57	0,33	0,33	0,00	4,25	4,00	0,12	1,79	0,15	0,27	0,30	1,26	1,73
3	0,09	0,50	0,20	0,24	4,44	4,88	0,94	0,36	0,97	0,01	0,57	0,39	0,44
4	0,32	0,23	0,21	0,45	4,71	4,44	0,29	1,86	0,52	0,29	0,00	0,69	1,23
5	0,17	0,21	0,12	0,91	4,85	4,75	0,26	1,26	0,54	0,33	0,31	0,90	1,76

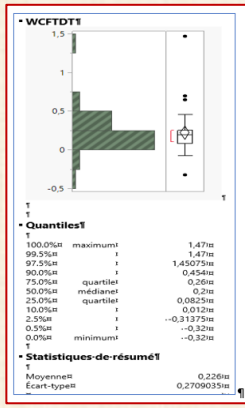
Y_RETCAP

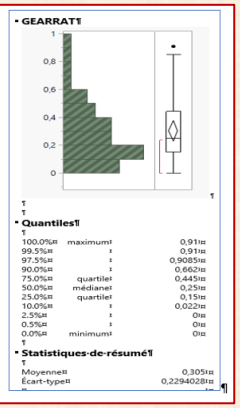


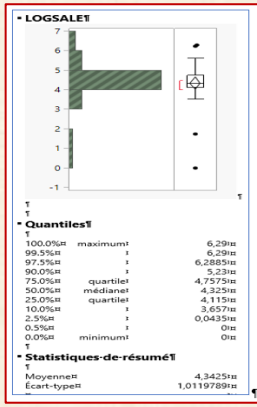
modèle Y_RETCAP vs X = ?

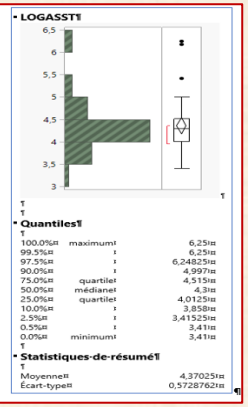
X

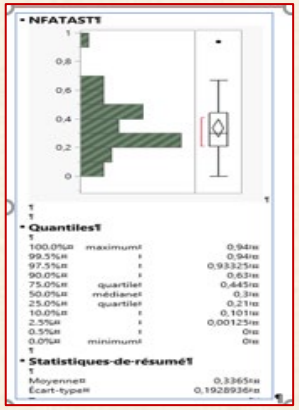


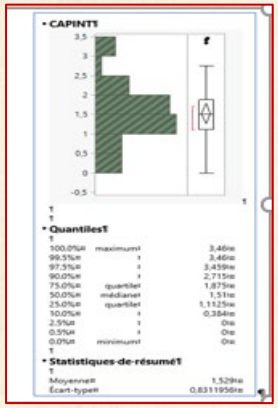


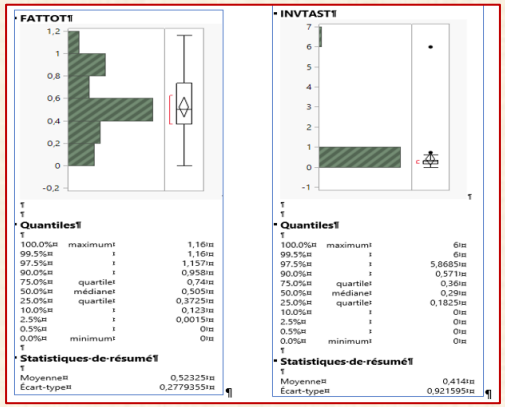


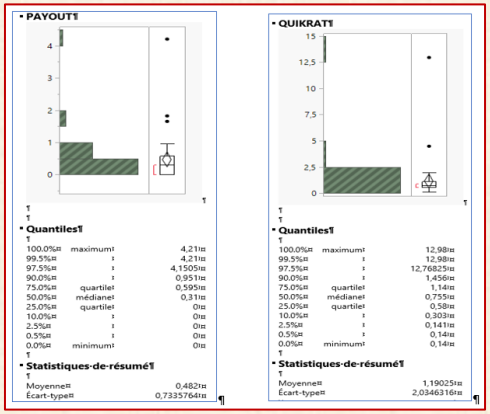


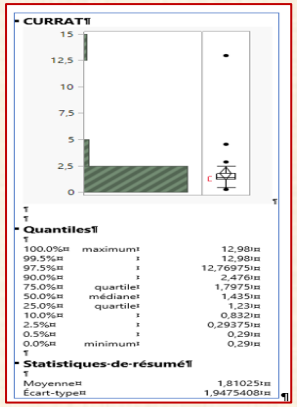












RÉGRESSION MARS

Ex2 : données simulées 1

Simulation **Regression MARS avec une seule variable X**

X : varie entre 8 et 20 50 valeurs : 8,00 8,25 ... 20,25

if1 = 1 si x plus petit que 13 et if1 = 0 si x plus grand que 13

if2 = 0 si x plus petit que 13 et if2 = 1 si x plus grand que 13

u1 = x - 13 max1 = max(u1,0) f1 = 25 + 5*u1;

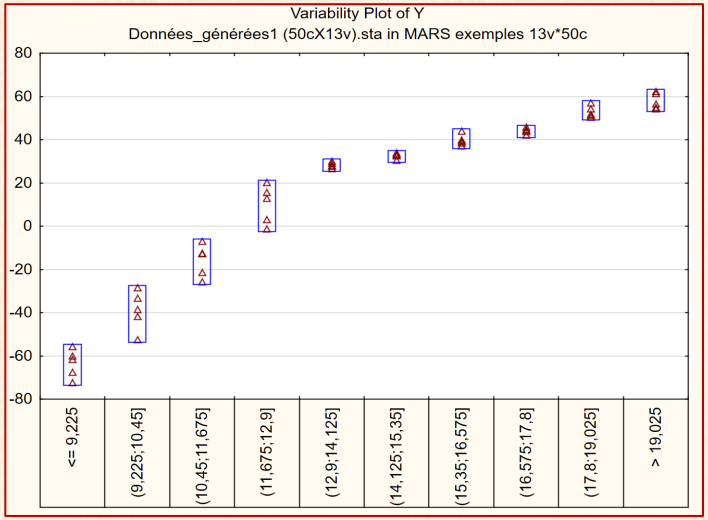
u2 = 13 -x max2 = max(u2,0) f2 = 25 - 20*u2;

fct = 25 + 5*max1 - 20*max2 = 25 + 5*max(0;x-13) - 20*max(0;13-x)

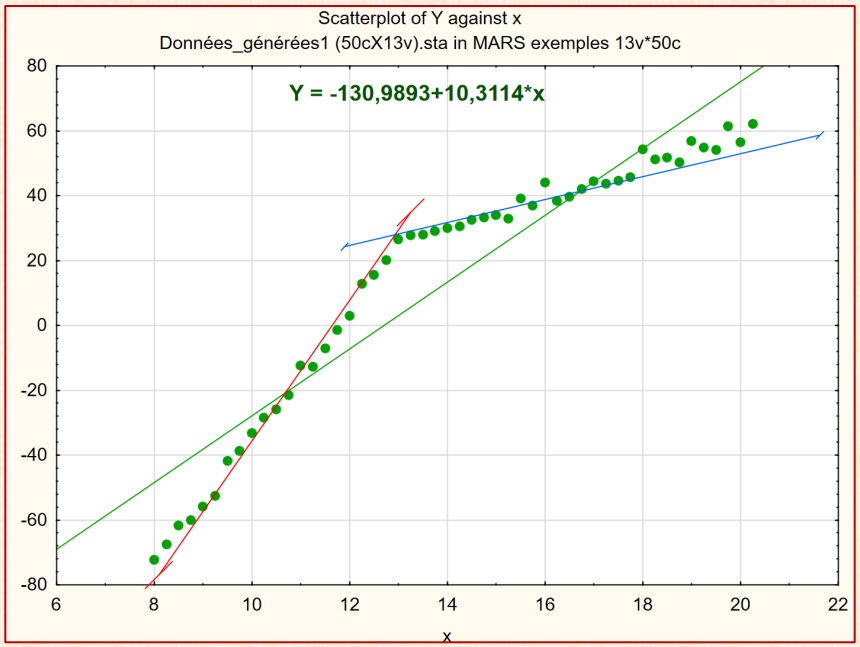
Y = fct + erreur erreur = (N(0, 2)

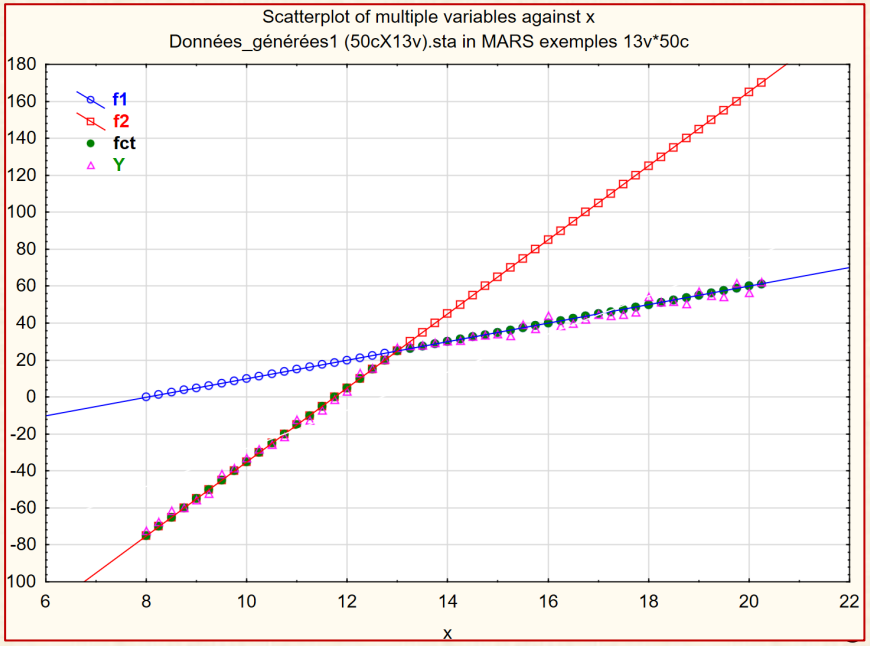
/

1 ID	2 x	3 if1	4 if2	5 u1	6 max1	7 u2	8 max2	9 f1	10 f2	11 fct	12 erreur	13 Y
1	8,00	1	0	-5,0	0,0	5,0	5,0	0,00	-75,00	-75,00	2,58	-72,42
2	8,25	1	0	-4,8	0,0	4,8	4,8	1,25	-70,00	-70,00	2,36	-67,64
3	8,50	1	0	-4,5	0,0	4,5	4,5	2,50	-65,00	-65,00	3,19	-61,81
4	8,75	1	0	-4,3	0,0	4,3	4,3	3,75	-60,00	-60,00	-0,15	-60,15
5	9,00	1	0	-4,0	0,0	4,0	4,0	5,00	-55,00	-55,00	-0,86	-55,86



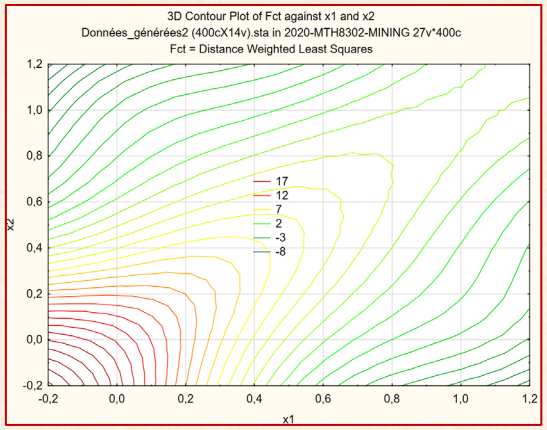
fct = 25 + 5*max1 - 20*max2
= 25 + 5*max(0;x-13) - 20*max(0;13-x)
Y = fct + erreur erreur = N(0, 2)

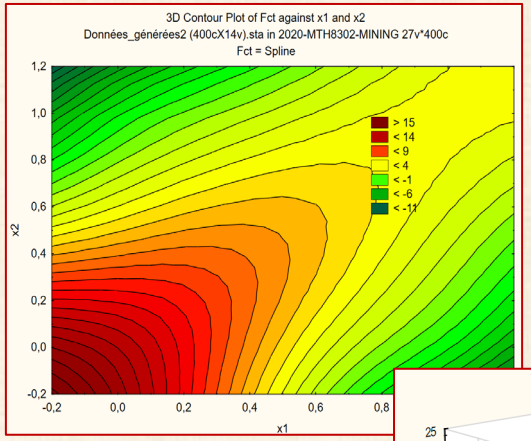




RÉGRESSION MARS

Ex3 : données simulées 2





Exemple provenant de SAS

Example 24.1 Surface Fitting with Many Noisy Variables

Consider a simulated data set that contains a response variable and 10 continuous predictors.
Each continuous predictor is sampled independently from the uniform distribution U(0,1)
Y = Fct + E = F + N(0,1)
Fct = A / B A = 40*exp(8*((X1-0,5)**2 + (X2-0,5)**2))
B = (exp(8*((X1-0,2)**2+(X2-0,7)**2)) + exp(8*((X1-0,7)**2 + (X2-0,2)**2)))
n = 400 generated observations p = 10 variables X1, X2,..., X10
x3 x4 ... x10 n'interviennenent pas dans Fct - idem pour les variables u1 u2 ...u10
seulement X1 et X2 sont actives et définissent la fonction fct ci-haut
toutes proviennent d'uéchantillon de la distribution uniforme U(0,1) sur l'intervalle (0,1)
la valeur finale de Y est perturbée en ajoutant un bruit E distribué N(0,1)
/

1 ID	2 x1	3 x2	4 x3	5 x4	6 x5	7 x6	8 x7	9 x8	10 x9	11 x10	12 A	13 B	14 Fct	15 E	16 Y	17 info	18 u1	19 u2	20 u3
1	0,97	0,41	0,57	0,48	0,02	0,64	0,54	0,03	0,30	0,07	247,42	223,06	1,11	0,80	1,91	ajout 10 variables	1,15	0,44	0,29
2	0,89	0,67	0,98	0,29	0,26	0,83	0,38	0,28	0,12	0,07	170,42	53,06	3,21	0,29	3,50	de bruit	-0,16	-0,46	0,46
3	0,34	0,59	0,87	0,88	0,46	0,91	0,99	0,02	0,00	0,08	52,05	10,59	4,92	1,01	5,92	u1 u2 ... u20	-1,52	-0,00	0,28
4	1,00	0,19	0,14	0,57	0,31	0,60	0,16	0,69	0,55	0,32	646,76	1374,47	0,47	-0,37	0,10	chaque u	-0,32	-0,19	-0,39
5	0,77	0,42	0,92	0,05	0,19	0,82	0,64	0,79	0,98	0,65	75,61	27,02	2,80	1,12	3,92	distribuée	0,19	-0,39	-0,48
6	0,82	0,43	0,50	0,20	0,84	0,92	0,65	0,77	0,70	0,00	92,80	39,06	2,38	-1,03	1,34	Norm (0,1)	1,46	-0,24	0,22
7	0,85	0,84	0,51	0,64	0,59	0,56	0,38	0,65	0,94	0,06	266,97	65,60	4,07	0,78	4,85	ou	1,25	-0,43	-0,05
8	0,44	0,47	0,77	0,43	0,62	0,92	0,04	0,58	0,67	0,32	41,27	5,50	7,51	-0,58	6,93	Unif (0,1) - 0,5	-0,91	0,14	-0,36
9	0,67	0,36	0,72	0,45	0,95	0,44	0,11	0,08	0,71	0,62	59,31	16,23	3,65	1,34	4,99		-0,67	-0,18	-0,37
10	0,44	0,68	0,29	0,71	0,43	0,21	0,45	0,30	0,79	0,62	53,48	12,54	4,27	0,05	4,32	moyenne = 0 pour N et U	-0,29	-0,33	0,38
11	0,51	0,19	0,86	0,88	0,59	0,49	0,67	0,37	0,83	0,87	86,88	18,79	4,62	1,39	6,01	ecart-type = 1 pour N	-2,34	-0,18	0,45
12	0,95	0,49	0,33	0,12	0,67	0,95	0,67	0,26	0,47	0,23	202,55	130,61	1,55	-0,26	1,29	ecart-type = 0,29 pour U	1,02	-0,30	0,48
13	0,22	0,76	0,33	0,61	0,65	0,47	0,76	0,23	0,90	0,74	125,55	75,14	1,67	-0,00	1,67		2,45	0,26	-0,33

Y = Fct + E = F + N(0,1) Fct = A / B

A = 40*exp(8*((X1-0,5)2 + (X2-0,5)**2))**

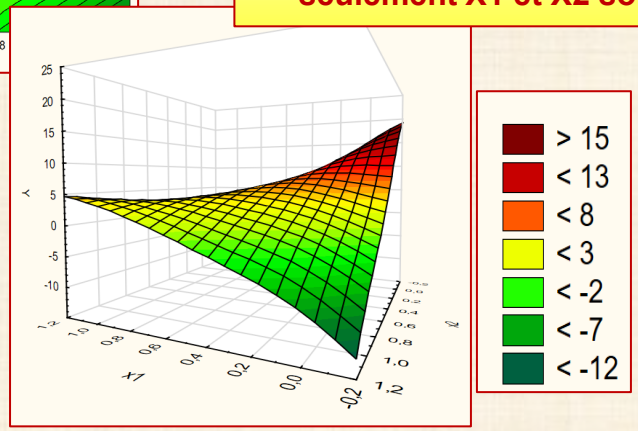
B = (exp(8*((X1-0,2)2+(X2-0,7)**2)) + exp(8*((X1-0,7)**2 + (X2-0,2)**2)))**

n = 400 observations p = 10 variables X1, X2,..., X10

x3 x4 ... x10 n'interviennent pas dans Fct

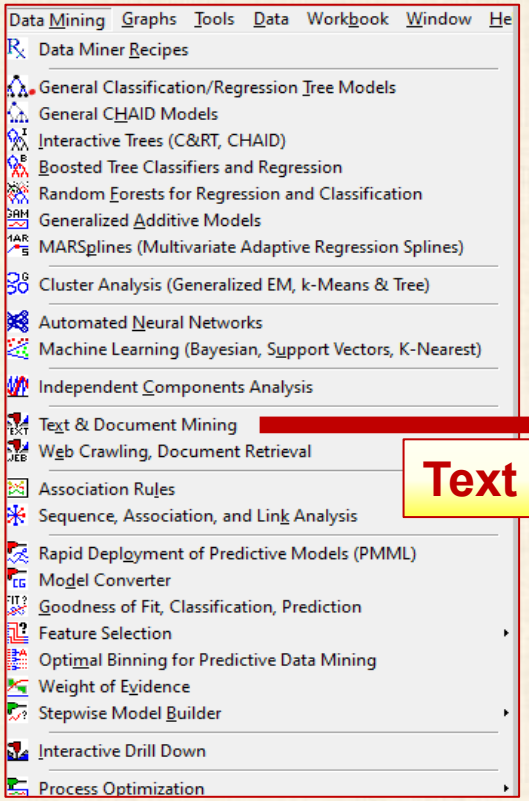
idem pour les variables u1 u2 ...u10

seulement X1 et X2 sont actives et définissent la fonction fct ci-haut



RÉGRESSION MARS

Ex4 : pourriels (spam)

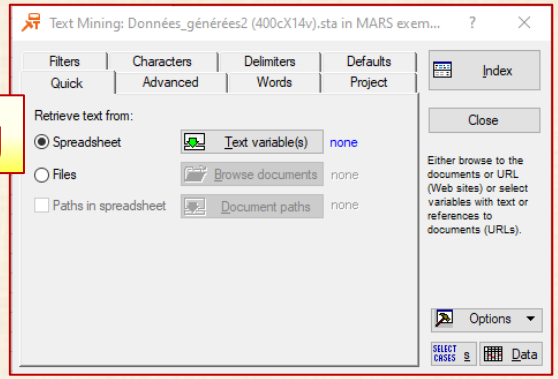


Text Mining

This corpus has been collected from free or free for research sources at the Web:
A collection of between 425 SMS spam messages extracted manually from the Grumbletext Web site.
This example concerns a study on classifying whether an e-mail is junk e-mail (coded as 1) or not (coded as 0).
The data were collected in Hewlett-Packard labs and donated by George Forman.
The data set contains 4,601 observations with 58 variables.
The response variable is a binary indicator of whether an e-mail is considered spam or not.
The 57 variables are continuous variables that record frequencies of some common words and characters in e-mails and lengths of uninterrupted sequences of capital letters.
The data set is publicly available at the UCI Machine Learning repository (Asuncion and Newman,
/

1 ID	2 SPAM?	3 MESSAGE
1	no	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	no	Ok lar... Joking wif u oni...
3	yes	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
4	no	U dun say so early hor... U c already then say...
5	no	Nah I don't think he goes to usf, he lives around here though
6	yes	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb

analyse mots clés



Address	Addresses	All	Bracket	Business
CS	CapAvg	CapLong	CapTotal	Conference
Credit	Data	Direct	Dollar	Edu
Email	Exclamation	Font	Free	George
HP	HPL	Internet	Lab	Labs
Mail	Make	Meeting	Money	Order
Original	Our	Over	PM	Paren
Parts	People	Pound	Project	RE
Receive	Remove	Report	Semicolon	Table
Technology	Telnet	Will	You	Your
_000	_85	_415	_650	_857

Variable Importance

Variable	Number of Bases	Importance
George	1	100.00
HP	1	78.35
Edu	3	61.25
Remove	2	49.21
Exclamation	3	44.14
Free	2	34.18
Meeting	3	32.57
_1999	2	29.71
Dollar	2	28.30
Money	3	26.39
CapLong	3	24.41
Our	2	19.46
Semicolon	2	14.98
RE	2	13.52
Business	3	13.48
Over	3	12.63
CapTotal	3	12.50
Will	1	10.81
Pound	2	9.73
Internet	1	5.88
_000	1	4.57
You	2	3.17

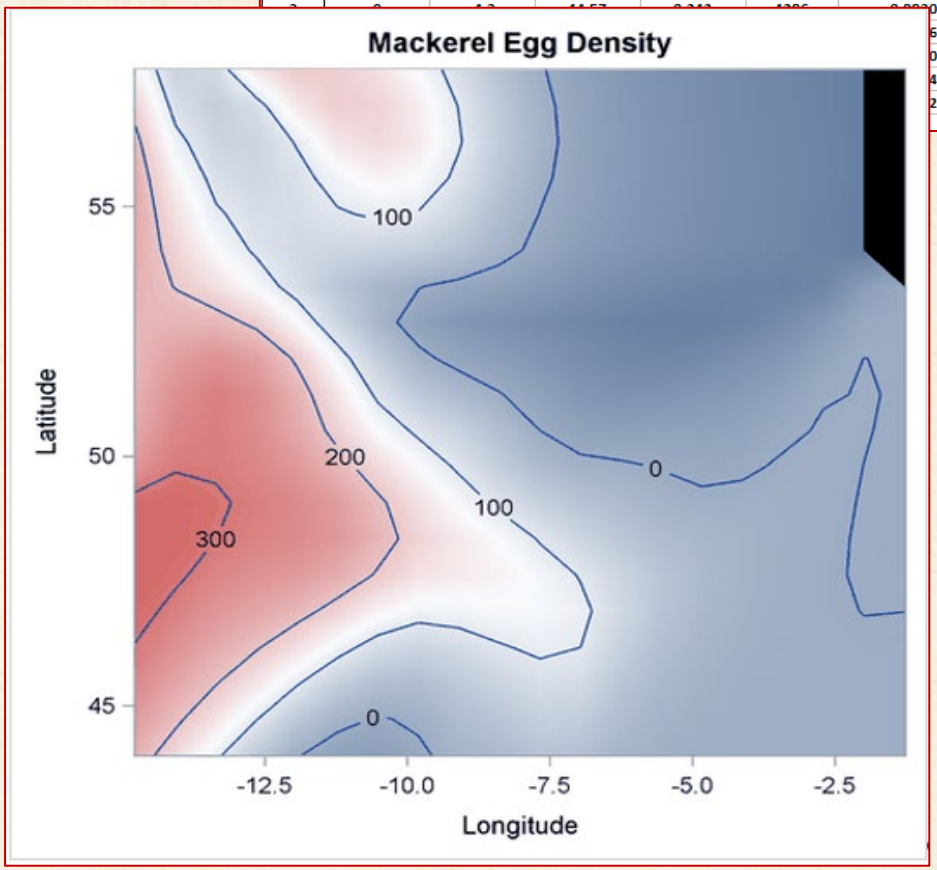
RÉGRESSION MARS

Ex5 : densité Œufs poisson maquereau

The example concerns a study of mackerel egg density

The example concerns a study of mackerel egg density
The data are a subset of the 1992 mackerel egg survey conducted over the Porcupine Bank west of Ireland.
The survey took place in the peak spawning area. Scientists took samples by hauling a net up from deep sea to the sea surface.
Then they counted the number of spawned mackerel eggs and used other geographic information
to estimate the sizes and distributions of spawning stocks.
The data set is used as an example in Bowman and Azzalini (1997)
This data set contains 634 observations and 7 variables.
response variable Egg_Count is the number of mackerel eggs collected from each sampling net.
Longitude and Latitude are the location values in degrees east and north, respectively, of each sample station.
Net_Area is the area of the sampling net in square meters.
Depth records the sea bed depth in meters at the sampling location.
Distance is the distance in geographic degrees from the sample location to the continental shelf edge.
/

1 ID	2 EggCount	3 longitude	4 latitude	5 Net_area	6 depth	7 distance	8 Y_density
1	0	-4.65	44.57	0.242	4342	0,83951	0,00000
2	0	-4.48	44.57	0.242	4334	0,85919	0,00000
						02	0,00000
						64	0,00980
						01	0,03922
						42	0,02941
						26	0,00000



Source : http://support.sas.com/documentation/cdl/en/statug/65328/HTML/default/viewer.htm#statug_adaptivereg_examples01.htm

RÉGRESSION MARS

Ex6 - pauvreté

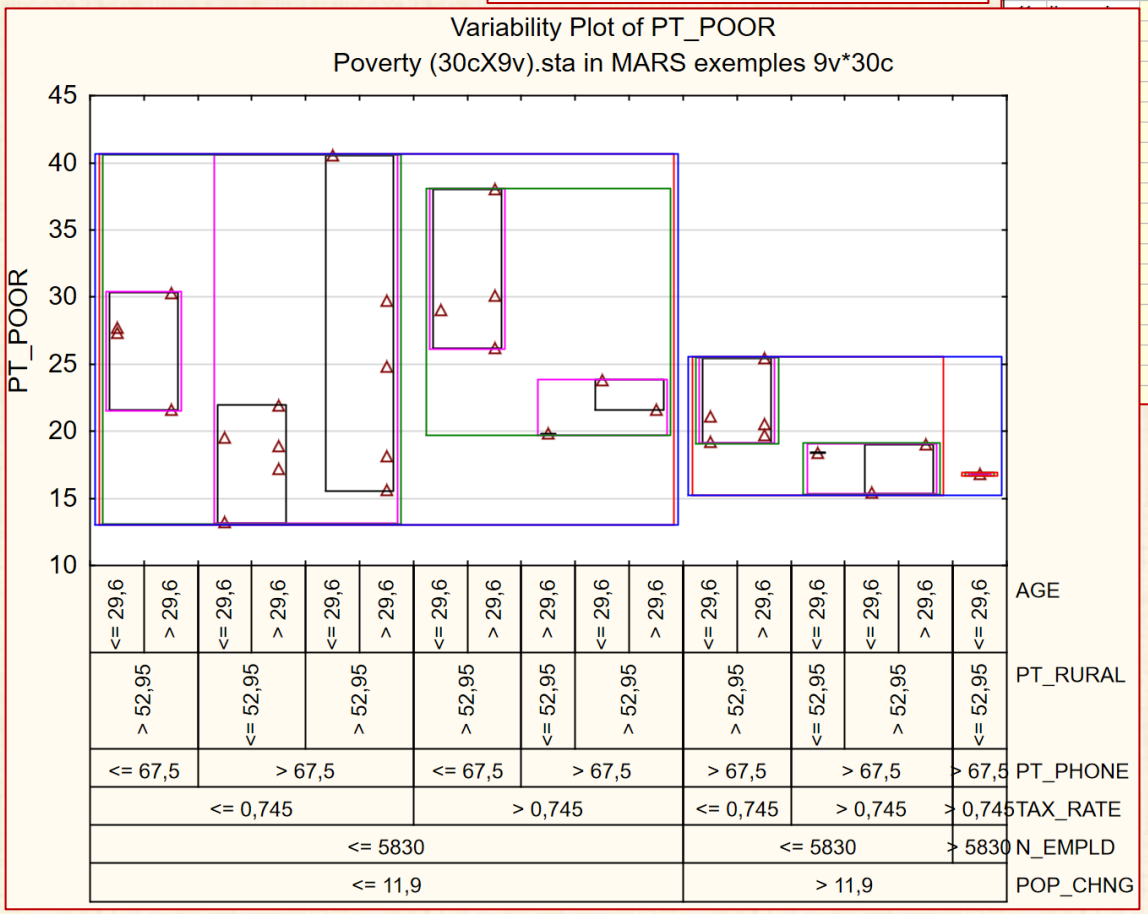
VARIABLES

X0 = County
Y = PT_POOR
= Percent of families below poverty level
X1 = POP_CHANGE
= Population change (1960-1970)
X2 = N_EMPLD
= No. of persons employed in agriculture
X3 = TAX_RATE
= Residential and farm property tax rate
X4 = PT_PHONE
= Percent residences with telephones
X5 = PT_RURAL = Percent rural population
X6 = AGE
Median age

prédiction de la pauvreté

X0 = County
Y = PT_POOR = Percent of families below poverty level
X1 = POP_CHANGE = Population change (1960-1970)
X2 = N_EMPLD = No. of persons employed in agriculture
X3 = TAX_RATE = Residential and farm property tax rate
X4 = PT_PHONE = Percent residences with telephones
X5 = PT_RURAL = Percent rural population
X6 = AGE = Median age
/

1 ID	2 County	3 PT_POOR	4 POP_CHNG	5 N_EMPLD	6 TAX_RATE	7 PT_PHONE	8 PT_RURAL	9 AGE
1	Benton	19,0	13,7	400	1,09	82	74,8	33,5
2	Cannon	26,2	-0,8	710	1,01	66	100,0	32,8
3	Carrol	18,1	9,6	1610	0,40	80	69,7	33,4
4	Cheatheam	15,4	40,0	500	0,93	74	100,0	27,8
5	Cumberland	29,0	8,4	640	0,92	65	74,0	27,9
6	DeKalb	21,6	3,5	920	0,59	64	73,1	33,2
7	Dyer	21,9	3,0	1890	0,63	82	52,3	30,8
8	Gibson	18,9	7,1	3040	0,49	85	49,6	32,4
9	Greene	21,1	13,0	2730	0,71	78	71,2	29,2
10	Hawkins	23,8	10,7	1850	0,93	74	70,6	28,7
		40,5	-16,2	2920	0,51	69	64,2	25,1
		21,6	6,6	1070	0,80	85	58,3	35,9
		25,4	21,9	160	0,74	69	100,0	31,4
		19,7	17,8	380	0,44	83	72,0	30,1
		38,0	-11,8	1140	0,81	54	100,0	34,1
		30,1	7,5	690	1,05	65	100,0	30,5
		24,8	3,7	1170	0,73	76	69,5	30,0
		30,3	1,6	1280	0,65	67	81,0	32,4
		19,5	8,4	2270	0,48	85	39,1	28,7
		15,6	2,7	960	0,72	84	58,4	33,4
		17,2	5,6	1710	0,62	84	42,4	29,9
		18,4	12,7	1410	0,84	86	36,4	23,3
		27,3	-4,8	200	0,73	66	99,8	27,5
		19,2	16,5	960	0,45	74	90,6	29,5
		16,8	15,2	11500	1,00	87	5,9	25,4
		13,2	11,6	1380	0,63	85	44,2	28,8
		29,7	4,9	530	0,54	70	100,0	33,1
		19,8	1,1	370	0,98	75	52,6	30,8
		27,7	3,8	440	0,46	48	100,0	28,4
		20,5	19,0	1630	0,68	83	72,1	30,4



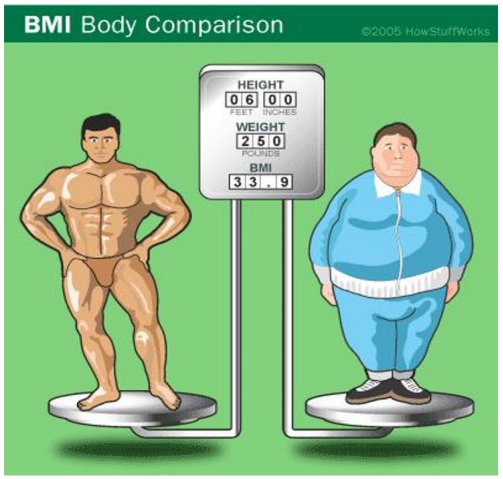
RÉGRESSION MARS

Ex7 : prédiction indice gras

Modelling Percent Body Fat in a Human Body Using Design of Experiments and Regression Analysis
Frank Deruyck, Dr. Sc., Lecturer, University College Ghent
http://lib.stat.cmu.edu/datasets/bodyfat
/

1 ID	2 Y_IndiceGras (%)	3 categori	4 genre	5 Age (an)	6 Poids kg)	7 Taille (m)	8 info	9 cou	10 poitrine	11 Abdomen	12 Biceps	13 avant bras	14 poignet	15 hanches	16 cuisses	17 genou	18 cheville
1	35,2	model	homme	46	164,7	1,82	mesures en cm	51,2	136,2	148,1	45,0	29,0	21,4	147,7	87,3	49,1	29,6
2	10,6	model	homme	57	67,0	1,66	circumference	35,2	99,6	86,4	31,7	27,3	16,9	90,1	53,0	35,0	21,3
3	24,2	model	homme	40	91,7	1,76		38,5	106,5	100,9	35,1	30,6	19,0	106,2	63,5	39,9	22,6
4	23,3	model	homme	52	75,7	1,71	1-cou	37,5	102,7	91,0	31,6	27,5	17,9	98,9	57,1	36,7	22,3
5	26,0	model	homme	54	104,3	1,82	2-poitrine	42,5	119,9	110,4	38,4	32,0	19,6	105,5	64,2	42,7	27,0
6	9,0	model	homme	47	83,6	1,88	3-Abdomen	37,3	99,6	88,8	30,3	27,9	17,8	101,4	57,4	39,6	24,6
7	22,1	model	homme	43	68,0	1,75	4-Biceps	35,2	91,1	85,7	29,4	26,6	17,4	96,9	55,5	35,7	22,0
8	9,4	model	homme	26	69,1	1,74	5-Avant bras	35,4	92,9	77,6	31,6	29,0	17,8	93,5	56,9	35,9	20,4
9	16,7	model	homme	40	71,7	1,75	6-poignet	36,3	97,0	86,6	29,8	26,3	17,3	92,6	55,9	36,3	22,1
10	29,9	model	homme	65	86,1	1,66	7-Hanches	40,8	106,4	100,5	35,9	30,5	19,1	100,5	59,2	38,1	24,0
11	11,7	model	homme	23	89,9	1,85	8-Cuisses	42,1	99,6	88,6	35,6	30,0	19,2	104,1	63,1	41,7	25,0
12	15,1	model	homme	34	63,5	1,78	9-Genou	36,0	89,2	83,4	28,3	26,2	16,5	89,6	52,4	35,6	20,4
13	18,7	model	homme	50	88,3	1,78	10-Cheville	39,0	103,7	97,6	32,7	30,0	19,0	104,2	60,0	40,9	25,5
14	17,5	model	homme	46	75,7	1,69		36,6	101,0	89,9	35,6	30,2	17,6	100,0	60,7	36,0	21,9

BODY FAT ≠ BMI



BMI = weight/length²

APPROACH

Calculation Body Volume by numerical integration → Time consuming!

%Fat = f(volume/weight)

+ Weight

Regression Analysis

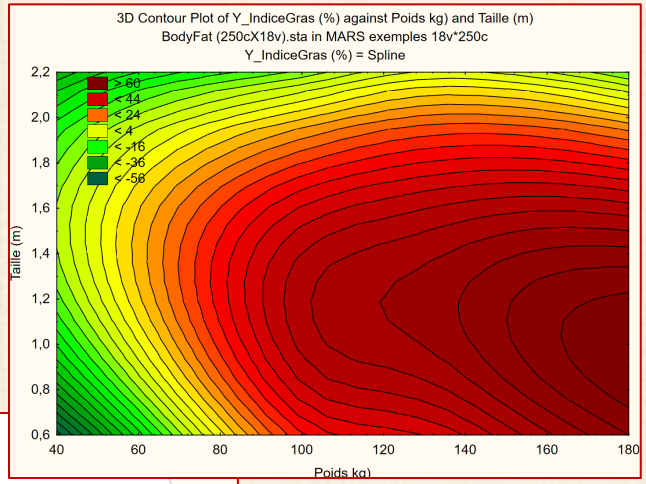
Easy, manual measurements

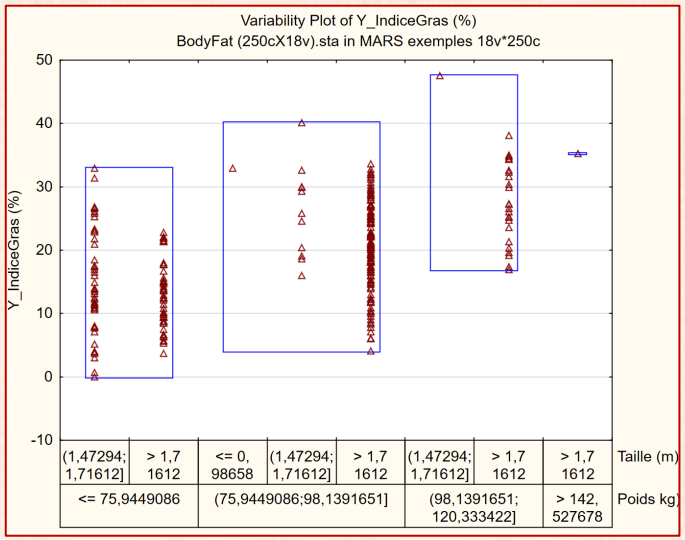
Body Volume = f(body measurements)

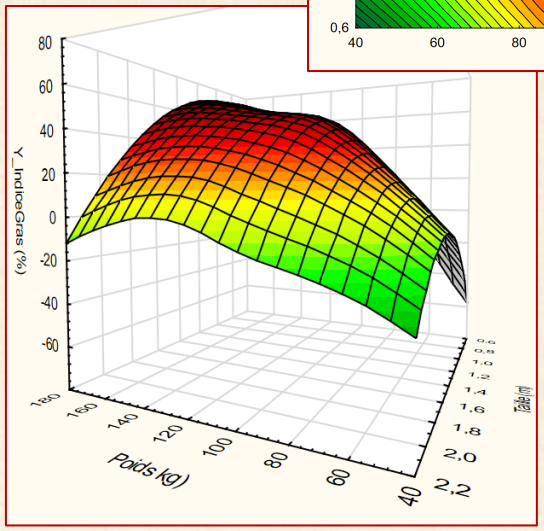
Model %Fat

SMARTFIT database
> 50 body measurements
n = 1200 adult records

Parameter selection → PCA
Sample extraction → DOE





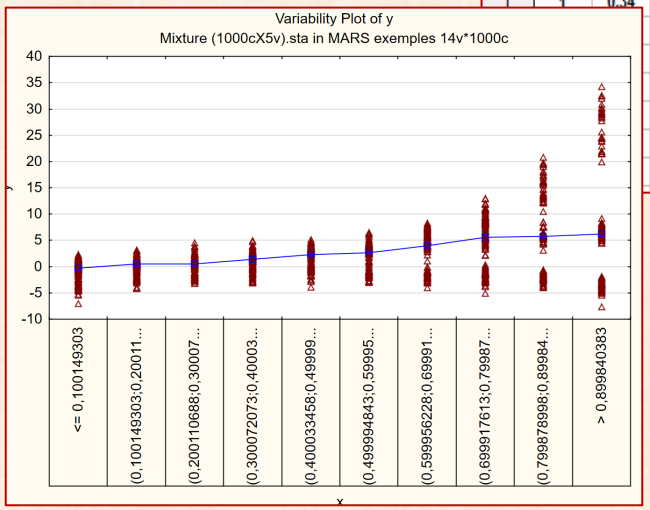


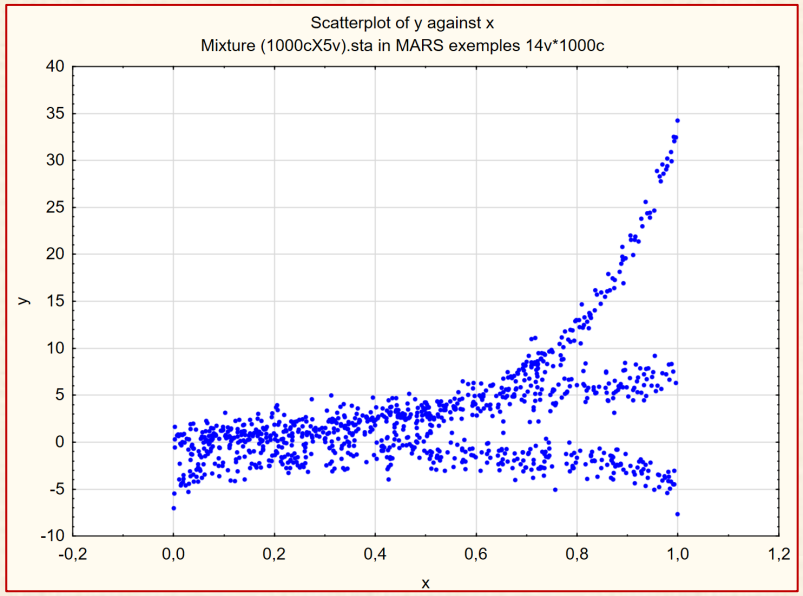
RÉGRESSION MARS

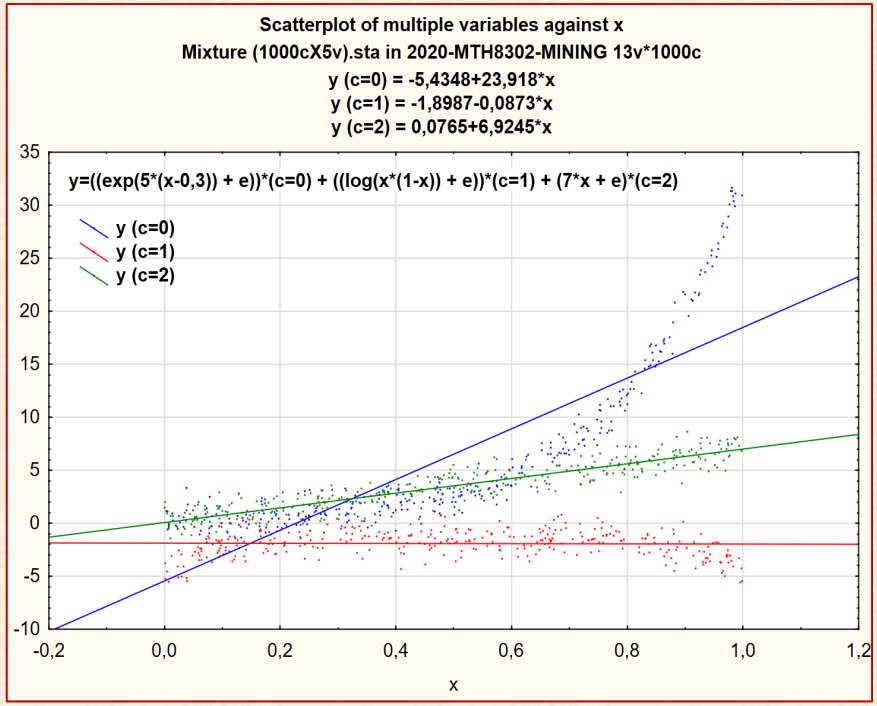
Ex8 : mélanges

Source : SAS SAS/STAT® 13.1 User's Guide
The ADAPTIVEREG Procedure (2013) p. 928
y=((exp(5*(x-0,3)) + e))*(c=0) + ((log(x*(1-x)) + e))*(c=1) + (7*x + e)*(c=2)

	1 ID	2 x	3 c1	4 c	5 e	6 y	7 c7	8 y (c=0)	9 y (c=1)	10 y (c=2)	11 c11	12 Y_pred (c=0)	13 Y_pred (c=1)	14 Y_pred (c=2)
1	1	0,34	2,62	2	0,47	2,84		12,39				10,81		
			0,59	0	-0,08	21,51		2,42				2,27		
			1,06	1	0,72	-0,70			-3,55				-3,62	
			1,09	1	-0,10	-3,09				0,18				0,67
			1,35	1	-0,48	-2,23		14,96				16,25		
			0,63	0	0,56	0,79			0,22				-1,19	
			0,60	0	0,34	1,11		5,22				4,45		







RÉGRESSION MARS

Ex9 : Diabètes

bon modèle
Y vs X = ?

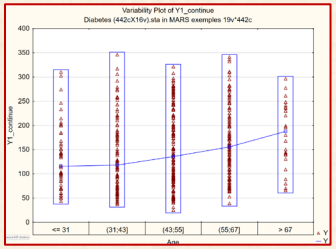
réponse Y Y1_continue

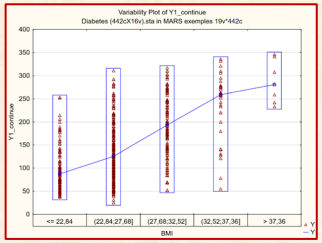
facteurs X

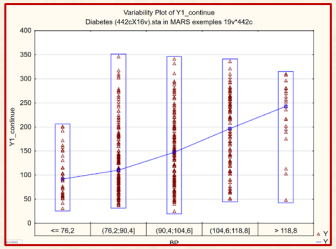
- X1_Age
- X2_Gender
- X3_BMI
- X4_BP
- X5_Total Cholesterol
- X6_LDL
- X7_HDL
- X8_TCH
- X9_LTG
- X10_Glucose

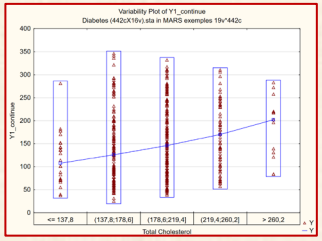
Efron, B., Hastie, T., Johnstone, J., and Tibshirani, R. (2004). Least Angle Regression.
Annals of Statistics (with discussion), vol. 32, pp. 407-499.
DATA Diabetes / n = 442 obs. X 13 variables : 10 X explicatives (v2 v3 ... v11) et 3 Y (continue, binaire, ordinale)
Y1_continue is a quantitative measure of disease progression one year after baseline. (25 à 346)
Y2_binaire et Y3_ordinale sont des recodages de Y1_continue
Y2_binaire = Low si Y_continue = 200 ou moins / = Hig si Y_continue 201 ou plus
Y3_ordinale = low si Y_continue = 150 ou moins / = Medium si Y_continue comprise entre 151 et 200 / = High si Y_continue = 201 ou plus
données séparées en 2 groupes : Training (309 obs.) Validation (133 obs.)
OBSERVATIONNELLES
/

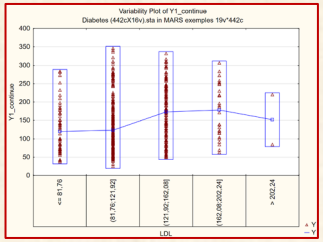
1 ID	2 Age	3 Gender	4 BMI	5 BP	6 Total Cholesterol	7 LDL	8 HDL	9 TCH	10 LTG	11 Glucose	12 Y1_continue	13 Y2_binaire	14 Y3_ordinale	15 Validation	16 Validation2
1	59	2	32,1	101,00	157	93,2	38	4,00	4,8598	87	151	Low	Medium	Training	1
2	48	1	21,6	87,00	183	103,2	70	3,00	3,8918	69	75	Low	Low	Validation	2
3	72	2	30,5	93,00	156	93,6	41	4,00	4,6728	85	141	Low	Low	Training	1
4	24	1	25,3	84,00	198	131,4	40	5,00	4,8903	89	206	High	High	Training	1
5	50	1	23,0	101,00	192	125,4	52	4,00	4,2905	80	135	Low	Low	Training	1
6	23	1	22,6	89,00	139	64,8	61	2,00	4,1897	68	97	Low	Low	Training	1
7	36	2	22,0	90,00	160	99,6	50	3,00	3,9512	82	138	Low	Low	Training	1
8	66	2	26,2	114,00	255	185,0	56	4,55	4,2485	92	63	Low	Low	Validation	2
9	60	2	32,1	83,00	179	119,4	42	4,00	4,4773	94	110	Low	Low	Training	1
10	29	1	30,0	85,00	180	93,4	43	4,00	5,3845	88	310	High	High	Training	1
11	22	1	18,6	97,00	114	57,6	46	2,00	3,9512	83	101	Low	Low	Validation	2

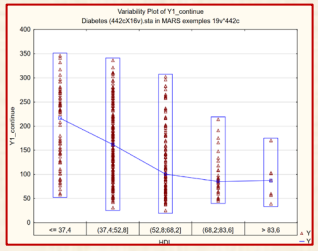


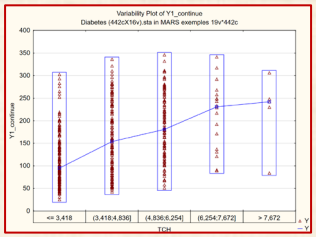


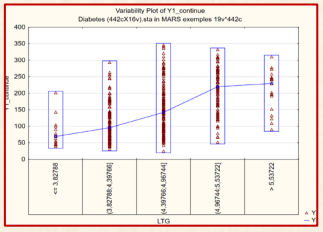








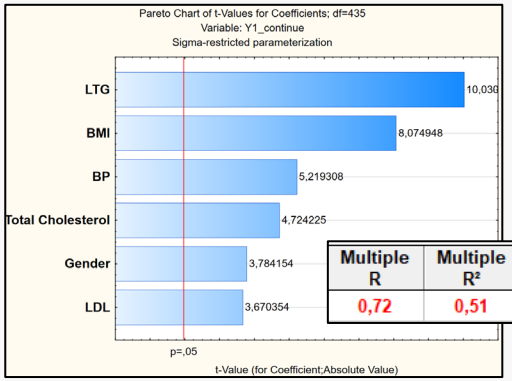




r>= -1 -0,80 -0,60 -0,40 -0,20 0 0,20 0,40 0,60 0,80 1

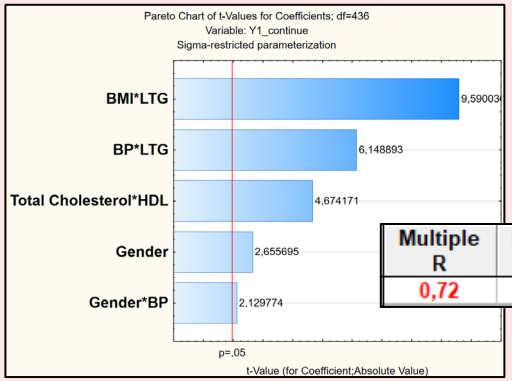
Variable	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose	Y1_continue
BMI	1,00	0,40	0,25	0,26	-0,37	0,41	0,45	0,39	0,59
BP		1,00	0,24	0,19	-0,18	0,26	0,39	0,39	0,44
Total Cholesterol			1,00	0,90	0,05	0,54	0,52	0,33	0,21
LDL				1,00	-0,20	0,66	0,32	0,29	0,17
HDL					1,00	-0,74	-0,40	-0,27	-0,39
TCH						1,00	0,62	0,42	0,43
LTG							1,00	0,46	0,57
Glucose								1,00	0,38
Y1_continue									1,00

modèle 1 : effets principaux



Multiple R	Multiple R²	Adjusted R²
0,72	0,51	0,51

modèle 2 : effets principaux + interactions + quadratiques



Multiple R	Multiple R²	Adjusted R²
0,72	0,52	0,51

Data Mining (ML) : défis … problèmes … difficultés

- Préparation des données, valeurs manquantes, …
- Choix, type, rôle, nature des variables (X, Y)
- Transformations / codages
- Détection **interactions**
- Grande quantité de données
- Complexité inconnue des relations
- Nécessaire : connaissance domaine d'application
- Critères d'évaluation / comparaison des modèles :
  peu ou pas des tests statistiques
- Nouveaux critères de performance :
  courbes ROC - Lift Chart
- Création ensembles : Test Validation

Régression Ordinaire (RO) : DIFFICULTÉS

- **Moindres carrés: minimise Erreur Quadratique Moyenne (MSE)**

$$MSE = \sum (y_i - y_i)^2$$

y_i valeurs observées y_i valeurs prédites

- **Manque de robustesse … instabilité … données aberrantes**
- **Sur ajustement avec les données d'apprentissage**

celles qui servent à estimer les paramètres du modèle

- **Instabilité en présence de la multicolinéarité**
- **Difficultés avec les transformations et codage des prédicteurs X**
- **Solution unique n'existe pas avec des milliers de variables X**

fléau de la dimensionalité « curse of dimensionality »

= faible densité données dans espaces de dimension élevée (p)

même si n est très grand

- **Solution ne tient pas compte de contraintes sur les coefficients**

par exemple: on veut des coefficients positifs

Régression Régularisée (RR) : contraintes coefficients modèle

Régression Ordinaire RO

minimiser

MSE + λ complexité modèle

minimiser

Régression Régularisée RR

$$MSE + \lambda \sum |\beta|^{k}$$

Ridge
norme L2
coefficients
$\sum \beta^2$

Lasso
norme L1
coefficients
$\sum |\beta|$

Comptage
norme L0
plus petit
sous ensemble
de variables
$\sum |\beta|^0$

k paramètre d'élasticité k = 0 ou 1 ou 2

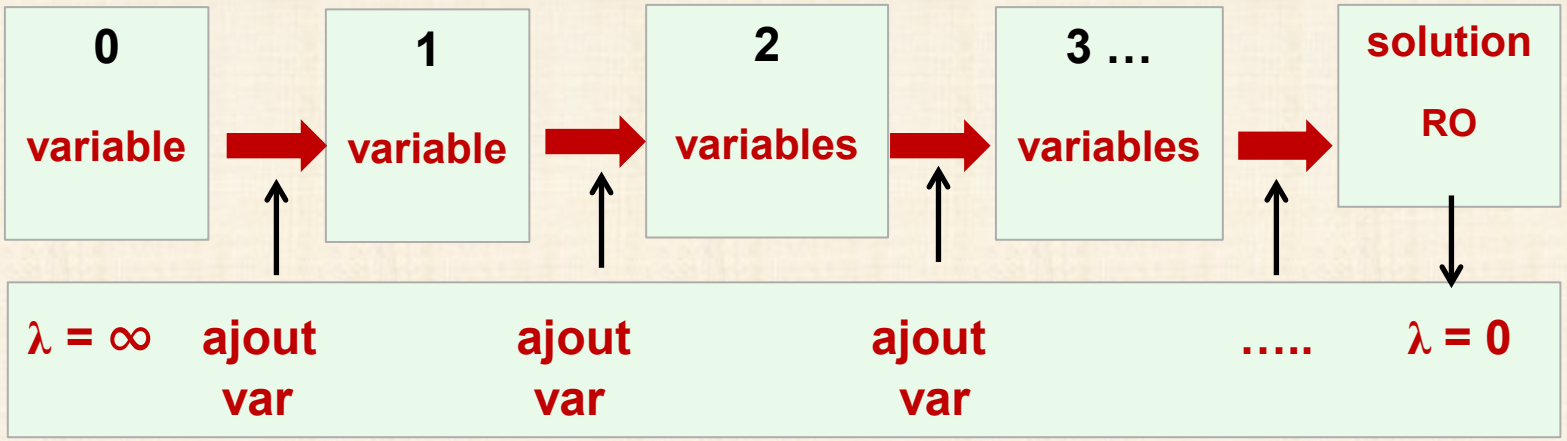
λ paramètre de régularisation

$\lambda = 0$ moindres carrés ordinaire

$0 < \lambda < \infty$ compromis à choisir

$\lambda = \infty$ cas limite avec coefficients = 0

processus itératif : General Path Seeker = GPS



Stratégie de sélection des variables

http://www.salford-systems.com

maintenant MINITAB

contrôle sélection variables k

k = 2 - Régression Ridge : minimise carrés - norme L2 $\Sigma\beta^2$

k = 1 - Régression Lasso: minimise valeurs absolues - norme L1 $\Sigma|\beta|$

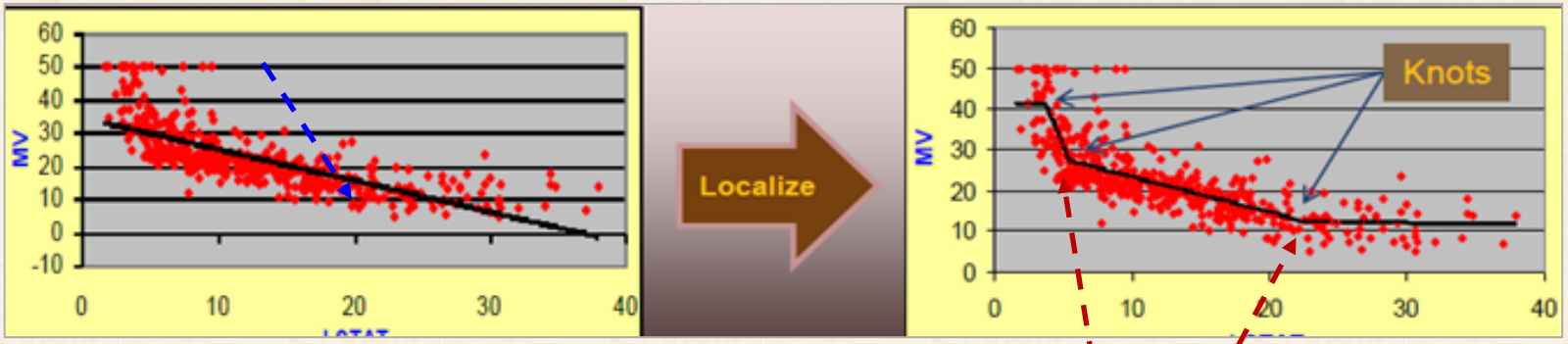
k = 0 - Régression Stepwise : comptage - norme L0 $\Sigma|\beta|^0$

Regression Ordinaire (RO) et GPS versus MARS

RO suite de modèles linéaires
modèles 1-variables / modèles 2-variables / modèles 3-variables …
variables retenues (critère): X1, X2, X3,…
variables ignorées X4 X5, …

GPS suite de modèles coefficients variables
modèles 1-variables / modèles 2-variables / …
sur un ensemble d'apprentissage
sélection variables et optimisation sur ensemble test

RO et PLS : incapable découvrir structure locale mais MARS oui



logiciel hautement recommandé
http://www.salford-systems.com

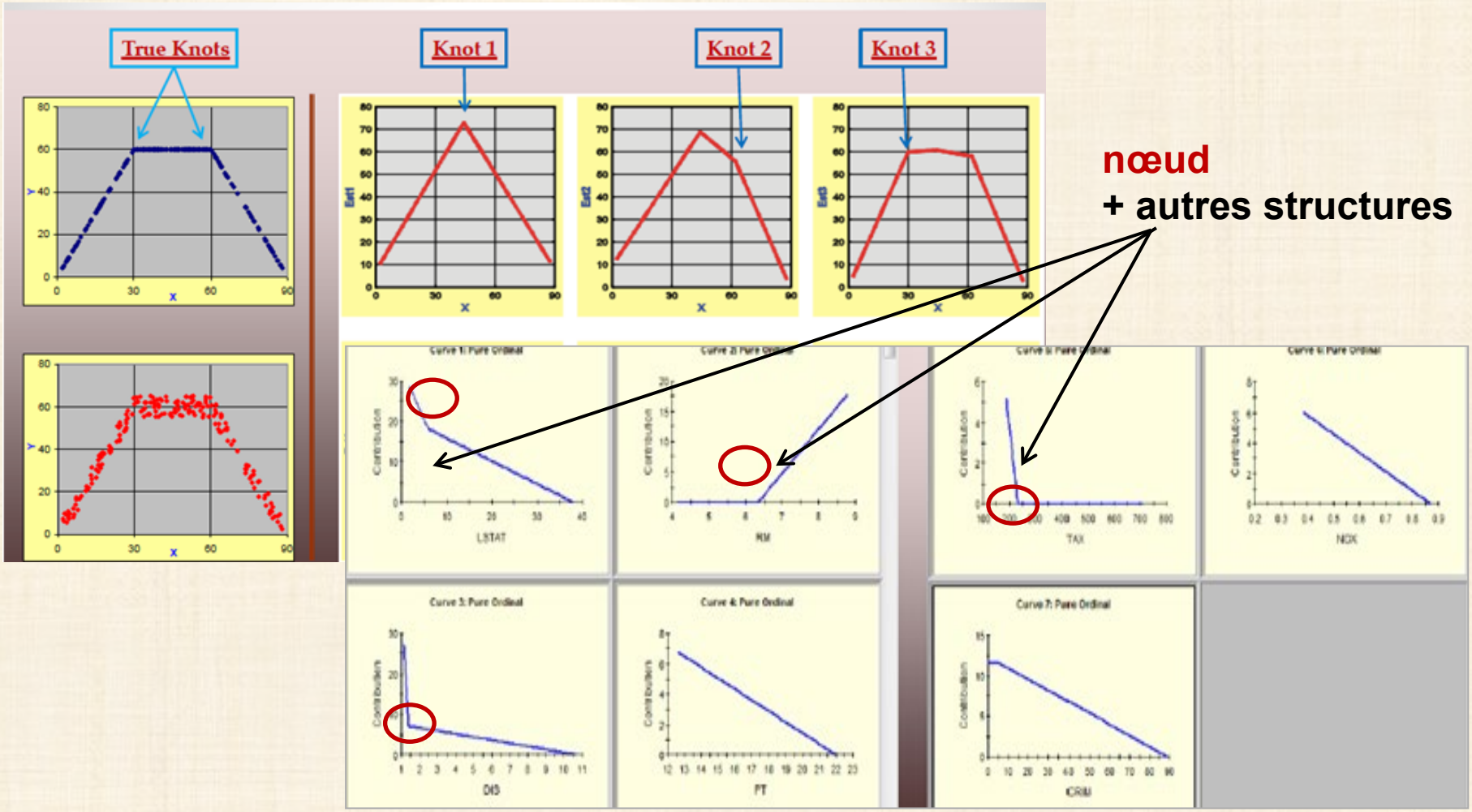
SPM® | CART | Random Forests® | MARS® | TREENET

vidéo de 15 minutes
https://www.salford-systems.com/resources/webinars-tutorials/how-to/how-to-build-a-model

présence nœuds naturels
mais aussi autres cas

nouvelle approche : RÉGRESSION MARS

- Développement **d'algorithmes itératif** (adaptatifs) **non-linéaire** qui construisent la fonction localement sur les données
- Utilisation de **splines** (*fonction définie par morceaux avec des polynômes*) d'ordre 1 avec **nœuds optimaux à déterminer**

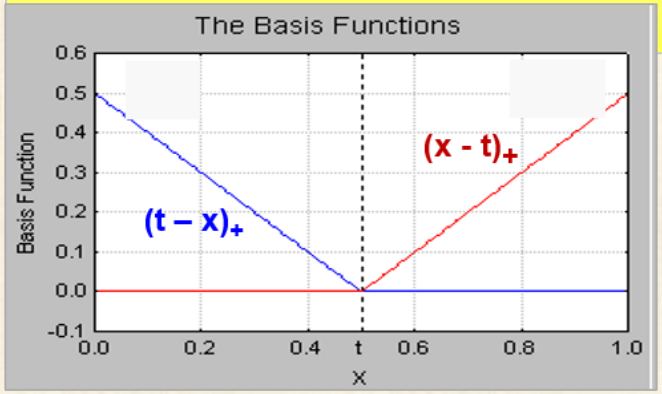


RÉGRESSION MARS

- Développée par Jerome Friedman (1991) (Stanford University)
- Applicable en régression (Y continue) et classification (Y catégorique)
- Peut traiter
  - variables X explicatives continues ou catégoriques
  - données manquantes
  - fonctions hautement non linéaires
  - interactions de tout ordre
- Performance aussi bonne / meilleure que les réseaux de neurones

  mais modèle plus facile à comprendre : pas une boite noire !
- Aucune hypothèse a priori sur relation $Y = f(X_1, X_2, \ldots, X_p)$
- Construction d'une fonction avec une série de coefficients β et les

  fonctions de base h_m



$$h_m(x;t) = (t-x)_+ = \begin{cases} t-x & \text{si } x \le t \\ 0 & \text{si } x \ge t \end{cases}$$

$$y = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

RÉGRESSION MARS :
modèle
régression non paramétrique

$$y = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

$$h_m(x; t) = (x - t)_+ = \begin{cases} x - t & \text{si } x \le t \\ 0 & \text{si } x \ge t \end{cases}$$

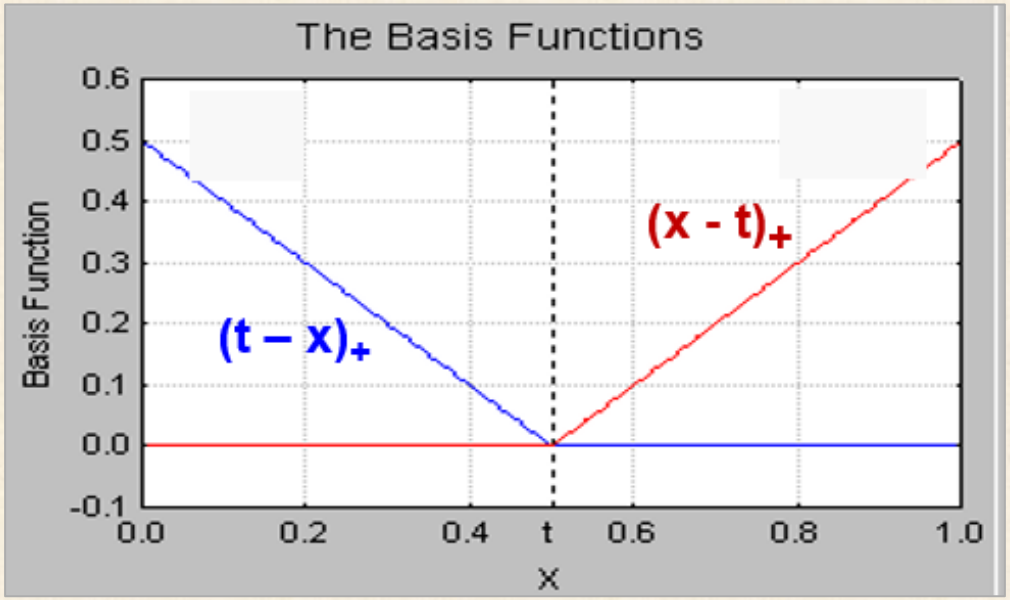
$$h_m(x; t) = (t - x)_+ = \begin{cases} t - x & \text{si } x \le t \\ 0 & \text{si } x \ge t \end{cases}$$

h_m **fonction de base**
h_m **fonction de base**

type bâtons de hockey

t : nœud

déterminés optimalement



RÉGRESSION MARS

$$y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m H_{km}\left(x_{v(k,m)}\right)$$

$$H_{km}\left(x_{v(k,m)}\right) = \prod_{k=1}^{K} h_{km}$$

$$GCV = \frac{\sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2}{\left(1 - \frac{C}{N}\right)^2}$$

$$C = 1 + c\,d$$

GVC Generalized Cross Validation error : mesure la qualité d'ajustement

N : nombre d'observations
d : degré de liberté effectif
= nombre de fonctions de base
c: pénalité d'ajout de fonction de base
valeur recommandée entre 2 et 3

Recommandations

- **Maximum fonctions de base**
- **Augmenter le degré des interactions**
- **Émondage (« prune »)**

RÉGRESSION MARS : processus

- Partition de l'espace X en régions avec chacune leurs équations / classification
- Algorithme itératif adaptatif détermine des régression linéaires par morceaux
- Variables les plus importantes et leurs interactions sont déterminées
- Exhibe un haut degré de flexibilité …. Mais possibilité de sur ajustement
- Méthodes pour éviter SUR AJUSTEMENT ('overfitting'):

  émondage ('pruning') + contrôle du nombre de fonctions de base
- Modèle MARS optimal : résultat de 2 phases : phase 1 (AVANT) phase 2 (ARRIÈRE)

  - phase 1 : construction du modèle en ajoutant des fonctions de base

    effets principaux, nœuds, interactions

  - phase 2 (pruning) : élimination des fonctions qui contribuent peu

RÉGRESSION MARS : processus construction modèle

- **Multivariate Adaptive Regression Splines (MARS) algorithme** (Friedman, 1991)
  algorithme prédictif combinant des transformations **non paramétriques** des variables explicatives avec un **partitionnement récursif** de l'espace.
- Extension des **modèles linéaires** qui modélise les non linéarités et les Interactions entre les variables.
- MARS construit modèle en **2 phases**: phase **AVANT** suivi phase **ARRIERE.**
- **Phase AVANT**
  MARS commence avec un effet général (intercepte) et ajoute des fonctions de base en paires au modèle.
- **Phase ARRIÈRE**
  phase AVANT fait usuellement un **SUR ajustement** ('overfitting') au modèle;
  phase ARRIERE fait **l'émondage** ('pruning') du modèle :
  enlèvement des termes un à un jusqu'à déterminer le meilleur sous modèle

RÉGRESSION MARS : processus construction modèle

- phase **AVANT** ajoute des termes en paires mais la phase **ARRIÈRE** typiquement on enlève un coté de la paire ; on ne retrouve pas souvent des termes en paires dans le modèle.
- phase **ARRIÈRE** utilize le critère **GVC** : **Generalized Cross Validation** pour comparer la performance des modèles afin de déterminer le meilleur modèle ; les petites valeurs de **GVC** sont préférables.
- **GCV** est une forme de **régularisation** : équilibre entre la qualité de l'ajustement et la complexité du modèle.

$$GCV = \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\left(1 - \frac{C}{N}\right)^2}$$

$$C = 1 + cd$$

N : nombre d'observations
d : degré de liberté effectif
= nombre de fonctions de base
c : pénalité d'ajout de fonction de base
recommandation : $2 \leq c \leq 3$

Recommandations

- Maximum fonctions de base
- Augmenter le degré des interactions (2 optimal ?)
- Émondage (« pruning »)

RÉGRESSION MARS : avantages

- MARS models are more flexible than linear regression models.
  MARS models are simple to understand and interpret.
  much simpler than neural network and random forest.

- MARS can handle both continuous and categorical data. MARS tends to be better than recursive partitioning for numeric data because hinges are more appropriate for numeric variables than the piecewise constant segmentation used by recursive partitioning.

- Building MARS models often requires little or no data preparation.
  The hinge functions automatically partition the input data, so the effect of outliers is contained.

- MARS is similar to recursive partitioning which also partitions the data into disjoint regions, although using a different method.
  Nevertheless, as with most statistical modeling techniques, known outliers should be considered for removal before training a MARS model

- MARS (like recursive partitioning) does automatic variable selection (meaning it includes important variables in the model and excludes unimportant ones).
  However, there can be some arbitrariness in the selection, especially when there are correlated predictors, and this can affect interpretability

RÉGRESSION MARS : avantages

- MARS models tend to have a good bias-variance trade-off. The models are flexible enough to model non-linearity and variable interactions (thus MARS models have fairly low bias), yet the constrained form of MARS basis functions prevents too much flexibility (thus MARS models have fairly low variance).

- MARS is suitable for handling fairly large datasets. It is a routine matter to build a MARS model from an input matrix with, say, 100 predictors and 10 000 obs.

  Such a model can be built in about a minute on a 1 GHz machine, assuming the maximum degree of interaction of MARS terms is limited to one (i. e. additive terms only). A degree two model with the same data on the same 1 GHz machine takes longer - about 12 minutes. Be aware that these times are highly data dependent. Recursive partitioning is much faster than MARS.

- With MARS models, as with any non-parametric regression, parameter confidence intervals and other checks on the model cannot be calculated directly (unlike linear regression models).

- Cross-validation and related techniques must be used for validating the model.

Forward Selection

The forward selection process in the multivariate adaptive regression splines algorithm is as follows:

1. Initialize by setting $\mathbf{B}_0 = \mathbf{1}$ and $M = 1$.
2. Repeat the following steps until the maximum number of bases $M_{\max}$ has been reached or the model cannot be improved by any combination of $\mathbf{B}_m$, $\mathbf{v}$, and t.
   a. Set the lack-of-fit criterion $\mathrm{LOF}^* = \infty$.
   b. For each selected basis: $\mathbf{B}_m, m \in \{0, \ldots, M-1\}$ do the following for each variable $\mathbf{v}$ that $\mathbf{B}_m$ does not consist of $\mathbf{v} \notin \{\mathbf{v}(k,m) | 1 \le k \le K_m\}$
      i. For each knot value (or a subset of categories) t of $\mathbf{v}: t \in \{\mathbf{v} | \mathbf{B}_m > 0\}$, form a model with all currently selected bases $\sum_{i=0}^{M-1} \mathbf{B}_i$ and two new bases: $\mathbf{B}_m \mathbf{T}_1(\mathbf{v}, t)$ and $\mathbf{B}_m \mathbf{T}_2(\mathbf{v}, t)$.
      ii. Compute the lack-of-fit criterion for the new model LOF.
      iii. If $\mathrm{LOF} < \mathrm{LOF}^*$, then update $\mathrm{LOF}^* = \mathrm{LOF}$, $m^* = m$, $\mathbf{v}^* = \mathbf{v}$, and $t^* = t$.
   c. Update the model by adding two bases that improve the most $\mathbf{B}_{m^*} \mathbf{T}_1(\mathbf{v}^*, t^*)$ and $\mathbf{B}_{m^*} \mathbf{T}_2(\mathbf{v}^*, t^*)$.
   d. Set $M = M + 2$.

The essential part of each iteration is to search a combination of $\mathbf{B}_m$, $\mathbf{v}$, and t such that adding two corresponding bases most improve the model. The objective of the forward selection step is to build a model that overfits the data. The lack-of-fit criterion for linear models is usually the residual sum of squares (RSS).

Source : SAS
http://support.sas.com/documentation/cdl/en/statug/65328/HTML/default/viewer.htm#statug_adaptivereg_details01.htm

SAS
MARS = PROC ADAPTIVREG

Algorithme

RÉGRESSION MARS

Backward Selection

The backward selection process in the multivariate adaptive regression splines algorithm is as follows:

1. Initialize by setting the overall lack-of-fit criterion: $\mathrm{LOF}^* = \infty$.
2. Repeat the following steps until the null model is reached. The final model is the best one that is found during the backward deletion process.
   a. For a selected basis $\mathbf{B}_m, m \in \{1, \ldots, M\}$:
      i. Compute the lack-of-fit criterion, LOF, for a model that excludes $\mathbf{B}_m$.
      ii. If $\mathrm{LOF} < \mathrm{LOF}^*$, save the model as the best one. Let $m^* = m$.
      iii. Delete $\mathbf{B}_{m^*}$ from the current model.
   b. Set $M = M - 1$.

The objective of the backward selection is to "prune" back the overfitted model to find the best model that has good predictive performance. So the lack-of-fit criteria that characterize model loyalty to original data are not appropriate. Instead, the multivariate adaptive regression splines algorithm uses a quantity similar to the generalized cross validation criterion. See the section Goodness-of-Fit Criteria for more information.

RÉGRESSION MARS

Ex1 : données financières

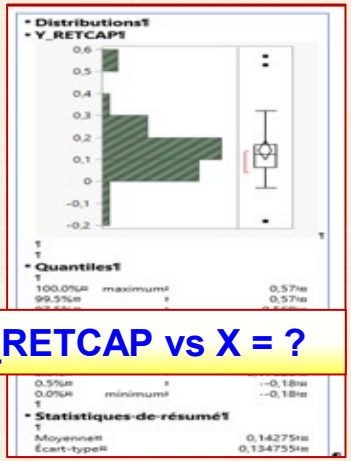
Financial data of 40 UK companies 1983

J. Jobson *Applied Multivariate Data Analysis*, vol 1, Regression and Experimental Design 1991, Springer-Verlag

Y_RETCAP	Return on capital employed
X1_WCFTCL	Ratio of working capital flow to total current liabilities
X2_WCFTDT	Ratio of working capital flow to total debt
X3_GEARRAT	Gearing ratio (debt-equity ratio)
X4_LOGSALE	Log to base 10 of total sales
X5_LOGASST	Log to base 10 of total assets
X6_NFATAST	Ratio of net fixed assets to total assets
X7_CAPINT	Capital intensity (ratio of total sales to total assets)
X8_FATTOT	Gross sixed assets to total assets
X9_INVTAST	Ratio of total inventories to total assets
X10_PAYOUT	Payout ratio
X11_QUIKRAT	Quick ratio
X12_CURRAT	Current ratio

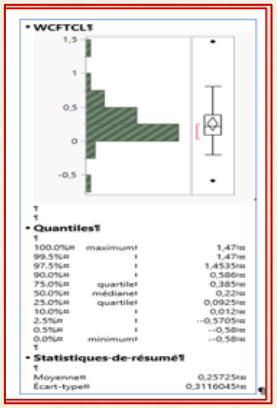
1 ID	2 Y_RETCAP	3 WCFTCL	4 WCFTDT	5 GEARRAT	6 LOGSALE	7 LOGASST	8 NFATAST	9 CAPINT	10 FATTOT	11 INVTAST	12 PAYOUT	13 QUIKRAT	14 CURRAT
1	0,26	0,25	0,25	0,46	4,11	4,30	0,10	0,64	0,12	0,74	0,07	0,18	1,53
2	0,57	0,33	0,33	0,00	4,25	4,00	0,12	1,79	0,15	0,27	0,30	1,26	1,73
3	0,09	0,50	0,20	0,24	4,44	4,88	0,94	0,36	0,97	0,01	0,57	0,39	0,44
4	0,32	0,23	0,21	0,45	4,71	4,44	0,29	1,86	0,52	0,29	0,00	0,69	1,23
5	0,17	0,21	0,12	0,91	4,85	4,75	0,26	1,26	0,54	0,33	0,31	0,90	1,76

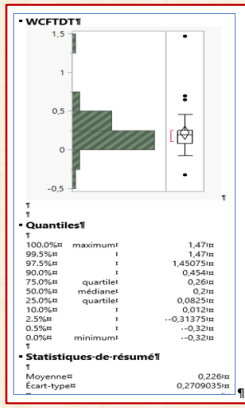
Y_RETCAP

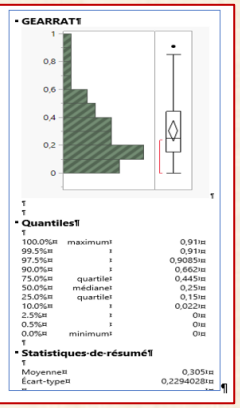


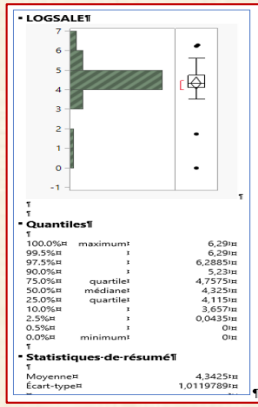
modèle Y_RETCAP vs X = ?

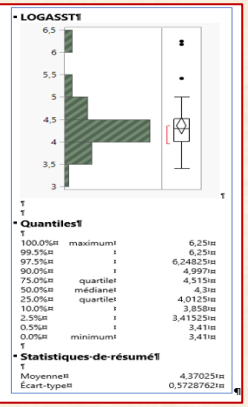
X

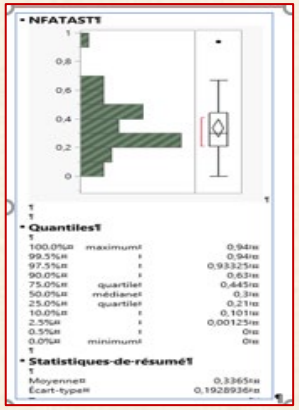


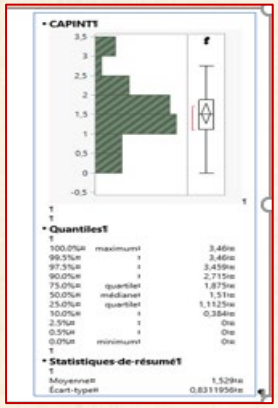


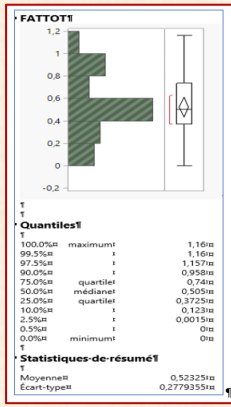


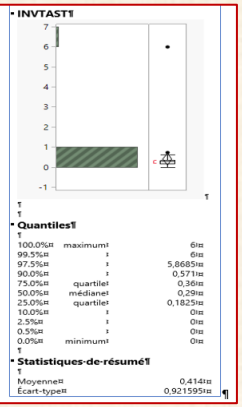


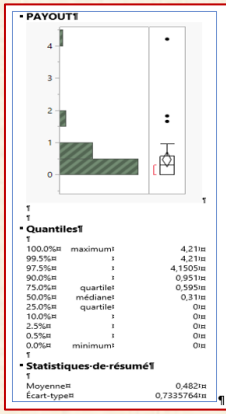


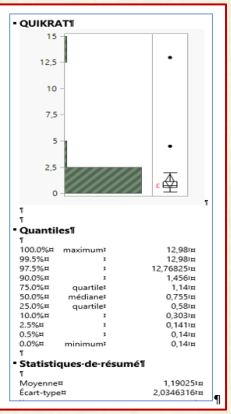


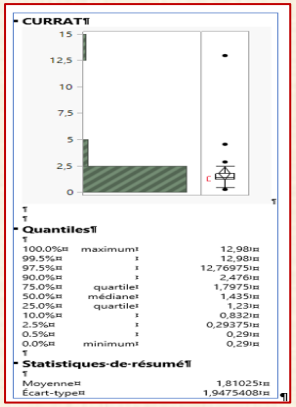












REGRESSION MARS : Ex1 - données financières

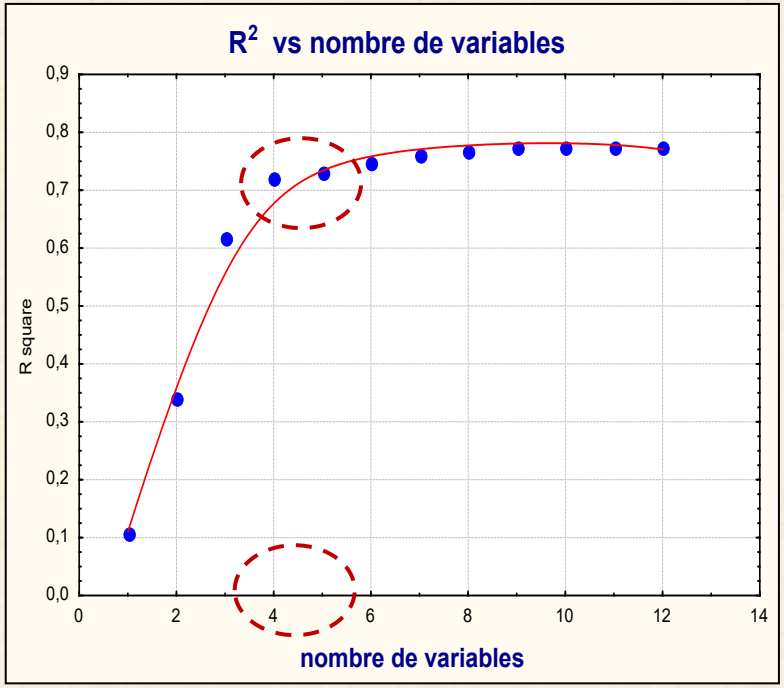
Exemple: données financières de 40 entreprises – 12 variables explicatives

Summary of best subsets; variable(s): Y_RETCAP
Max R square and standardized regression coefficients for each sub model

R square	No. of - Effects	WCF TCL	WCF TDT	GEA RRA T	LOG SAL E	LOGA SST	NFAT AST	CAPI NT	FAT TOT	INVT AST	PAY OUT	QUIK RAT	CUR RAT
0,775	12	0,42	0,72	-0,01	0,84	-0,35	-0,54	-0,06	-0,23	0,00	-0,09	1,07	-1,76
0,775	11	0,42	0,72	-0,01	0,84	-0,35	-0,54	-0,06	-0,23		-0,09	1,07	-1,76
0,775	10	0,42	0,74		0,84	-0,35	-0,54	-0,06	-0,23		-0,09	1,06	-1,76
0,773	9	0,43	0,71		0,76	-0,29	-0,52		-0,23		-0,09	0,93	-1,64
0,766	8		1,22		0,75	-0,25	-0,51		-0,25		-0,08	0,84	-1,67
0,760	7		1,23		0,74	-0,24	-0,53		-0,23			0,92	-1,76
0,747	6		1,21		0,71	-0,19	-0,73					1,02	-1,85
0,732	5		1,21		0,51		-0,73					0,68	-1,63
0,721	4		1,23		0,46		-0,68						-1,00
0,615	3	1,04					-0,61						-0,99
0,341	2		1,01									-0,95	
0,106	1	0,32											

WCF TDT LOG SALE NFAT AST CUR RAT

Max R^2



REGRESSION MARS : Ex1 - données financières

Regression Summary for Dependent Variable: Y_RETCAP

R = 0,88 R^2 = 0,775 Adjusted R^2 = 0,675 F(12,27) = 7,7355

	Beta	Std.Err. - of Beta	B	Std.Err. - of B	t(27)	p-level
Intercept			0,230	0,135	1,706	0,0995
WCFTCL	0,422	0,471	0,183	0,204	0,897	0,3775
WCFTDT	0,720	0,615	0,358	0,306	1,170	0,2522
GEARRAT	-0,011	0,128	-0,007	0,075	-0,088	0,9305
LOGSALE	0,845	0,282	0,112	0,038	2,991	0,0059
LOGASST	-0,346	0,198	-0,081	0,047	-1,749	0,0916
NFATAST	-0,545	0,204	-0,380	0,142	-2,672	0,0126
CAPINT	-0,062	0,148	-0,010	0,024	-0,418	0,6790
FATTOT	-0,227	0,186	-0,110	0,090	-1,217	0,2343
INVTAST	0,001	0,097	0,000	0,014	0,010	0,9922
PAYOUT	-0,093	0,100	-0,017	0,018	-0,938	0,3568
QUIKRAT	1,072	0,749	0,071	0,050	1,431	0,1639
CURRAT	-1,764	0,683	-0,122	0,047	-2,584	0,0155

coefficients de régression

coefficients de régression en variables centrées-réduites

Analysis of Variance DV: Y_RETCAP

	Sums of - Squares	df	Mean - Squares	F	p-level
Regress.	0,5486	12	0,0457	7,73	0,00001
Residual	0,1596	27	0,0059		
Total	0,7082				

4 variables importantes pour expliquer Y_RETCAP

LOGSALE - LOGASST

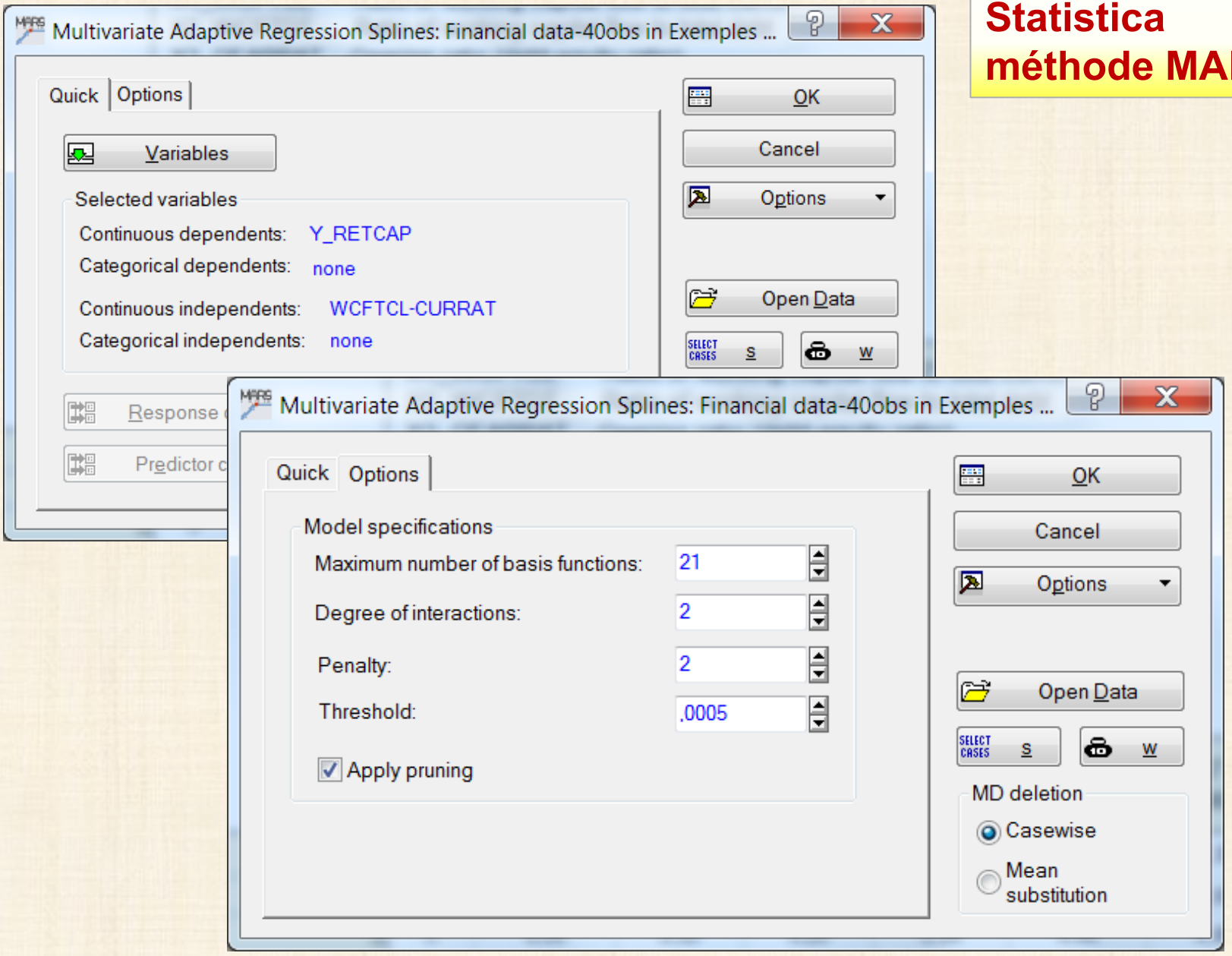
NFATAST - CURRAT

R^2 = 0,775

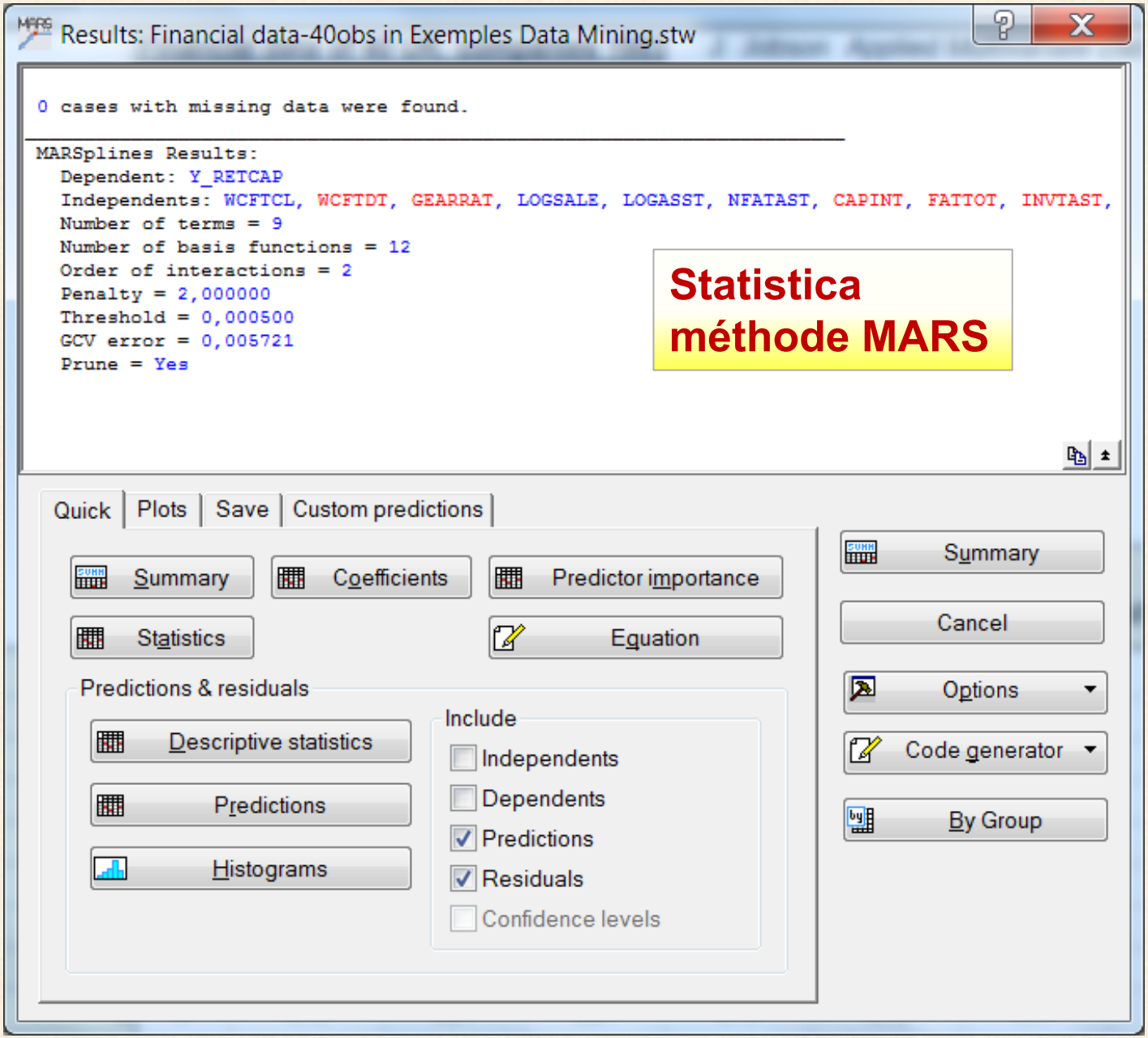
R^2 ad = 0,675

REGRESSION MARS : Ex1 - données financières

Statistica méthode MARS



REGRESSION MARS : Ex1 - données financières



REGRESSION MARS : Ex1 - données financières

Statistica
méthode MARS

Model specifications	Model Summary Value
Independents	12
Dependents	1
Number of terms	9
Number of basis functions	12
Order of interactions	2
Penalty	2,000000
Threshold	0,000500
GCV error	0,005721
Prune	Yes

Regression statistics	Regression statistics Y_RETCAP
Mean (observed)	0,142750
Standard deviation (observed)	0,134755
Mean (predicted)	0,142750
Standard deviation (predicted)	0,127353
Mean (residual)	-0,000000
Standard deviation (residual)	0,044046
R-square	0,893164
R-square adjusted	0,861113

Dependents	Number of References to Each Predictor Number of times each predictor is used References (to Basis Functions)
WCFTCL	0
WCFTDT	3
GEARRAT	1
LOGSALE	0
LOGASST	0
NFATAST	0
CAPINT	1
FATTOT	4
INVTAST	2
PAYOUT	1
QUIKRAT	0
CURRAT	0

MARS

R-square = 0,89

R-square ajusted = 0,86

versus

meilleur modèle avec sélection variables

R-square = 0,77 (12 var)

= 0,74 (6 var)

R-square ajusted = 0,67 (12 var)

REGRESSION MARS : Ex1 - données financières

Statistica méthode MARS

Équation de prédiction … sortie de Statistica ….. assez illisible

Y_RETCAP = 1,61467455597378e-001 + 1,09830689920240e+000*max(0; WCFTDT-2,00000000000000e-001)
- 8,85189400626038e-001*max(0; 2,00000000000000e-001-WCFTDT) + 3,95394205024779e-001*max(0; 5,00000000000000e-001-FATTOT) - 6,86221641809597e-001*max(0; 1,65000000000000e+000-CAPINT)*max(0; 5,00000000000000e-001-FATTOT) - 1,65400960342744e+000*max(0; 3,50000000000000e-001-GEARRAT)*max(0; FATTOT-5,00000000000000e-001) - 4,27901294646320e+000*max(0; WCFTDT-2,00000000000000e-001)*max(0; 2,90000000000000e-001-INVTAST) + 2,11780184468778e+000*max(0; 5,00000000000000e-001-FATTOT)*max(0; 3,30000000000000e-001-INVTAST) + 1,40994840889384e-001*max(0; 4,60000000000000e-001-PAYOUT)

édition (un peu pénible …) pour lisibilité .. édition

Y_RETCAP = 0,161
+ 1,098*max(0; WCFTDT - 0,2)
- 0,885*max(0; 0,2 - WCFTDT)
+ 0,395*max(0; 0,5 - FATTOT)
+ 0,140*max(0; 0,46 - PAYOUT)
- 0,686*max(0; 1,65 - CAPINT)*max(0; 0,5 - FATTOT)
- 1,654*max(0; 0,35 - GEARRAT)*max(0; FATTOT – 0,5)
- 4,279*max(0; WCFTDT – 0,2)*max(0; 029 - INVTAST)
+ 2,118*max(0; 0,5 - FATTOT)*max(0; 0,33 - INVTAST)

interactions

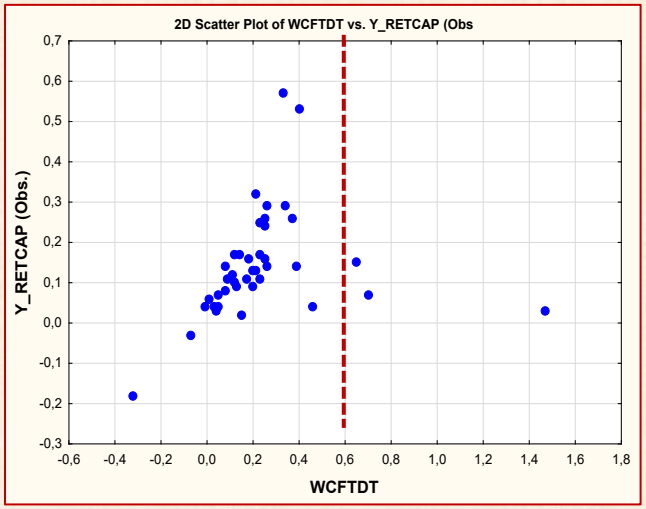
REGRESSION MARS : Ex1 - données financières

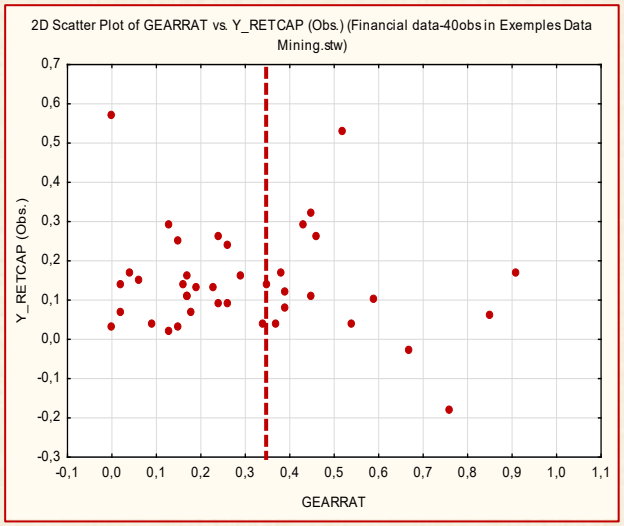
Coefficients, knots and basis functions	Model coefficients NOTE: Highlighted cells indicate basis functions of type max(0, independent-knot), otherwise max(0, knot-independent)						
	Coefficients Y_RETCAP	Knots WCFTCL	Knots WCFTDT	Knots GEARRAT	Knots LOGSALE	Knots LOGASST	Knots NFATAST
Intercept	0,16147						
Term.1	1,09831		0,200000				
Term.2	-0,88519		0,200000				
Term.3	0,39539						
Term.4	-0,68622						
Term.5	-1,65401			0,350000			
Term.6	-4,27901		0,200000				
Term.7	2,11780						
Term.8	0,14099						

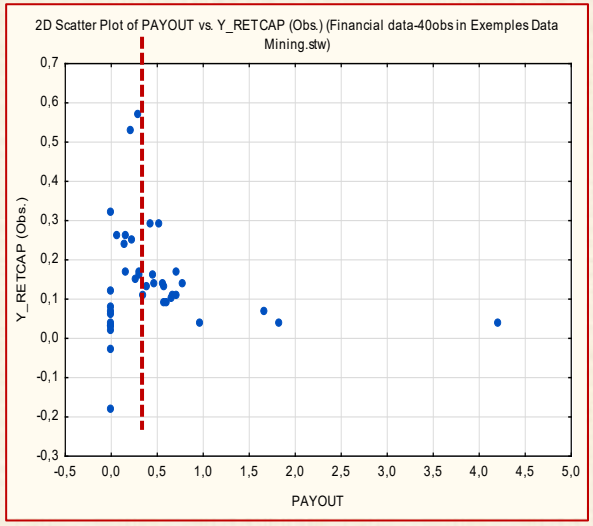
Statistica méthode MARS

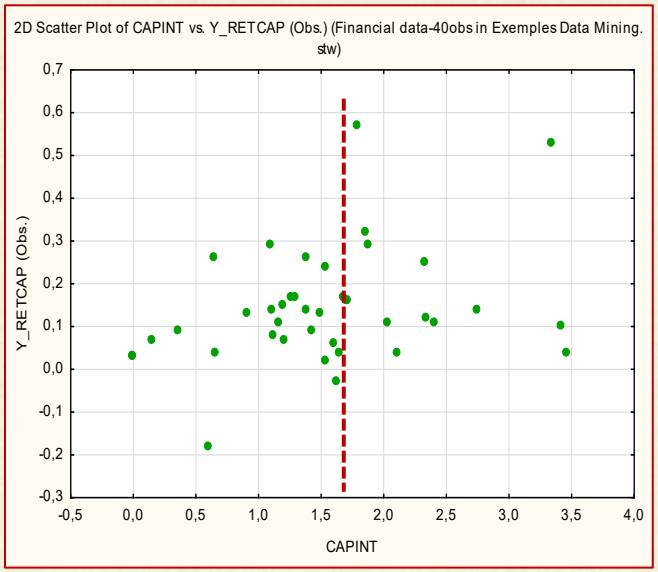
Coefficients, knots and basis functions	Model coefficients NOTE: Highlighted cells indicate basis functions of type max(0, independent-knot), otherwise max(0, knot-independent)					
	Knots CAPINT	Knots FATTOT	Knots INVTAST	Knots PAYOUT	Knots QUIKRAT	Knots CURRAT
Intercept						
Term.1						
Term.2						
Term.3		0,500000				
Term.4	1,650000	0,500000				
Term.5		0,500000				
Term.6			0,290000			
Term.7		0,500000	0,330000			
Term.8				0,460000		

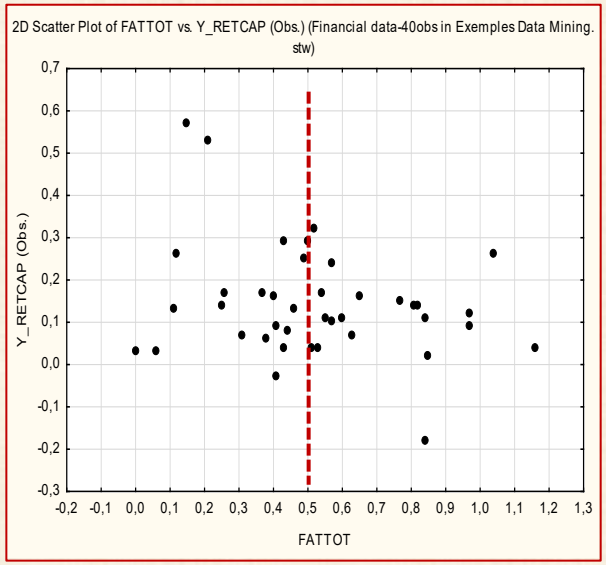
REGRESSION MARS : localisation des nœuds

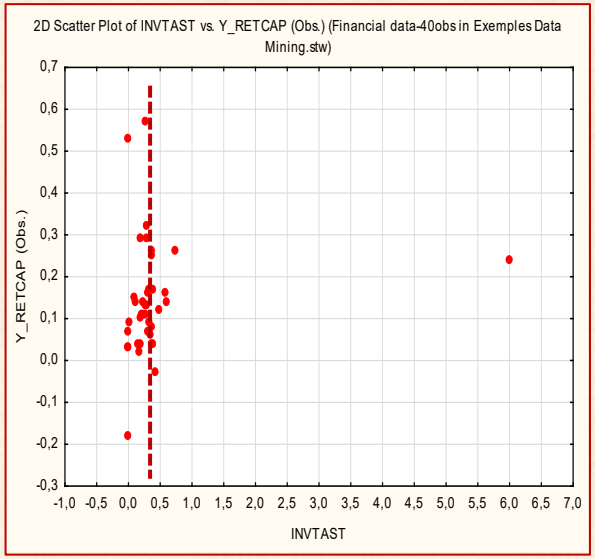








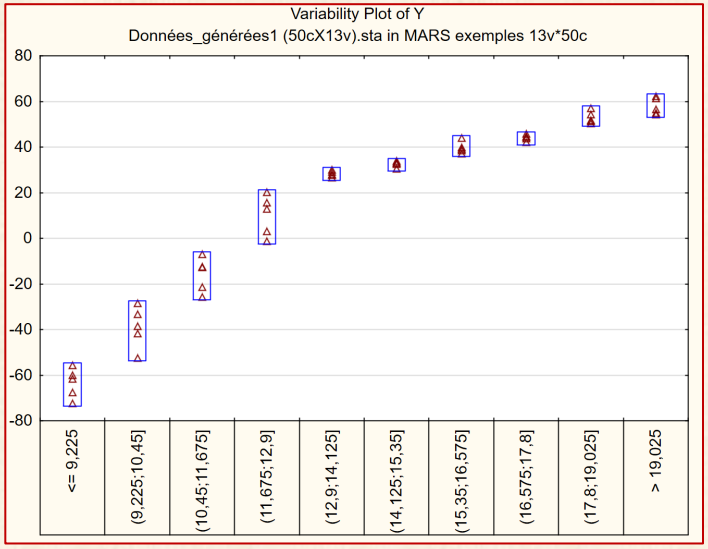




nœuds ? interprétation ? poids ?

RÉGRESSION MARS

Ex2 : données simulées 1



Simulation **Regression MARS avec une seule variable X**

X : varie entre 8 et 20 50 valeurs : 8,00 8,25 ... 20,25

if1 = 1 si x plus petit que 13 et if1 = 0 si x plus grand que 13

if2 = 0 si x plus petit que 13 et if2 = 1 si x plus grand que 13

u1 = x - 13 max1 = max(u1,0) f1 = 25 + 5*u1;

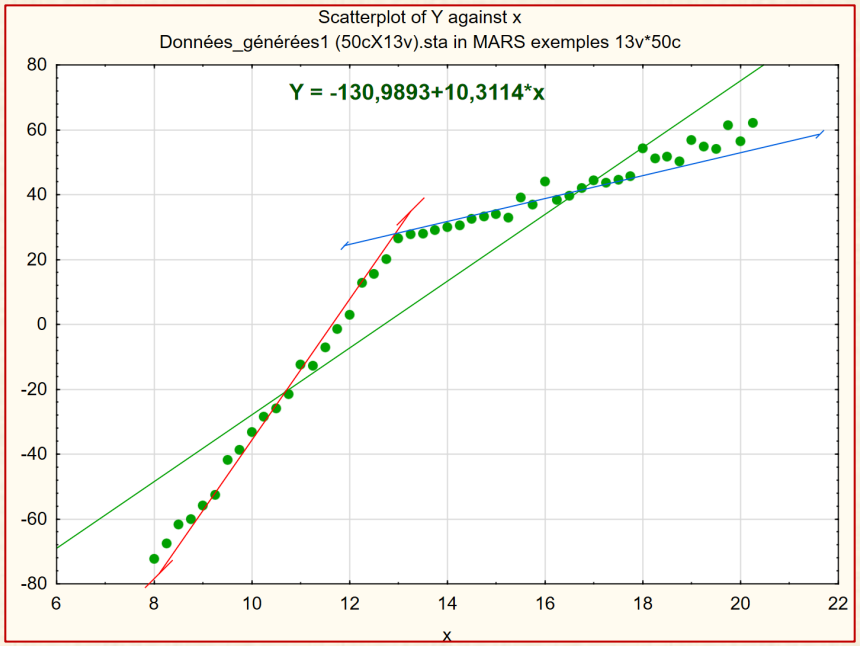
u2 = 13 -x max2 = max(u2,0) f2 = 25 - 20*u2;

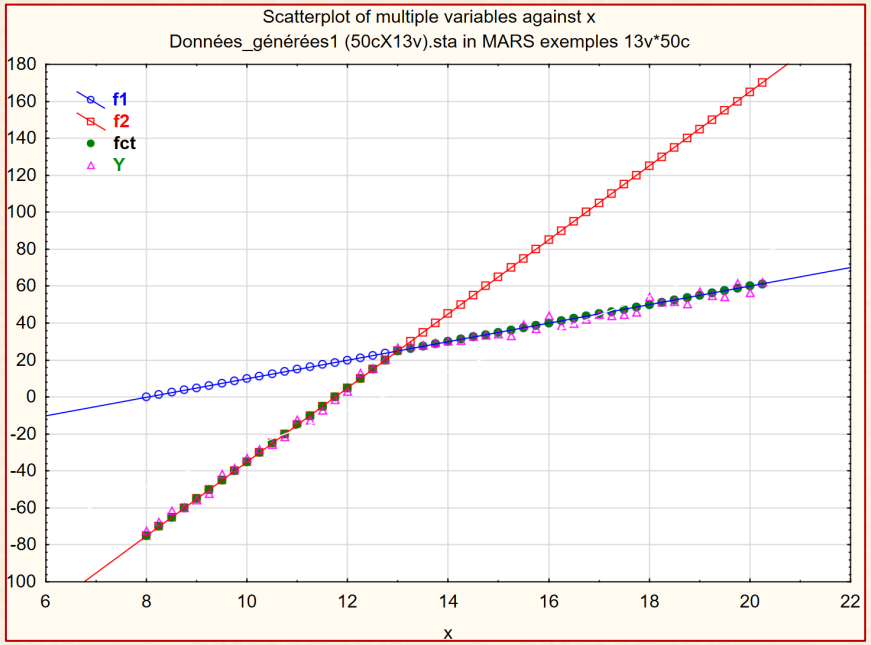
fct = 25 + 5*max1 - 20*max2 = 25 + 5*max(0;x-13) - 20*max(0;13-x)

Y = fct + erreur erreur = (N(0, 2)

/

1 ID	2 x	3 if1	4 if2	5 u1	6 max1	7 u2	8 max2	9 f1	10 f2	11 fct	12 erreur	13 Y
1	8,00	1	0	-5,0	0,0	5,0	5,0	0,00	-75,00	-75,00	2,58	-72,42
2	8,25	1	0	-4,8	0,0	4,8	4,8	1,25	-70,00	-70,00	2,36	-67,64
3	8,50	1	0	-4,5	0,0	4,5	4,5	2,50	-65,00	-65,00	3,19	-61,81
4	8,75	1	0	-4,3	0,0	4,3	4,3	3,75	-60,00	-60,00	-0,15	-60,15
5	9,00	1	0	-4,0	0,0	4,0	4,0	5,00	-55,00	-55,00	-0,86	-55,86





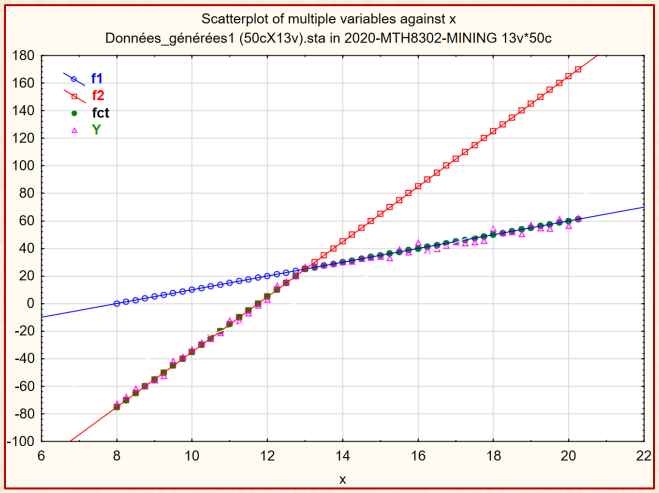
REGRESSION MARS : Ex2 - données simulées

Simulation Regression MARS avec une seule variable X
X : varie entre 8 et 20 50 valeurs : 8,00 8,25 ... 20,25
if1 = 1 si x plus petit que 13 et if1 = 0 si x plus grand que 13
if2 = 0 si x plus petit que 13 et if2 = 1 si x plus grand que 13
u1 = x - 13 max1 = max(u1,0) f1 = 25 + 5*u1;
u2 = 13 -x max2 = max(u2,0) f2 = 25 - 20*u2;

fct = 25 + 5*max1 - 20*max2 = 25 + 5*max(0;x-13) - 20*max(0;13-x)
Y = fct + erreur erreur = (N(0, 2)
/

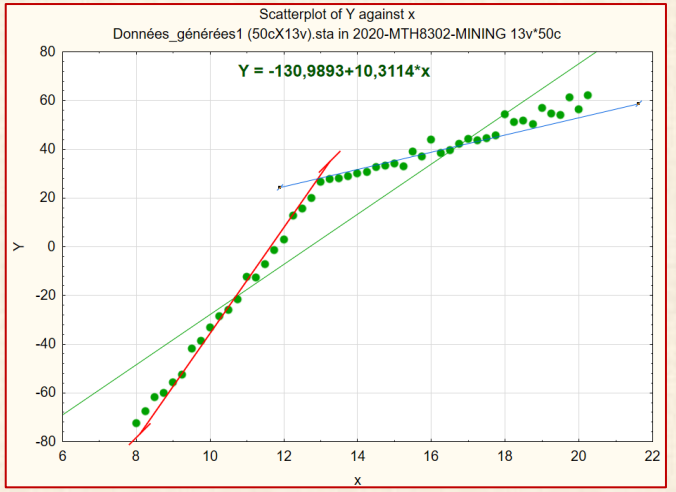
1 ID	2 x	3 if1	4 if2	5 u1	6 max1	7 u2	8 max2	9 f1	10 f2	11 fct	12 erreur	13 Y
1	8,00	1	0	-5,0	0,0	5,0	5,0	0,00	-75,00	-75,00	2,58	-72,42
2	8,25	1	0	-4,8	0,0	4,8	4,8	1,25	-70,00	-70,00	2,36	-67,64
3	8,50	1	0	-4,5	0,0	4,5	4,5	2,50	-65,00	-65,00	3,19	-61,81
4	8,75	1	0	-4,3	0,0	4,3	4,3	3,75	-60,00	-60,00	-0,15	-60,15
5	9,00	1	0	-4,0	0,0	4,0	4,0	5,00	-55,00	-55,00	-0,86	-55,86

Regression statistics	Y
Mean (observed)	14,65986
Standard deviation (observed)	39,76769
Mean (predicted)	14,65986
Standard deviation (predicted)	39,69857
Mean (residual)	-0,00000
Standard deviation (residual)	2,34361
R-square	0,99653
R-square adjusted	0,99613



Model specifications	Value
Independents	1
Dependents	1
Number of terms	4
Number of basis functions	3
Order of interactions	1
Penalty	2,000000
Threshold	0,000500
GCV error	8,077343
Prune	Yes

Coefficients, knots and basis functions	Coefficients Y	Knots x
Intercept	40,1731	
Term.1	7,9790	13,75000
Term.2	-19,8361	13,75000
Term.3	-5,7884	12,25000
Term.4	2,9895	15,25000



x:f1: y = -40 + 5*x; r = 1,0000; p = ---; r^2 = 1,0000
x:f2: y = -235 + 20*x; r = 1,0000; p = ---; r^2 = 1,0000

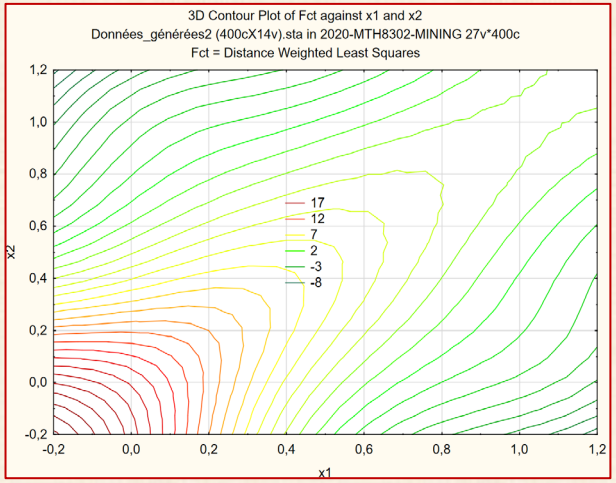
x:Y: y = -130,9893 + 10,3114*x; r = 0,9449; p = 0.0000; r^2 = 0,89

Modèle régression usuel : Y = -130,99 + 10,31*x r^2 = 0,89

Modèle MARS : Y = 40,17 + 7,98*max(0; x-13,75)
- 19,84*max(0; 13,75-x)
- 5,799*max(0; x-12,25)
+ 2,989*max(0; x-15,25) r^2 = 0,99

RÉGRESSION MARS

Ex3 : données simulées 2

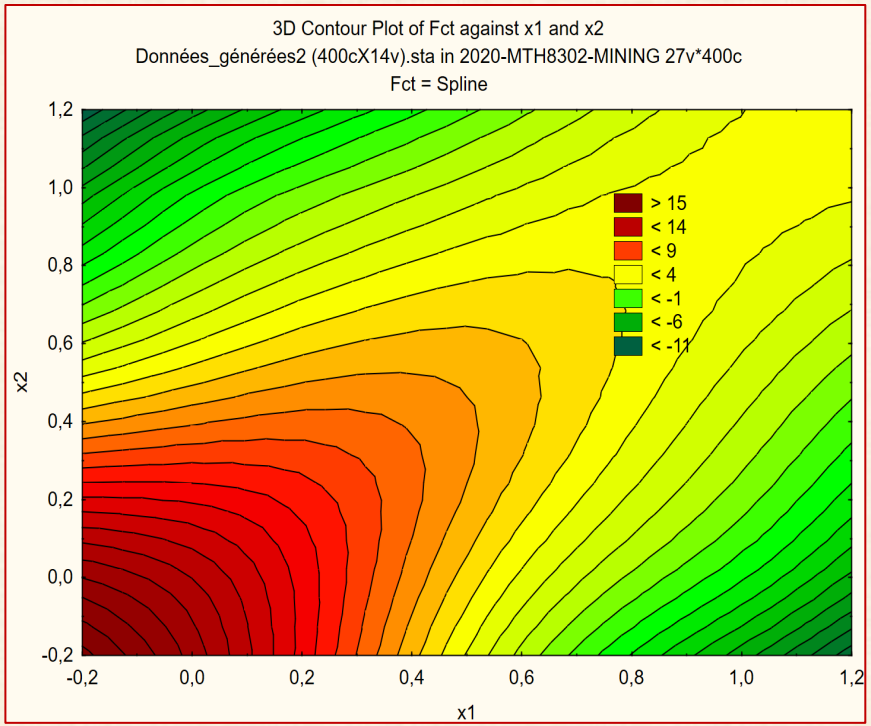


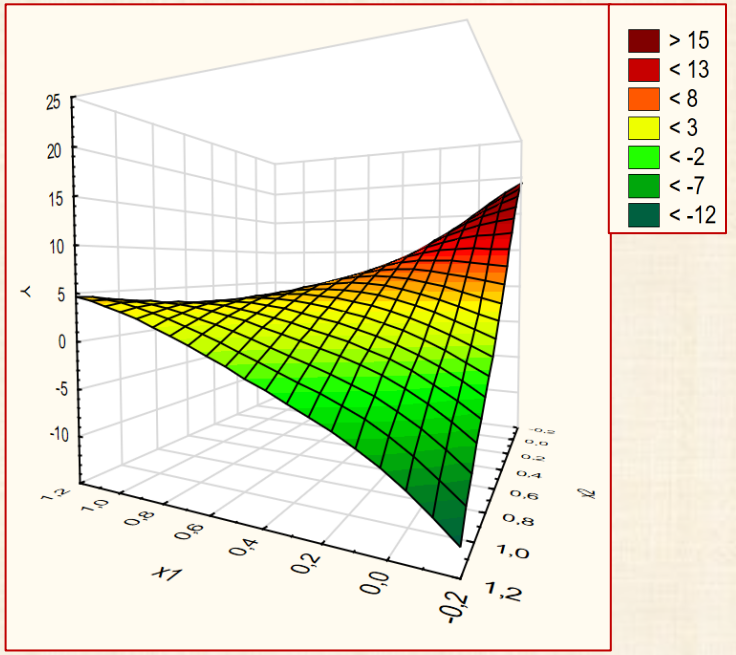
Exemple provenant de SAS

Example 24.1 Surface Fitting with Many Noisy Variables

Consider a simulated data set that contains a response variable and 10 continuous predictors.
Each continuous predictor is sampled independently from the uniform distribution U(0,1)
Y = Fct + E = F + N(0,1)
Fct = A / B A = 40*exp(8*((X1-0,5)**2 + (X2-0,5)**2))
B = (exp(8*((X1-0,2)**2+(X2-0,7)**2)) + exp(8*((X1-0,7)**2 + (X2-0,2)**2)))
n = 400 generated observations p = 10 variables X1, X2,..., X10
x3 x4 ... x10 n'interviennenent pas dans Fct - idem pour les variables u1 u2 ...u10
seulement X1 et X2 sont actives et définissent la fonction fct ci-haut
toutes proviennent d'unéchantillon de la distribution uniforme U(0,1) sur l'intervalle (0,1)
la valeur finale de Y est perturbée en ajoutant un bruit E distribué N(0,1)
/

1 ID	2 x1	3 x2	4 x3	5 x4	6 x5	7 x6	8 x7	9 x8	10 x9	11 x10	12 A	13 B	14 Fct	15 E	16 Y	17 info	18 u1	19 u2	20 u3
1	0,97	0,41	0,57	0,48	0,02	0,64	0,54	0,03	0,30	0,07	247,42	223,06	1,11	0,80	1,91	ajout 10 variables	1,15	0,44	0,29
2	0,89	0,67	0,98	0,29	0,26	0,83	0,38	0,28	0,12	0,07	170,42	53,06	3,21	0,29	3,50	de bruit	-0,16	-0,46	0,46
3	0,34	0,59	0,87	0,88	0,46	0,91	0,99	0,02	0,00	0,08	52,05	10,59	4,92	1,01	5,92	u1 u2 ... u20	-1,52	-0,00	0,28
4	1,00	0,19	0,14	0,57	0,31	0,60	0,16	0,69	0,55	0,32	646,76	1374,47	0,47	-0,37	0,10	chaque u	-0,32	-0,19	-0,39
5	0,77	0,42	0,92	0,05	0,19	0,82	0,64	0,79	0,98	0,65	75,61	27,02	2,80	1,12	3,92	distribuée	0,19	-0,39	-0,48
6	0,82	0,43	0,50	0,20	0,84	0,92	0,65	0,77	0,70	0,00	92,80	39,06	2,38	-1,03	1,34	Norm (0,1)	1,46	-0,24	0,22
7	0,85	0,84	0,51	0,64	0,59	0,56	0,38	0,65	0,94	0,06	266,97	65,60	4,07	0,78	4,85	ou	1,25	-0,43	-0,05
8	0,44	0,47	0,77	0,43	0,62	0,92	0,04	0,58	0,67	0,32	41,27	5,50	7,51	-0,58	6,93	Unif (0,1) - 0,5	-0,91	0,14	-0,36
9	0,67	0,36	0,72	0,45	0,95	0,44	0,11	0,08	0,71	0,62	59,31	16,23	3,65	1,34	4,99		-0,67	-0,18	-0,37
10	0,44	0,68	0,29	0,71	0,43	0,21	0,45	0,30	0,79	0,62	53,48	12,54	4,27	0,05	4,32	moyenne = 0 pour N et U	-0,29	-0,33	0,38
11	0,51	0,19	0,86	0,88	0,59	0,49	0,67	0,37	0,83	0,87	86,88	18,79	4,62	1,39	6,01	ecart-type = 1 pour N	-2,34	-0,18	0,45
12	0,95	0,49	0,33	0,12	0,67	0,95	0,67	0,26	0,47	0,23	202,55	130,61	1,55	-0,26	1,29	ecart-type = 0,29 pour U	1,02	-0,30	0,48
13	0,22	0,76	0,33	0,61	0,65	0,47	0,76	0,23	0,90	0,74	125,55	75,14	1,67	-0,00	1,67		2,45	0,26	-0,33





Ex3 : données simulées-2 multidimensionnelles

Exemple provenant de SAS

Example 24.1 Surface Fitting with Many Noisy Variables

Consider a simulated data set that contains a response variable and 10 continuous predictors.
Each continuous predictor is sampled independently from the uniform distribution U(0,1)
Y = Fct + E = F + N(0,1)
Fct = A / B A = 40*exp(8*((X1-0,5)**2 + (X2-0,5)**2))
B = (exp(8*((X1-0,2)**2+(X2-0,7)**2)) + exp(8*((X1-0,7)**2 + (X2-0,2)**2)))
n = 400 generated observations p = 10 variables X1, X2,..., X10
x3 x4 ... x10 n'interviennenent pas dans Fct - idem pour les variables u1 u2 ...u10
seulement X1 et X2 sont actives et définissent la fonction fct ci-haut
toutes proviennent d'unéchantillon de la distribution uniforme U(0,1) sur l'intervalle (0,1)
la valeur finale de Y est perturbée en ajoutant un bruit E distribué N(0,1)
/

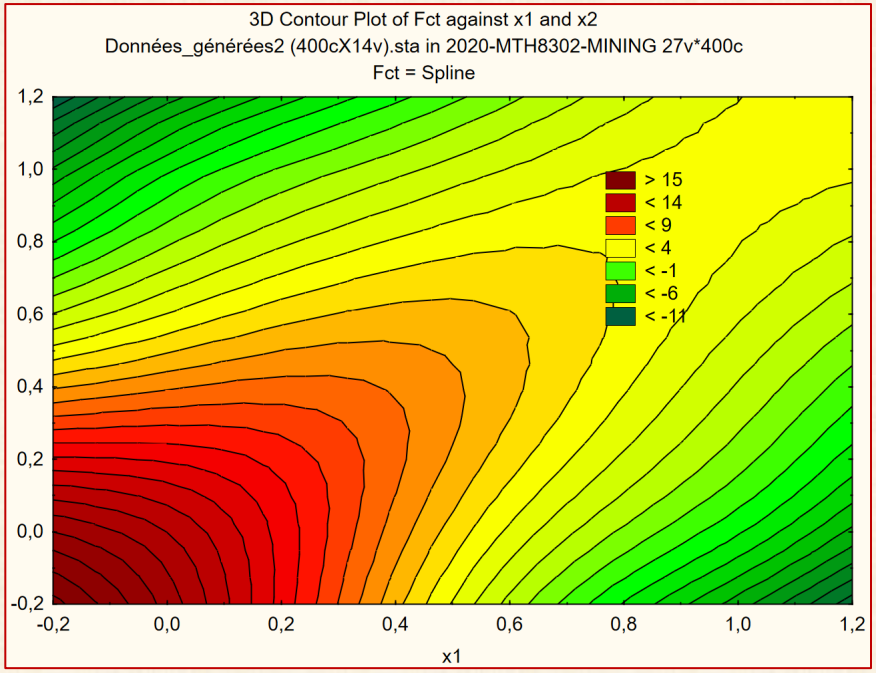
1 ID	2 x1	3 x2	4 x3	5 x4	6 x5	7 x6	8 x7	9 x8	10 x9	11 x10	12 A	13 B	14 Fct	15 E	16 Y	17 info	18 u1	19 u2	20 u3
1	0,97	0,41	0,57	0,48	0,02	0,64	0,54	0,03	0,30	0,07	247,42	223,06	1,11	0,80	1,91	ajout 10 variables	1,15	0,44	0,29
2	0,89	0,67	0,98	0,29	0,26	0,83	0,38	0,28	0,12	0,07	170,42	53,06	3,21	0,29	3,50	de bruit	-0,16	-0,46	0,46
3	0,34	0,59	0,87	0,88	0,46	0,91	0,99	0,02	0,00	0,08	52,05	10,59	4,92	1,01	5,92	u1 u2 ... u20	-1,52	-0,00	0,28
4	1,00	0,19	0,14	0,57	0,31	0,60	0,16	0,69	0,55	0,32	646,76	1374,47	0,47	-0,37	0,10	chaque u	-0,32	-0,19	-0,39
5	0,77	0,42	0,92	0,05	0,19	0,82	0,64	0,79	0,98	0,65	75,61	27,02	2,80	1,12	3,92	distribuée	0,19	-0,39	-0,48
6	0,82	0,43	0,50	0,20	0,84	0,92	0,65	0,77	0,70	0,00	92,80	39,06	2,38	-1,03	1,34	Norm (0,1)	1,46	-0,24	0,22
7	0,85	0,84	0,51	0,64	0,59	0,56	0,38	0,65	0,94	0,06	266,97	65,60	4,07	0,78	4,85	ou	1,25	-0,43	-0,05
8	0,44	0,47	0,77	0,43	0,62	0,92	0,04	0,58	0,67	0,32	41,27	5,50	7,51	-0,58	6,93	Unif (0,1) - 0,5	-0,91	0,14	-0,36
9	0,67	0,36	0,72	0,45	0,95	0,44	0,11	0,08	0,71	0,62	59,31	16,23	3,65	1,34	4,99		-0,67	-0,18	-0,37
10	0,44	0,68	0,29	0,71	0,43	0,21	0,45	0,30	0,79	0,62	53,48	12,54	4,27	0,05	4,32	moyenne = 0 pour N et U	-0,29	-0,33	0,38
11	0,51	0,19	0,86	0,88	0,59	0,49	0,67	0,37	0,83	0,87	86,88	18,79	4,62	1,39	6,01	ecart-type = 1 pour N	-2,34	-0,18	0,45
12	0,95	0,49	0,33	0,12	0,67	0,95	0,67	0,26	0,47	0,23	202,55	130,61	1,55	-0,26	1,29	ecart-type = 0,29 pour U	1,02	-0,30	0,48
13	0,22	0,76	0,33	0,61	0,65	0,47	0,76	0,23	0,90	0,74	125,55	75,14	1,67	-0,00	1,67		2,45	0,26	-0,33

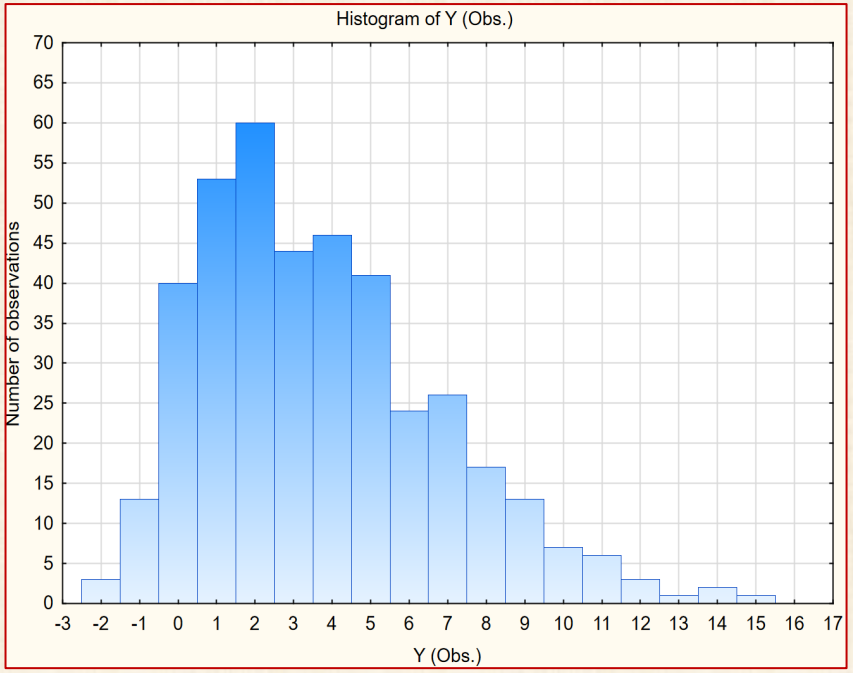
$$Y = F + E = F + N(0,1)$$

$$F = A / B$$

$$A = 40*\exp(8*((X1-0,5)**2 + (X2-0,5)**2))$$

$$B = \exp(8*((X1-0,2)**2 + (X2-0,7)**2) + \exp(8*((X1-0,7)**2 + (X2-0,2)**2)))$$





Ex3 : données simulées-2 multidimensionnelles

Analyse classique : Response Surface Methodology (modèle = polynôme second degré)

effect	Degr. of Freedom	Y SS	Y MS	Y F	Y p
"x1"*"x2"	1	1484,273	1484,273	598,6059	0,000000
Intercept	1	123,768	123,768	49,9154	0,000000
"x1"	1	99,447	99,447	40,1067	0,000000
"x2"^2	1	94,918	94,918	38,2804	0,000000
"x2"	1	88,310	88,310	35,6154	0,000000
"x1"^2	1	80,242	80,242	32,3615	0,000000
"x3"*"x4"	1	41,583	41,583	16,7702	0,000053
"x7"	1	10,182	10,182	4,1063	0,043519
"x4"*"x6"	1	9,468	9,468	3,8183	0,051530
"x1"*"x10"	1	8,734	8,734	3,5225	0,061413
"x1"*"x7"	1	8,445	8,445	3,4058	0,065854
"x7"*"x10"	1	7,186	7,186	2,8982	0,089606
"x6"	1	6,716	6,716	2,7087	0,100742
"x5"*"x6"	1	6,043	6,043	2,4371	0,119446
"x4"*"x10"	1	6,010	6,010	2,4237	0,120458
"x8"*"x9"	1	5,075	5,075	2,0466	0,153485
"x4"*"x8"	1	5,040	5,040	2,0326	0,154895
"x3"*"x10"	1	4,871	4,871	1,9643	0,161979
"x4"^2	1	4,645	4,645	1,8732	0,172026
"x10"^2	1	4,427	4,427	1,7854	0,182398
"x2"*"x6"	1	4,323	4,323	1,7433	0,187623
"x10"	1	4,234	4,234	1,7076	0,192194
"x5"*"x10"	1	3,953	3,953	1,5943	0,207590
"x2"*"x5"	1	3,697	3,697	1,4908	0,222949
"x7"^2	1	2,857	2,857	1,1524	0,283829
"x2"*"x7"	1	2,558	2,558	1,0316	0,310511
"x7"*"x8"	1	2,241	2,241	0,9036	0,342497
"x6"*"x10"	1	2,134	2,134	0,8606	0,354251
"x2"*"x9"	1	2,052	2,052	0,8278	0,363577
"x5"^2	1	1,881	1,881	0,7588	0,384339
"x6"*"x7"	1	1,674	1,674	0,6750	0,411892
"x6"*"x8"	1	1,533	1,533	0,6185	0,432179
"x9"^2	1	1,479	1,479	0,5963	0,440547
"x3"*"x5"	1	1,426	1,426	0,5752	0,448740
"x9"*"x10"	1	1,288	1,288	0,5193	0,471664

Dependent Variable	Multiple R	Multiple R^2	Adjusted R^2
Y	0,897206	0,804979	0,767026

suite

"x3"*"x7"	1	1,247	1,247	0,5027	0,478800
"x3"*"x6"	1	1,178	1,178	0,4751	0,491131
"x2"*"x10"	1	1,126	1,126	0,4542	0,500801
"x8"*"x10"	1	0,954	0,954	0,3849	0,535438
"x3"^2	1	0,921	0,921	0,3713	0,542714
"x3"*"x8"	1	0,816	0,816	0,3293	0,566480
"x7"*"x9"	1	0,698	0,698	0,2813	0,596187
"x5"	1	0,659	0,659	0,2659	0,606440
"x2"*"x4"	1	0,644	0,644	0,2596	0,610698
"x1"*"x4"	1	0,628	0,628	0,2535	0,614986
"x6"^2	1	0,584	0,584	0,2355	0,627780
"x6"*"x9"	1	0,567	0,567	0,2286	0,632902
"x4"	1	0,499	0,499	0,2014	0,653884
"x2"*"x8"	1	0,442	0,442	0,1783	0,673136
"x5"*"x8"	1	0,434	0,434	0,1749	0,676097
"x8"^2	1	0,405	0,405	0,1633	0,686408
"x1"*"x3"	1	0,323	0,323	0,1301	0,718526
"x8"	1	0,306	0,306	0,1233	0,725662
"x3"	1	0,258	0,258	0,1042	0,746992
"x5"*"x7"	1	0,162	0,162	0,0655	0,798219
"x3"*"x9"	1	0,150	0,150	0,0605	0,805897
"x1"*"x5"	1	0,137	0,137	0,0554	0,814104
"x4"*"x5"	1	0,134	0,134	0,0542	0,815990
"x2"*"x3"	1	0,112	0,112	0,0452	0,831842
"x4"*"x9"	1	0,104	0,104	0,0418	0,838036
"x1"*"x6"	1	0,100	0,100	0,0403	0,841000
"x4"*"x7"	1	0,091	0,091	0,0367	0,848135
"x5"*"x9"	1	0,088	0,088	0,0355	0,850655
"x1"*"x9"	1	0,017	0,017	0,0068	0,934118
"x9"	1	0,006	0,006	0,0025	0,960439
"x1"*"x8"	1	0,001	0,001	0,0004	0,983409
Error	334	828,169	2,480		
Total	399	4246,568			

Ex3 : données simulées-2 multidimensionnelles

Analyse MARS - Modèle 1 : sans inter

Model specifications	Value
Independents	10
Dependents	1
Number of terms	8
Number of basis functions	7
Order of interactions	1
Penalty	2,000000
Threshold	0,000500
GCV error	6,410808
Prune	Yes

Coefficients, knots and basis functions	Coefficients Y	Knots x1	Knots x2	Knots x3	Knots x4	Knots x5	Knots x6	Knots x7	Knots x8	Knots x9	Knots x10
Intercept	8,2722										
Term.1	-6,8740	0,247									
Term.2	-11,1194	0,247									
Term.3	-6,4876		0,329								
Term.4	-6,0680		0,329								
Term.5	8,7056						0,167				
Term.6	-2,5229							0,405			
Term.7	10,5187								0,864		

Regression statistics	Y
Mean (observed)	4,182
Standard deviation (observed)	3,113
Mean (predicted)	4,182
Standard deviation (predicted)	1,933
Mean (residual)	0,000
Standard deviation (residual)	2,440
R-square	0,386
R-square adjusted	0,373

Modèle 1 : sans terme d'interaction - Rsquare ajusted = 0,373 … faible !

Y = 8,27 - 6,87*max(0; x1-0,247) - 11,11*max(0; 0,247-x1) - 6,49*max(0; x2- 0,329)
- 6,068*max(0; 0,329-x2) + 8,705*max(0; 0,167-x6)-2,52*max(0; 0,404-x7)
+ 10,52*max(0; x8-0,864)

Analyse MARS - Modèle 2 : avec inter

Model specifications	Value
Independents	10
Dependents	1
Number of terms	13
Number of basis functions	19
Order of interactions	2
Penalty	2,000000
Threshold	0,000500
GCV error	1,552321
Prune	Yes

Coefficients, knots and basis functions	Coefficients Y	Knots x1	Knots x2	Knots x3	Knots x4	Knots x5	Knots x6	Knots x7	Knots x8	Knots x9	Knots x10
Intercept	10,03										
Term.1	-10,61	0,19									
Term.2	-49,23	0,19									
Term.3	110,40	0,19	0,54								
Term.4	-18,75		0,33								
Term.5	16,33		0,33								
Term.6	-8,83	0,19	0,61								
Term.7	36,80	0,19	0,29								
Term.8	97,51	0,19	0,32								
Term.9	55,53	0,67	0,33								
Term.10	-73,17	0,06	0,33								
Term.11	-8,13	0,63									
Term.12	78,40	0,47	0,33								

Regression statistics	Y
Mean (observed)	4,311
Standard deviation (observed)	3,262
Mean (predicted)	4,311
Standard deviation (predicted)	3,046
Mean (residual)	0,000
Standard deviation (residual)	1,170
R-square	0,871
R-square adjusted	0,867

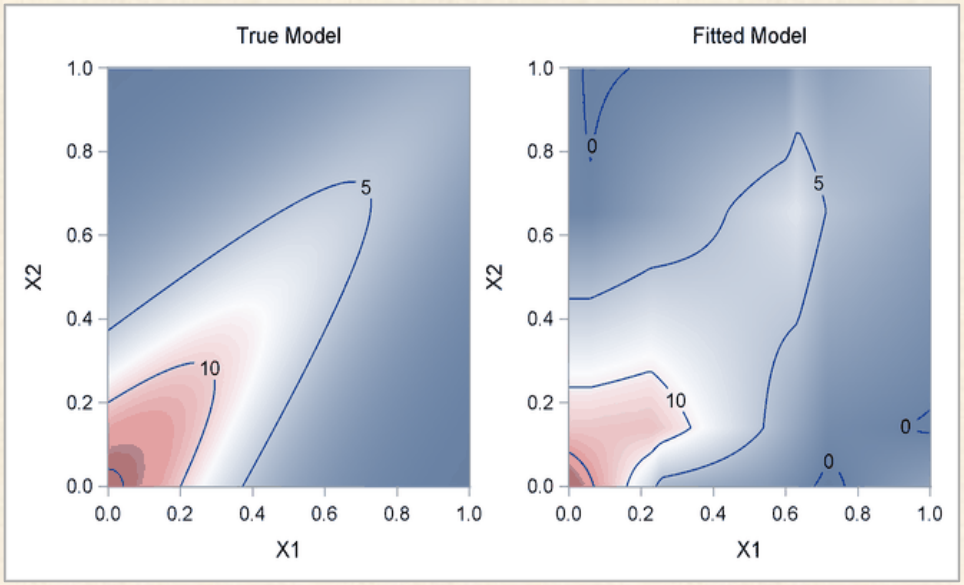
Modèle 2 : avec interaction - Rsquare ajusted = 0,867 … seulement x1 et x2 sont identifiées

Y = 10,03 - 10,61*max(0; x1-0,19) - 49,23*max(0; 0,192-x1) + 110,39*max(0; 0,19-x1)*max(0; 0,537-x2)
- 18,746*max(0; x2-0,329) + 16,33*max(0; 0,3-x2) - 8,83*max(0; x1-0,19)*max(0; x2-0,606)
+ 36,80*max(0; x1-0,19)*max(0; x2-0,29) + 97,507*max(0; 0,19-x1)*max(0; x2-0,32)
+ 55,53*max(0; x1-0,67)*max(0; 0,329-x2) - 73,167*max(0; x1-0,061)*max(0; 0,329-x2)
- 8,13*max(0; x1-0,629) + 78,398*max(0; x1-0,4715)*max(0; 0,329-x2)

Ex3 : données simulées-2 multidimensionnelles

Analyse avec SAS : résultat légèrement différent mais avec plus de fonctions de base

Fit Statistics	
GCV	1.55656
GCV R-Square	0.86166
Effective Degrees of Freedom	27
R-Square	0.87910
Adjusted R-Square	0.87503
Mean Square Error	1.40260
Average Square Error	1.35351



Regression Spline Model after Backward Selection

Name	Coefficient	Parent	Variable	Knot
Basis0	12.3031		Intercept	
Basis1	13.1804	Basis0	X1	0.05982
Basis3	-23.4892	Basis0	X2	0.1387
Basis4	-171.03	Basis0	X2	0.1387
Basis5	-86.1867	Basis3	X1	0.6333
Basis7	-436.86	Basis4	X1	0.5488
Basis8	397.18	Basis4	X1	0.5488
Basis9	11.4682	Basis1	X2	0.6755
Basis10	-19.1796	Basis1	X2	0.6755
Basis13	126.84	Basis11	X1	0.6018
Basis14	40.8134	Basis11	X1	0.6018
Basis15	22.2884	Basis0	X1	0.7170
Basis17	-53.8746	Basis12	X1	0.2269
Basis19	598.89	Basis4	X1	0.2558

RÉGRESSION MARS

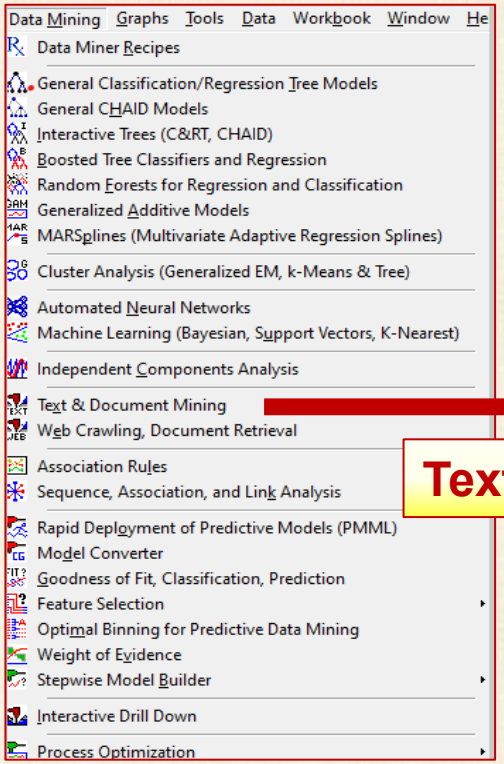
Ex4 : pourriels (spam)

This corpus has been collected from free or free for research sources at the Web:

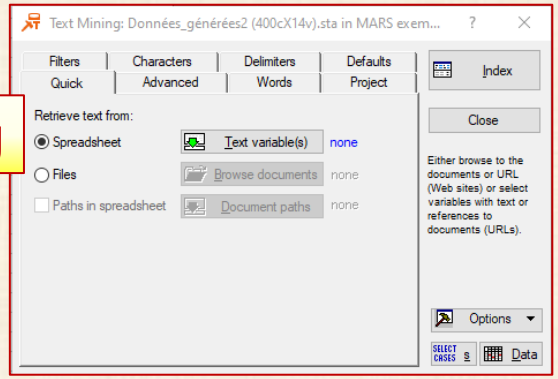
A collection of between 425 SMS spam messages extracted manually from the Grumbletext Web site.
This example concerns a study on classifying whether an e-mail is junk e-mail (coded as 1) or not (coded as 0).
The data were collected in Hewlett-Packard labs and donated by George Forman.
The data set contains 4,601 observations with 58 variables.
The response variable is a binary indicator of whether an e-mail is considered spam or not.
The 57 variables are continuous variables that record frequencies of some common words and characters in e-mails and lengths of uninterrupted sequences of capital letters.
The data set is publicly available at the UCI Machine Learning repository (Asuncion and Newman,
/

1 ID	2 SPAM?	3 MESSAGE
1	no	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	no	Ok lar... Joking wif u oni...
3	yes	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
4	no	U dun say so early hor... U c already then say...
5	no	Nah I don't think he goes to usf, he lives around here though
6	yes	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb

analyse mots clés



Text Mining



Address	Addresses	All	Bracket	Business
CS	CapAvg	CapLong	CapTotal	Conference
Credit	Data	Direct	Dollar	Edu
Email	Exclamation	Font	Free	George
HP	HPL	Internet	Lab	Labs
Mail	Make	Meeting	Money	Order
Original	Our	Over	PM	Paren
Parts	People	Pound	Project	RE
Receive	Remove	Report	Semicolon	Table
Technology	Telnet	Will	You	Your
_000	_85	_415	_650	_857

Variable Importance

Variable	Number of Bases	Importance
George	1	100.00
HP	1	78.35
Edu	3	61.25
Remove	2	49.21
Exclamation	3	44.14
Free	2	34.18
Meeting	3	32.57
_1999	2	29.71
Dollar	2	28.30
Money	3	26.39
CapLong	3	24.41
Our	2	19.46
Semicolon	2	14.98
RE	2	13.52
Business	3	13.48
Over	3	12.63
CapTotal	3	12.50
Will	1	10.81
Pound	2	9.73
Internet	1	5.88
_000	1	4.57
You	2	3.17

Ex4 : predicting E-Mail Spam

choix de certains mots (text mining)

000 _85 _415 _650
_857 _1999 _3d address
addresses all bang bracket
business cap_avg cap_long
cap_total conference credit cs
data direct dollar edu email
font free george hp hpl
internet lab labs mail
make meeting money order
original our over paren parts
people pm pound project re
receive remove report semicol
table technology telnet will
you your

Variable Importance

Variable	Number of Bases	Importance
George	1	100.00
Hp	1	78.35
Edu	3	61.25
Remove	2	49.21
Bang	3	44.14
Free	2	34.18
Meeting	3	32.57
_1999	2	29.71
Dollar	2	28.30
Money	3	26.39
Cap_Long	3	24.41
Our	2	19.46
Semicol	2	14.98
Re	2	13.52

Table of Class by Error

Class	Error 0	Error 1	Total
0	885	59	944
	56.26	3.75	60.01
	93.75	6.25	
1	592	37	629
	37.64	2.35	39.99
	94.12	5.88	
Total	1477	96	1573
	93.90	6.10	100.00

Frequency
Percent
Row Pct

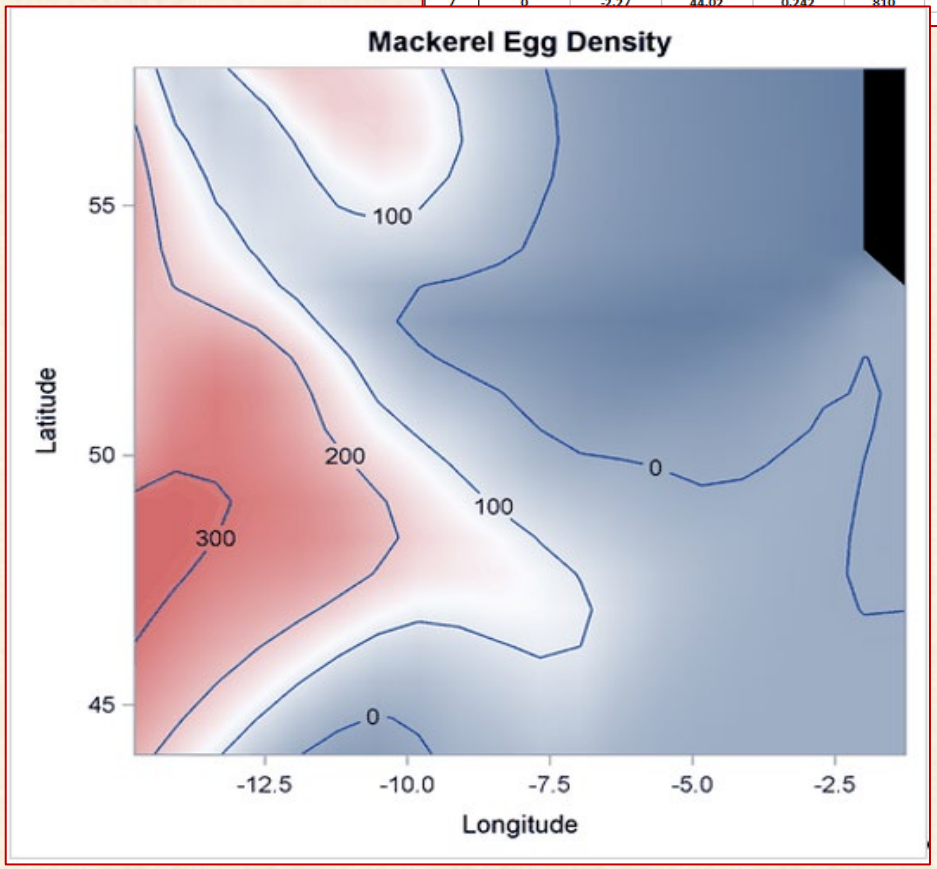
RÉGRESSION MARS

Ex5 : densité Œufs poisson maquereau

The example concerns a study of mackerel egg density

The example concerns a study of mackerel egg density
The data are a subset of the 1992 mackerel egg survey conducted over the Porcupine Bank west of Ireland.
The survey took place in the peak spawning area. Scientists took samples by hauling a net up from deep sea to the sea surface.
Then they counted the number of spawned mackerel eggs and used other geographic information
to estimate the sizes and distributions of spawning stocks.
The data set is used as an example in Bowman and Azzalini (1997)
This data set contains 634 observations and 7 variables.
response variable Egg_Count is the number of mackerel eggs collected from each sampling net.
Longitude and Latitude are the location values in degrees east and north, respectively, of each sample station.
Net_Area is the area of the sampling net in square meters.
Depth records the sea bed depth in meters at the sampling location.
Distance is the distance in geographic degrees from the sample location to the continental shelf edge.

1 ID	2 EggCount	3 longitude	4 latitude	5 Net_area	6 depth	7 distance	8 Y_density
1	0	-4.65	44.57	0.242	4342	0,83951	0,00000
2	0	-4.48	44.57	0.242	4334	0,85919	0,00000
3	0	-4.3	44.57	0.242	4286	0,89302	0,00000
4	1	-2.87	44.02	0.242	1438	0,39564	0,00980
5	4	-2.07	44.02	0.242	166	0,04001	0,03922
6	3	-2.13	44.02	0.242	460	0,09742	0,02941
7	0	-2.27	44.02	0.242	810	0,23626	0,00000



Source : http://support.sas.com/documentation/cdl/en/statug/65328/HTML/default/viewer.htm#statug_adaptivereg_examples01.htm

Ex5 Nonparametric Poisson Model for Mackerel Egg Density

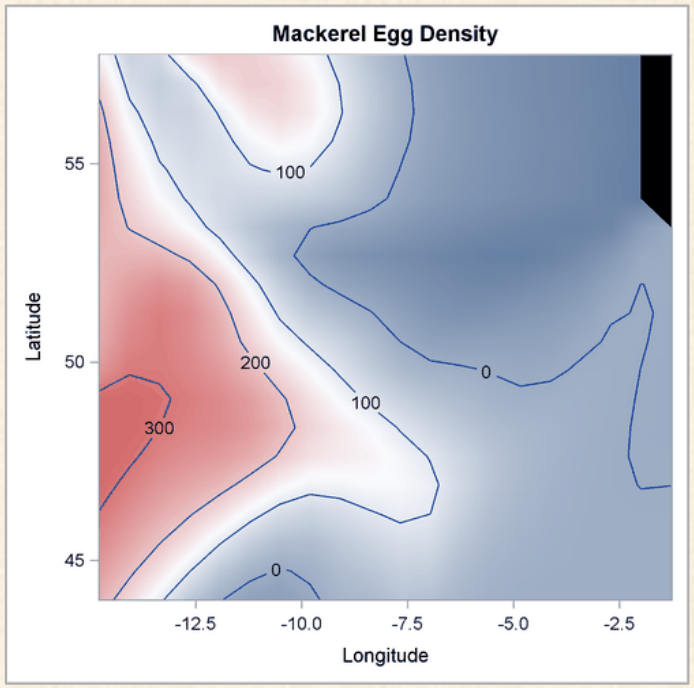
e example concerns a study of mackerel egg density

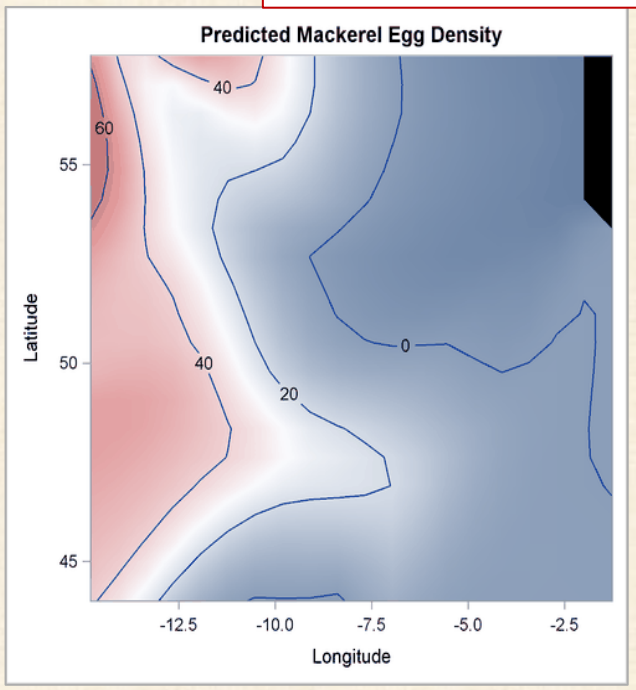
e example concerns a study of mackerel egg density
e data are a subset of the 1992 mackerel egg survey conducted over the Porcupine Bank west of Ireland.
e survey took place in the peak spawning area. Scientists took samples by hauling a net up from deep sea to the sea surface.
en they counted the number of spawned mackerel eggs and used other geographic information
estimate the sizes and distributions of spawning stocks.
e data set is used as an example in Bowman and Azzalini (1997)
is data set contains 634 observations and 5 variables.
sponse variable Egg_Count is the number of mackerel eggs collected from each sampling net.
ngitude and Latitude are the location values in degrees east and north, respectively, of each sample station.
t_Area is the area of the sampling net in square meters.
pth records the sea bed depth in meters at the sampling location.
stance is the distance in geographic degrees from the sample location to the continental shelf edge.

1 ID	2 EggCount	3 longitude	4 latitude	5 Net_area	6 depth	7 distance	8 density
1	0	-4.65	44.57	0.242	4342	0,83951	0,00000
2	0	-4.48	44.57	0.242	4334	0,85919	0,00000
3	0	-4.3	44.57	0.242	4286	0,89302	0,00000
4	1	-2.87	44.02	0.242	1438	0,39564	0,00980
5	4	-2.07	44.02	0.242	166	0,04001	0,03922
6	3	-2.13	44.02	0.242	460	0,09742	0,02941
7	0	-2.27	44.02	0.242	810	0,23626	0,00000
8							
9							
10							

Variable Importance

Variable	Number of Bases	Importance
Longitude	7	100.00
Depth	8	30.26
Latitude	5	18.93
Distance	3	8.56



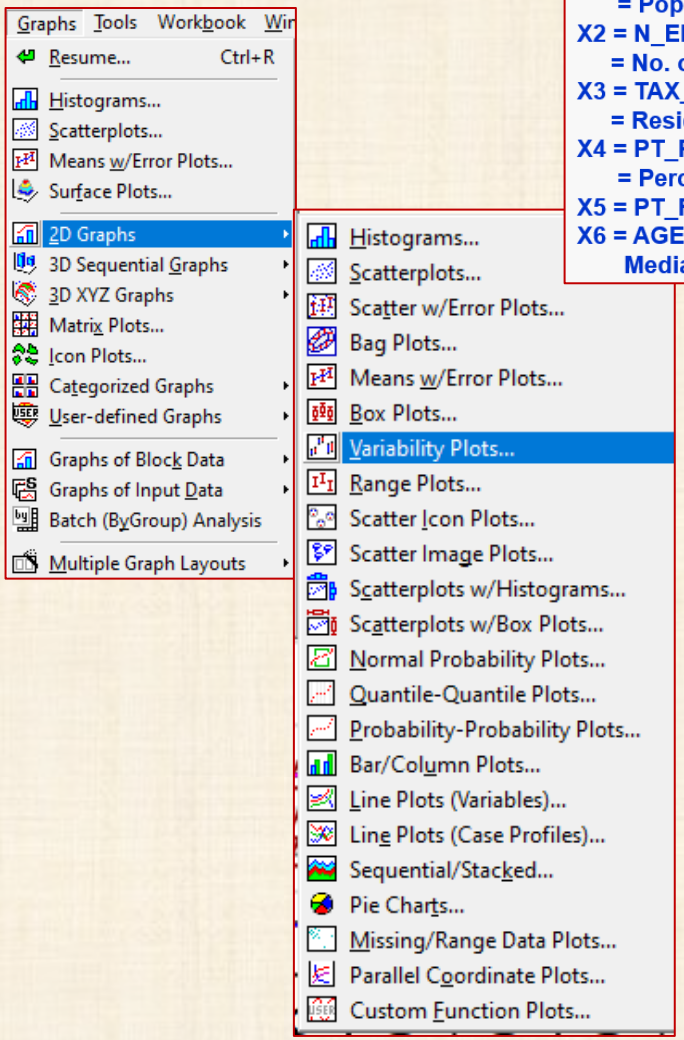


RÉGRESSION MARS

Ex6 - pauvreté

VARIABLES

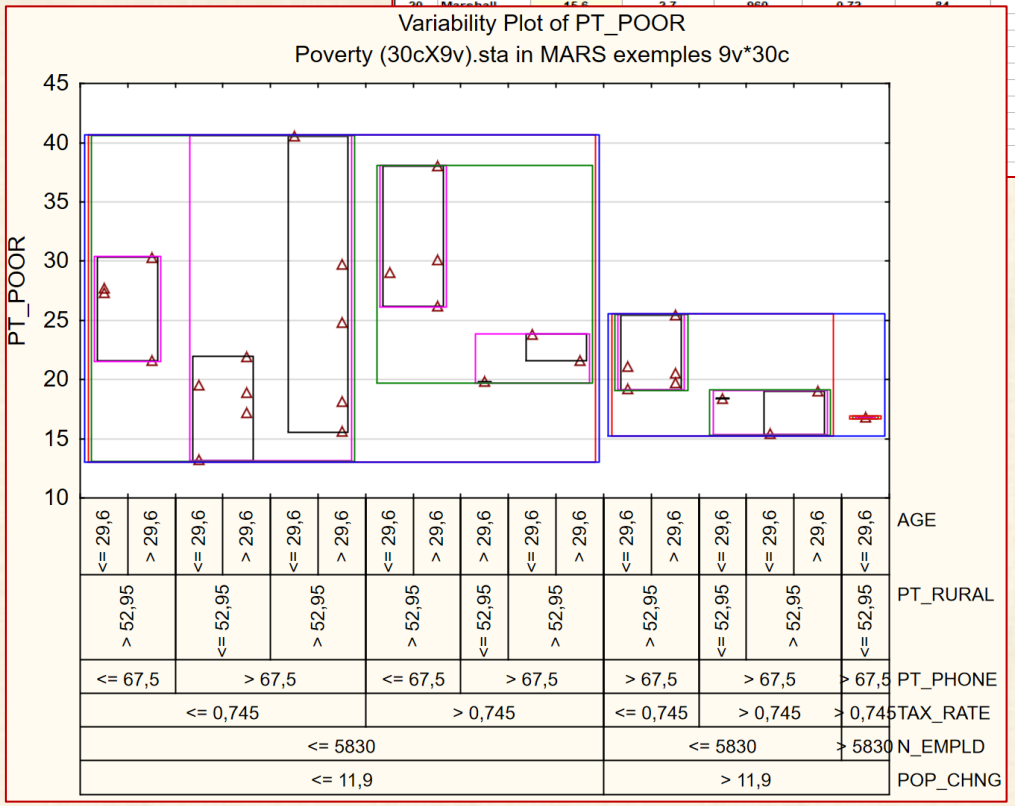
X0 = County
Y = PT_POOR
= Percent of families below poverty level
X1 = POP_CHANGE
= Population change (1960-1970)
X2 = N_EMPLD
= No. of persons employed in agriculture
X3 = TAX_RATE
= Residential and farm property tax rate
X4 = PT_PHONE
= Percent residences with telephones
X5 = PT_RURAL = Percent rural population
X6 = AGE
Median age



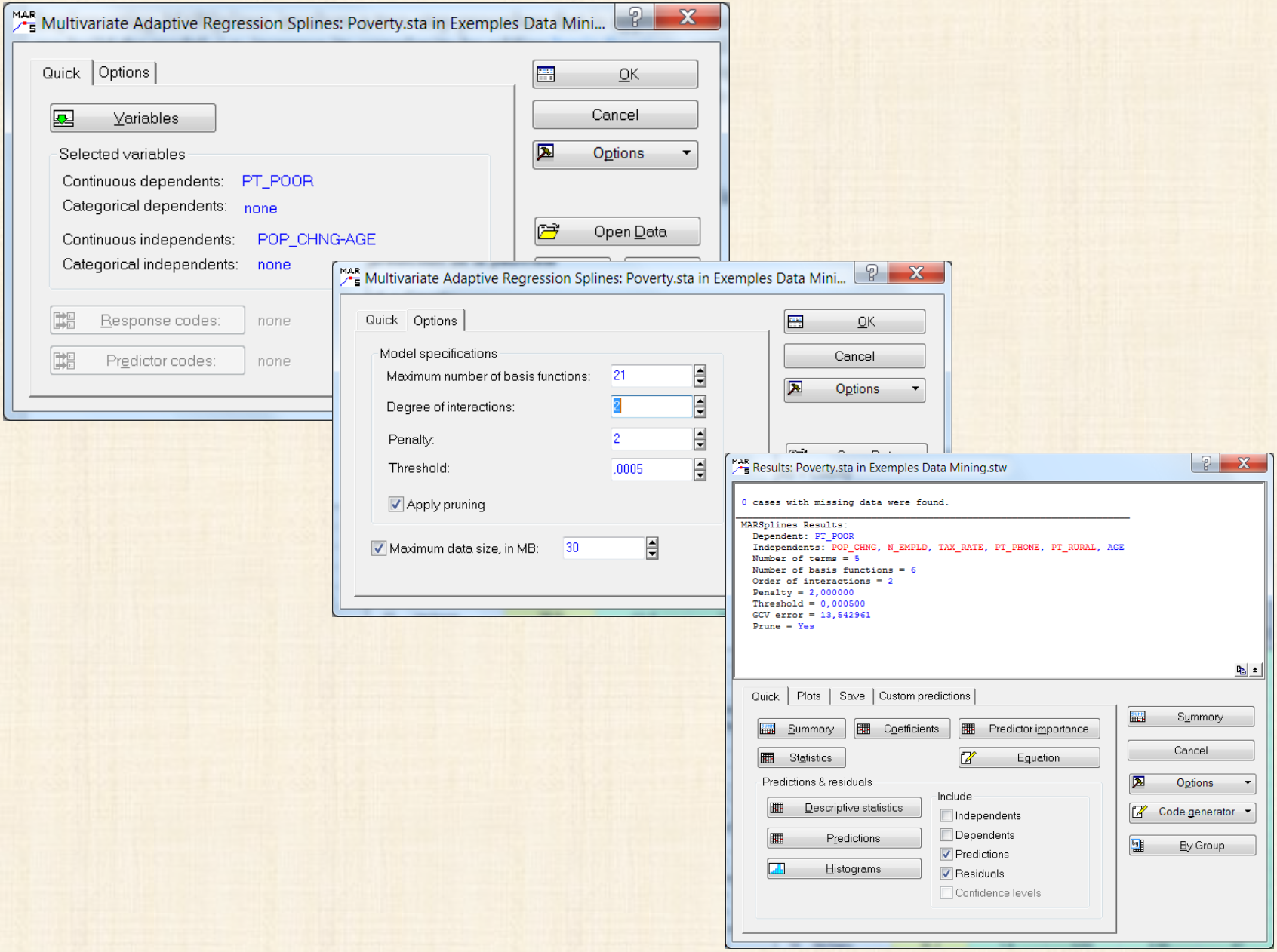
prédiction de la pauvreté

X0 = County
Y = PT_POOR = Percent of families below poverty level
X1 = POP_CHANGE = Population change (1960-1970)
X2 = N_EMPLD = No. of persons employed in agriculture
X3 = TAX_RATE = Residential and farm property tax rate
X4 = PT_PHONE = Percent residences with telephones
X5 = PT_RURAL = Percent rural population
X6 = AGE = Median age
/

1 ID	2 County	3 PT_POOR	4 POP_CHNG	5 N_EMPLD	6 TAX_RATE	7 PT_PHONE	8 PT_RURAL	9 AGE
1	Benton	19,0	13,7	400	1,09	82	74,8	33,5
2	Cannon	26,2	-0,8	710	1,01	66	100,0	32,8
3	Carrol	18,1	9,6	1610	0,40	80	69,7	33,4
4	Cheatheam	15,4	40,0	500	0,93	74	100,0	27,8
5	Cumberland	29,0	8,4	640	0,92	65	74,0	27,9
6	DeKalb	21,6	3,5	920	0,59	64	73,1	33,2
7	Dyer	21,9	3,0	1890	0,63	82	52,3	30,8
8	Gibson	18,9	7,1	3040	0,49	85	49,6	32,4
9	Greene	21,1	13,0	2730	0,71	78	71,2	29,2
10	Hawkins	23,8	10,7	1850	0,93	74	70,6	28,7
11	Haywood	40,5	-16,2	2920	0,51	69	64,2	25,1
12	Henry	21,6	6,6	1070	0,80	85	58,3	35,9
13	Houston	25,4	21,9	160	0,74	69	100,0	31,4
14	Humphreys	19,7	17,8	380	0,44	83	72,0	30,1
15	Jackson	38,0	-11,8	1140	0,81	54	100,0	34,1
16	Johnson	30,1	7,5	690	1,05	65	100,0	30,5
17	Lawrence	24,8	3,7	1170	0,73	76	69,5	30,0
18	McNairy	30,3	1,6	1280	0,65	67	81,0	32,4
19	Madison	19,5	8,4	2270	0,48	85	39,1	28,7
							58,4	33,4
							42,4	29,9
							36,4	23,3
							99,8	27,5
							90,6	29,5
							5,9	25,4
							44,2	28,8
							100,0	33,1
							52,6	30,8
							100,0	28,4
							72,1	30,4



Ex6 – indicateur pauvreté



Ex6 – indicateur pauvreté

Model coefficients

NOTE: Highlighted cells indicate basis functions of type
max(0, independent-knot),
otherwise max(0, knot-independent)

	Coefficients PT_POOR	Knots POP CHNG	Knots - N_EMPLD	Knots - TAX RATE	Knots - PT PHONE	Knots - PT RURAL	Knots AGE
Intercept	20,2819						
Term.1	-0,25678	7,1000					
Term.2	0,00183		1070,0		75,00		
Term.3	1,07462			0,400	75,00		
Term.4	0,16003					71,20	

Model Summary

	Value
Independents	6
Dependents	1
Number of terms	5
Number of basis functions	6
Order of interactions	2
Penalty	2,0000
Threshold	0,00050
GCV error	13,5429
Prune	Yes

Ex6 – indicateur pauvreté

Regression statistics	
	PT_POOR
Mean (observed)	23,01000
Standard deviation (observed)	6,42658
Mean (predicted)	23,01000
Standard deviation (predicted)	5,86822
Mean (residual)	-0,00000
Standard deviation (residual)	2,62009
R-square	0,83378
R-square adjusted	0,79916

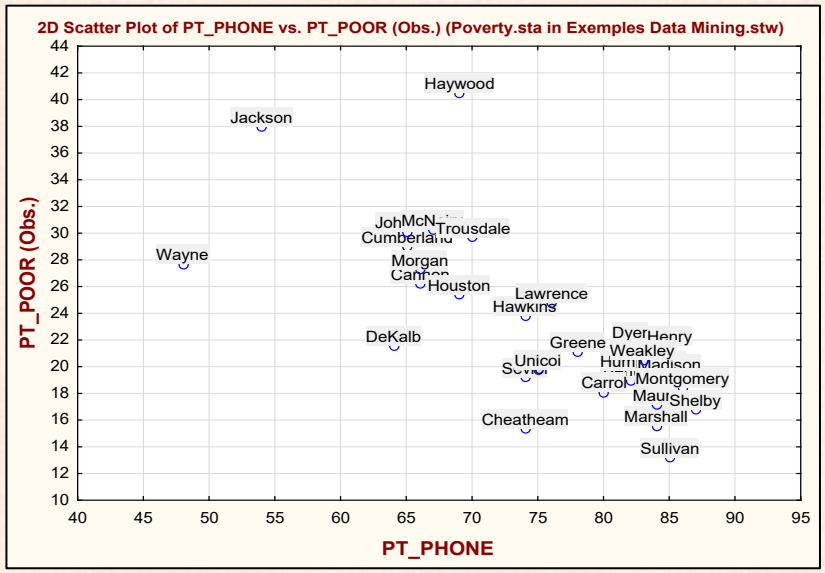
PT_POOR = 2,02819296499116e+001
- 2,56780388292518e-001*
max(0; POP_CHNG-7,10000000000000e+000)
+ 1,83090026675797e-003
*max(0; N_EMPLD-1,07000000000000e+003)
*max(0; 7,50000000000000e+001-PT_PHONE)
+ 1,07462185377171e+000
*max(0; TAX_RATE-4,00000000000000e-001)
*max(0; 7,50000000000000e+001-PT_PHONE)
+ 1,60029686783812e-001
*max(0; PT_RURAL-7,12000000000000e+001)

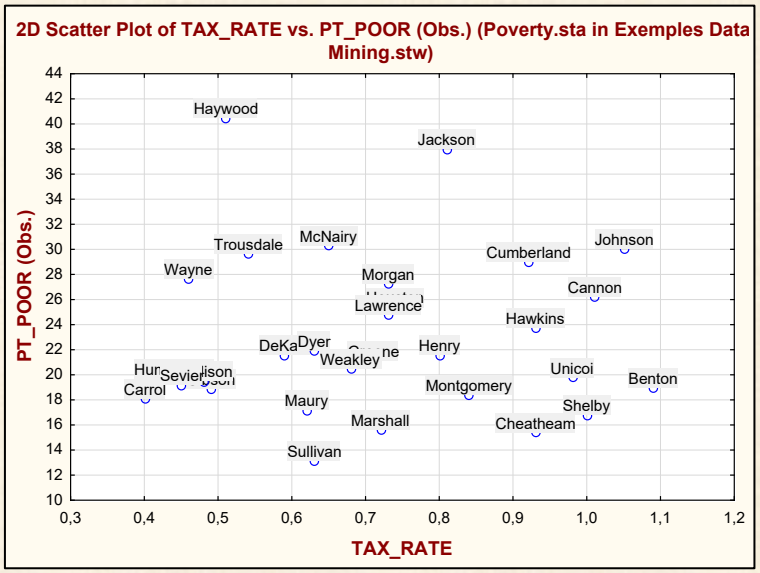
laborieux à lire

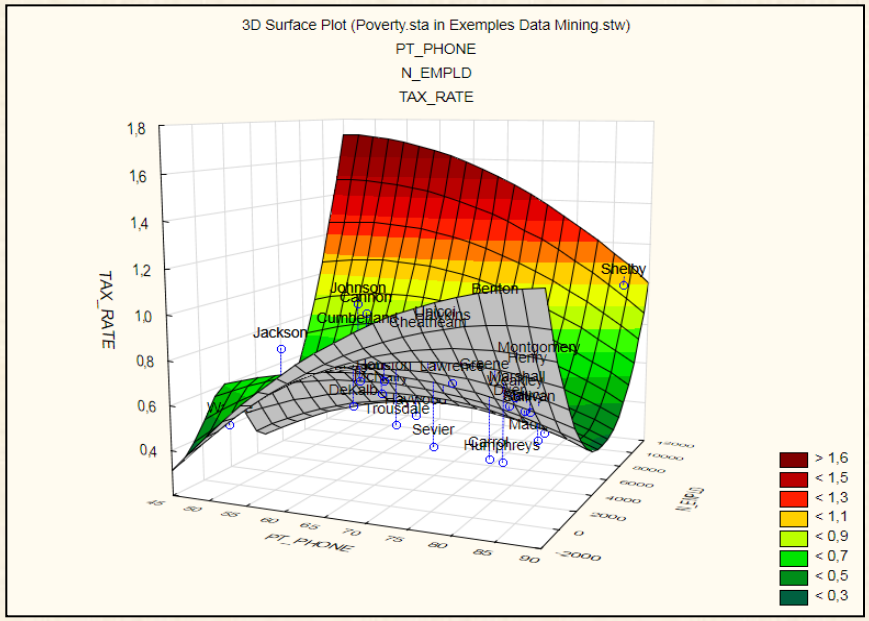
plus lisible

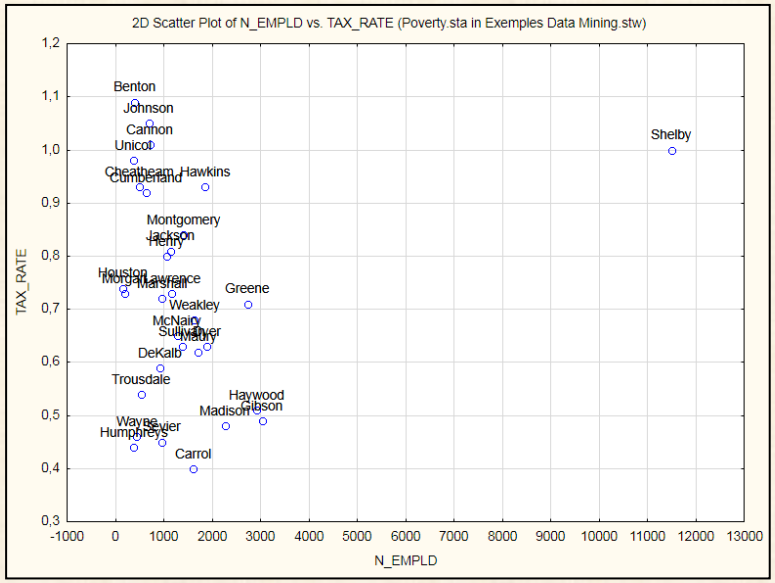
PT_POOR = 20,28
- 0,257 * max(0; POP_CHNG-7,1)
+ 0,0018 * max(0; N_EMPLD-1070) * max(0; 75-PT_PHONE)
+ 1,07 * max(0; TAX_RATE-0,4) * max(0; 75-PT_PHONE)
+ 0,16 * max(0; PT_RURAL-71,2)

Ex6 – indicateur pauvreté



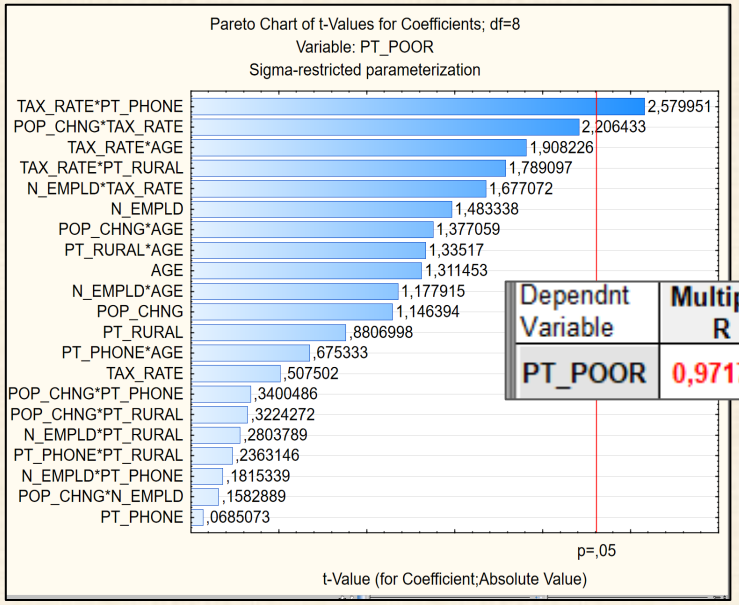






Ex6 – indicateur pauvreté

analyse régression classique et stepwise



Dependnt Variable	Multiple R	Multiple R^2	Adjusted R^2
PT_POOR	0,971703	0,944206	0,797748

Effect	PT_POOR Param.	PT_POOR Std.Err	PT_POOR t	PT_POOR p
TAX_RATE*PT_PHONE	-3,3548	1,3003	-2,57995	0,032619
POP_CHNG*TAX_RATE	2,7675	1,2543	2,20643	0,058405
TAX_RATE*AGE	7,4319	3,8947	1,90823	0,092785
TAX_RATE*PT_RURAL	-0,9368	0,5236	-1,78910	0,111393
N_EMPLD*TAX_RATE	0,0177	0,0105	1,67707	0,132051
N_EMPLD	-0,0711	0,0479	-1,48334	0,176274
POP_CHNG*AGE	0,0926	0,0673	1,37706	0,205794
PT_RURAL*AGE	0,0884	0,0662	1,33517	0,218563
AGE	-18,1920	13,8716	-1,31145	0,226090
N_EMPLD*AGE	0,0016	0,0014	1,17792	0,272680
POP_CHNG	-6,1996	5,4079	-1,14639	0,284762
Intercept	435,1992	455,0209	0,95644	0,366857
PT_RURAL	-2,2153	2,5154	-0,88070	0,404161
PT_PHONE*AGE	0,0610	0,0903	0,67533	0,518504
TAX_RATE	53,9447	106,2945	0,50750	0,625495
POP_CHNG*PT_PHONE	0,0085	0,0250	0,34005	0,742576
POP_CHNG*PT_RURAL	0,0067	0,0208	0,32243	0,755394
N_EMPLD*PT_RURAL	0,0000	0,0001	0,28038	0,786301
PT_PHONE*PT_RURAL	0,0022	0,0091	0,23631	0,819127
N_EMPLD*PT_PHONE	0,0001	0,0006	0,18153	0,860464
POP_CHNG*N_EMPLD	0,0001	0,0004	0,15829	0,878152
PT_PHONE	-0,2162	3,1560	-0,06851	0,947063

Effect	Degr. of Freedom	PT_POOR SS	PT_POOR MS	PT_POOR F	PT_POOR p
TAX_RATE*PT_PHONE	1	55,600	55,59999	6,656145	0,032619
POP_CHNG*TAX_RATE	1	40,666	40,66620	4,868348	0,058405
TAX_RATE*AGE	1	30,417	30,41667	3,641326	0,092785
TAX_RATE*PT_RURAL	1	26,737	26,73743	3,200868	0,111393
N_EMPLD*TAX_RATE	1	23,494	23,49391	2,812570	0,132051
N_EMPLD	1	18,379	18,37944	2,200291	0,176274
POP_CHNG*AGE	1	15,840	15,84007	1,896292	0,205794
PT_RURAL*AGE	1	14,891	14,89105	1,782680	0,218563
AGE	1	14,367	14,36672	1,719910	0,226090
N_EMPLD*AGE	1	11,590	11,58991	1,387484	0,272680
POP_CHNG	1	10,978	10,97791	1,314219	0,284762
Intercept	1	7,641	7,64127	0,914773	0,366857
PT_RURAL	1	6,479	6,47900	0,775632	0,404161
PT_PHONE*AGE	1	3,810	3,80968	0,456075	0,518504
TAX_RATE	1	2,151	2,15143	0,257558	0,625495
POP_CHNG*PT_PHONE	1	0,966	0,96590	0,115633	0,742576
POP_CHNG*PT_RURAL	1	0,868	0,86839	0,103959	0,755394
N_EMPLD*PT_RURAL	1	0,657	0,65666	0,078612	0,786301
PT_PHONE*PT_RURAL	1	0,466	0,46648	0,055845	0,819127
N_EMPLD*PT_PHONE	1	0,275	0,27528	0,032955	0,860464
POP_CHNG*N_EMPLD	1	0,209	0,20929	0,025055	0,878152
PT_PHONE	1	0,039	0,03920	0,004693	0,947063
Error	8	66,825	8,35318		
Total	29	1197,727			

PT_PHONE	Step Number 3	1	24,93030	0,000031			In
POP_CHNG		1	14,46369	0,000743			In
TAX_RATE		1			0,17407	0,679945	Out
N_EMPLD		1			0,27568	0,603997	Out
PT_RURAL		1			1,56624	0,221897	Out
AGE		1			0,69720	0,411333	Out
POP_CHNG*N_EMPLD		1			0,37007	0,548243	Out
POP_CHNG*TAX_RATE		1			1,90508	0,179264	Out
N_EMPLD*TAX_RATE		1			0,13911	0,712191	Out
POP_CHNG*PT_PHONE		1			0,68908	0,414033	Out
N_EMPLD*PT_PHONE		1			0,15606	0,696028	Out
TAX_RATE*PT_PHONE		1			0,07543	0,785760	Out
POP_CHNG*PT_RURAL		1			1,39097	0,248921	Out
N_EMPLD*PT_RURAL		1			3,62637	0,067992	Out
TAX_RATE*PT_RURAL		1			1,33325	0,258731	Out
PT_PHONE*PT_RURAL		1			1,76890	0,195068	Out
POP_CHNG*AGE		1			0,88876	0,354492	Out
N_EMPLD*AGE		1			0,20766	0,652388	Out
TAX_RATE*AGE		1			0,02898	0,866136	Out
PT_PHONE*AGE		1			0,70741	0,407975	Out
PT_RURAL*AGE		1			0,72250	0,403085	Out

stepwise

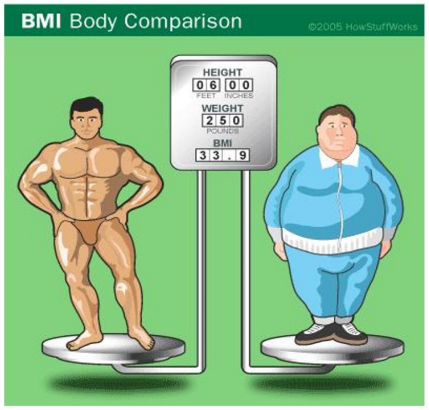
RÉGRESSION MARS

Ex7 : prédiction indice gras

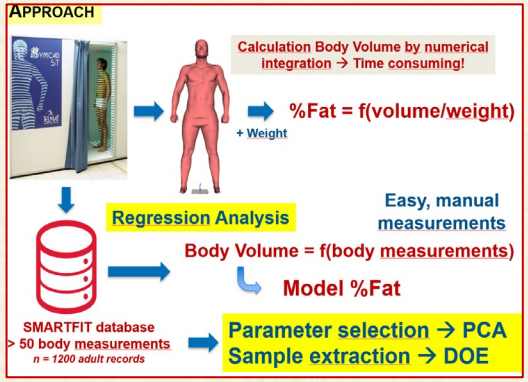
Modelling Percent Body Fat in a Human Body Using Design of Experiments and Regression Analysis
Frank Deruyck, Dr. Sc., Lecturer, University College Ghent
http://lib.stat.cmu.edu/datasets/bodyfat
/

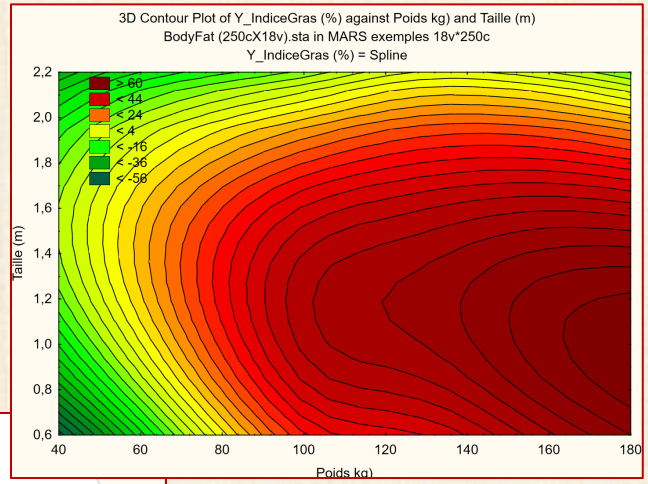
1 ID	2 Y_IndiceGras (%)	3 categori	4 genre	5 Age (an)	6 Poids kg)	7 Taille (m)	8 info	9 cou	10 poitrine	11 Abdomen	12 Biceps	13 avant bras	14 poignet	15 hanches	16 cuisses	17 genou	18 cheville
1	35,2	model	homme	46	164,7	1,82	mesures en cm	51,2	136,2	148,1	45,0	29,0	21,4	147,7	87,3	49,1	29,6
2	10,6	model	homme	57	67,0	1,66	circumference	35,2	99,6	86,4	31,7	27,3	16,9	90,1	53,0	35,0	21,3
3	24,2	model	homme	40	91,7	1,76		38,5	106,5	100,9	35,1	30,6	19,0	106,2	63,5	39,9	22,6
4	23,3	model	homme	52	75,7	1,71	1-cou	37,5	102,7	91,0	31,6	27,5	17,9	98,9	57,1	36,7	22,3
5	26,0	model	homme	54	104,3	1,82	2-poitrine	42,5	119,9	110,4	38,4	32,0	19,6	105,5	64,2	42,7	27,0
6	9,0	model	homme	47	83,6	1,88	3-Abdomen	37,3	99,6	88,8	30,3	27,9	17,8	101,4	57,4	39,6	24,6
7	22,1	model	homme	43	68,0	1,75	4-Biceps	35,2	91,1	85,7	29,4	26,6	17,4	96,9	55,5	35,7	22,0
8	9,4	model	homme	26	69,1	1,74	5-Avant bras	35,4	92,9	77,6	31,6	29,0	17,8	93,5	56,9	35,9	20,4
9	16,7	model	homme	40	71,7	1,75	6-poignet	36,3	97,0	86,6	29,8	26,3	17,3	92,6	55,9	36,3	22,1
10	29,9	model	homme	65	86,1	1,66	7-Hanches	40,8	106,4	100,5	35,9	30,5	19,1	100,5	59,2	38,1	24,0
11	11,7	model	homme	23	89,9	1,85	8-Cuisses	42,1	99,6	88,6	35,6	30,0	19,2	104,1	63,1	41,7	25,0
12	15,1	model	homme	34	63,5	1,78	9-Genou	36,0	89,2	83,4	28,3	26,2	16,5	89,6	52,4	35,6	20,4
13	18,7	model	homme	50	88,3	1,78	10-Cheville	39,0	103,7	97,6	32,7	30,0	19,0	104,2	60,0	40,9	25,5
14	17,5	model	homme	46	75,7	1,69		36,6	101,0	89,9	35,6	30,2	17,6	100,0	60,7	36,0	21,9

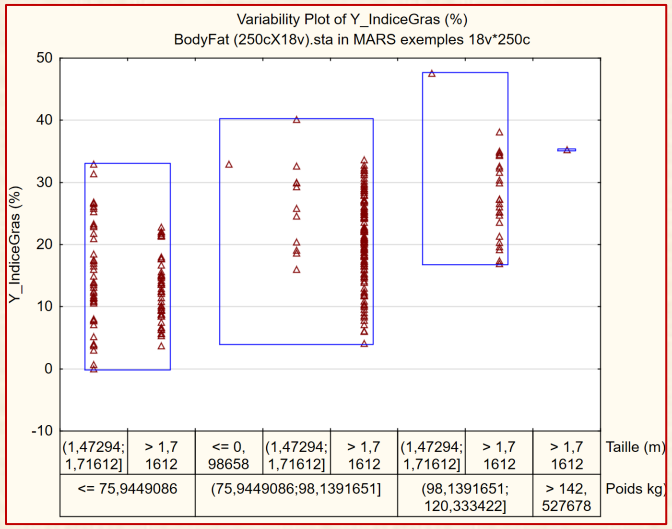
BODY FAT ≠ BMI

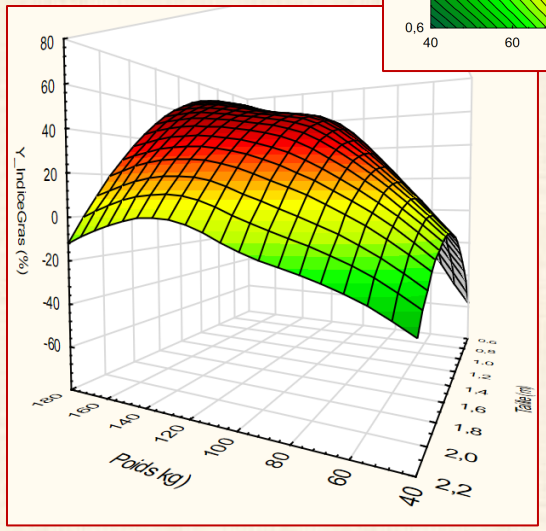


BMI = weight/length²





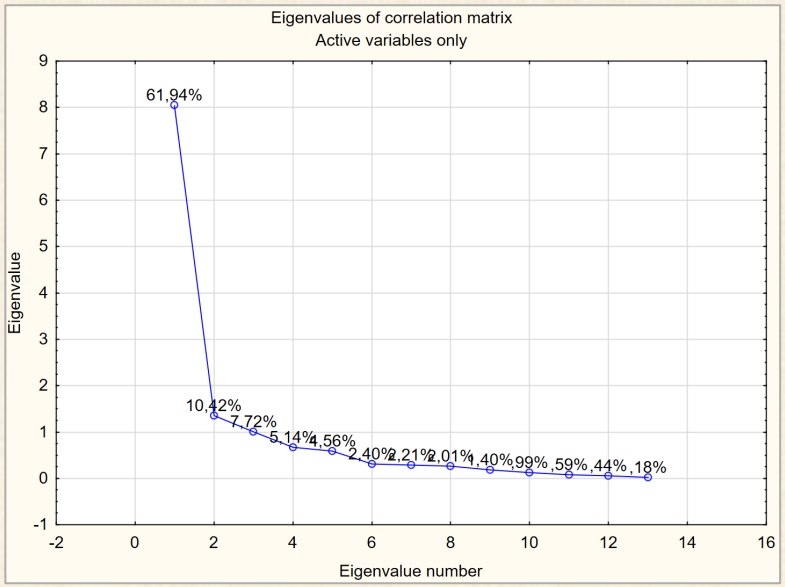


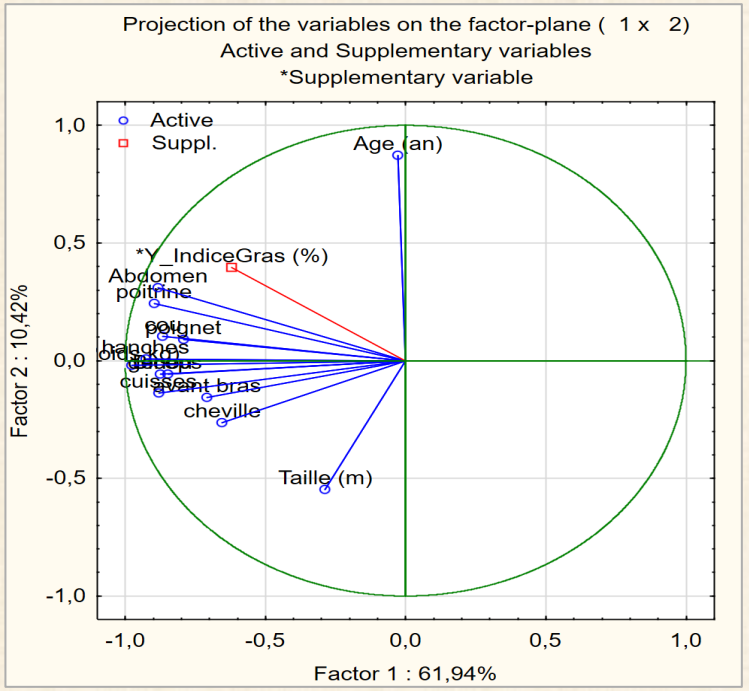


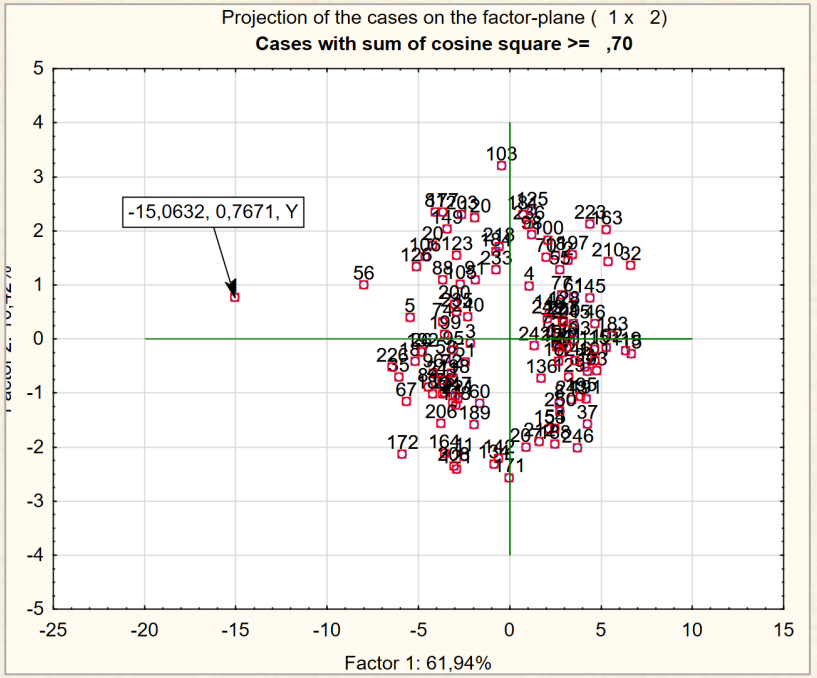
Ex7 : prédiction indice gras

ACP

Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %	IC
1	8,052	61,94	8,05	61,94	1,00
2	1,355	10,42	9,41	72,36	5,94
3	1,004	7,72	10,41	80,09	8,02
4	0,668	5,14	11,08	85,22	12,06
5	0,593	4,56	11,67	89,78	13,59
6	0,313	2,40	11,98	92,19	25,76
7	0,287	2,21	12,27	94,39	28,05
8	0,261	2,01	12,53	96,40	30,86
9	0,182	1,40	12,71	97,80	44,26
10	0,129	0,99	12,84	98,79	62,47
11	0,077	0,59	12,92	99,38	105,20
12	0,057	0,44	12,98	99,82	141,33
13	0,024	0,18	13,00	100,00	340,70



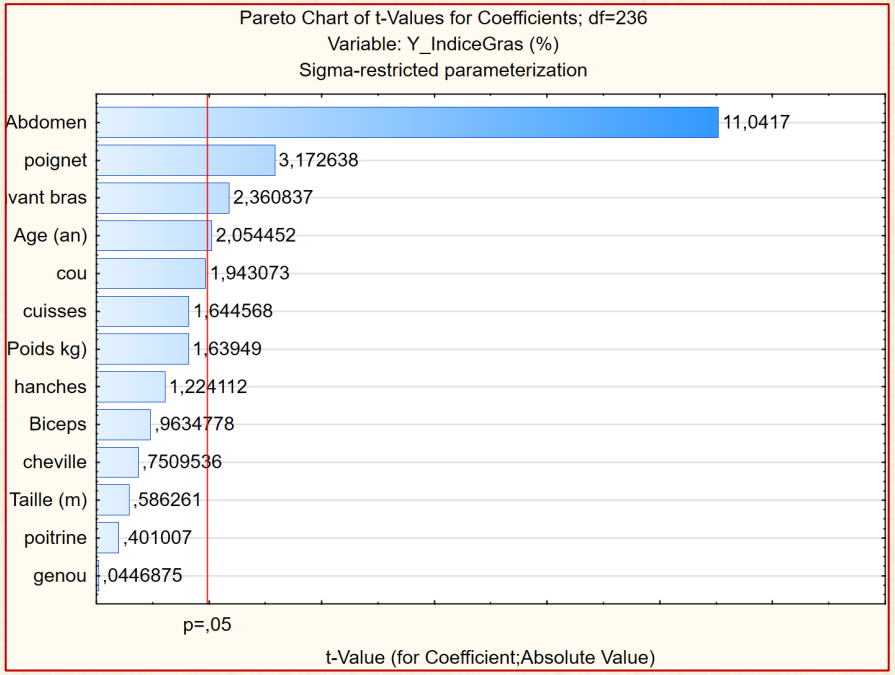




Ex7 : prédiction indice gras

REG ordinaire

Effect	SS	Degr. of Freedom	MS	F	p
Abdomen	2258,968	1	2258,968	121,9190	0,000000
poignet	186,500	1	186,500	10,0656	0,001711
avant bras	103,269	1	103,269	5,5736	0,019048
Age (an)	78,204	1	78,204	4,2208	0,041033
cou	69,955	1	69,955	3,7755	0,053197
cuisses	50,112	1	50,112	2,7046	0,101390
Poids kg)	49,803	1	49,803	2,6879	0,102444
hanches	27,764	1	27,764	1,4984	0,222131
Intercept	21,153	1	21,153	1,1417	0,286392
Biceps	17,200	1	17,200	0,9283	0,336294
cheville	10,449	1	10,449	0,5639	0,453429
Taille (m)	6,368	1	6,368	0,3437	0,558260
poitrine	2,979	1	2,979	0,1608	0,688778
genou	0,037	1	0,037	0,0020	0,964394
Error	4372,708	236	18,528		



Dependent Variable	Multiple R	Multiple R^2	Adjusted R^2
Y_IndiceGras (%)	0,866570	0,750943	0,737224

REG stepwise

Abdomen	Step Number 5	1	316,1586	0,000000			In
Poids kg)		1	29,1866	0,000000			In
poignet		1	12,5138	0,000483			In
avant bras		1	6,7960	0,009697			In
poitrine		1			0,1417	0,706921	Out
Age (an)		1			2,3233	0,128747	Out
Biceps		1			1,4986	0,222067	Out
cou		1			2,5600	0,110895	Out
Taille (m)		1			0,7517	0,386785	Out
hanches		1			0,0692	0,792687	Out
cuisses		1			1,4465	0,230251	Out
genou		1			0,9400	0,333248	Out
cheville		1			0,8231	0,365171	Out

Ex7 : prédiction indice gras

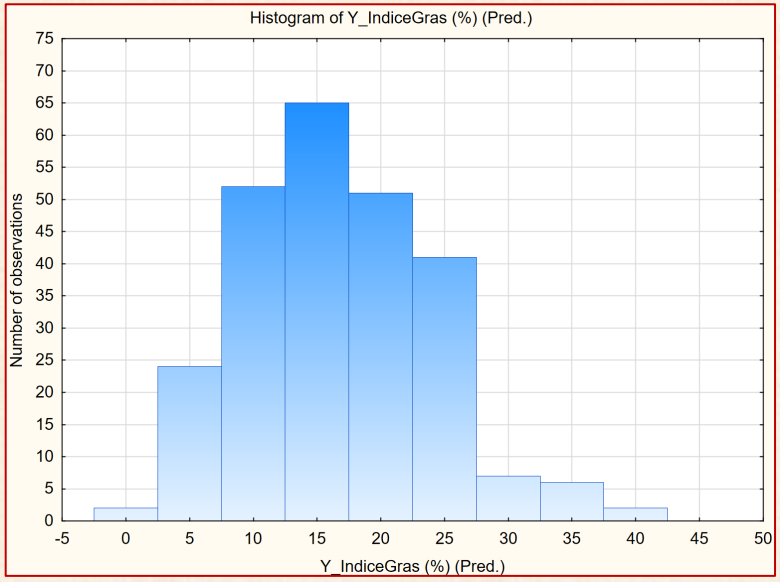
REG MARS

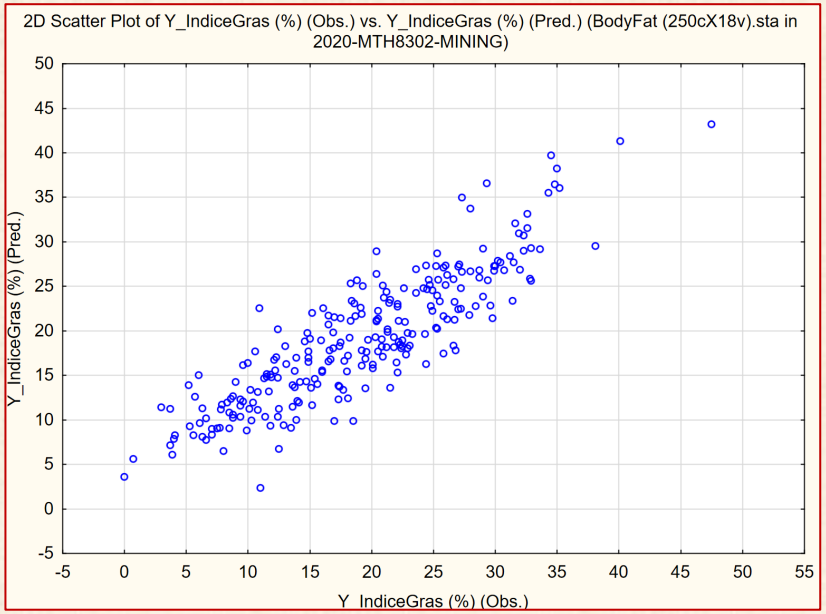
Model specifications	Value
Independents	13
Dependents	1
Number of terms	7
Number of basis functions	6
Order of interactions	1
Penalty	2,000000
Threshold	0,000500
GCV error	18,17612
Prune	Yes

Coefficients, knots and basis functions	Coefficients Y_IndiceGras (%)	Knots Age (an)	Knots Poids kg)	Knots Taille (m)	Knots cou	Knots poitrine	Knots Abdomen	Knots Biceps	Knots avant bras	Knots poignet	Knots hanches	Knots cuisses	Knots genou	Knots cheville
Intercept	12,66421													
Term.1	-0,37983		75,29627											
Term.2	3,68097									18,00000				
Term.3	-1,00070							35,80000						
Term.4	-0,43532							35,80000						
Term.5	-1,70706											53,70000		
Term.6	1,01419						82,50000							

Y_IndiceGras (%) = 1,26642146508039e+001 - 3,79828878462834e-001*max(0; Poids kg) -7,52962720000000e+001) + 3,68096788214241e+000*max(0; 1,80000000000000e+001-poignet) - 1,00069671220694e+000*max(0; Biceps-3,58000000000000e+001) - 4,35317413634945e-001*max(0; 3,58000000000000e+001-Biceps) - 1,70706133258035e+000*max(0; 5,37000000000000e+001-cuisses) + 1,01419202296201e+000*max(0; Abdomen-8,25000000000000e+001)

Regression statistics	Y_IndiceGras (%)
Mean (observed)	19,13880
Standard deviation (observed)	8,39704
Mean (predicted)	19,13880
Standard deviation (predicted)	7,35593
Mean (residual)	-0,00000
Standard deviation (residual)	4,04976
R-square	0,76740
R-square adjusted	0,76067



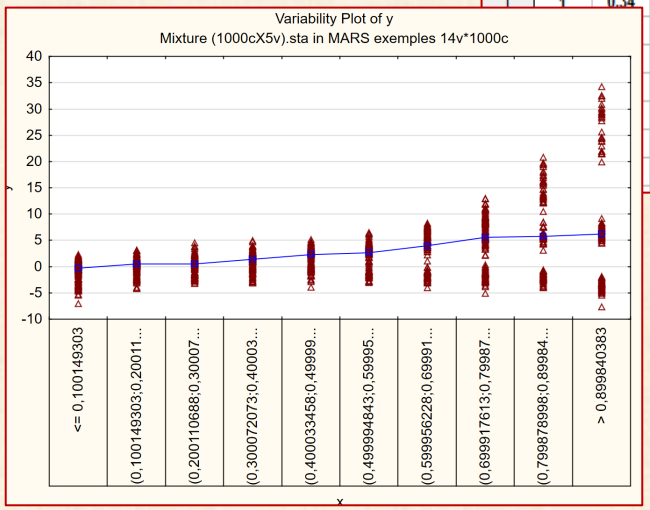


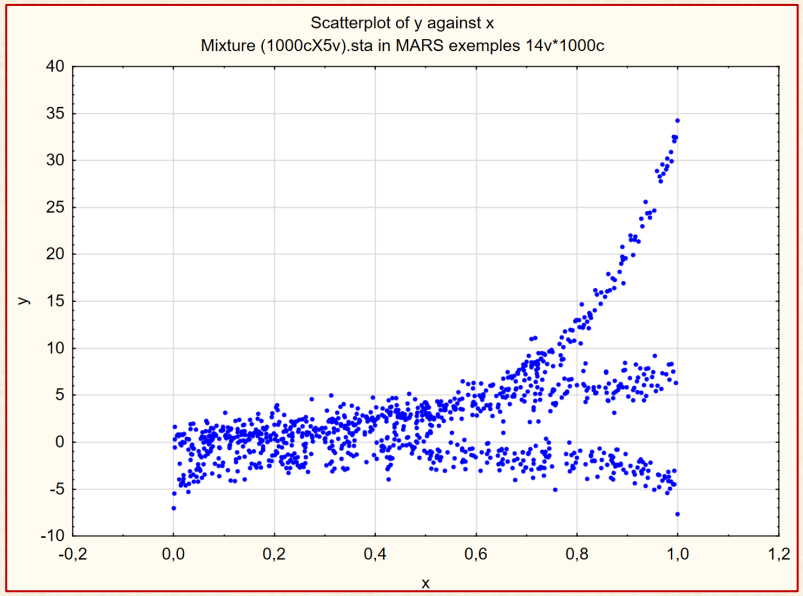
RÉGRESSION MARS

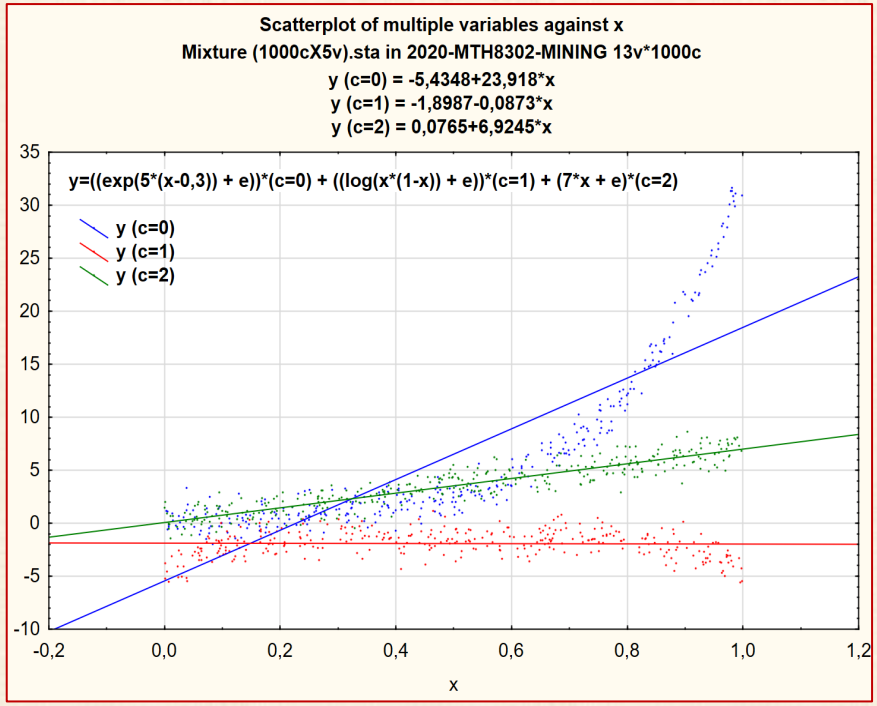
Ex8 : mélanges

Source : SAS SAS/STAT® 13.1 User's Guide
The ADAPTIVEREG Procedure (2013) p. 928
y=((exp(5*(x-0,3)) + e))*(c=0) + ((log(x*(1-x)) + e))*(c=1) + (7*x + e)*(c=2)

	1 ID	2 x	3 c1	4 c	5 e	6 y	7 c7	8 y (c=0)	9 y (c=1)	10 y (c=2)	11 c11	12 Y_pred (c=0)	13 Y_pred (c=1)	14 Y_pred (c=2)
1	1	0,34	2,62	2	0,47	2,84		12,39				10,81		
			0,59	0	-0,08	21,51		2,42				2,27		
			1,06	1	0,72	-0,70			-3,55				-3,62	
			1,09	1	-0,10	-3,09				0,18				0,67
			1,35	1	-0,48	-2,23		14,96				16,25		
			0,63	0	0,56	0,79			0,22				-1,19	
			0,60	0	0,34	1,11		5,22				4,45		







Ex8 - Data with mixture

analyse avec SAS

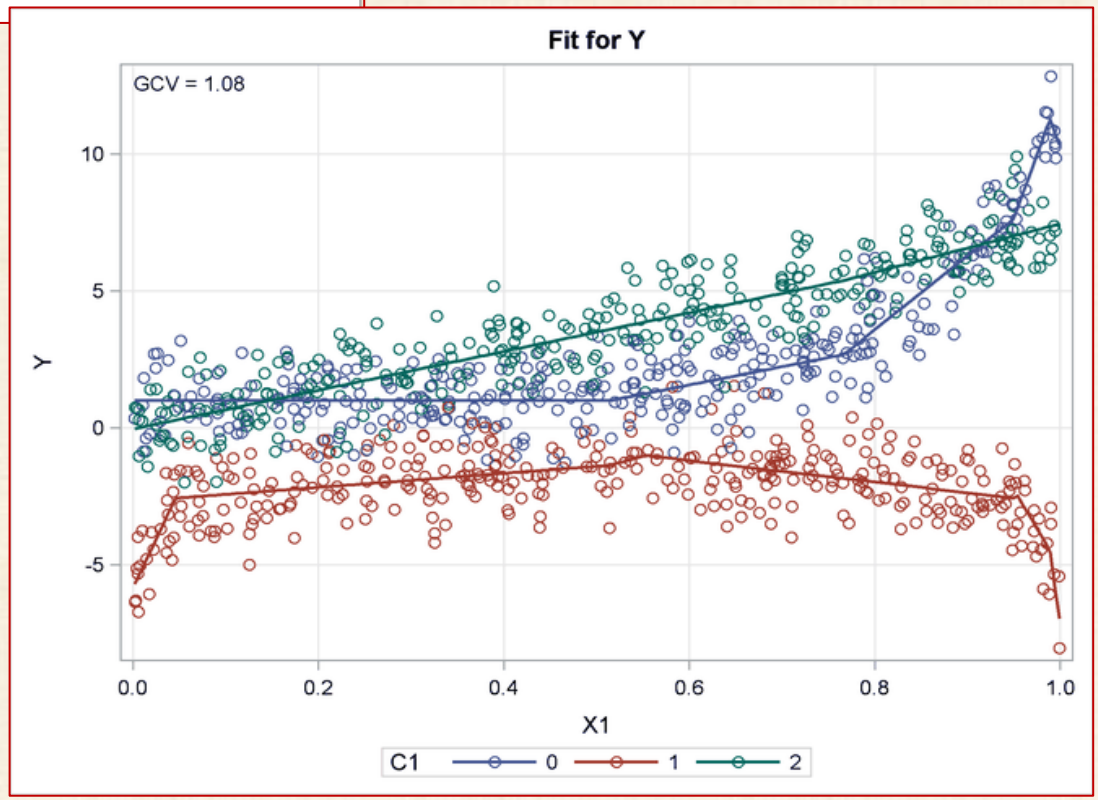
This example shows how you can use PROC ADAPTIVEREG to fit a model from a data set that contains mixture structures. It also demonstrates how to use the CLASS statement.

Consider a simulated data set that contains a response variable and two predictors, one continuous and the other categorical. The continuous predictor is sampled from the uniform distribution $U(0,1)$, and the classification variable is sampled from $U(0,3)$ and then rounded to integers. The response variable is constructed from three different models that depend on the CLASS variable levels, with error sampled from the standard normal distribution.

$$y = \begin{cases} \exp(5(x-0.3)^2), & \text{if } c = 0 \\ \log(x-x^2), & \text{if } c = 1 \\ 7x, & \text{if } c = 2 \end{cases}$$

Regression Spline Model after Backward Selection

Name	Coefficient	Parent	Variable	Knot	Levels
Basis0	5.3829		Intercept		
Basis1	-4.3871	Basis0	C1		1·0
Basis3	32.7761	Basis0	C1		1
Basis5	20.2859	Basis4	X1	0.7665	
Basis7	-11.4183	Basis2	X1	0.7665	
Basis8	-7.0758	Basis2	X1	0.7665	
Basis9	58.4911	Basis3	X1	0.5531	
Basis10	-71.6388	Basis3	X1	0.5531	
Basis11	-69.0764	Basis3	X1	0.04580	
Basis13	-119.71	Basis3	X1	0.9526	
Basis15	66.5733	Basis1	X1	0.9499	
Basis17	6.6681	Basis1	X1	0.5143	
Basis19	-185.21	Basis1	X1	0.9890	

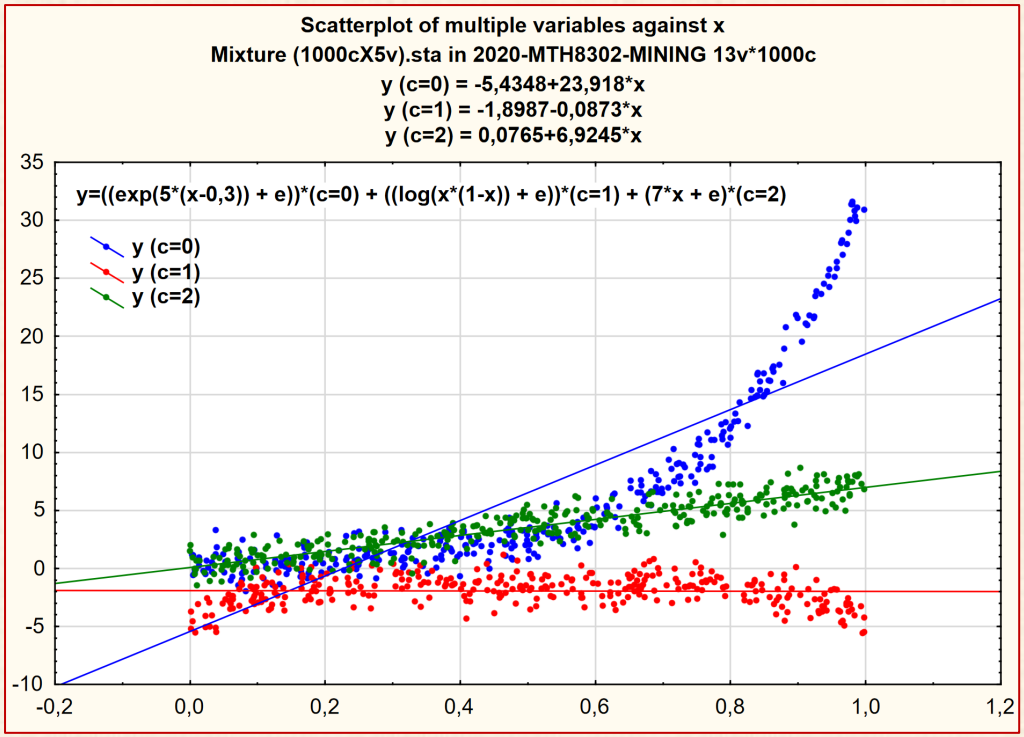


Ex8- Data with mixture

analyse avec Statistica

Source : SAS SAS/STAT® 13.1 User's Guide
The ADAPTIVEREG Procedure (2013) p. 928
y=((exp(5*(x-0,3)) + e))*(c=0) + ((log(x*(1-x)) + e))*(c=1)
+ (7*x + e)*(c=2)
/

	1 ID	2 x	3 c1	4 c	5 e	6 y
1	1	0,79	0,91	0	0,95	12,39
2	2	0,46	0,19	0	0,21	2,42
3	3	0,98	1,72	1	0,38	-3,55
4	4	0,05	2,23	2	-0,16	0,18
5	5	0,85	0,62	0	-0,74	14,96
...						
999	999	0,01	2,19	2	-0,18	-0,11
1000	1000	0,58	1,83	1	-1,13	-2,54

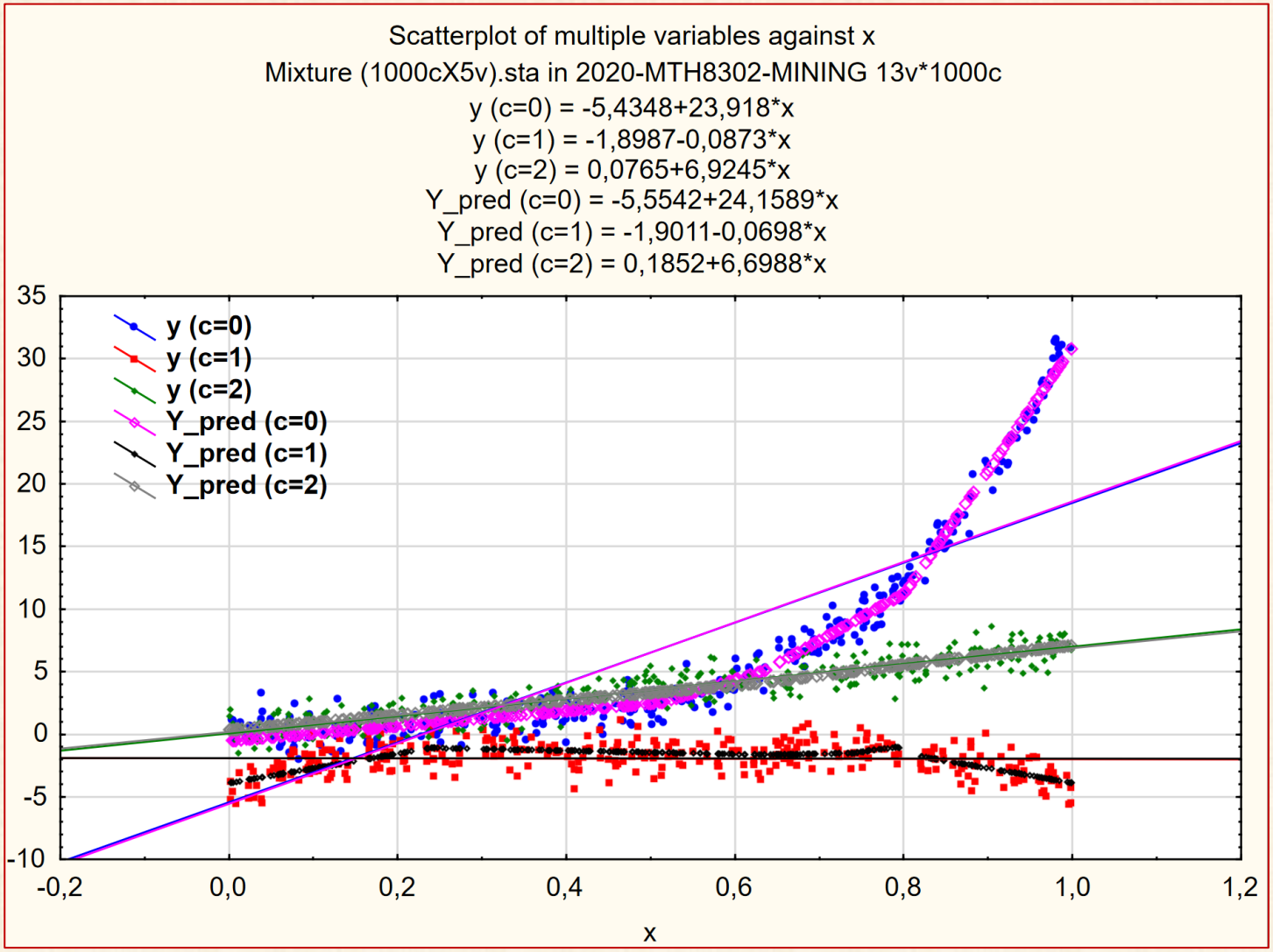


Regression statistics	y
Mean (observed)	2,660
Standard deviation (observed)	5,937
Mean (predicted)	2,660
Standard deviation (predicted)	5,852
Mean (residual)	-0,000
Standard deviation (residual)	1,002
R-square	0,972
R-square adjusted	0,971

y = 4,17 + 8,58*max(0; x-0,63) - 5,98*max(0; 0,63-x) + 13,92*max(0; x-0,63*max(0; c_0-0,00
+ 13,958*max(0; x-0,738*max(0; c_1-0,00) - 5,74*max(0; 0,70-x)*max(0; c_1-0,00)
+ 61,53*max(0; x-0,80) - 62,435*max(0; x-0,80*max(0; c_2) - 82,66*max(0; x-0,79)*max(0; c_1-0,00)
- 13,13*max(0; x-0,239)*max(0; c_1-0,00 - 0,858*max(0; c_0-0,00)
+ 14,79*max(0; x-0,51*max(0; c_0-0,00)

Ex8 - Data with mixture

analyse avec Statistica



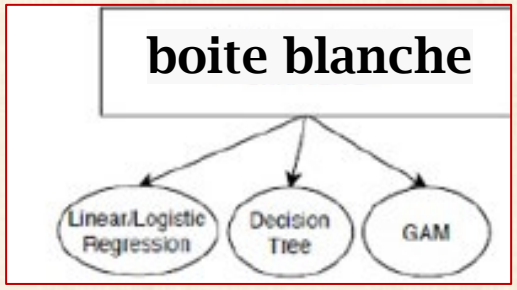
CONCLUSION

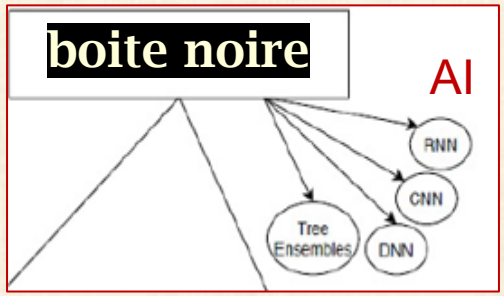
méthode MARS est démontrée (milliers d'articles depuis 1991)

- **supérieure à presque toutes les méthodes de modélisation incluant les réseaux de neurones et les arbres de classification**
- **a été appliquée dans tous les domaines:**
  économie, assurances, sciences de la vie,…

Briand_Freimut_Vollei-MARS.pdf
Chen-Otimal Airline Yield & Statistical Learning.pdf
Francis-MARS versus ANN.pdf
Friedman-article original MARS.pdf
Jesus&all-Comparison Statistical Methods.pdf
Kolyshkina-DataMining-Modeling Insurance Risk.pdf
Quiros&all-Satellite Images-MARS.pdf
Sephton-Forecasting recessions-MARS.pdf

- **robuste vis-à-vis des données aberrantes et manquantes**
- **modélise des données non linéaires de dimension élevée**
- **MARS pas une boite noire mais une boite blanche :**
  relation directe entre les entrées X et la sortie Y





Modèle MARS : avantages et inconvénients

Aucune technique de modélisation de régression n'est la meilleure pour toutes les situations.

- Les modèles MARS sont plus flexibles que les modèles de régression linéaire.
- Les modèles MARS sont simples à comprendre et à interpréter.
- Les modèles MARS sont souvent supérieurs aux modèles générés par un réseau neuronal entraîné ou une forêt aléatoire.
- MARS peut gérer à la fois des données continues et catégorielles .
- MARS a tendance à être meilleur que le partitionnement récursif pour les données numériques car les charnières sont plus appropriées pour les variables numériques que la segmentation constante par morceaux utilisée par le partitionnement récursif.
- La création de modèles MARS nécessite souvent peu ou pas de préparation des données.
- Les fonctions de charnière partitionnent automatiquement les données d'entrée, de sorte que l'effet des valeurs aberrantes est contenu. MARS est similaire au partitionnement récursif qui divise également les données en régions disjointes, bien qu'en utilisant une méthode différente.
- MARS (comme le partitionnement récursif) effectue une sélection automatique des variables (ce qui signifie qu'il inclut les variables importantes dans le modèle et exclut les variables sans importance). Cependant, il peut y avoir un certain arbitraire dans la sélection, en particulier lorsqu'il existe des prédicteurs corrélés, ce qui peut affecter l'interprétabilité. [1]
- Les modèles MARS ont tendance à présenter un bon compromis biais-variance. Les modèles sont suffisamment flexibles pour modéliser la non-linéarité et les interactions variables (les modèles MARS ont donc un biais assez faible), mais la forme contrainte des fonctions de base MARS empêche une trop grande flexibilité (les modèles MARS ont donc une variance assez faible).
- MARS est adapté à la gestion de grands ensembles de données et les implémentations s'exécutent très rapidement.
- Avec les modèles MARS, comme avec toute régression non paramétrique, les intervalles de confiance des paramètres et autres contrôles sur le modèle ne peuvent pas être calculés directement (contrairement aux modèles de régression linéaire). La validation croisée et les techniques associées doivent plutôt être utilisées pour valider le modèle.
- Les modèles MARS peuvent faire des prédictions très rapidement, car ils nécessitent uniquement l'évaluation d'une fonction linéaire des prédicteurs.
- La fonction ajustée résultante est continue, contrairement au partitionnement récursif, qui peut donner un modèle plus réaliste dans certaines situations. Le modèle n'est ni lisse ni différentiable.

https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_spline

Monographies & Articles

Berry, M., J., A., & Linoff, G., S., (2000). *Mastering Data Mining*. New York: Wiley

D.Hand (1999): *Why data mining is more than statistics write large*, ISI,Helsinki, http://www.stat.fi/isi99/index.html

D.Hand (2000): *Methodological Issues in Data Mining*, in Compstat 2000, Physica-Verlag, 77-85, 2000

Edelstein, H., A. (1999). *Introduction to Data Mining and Knowledge Discovery (3rd ed)*. Potomac, MD: Two Crows Corp.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances In Knowledge Discovery & Data Mining*. Cambridge, MA: MIT Press.

Friedman J. (1997): *Data Mining and Statistics, What's the Connection?* http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps

Friedman J. (1999): *The role of Statistics in Data Revolution*, ISI, Helsinki, http://www.stat.fi/isi99/index.html

Friedman J. (2009): première heure du cours STAT315B (Stanford Univ.) sur le Data Mining (donné à l'hiver 2009) http://myvideos.stanford.edu/player/slplayer.aspx?course=STATS315B&p=true

Gaudard, M. Ramsey, P., Stephens, M. ((2006). *Interactive Data Mining and Design of Experiments: The JMP Partition and Custom Design Platforms*. North Haven Group, LLC

Giudici, P. (2003*). Applied Data Mining: Statistical Methods for Industry*, John Wiley & Sons.

Han, J., Kamber, M. (2000). *Data Mining: Concepts and Techniques*. New York: Morgan-Kaufman.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of Statistical Learning : Data Mining, Inference, and Prediction*. New York: Springer.

Kantardzic, M.M., Zurada, J. (editors) (2005). *Next Generation of Data-Mining Applications*. John Wiley & Sons, Copyright the Institute of electrical and Electronic Engineers (IEEE).

Larose, Daniel T, (2005*) Discovering Knowledge in Data : An Introduction to Data Mining* . John Wiley & Sons.

Nisbet R., Elder, J., Miner, G. (2009*) Handbook of Statistical Analysis & Data Mining Apllications*, Academic Press. ISBN 978-0-12-374765-5

Pregibon, D. (1997). *Data Mining*. Statistical Computing and Graphics, 7, 8.

StatSoft : 35 vidéos de 8-10 minutes sur YouTube http://www.statsoft.com/support/download/video-tutorials/

Tufféry, S. (2007). *Data Mining et statistique décisionnelle*, Éditions TECHNIP, Paris.

Weiss, S. M., & Indurkhya, N. (1997). *Predictive Data Mining: A practical Guide*. New York: Morgan-Kaufman.

Westphal, C., Blaxton, T. (1998). *Data Mining Solutions*. New York: Wiley.

Witten, I. H., & Frank, E. (2000). *Data Mining*. New York: Morgan-Kaufmann.

Sites Internet

http://www.kdnuggets.com/

http://www.ccsu.edu/datamining/

http://www.math.ccsu.edu/dm/dm%20resources.htm

http://www.dmreview.com

http://www.scd.ucar.edu/hps/GROUPS/dm/dm.html

http://www.infogoal.com/dmc/dmcdwh.htm

http://www.salford-systems.com

http://support.sas.com/documentation/cdl/en/statug/65328/HTML/default/viewer.htm#statug_adaptivereg_overview.htm

https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_spline