Chapitre 3 Regression Multiple-1

ALGORITHMES (méthodes) (Machine Learning)

SUPERVISÉES : X et Y

- Régression multiple ordinaire
- Régression non linéaire
- Régression linéaire généralisée
- Régression avec contraintes: Ridge, Lasso
- Régression splines (MARS)
- Régression généralisée additive
- Régression réseaux neuronaux
- Flux Tenseur
- Arbres de classification (CRT)
- Forêts Aléatoires
- Méthodes gradient non-convexe
- Algorithmes génétiques
- Méthodes ensemblistes
- Régression boosted
- XGBoost
- ……

NON SUPERVISÉES : X

- Réduction dimension (PCA)
- Clustering
- K-Means
- K-Neighbour
- Classification hiérarchique
- Réseaux Baysiens
- Modèle de Markov
- ….

SÉRIES CHRONOLOGIQUES

Deep Learning

- Apprentissage profond
- réseaux neurones multicouches
- intelligence artificielle (AI)

MODÈLE $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \bullet\bullet\bullet + \beta_m X_m + \varepsilon$ $\qquad \varepsilon \sim N(0, \sigma^2)$

DONNÉES

#	X_0	X_1	X_2	X_3	.....	X_m	Y
1	1	x_{11}	x_{12}	x_{13}	.....	x_{1m}	y_1
2	1	x_{21}	x_{22}	x_{23}	.....	x_{2m}	y_2
.	.	.	.	.	.....	.	.
i	1	x_{i1}	x_{i2}	x_{i3}	.....	x_{im}	y_i
.	.	.	.	.	.....	.	.
N	1	x_{N1}	x_{N2}	x_{N3}	....	x_{Nm}	y_N

$X_{i0} = 1$: effet général

matrice des effets : colonnes types X_j , $X_j X_k$, X_j^2

écriture matricielle $X_{N \times p} = [\, x_{ij} \,]$ $\qquad p = 1 + m$ $\qquad Y_{N \times 1}$: vecteur N x 1

$\beta_{p \times 1} = (\beta_0 , \beta_1 , \beta_2 , \beta_3 , ... , \beta_m)'$: vecteur p x 1

remarque : l'opération de transposition de vecteurs / matrices est notée par le symbole '

ESTIMATION (principe de moindres carrés) : $\underset{\beta}{\text{Min}} \sum_i (y_i - \sum_j \beta_j x_{ij})^2$

système d'équations linéaires à résoudre : $(X' X)\, \beta = X' Y$

solution : $\hat{\beta} = (X' X)^{-1} X' Y = C\, Y$

$C = (X' X)^{-1} X'$ est une matrice p x N de valeurs fixes connues

ÉQUATION de PRÉDICTION $\hat{y} = X \hat{\beta}$

PROPRIÉTÉS ESTIMATEURS $\hat{\beta}$

- combinaisons linéaires des y_i
- sans biais : $E(\hat{\beta}_j) = \beta_j$ pas d'erreur systématique
- $var(\hat{\beta}) = (X'X)^{-1}\sigma^2$ variance minimale (« meilleurs »)

ESTIMATION de σ^2

résidu : $e_i = \hat{y}_i - y_i$

somme de carrés résiduels : $SSresid = \sum e_i^2$

carré résiduel moyen : $MSresid = SSresid / N - m - 1$

estimation : $\hat{\sigma}^2 = MSresid$ $\hat{\sigma} = (MSresid)^{0.5}$

DÉCOMPOSITION DE LA VARIABILITÉ : tableau d'analyse de la variance

$SStot = \sum (y_i - \bar{y})^2$: somme TOTALE des carrés

$SSreg = \sum (\hat{y}_i - \bar{y})^2$: somme des carrés du MODÈLE (régression)

$SSresid = \sum (\hat{y}_i - y_i)^2$: somme des carrés RÉSIDUELS

ÉQUATION FONDAMENTALE

somme de carrés (SS) : SStot = SSreg + SSresid

variabilité : totale = modèle + résiduelle

degrés de liberté (df) : N – 1 = m + (N – m - 1)

TABLEAU D'ANALYSE VARIANCE : modèle de régression linéaire multiple

SOURCE	df	SS	MS = SS / df	F-ratio	p-valeur
régression	m	SSreg	MSreg = SSreg / m	F_o = MSreg / MSresid	P(F ≥ F_o)
résiduelle	N – m - 1	SSresid	MSresid = SSresid /(N – m - 1) = $\hat{\sigma}^2$	----------	--------
totale	N – 1	SStot	----------	---------	--------

R^2 = SSreg / SStot : coefficient de détermination

$0 \leq R^2 \leq 1$: fraction de la variabilité de Y expliquée par les variables X

$R^2_{ajusté} = 1 - [(N - 1)/(N - m)] (1 - R^2)$: coefficient de détermination ajusté

remarques

- ajouter une variable explicative additionnelle dans un modèle augmente SSreg et R^2
- l'augmentation de R^2 pas toujours importante et significative
- $R^2_{ajusté}$ est préférable à R^2 pour comparer deux modèles

Test global $H_{0G} : \beta_1 = \beta_2 = \ldots = \beta_m = 0$

vs H_{1G} : non H_{0G} ... au moins un $\beta \neq 0$

rejeter H_0 au seuil α si $F_o > F_{m,\ N-m-1,\ 1-\alpha}$

Distribution d'échantillonnage de $\hat{\beta}_j$ $\quad j = 0, 1, 2, \ldots, k$

$[(\hat{\beta}_j - \beta_j) / \hat{\sigma} (c_{jj})^{0.5}] \sim t_{N-m-1}$ $\quad c_{jj}$: j-ème élément diagonal de $(X' X)^{-1}$

Applications

(a) test $H_{0j} : \beta_j = 0$ vs $H_{1j} : \beta_j \neq 0$

rejeter H_{0j} au seuil α si $| \hat{\beta}_j | / \hat{\sigma} (c_{jj})^{0.5} > t_{N-m-1, 1-\alpha/2}$

(b) Intervalle de confiance $\beta_j : \hat{\beta}_j \pm t_{N-m-1, 1-\alpha/2} \hat{\sigma} (c_{jj})^{0.5}$

(c) INTERVALLE de CONFIANCE MOYENNE de Y à $X_1 = x_1^*, X_2 = x_2^*, .., X_k = x_k^*$

$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^*$ $\quad x^* = (x_1^*, x_2^*, \ldots, x_k^*)$

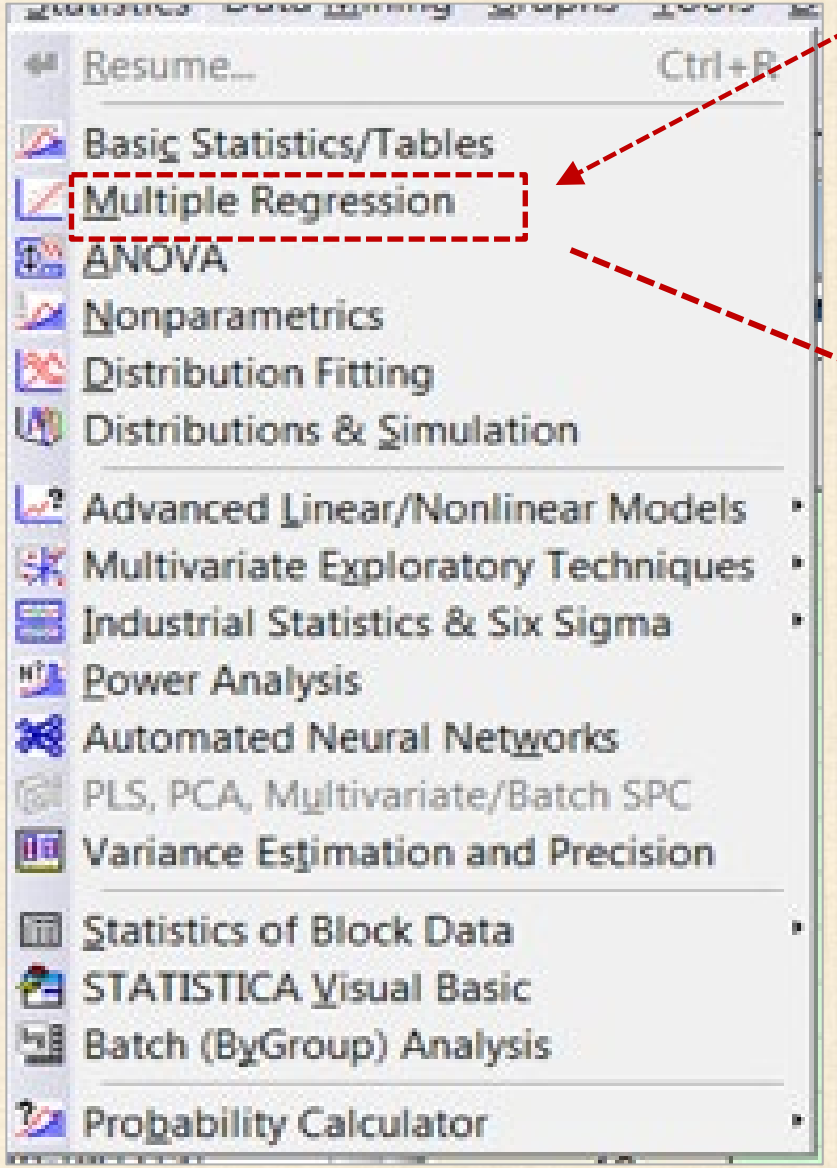
$E (Y \mid X = x^*) : \hat{y}^* \pm t_{N-m-1, 1-\alpha/2} \hat{\sigma} [x^* (X' X)^{-1} x^{*\prime}]^{0.5}$

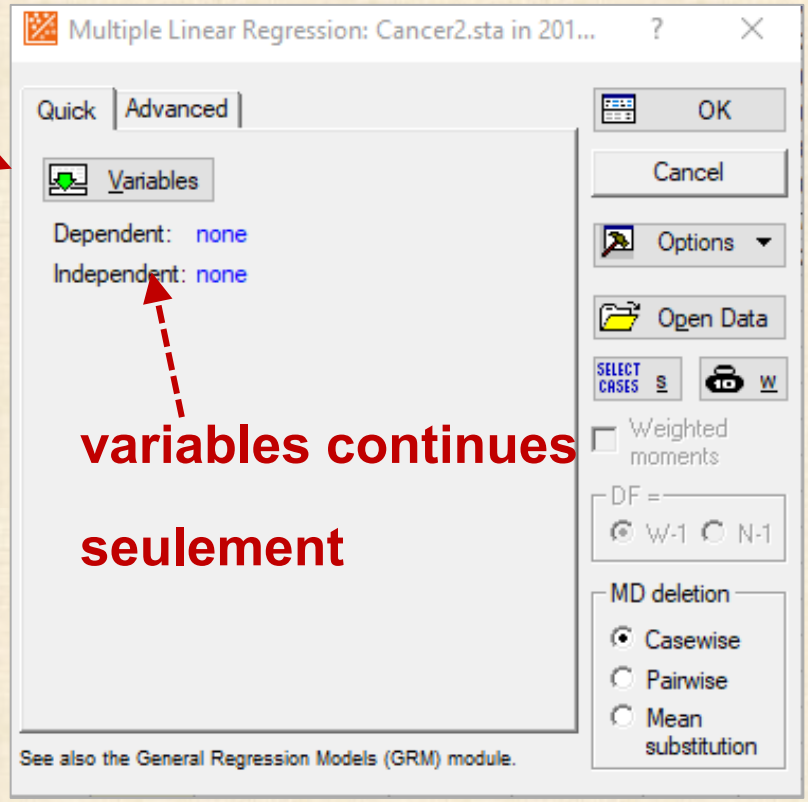
(d) INTERVALLE de PRÉDICTION VALEUR de Y à $X_1 = x_1^*, X_2 = x_2^*, .., X_k = x_k^*$

$Y \mid X = x^* : \hat{y}^* \pm t_{N-m-1, 1-\alpha/2} \hat{\sigma} [1 + x^* (X' X)^{-1} x^{*\prime}]^{0.5}$

Mise en œuvre avec *STATISTICA*

Statistics… Multiple Linear Regression

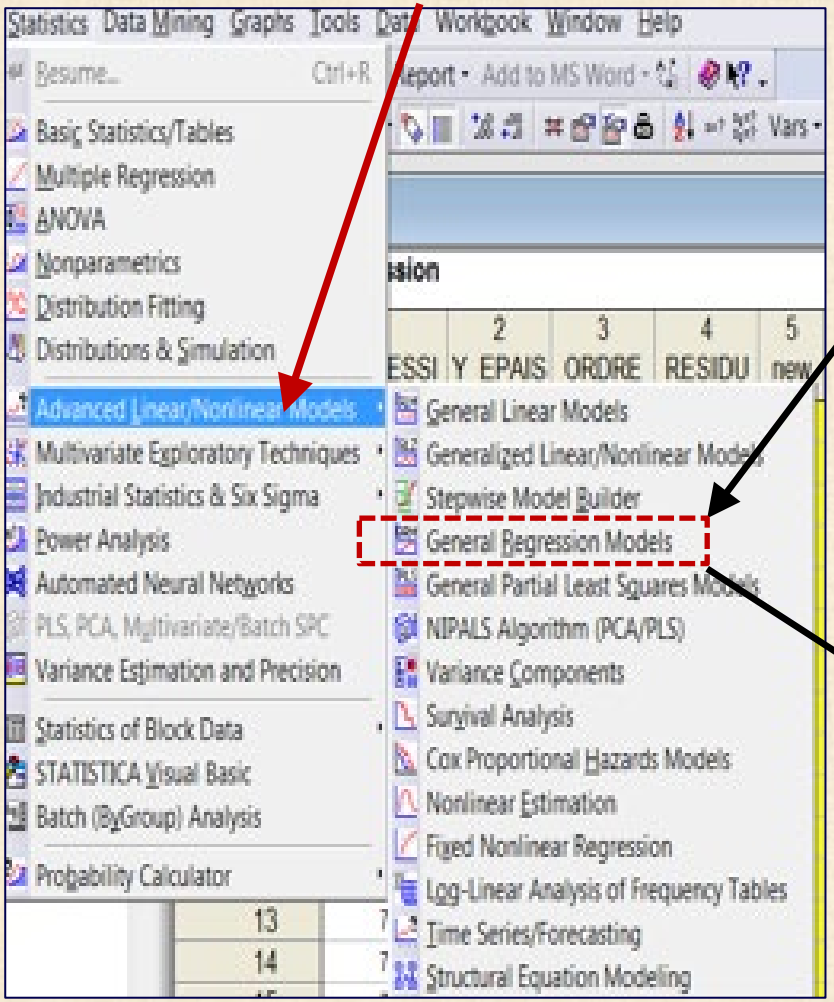


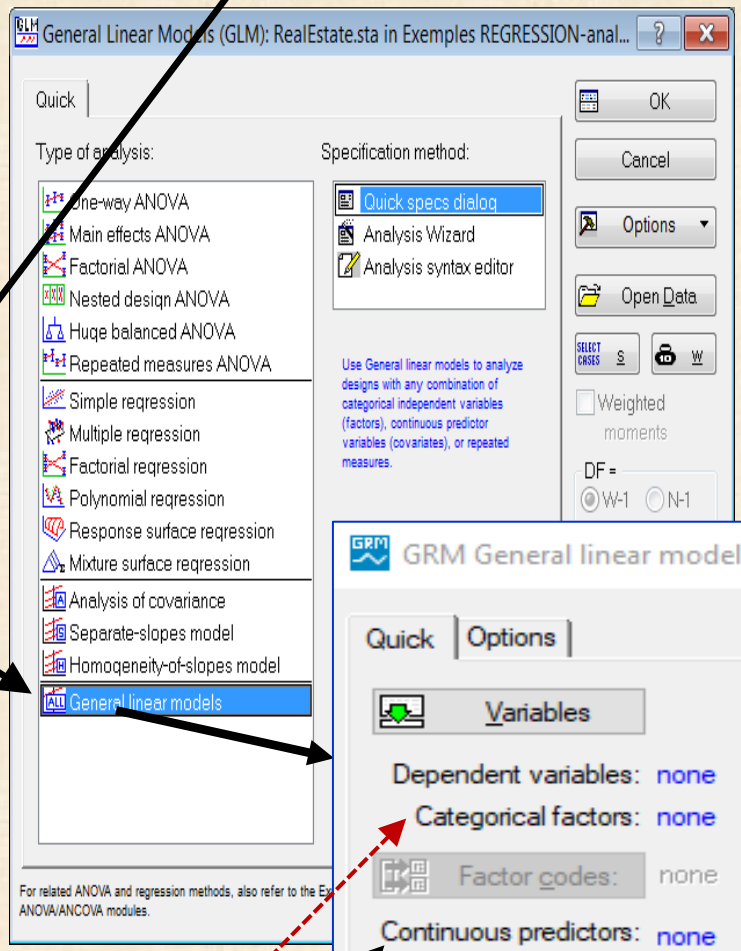


variables continues seulement

Mise en œuvre avec *STATISTICA*

Statistics … Advanced Linear/ Nonlinear Models … General Regression Models





Statistica fait un codage à effet

variables catégoriques

variables continues

EXEMPLE RÉGRESSION MULTIPLE

Exemple: production de gazoline avec huiles brutes (données historiques)

N. H. Prater, *Petroleum Refiner - Experimental Designs in Industry* (ed. Chew) Wiley 1956 pp109-137

Y : rendement production gazoline (% conversion de l'huile brute)

X1 : gravité huile brute (deg. API) X2 : pression vapeur (PSIA)

X3 : ASTM point 10% (deg. F) X4 : point sortie gazoline (deg. F)

#	X1	X2	X3	X4	Y
1	38.4	6.1	220	235	6.9
2	40.3	4.8	231	307	14.4
3	40.0	6.1	217	212	7.4
4	31.8	0.2	316	365	8.5
5	40.8	3.5	210	218	8.0
6	41.3	1.8	267	235	2.8
7	38.1	1.2	274	285	5.0
8	50.8	8.6	190	205	12.2
9	32.2	5.2	236	267	10.0
10	38.4	6.1	220	300	15.2
11	40.3	4.8	231	367	26.8
12	32.2	2.4	284	351	14.0
13	31.8	0.2	316	379	14.7
14	41.3	1.8	267	275	6.4
15	38.1	1.2	274	365	17.6
16	50.8	8.6	190	275	22.3

data = gazoline.sta

17	32.2	5.2	236	360	24.8
18	38.4	6.1	220	365	26.0
19	40.3	4.8	231	395	34.9
20	40.0	6.1	217	272	18.2
21	32.2	2.4	284	424	23.2
22	31.4	0.2	316	428	18.0
23	40.8	3.5	210	273	13.1
24	41.3	1.8	267	358	16.1
25	38.1	1.2	274	444	32.1
26	50.8	8.6	190	345	34.7
27	32.2	5.2	236	402	31.7
28	38.4	6.1	220	410	33.6
29	40.0	6.1	217	340	30.4
30	40.8	3.5	210	347	26.6
31	41.3	1.8	267	416	27.8
32	50.8	8.6	190	407	45.7

présence d'une structure dans ces données?

réponse : oui

autre analyse proposée plus loin

Examen des données

	1 ID	2 X1	3 X2	4 X3	5 X4	6 Y	7 new	8 X1cr	9 X2cr	10 X3cr	11 X4cr	12 Ycr	13 new	14 X1n	15 X2n	16 X3n	17 X4n	18 Yn	19 new	20 ID2	21 X1cop	22 X2cop	23 X3cop	24 groupe	25 X4cop	26 Ycop
1	1	38,4	6,1	220	235	6,9	variables	-0,1	0,7	-1	-1	-1,2	variables	-0,28	0,40	-0,52	-0,75	6,9	X1cop = X1	8	50,8	8,6	190	1	205	12,2
2	2	40,3	4,8	231	307	14,4	centrées-réduites	0,2	0,2	0	0	-0,5	normalisées	-0,08	0,10	-0,35	-0,15	14,4	idem pour X1 X2 X3	16	50,8	8,6	190	1	275	22,3
3	3	40,0	6,1	217	212	7,4		0,1	0,7	-1	-2	-1,1		-0,11	0,40	-0,57	-0,94	7,4		26	50,8	8,6	190	1	345	34,7
4	4	31,8	0,2	316	365	8,5	X1cr =	-1,3	-1,5	2	0	-1,0	x1n =	-0,96	-1,00	1,00	0,34	8,5	groupe selon les	32	50,8	8,6	190	1	407	45,7
5	5	40,8	3,5	210	218	8,0	(X1 - X1moy) / X1sd	0,3	-0,3	-1	-2	-1,1	(x1 - m) / b	-0,03	-0,21	-0,68	-0,89	8,0	valeus distinctes de	6	41,3	1,8	267	2	235	2,8
6	6	41,3	1,8	267	235	2,8		0,4	-0,9	1	-1	-1,6		0,02	-0,62	0,22	-0,75	2,8	X1 X2 X3	14	41,3	1,8	267	2	275	6,4
7	7	38,1	1,2	274	285	5,0	moyenne X1cr = 0	-0,2	-1,1	1	-1	-1,4	m = (Xmax - Xmin) / 2	-0,31	-0,76	0,33	-0,33	5,0		24	41,3	1,8	267	2	358	16,1
8	8	50,8	8,6	190	205	12,2	écart-type X1cr = 1	2,0	1,7	-1	-2	-0,7	b = (Xmax + Xmin) / 2	1,00	1,00	-1,00	-1,00	12,2	en ordre croissante	31	41,3	1,8	267	2	416	27,8
9	9	32,2	5,2	236	267	10,0		-1,2	0,4	0	-1	-0,9		-0,92	0,19	-0,27	-0,48	10,0	de X3	5	40,8	3,5	210	3	218	8,0
10	10	38,4	6,1	220	300	15,2	idem pour	-0,1	0,7	-1	0	-0,4	idem pour	-0,28	0,40	-0,52	-0,21	15,2		23	40,8	3,5	210	3	273	13,1
11	11	40,3	4,8	231	367	26,8	X2cr X3cr X4cr Ycr	0,2	0,2	0	1	0,7	X2n X3n X4n Yn	-0,08	0,10	-0,35	0,36	26,8	10 groupes distincts	30	40,8	3,5	210	3	347	26,6
12	12	32,2	2,4	284	351	14,0		-1,2	-0,7	1	0	-0,5	Yn = Y	-0,92	-0,48	0,49	0,22	14,0		2	40,3	4,8	231	4	307	14,4
13	13	31,8	0,2	316	379	14,7		-1,3	-1,5	2	1	-0,5		-0,96	-1,00	1,00	0,46	14,7	interprétation de groupe ?	11	40,3	4,8	231	4	367	26,8
14	14	41,3	1,8	267	275	6,4		0,4	-0,9	1	-1	-1,2	varient sur	0,02	-0,62	0,22	-0,41	6,4	disons	19	40,3	4,8	231	4	395	34,9
15	15	38,1	1,2	274	365	17,6		-0,2	-1,1	1	0	-0,2	l'intervalle -1 a 1	-0,31	-0,76	0,33	0,34	17,6	pays d'origine	3	40,0	6,1	217	5	212	7,4
16	16	50,8	8,6	190	275	22,3		2,0	1,7	-1	-1	0,2		1,00	1,00	-1,00	-0,41	22,3	de l'huile brute	20	40,0	6,1	217	5	272	18,2
17	17	32,2	5,2	236	360	24,8		-1,2	0,4	0	0	0,5		-0,92	0,19	-0,27	0,30	24,8	groupe = pays	29	40,0	6,1	217	5	340	30,4
18	18	38,4	6,1	220	365	26,0		-0,1	0,7	-1	0	0,6		-0,28	0,40	-0,52	0,34	26,0	1 = paya1	1	38,4	6,1	220	6	235	6,9
19	19	40,3	4,8	231	395	34,9		0,2	0,2	0	1	1,4		-0,08	0,10	-0,35	0,59	34,9	2 = pays2	10	38,4	6,1	220	6	300	15,2
20	20	40,0	6,1	217	272	18,2		0,1	0,7	-1	-1	-0,1		-0,11	0,40	-0,57	-0,44	18,2	.	18	38,4	6,1	220	6	365	26,0
21	21	32,2	2,4	284	424	23,2		-1,2	-0,7	1	1	0,3		-0,92	-0,48	0,49	0,83	23,2	.	28	38,4	6,1	220	6	410	33,6
22	22	31,4	0,2	316	428	18,0		-1,4	-1,5	2	1	-0,2		-1,00	-1,00	1,00	0,87	18,0	10 = pays10	7	38,1	1,2	274	7	285	5,0
23	23	40,8	3,5	210	273	13,1		0,3	-0,3	-1	-1	-0,6		-0,03	-0,21	-0,68	-0,43	13,1		15	38,1	1,2	274	7	365	17,6
24	24	41,3	1,8	267	358	16,1		0,4	-0,9	1	0	-0,3		0,02	-0,62	0,22	0,28	16,1		25	38,1	1,2	274	7	444	32,1
25	25	38,1	1,2	274	444	32,1		-0,2	-1,1	1	2	1,2		-0,31	-0,76	0,33	1,00	32,1		9	32,2	5,2	236	8	267	10,0
26	26	50,8	8,6	190	345	34,7		2,0	1,7	-1	0	1,4		1,00	1,00	-1,00	0,17	34,7		17	32,2	5,2	236	8	360	24,8
27	27	32,2	5,2	236	402	31,7		-1,2	0,4	0	1	1,1		-0,92	0,19	-0,27	0,65	31,7		27	32,2	5,2	236	8	402	31,7
28	28	38,4	6,1	220	410	33,6		-0,1	0,7	-1	1	1,3		-0,28	0,40	-0,52	0,72	33,6		12	32,2	2,4	284	9	351	14,0
29	29	40,0	6,1	217	340	30,4		0,1	0,7	-1	0	1,0		-0,11	0,40	-0,57	0,13	30,4		21	32,2	2,4	284	9	424	23,2
30	30	40,8	3,5	210	347	26,6		0,3	-0,3	-1	0	0,6		-0,03	-0,21	-0,68	0,19	26,6		4	31,8	0,2	316	10	365	8,5
31	31	41,3	1,8	267	416	27,8		0,4	-0,9	1	1	0,8		0,02	-0,62	0,22	0,77	27,8		13	31,8	0,2	316	10	379	14,7
32	32	50,8	8,6	190	407	45,7		2,0	1,7	-1	1	2,4		1,00	1,00	-1,00	0,69	45,7		22	31,4	0,2	316	10	428	18,0

examen
des
données
et variables

variables
originelles X

vs

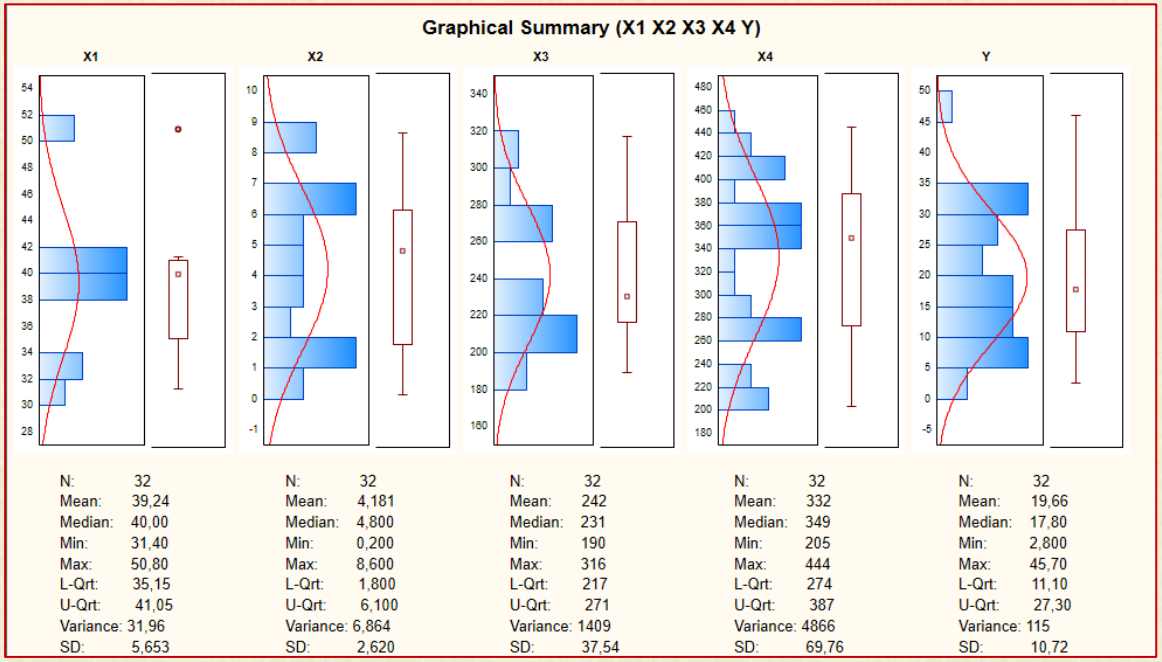
variables
centrées-réduites
Xcr

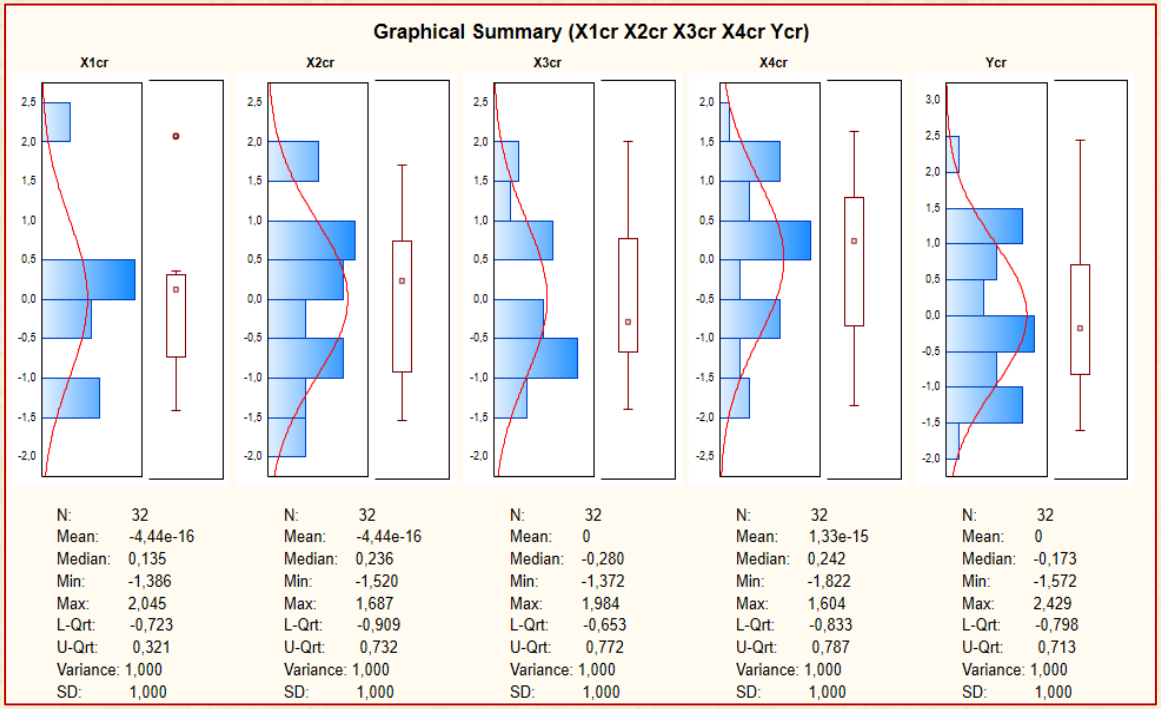
avec

Basic Stat

et

graphical
summary



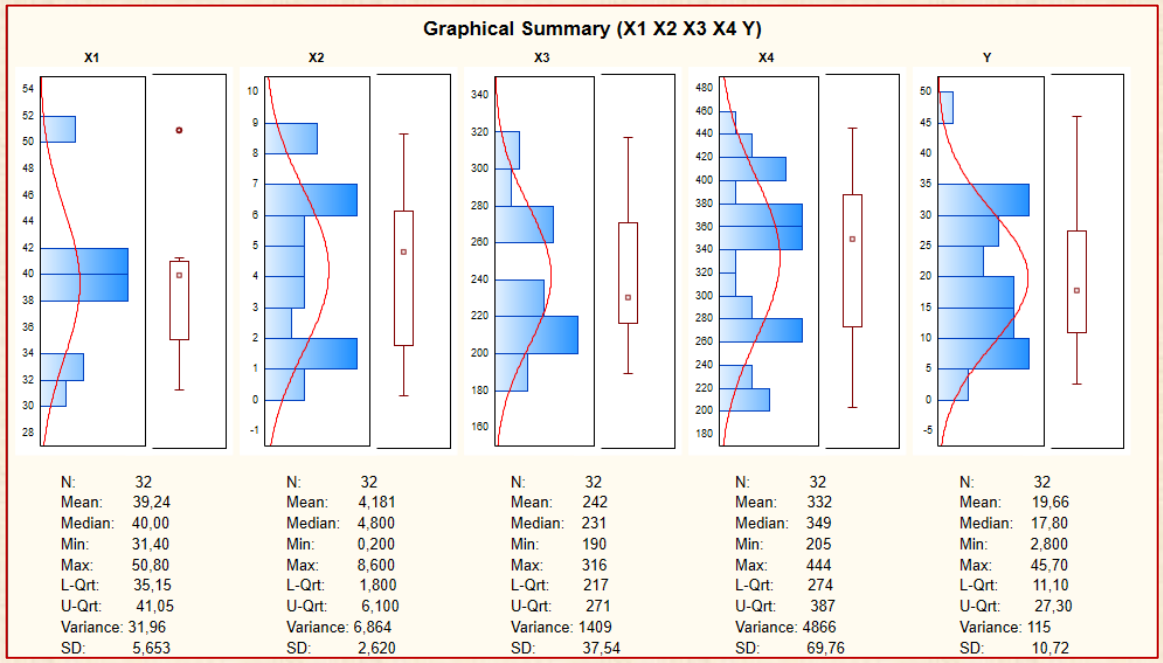


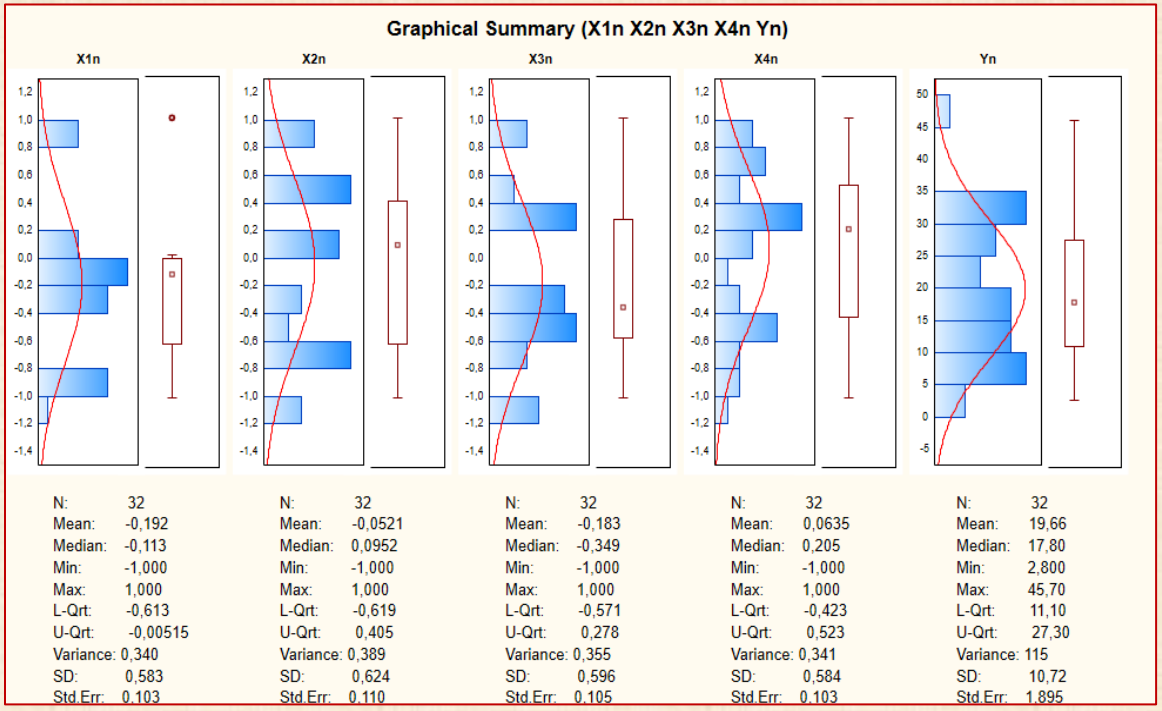
examen
des
données
et variables

variables
originelles X

vs

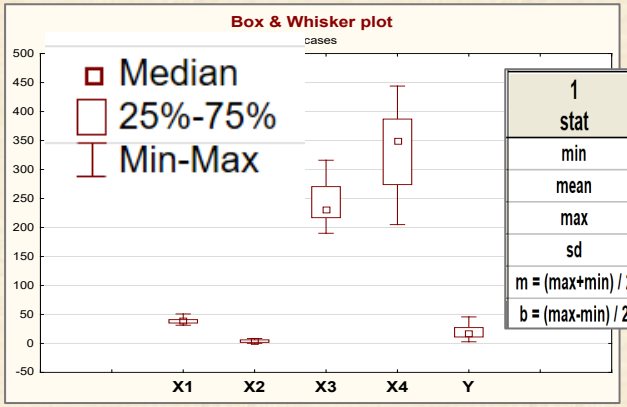
variables
normalisées Xn



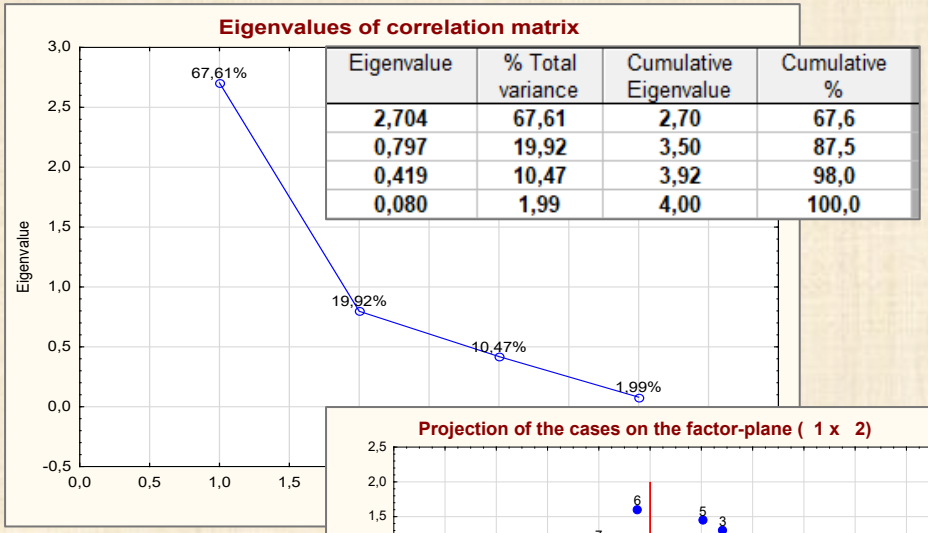


EXEMPLE RÉGRESSION MULTIPLE

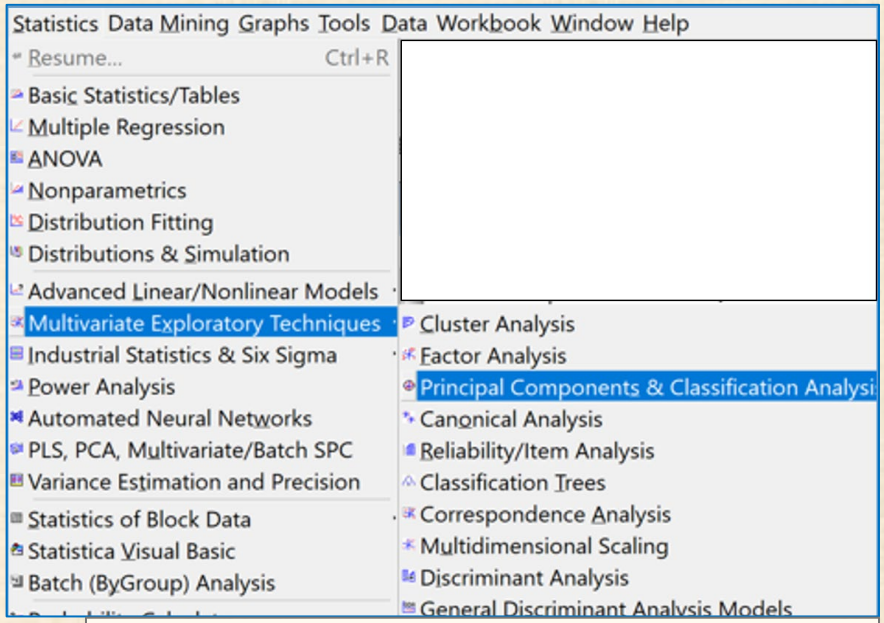
Examen et visualisation des données : Analyse en Composants Principales (ACP)

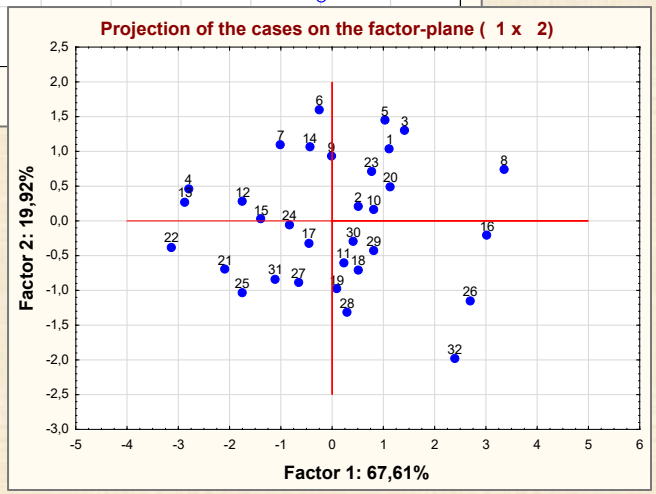


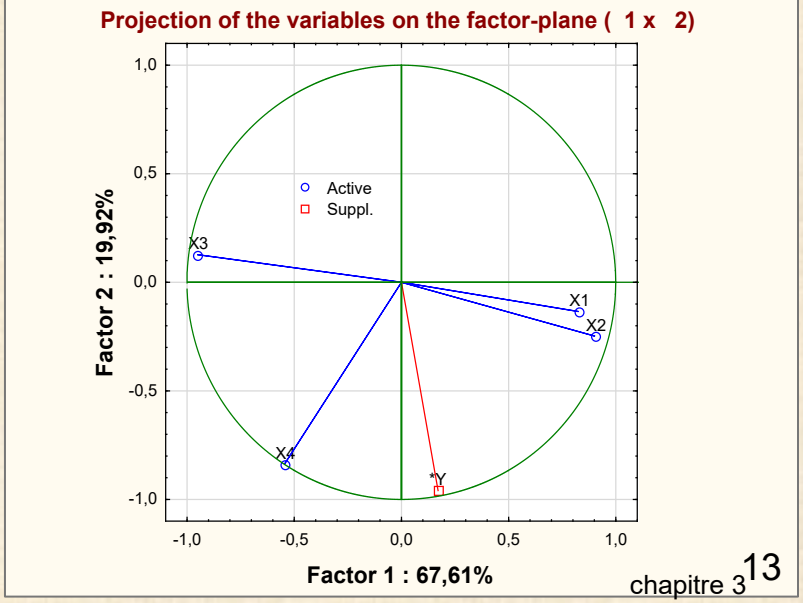
1 stat	2 X1	3 X2	4 X3	5 X4	6 Y
min	31,40	0,20	190,00	205,00	2,80
mean	39,24	4,18	241,50	332,09	19,66
max	50,80	8,60	316,00	444,00	45,70
sd	5,65	2,62	37,54	69,76	10,72
m = (max+min) / 2	41,10	4,40	253,00	324,50	24,25
b = (max-min) / 2	9,70	4,20	63,00	119,50	21,45



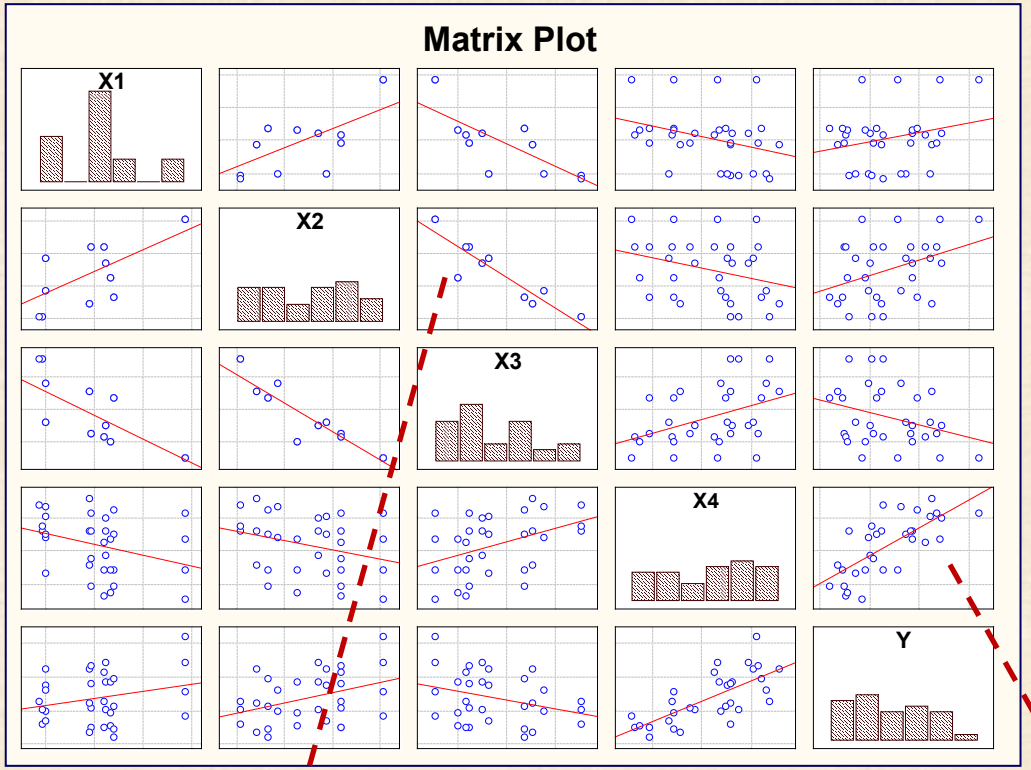
Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
2,704	67,61	2,70	67,6
0,797	19,92	3,50	87,5
0,419	10,47	3,92	98,0
0,080	1,99	4,00	100,0





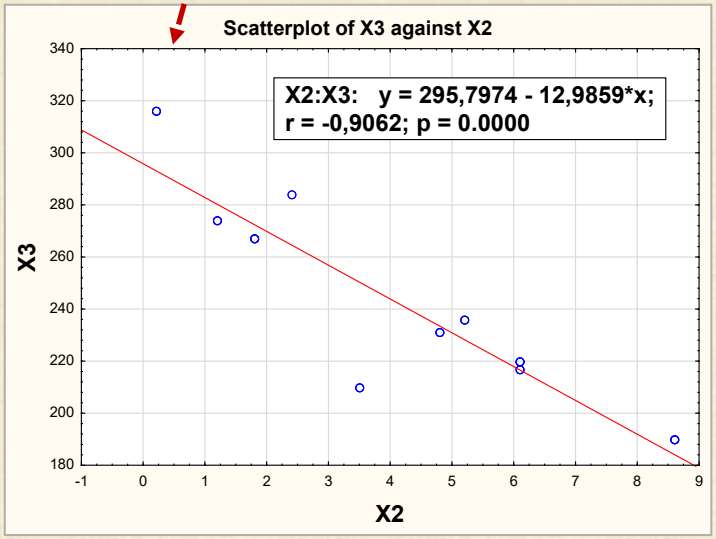


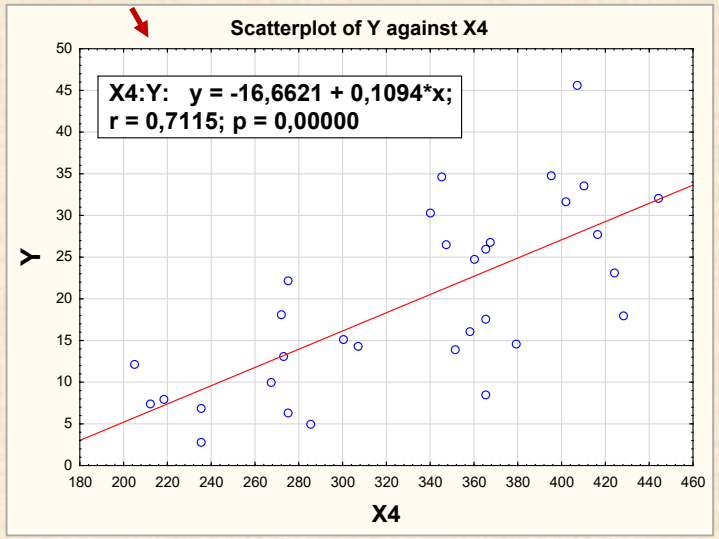
EXEMPLE RÉGRESSION MULTIPLE



matrice corrélations

	X1	X2	X3	X4	Y
X1	1.00	0.62	-0.70	-0.32	0.25
X2	0.62	1.00	-0.91	-0.30	0.38
X3	-0.70	-0.91	1.00	0.41	-0.32
X4	-0.32	-0.30	0.41	1.00	0.71
Y	0.25	0.38	-0.32	0.71	1.00





EXEMPLE RÉGRESSION MULTIPLE avec Multiple Regression

Regression Summary for Dependent Variable: Y

$R = 0.981$ $R^2 = 0.962$ Adjusted $R^2 = 0.957$

$F(4,27) = 172.06$ $p < 0.00000$ Std.Error of estimate: 2.23

	b*	Std.Err. b*	b	Std.Err. b	t(27)	p-level
Inter cept			-6.97	10.13	-0.69	0.497
X1	0.120	0.053	0.23	0.10	2.29	0.030
X2	0.136	0.090	0.56	0.37	1.50	0.144
X3	-0.522	0.102	-0.15	0.03	-5.10	0.000
X4	1.006	0.042	0.15	0.01	24.02	0.000

t(27) = b / Std.Err. (b)
= b* / Std.Err.(b*)

variables significatives p-level ≤ 0,05

test de signification des variables X

coefficients de l'équation de prédiction avec toutes les variables X et Y en format centrées-réduites :

$$Xcr = (X - \bar{X}) / s_X$$

$\bar{X}$: moyenne de X

s_X : écart-type de X

variables centrées-réduites

$\bar{X}_{cr} = 0$

$s_{XCR} = 1$

coefficients bruts de l'équation de prédiction avec toutes les variables X et Y dans leurs unités d'origine

Utilisation pour diagramme Pareto page suivante

X2 pas significative mais domaine de variation le plus faible des 4 variables X

EXEMPLE RÉGRESSION MULTIPLE

avec General Regression Models

Regression Summary for Dependent Variable: Y

$R = 0.981$ $R^2 = 0.962$ Adjusted $R^2 = 0.957$

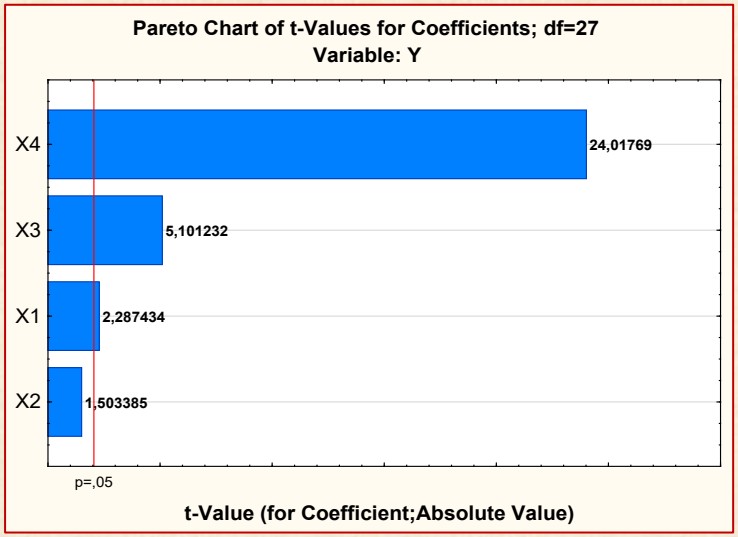
$F(4,27) = 172.06$ $p < 0.00000$ Std. Error of estimate: $2.23 = \hat{\sigma}$

	Sums of Square SS	df	Mean Square MS	F	p-level
Regress.	3429.53	4	857.38	172.06	0.000000
Residual	134.55	27	4.98		
Total	3564.08				

	DDF	Y SS	Y MS	Y F	Y p
Interce	1	2.357	2.36	0.47	0.497
"X1"	1	26.07	26.07	5.23	0.030
"X2"	1	11.26	11.26	2.26	0.144
"X3"	1	129.68	129.68	26.02	0.00002
"X4"	1	2874.54	2874.54	576.85	0.000000
Error	27	134.55	4.98		
Total	31	3564.08			

$F = t^2$

Diagramme Pareto



X4 explique

(2874 /3564) *100 = 80,6 %

de la variabilité de Y

EXEMPLE RÉGRESSION MULTIPLE

observations triées ordre décroissant résidu standard en valeur absolue	Valeur Obs Y	Valeur Préd $\hat{Y}$	Résidu r $Y - \hat{Y}$	Valeur prédite en centrée réduite	Résidu standard r / sigma	Ecart type Préd ET ($\hat{Y}$)
29 . . . . . *	30.40	25.78	4.62	0.58	2.07	0.540
10 .*	15.20	18.78	-3.58	-0.08	-1.60	0.603
2 .*	14.40	17.93	-3.53	-0.16	-1.58	0.424
19 . . . . . * .	34.90	31.54	3.36	1.13	1.50	0.617
20 . . . . . *. .	18.20	15.26	2.94	-0.42	1.32	0.603
24 . *	16.10	19.01	-2.91	-0.06	-1.30	0.731
...	...	...	...	...	...	...
12 . . . *. . . .	14.00	13.65	0.35	-0.57	0.16	0.744
26 . . . *. . . .	34.70	34.43	0.27	1.40	0.12	1.054
14 . . . *. . . .	6.40	6.17	0.23	-1.28	0.10	0.833
31 . . . * . . .	27.80	27.98	-0.18	0.79	-0.08	0.872
30 . . . *	26.60	26.64	-0.04	0.66	-0.02	1.202
Minimum .*	2.80	-0.01	-3.58	-1.87	-1.60	0.424
Maximum . . .. . *	45.70	44.02	4.62	2.32	2.07	1.223
Moyenne . . . * .	19.66	19.66	0.00	0.00	0.00	0.857

résidu standard plus grand que 3 en valeur absolue
un critère d'une observation douteuse ; autres critères à venir

analyse diagnostique après modélisation : étape absolument nécessaire pour l'examen qualité du modèle

VÉRIFICATION (« model checking »)

- variance constante ?
- données douteuses / aberrantes ? / outliers ? comment identifier ?
- 2 types de variabilité : variabilité homogène, variabilité hétérogène
- SPC (Statistical Process Control) : graphique comportement de processus

https://cours.polymtl.ca/mth6301/WEB-mth8302/mth8302-Cours&Plus/Clement-ControleStatistiqueProcessus.pdf

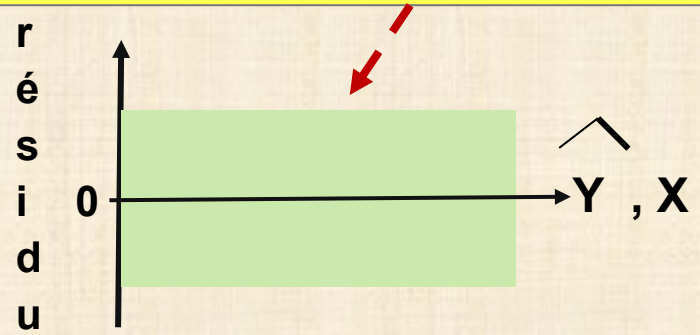
- Y: observations indépendantes ? dépend du plan de collecte
- critères de « bon » modèle ? - critères page suivante
- normalité résidus ? (surtout pour identifier données douteuses / aberrantes)

Si PROBLÈMES

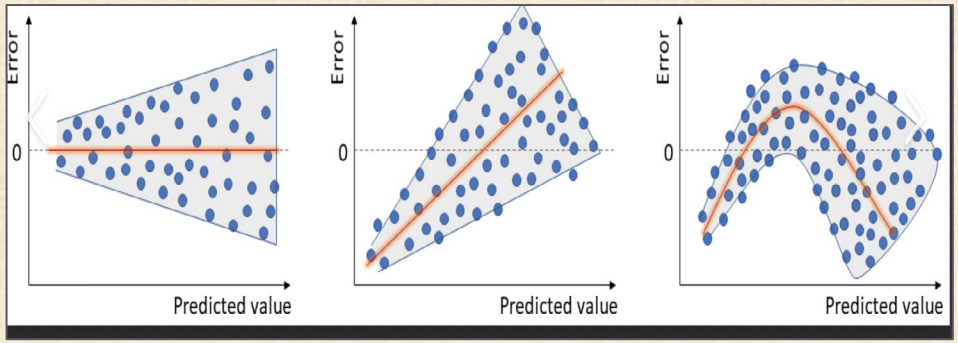
- ré examen données / élimination de données aberrantes / douteuses
- transformer Y avec la transformation Box-Cox : p. 20-21-22

critères « bon modèle »

- test global F significatif
- tests individuels significatifs : chacun des coefficients du modèle ajusté
- R^2 élevé (au moins 0.70) et R^2_{adj} légèrement inférieur à R^2
- analyse de sensibilité : pas d'observations ayant une influence prépondérante
- absence de colinéarité forte entre les variables X
- analyse des résidus : pas d'anomalies si
  - indépendance des observations de Y
  - distribution gaussienne : alignement sur un q-q plot (graphique quantile-quantile)
    pas le critère le plus important – utile pour identifier valeurs aberrantes
  - variance de Y constante : critère important
  - graphiques des résidus avec $\hat{y}$, y observés , chaque X :
    bande horizontale



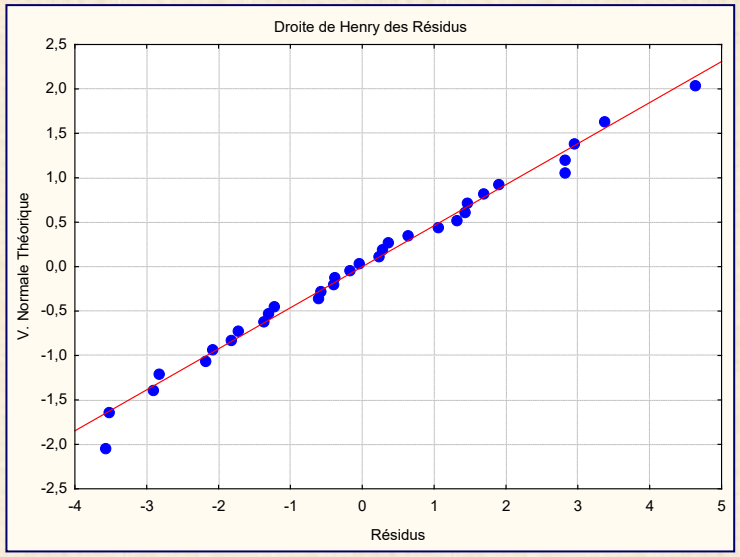
mauvais comportement



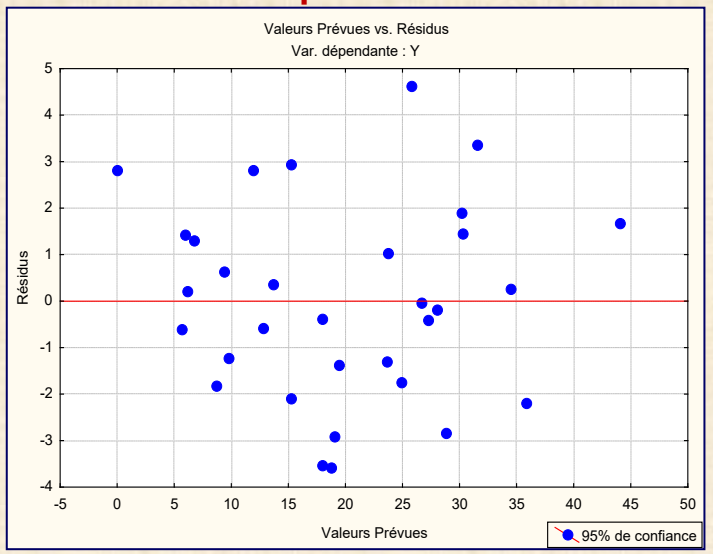
jugement visuel : il faut pratiquer !
« practice make perfect »

analyse des résidus : exemple

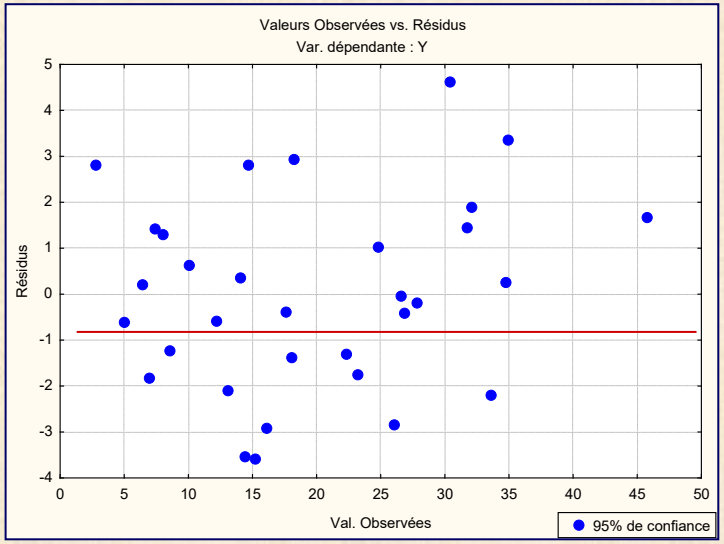
résidus sur échelle gaussienne



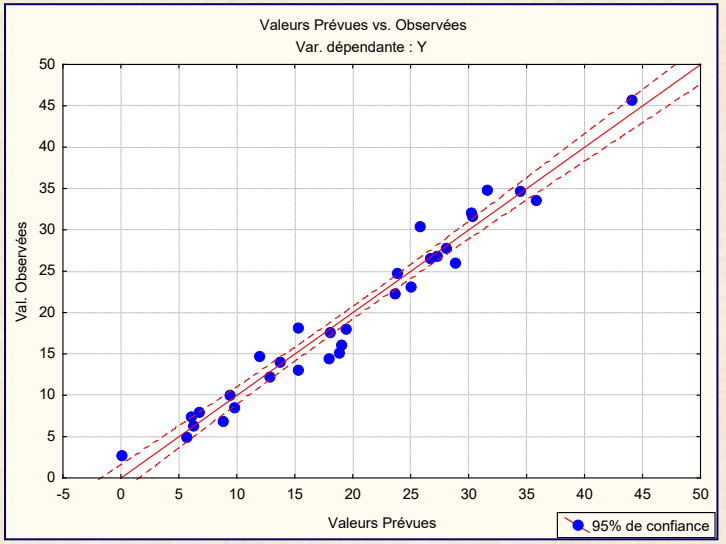
résidus vs prédictions



résidus vs observées



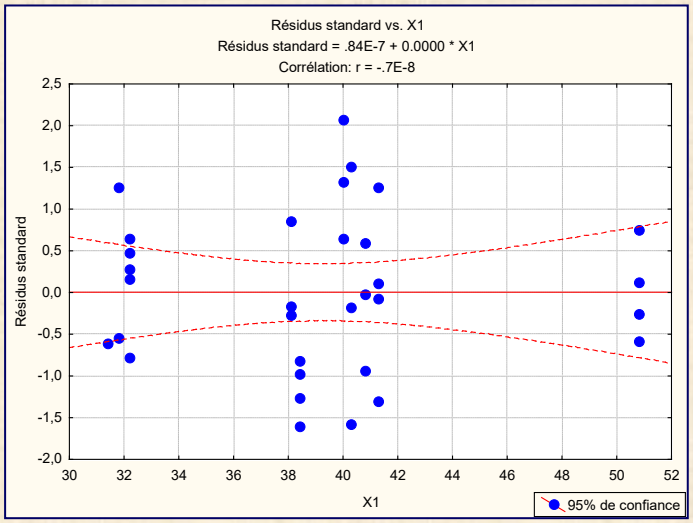
prédictions vs observées



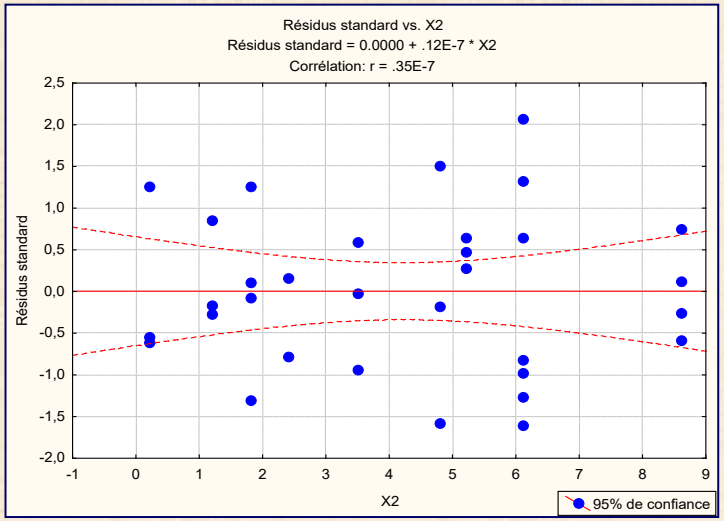
conclusion : OK ? **réponse = oui**

analyse des résidus : exemple données gazoline

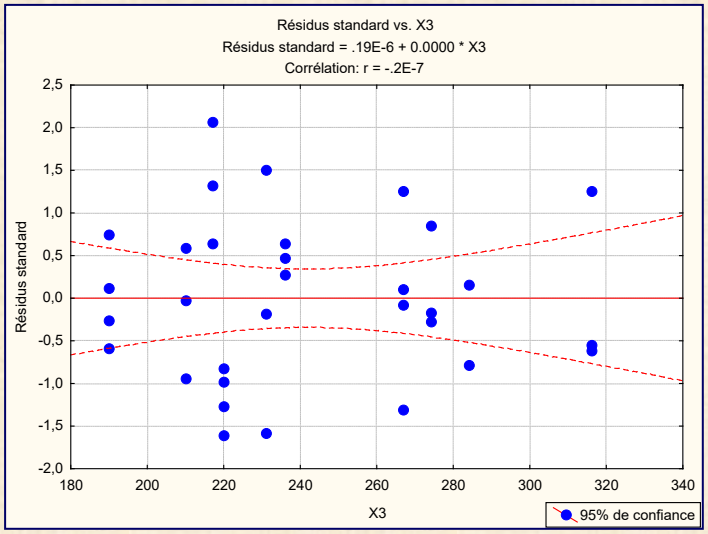
résidu vs X1



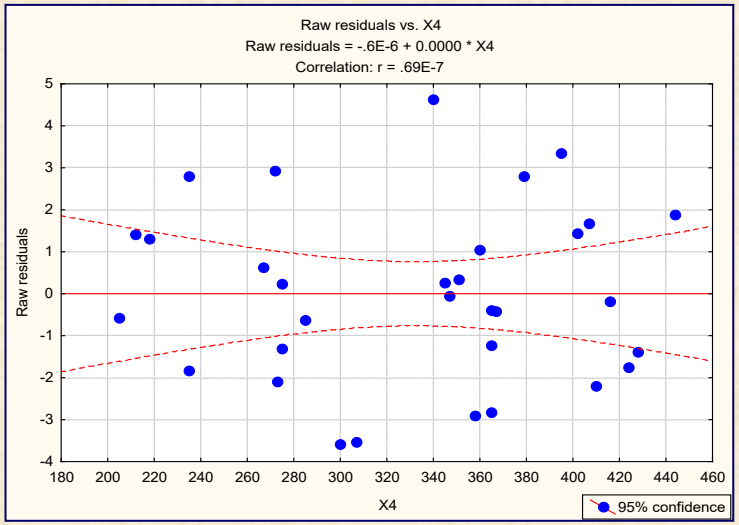
résidu vs X2



résidu vs X3



résidu vs X4

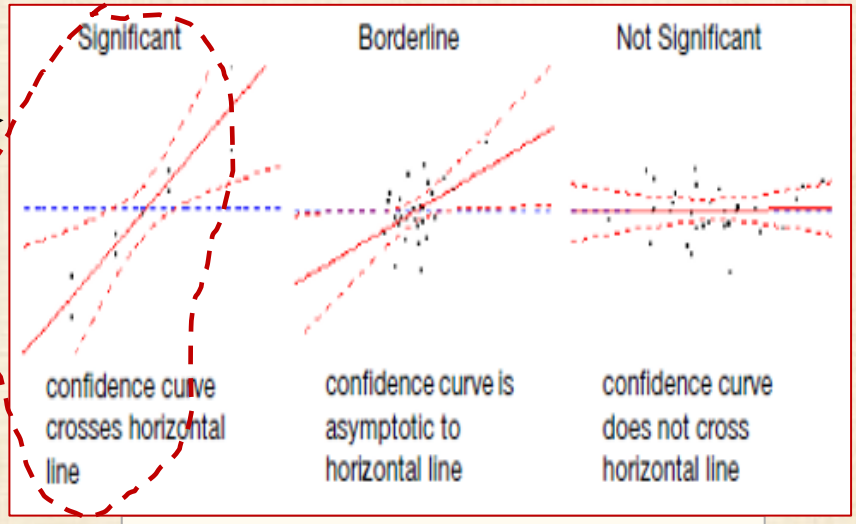


conclusion : OK ?

Residu partiel : influence de Xi

résidu avec / sans la variable Xi dans le modèle

permet de savoir si la variable est importante

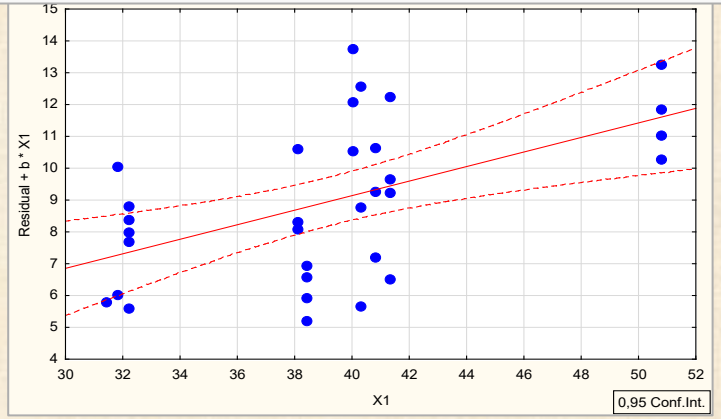


résidu partiel Xi = résidu + bi *Xi

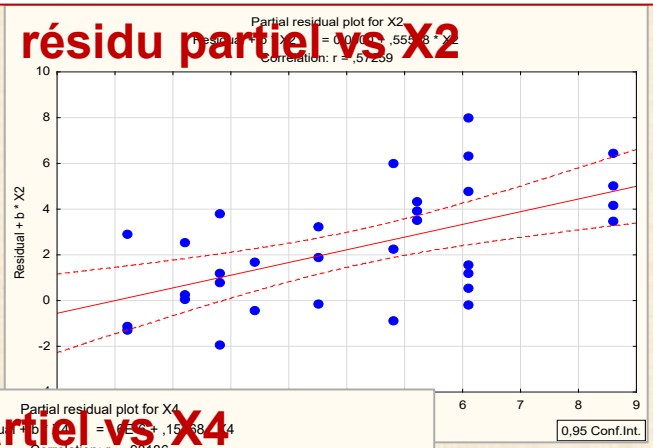
graphique : résidu partiel VS Xi

résidu partiel de X1

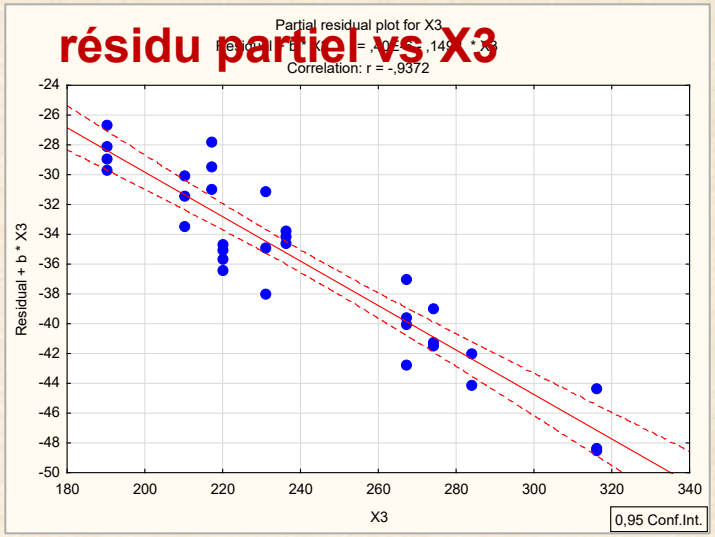
= residu + b * X1 = 0,26E-6 + 0,23 * X1



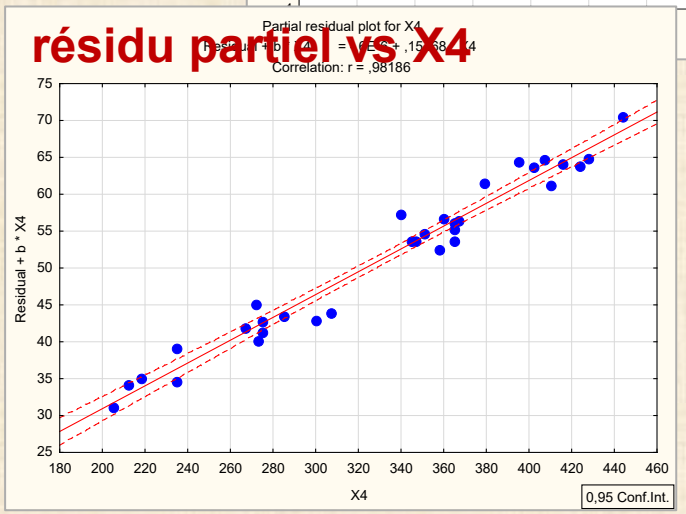
résidu partiel vs X2



résidu partiel vs X3



résidu partiel vs X4



CORRECTIFS si anomalies des résidus

- **élimination** variables colinéaires redondantes :
  méthode de sélection de variables pas à pas (stepwise)
- **ajout** termes additionnels dans modèle : X_i^2 , $X_i X_{i'}$ $(i \neq i')$
- **élimination** d'observations influentes / douteuses / aberrantes
- **ajout** de nouvelles variables explicatives : si possible ….
- **recherche** de nouveaux modèles / formes fonctionnelles

Transformation de Box-Cox de Y : transformation de puissance

nouvelle variable Yt $\quad Yt = (Y + \delta)^{**} \lambda$ si $\lambda \neq 0$
$\quad = \ln(Y + \delta)$ si $\lambda = 0$

méthode pour obtenir λ

Algorithme Box-Cox

Estimation Vraisemblance Maximale

Tansformation de Box-Cox

but = stabiliser la variance + s'approcher distribution normale

Yt nouvelle variable de transformation de puissance

$$Yt = (Y + \delta)^{**} \lambda \quad si \quad \lambda > 0$$
$$= \ln(Y + \delta) \quad si \quad \lambda = 0$$

δ : constante de translation si Y prend des valeurs négatives ou 0

valeurs commodes de λ à employer : -2 / -1,5 / -1 / 0,5 / 0 / 0.5 / 1 / 1,5 / 2

méthode pour obtenir λ : algorithme Box-Cox

Algorithme Box-Cox data : y1, y2,…, yn transformation y_{ti}

$$Y_{ti} = [(yi + \delta)^{**} \lambda - 1] / [\lambda G^{**}(\lambda - 1)] \quad si \quad \lambda \neq 0 \quad i = 1, 2, \ldots, n$$
$$= G * \ln(yi + \delta) \quad si \quad \lambda = 0$$

$G = [\Pi (yi + \delta)]^{**}1/n$ = moyenne géométrique des yi

ET(W) : notation pour représenter l'Écart-Type d'une variable W

facteur G : pour que ET(Yt) soit approximativement égal à ET(Y)

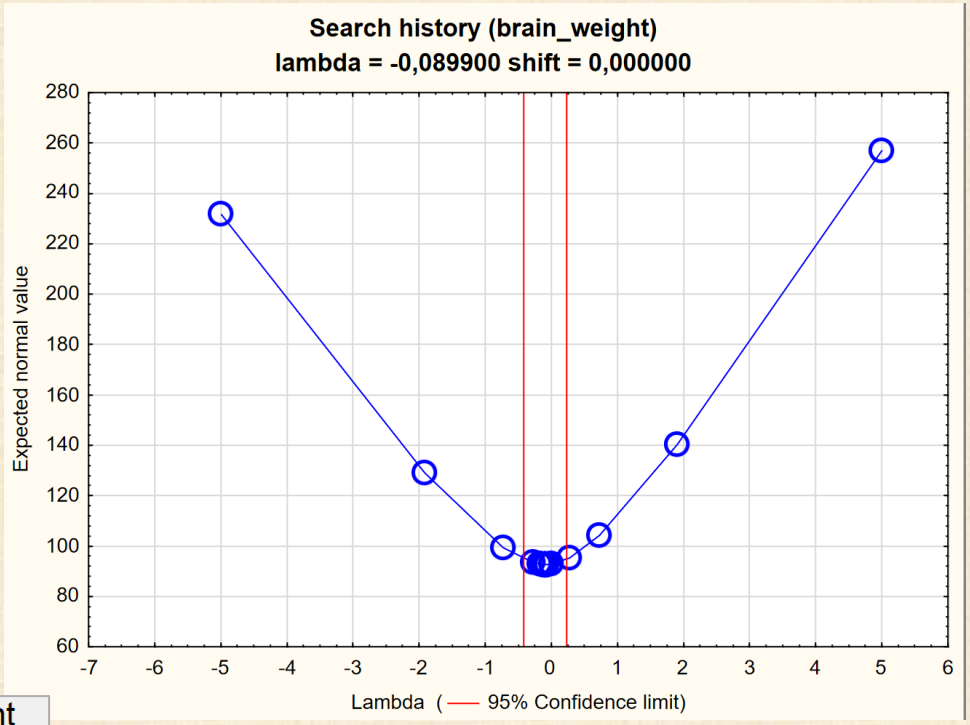
comparaison des ET(Yt) à différentes valeurs de λ

Estimation de λ : V(λ) fonction vraisemblance de λ V (λ) = constante / ET(Yt)

maximum de V s'obtient avec minimum de ET(Yt)

Exemple : transformation Box-Cox

2 nom_mamifère	3 body_weight	4 brain_weight	5 percent
African Elephant	6654	5712	0,09
Asian Elephant	2547	4603	0,18
Giraffe	529	680	0,13
Horse	521	655	0,13
Cow	465	423	0,09
Gorilla	207	406	0,20
Pig	192	180	0,09
Jaguar	100	157	0,16
Man	62	1320	2,13
Chimpanzee	52	440	0,85
Gray Wolf	36	120	0,33
Kangaroo	35	56	0,16
Baboon	11	179	1,63
Red Fox	4	50	1,25
Cat	3	26	0,87



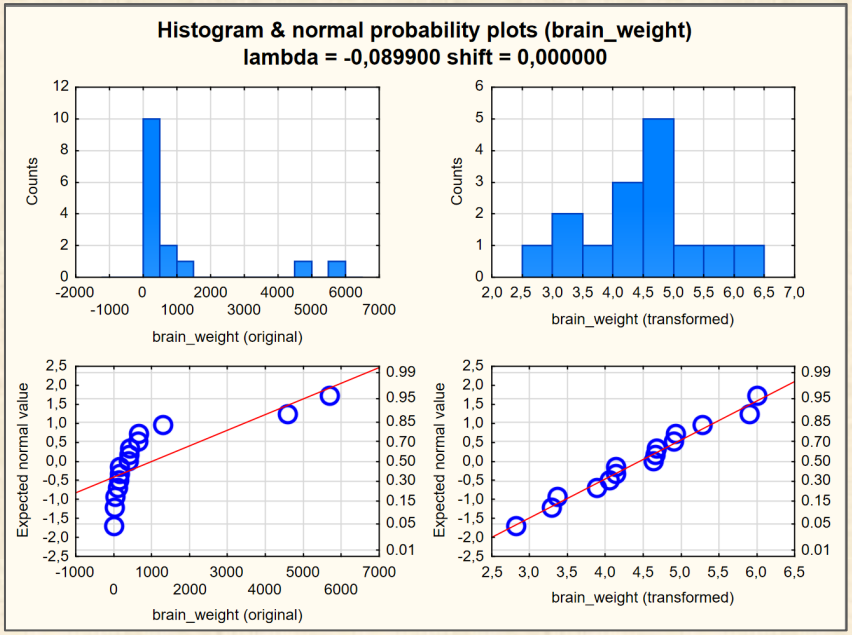
Formula for Box-Cox transformation
((v4^(-0,089900))-1)/(-0,089900)

brain_weight original	brain_weight transformed
5712	6,01
4603	5,91
680	4,93
655	4,91
423	4,66
406	4,64
180	4,15
157	4,06
1320	5,29
440	4,69
120	3,89
56	3,38
179	4,15
50	3,30
26	2,82

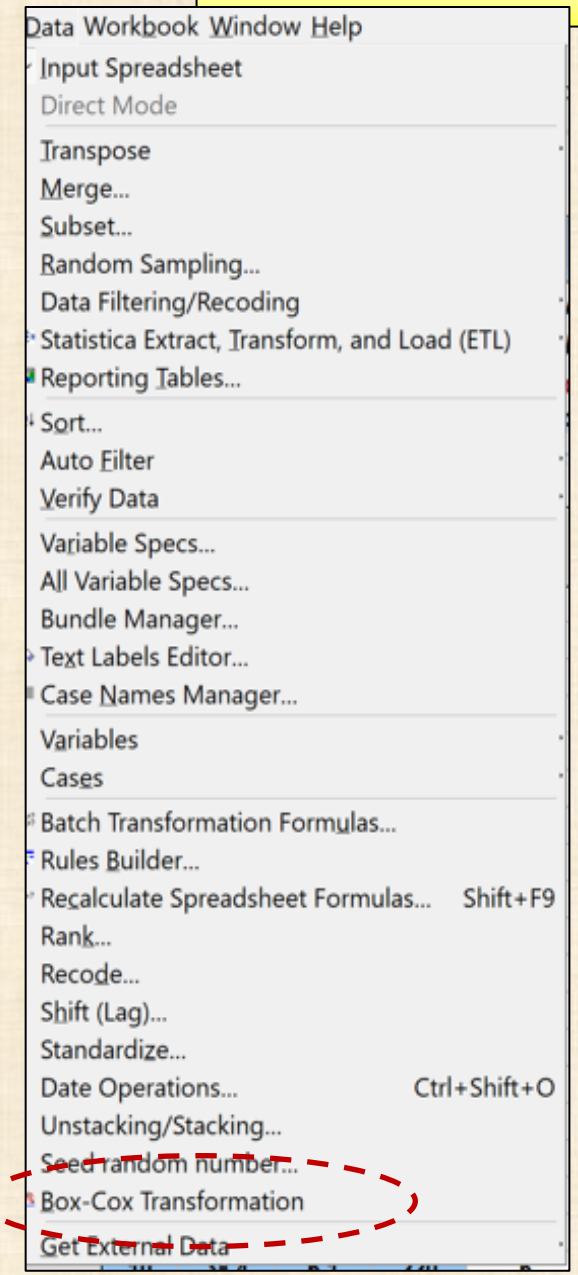
recommendation

Yt = log(Y)

valeur optimale
λ = - 0,0899
remplacée par 0

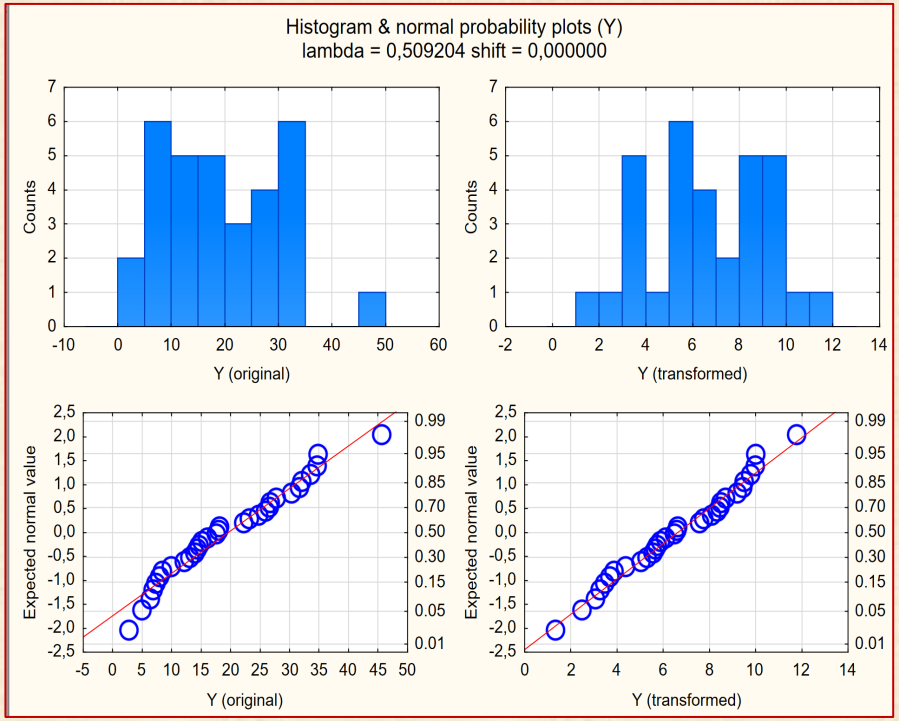


Exemple 2 : voir Data …. Box-Cox Transformation
data = gazoline variable Y



Y original	Y transformed
14,40000	5,67365
7,40000	3,47772
8,50000	3,87560
8,00000	3,69809
2,80000	1,35359
5,00000	2,49299
12,20000	5,05534
10,00000	4,37941
15,20000	5,88684
26,80000	8,51516
14,00000	5,56487
14,70000	5,75426
6,40000	3,08995
17,60000	6,49533
22,30000	7,57884
24,80000	8,10938
26,00000	8,35470
34,90000	10,02345
18,20000	6,64097
23,20000	7,77304
18,00000	6,59269
13,10000	5,31441
16,10000	6,12020
32,10000	9,52369
34,70000	9,98842
31,70000	9,45057
33,60000	9,79396
30,40000	9,20976
26,60000	8,47527
27,80000	8,71248
45,70000	11,78747

Transformed variable(s)	Data statistics Lambda
Y	0,509204



recommendation : pas de changement

distinction : effet significatif ne veut pas dire effet important

RELATION entre F et R^2 R^2 = df_modèle * F_modèle / (df_modèle * F_modèle + df_erreur)

F = df_erreur * R^2 / (df_modèle * (1 – R^2))

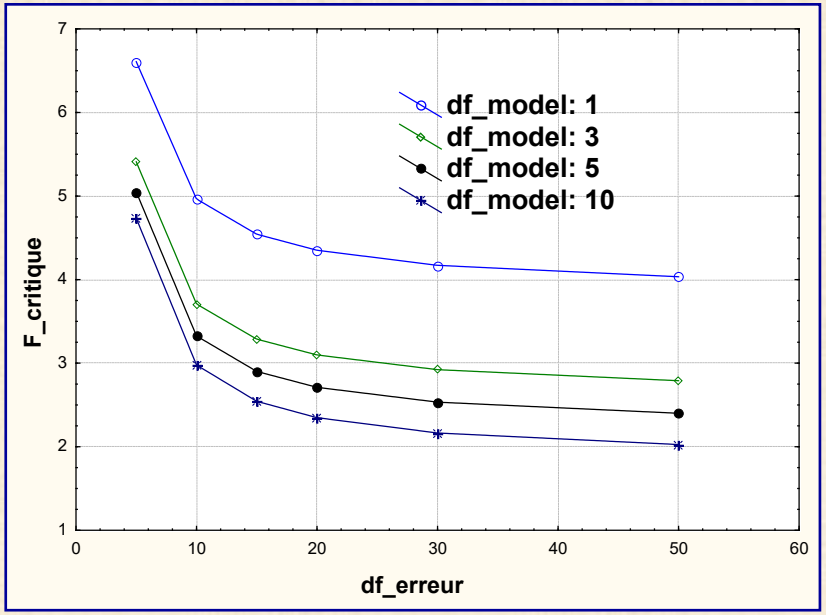
si F_o (modèle) ≥ $F_{m, N-m-1, 0,95}$ (F_critique à 5%) : on REJETTE l'hypothèse nulle

cela n'implique pas nécessairement un R^2 élevé : voir colonne 4 du tableau

distinction à faire : effet SIGNIFICATIF et effet IMPORTANT (= explique 50% et plus disons)

ratio F doit être au moins 4 fois plus grand (F_gros = 4* $F_{m, N-m-1, 0,95}$) $\alpha = 0,05$ pour que $R^2 \geq 50\%$ c-a-d α doit être plus petite que 0,001 selon le tableau

df modèle	df erreur	F critique α = 0,05	R2 α = 0,05	F gros 4 fois F critique	R2 F gros	α F gros
1	5	6,61	0,57	26,43	0,84	0,0036
1	10	4,96	0,33	19,86	0,67	0,0012
1	15	4,54	0,23	18,17	0,55	0,0007
1	20	4,35	0,18	17,4	0,47	0,0005
1	30	4,17	0,12	16,68	0,36	0,0003
1	50	4,03	0,07	16,14	0,24	0,0002
3	5	5,41	0,76	21,64	0,93	0,0027
3	10	3,71	0,53	14,83	0,82	0,0005
3	15	3,29	0,4	13,15	0,72	0,0002
3	20	3,1	0,32	12,39	0,65	0,00008
3	30	2,92	0,23	11,69	0,54	0,00003
3	50	2,79	0,14	11,16	0,4	0,00001
5	5	5,05	0,83	20,2	0,95	0,0025
5	10	3,33	0,62	13,3	0,87	0,0004
5	15	2,9	0,49	11,61	0,79	0,0001
5	20	2,71	0,4	10,84	0,73	0,00004
5	30	2,53	0,3	10,13	0,63	0,000009
5	50	2,4	0,19	9,6	0,49	0,000001
10	5	4,74	0,9	18,94	0,97	0,0027
10	10	2,98	0,75	11,91	0,92	0,00027
10	15	2,54	0,63	10,17	0,87	0,00005
10	20	2,35	0,54	9,39	0,82	0,00001
10	30	2,16	0,42	8,66	0,74	0,000002
10	50	2,03	0,29	8,1	0,62	0,000001



recommandation Clément

effet est important si p-value est très petit disons p-value < 0,001

EXEMPLE RÉGRESSION MULTIPLE

Exemple: production de gazoline avec huiles brutes (données historiques)

N. H. Prater, *Petroleum Refiner - Experimental Designs in Industry* (ed. Chew) Wiley 1956 pp109-137

Y : rendement production gazoline (% de l'huile brute)

X1 : gravité huile brute (deg. API) X2 : pression vapeur (PSIA)

X3 : ASTM point 10% (deg. F) X4 : point sortie gazoline (deg. F)

#	X1	X2	X3	X4	Y
1	38.4	6.1	220	235	6.9
2	40.3	4.8	231	307	14.4
3	40.0	6.1	217	212	7.4
4	31.8	0.2	316	365	8.5
5	40.8	3.5	210	218	8.0
6	41.3	1.8	267	235	2.8
7	38.1	1.2	274	285	5.0
8	50.8	8.6	190	205	12.2
9	32.2	5.2	236	267	10.0
10	38.4	6.1	220	300	15.2
11	40.3	4.8	231	367	26.8
12	32.2	2.4	284	351	14.0
13	31.8	0.2	316	379	14.7
14	41.3	1.8	267	275	6.4
15	38.1	1.2	274	365	17.6
16	50.8	8.6	190	275	22.3
17	32.2	5.2	236	360	24.8
18	38.4	6.1	220	365	26.0
19	40.3	4.8	231	395	34.9
20	40.0	6.1	217	272	18.2
21	32.2	2.4	284	424	23.2
22	31.4	0.2	316	428	18.0
23	40.8	3.5	210	273	13.1
24	41.3	1.8	267	358	16.1
25	38.1	1.2	274	444	32.1
26	50.8	8.6	190	345	34.7
27	32.2	5.2	236	402	31.7
28	38.4	6.1	220	410	33.6
29	40.0	6.1	217	340	30.4
30	40.8	3.5	210	347	26.6
31	41.3	1.8	267	416	27.8
32	50.8	8.6	190	407	45.7

structure dans ces données?

réponse

oui

EXEMPLE RÉGRESSION MULTIPLE

#	X1	X2	X3	groupe	X4	Y
8	50,8	8,6	190	1	205	12,2
16	50,8	8,6	190	1	275	22,3
26	50,8	8,6	190	1	345	34,7
32	50,8	8,6	190	1	407	45,7
6	41,3	1,8	267	2	235	2,8
14	41,3	1,8	267	2	275	6,4
24	41,3	1,8	267	2	358	16,1
31	41,3	1,8	267	2	416	27,8
5	40,8	3,5	210	3	218	8,0
23	40,8	3,5	210	3	273	13,1
30	40,8	3,5	210	3	347	26,6
2	40,3	4,8	231	4	307	14,4
11	40,3	4,8	231	4	367	26,8
19	40,3	4,8	231	4	395	34,9
3	40,0	6,1	217	5	212	7,4
20	40,0	6,1	217	5	272	18,2
29	40,0	6,1	217	5	340	30,4
1	38,4	6,1	220	6	235	6,9
10	38,4	6,1	220	6	300	15,2
18	38,4	6,1	220	6	365	26,0
28	38,4	6,1	220	6	410	33,6
7	38,1	1,2	274	7	285	5,0
15	38,1	1,2	274	7	365	17,6
25	38,1	1,2	274	7	444	32,1
9	32,2	5,2	236	8	267	10,0
17	32,2	5,2	236	8	360	24,8
27	32,2	5,2	236	8	402	31,7
12	32,2	2,4	284	9	351	14,0
21	32,2	2,4	284	9	424	23,2
4	31,8	0,2	316	10	365	8,5
13	31,8	0,2	316	10	379	14,7
22	31,4	0,2	316	10	428	18,0

Données : structure de groupe

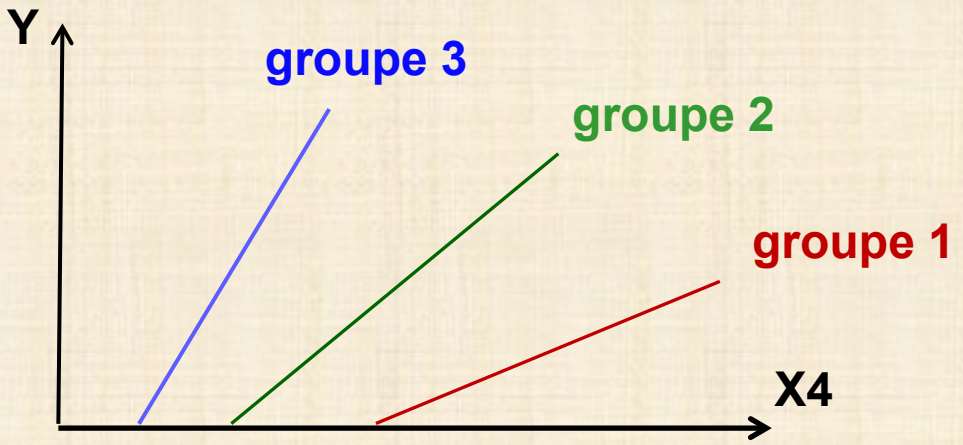
10 groupes d'huile brute définis par X1 X2 X3

signification du groupe ? provenance ? autre ?

régression de Y sur X4 : X4 rôle de covariable

1 modèle pour chaque groupe

analyse de covariance : présence de variables continues + variables catégoriques



modèle 1: pentes distinctes - Général

modèle 2: pente égales - ANCOVA

modèle d'analyse de covariance

ANALYSE COVARIANCE : variable continue + variable catégorique

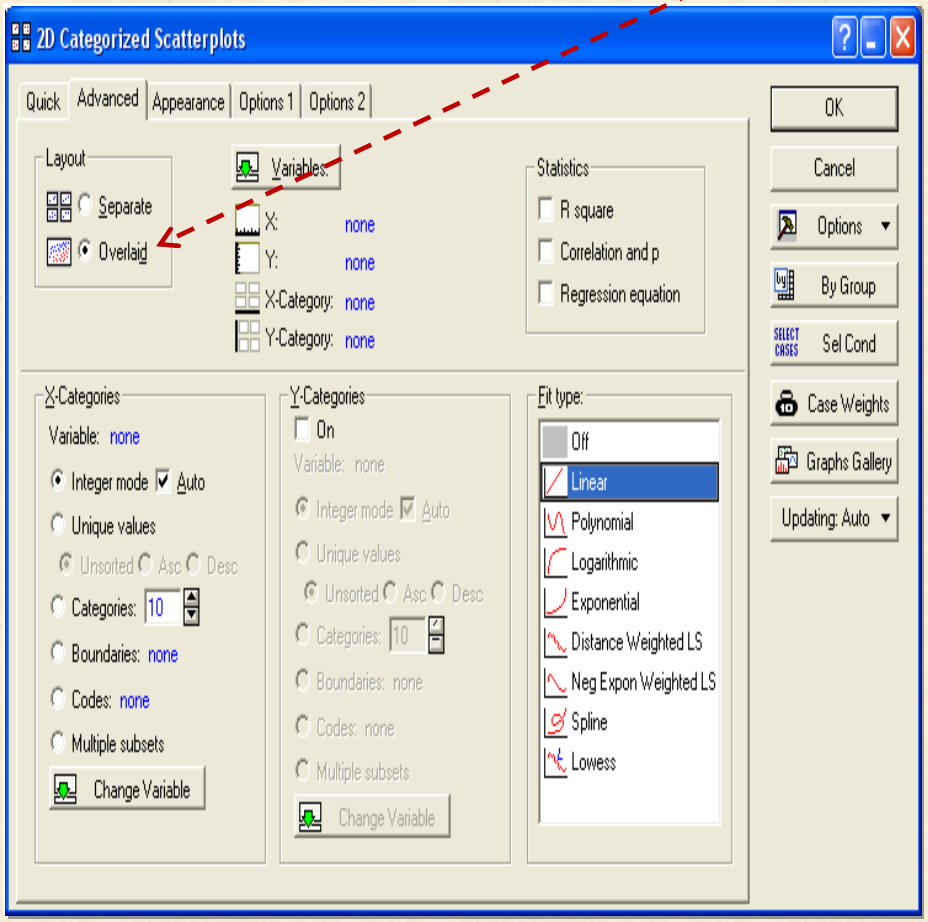
Utilisation de *STATISTICA*

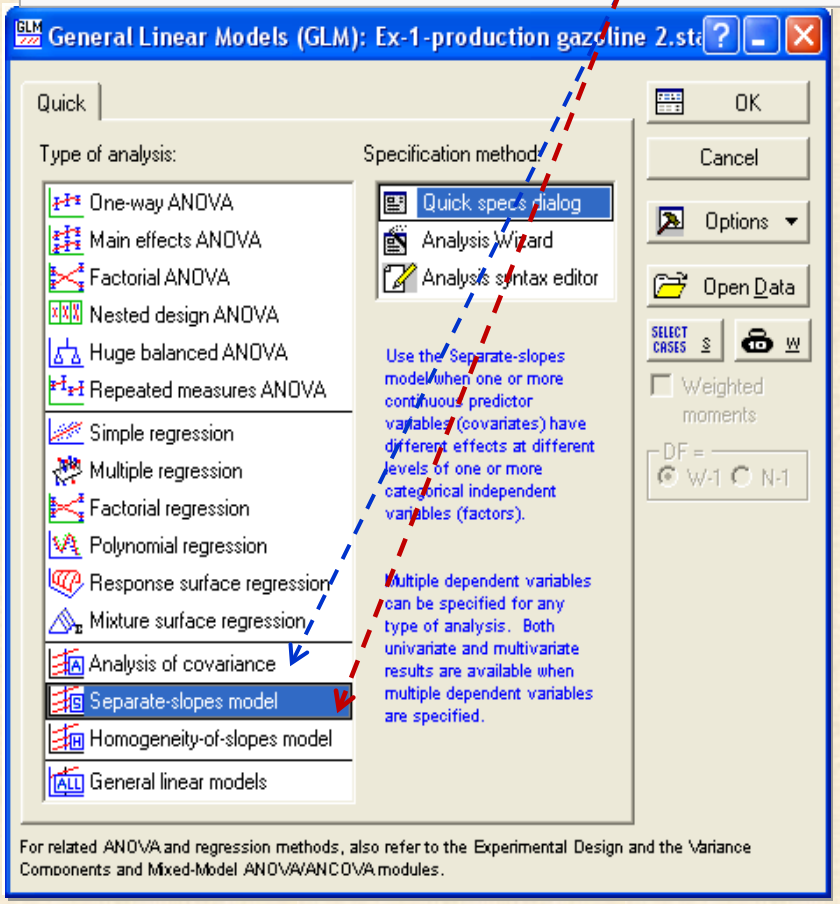
Graphs …

2D Categorized Scatterplots …. Overlaid

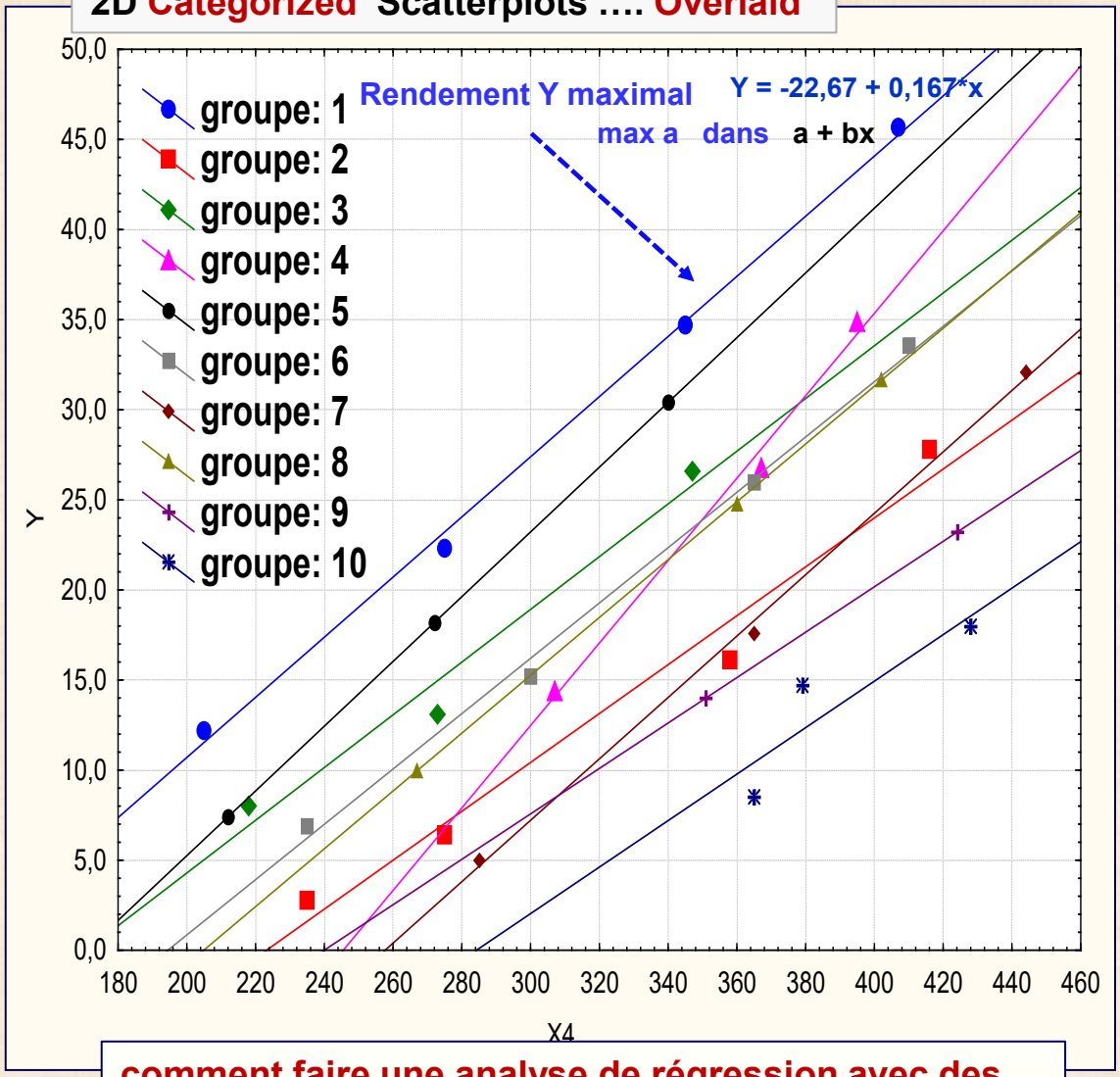
Statistics …

Advanced Linear / Nonlinear Models …
General Linear Models (GLM) …





ANALYSE COVARIANCE : groupes distincts?
après avoir enlevé l'effet de X4



comment faire une analyse de régression avec des variables continues + variables catégoriques ?

groupe: 1 Y = -22,67 + 0,167*x

r = 0,999; p = 0,0012; r^2 = 0,99

groupe: 2 Y = -30,29 + 0,136*x

r = 0,9875; p = 0,0125; r^2 = 0,98

groupe: 3 Y = -24,97 + 0,146*x

r = 0,9855; p = 0,1084; r^2 = 0,97

groupe: 4 Y = -56,16 + 0,229*x

r = 0,9963; p = 0,0550; r^2 = 0,99

groupe: 5 Y = -30,69 + 0,180*x

r = 1,0000; p = 0,0006; r^2 = 1,0000

groupe: 6 Y = -29,86 + 0,153*x

r = 0,9979; p = 0,0021; r^2 = 0,99

groupe: 7 Y = -43,91 + 0,170*x

r = 0,9990; p = 0,0281; r^2 = 0,998

groupe: 8 Y = -32,88 + 0,160*x

r = 1,0000; p = 0,0048; r^2 = 0,9999

groupe: 9 Y = -30,24 + 0,126*x

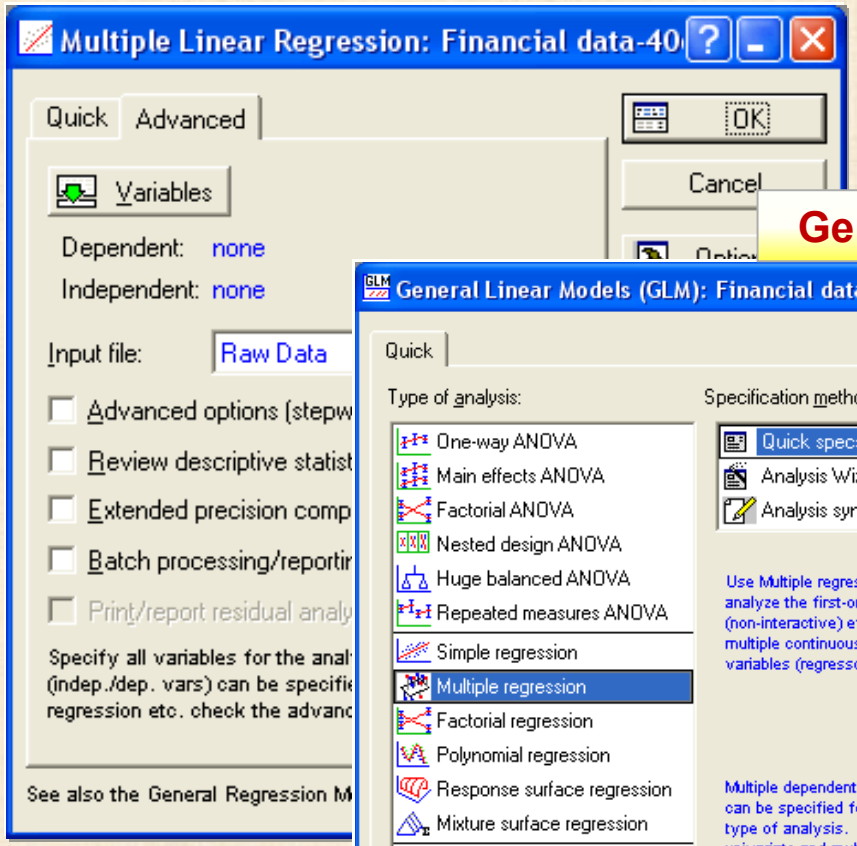
r = 1,0000; p = ---; r^2 = 1,0000

groupe: 10 Y = -36,66 +0,129*x

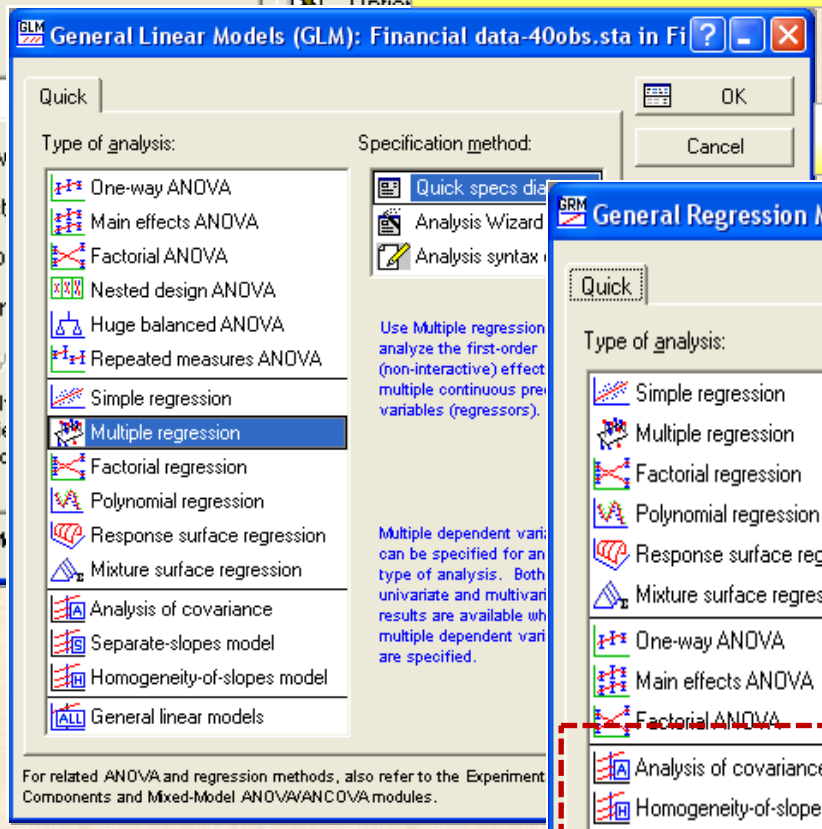
r = 0,8848; p =0,0387; r^2 = 0,78

ANALYSE COVARIANCE

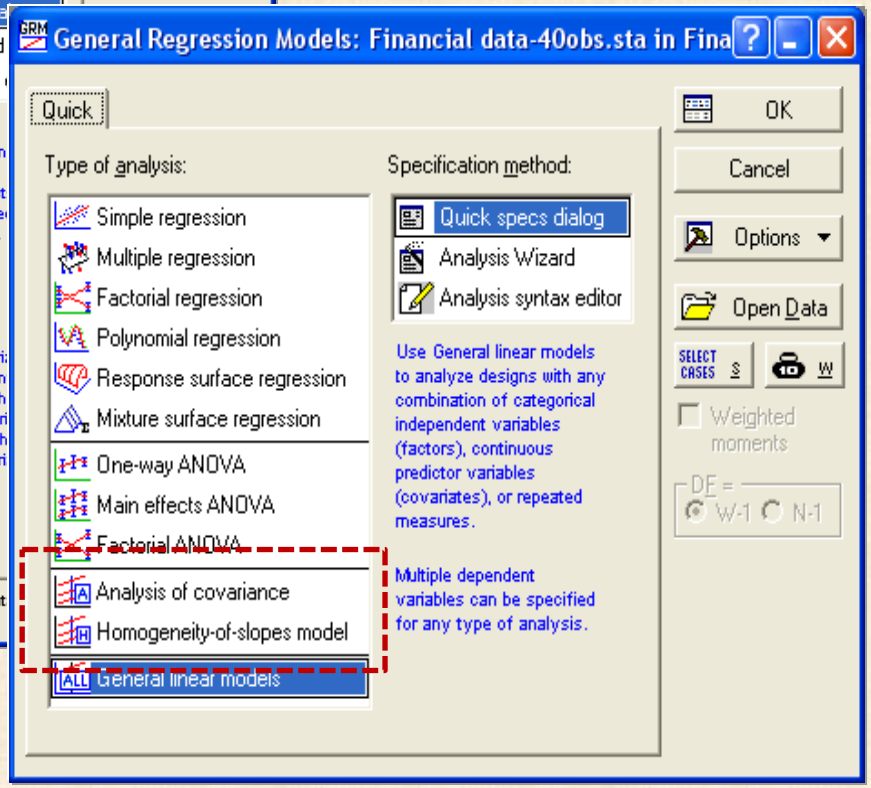
Multiple Linear Regression



General Linear Models (GLM)



General Regression Models



EXEMPLE RÉGRESSION MULTIPLE

Financial data of 40 UK companies 1983

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

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

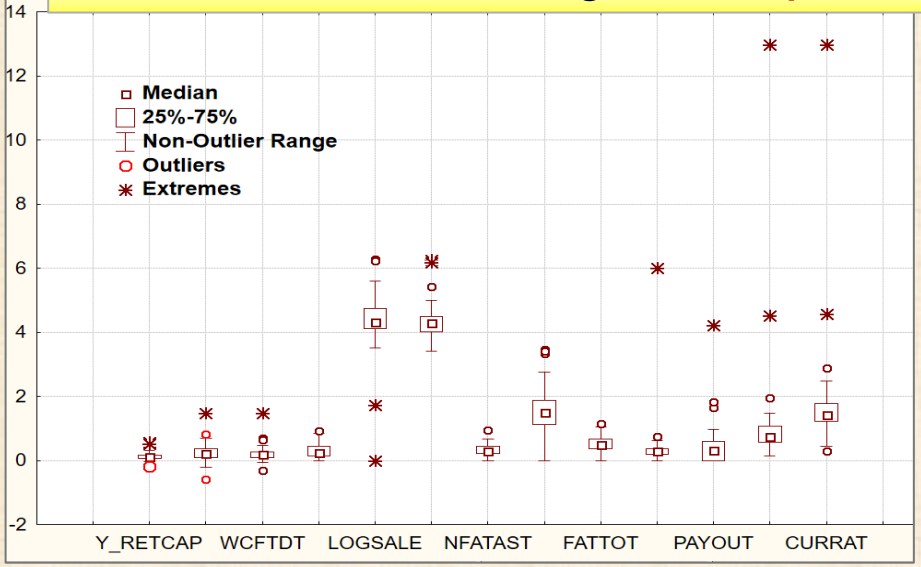
données : Financial data-40obs.sta

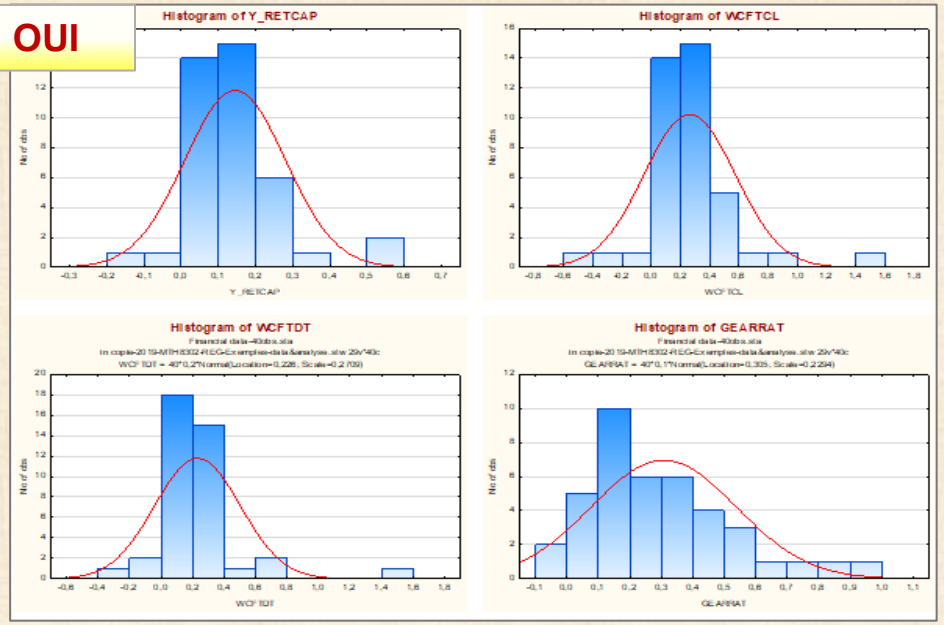
	Y RETCAP	WCFTCL	WCFTDT	GEARRAT	LOG SALE	LOG ASST	NFATAST	CAPINT	FATTOT	INVTAST	PAYOUT	QUIKRAT	CURRAT
1	0,26	0,25	0,25	0,46	4,11	4,30	0,10	0,64	0,12	0,74	0,07	0,18	1,53
2	0,57	0,33	0,33	0,00	4,25	4,00	0,12	1,79	0,15	0,27	0,30	1,26	1,73
3	0,09	0,50	0,20	0,24	4,44	4,88	0,94	0,36	0,97	0,01	0,57	0,39	0,44
.	.	.	.	.	.	.	.	.	.	.	.	.	.
40	0,13	0,30	0,20	0,23	4,54	4,37	0,32	1,49	0,46	0,28	0,58	1,47	2,49

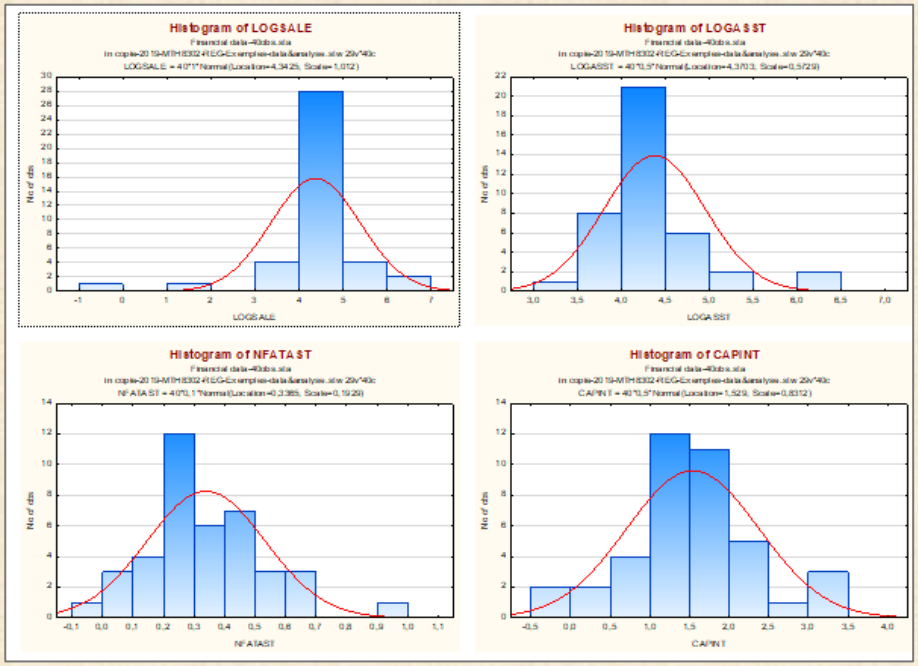
visualisation des données

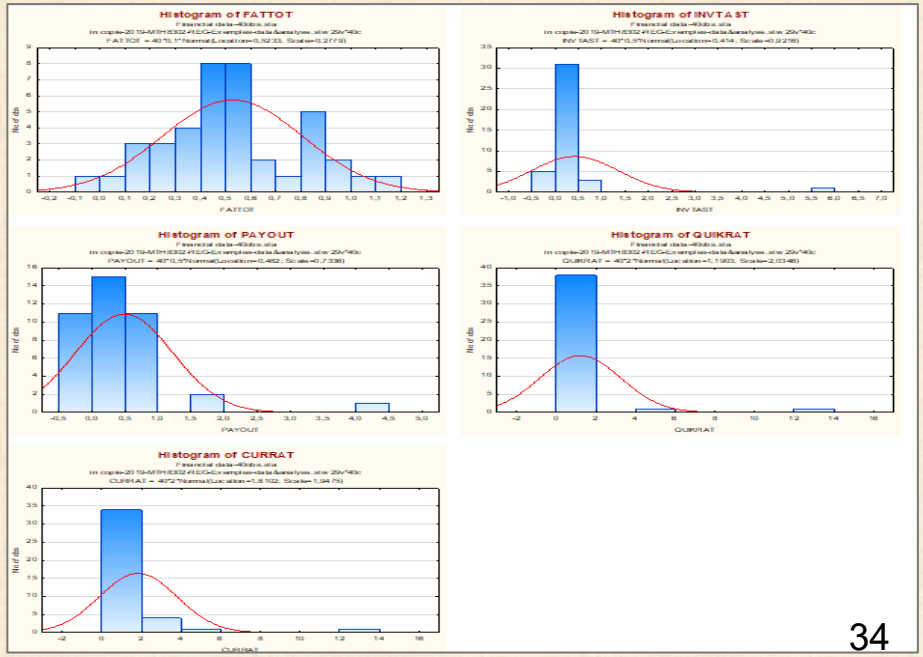
EXEMPLE RÉGRESSION MULTIPLE

données aberrantes ? hétérogénéité? réponse : OUI





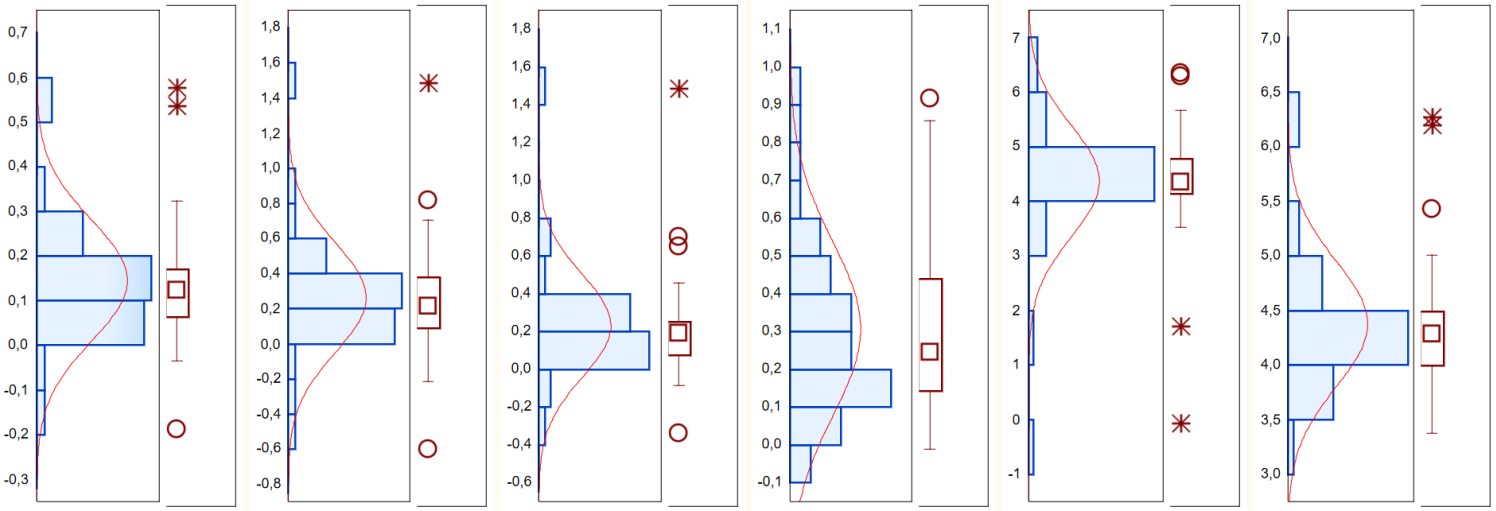




visualisation des données
Avec
Basic Statistics
and Tables

- Basic Statistics/Tables (Financial data-40obs.sta in 2021-MTH8302-Exemples-RE
  - Descriptive statistics dialog
    - Graphical Summary (Y_RETCAP WCFTCL WCFTDT GEARRAT LOGSALE...)
    - Graphical Summary (NFATAST CAPINT FATTOT INVTAST PAYOUT...)
    - Graphical Summary (CURRAT)

Graphical Summary (Y_RETCAP WCFTCL WCFTDT GEARRAT LOGSALE...)

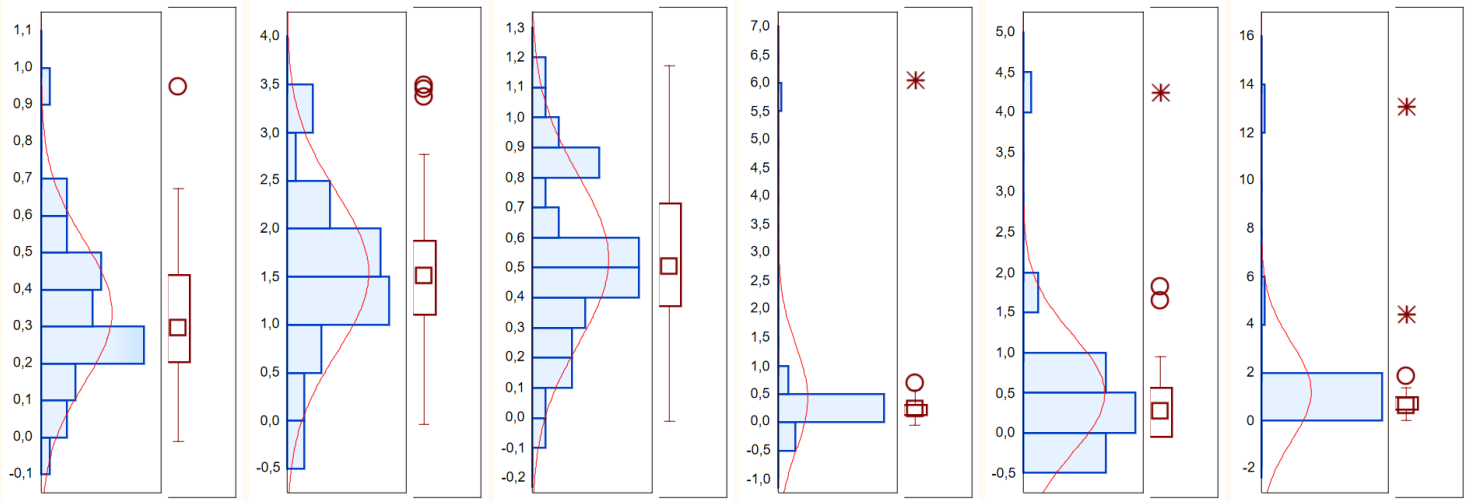


	Y_RETCAP	WCFTCL	WCFTDT	GEARRAT	LOGSALE	LOGASST
N:	40	40	40	40	40	40
Mean:	0,143	0,257	0,226	0,305	4,343	4,370
Median:	0,125	0,220	0,200	0,250	4,325	4,300
Min:	-0,180	-0,580	-0,320	0	0	3,410
Max:	0,570	1,470	1,470	0,910	6,290	6,250
L-Qrt:	0,0650	0,0950	0,0850	0,150	4,120	4,015
U-Qrt:	0,170	0,380	0,260	0,440	4,745	4,500
Variance:	0,0182	0,0971	0,0734	0,0526	1,024	0,328
SD:	0,135	0,312	0,271	0,229	1,012	0,573
Std.Err:	0,0213	0,0493	0,0428	0,0363	0,160	0,0906
Skw:	1,142	1,195	2,558	0,909	-2,143	1,733
Kurt:	3,214	5,901	11,14	0,454	9,117	4,070
95% Conf SD						
Lower:	0,110	0,255	0,222	0,188	0,829	0,469
Upper:	0,173	0,400	0,348	0,295	1,299	0,736
95% Conf Mean						
Lower:	0,0997	0,158	0,139	0,232	4,019	4,187
Upper:	0,186	0,357	0,313	0,378	4,666	4,553

visualisation des données avec Basic Statistics and Tables .... suite

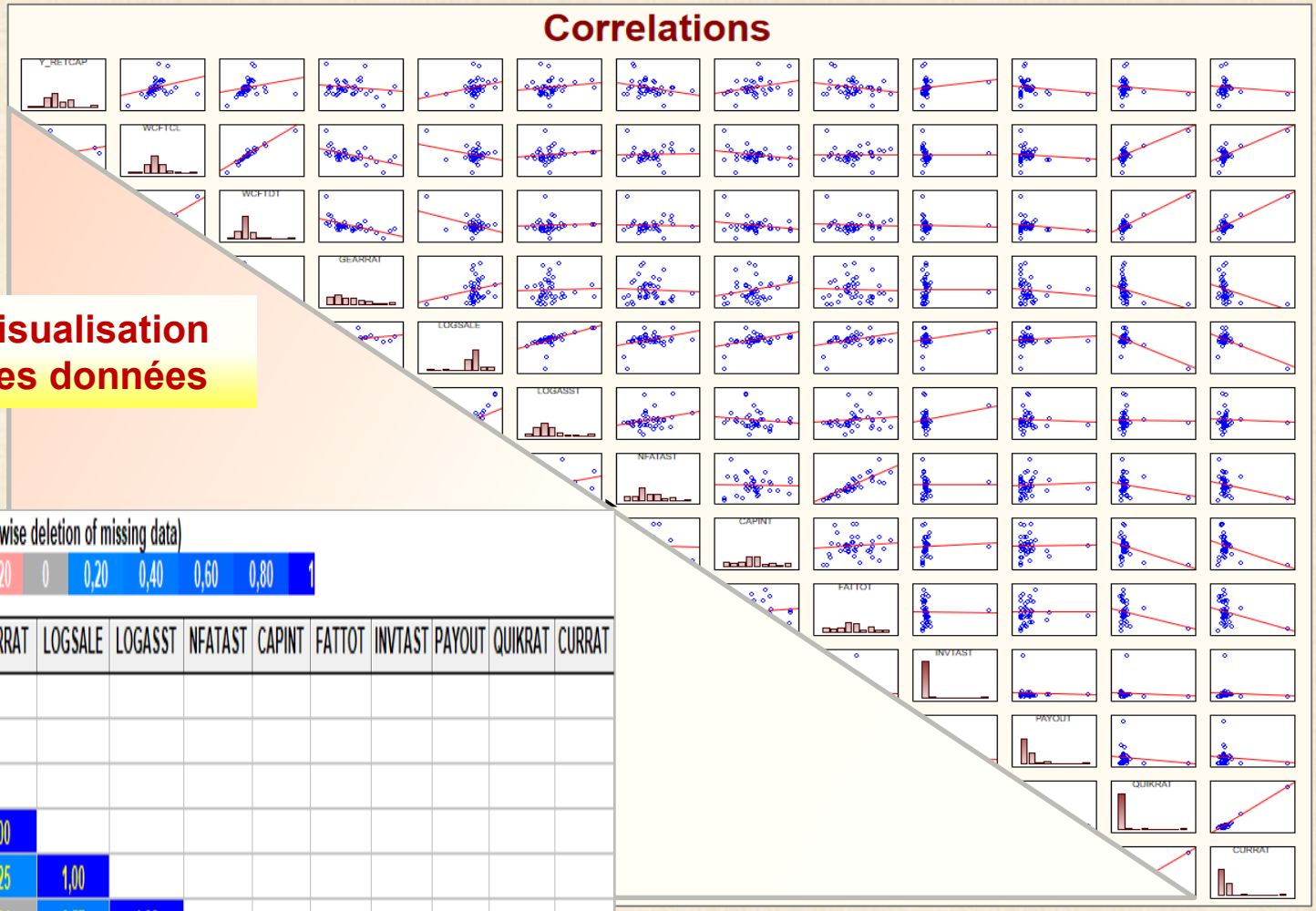
- Basic Statistics/Tables (Financial data-40obs.sta in 2021-MTH8302-Exemples-RE
  - Descriptive statistics dialog
    - Graphical Summary (Y_RETCAP WCFTCL WCFTDT GEARRAT LOGSALE...)
    - Graphical Summary (NFATAST CAPINT FATTOT INVTAST PAYOUT...)
    - Graphical Summary (CURRAT)

Graphical Summary (NFATAST CAPINT FATTOT INVTAST PAYOUT...)



	NFATAST	CAPINT	FATTOT	INVTAST	PAYOUT	QUIKRAT
N:	40	40	40	40	40	40
Mean:	0,337	1,529	0,523	0,414	0,482	1,190
Median:	0,300	1,510	0,505	0,290	0,310	0,755
Min:	0	0	0	0	0	0,140
Max:	0,940	3,460	1,160	6,000	4,210	12,98
L-Qrt:	0,210	1,115	0,375	0,185	0	0,580
U-Qrt:	0,440	1,870	0,710	0,360	0,590	1,110
Variance:	0,0372	0,691	0,0772	0,849	0,538	4,140
SD:	0,193	0,831	0,278	0,922	0,734	2,035
Std.Err:	0,0305	0,131	0,0439	0,146	0,116	0,322
Skw:	0,790	0,457	0,274	5,992	3,733	5,345
Kurt:	1,192	0,511	-0,311	37,13	17,32	30,70
95% Conf SD						
Lower:	0,158	0,681	0,228	0,755	0,601	1,667
Upper:	0,248	1,067	0,357	1,183	0,942	2,613
95% Conf Mean						
Lower:	0,275	1,263	0,434	0,119	0,247	0,540
Upper:	0,398	1,795	0,612	0,709	0,717	1,841

Visualisation des données



Color map of correlations N=40 (Casewise deletion of missing data)

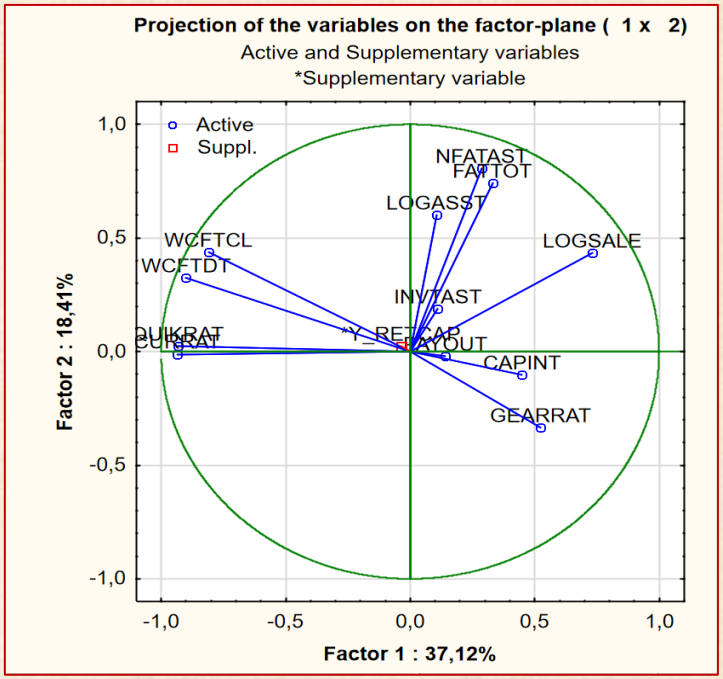
r>= -1 -0,80 -0,60 -0,40 -0,20 0 0,20 0,40 0,60 0,80 1

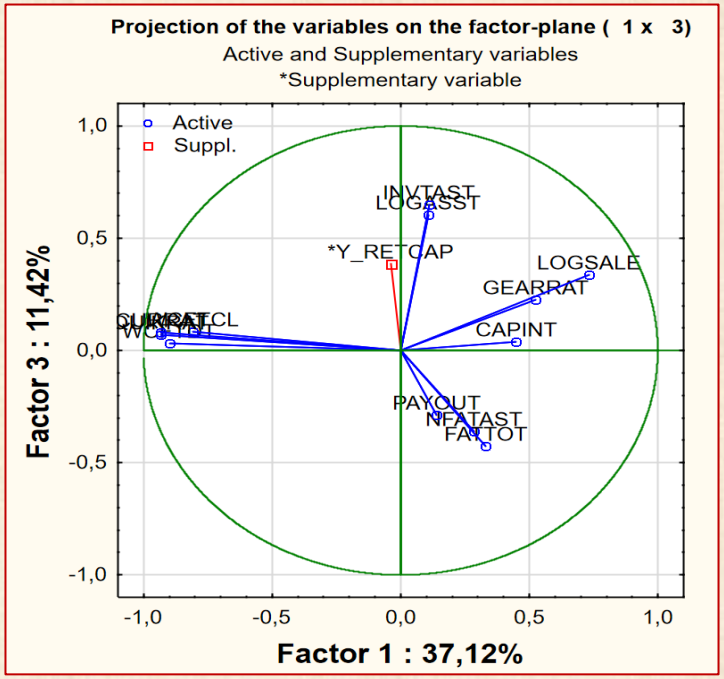
Variable	Y_RETCAP	WCFTCL	WCFTDT	GEARRAT	LOGSALE	LOGASST	NFATAST	CAPINT	FATTOT	INVTAST	PAYOUT	QUIKRAT	CURRAT
Y_RETCAP	1,00												
WCFTCL	0,32	1,00											
WCFTDT	0,23	0,96	1,00										
GEARRAT	-0,17	-0,55	-0,56	1,00									
LOGSALE	0,29	-0,31	-0,45	0,25	1,00								
LOGASST	0,14	0,18	0,06	0,04	0,57	1,00							
NFATAST	-0,30	0,04	-0,04	-0,07	0,36	0,28	1,00						
CAPINT	0,31	-0,24	-0,25	0,25	0,44	-0,16	-0,03	1,00					
FATTOT	-0,26	-0,01	-0,07	-0,05	0,36	0,14	0,84	0,10	1,00				
INVTAST	0,14	-0,00	-0,05	0,00	0,24	0,25	0,00	0,04	-0,01	1,00			
PAYOUT	-0,14	-0,12	-0,15	-0,11	0,09	-0,05	0,05	0,02	0,00	-0,05	1,00		
QUIKRAT	-0,11	0,71	0,83	-0,32	-0,66	-0,02	-0,21	-0,36	-0,27	-0,06	-0,13	1,00	
CURRAT	-0,10	0,70	0,82	-0,33	-0,64	-0,05	-0,27	-0,35	-0,30	-0,07	-0,11	0,98	1,00

Visualisation des données

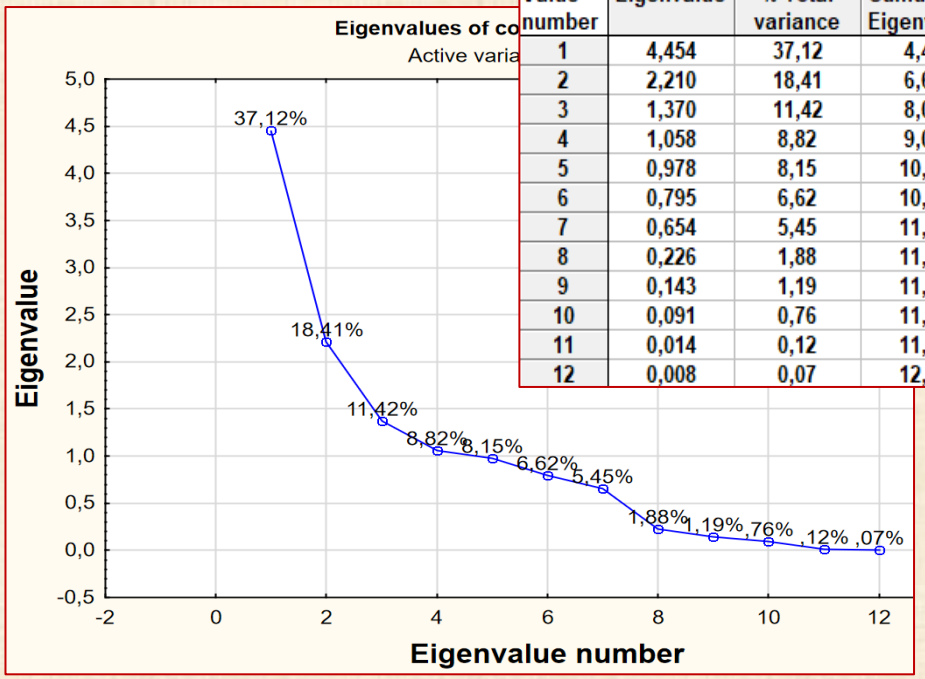
analyse avec ACP

Analyse Composantes Principales



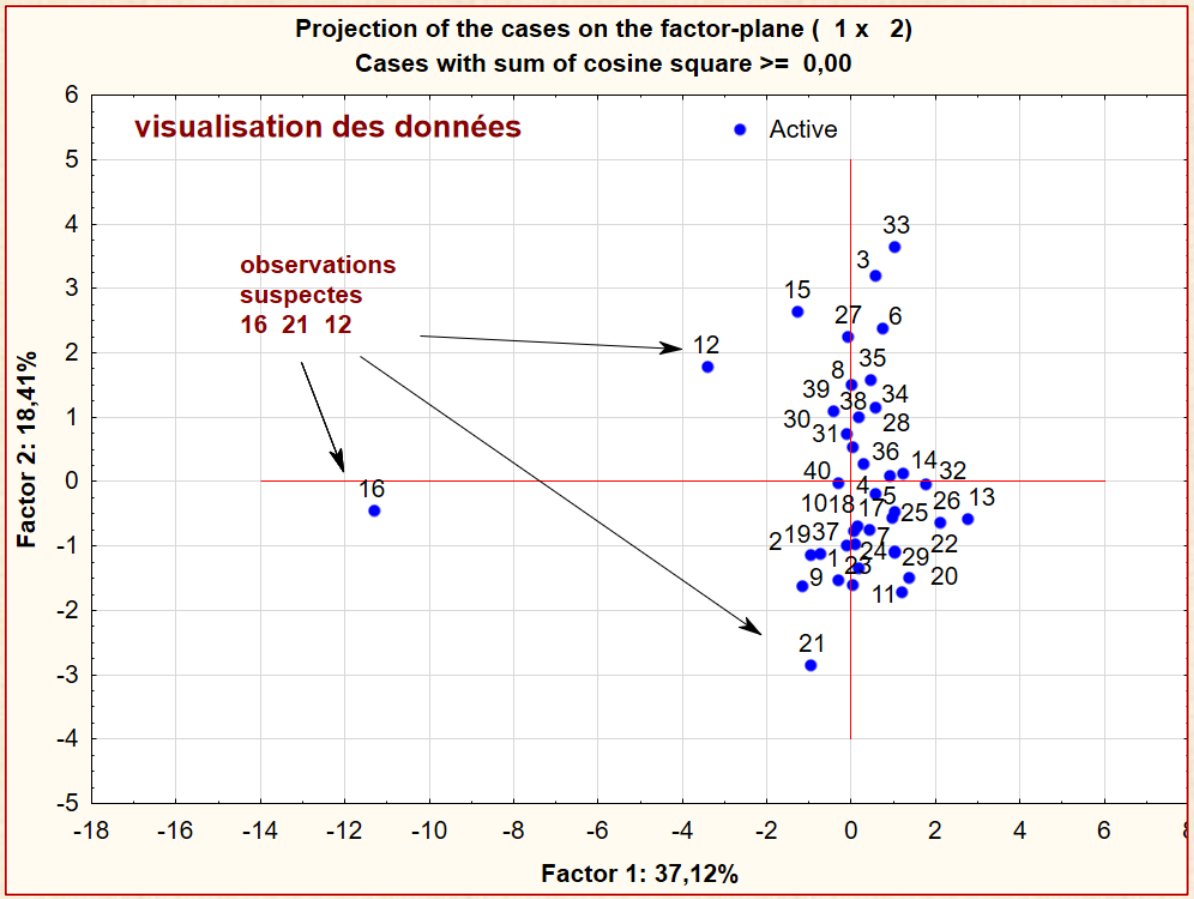


Variable	Factor 1	Factor 2	Factor 3
WCFTCL	-0,81	0,44	0,08
WCFTDT	-0,90	0,32	0,03
GEARRAT	0,52	-0,34	0,23
LOGSALE	0,73	0,44	0,34
LOGASST	0,11	0,60	0,60
NFATAST	0,29	0,81	-0,36
CAPINT	0,45	-0,10	0,04
FATTOT	0,33	0,74	-0,43
INVTAST	0,11	0,19	0,65
PAYOUT	0,14	-0,02	-0,29
QUIKRAT	-0,93	0,02	0,07
CURRAT	-0,93	-0,01	0,08
*Y_RETCAP	-0,04	0,02	0,38



Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	4,454	37,12	4,45	37,12
2	2,210	18,41	6,66	55,53
3	1,370	11,42	8,03	66,95
4	1,058	8,82	9,09	75,77
5	0,978	8,15	10,07	83,92
6	0,795	6,62	10,87	90,54
7	0,654	5,45	11,52	95,99
8	0,226	1,88	11,74	97,87
9	0,143	1,19	11,89	99,06
10	0,091	0,76	11,98	99,82
11	0,014	0,12	11,99	99,93
12	0,008	0,07	12,00	100,00

visualisation des données dans l'espace des 2 facteurs principaux



données aberrantes ? hétérogénéité ?

observation 16 et 21

Case	Factor 1	Factor 2
1	-0,30	-1,53
2	-0,97	-1,15
3	0,60	3,20
4	0,59	-0,20
5	1,03	-0,48
6	0,75	2,38
7	0,11	-0,98
8	0,01	1,49
9	-1,14	-1,63
10	0,07	-0,77
11	1,22	-1,72
12	-3,41	1,77
13	2,79	-0,58
14	1,24	0,13
15	-1,28	2,63
16	-11,31	-0,46
17	0,43	-0,75
18	0,15	-0,70
19	-0,10	-0,99
20	1,37	-1,49
21	-0,96	-2,86
22	1,03	-1,10
23	0,04	-1,61
24	0,19	-1,34
25	0,98	-0,57
26	2,12	-0,65
27	-0,08	2,24
28	0,30	0,26
29	1,04	-1,10
30	-0,09	0,73
31	0,03	0,52
32	1,78	-0,05
33	1,02	3,63
34	0,57	1,15
35	0,48	1,56
36	0,91	0,09
37	-0,73	-1,13
38	0,18	0,99
39	-0,41	1,10
40	-0,30	-0,03

EXEMPLE RÉGRESSION MULTIPLE

matrice de corrélation

	WCFTCL	WCFTDT	GEARRAT	LOGSALE	LOGASST	NFATAST	CAPINT	FATTOT	INVTAST	PAYOUT	QUIKRAT	CURRAT	Y RETCAP
WCFTCL	1												
WCFTDT	**0,962**	1											
GEARRAT	-0,552	-0,561	1										
LOGSALE	-0,310	-0,453	0,250	1									
LOGASST	0,183	0,064	0,039	0,568	1								
NFATAST	0,038	-0,042	-0,067	0,359	0,281	1							
CAPINT	-0,238	-0,252	0,253	0,437	-0,165	-0,029	1						
FATTOT	-0,013	-0,073	-0,048	0,363	0,139	**0,844**	0,103	1					
INVTAST	-0,002	-0,045	0,004	0,239	0,251	0,001	0,042	-0,010	1				
PAYOUT	-0,122	-0,152	-0,109	0,092	-0,050	0,052	0,018	0,004	-0,047	1			
QUIKRAT	0,707	**0,825**	-0,320	-0,662	-0,025	-0,212	-0,358	-0,267	-0,062	-0,133	1		
CURRAT	0,701	**0,821**	-0,331	-0,641	-0,046	-0,270	-0,353	-0,296	-0,069	-0,108	**0,985**	1	
Y_RETCAP	**0,325**	0,233	-0,168	**0,295**	0,141	**- 0,297**	**0,310**	-0,256	0,141	-0,140	-0,110	-0,097	1

4 paires corrélations fortes dans X (soulignées)

WCFTCL	avec : WCFTDT	0,96
WCFTDT	avec : QUIKRAT, CURRAT	0,82
NFATAST	avec : FATTOT	0,84
QUICKRAT	avec : CURRAT	0,98

Y_RETCAP corrélée avec: WCFCL / LOGSALE / NFATAST / CAPINT
corrélation relativement faible (entre -0,3 et 0,3)
très souvent le cas dans les matrices de corrélation

EXEMPLE RÉGRESSION MULTIPLE

Regression Summary for Dependent Variable: Y_RETCAP

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

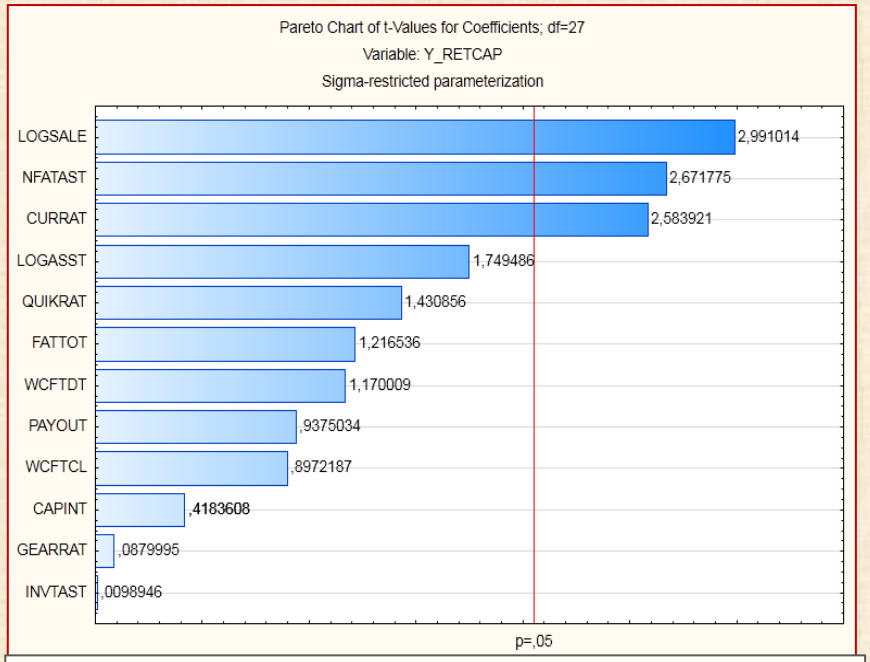
	Beta	Std.Err. - of Beta	B	Std.Err. B	t(27)	p-level
Intercept			0,230	0,135	1,706	0,0995
WCFTCL	0,422	0,471	0,183	0,204	0,897	0,3775
WCFTDT	0,720	0,615	0,358	0,306	1,170	0,2522
GEARRAT	-0,011	0,128	-0,007	0,075	-0,088	0,9305
LOGSALE	0,845	0,282	0,112	0,038	2,991	0,0059
LOGASST	-0,346	0,198	-0,081	0,047	-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 en variables centrées-réduites

coefficients de régression

Analysis of Variance DV: Y_RETCAP

	SS	df	MS	F	p-level
Regress.	0,5486	12	0,0457	7,73	0,00001
Residual	0,1596	27	0,0059		
Total	0,7082				



variables significatives de Y_RETCAP

LOGSALE – NFATAST – CURRAT

LOGASST p-value = 0,09

Comparaison de modèles : BASE de tous les tests d'hypothèses

THÉORIE

Situtation fréquente ratio F global significatif - R^2 assez élevé
certaines variables (effets) non significatives

Solution: nouveau modèle avec les variables importantes - modèle réduit

MC : Modèle Complet : $y_i = \beta_0 + \beta_1 x_{i\,1} + \beta_2 x_{i\,2} + \ldots + \beta_p x_{i\,p} + e_i \qquad i = 1,2,\ldots, n$

Supposons variables $X_1, X_2, \ldots, X_q$ $(q < p)$ considérées pour élimination
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$ sous modèle du modèle complet (modèle réduit)
MR : Modèle Réduit : $y_i = \beta_0 + \beta_{q+1} x_{i\,q+1} + \beta_{q+2}\, x_{i\,q+2} + \ldots + \beta_p x_{i\,p} + e_i$

SSMC : somme des carrés pour le modèle complet
SSMR : somme des carrés pour le modèle réduit SSMR < SSMC
SSE : somme des carrés résiduels pour le modèle complet

ratio $F = [\,(SSMC - SSMR)/q\,] \div [SSE / (n - p - 1)] \sim F(q, n - p - 1)$

On rejette H_0 si le ratio $F > F(q, n - p - 1, 1 - \alpha)$: modèle complet conservé

Si H_0 n'est pas rejetée : modèle réduit est retenu

remarque: le test de nullité sur un coefficient individuel β_j est une cas particulier de ce test général de comparaison de 2 modèles

EXEMPLE RÉGRESSION MULTIPLE

modèle complet avec 12 variables explicatives
modèle réduit avec 6 variables explicatives :
LOGSALE – LOGASST – NFATAST – CURRAT – CURRAT - QUIKRAT

Regression Summary for Dependent Variable: Y_RETCAP
R = 0,86 R^2 = 0,74 Adjusted R^2 = 0,694
F(6,33) = 15,738

	Beta	Std.Err. of Beta	B	Std.Err. of B	t(33)	p-level
Intercept			0,211	0,103	2,055	0,0479
WCFTCL	0,982	0,136	0,425	0,059	7,243	0,0000
LOGSALE	0,733	0,188	0,098	0,025	3,896	0,0005
LOGASST	-0,302	0,137	-0,071	0,032	-2,194	0,0354
NFATAST	-0,723	0,107	-0,505	0,075	-6,762	0,0000
QUIKRAT	1,258	0,624	0,083	0,041	2,018	0,0518
CURRAT	-1,765	0,587	-0,122	0,041	-3,006	0,0050

Analysis of Variance
DV: Y_RETCAP

	Sums of Squares	df	Mean Square	F	p-level
Regress.	0,5248	6	0,0875	15,74	0,00000
Residual	0,1834	33	0,0056		
Total	0,7082				

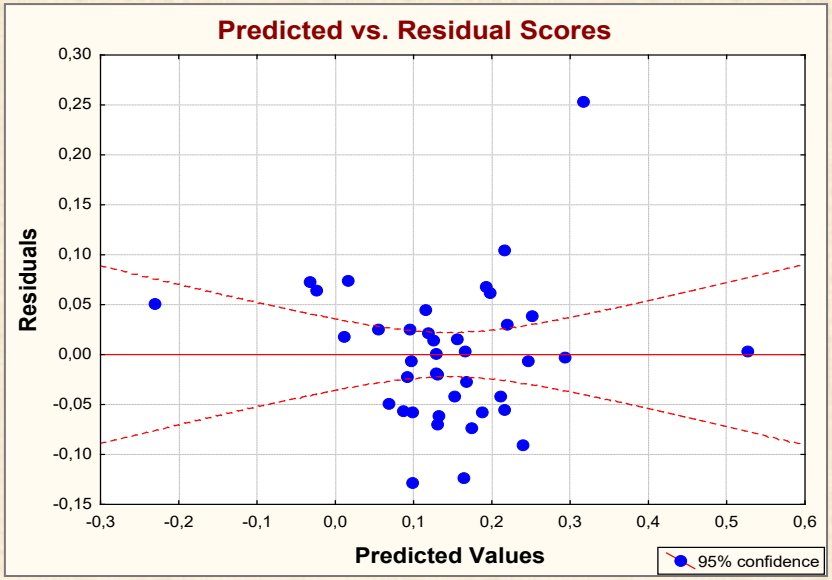
Test: comparaison des modèles

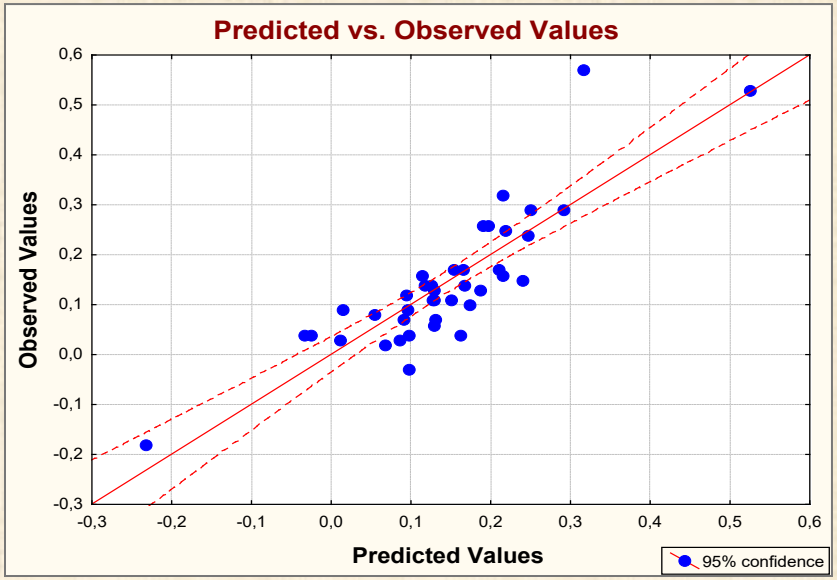
$$F = \frac{(0{,}5486 - 0{,}5248) / 6}{0{,}1596 / 27} = 1{,}07$$

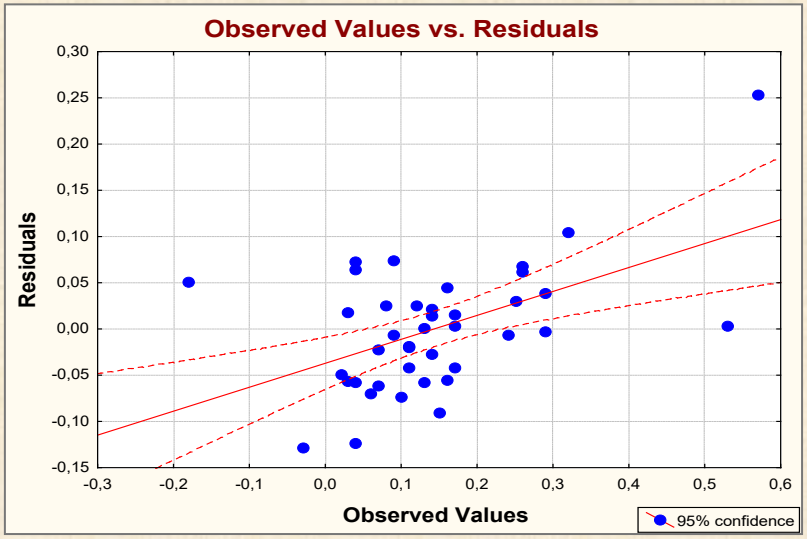
H_0 pas rejetée
modèle réduit retenu

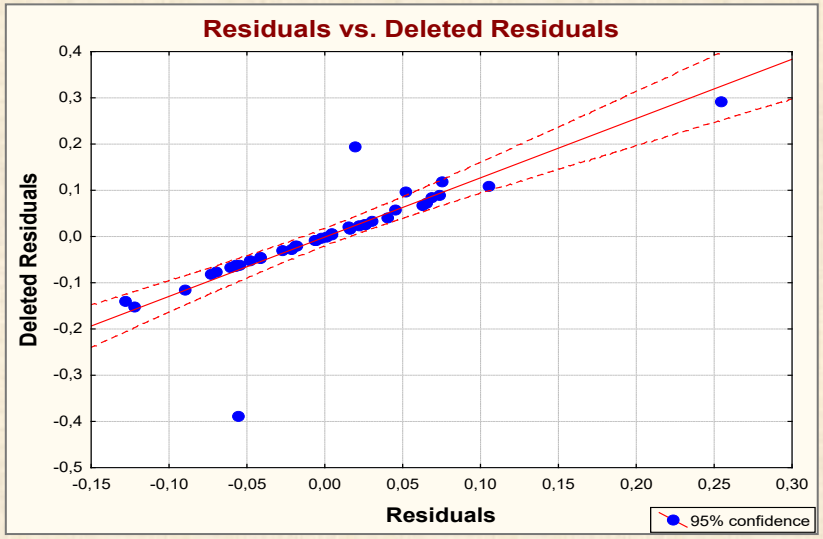
12 variables: R^2 = 0,77 R^2 ajust. = 0,67
6 variables: R^2 = 0,74 R^2 ajust. = 0,69

Analyse des résidus : modèle réduit









conclusion : ?

EXEMPLE RÉGRESSION MULTIPLE

Predicted & Residual Values – modèle à 6 variables réduit Y_RETCAP

	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard resid	Std.Err. Pred.Val	Mahalanobis distance	Deleted residual	Cook's distance
1	0,2600	0,1914	0,0686	0,420	0,920	0,0341	7,169	0,0867	0,0403
2	0,5700	0,3159	0,2541	1,492	3,409	0,0269	4,116	0,2923	0,2866
3	0,0900	0,0150	0,0750	-1,101	1,006	0,0457	13,682	0,1202	0,1395

etc . . . . . .

Observed value The observed value for the dependent variable.

Predicted value The predicted value given the current regression equation.

Residual value The observed value minus the predicted value.

Standard predicted value The standardized predicted value of the dependent variable.

Standard residual value The standardized residual value (observed minus predicted divided by the square root of the residual mean square.

Standard error of predicted value The standard error of the unstandardized predicted value discarding any outliers.

étude approfondie chapitre suivant

critères identification
données influentes
valeurs aberrantes (outliers)

Mahalanobis distance

the independent variables (in the equation) as defining a multidimensional space in which each observation can be plotted.

Also, one can plot a point representing the means for all independent variables. This "mean point" in the multidimensional space is also called the centroid. The Mahalanobis distance is the distance of a case from the centroid in the multidimensional space, defined by the correlated independent variables (if the independent variables are uncorrelated, it is the same as the simple Euclidean distance).

Thus, this measure provides an indication of whether or not an observation is an outlier with respect to the independent variable values.

Deleted residual

The deleted residual is the residual value for the respective case, had it not been included in the regression analysis, that is, if one would exclude this case from all computations.

If the deleted residual differs greatly from the respective standardized residual value, then this case is possibly an outlier because its exclusion changed the regression equation.

Cook's distance

This is another measure of the impact of the respective case on the regression equation. It indicates the difference between the computed B values and the values one would have obtained, had the respective case been excluded.

All distances should be of about equal magnitude;

if not, then there is reason to believe that the respective case(s) biased the estimation of the regression coefficients.

Remarques

With small N (less than 100), multiple regression estimates (the B coefficients) are not very stable

Single extreme observations can greatly influence the final estimates.

Advise: review these statistics and repeat crucial analyses after discarding any outliers.

critères identification données influentes
valeurs aberrantes (outliers)

distance Mahalanobis
distance Cook

Mahalanobis distance (MD) = distance of a case from the centroid of all cases in the space defined by the independent variables

$$MD = (n-1)*(X_{raw} - X_{mean})*C^{-1}*(X_{raw} - X_{mean})'$$

X_{raw}	vector of raw data for the independent variables
X_{mean}	vector of means for the independent variables
C^{-1}	inverse of the matrix of crossproducts of deviations for the independent variables
n	number of valid cases
RMS	Residual Mean Square
DR	Deleted Residual
MD	Mahalanobis Distance
CD	Cook Distance

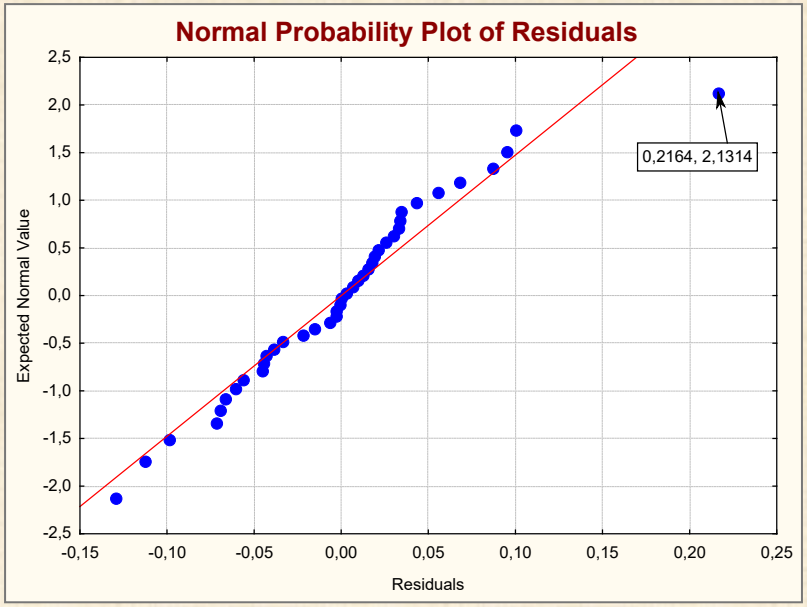
Deleted Residual (DR) = residual obtained had the case not been included in the estimation of the regression equation.

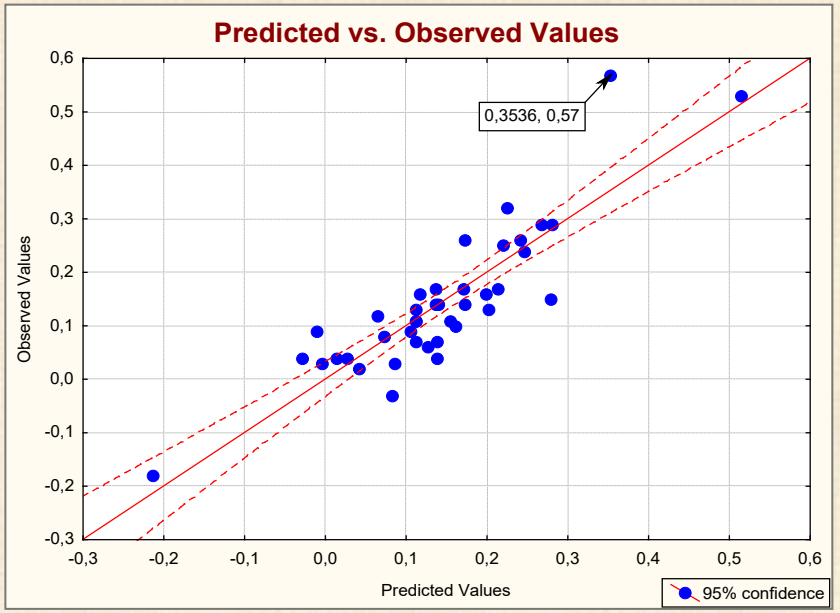
$$DR = 1 - (1/n) - (X_{raw} - X_{mean})*C^{-1}*(X_{raw} - X_{mean})'$$

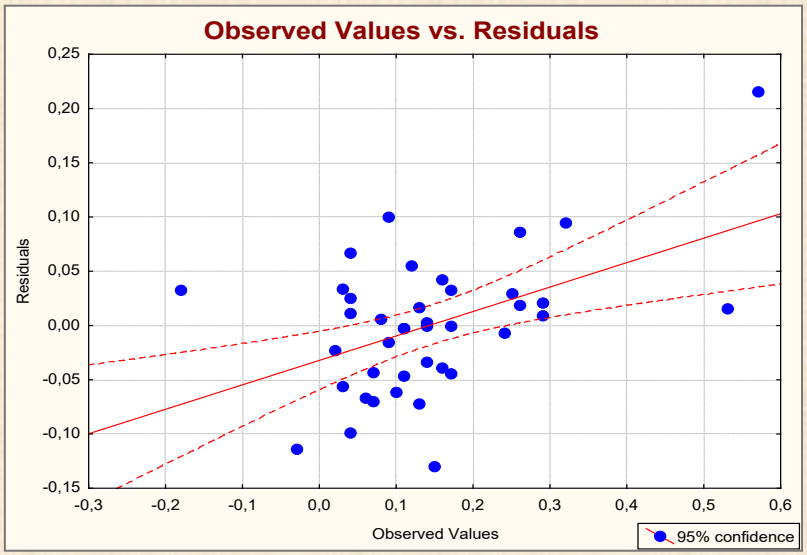
Cook's distance (CD) = for assessing the changes that would result in all residuals if the respective case were to be omitted from the regression analysis.

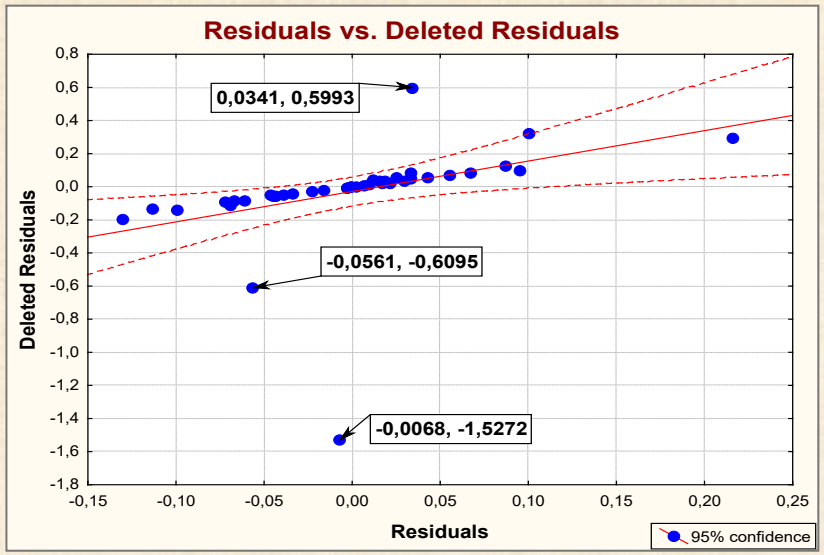
$$CD = \{DR^2 *[1/n + MD/(n-1)]\}/[(No.\ of\ vars + 1)*RMS]$$

Analyse des résidus : modèle réduit









conclusion : ?

Variables centrées- réduites

But: comparer les coefficients de régression
impact des variables explicatives

difficile si on utilise les variables mesurées dans leurs unités

données: $y_i \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ip}$ $\quad i = 1, 2, .., n$

moyennes: $\bar{y} \quad \bar{x}_1 \quad \bar{x}_2 \quad \dots \quad \bar{x}_p$

écart types: $s_y \quad s_1 \quad s_2 \quad \dots \quad s_p$

corrélations: $r_{i\,i'} = s_{i\,i'} / (s_i\, s_{i'})$ $\quad s_{i\,i'}$: covariance de X_i et $X_{i'}$ $\quad i \neq i'$

modèle: $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_p x_{ip}$ avec unités réelles

alors $b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \dots - b_p \bar{x}_p$

modèle en variables centrées-réduites:

$$\frac{\hat{y}_i - \bar{y}}{s_y} = \frac{b_1 s_1}{s_y} \frac{(x_{i1} - \bar{x}_1)}{s_1} + \frac{b_2 s_2}{s_y} \frac{(x_{i2} - \bar{x}_2)}{s_2} + \dots + \frac{b_p s_p}{s_y} \frac{(x_{ip} - \bar{x}_p)}{s_p}$$

$b^*_j = b_j s_j / s_y$: mesure l'impact de la variable X_j sur Y - coefficients peuvent être comparés

mais ils ne tiennent pas en compte l'écart type (précision) de l'estimation

effet général = 0 dans le modèle en variables centrées-réduites

EXEMPLE RÉGRESSION MULTIPLE

Regression Summary for Dependent Variable: Y_RETCAP
Modèle à 12 variables
R = 0,88 R² = 0,775 Adjusted R² = 0,675 F(12,27) = 7,7355

	b*	Std.Err. of Beta	b	Std.Err. of b	t(27)	p-level
Intercept	0		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

b*
coefficients régression en variables centrées-réduites

b
coefficients régression en unités réelles

coefficients de régression en variables centrées-réduites : quelles variables sont les plus importantes?

Regression Summary Dependent Variable: Y_RETCAP
Modèle à 6 variables
R = 0,86 R² = 0,74 Adjusted R² = 0,694 F(6,33) = 15,738

	b*	Std.Er of b*	b	Std.Er of b	t(33)	p-level
Intercept	0		0,211	0,103	2,055	0,0479
WCFTCL	0,982	0,136	0,425	0,059	7,243	0,0000
LOGSALE	0,733	0,188	0,098	0,025	3,896	0,0005
LOGASST	-0,302	0,137	-0,071	0,032	-2,194	0,0354
NFATAST	-0,723	0,107	-0,505	0,075	-6,762	0,0000
QUIKRAT	1,258	0,624	0,083	0,041	2,018	0,0518
CURRAT	-1,765	0,587	-0,122	0,041	-3,006	0,0050

EXEMPLE RÉGRESSION MULTIPLE

Regression Summary for Dependent Variable: Y_RETCAP

Modèle à 6 variables

R = 0,86 R^2 = 0,74 Adjusted R^2 = 0,694 F(6,33) = 15,738

	b*	Std.Er of b*	B	Std.Er of b	t(33)	p-level
Intercept	0		0,211	0,103	2,055	0,0479
WCFTCL	0,982	0,136	0,425	0,059	7,243	0,0000
LOGSALE	0,733	0,188	0,098	0,025	3,896	0,0005
LOGASST	-0,302	0,137	-0,071	0,032	-2,194	0,0354
NFATAST	-0,723	0,107	-0,505	0,075	-6,762	0,0000
QUIKRAT	1,258	0,624	0,083	0,041	2,018	0,0518
CURRAT	-1,765	0,587	-0,122	0,041	-3,006	0,0050

Regression Summary for Dependent Variable: Y_RETCAPcr

Modèle en variables centrées-réduites CR

R= 0,86 R^2 = 0,74 Adjusted R^2 = 0,694 F(6,33)=15,738

=

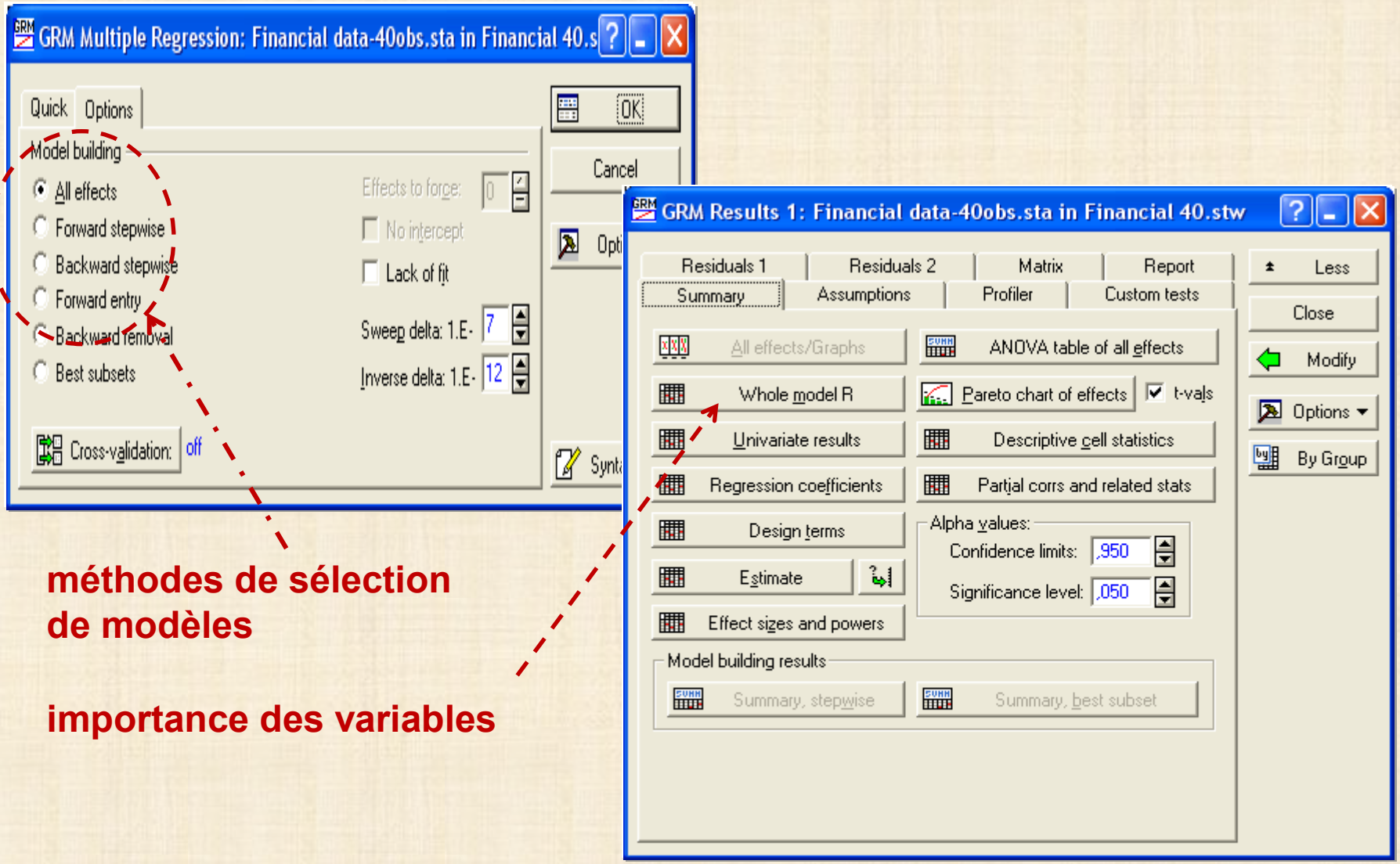
	b*	Std.Er of b*	b	Std.Err of b	t(33)	p-level	Rang
Intercept	0		0,000	0,0875	0,000	1,0000	
WCFTCL cr	0,982	0,136	0,982	0,136	7,243	0,0000	1
LOGSALE cr	0,733	0,188	0,733	0,188	3,896	0,0005	3
LOGASST cr	-0,302	0,137	-0,302	0,137	-2,194	0,0354	5
NFATAST cr	-0,723	0,107	-0,723	0,107	-6,762	0,0000	2
QUIKRAT cr	1,258	0,624	1,258	0,625	2,018	0,0518	6
CURRAT cr	-1,765	0,587	-1,765	0,587	-3,006	0,0050	4

coefficients de régression en variables centrées-réduites; variables les plus importantes = ?

Importance des X sur Y en ordre décroissant

(1)=WCFTCL (2)=NFATAST (3)=LOGSALE
(4)=CURRAT (5)=LOGASST (6)=QUICKRAT

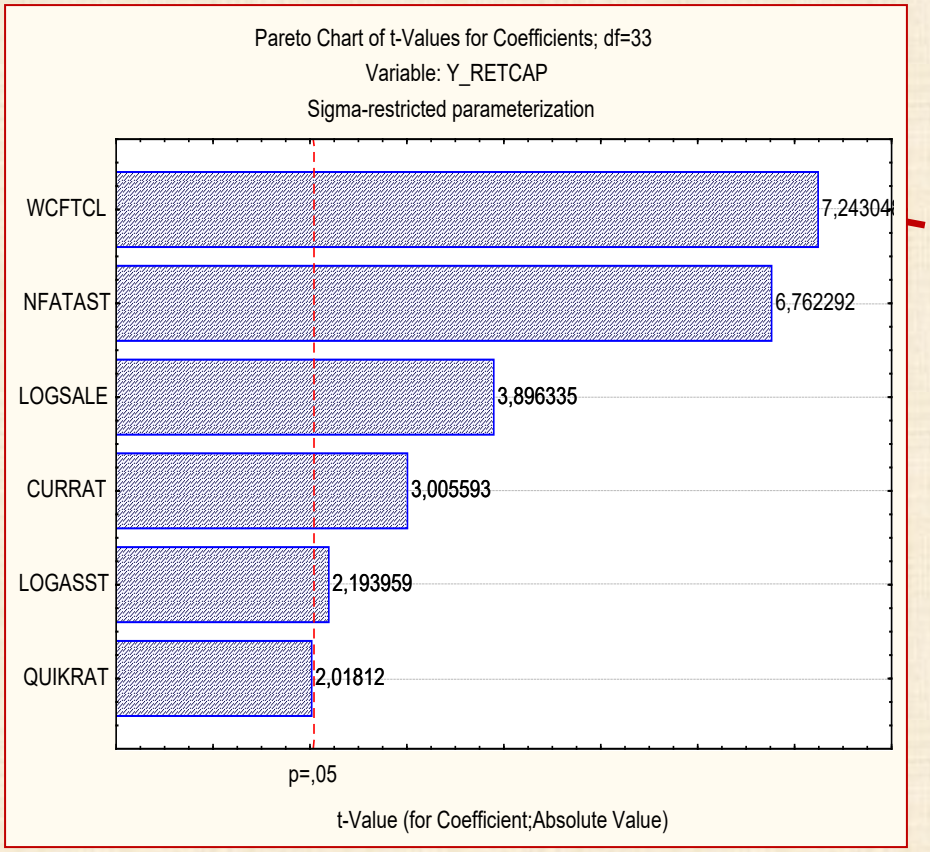
Exemple: données financières de 40 entreprises modèle à 6 variables



méthodes de sélection de modèles

importance des variables

Exemple: données financières de 40 entreprises modèle à 6 variables



coefficients du modèle : param

	Y_RETCAP param	Y_RETCAP Std.Err	Y_RETCAP t	Y_RETCAP p
Intercept	0,211	0,103	2,05	0,0479
WCFTCL	0,425	0,059	7,24	0,0000
LOGSALE	0,098	0,025	3,90	0,0005
LOGASST	-0,071	0,032	- 2,19	0,0354
NFATAST	-0,505	0,075	- 6,76	0,0000
QUIKRAT	0,083	0,041	2,02	0,0518
CURRAT	-0,122	0,041	- 3,01	0,0050

Le test t d'une variable explicative X (unités d'origine)

= bonne indication importance sur Y

Sélection de variables : critères

Objectifs de la modélisation statistique

- descriptif: recherche *exploratoire* des liaisons entre Y et les variables explicatives X_j
- explicatif: recherche *confirmatoire* employant des connaissances a priori du domaine d'application concerné – prise de décisions
- prédictif: recherche de modèles *parcimonieux* – nombre volontairement restreint de variables explicatives - meilleur modèle avec prédictions les plus fiables

CRITÈRES

- **F de Fisher** pour comparer des séquences de *modèles emboités*

$$F = [(SSR - SSR_q)/q] \div [SSE/(n - p - 1)] \sim F(q, n - p - 1) \qquad q < q$$

$$= [(R^2_p - R^2_q)(n - p - 1)] \div [(1 - R^2)q]$$

SSR_p : modèle avec p variables explicatives et R^2_p associé

SSR_q : modèle réduit avec q variable explicatives et R^2_q associé

$$\text{Si } R^2_p - R^2_q > [q(1 - R^2)/(n - p - 1)]\ F(q ; n - p - 1; 1 - \alpha)$$

l'ajout de q variables est justifié

- $R^2_{ajusté} = 1 - (n-1)(1 - R^2)/(n - p - 1) = 1 - [(n-1)MSE / SST]$
- **MSE** : erreur quadratique moyenne du modèle complet avec p variables

maximiser le $R^2_{ajusté}$ revient à minimiser MSE

CRITÈRES (suite)

- **C_q de Mallows** $C_q = (n-q-1)(MSE_q/MSE) - (n - 2q - 2)$

  MSE_q erreur quadratique pour un modèle à q variables

  on recherche une valeur inférieure et **proche de q + 1**

- **PRESS** $PRESS = \Sigma(y_i - \widehat{y}_{(i)})^2$

  y_i : valeur observée de y à la i-ème observation $(x_{i1}, x_{i2},\ldots,x_{ip})$

  $\widehat{y}_{(i)}$: prédiction de y_i sans tenir compte de la i-ème observation

permet de comparer les capacités prédictives de deux modèles

Algorithme de sélection

Global

on compare TOUS les 2^p modèles possibles en cherchant à optimiser

critères: max R^2 max $R^2_{ajusté}$ min C_q (Mallows) proche de 1 + q

Sélection avant (forward)

à chaque pas, une variable est ajoutée au modèle – celle dont la valeur p-value associée à la statistique partielle du test F de Fisher qui compare les 2 modèles est minimum.

Arrêt: si p-value est plus grande qu'une valeur seuil (e.g. 0,50)
équivalent au plus petit F

Élimination arrière (backward)

on démarre avec le modèle complet; à chaque pas, la variable associée à la plus grande valeur p-value est éliminée du modèle.

Arrêt: si p-valeur et plus petite qu'un seuil (souvent 0,10)
équivalent au plus grand F

Mixte (stepwise - pas à pas) / sélection avant + élimination arrière

on introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle des variables devenues moins indispensables du fait de la présence de celles nouvellement introduites

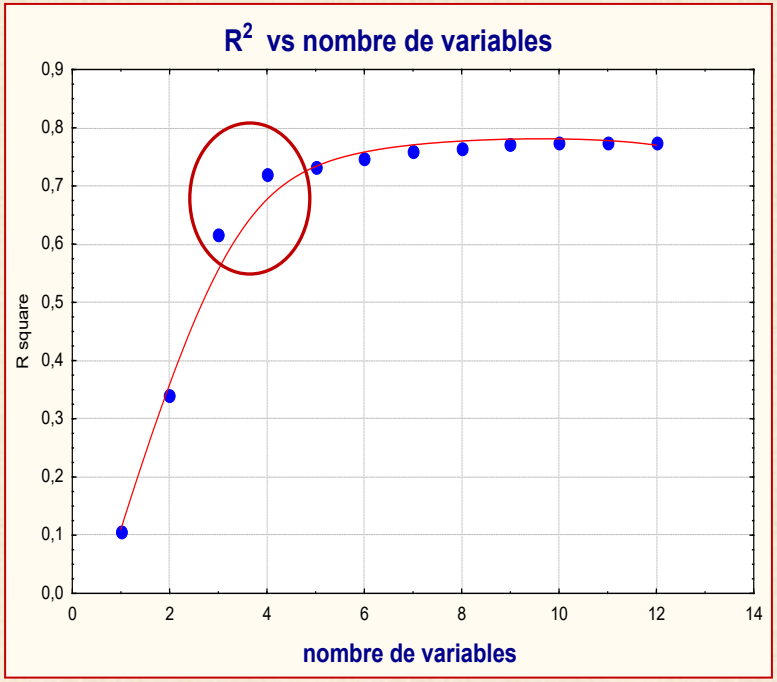
Exemple: données financières de 40 entreprises
12 variables explicatives X

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	GEARRAT	LOGSALE	LOGASST	NFATAST	CAPINT	FATTOT	INVTAST	PAYOUT	QUIKRAT	CURRAT
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



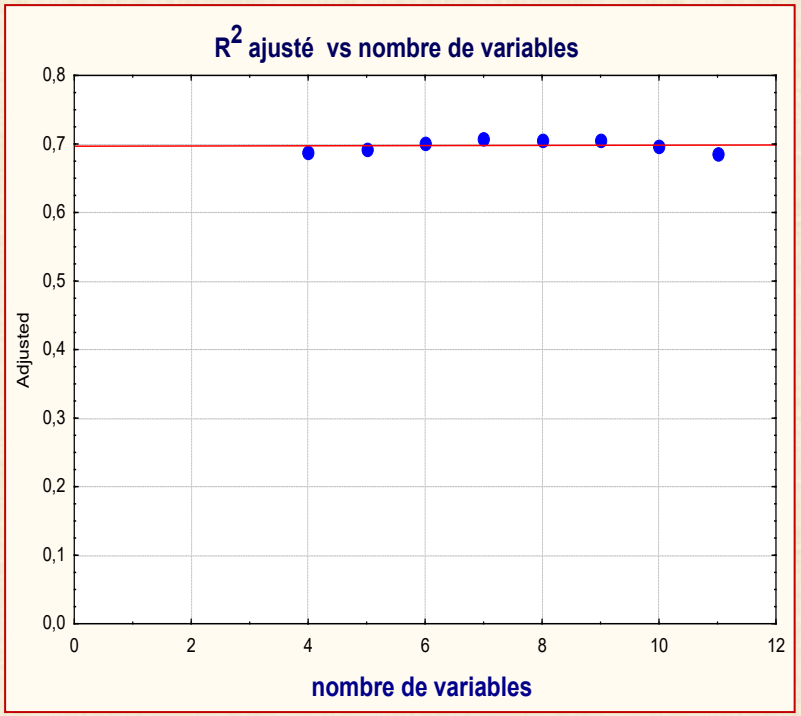
modèle retenu: WCFTDT – LOGSALE- NFATAST - CURRAT

Exemple: données financières de 40 entreprises
12 variables explicatives X

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

Adjusted R square	No Of Effect	WC FTCL	WCF TDT	GEAR RAT	LOG SALE	LOG ASST	NFA TAST	CAPI NT	FAT TOT	INVT AST	PAY OUT	QUI KRAT	CUR RAT
0,686	11	0,42	0,72	-0,01	0,84	-0,35	-0,54	-0,06	-0,23		-0,09	1,07	-1,76
0,697	10	0,42	0,74		0,84	-0,35	-0,54	-0,06	-0,23		-0,09	1,06	-1,76
0,705	9	0,43	0,71		0,76	-0,29	-0,52		-0,23		-0,09	0,93	-1,64
0,705	8		1,22		0,75	-0,25	-0,51		-0,25		-0,08	0,84	-1,67
0,708	7		1,23		0,74	-0,24	-0,53		-0,23			0,92	-1,76
0,701	6		1,21		0,71	-0,19	-0,73					1,02	-1,85
0,692	5		1,21		0,51		-0,73					0,68	-1,63
0,689	4		1,23		0,46		-0,68						-1,00

Max R^2 ajusté



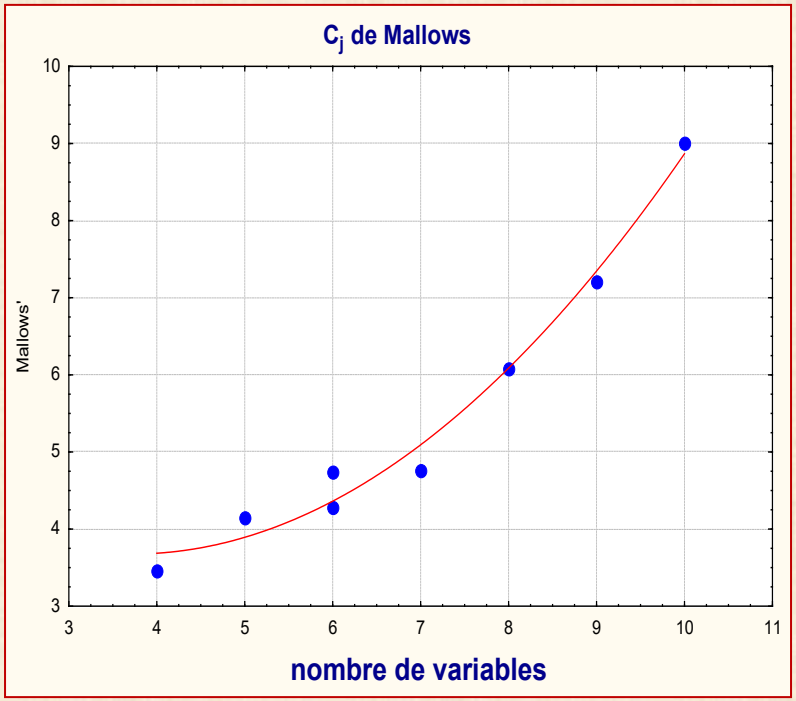
modèle retenu: WCFTDT – LOGSALE- NFATAST - CURRAT

Exemple: données financières de 40 entreprises
12 variables explicatives X

Summary of best subsets; variable(s): Y_RETCAP
Mallow's Cp — standardized regression coefficients for each sub model

Mallows' - Cp	No. Of Effect	WCF TCL	WCF TDT	GEA RRAT	LOGS ALE	LOG ASST	NFAT AST	CAPI NT	FATT OT	INVT AST	PAYO UT	QUIK RAT	CUR RAT
3,463	4		1,23		0,46		-0,68						-1,00
4,142	5		1,21		0,51		-0,73					0,68	-1,63
4,289	6		1,21		0,71	-0,19	-0,73					1,02	-1,85
4,751	6		1,26		0,61	-0,17	-0,45		-0,26				-0,95
4,756	7		1,23		0,74	-0,24	-0,53		-0,23			0,92	-1,76
6,080	8		1,22		0,75	-0,25	-0,51		-0,25		-0,08	0,84	-1,67
7,206	9	0,43	0,71		0,76	-0,29	-0,52		-0,23		-0,09	0,93	-1,64
9,008	10	0,42	0,74		0,84	-0,35	-0,54	-0,06	-0,23		-0,09	1,06	-1,76

Min C_j de Mallows



modèle retenu: WCFTDT – LOGSALE- NFATAST - CURRAT

Exemple: données financières de 40 entreprises 12 variables explicatives X

Summary of stepwise regression; variable:
Forward stepwise P to enter: ,05, P to remove: ,05

	Steps	Degr. of - Freedom	F to - remove	P to - remove	F to - enter	P to - enter	Effect - status
WCFTCL	Step 1	1			4,484	0,041	Entered
WCFTDT		1			2,187	0,147	Out
GEARRAT		1			1,103	0,300	Out
LOGSALE		1			3,617	0,065	Out
LOGASST		1			0,772	0,385	Out
NFATAST		1			3,688	0,062	Out
CAPINT		1			4,030	0,052	Out
FATTOT		1			2,658	0,111	Out
INVTAST		1			0,774	0,384	Out
PAYOUT		1			0,765	0,387	Out
QUIKRAT							Out
CURRAT							Out

Summary of stepwise regression; variable: Y_RETCAP
Forward stepwise P to enter: ,05, P to remove: ,05

	Steps	Degr. of - Freedom	F to - remove	P to - remove	F to - enter	P to - enter	Effect - status
WCFTCL	Step 8	1	46,090	0,0000			In
CURRAT		1	16,795	0,0002			In
NFATAST		1	39,427	0,0000			In
LOGSALE		1	8,468	0,0062			In
LOGASST		1			2,187	0,1484	Out
GEARRA		1			0,012	0,9132	Out
CAPINT		1			1,664	0,2058	Out
FATTOT		1			0,411	0,5256	Out
INVTAST		1			0,003	0,9578	Out
PAYOUT		1			1,166	0,2878	Out
WCFTDT		1			3,894	0,0566	Out
QUIKRAT		1			1,478	0,2324	Out

Sélection avant

Summary of stepwise regression; variable:
Forward stepwise P to enter: ,05, P to remove: ,05

	Steps	Degr. of - Freedom	F to - remove	P to - remove	F to - enter	P to - enter	Effect - status
WCFTCL	Step 2	1	4,484	0,0408			In
WCFTDT		1			3,844	0,0575	Out
GEARRAT		1			0,008	0,9305	Out
LOGSALE		1			8,876	0,0051	Out
LOGASST		1			0,288	0,5948	Out
NFATAST		1			4,457	0,0416	Out
CAPINT		1			7,973	0,0076	Out
FATTOT		1			2,819	0,1016	Out
INVTAST		1			0,852	0,3621	Out
PAYOUT		1			0,433	0,5147	Out
QUIKRAT		1			12,846	0,0010	Entered
CURRAT		1			11,188	0,0019	Out

modèle retenu

WCFTDT – LOGSALE - NFATAST - CURRAT

Exemple : données financières de 40 entreprises - 12 variables X

Backward stepwise — Summary of stepwise regression; variable: P to enter: ,05, P to remove: ,05

	Steps	Degr. of - Freedom	F to - remove	P to - remove	F to - enter	P to - enter	Effect - status
WCFTCL	Step 1	1	0,805	0,378			In
WCFTDT		1	1,369	0,252			In
GEARRAT		1	0,008	0,931			In
LOGSALE		1	8,946	0,006			In
LOGASST		1	3,061	0,092			In
NFATAST		1	7,138	0,013			In
CAPINT		1	0,175	0,679			In
FATTOT		1	1,480	0,234			In
INVTAST		1	0,000	0,992			Removed
PAYOUT		1	0,879	0,357			In
QUIKRAT		1	2,047	0,164			In
CURRAT		1	6,677	0,016			In

Élimination arrière

Backward stepwise — Summary of stepwise regression; variable: Y_RETCAP, P to enter: ,05, P to remove: ,05

	Steps	Degr. of - Freedom	F to - remove	P to - remove	F to - enter	P to - enter	Effect - status
WCFTCL	Step 2	1	0,841	0,367			In
WCFTDT		1	1,430	0,242			In
GEARRAT		1	0,008	0,928			Removed
LOGSALE		1	9,573	0,004			In
LOGASST		1	3,175	0,086			In
NFATAST		1	7,480	0,011			In
CAPINT		1	0,182	0,673			In
FATTOT		1	1,535	0,226			In
CURRAT		1	7,019	0,013			In
PAYOUT		1	0,918	0,346			In
QUIKRAT		1	2,177	0,151			In
INVTAST		1			0,000	0,992	Out

Backward stepwise — Summary of stepwise regression; variable: Y_RETCAP, P to enter: ,05, P to remove: ,05

	Steps	Degr. of - Freedom	F to - remove	P to - remove	F to - enter	P to - enter	Effect - status
CURRAT	Step 9	1	27,239	0,000			In
WCFTDT		1	54,914	0,000			In
NFATAST		1	45,737	0,000			In
LOGSALE		1	14,393	0,001			In
QUIKRAT		1			1,397	0,245	Out
LOGASST		1			0,685	0,414	Out
FATTOT		1			1,299	0,262	Out
PAYOUT		1			0,588	0,448	Out
WCFTCL		1			0,175	0,678	Out
CAPINT		1			0,373	0,546	Out
GEARRAT		1			0,212	0,648	Out
INVTAST		1			0,049	0,826	Out

modèle retenu

WCFTDT – LOGSALE - NFATAST - CURRAT