Chapitre 4 Multiple -2

Influence - résidus - validation

Avant modélisation: conduire des analyses uni et bi-variées et multivariées (ACP) pour identifier les problèmes sur les distributions de chacune des variables : dissymétrie, valeurs atypiques (« outliers »), aberrantes, corrélations élevées entre les variables 2 à 2, utilisation du SPC (pas/peu employé … dommage!)

Après modélisation: aides et diagnostics associés à la régression multiple pour détecter violations: variance non constante, identification de points influents

méthode de moindres carrés: méthode pas très robuste

l'estimation des paramètres est sensible en présence de points influents : lesquels?

Effet de LEVIER

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

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

Y prédites $\hat{Y} = X \hat{\beta} = X C Y = H Y$ $H = X (X' X)^{-1} X'$ « hat matrix »

résidus bruts $e = \hat{Y} - Y = (I - H)Y$ I : matrice identité

observation i influente si h_{ii} (levier) de la diagonale de H est « grand »

$$h_{ii} = (1/n) + (x_i - \bar{x})' (Z' Z)^{-1} (x_i - \bar{x}) = (1/n) + \Sigma [v_j (x_i - \bar{x}) / \lambda_j^{0,5}]^2$$

Z : matrice des données observées privée de première colonne de 1 et dont on a retranché à chaque ligne le vecteur moyen $\bar{x}$ et

λ_j , v_j : valeurs et vecteurs propres de Z'Z

effet de levier important de l'observation x_i : si éloignée du barycentre $\bar{x}$ ET dans la direction d'un vecteur propre v_j associée à une valeur propre λ_j petite

Influence - résidus - validation

TYPES de RÉSIDUS

résidu brut $\hat{e}_i = Y_i - \hat{Y}_i = (I - H)Y_i$ I : matrice identité

résidu standardisé $r_i = e_i / \hat{\sigma}\,(1 - h_{ii})^2$ résidu brut divisé par son écart type

résidu deleted $t_i = e_i / \hat{\sigma}_{(i)}\,(1 - h_{ii})^2$ résidu privé (deleted) de l'observation i

signal d'alerte $|r| > 2{,}5$ ou $|t| > 2{,}5$

notation $\hat{b}_{(i)}$ $\hat{y}_{(i)}$ $\hat{e}_{(i)}$ $\hat{\sigma}_{(i)}$: estimations réalisées SANS l' observation i

MESURES D'INFLUENCE

effet levier: observations x_i éloignées du barycentre $\bar{x}$

ET grand résidu : valeurs atypiques de Y

MESURES SYNTHÉTIQUES (globales)

Distance de Cook = $CD_i = [h_{ii} / (1- h_{ii})]\, r_i^2 /(1+p) = (y - \hat{y}_{(i)})(y - \hat{y}_{(i)}) / \hat{\sigma}(1+p)$

observation influente si $D_i > 1$

DFFITS = $t_i\ [h_{ii} / (1- h_{ii})]^{0,5} = (\hat{y}_i - \hat{y}_{(i)})(\hat{y}_i - \hat{y}_{(i)}) / [(h_{ii})^{0,5}\ \hat{\sigma}_{(i)}]$

différence de prédiction avec et sans l'observation i

Influence - résidus - validation

Exemple 1: données financières - 40 entreprises et 12 variables explicatives potentielles
modèle avec 4 variables explicatives WCFTCL - LOGSALE – NFATAST – CURRAT

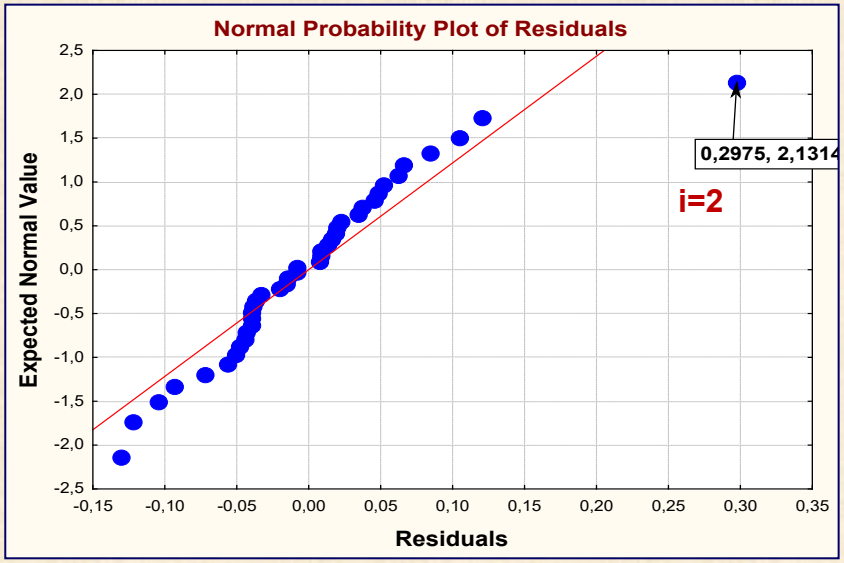
Regression Summary for Y_RETCAP

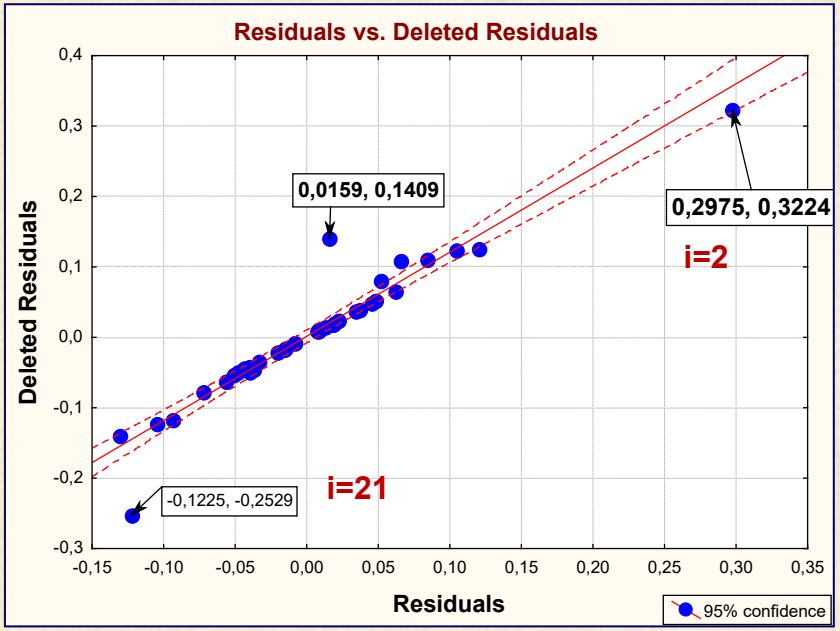
R = 0,83 R^2 = 0,69 Adjusted R^2 = 0,655

	b*	Std. b*	b	Std. Er b	t(35)	P level
Inter			0,063	0,0852	0,742	0,4631
WCF TCL	0,966	0,142	0,418	0,0615	6,789	0,0000
LOG SALE	0,375	0,129	0,050	0,0172	2,910	0,0062
NFA FAST	-0,661	0,105	-0,462	0,0736	-6,279	0,0000
CURR	-0,713	0,174	-0,049	0,0120	-4,098	0,0002

Analysis of Variance; Y_RETCAP

	SS	df	MS	F	P level
Regr	0,4889	4	0,1222	19,51	0,0000
Resi	0,2193	35	0,0063		
Total	0,7082				





Influence - résidus - validation

Predicted & Residual Values Dependent variable: Y_RETCAP

	Observ Value	Predicted Value	Residual	Standard Pred. v.	Stand Residu	Std. Err. Pred. Val	Mahalanobis - Distance	Deleted Residual	Cook's - Distance
1	0,260	0,251	0,009	0,97	0,11	0,022	2,12	0,010	0,000
2	0,570	0,273	0,297	1,16	3,76	0,022	2,03	0,322	0,256
3	0,090	0,038	0,052	-0,94	0,66	0,047	12,58	0,080	0,071
.	.	.	.	.	.	.	.	.	.
21	0,0300	0,1525	-0,1225	0,087	-1,548	0,0568	19,137	-0,2529	1,0531
.	.	.	.	.	.	.	.	.	.
39	0,140	0,118	0,022	-0,22	0,28	0,023	2,25	0,024	0,002
40	0,130	0,145	-0,015	0,02	-0,18	0,015	0,51	-0,015	0,000
Min	-0,180	-0,246	-0,130	-3,47	-1,64	0,014	0,18	-0,253	0,000
Max	0,570	0,445	0,297	2,70	3,76	0,075	33,62	0,322	1,053
Mea	0,143	0,143	-0,000	0,00	-0,00	0,025	3,90	0,001	0,065
Median	0,125	0,145	-0,008	0,02	-0,11	0,020	1,53	-0,009	0,004

2 observations sont identifiées comme influentes: i = 2 et i = 21

raisons différentes: i = 2 valeur atypique de Y

i = 21 x_i éloignée de $\overline{x}$

Exemple 1 : données financières - 40 entreprises et 12 variables explicatives potentielles
ajustement modèle 4 variables explicatives WCFTDT - LOGSALE – LOGASST – CURRAT
observation 2 et observation 21 enlevées: 38 entreprises sur 40

	Param.	Std Er	t	p	Beta (ß)	St.Err ß
Intercept	0,179	0,088	2,035	0,0500		
WCFTCL	0,416	0,088	9,00	0,0000	1,1266	0,1252
LOG SALE	0,025	0,046	1,52	0,1372	0,1998	0,1311
NFA TAST	-0,440	0,057	-7,72	0,0000	-0,6974	0,0904
CURRAT	-0,059	0,011	-5,46	0,0000	-0,9920	0,1818

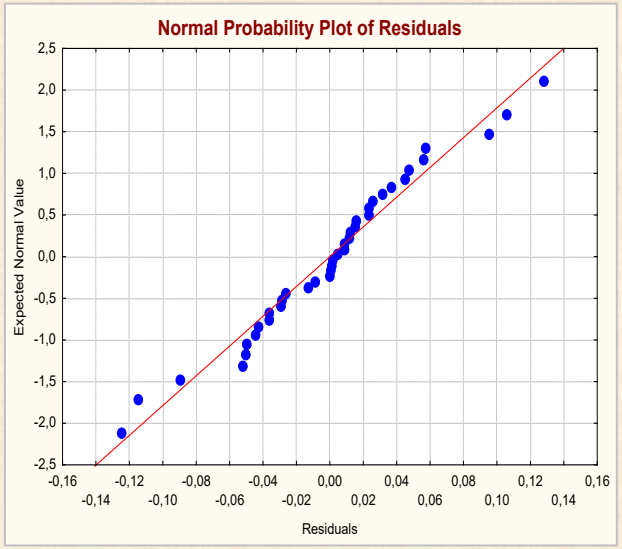
Analysis of Variance; DV: Y_RETCAP
R^2 = 0,79 R^2 ajusté = 0,77

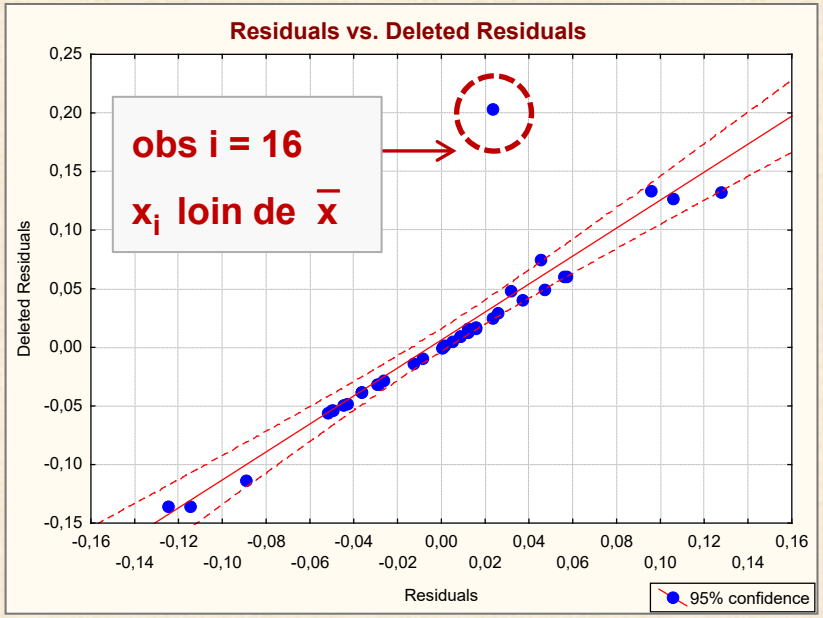
	Sums of Squares	df	Mean - Squares	F	p-level
Regress.	0,4044	4	0,1011	31,51	0,00000
Residual	0,1059	33	0,0032		
Total	0,5103				

avec obs = 2 et 21

	param.
Intercept	0,063
WCFTCL	0,418
LOGSALE	0,050
NFATAST	- 0,462
CURRAT	- 0,049

R^2 = 0,69
R^2_{AJUS} = 0,655





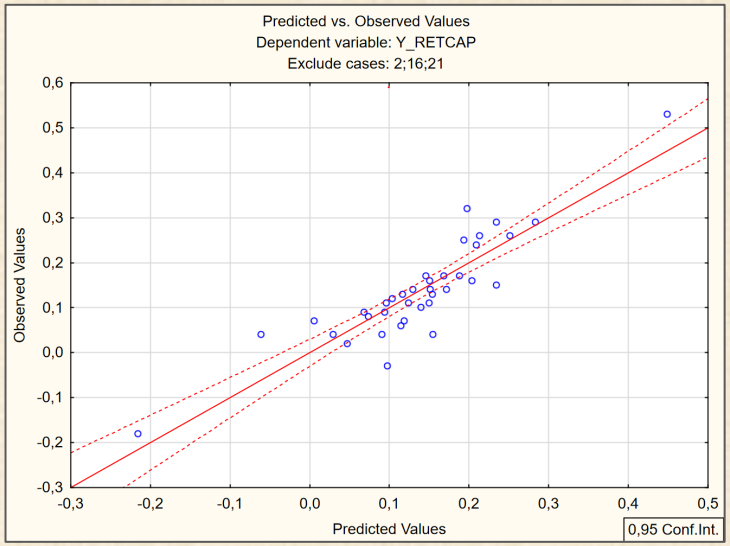
Exemple 1: données financières - 40 entreprises et 12 variables explicatives potentielles
ajustement modèle 4 variables explicatives WCFTDT - LOGSALE – LOGASST – CURRAT
observations 2 - 16 - 21 enlevées : 37 entreprises

Regression Summary for Dependent Variable: Y_RETCAP

(Financial-37obs.sta in Exemples REGRESSION-analyses.stw)

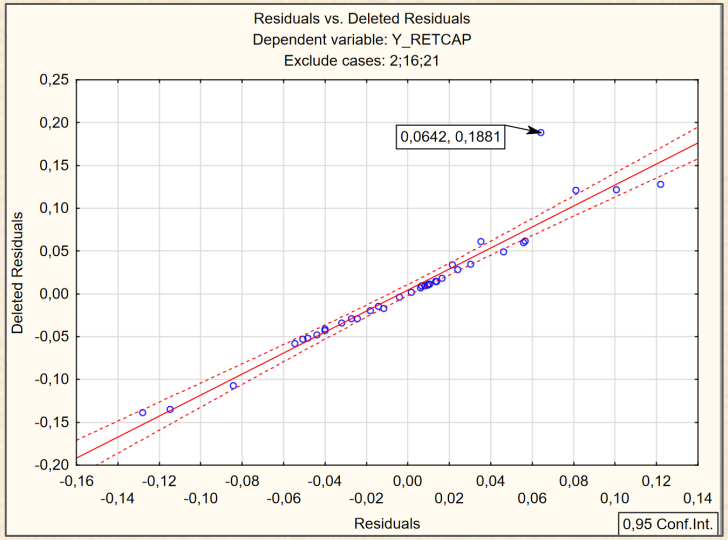
$R = 0,893$ $R^2 = 0,797$ Adjusted $R^2 = 0,772$

	b*	Std.Err b*	b	Std.Err. b	t(32)	p-value
Intercept			0,181	0,087	2,072	0,04644
WCFTCL	0,886	0,097	0,419	0,046	9,104	0,00000
LOGSAL	0,152	0,085	0,030	0,017	1,785	0,08376
NFATAST	-0,694	0,088	-0,449	0,057	-7,865	0,00000
CURRAT	-0,443	0,096	-0,073	0,016	-4,593	0,00006



Analysis of Variance; DV: Y_RETCAP

	Sums of Squares	df	Mean Squares	F	p-value
Regress.	0,397908	4	0,099477	31,4485	0,000000
Residual	0,101222	32	0,003163		
Total	0,499130				



Exemple 1: données financières - 40 entreprises et 12 variables explicatives potentielles
ajustement modèle 4 variables explicatives WCFTDT - LOGSALE - LOGASST - CURRAT
observation 2 – 16 - 21 enlevées : 37 entreprises

Predicted & Residual Values									
	Observed - Value	Predicted - Value	Residual	Standard - Pred. v.	Standard - Residual	Std.Err. - Pred.Val	Mahalanobis - Distance	Deleted - Residual	Cook's Distance
10	0,060000	0,114606	-0,054606	-0,21584	-0,97091	0,013597	1,13099	-0,057995	0,012429
11	0,070000	0,005819	0,064181	-1,25059	1,14116	0,045647	22,74089		
12	-0,180000	-0,215418	0,035418	-3,35493	0,62973	0,036448	14,14587	0,061061	0,099005
13	0,120000	0,103557	0,016443	-0,32093	0,29236	0,015651	1,81471	0,017823	0,001555
14	0,150000	0,234223	-0,084223	0,92193	-1,49751	0,026181	6,82789	-0,107522	0,158396
15	0,080000	0,073779	0,006221	-0,60417	0,11061	0,012585	0,82964	0,006549	0,000136
16	0,090000	0,094085	-0,004085	-0,41103	-0,07263	0,011346	0,49216	-0,004258	0,000047
17	0,250000	0,193320	0,056680	0,53287	1,00779	0,016148	1,99486	0,061773	0,019890
18	-0,030000	0,097923	-0,127923	-0,37451	-2,27451				
19	0,040000	0,090693	-0,050693	-0,44329	-0,90133	0,012475	0,79813	-0,053316	0,008842
20	0,170000	0,145883	0,024117	0,08167	0,42880	0,021978	4,52436	0,028463	0,007822
Min	-0,180000	-0,215418	-0,127923	-3,35493	- 2,274				0,000009
Max	0,530000	0,44870	0,12215	2,96274	2,171		22,740		1,472977
Mean	0,137297	0,13729	-0,0000	0,00000	-0,0000	0,01894	3,89189	0,00457	0,072628
med	0,130000	0,14583	0,00661	0,08167	0,11764	0,01576	1,85651	0,00920	0,007822

modèle acceptable avec cette petite déviation

PROBLÈME de MULTICOLINÉARITÉ

But et objectifs de l'analyse de régression multiple

- Identifier les effets relatifs des variables explicatives X sur une variable de réponse Y
- Faire des prédictions de Y avec de nouvelles valeurs de X
- Sélectionner un ensemble approprié de variables X pour modéliser Y

Problématique fréquente

- Variables X présentent un degré de dépendance linéaire fort (à définir) rendant les objectifs ci-haut impossible à réaliser correctement
- Problème fréquent et important Situation connue sous le nom de *Multicolinéarité*

Sources de la multicollinéarité Exemple Cause

- Méthode de collecte des données .. étude de cas Acétylène …. espace observationnel restreint
- Contraintes sur le modèle ……. étude de cas Acétylène ….. variables quasi redondantes
- Spécification du modèle …… étude de cas Acétylène ….. nécessité forme spécifique
- Modèle surdéfini …… médecine / sc. humaines … # variables > # observations en génomique

Méthodes (critères) pour détecter la multicolinéarité

coefficient de corrélation / variance inflation factor (VIF) / indice de condition (IC)

Solutions pour contrer la multicolinéarité

sélection variables / régression biaisée (ridge) / régression composantes principales

Exemples traités

Exemple 1 : Voitures - 6 variables continues X / 2 variables catégoriques X / 1 variable continue Y

Exemple 2 : Acétylène (étude de cas développée) / 3 variables continues X / 1 variable continue Y

Exemple 2 : Voitures / 6 variables continues X / 2 variables catégoriques X 1 variable continue Y

NOM	CYL	PUIS	LON	LAR	POIDS	VITESSE	NAT	FINITION	Y_PRIX
ALFASUD-TI-1350	1350	79	393	161	870	165	I	B	30570
AUDI-100-L	1588	85	468	177	1110	160	D	TB	39990
SIMCA-1307-GLS	1294	68	424	168	1050	152	F	M	29600
CITROEN-GS-CLUB	1222	59	412	161	930	151	F	M	28250
FIAT-132-1600GLS	1585	98	439	164	1105	165	I	B	34900
LANCIA-BETA-1300	1297	82	429	169	1080	160	I	TB	35480
PEUGEOT-504	1796	79	449	169	1160	154	F	B	32300
RENAULT-16-TL	1565	55	424	163	1010	140	F	B	32000
RENAULT-30-TS	2664	128	452	173	1320	180	F	TB	47700
TOYOTA-COROLLA	1166	55	399	157	815	140	J	M	26540
ALFETTA-1.66	1570	109	428	162	1060	175	I	TB	42395
PRINCESS-1800-HL	1798	82	445	172	1160	158	GB	B	33990
DATSUN-200L	1998	115	469	169	1370	160	J	TB	43980
TAUNUS-2000-GL	1993	98	438	170	1080	167	D	B	35010
RANCHO	1442	80	431	166	1129	144	F	TB	39450
MAZDA-9295	1769	83	440	165	1095	165	J	M	27900
OPEL-REKORD-L	1979	100	459	173	1120	173	D	B	32700
LADA-1300	1294	68	404	161	955	140	U	M	22100

Variables catégoriques

NAT

I : Italie
D: Allemagne
F : France
J : Japon
GB : Grande Bretagne
U : Russie

FINITION

B : Bien
M : Moyenne
TB : Très Bien

Modèle 1 : Y_Prix vs 6 variables continues
Modèle 2 : Y_Prix vs 6 variables continues X + 2 variables catégoriques

QUESTION : comment incorporer les variables catégoriques dans un modèle de régression ?

Comment incorporer les variables catégoriques dans un modèle de régression ?

méthode 1 : codage disjonctif (variables à valeurs 1 / 0) pour chacune des modalités
méthode 2 : codage à effet (variables à valeurs 1 / 0 / -1) pour chacune des modalités
-1 correspond à une modalité de référence (à choisir)
méthode 2 préférable à méthode 1 méthode 2 utilisée par Statistica
modalité de référence : dernière modalité distincte de la variable catégorique
https://cours.polymtl.ca/mth6301/mth8302-Cours&Plus/Clement/Clement-Codage_Variables_Modeles_Statistiques.pdf

X: variable quantitative variant dans l'intervalle [a, b]

a = min (A) b = max(A)
W variable de codage associée à X
$W = (X - c) / d$
c = (a + b) /2 : point milieu de l'intervalle [a, b]
d = (b - a) /2 : demi longueur de l'intervalle [a, b]
l'intervalle de variation de W [-1, + 1].

Xcr : codage centrage-réduction

observations $x_1, x_2, \ldots, x_n$ de X
$Xbar = (1/n) \sum x_i$ $ET(X) = [(1/ (n-1)) \sum (x_i - Xbar)^2]^{0,5}$
Xcr forme centrée-réduite de X
$Xcr = (X - Xbar) / ET(X)$
moyenne (Xcr) = 0 écart-type (Xcr) = 1

X : variable catégorique variant à k modalités (k ≥ 3) : codage à effet

modalités de X : $m_1, m_2, \ldots, m_k$
création de k – 1 variables $U_1, U_2, \ldots, U_{k-1}$ à valeurs -1, 0, 1
choix d'une modalité de référence disons m_k
m_k choix arbitraire

X	U_1	U_2	$U_3 \ldots U_{k-2}$	U_{k-1}
m_1	1	0	0 …… 0	0
m_2	0	1	0 …… 0	0
……				
m_{k-1}	0	0	0 …… 0	1
m_k	-1	-1	-1 ….. -1	-1

codage disjonctif complet

exemple avec k = 5 modalités

X	U_1	U_2	U_3	U_4	U_5
m1	1	0	0	0	0
m2	0	1	0	0	0
m3	0	0	1	0	0
m4	0	0	0	1	0
m5	0	0	0	0	1

problème : multi colinéarité
$U_1 + U_2 + U_3 + U_4 + U_5 = 1$
alternative : disjonctif complet réduit
enlèvement d'une variable, disons U_5
variables de codage retenues: $U_1\ U_2\ U_3\ U_4$

modèle : $Y = \beta_0 + \beta_1 * U_1 + \beta_2 * U_2 + \beta_3 * U_3 + \beta_4 * U_4$

Si $U_1 = U_2 = U_3 = U_4 = 0$ $(X = m_5)$ alors $Y = \beta_0$

β_0 effet général confondu avec l'effet de la modalité m_5

X : variable catégorique variant à 2 modalités m1 et m2

X remplacée par variable U définit par :
U = - 1 si X = m1
U = +1 si X = m2
cas particulier : codage à effet
L'assignation de m1 à -1 est arbitraire.

STATISTICA utilise le codage à effet

Exemple : codage variables catégoriques - modèle de régression – module GRM

data = voitures

NOM	CYL	PUIS	LON	LAR	POIDS	VITESSE	NAT	FINITION	Y_PRIX
ALFASUD-TI-1350	1350	79	393	161	870	165	I	B	30570
AUDI-100-L	1588	85	468	177	1110	160	D	TB	39990
SIMCA-1307-GLS	1294	68	424	168	1050	152	F	M	29600
CITROEN-GS-CLUB	1222	59	412	161	930	151	F	M	28250
FIAT-132-1600GLS	1585	98	439	164	1105	165	I	B	34900
LANCIA-BETA-1300	1297	82	429	169	1080	160	I	TB	35480
PEUGEOT-504	1796	79	449	169	1160	154	F	B	32300
RENAULT-16-TL	1565	55	424	163	1010	140	F	B	32000
RENAULT-30-TS	2664	128	452	173	1320	180	F	TB	47700
TOYOTA-COROLLA	1166	55	399	157	815	140	J	M	26540
ALFETTA-1.66	1570	109	428	162	1060	175	I	TB	42395
PRINCESS-1800-HL	1798	82	445	172	1160	158	GB	B	33990
DATSUN-200L	1998	115	469	169	1370	160	J	TB	43980
TAUNUS-2000-GL	1993	98	438	170	1080	167	D	B	35010
RANCHO	1442	80	431	166	1129	144	F	TB	39450
MAZDA-9295	1769	83							
OPEL-REKORD-L	1979	100							
LADA-1300	1294	68							

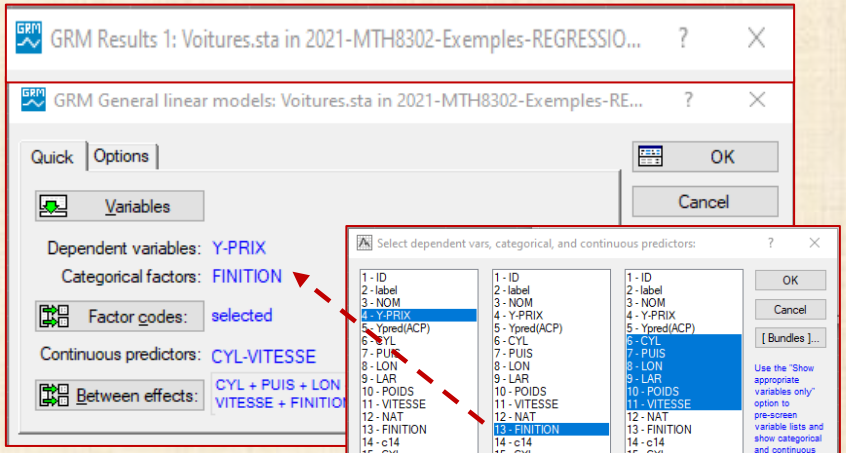
Variables catégoriques

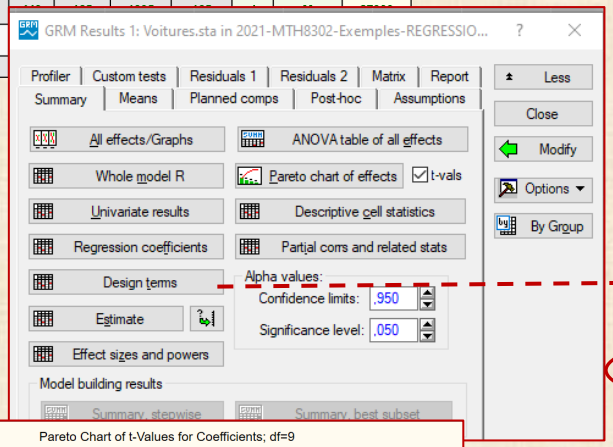
NAT

I : Italie
D: Allemagne
F : France
J : Japon
GB : Grande Bretagne
U : Russie

FINITION

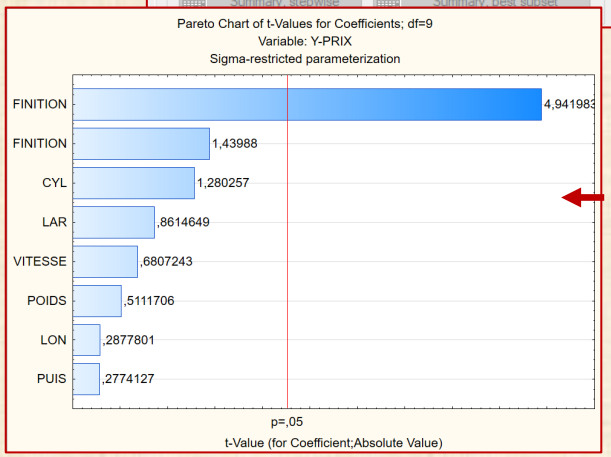
B : Bien
M : Moyenne
TB : Très Bien





Label	Column	Variable	Level of Variable	versus Level
Intercept	1			
CYL	2	CYL		
PUIS	3	PUIS		
LON	4	LON		
LAR	5	LAR		
POIDS	6	POIDS		
VITESSE	7	VITESSE		
FINITION	8	FINITION	B	M
FINITION	9	FINITION	TB	M

FINITION : codage à effet
création de 2 variables 1 / 0 / -1



Effect	Level of Effect	Column	Y-PRIX Param.	Y-PRIX Std.Err	Y-PRIX t	Y-PRIX p
Intercept		1	33319,83	26018,53	1,28062	0,232344
CYL		2	4,72	3,69	1,28026	0,232465
PUIS		3	-32,64	117,64	-0,27741	0,787730
LON		4	21,35	74,20	0,28778	0,780032
LAR		5	-214,04	248,46	-0,86146	0,411346
POIDS		6	7,25	14,17	0,51117	0,621535
VITESSE		7	90,24	132,56	0,68072	0,513176
FINITION	B	8	-1319,99	916,74	-1,43988	0,183764
FINITION	TB	9	6591,59	1333,79	4,94198	0,000800

Exemple 2: prix voiture vs 6 variables continues

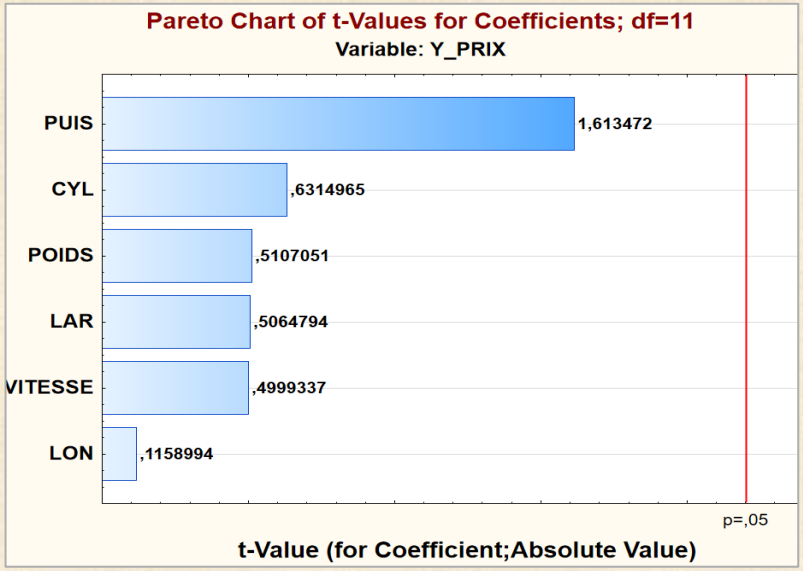
Regression Summary for Dependent Variable: y_PRIX

R = 0.84 R^2 = 0.71 Adjusted R^2 = 0.55 F(6,11) = 4.4690 p = 0,0156

	Beta	Std.Err. - of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			-8239,4	42718,4	-0,193	0,8506
CYL	-0,199	0,316	-3,5	5,6	-0,631	0,5406
PUIS	0,875	0,542	282,2	174,9	1,613	0,1349
LON	-0,051	0,436	-15,0	129,7	-0,116	0,9098
LAR	0,169	0,333	208,7	412,0	0,506	0,6225
POIDS	0,262	0,513	12,6	24,6	0,511	0,6197
VITESSE	-0,205	0,411	-111,1	222,3	-0,500	0,6270

ANOVA – DV = Y_PRIX

	Sums of Squares	df	Mean - Square	F	p-level
Regres	520591932	6	86765322	4,469	0,0156
Residu	213563858	11	19414896		
Total	734155790				



- test global F significatif
- aucun coefficient significatif !
- explication: multicolinéarité !
- matrice $(X'X)^{-1}$ mal conditionnée
  déterminant $X'X \approx 0$
- corrélations fortes entre les X

détection multicolinéarité

- ✓ étude de la matrice corrélations
- ✓ calcul des facteurs d'inflation de la variance (VIF)
- ✓ indice de conditionnement (IC) basé sur valeurs propres de X' X

3 Critères de détection multicolinéarité

- $R = (r_{ij})$ matrice de corrélation des variables X

indicateur

indicateur potentiel : $r_{ij} \geq 0.95$ mais non suffisant

- $VIF_j = 1 / (1 - R^2_j)$ R_j : coefficient de corrélation multiple entre X_j et toutes les autres variables X_i $j \neq i$

critère 1

critère 1 : max $VIF_j \geq 10$ c-à-d $R^2_j \geq 0{,}90$

- $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0$ valeurs propres de matrice corrélation R

critère 2

critère 2 : Indice Conditionnement = IC = $\lambda_1/\lambda_k > 100$ k = 2, 3, ...

faire avec Analyse en Composantes Principales (ACP)

Stratégies pour la modélisation

- méthodes de sélection de variables
- si on veut conserver toutes les variables
  - régression ACP composantes principales
  - régression pénalisée : RIDGE LASSO ELASTIC NET
  - régression PLS Partial Least Square

Stratégies pour la modélisation

- méthodes de sélection de variables
- méthodes de régression modifiée / pénalisée
  - régression ACP — Analyse Composantes Principales
  - régression pénalisée : RIDGE LASSO ELASTIC NET
  - régression PLS — Partial Least Square

Régression pénalisée : contrainte sur les coefficients de régression

l_1 : pénalité en valeur absolue — l_2 : pénalité en valeur quadratique
λ : paramètre d'ajustement (tuning) — α : paramètre de mélange de l_1 et l_2
N : nombre d'observations — p : nombre de variables

RIDGE

$$\hat{\beta}^{ridge} = \text{argmin}_{\beta}\left\{\sum_{i=1}^{N} -LogLikelihood(\beta;y_i) + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2\right\}$$

LASSO

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N} -LogLikelihood(\beta;y_i) + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

ELASTIC NET

$$\hat{\beta}^{enet} = \text{argmin}_{\beta}\left\{\sum_{i=1}^{N} -LogLikelihood(\beta;y_i) + \lambda\sum_{j=1}^{p}\left(\alpha|\beta_j| + \frac{(1-\alpha)}{2}\beta_j^2\right)\right\}$$

Exemple 2: voiture

multicolinéarité ?

Correlations N =18 (Casewise deletion of missing data)							
	CYL	PUIS	LON	LAR	POIDS	VITESSE	Y-PRIX
CYL	1,00						
PUIS	0,80	1,00					
LON	0,70	0,64	1,00				
LAR	0,63	0,52	0,85	1,00			
POIDS	0,79	0,77	0,87	0,72	1,00		
VITESSE	0,66	0,84	0,48	0,47	0,48	1,00	
Y-PRIX	0,64	0,80	0,64	0,55	0,75	0,58	1,00

0 corrélation ≥ 0,95

maximum r = 0,87

8 corrélations ≥ 0,70

sur 21 corrélations

indicateur 1 pas satisfait

Collinearity statistics for Terms in the equation			
	Tolerance	VIF	R square
CYL	0,265	3,77	0,735
PUIS	0,090	11,12	0,910
LON	0,139	7,20	0,861
LAR	0,238	4,20	0,762
POIDS	0,100	9,96	0,900
VITESSE	0,157	6,38	0,843

critère 2 satisfait

2 VIF ≥ 10 sur 6

multicolinéarité présente

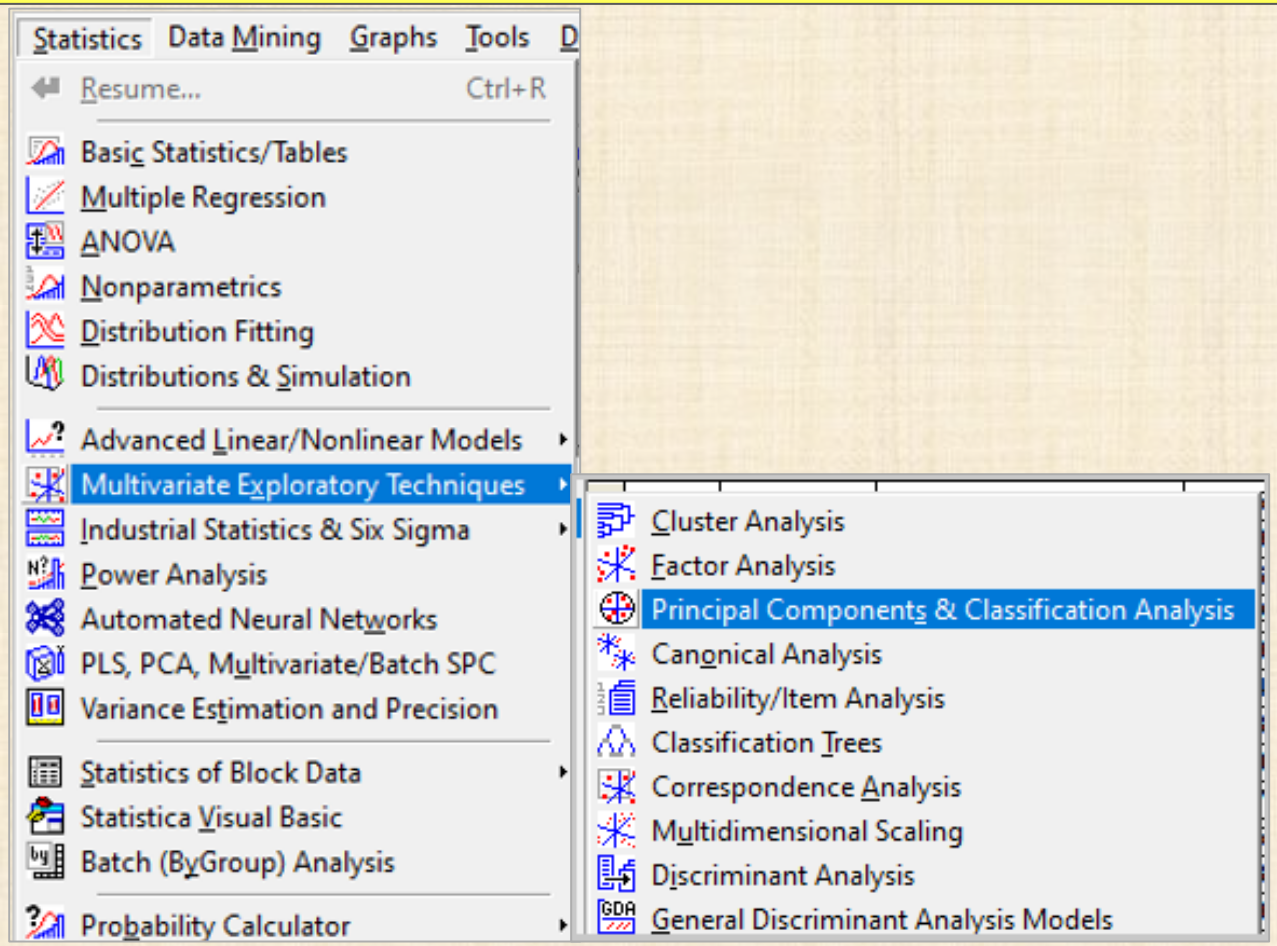
et le critère 3 ? …

Exemple 2: voiture

critère 3 : indice IC - Analyse Composantes Principales (ACP)

Statistics … Multivariate Exploratory Techniques

…. Principal Components & Classification



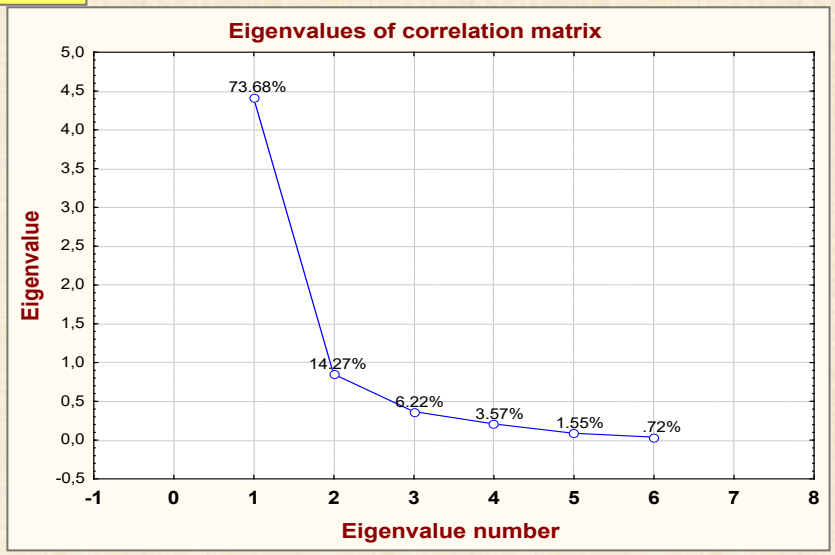
Exemple 2: voiture

critère 3 : satisfait

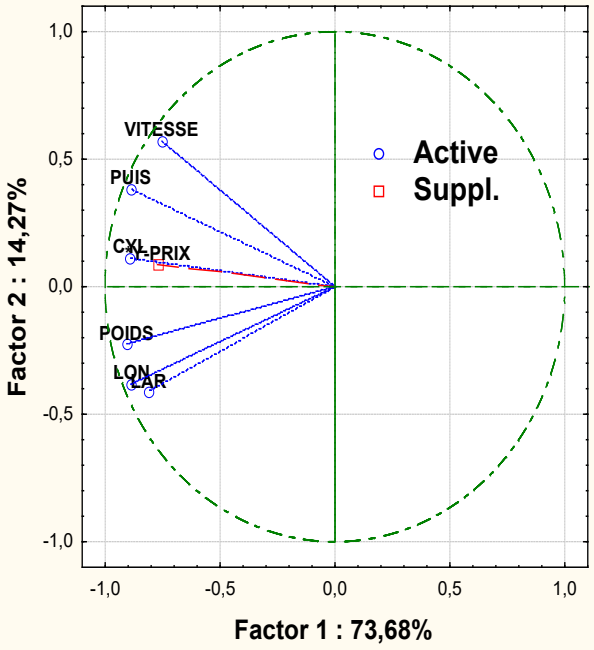
composantes principales

Principal Components & Classification

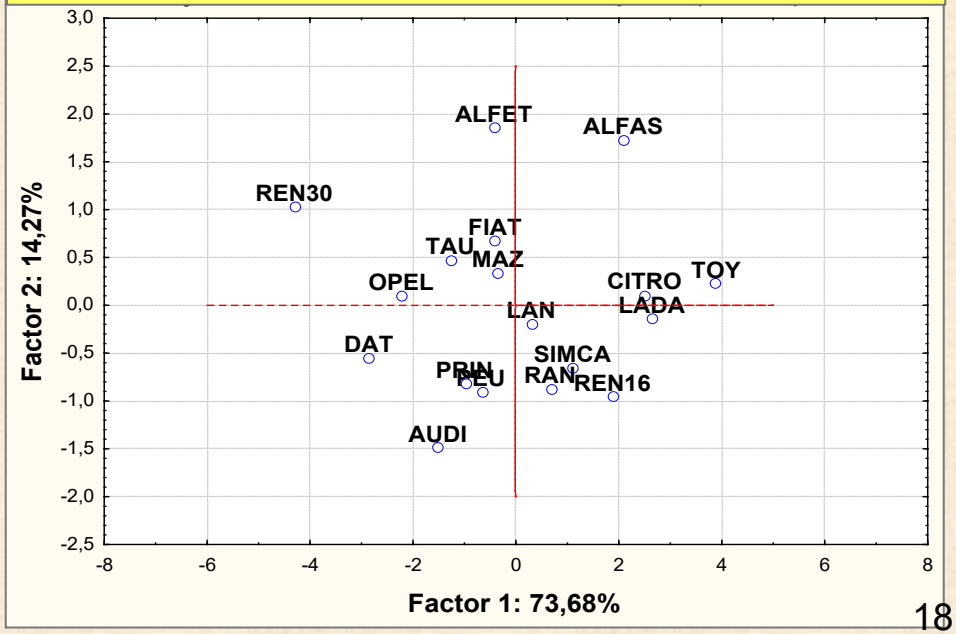
Eigenvalue	% Total variance	Cumulative eigenvalue	Cumulative %	Indice IC
4,42	73,7	4,42	73,7	1,00
0,86	14,3	5,28	87,9	5,16
0,37	6,2	5,65	94,2	11,85
0,21	3,6	5,86	97,7	20,67
0,09	1,5	5,96	99,3	47,64
0,04	0,7	6,00	100,0	102,12



variables selon 2 premières composantes



données selon 2 premières composantes



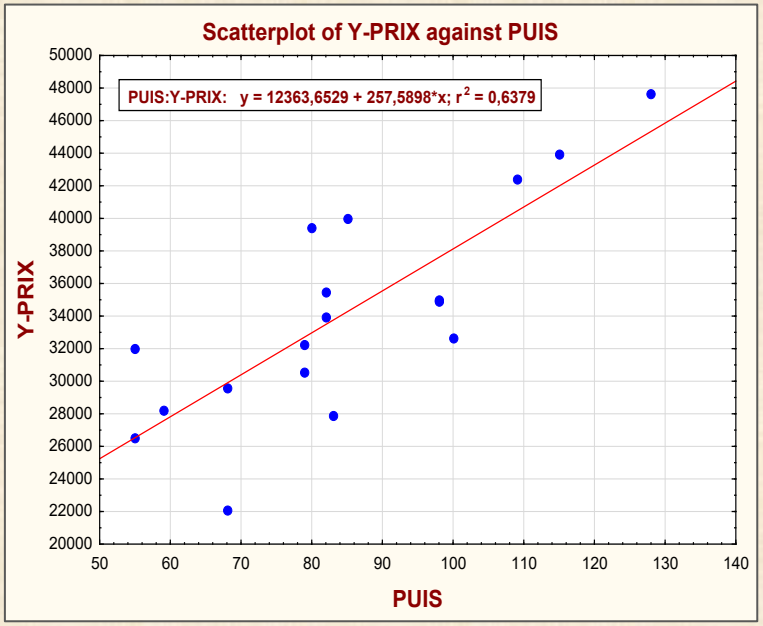
Modélisation méthode de sélection : forward stepwise : PUIS (étape 1) et POIDS (étape 2)

Variable	Step	Multiple R	Multiple R^2	R-square change	F to enter	P level
PUIS	1	0,7987	0,6379	0,6379	28,189	0,0001
POIDS	2	0,8286	0,6866	0,0487	2,331	0,1476

b*	Std.Err. b*	b	Std.Err. b	t(15)	p-level
		1775,60	8030,95	0,221	0,8280
0,5363	0,2246	172,97	72,42	2,388	0,0305
0,3429	0,2246	16,45	10,77	1,527	0,1476

POIDS pas significatif

modèle retenu :
Y_PRIX = 12363,65 + 257,59*PUIS
mais une seule variable ….



N=18	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			12363,65	4215,923	2,932609	0,009757
PUIS	0,798700	0,150432	257,59	48,516	5,309370	0,000071

si modèle standard n'est pas satisfaisant
(plusieurs raisons) : tests, mauvais signes des coefficients,…

SI on veut avoir TOUTES les variables dans le modèle

méthodes de régression alternatives

- régression ACP (composantes principales)
- régression RIDGE
- régression LASSO
- régression ELASTIC NET
- régression PLS (Partial Least Square)

Méthode régression sur composantes principales : régression ACP

Méthodologie

1. Effectue une ACP (Analyse en Composantes Principales) sur les X
2. Régression ascendante sur la première composante Z1,
   puis sur les deux premières Z1, Z2,…etc. (détails page suivante)

Propriétés

a) chaque composante Z est une combinaison linéaire de toutes les X
b) les composantes principales Z sont non corrélées entre elles
c) tous les prédicteurs X sont présents dans la modélisation de Y avec Z

modèle ACP : basé sur les composantes principales Z de X

modèle avec X : $Y = X\beta + \varepsilon$ modèle avec Z (comp. princ.) : $Y = Z\alpha + \varepsilon$

T : matrice p x p des p vecteurs propres T_j associés aux
p valeurs propres $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$ de X'X $T'T = I$ identité

$\Lambda = diag(\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p)$: matrice diagonale des p valeurs propres

$Z = [Z_1, Z_2, \ldots Z_P]$: composantes principales de X dans l'espace T $Z'Z = \Lambda$

liens : $X = ZT'$ $Z = XT$ $\alpha = T'\beta$ $\beta = T\alpha$ $T'X'XT = (XT)'XT = Z'Z = \Lambda$

modèle : $Y = X\beta = ZT'\beta = Z\alpha$ $\alpha = T'\beta$ $\beta = T\alpha$

estimation avec Z : $\hat{\alpha} = (Z'Z)^{-1}Z'Y = \Lambda^{-1}Z'Y$ $Var(\hat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1}$

$Var(\hat{\beta}) = Var(T\hat{\alpha}) = T var(\hat{\alpha})T' = T\Lambda^{-1}T'\sigma^2$ $Var(\hat{\beta}_j) = \sigma^2 (\Sigma\, t^2_{j\,i} / \lambda_i)$

$T_j = (t_{j1}, t_{j2}, \ldots, t_{jn})$ j-ème vecteur propre (format ligne) associé à λ_j

remarque : on fait les calculs avec les variables sous forme centrées-réduites Xcr car leurs valeurs propres et vecteurs propres sont identiques à ceux des variables originelles X

Mise en œuvre Régression en composantes principales

approprié si on - a diagnostiqué un problème de multicolinéarité
- veut obtenir un modèle contenant toutes les variables X

1. Ajouter au fichier initial: variables initiales X sous forme centrées-réduites Xcr.
   Avec Statistica : copier les variables initiales X dans une autre portion du fichier
   appliquer sur ces nouvelles colonnes, une opération de centrage-réduction.
   Elle est accessible via le bouton droit de la souris. (*standardized*).

2. Exécuter une analyse en composantes principales (ACP) sur les variables Xcr
   résultat: valeurs propres λ_1, λ_2,… et les vecteurs propres correspondants T1, T2, …

3. Les composantes de chacun des vecteurs Tj sont des coefficients de la combinaison linéaire des variables initiales Xcr. L'ACP exprime les données initiales Xcr dans l'espace des facteurs Factor1 = Z1, Factor2 = Z2,.. .
   L'étape 3 définit de nouvelles variables Z1, Z2,…, que l'on ajoute au fichier initial.

4. Le fichier initial contient maintenant les données dans 3 espaces:
   - initial (X) - centrées-réduites (Xcr) - facteurs (Z)

5. Régression sur composantes principales

5a) régresser Y sur les facteurs F avec la méthode sélection avant (forward)
   Régresser Y sur F1, régresser Y sur F1 et F2,…, etc
   Arrêter lorsqu'un nouveau F devient non significatif.

5b) Employer le modèle de l'étape précédente 5a) avec tous les F significatifs.
   L'équation de prédiction est Yprédit = a + b1F1 + b2F2 +….
   Les Fi peuvent s'exprimer en termes des Xcr.
   L'équation de prédiction Yprédit peut s'écrire en termes des Xcr
   Yprédit = c0 + c1*X1cr + c2*X2cr +…

5c) Écrire cette équation avec les variables originelles X : Yprédit = d + d1X1 + d2X2 +….

5d) Ajouter une nouvelle colonne au fichier initial avec ces prédictions.

5e) Comparer Yprédit avec Yobservé et les valeurs prédites avec régression ordinaire.

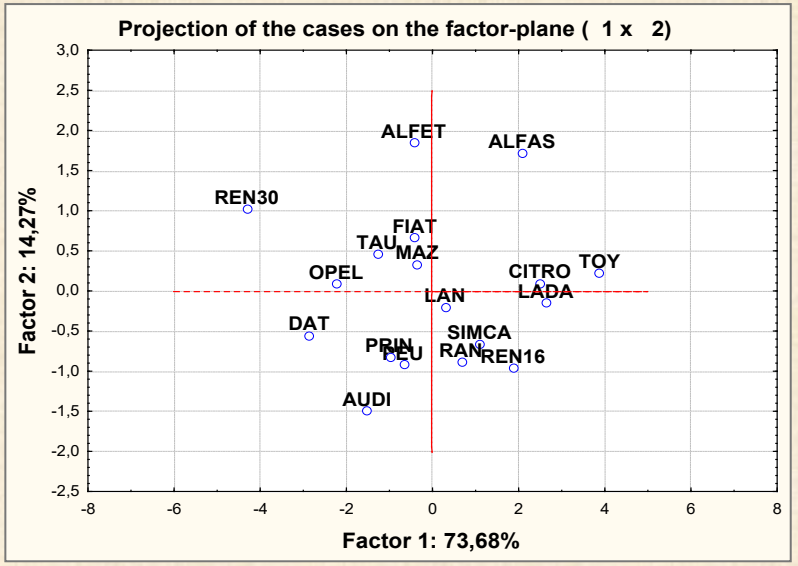
étude de cas détaillée : plus loin

données dans espace des facteurs

régression sur composantes principales

voiture	Factor 1 = Z1	Factor 2 = Z2	Factor 3 = Z3	Factor 4 = Z4	Factor 5 = Z5	Factor 6 =Z6
ALFASUD-TI-1350	2,08	1,74	0,56	0,196	-0,293	0,052
AUDI-100-L	-1,52	-1,48	1,28	-0,205	0,144	-0,318
SIMCA-1307-GLS	1,09	-0,66	0,44	-0,163	-0,365	0,264
CITROEN-GS-CLUB	2,50	0,11	0,14	-0,017	0,220	0,256
FIAT-132-1600GLS	-0,42	0,68	-0,19	-0,610	0,256	-0,036
LANCIA-BETA-1300	0,30	-0,19	0,66	-0,540	-0,433	0,194
PEUGEOT-504	-0,66	-0,91	-0,25	0,197	0,203	0,150
RENAULT-16-TL	1,89	-0,95	-0,60	0,613	0,285	0,106
RENAULT-30-TS	-4,29	1,03	-0,58	0,823	-0,364	0,043
TOYOTA-COROLLA	3,87	0,23	-0,29	0,258	0,271	-0,320
ALFETTA-1.66	-0,43	1,86	0,02	-0,738	0,163	-0,053
PRINCESS-1800-HL	-0,99	-0,82	0,21	0,295	-0,180	0,180
DATSUN-200L	-2,86	-0,54	-1,21	-0,750	0,053	-0,056
TAUNUS-2000-GL	-1,28	0,47	0,27	0,566	-0,065	-0,246
RANCHO	0,67	-0,87	-0,61	-0,348	-0,366	-0,118
MAZDA-9295	-0,37	0,35	-0,07	0,100	0,512	0,329
OPEL-REKORD-L	-2,23	0,10	0,77	0,230	0,329	-0,152
LADA-1300	2,63	-0,14	-0,56	0,093	-0,371	-0,275

Variables dans espace des facteurs



Z1 = Factor 1 Z2 = Factor 2 , …
sont orthogonales
donc pas de multicolinéarité

Relations entre variables centrée-réduites et facteurs

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
CYLcr	-0,893	0,115	-0,216	0,374	0,046	-0,012
PUIScr	-0,887	0,385	-0,113	-0,165	-0,089	-0,132
LONcr	-0,886	-0,381	0,041	-0,129	0,223	-0,040
LARcr	-0,814	-0,413	0,369	0,098	-0,146	-0,023
POIDScr	-0,905	-0,225	-0,296	-0,140	-0,093	0,121
VITESSEcr	-0,755	0,574	0,297	-0,034	0,057	0,095

ensuite :
régression sur Z1, Z2 …
avec méthode sélection variables
méthode = forward

Régression sur composantes principales

régression basée composante principale Z1 est significative

régression Z1, Z2
Z2 n'est pas significative

Regression Summary for Dependent Variable: Y-PRIX
R = 0,77 R² = 0,60 Adjusted R² = 0,57

	b*	Std.Err. b*	b	Std.Err b	t(16)	p-level
Interc			34158,6	1013,92	33,69	0,00000
Factor 1	- 0,772	0,159	- 2414,4	496,20	- 4,87	0,00017

(1) Yprédit = 34158,6 - 2414,4*Z1

Z1 en fonction des variables centrées-réduites Xcr

(2) Z1 = - 0,893*CYLcr - 0,887*PUIScr - 0,886*LONcr - 0,814*LARcr - 0,905*POIDScr - 0,359*VITESSEcr

(2) dans (1) donne (3)

(3) Yprédit = 34158,6 + 2156,1*CYLcr + 2141,6*PUIScr + 1965,3*LONcr + 1965,3*LARcr + 2185,0*POIDScr + 866,7*VITESSEcr

réécrire (3) avec les variables originelles X

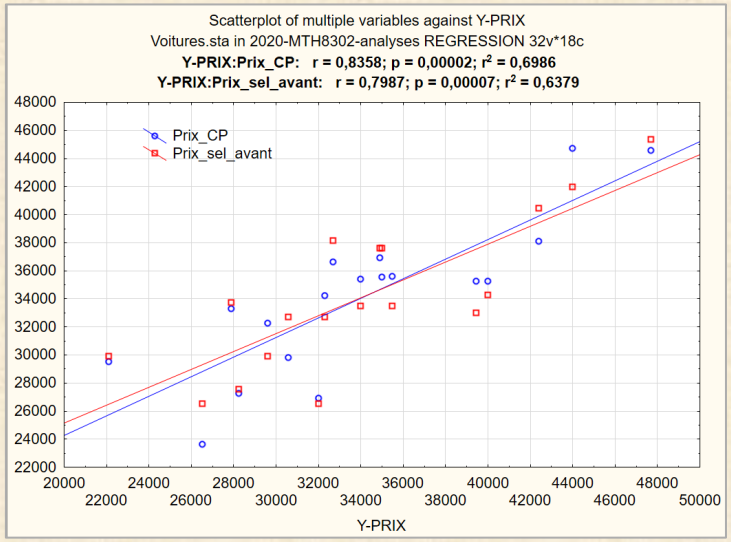
(4) Yprédit = a + b1*CYL + b2*PUIS + b3*LON + b4*LAR + b5*POIDS + b6*VITESSE
= - 9856,87 - 4,02*CYL + 181,54*PUIS - 42,92*LON + 141,91*LAR + 26,31*POIDS + 11,22*VITESSE

Comparer éq. (4) avec éq. (5) obtenu par la méthode de sélection avant

(5) Yprédit = 12363,65 + 257,59*PUIS

modèle retenu

basé sur Z1 seulement



Régression « RIDGE »

Méthode Hoerl et Kennard

$b_{MC} = \widehat{\beta} = (X'X)^{-1} X'y$ estimateur de moindres carrés (MC)

nouvelle classe d'estimateurs appelés RIDGE (R)

$b_R(k) = (X'X + k\, I\,)^{-1} X'y$ $\quad k > 0$ $\quad b_R(0) = b_{MC}$

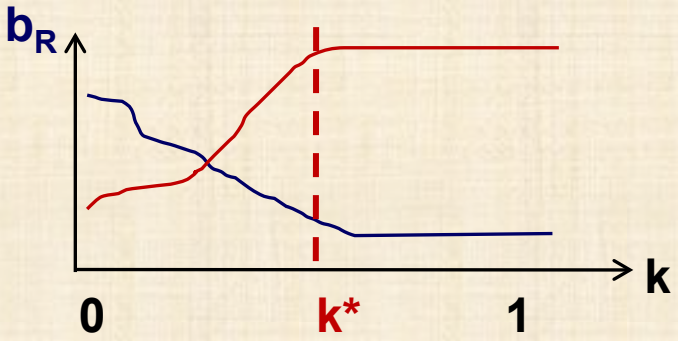
k : paramètre biaisant - à déterminer

$b_R = c\, b_{MC}$ $\quad 0 < c < 1$ **interprétation**

- estimateur biaisé rétréci d'erreur minimale
- évite l'instabilité des coefficients b_{MC} en multicolinéarité

Choix de k graphique de b_R en fonction de k («Ridge Trace»)

recherche d'une petite valeur k* avec b_R quasi constants

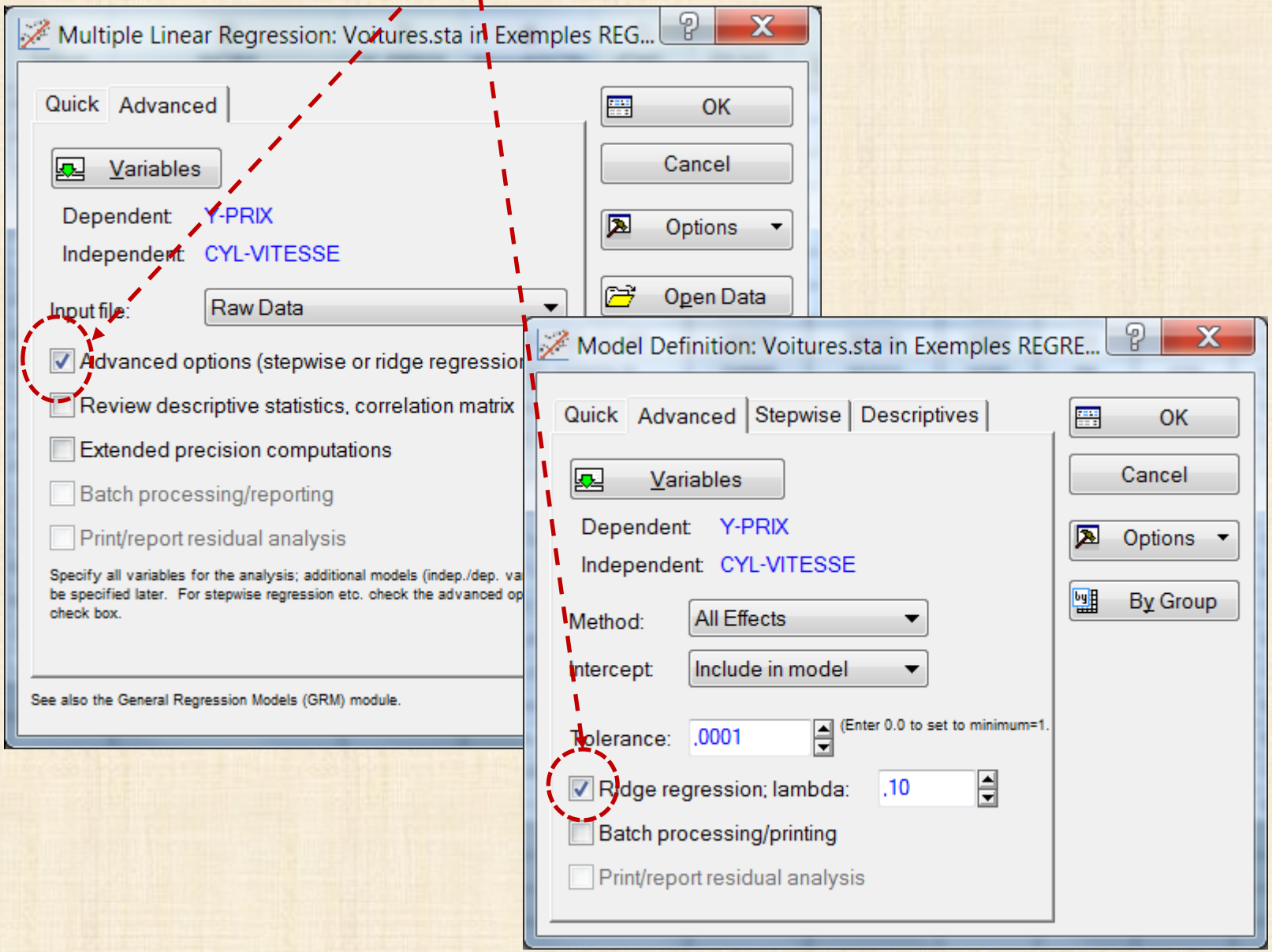


k * : choix subjectif !

recommandation

préférable k* le plus près de 0
même si la stabilité n'est pas parfaite
…. elle ne le saura jamais

Régression « RIDGE » : mise en œuvre avec Statistica



Régression « RIDGE » : mise en oeuvre

k = 0,1

Ridge Regression Summary for Dependent Variable: Y_PRIX
l=,10000 R= ,80904640 R²= ,65455607 Adjusted R²= ,4661321
F(6,11)=3,4738 p<,03544 Std.Error of estimate: 4801,6

N=18	b*	Std.Err. of b*	b	Std.Err. of b	t(11)	p-value
Intercept			-12537,5	42313,45	-0,296301	0,772519
CYL	-0,073973	0,282561	-1,3	4,97	-0,261794	0,798318
PUIS	0,505704	0,351951	163,1	113,51	1,436858	0,178586
LON	0,042527	0,331374	12,6	98,50	0,128334	0,900201
LAR	0,063404	0,276077	78,4	341,43	0,229662	0,822569
POIDS	0,308675	0,341974	14,8	16,41	0,902627	0,386063
VITESSE	0,005721	0,285618	3,1	154,60	0,020029	0,984379

k = 0,2

Ridge Regression Summary for Dependent Variable: Y_PRIX)
l=,20000 R= ,79053290 R²= ,62494227 Adjusted R²= ,42036533
F(6,11)=3,0548 p<,05190 Std.Error of estimate: 5003,2

N=18	b*	Std.Err. of b*	b	Std.Err. of b	t(11)	p-value
Intercept			-15135,9	40946,82	-0,369648	0,718662
CYL	-0,014560	0,257258	-0,3	4,52	-0,056597	0,955881
PUIS	0,400763	0,292714	129,3	94,40	1,369129	0,198265
LON	0,073289	0,284133	21,8	84,46	0,257938	0,801217
LAR	0,051343	0,247578	63,5	306,19	0,207380	0,839502
POIDS	0,277734	0,287761	13,3	13,81	0,965154	0,355207
VITESSE	0,051032	0,247858	27,6	134,17	0,205890	0,840638

k = 0,1 à 0,7

k	inter	cyl	puis	lon	lar	poids	vitesse
0,0	-8239,36	-3,51	282,17	-15,04	208,69	12,57	-111,11
0,1	-12537,53	-1,30	163,10	12,64	78,41	14,81	3,10
0,2	-15135,91	-0,26	129,25	21,79	63,50	13,33	27,62
0,3	-16723,11	0,33	111,21	25,73	63,83	12,14	38,24
0,4	-17652,15	0,68	99,50	27,73	67,14	11,24	43,86
0,5	-18141,40	0,92	91,08	28,82	70,73	10,53	47,08
0,6	-18328,20	1,08	84,63	29,41	73,91	9,95	48,97
0,7	-18301,99	1,20	79,46	29,71	76,54	9,47	50,07

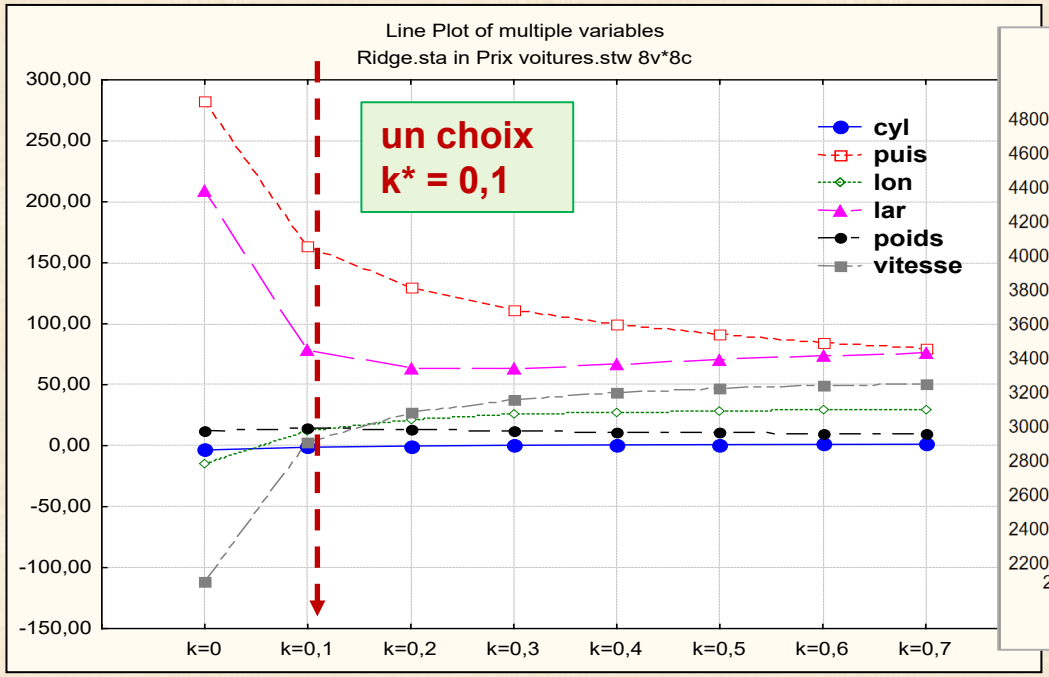
Régression « RIDGE » : mise en oeuvre

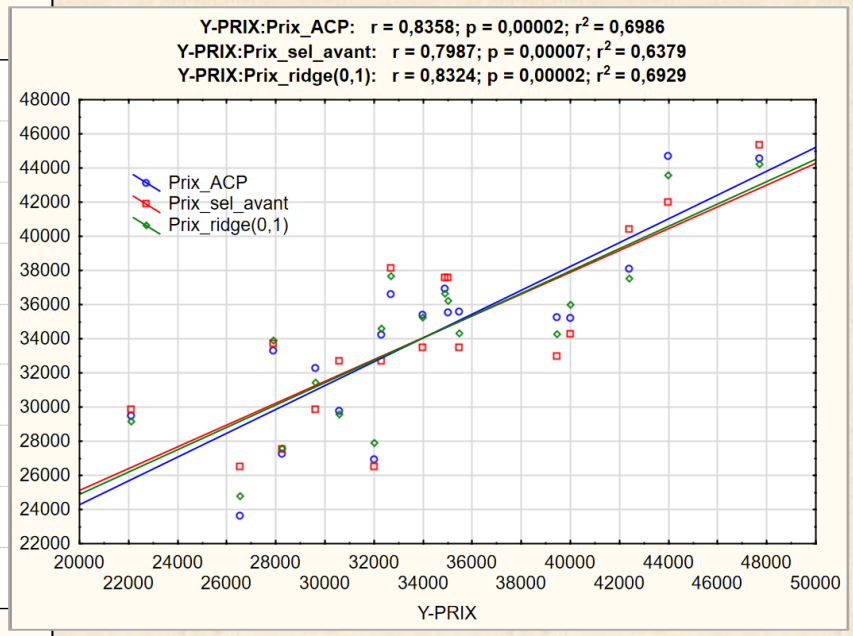
k	inter	cyl	puis	lon	lar	poids	vitesse
0,0	-8239,36	-3,51	282,17	-15,04	208,69	12,57	-111,11
0,1	-12537,53	-1,30	163,10	12,64	78,41	14,81	3,10
0,2	-15135,91	-0,26	129,25	21,79	63,50	13,33	27,62
0,3	-16723,11	0,33	111,21	25,73	63,83	12,14	38,24
0,4	-17652,15	0,68	99,50	27,73	67,14	11,24	43,86
0,5	-18141,40	0,92	91,08	28,82	70,73	10,53	47,08
0,6	-18328,20	1,08	84,63	29,41	73,91	9,95	48,97
0,7	-18301,99	1,20	79,46	29,71	76,54	9,47	50,07

Prix_ridge (0,1) = - 12537,53 - 1,30* CYL

+ 163,1*PUIS + 12,64*LON + 78,41*LAR

+ 14,81*POIDS + 3.1*VITESSE;





Exemple 3 : Acétylène (étude de cas) - 3 variables continues X / 1 variable continue Y

Source : Montgomery, Peck, Vining Introducion to Linear Regression Analysis, 4 ed, Wiley 2006 p. 328-364

L'étude originale : Kinugi, Tamara, Naito (1961) *New acetylene process uses hydrogen dillution,* Chem. Eng. Prog. vol 57 pp 43-49
Information sur l'acétylène (C2H2) : https://en.wikipedia.org/wiki/Acetylene
Acetylene (systematic name: ethyne) is the chemical compound with the formula C2H2.
It is a hydrocarbon and the simplest alkyne.
This colorless gas is widely used as a fuel and a chemical building block.
It is unstable in its pure form and thus is usually handled as a solution.
Pure acetylene is odorless, but commercial grades usually have a marked odor due to impurities.

Montgomery &all présentent une étude cas très détaillée de ces données.
Plusieurs méthodes sont présentées pour le traitement de variables explicatives présentant de la multicolinéarité.
Modèles alternatifs à la régression ordinaire : régression biaisée ridge, régression sur composantes principales.

VARIABLES

variable de réponse P : Pourcentage (%) de conversion de n-heptane en acetylene en fonction de
3 variables explicatives X : Temp, H2, Contact

facteurs Temp : température réaction (deg. C) : 3 valeurs 1100, 1200, 1300

H2 : ratio de H2 à n-Heptane (mole ratio) : 7 valeurs distinctes 5,3 7,5 9,0 ,,, 23,0

Cont : temps contact (sec) : 14 valeurs distinctes 0,0115 0,0120 0,0130 0,0135 .... 0,0980 0,3200 0,3800

versions centrés-réduites de Temp, H2, Cont : T H C

PLAN 16 essais 1 à 16 - Ce plan de collecte des données est une combinaison expérimental-observationnel.
Le facteur Température est contrôlé à 3 valeurs mais les 2 autres facteurs (H2 et Cont) sont observés et varient peu.
C'est typique dans la collecte de données avec un procédé en cours d'opération.
Les facteurs Temp et Cont sont corrélés dans la zone de valeurs Cont inférieures à 0,10
Les points ajoutés C D G H constituent de légères extrapolation de l'ensemble des 16 essais.

MODÈLE : on veut un modèle polynômial du second degré (surface de réponse) basé sur les 3 variables.

$$P = \beta_0 + \beta_1 * Temp + \beta_2 * H2 + \beta_3 * Cont + \beta_{12} * Temp * H2 + \beta_{13} * Temp * Cont + \beta_{23} * H2 * Cont + \beta_{11} * Temp * Temp + \beta_{22} * H2 * H\,2 + \beta_{33} * Cont * Cont$$

Le modèle sera ajusté dans les variables originelles Temp, H2, Cont et dans leur version centrés-réduites T, H et C.

DIFFICULTÉS : prédictions erronées, multicolinarités des variables explicatives.

MÉITHODES ALTERNATIVES : sélection de variables , régression biaisée ridge, régression sur composantes principales.

Exemple 3 : Acétylène (étude de cas) - 3 variables continues X / 1 variable continue Y

Source : Montgomery, Peck, Vining Introducion to Linear Regression Analysis, 4 ed, Wiley 2006 p. 328-364

L'étude originale : Kinugi, Tamara, Naito (1961) *New acetylene process uses hydrogen dillution,* Chem. Eng. Prog. vol 57 pp 43-49
Information sur l'acétylène (C2H2) : https://en.wikipedia.org/wiki/Acetylene
Acetylene (systematic name: ethyne) is the chemical compound with the formula C2H2.
It is a hydrocarbon and the simplest alkyne.
This colorless gas is widely used as a fuel and a chemical building block.
It is unstable in its pure form and thus is usually handled as a solution.
Pure acetylene is odorless, but commercial grades usually have a marked odor due to impurities.

Montgomery &all présentent une étude cas très détaillée de ces données.
Plusieurs méthodes sont présentées pour le traitement de variables explicatives présentant de la multicolinéarité.
Modèles alternatifs à la régression ordinaire : régression biaisée ridge, régression sur composantes principales.

VARIABLES

variable de réponse P : Pourcentage (%) de conversion de n-heptane en acetylene en fonction de
3 variables explicatives X : Temp, H2, Contact

facteurs Temp : température réaction (deg. C) : 3 valeurs 1100, 1200, 1300

H2 : ratio de H2 à n-Heptane (mole ratio) : 7 valeurs distinctes 5,3 7,5 9,0 ,,, 23,0

Cont : temps contact (sec) : 14 valeurs distinctes 0,0115 0,0120 0,0130 0,0135 .... 0,0980 0,3200 0,3800

versions centrés-réduites de Temp, H2, Cont : T H C

PLAN 16 essais 1 à 16 - Ce plan de collecte des données est une combinaison expérimental-observationnel.
Le facteur Température est contrôlé à 3 valeurs mais les 2 autres facteurs (H2 et Cont) sont observés et varient peu.
C'est typique dans la collecte de données avec un procédé en cours d'opération.
Les facteurs Temp et Cont sont corrélés dans la zone de valeurs Cont inférieures à 0,10
Les points ajoutés C D G H constituent de légères extrapolations de l'ensemble des 16 essais.

MODÈLE : on veut un modèle polynômial du second degré (surface de réponse) basé sur les 3 variables.

$$P = \beta_0 + \beta_1 * Temp + \beta_2 * H2 + \beta_3 * Cont + \beta_{12} * Temp * H2 + \beta_{13} * Temp * Cont + \beta_{23} * H2 * Cont + \beta_{11} * Temp * Temp + \beta_{22} * H2 * H2 + \beta_{33} * Cont * Cont$$

Le modèle sera ajusté dans les variables originelles Temp, H2, Cont et dans leur version centrés-réduites T, H et C.

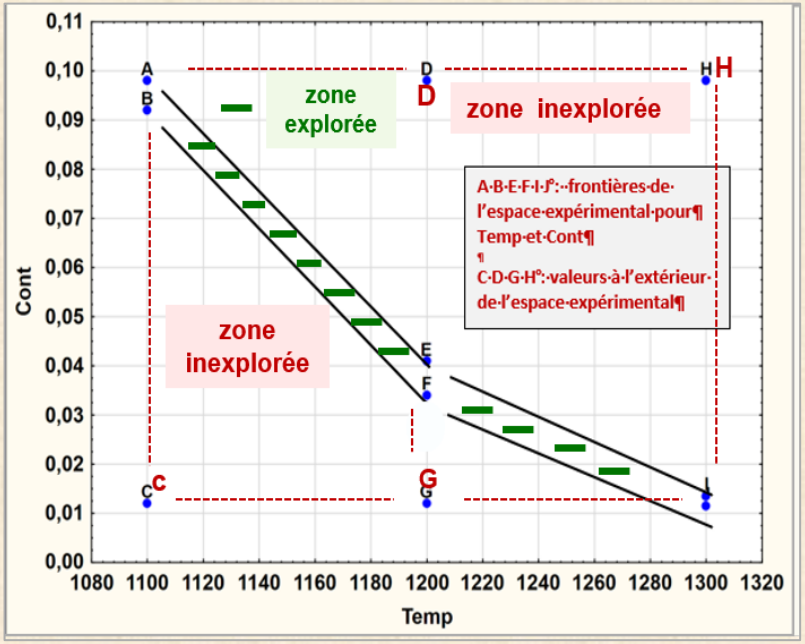
DIFFICULTÉS : prédictions erronées, **cause** : multicolinarités des variables explicatives. …. problème fréquent en REG MULT

MÉTHODES ALTERNATIVES : sélection de variables , régression biaisée ridge, régression sur composantes principales.

Exemple 3 : Acétylène (étude de cas développée)
3 variables continues X / 1 variable continue Y

1 new	2 ID_essai	3 ID2	4 P	5 Temp	6 H2	7 Cont	8 TempH2	9 TempCont	10 H2Cont	11 TempTemp	12 H2H2	13 ContCont	14 new12	15 T	16 H	17 C	18 TH	19 TC	20 HC	21 TT	22 HH	23 CC
plan	1		49,0	1300	7,5	0,0120	9750,0	15,6	0,09	1690000	56,25	0,0001	recodage	1,0853	-0,8731	-0,8949	-0,9476	-0,9712	0,7814	1,1779	0,7624	0,8008
de	2		50,2	1300	9,0	0,0120	11700,0	15,6	0,11	1690000	81,00	0,0001	de	1,0853	-0,6082	-0,8949	-0,6601	-0,9712	0,5443	1,1779	0,3699	0,8008
collecte	3	J	50,5	1300	11,0	0,0115	14300,0	15,0	0,13	1690000	121,00	0,0001	Temp	1,0853	-0,2550	-0,9107	-0,2767	-0,9884	0,2322	1,1779	0,0650	0,8293
des	4		48,5	1300	13,5	0,0130	17550,0	16,9	0,18	1690000	182,25	0,0002	H2	1,0853	0,1865	-0,8633	0,2025	-0,9369	-0,1610	1,1779	0,0348	0,7452
données	5	I	47,5	1300	17,0	0,0135	22100,0	17,6	0,23	1690000	289,00	0,0002	Cont	1,0853	0,8047	-0,8475	0,8733	-0,9198	-0,6820	1,1779	0,6475	0,7182
	6		44,5	1300	23,0	0,0120	29900,0	15,6	0,28	1690000	529,00	0,0001	en	1,0853	1,8644	-0,8949	2,0234	-0,9712	-1,6684	1,1779	3,4760	0,8008
	7		28,0	1200	5,3	0,0400	6360,0	48,0	0,21	1440000	28,09	0,0016	variables	-0,1550	-1,2617	-0,0099	0,1956	0,0015	0,0125	0,0240	1,5919	0,0001
	8		31,5	1200	7,5	0,0380	9000,0	45,6	0,29	1440000	56,25	0,0014	centrées	-0,1550	-0,8731	-0,0731	0,1354	0,0113	0,0638	0,0240	0,7624	0,0053
	9		34,5	1200	11,0	0,0320	13200,0	38,4	0,35	1440000	121,00	0,0010	réduites	-0,1550	-0,2550	-0,2627	0,0395	0,0407	0,0670	0,0240	0,0650	0,0690
	10		35,0	1200	13,5	0,0260	16200,0	31,2	0,35	1440000	182,25	0,0007	T	-0,1550	0,1865	-0,4524	-0,0289	0,0701	-0,0844	0,0240	0,0348	0,2046
	11	F	38,0	1200	17,0	0,0340	20400,0	40,8	0,58	1440000	289,00	0,0012	H	-0,1550	0,8047	-0,1995	-0,1248	0,0309	-0,1606	0,0240	0,6475	0,0398
	12	E	38,5	1200	23,0	0,0410	27600,0	49,2	0,94	1440000	529,00	0,0017	C	-0,1550	1,8644	0,0217	-0,2891	-0,0034	0,0405	0,0240	3,4760	0,0005
	13		15,0	1100	5,3	0,0840	5830,0	92,4	0,45	1210000	28,09	0,0071		-1,3954	-1,2617	1,3808	1,7606	-1,9268	-1,7422	1,9471	1,5919	1,9067
	14	A	17,0	1100	7,5	0,0980	8250,0	107,8	0,74	1210000	56,25	0,0096		-1,3954	-0,8731	1,8233	1,2184	-2,5443	-1,5920	1,9471	0,7624	3,3245
	15	B	20,5	1100	11,0	0,0920	12100,0	101,2	1,01	1210000	121,00	0,0085		-1,3954	-0,2550	1,6337	0,3558	-2,2796	-0,4166	1,9471	0,0650	2,6689
	16		29,5	1100	17,0	0,0860	18700,0	94,6	1,46	1210000	289,00	0,0074		-1,3954	0,8047	1,4440	-1,1229	-2,0150	1,1620	1,9471	0,6475	2,0853

37 essai	38 ID2	39 P	40 Temp	41 H2	42 Cont
1		49,0	1300	7,5	0,0120
2		50,2	1300	9,0	0,0120
3	J	50,5	1300	11,0	0,0115
4		48,5	1300	13,5	0,0130
5	I	47,5	1300	17,0	0,0135
6		44,5	1300	23,0	0,0120
7		28,0	1200	5,3	0,0400
8		31,5	1200	7,5	0,0380
9		34,5	1200	11,0	0,0320
10	G	35,(	1200	13,5	0,0260
11	F	38,0	1200	17,0	0,0340
12	E	38,5	1200	23,0	0,0410
13		15,0	1100	5,3	0,0840
14	A	17,0	1100	7,5	0,0980
15	B	20,5	1100	11,0	0,0920
16		29,5	1100	17,0	0,0860
extrapol	C		1100	23,0	0,0115
extrapol	D		1200	7,5	0,0980
extrapol	G		1200	23,0	0,0115
extrapol	H		1300	7,5	0,0980



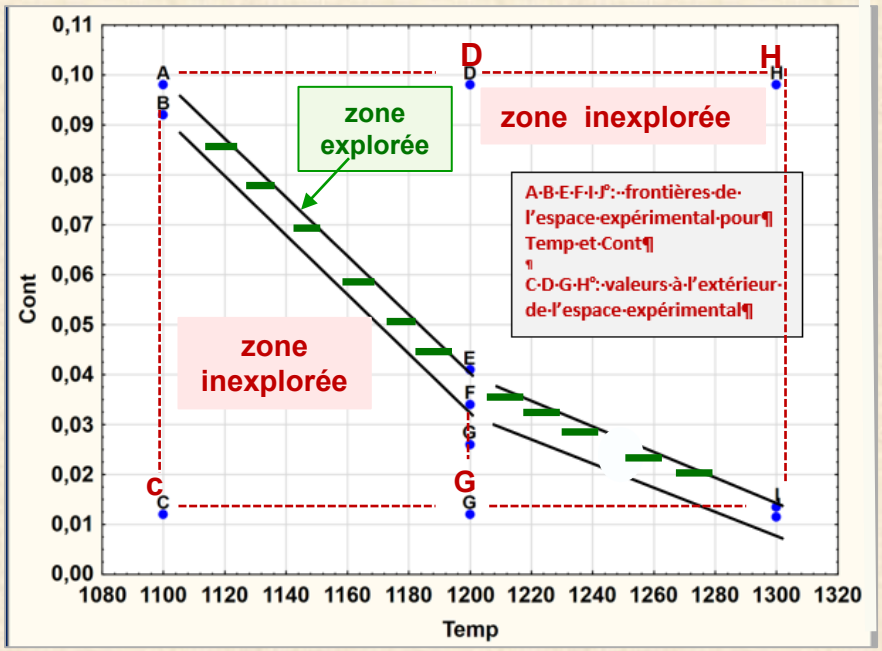
Exemple 3 : Acétylène (étude de cas développée)
3 variables continues X / 1 variable continue Y

1 new	2 ID_essai	3 ID2	4 P	5 Temp	6 H2	7 Cont	8 TempH2	9 TempCont	10 H2Cont	11 TempTemp	12 H2H2	13 ContCont	14 new12	15 T	16 H	17 C	18 TH	19 TC	20 HC	21 TT	22 HH	23 CC
plan	1		49,0	1300	7,5	0,0120	9750,0	15,6	0,09	1690000	56,25	0,0001	recodage de	1,0853	-0,8731	-0,8949	-0,9476	-0,9712	0,7814	1,1779	0,7624	0,8008
de	2		50,2	1300	9,0	0,0120	11700,0	15,6	0,11	1690000	81,00	0,0001	Temp	1,0853	-0,6082	-0,8949	-0,6601	-0,9712	0,5443	1,1779	0,3699	0,8008
collecte	3	J	50,5	1300	11,0	0,0115	14300,0	15,0	0,13	1690000	121,00	0,0001	H2	1,0853	-0,2550	-0,9107	-0,2767	-0,9884	0,2322	1,1779	0,0650	0,8293
des	4		48,5	1300	13,5	0,0130	17550,0	16,9	0,18	1690000	182,25	0,0002	Cont	1,0853	0,1865	-0,8633	0,2025	-0,9369	-0,1610	1,1779	0,0348	0,7452
données	5	I	47,5	1300	17,0	0,0135	22100,0	17,6	0,23	1690000	289,00	0,0002	en variables	1,0853	0,8047	-0,8475	0,8733	-0,9198	-0,6820	1,1779	0,6475	0,7182
	6		44,5	1300	23,0	0,0120	29900,0	15,6	0,28	1690000	529,00	0,0001	centrées réduites (cr)	1,0853	1,8644	-0,8949	2,0234	-0,9712	-1,6684	1,1779	3,4760	0,8008
	7		28,0	1200	5,3	0,0400	6360,0	48,0	0,21	1440000	28,09	0,0016		-0,1550	-1,2617	-0,0099	0,1956	0,0015	0,0125	0,0240	1,5919	0,0001
	8		31,5	1200	7,5	0,0380	9000,0	45,6	0,29	1440000	56,25	0,0014	T = Temp_cr	-0,1550	-0,8731	-0,0731	0,1354	0,0113	0,0638	0,0240	0,7624	0,0053
	9		34,5	1200	11,0	0,0320	13200,0	38,4	0,35	1440000	121,00	0,0010	H = H2_cr	-0,1550	-0,2550	-0,2627	0,0395	0,0407	0,0670	0,0240	0,0650	0,0690
	10		35,0	1200	13,5	0,0260	16200,0	31,2	0,35	1440000	182,25	0,0007	C = Cont_cr	-0,1550	0,1865	-0,4524	-0,0289	0,0701	-0,0844	0,0240	0,0348	0,2046
	11	F	38,0	1200	17,0	0,0340	20400,0	40,8	0,58	1440000	289,00	0,0012		-0,1550	0,8047	-0,1995	-0,1248	0,0309	-0,1606	0,0240	0,6475	0,0398
	12	E	38,5	1200	23,0	0,0410	27600,0	49,2	0,94	1440000	529,00	0,0017		-0,1550	1,8644	0,0217	-0,2891	-0,0034	0,0405	0,0240	0,3760	0,0005
	13		15,0	1100	5,3	0,0840	5830,0	92,4	0,45	1210000	28,09	0,0071		-1,3954	-1,2617	1,3808	1,7606	-1,9268	-1,7422	1,9471	1,5919	1,9067
	14	A	17,0	1100	7,5	0,0980	8250,0	107,8	0,74	1210000	56,25	0,0096		-1,3954	-0,8731	1,8233	1,2184	-2,5443	-1,5920	1,9471	0,7624	3,3245
	15	B	20,5	1100	11,0	0,0920	12100,0	101,2	1,01	1210000	121,00	0,0085		-1,3954	-0,2550	1,6337	0,3558	-2,2796	-0,4166	1,9471	0,0650	2,6689
	16		29,5	1100	17,0	0,0860	18700,0	94,6	1,46	1210000	289,00	0,0074		-1,3954	0,8047	1,4440	-1,1229	-2,0150	1,1620	1,9471	0,6475	2,0853

Extrapolation : points C D G H

prédiction avec le modèle développé avec les 16 essais ; valeurs de P =?

37 essai	38 ID2	39 P	40 Temp	41 H2	42 Cont
1		49,0	1300	7,5	0,0120
2		50,2	1300	9,0	0,0120
3	J	50,5	1300	11,0	0,0115
4		48,5	1300	13,5	0,0130
5	I	47,5	1300	17,0	0,0135
6		44,5	1300	23,0	0,0120
7		28,0	1200	5,3	0,0400
8		31,5	1200	7,5	0,0380
9		34,5	1200	11,0	0,0320
10	G	35,0	1200	13,5	0,0260
11	F	38,0	1200	17,0	0,0340
12	E	38,5	1200	23,0	0,0410
13		15,0	1100	5,3	0,0840
14	A	17,0	1100	7,5	0,0980
15	B	20,5	1100	11,0	0,0920
16		29,5	1100	17,0	0,0860
extrapol	C		1100	23,0	0,0115
extrapol	D		1200	7,5	0,0980
extrapol	G		1200	23,0	0,0115
extrapol	H		1300	7,5	0,0980



pourquoi une telle zone ?

réponse = ?

Exemple 3 : Acétylène - étude de cas

Modèle quadratique

Variables centrés-réduites

effet	SS	Degr. of Freedom	MS	F	p
Intercept	878,7179	1	878,7179	1081,369	0,000000
T	0,6408	1	0,6408	0,789	0,408719
H	66,5192	1	66,5192	81,860	0,000102
C	1,4262	1	1,4262	1,755	0,233461
TH	15,7623	1	15,7623	19,397	0,004547
TC	1,3386	1	1,3386	1,647	0,246663
HC	4,2107	1	4,2107	5,182	0,063116
TT	0,8392	1	0,8392	1,033	0,348741
HH	5,4791	1	5,4791	6,743	0,040844
CC	1,8391	1	1,8391	2,263	0,183182
Error	4,8756	6	0,8126		

effet	P Param.	P Std.Err	P t	P p
Intercept	35,8958	1,09158	32,88418	0,000000
T	4,0038	4,50870	0,88801	0,408719
H	2,7783	0,30708	9,04765	0,000102
C	-8,0423	6,07066	-1,32479	0,233461
TH	-6,4568	1,46603	-4,40425	0,004547
TC	-26,9804	21,02129	-1,28348	0,246663
HC	-3,7681	1,65535	-2,27634	0,063116
TT	-12,5236	12,32380	-1,01621	0,348741
HH	-0,9727	0,37460	-2,59668	0,040844
CC	-11,5932	7,70628	-1,50439	0,183182

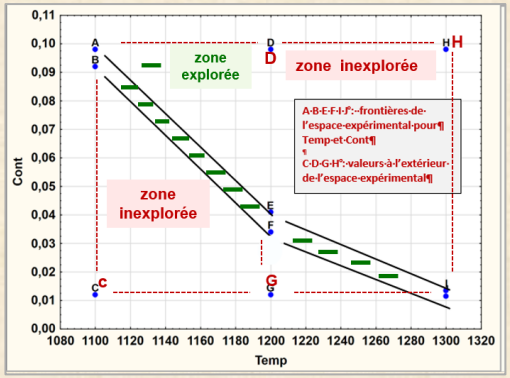
- prédictions identiques à celles avec les variables d'origine
- prédictions négatives aux limites extrapolées de l'espace d'observation
  C D G H

Collinearity statistics for terms in the equation
Sigma-restricted parameterization

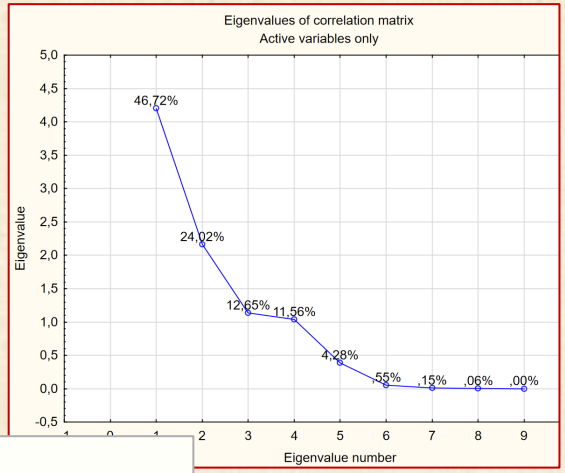
effet	Tolernce	Variance (VIF) Infl fac	R square
T	0,0026649	375,2478	0,9973351
H	0,5745043	1,7406	0,4254957
C	0,0014700	680,2800	0,9985300
TH	0,0322195	31,0371	0,9677805
TC	0,0001524	6563,3452	0,9998476
HC	0,0280810	35,6113	0,9719190
TT	0,0005674	1762,5754	0,9994326
HH	0,3160238	3,1643	0,6839762
CC	0,0008645	1156,7663	0,9991355

Matrice de corrélation

facteur	T	H	C	TH	TC	HC	TT	HH	CC	*P
T	1,000	0,224	-0,958	-0,132	0,443	0,206	-0,271	0,031	-0,577	0,945
H	0,224	1,000	-0,240	0,039	0,192	-0,023	-0,148	0,498	-0,224	0,370
C	-0,958	-0,240	1,000	0,195	-0,661	-0,274	0,501	-0,018	0,765	-0,914
TH	-0,132	0,039	0,195	1,000	-0,265	-0,974	0,246	0,398	0,275	-0,366
TC	0,443	0,192	-0,661	-0,265	1,000	0,324	-0,972	0,126	-0,972	0,421
HC	0,206	-0,023	-0,274	-0,974	0,324	1,000	-0,279	-0,375	-0,359	0,419
TT	-0,271	-0,148	0,501	0,246	-0,972	-0,279	1,000	-0,124	0,894	-0,249
HH	0,031	0,498	-0,018	0,398	0,126	-0,375	-0,124	1,000	-0,158	-0,038
CC	-0,577	-0,224	0,765	0,275	-0,972	-0,359	0,894	-0,158	1,000	-0,555
*P	0,945	0,370	-0,914	-0,366	0,421	0,419	-0,249	-0,038	-0,555	1,000



Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %	IC =4,205230/Eigenvalue;
4,20523	46,725	4,2052	46,72	1,00
2,16200	24,022	6,3672	70,75	1,95
1,13868	12,652	7,5059	83,40	3,69
1,04048	11,561	8,5464	94,96	4,04
0,38523	4,280	8,9316	99,24	10,92
0,04954	0,550	8,9811	99,79	84,89
0,01363	0,151	8,9948	99,94	308,63
0,00513	0,057	8,9999	100,00	820,08
0,00010	0,001	9,0000	100,00	43381,31



multicolinéarité ?

réponse : oui et sévère

modèles de régression à développer:

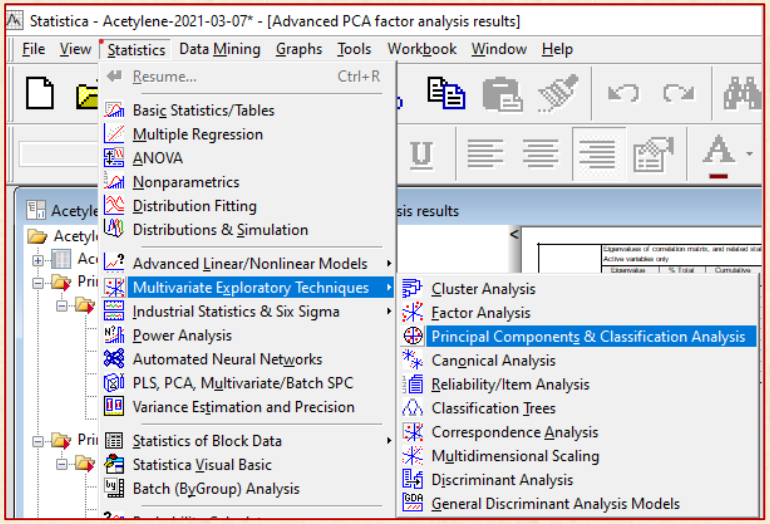
- régression sur composantes principales ACP
- régression Ridge

Exemple 3 : Acétylène - étude de cas | Analyse sur composantes principales (ACP)

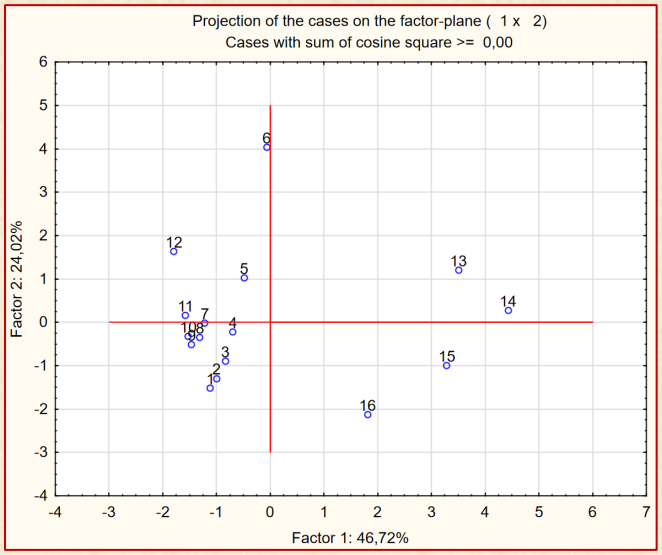
étape 1 : création des variables centrées-réduites

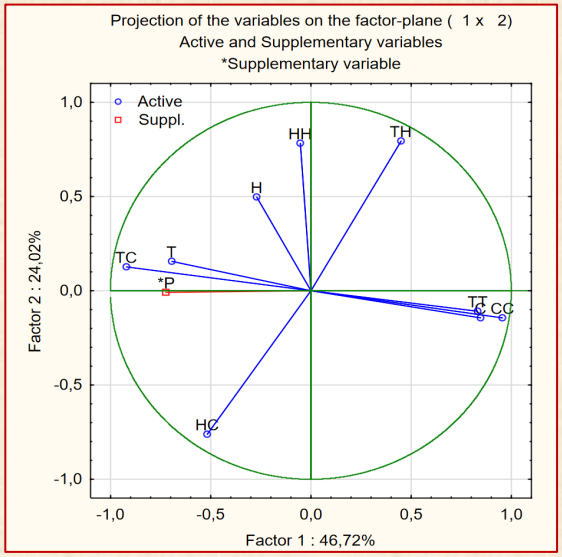
T H C TH TC HC TT HH CC

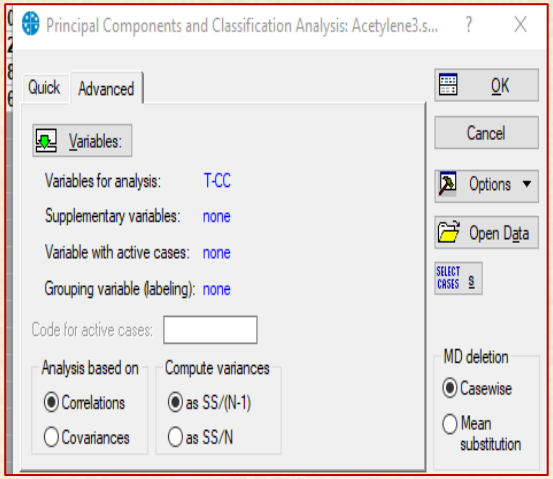
14 new12	15 T	16 H	17 C	18 TH	19 TC	20 HC	21 TT	22 HH	23 CC
recodage de	1,0853	-0,8731	-0,8949	-0,9476	-0,9712	0,7814	1,1779	0,7624	0,8008
Temp	1,0853	-0,6082	-0,8949	-0,6601	-0,9712	0,5443	1,1779	0,3699	0,8008
H2	1,0853	-0,2550	-0,9107	-0,2767	-0,9884	0,2322	1,1779	0,0650	0,8293
Cont	1,0853	0,1865	-0,8633	0,2025	-0,9369	-0,1610	1,1779	0,0348	0,7452
en variables	1,0853	0,8047	-0,8475	0,8733	-0,9198	-0,6820	1,1779	0,6475	0,7182
centrées réduites (cr)	1,0853	1,8644	-0,8949	2,0234	-0,9712	-1,6684	1,1779	3,4760	0,8008
	-0,1550	-1,2617	-0,0099	0,1956	0,0015	0,0125	0,0240	1,5919	0,0001
T = Temp_cr	-0,1550	-0,8731	-0,0731	0,1354	0,0113	0,0638	0,0240	0,7624	0,0053
H = H2_cr	-0,1550	-0,2550	-0,2627	0,0395	0,0407	0,0670	0,0240	0,0650	0,0690
C = Cont_cr	-0,1550	0,1865	-0,4524	-0,0289	0,0701	-0,0844	0,0240	0,0348	0,2046
	-0,1550	0,8047	-0,1995	-0,1248	0,0309	-0,1606	0,0240	0,6475	0,0398
	-0,1550	1,8644	0,0217	-0,2891	-0,0034	0,0405	0,0240	3,4760	0,0005
	-1,3954	-1,2617	1,3808	1,7606	-1,9268	-1,7422	1,9471	1,5919	1,9067
	-1,3954	-0,8731	1,8233	1,2184	-2,5443	-1,5920	1,9471	0,7624	3,3245
	-1,3954	-0,2550	1,6337	0,3558	-2,2796	-0,4166	1,9471	0,0650	2,6689
	-1,3954	0,8047	1,4440	-1,1229	-2,0150	1,1620	1,9471	0,6475	2,0853



étape 2 : Analyse en Composantes Principales (ACP) des variables T H C TH TC HC TT HH CC







Exemple 3 : Acétylène - étude de cas — régression sur composantes principales

résultat de ACP :
relation entre les variables centrées-réduites et les facteurs (vecteurs propres)

Factor 1 = FS1 … Factor 9 = FS9

27 c26	28 Factor 1	29 Factor 2	30 Factor 3	31 Factor 4	32 Factor 5	33 Factor 6	34 Factor 7	35 Factor 8	36 Factor 9
matrice des	-0,338691	0,105762	0,649400	-0,012103	-0,142553	0,249490	0,221473	-0,538744	-0,174280
vecteurs propres	-0,132359	0,339016	-0,001450	0,724723	0,583891	-0,020702	0,011304	-0,028780	0,003324
=	0,413622	-0,097999	-0,469273	0,075249	0,018469	-0,016032	0,168848	-0,712893	-0,236874
Factor 1 ... Factor 9	0,219512	0,540056	0,087054	-0,361627	0,166331	-0,366240	0,589696	0,110707	-0,002548
	-0,449170	0,086411	-0,287945	-0,189378	0,094467	-0,029220	-0,062038	0,150412	-0,797341
définies par des	-0,252758	-0,517197	-0,054713	0,344885	-0,201039	-0,315494	0,620188	0,148413	-0,008333
combinaisons des	0,405575	-0,074762	0,441687	0,219505	-0,144309	-0,540689	-0,318931	0,047691	-0,411690
variables initiales	-0,025556	0,531633	-0,221360	0,342909	-0,734417	0,070755	0,002471	0,075636	-0,005021
H T C ... CC	0,466601	-0,097142	0,143135	0,132722	0,034777	0,636550	0,290459	0,368523	-0,328882

résultat de ACP : données pour la régression de P sur les factor Scores FS

37 c36	38 FS1	39 FS2	40 FS3	41 FS4	42 FS5	43 FS6	44 FS7	45 FS8	46 FS9	4 P
matrice Z=XT	-0,541858	-1,03483	1,04940	0,17997	-1,73933	0,65799	-0,67570	0,73700	-0,22282	49,0
factor scores	-0,484909	-0,88335	1,16370	0,03805	-0,89133	0,38607	-0,51651	0,17246	0,20272	50,2
	-0,404711	-0,61354	1,29071	-0,07714	0,00220	0,16259	-0,21812	-0,10152	1,66613	50,5
FS1 - FS2 - FS3	-0,338737	-0,15215	1,31670	-0,14097	0,75248	-0,35943	-0,09467	-1,21447	-0,96562	48,5
FS4 - FS5- FS6	-0,234819	0,68959	1,27893	-0,00003	1,08427	-0,68770	0,45752	-1,24399	-1,76008	47,5
FS7 - FS8 - FS9	-0,029714	2,74477	0,96010	0,77229	-0,22326	-0,20498	1,09898	1,37214	1,05103	44,5
	-0,594157	-0,01498	-1,09689	-1,14777	-1,57877	-0,18586	1,35056	-0,39664	-0,56749	28,0
expression des	-0,638700	-0,23877	-0,92489	-1,08519	-0,36320	-0,41755	1,26886	-0,64541	0,37771	31,5
données dans	-0,714124	-0,35504	-0,72101	-0,83042	0,93745	-0,31602	0,65059	0,55591	0,77003	34,5
l'espace des facteurs	-0,743721	-0,22224	-0,62073	-0,56237	1,42962	0,40425	-0,62007	2,49222	-1,39618	35,0
	-0,766876	0,10404	-0,86170	0,07681	1,34739	0,35856	-1,58847	-0,96120	1,45076	38,0
nouvelles variables	-0,872296	1,10665	-1,51355	1,85531	-0,81255	0,91907	-0,85182	-0,95464	-0,59830	38,5
pour la régression en	1,711229	0,81627	-0,37880	-1,20224	-0,88810	-1,92494	-2,02909	0,14295	-0,07740	15,0
composantes principales	2,161909	0,18527	-0,10654	-0,56061	0,12897	2,56130	0,28717	-0,13059	-0,56190	17,0
	1,604726	-0,67915	-0,20970	0,33406	0,74545	0,07182	0,83091	-0,42657	1,02548	20,5
	0,886759	-1,45255	-0,62574	2,35026	0,06872	-1,42518	0,64987	0,60236	-0,39407	29,5

étape 3 : régression « stepwise forward » de P sur FS1, FS2,…, FS9

Exemple 3 : Acétylène - étude de cas | régression en composante principales

étape 3 : Régression « stepwise forward » de P sur FS1, FS2,…, FS9

Summary of stepwise regression; variable: P (Acetylene2.sta in Acetylene-2021-03-06)
Forward stepwise P to enter: ,05, P to remove: ,05

Effect	Steps	Degr. of Freedom	F to remove	P to remove	F to enter	P to enter	Effect status
"FS1"	Step Number 1	1			15,26626	0,001579	Entered
"FS2"		1			0,00075	0,978585	Out
"FS3"		1			9,83163	0,007298	Out
"FS4"		1			0,84880	0,372495	Out
"FS5"		1			0,00047	0,983045	Out
"FS6"		1			0,03265	0,859189	Out
"FS7"		1			0,02809	0,869289	Out
"FS8"		1			0,01945	0,891074	Out
"FS9"		1			0,00783	0,930763	Out
"FS1"	Step Number 2	1	15,2663	0,001579			In
"FS2"		1			0,00145	0,970207	Out
"FS3"		1			81,47973	0,000001	Entered
"FS4"		1			1,76428	0,206940	Out
"FS5"		1			0,00091	0,976411	Out
"FS6"		1			0,06355	0,804915	Out
"FS7"		1			0,05465	0,818802	Out
"FS8"		1			0,03781	0,848831	Out
"FS9"		1			0,01520	0,903767	Out
"FS1"	Step Number 3	1	103,0251	0,000000			In
"FS3"		1	81,4797	0,000001			In
"FS2"		1			0,00973	0,923045	Out
"FS4"		1			79,22761	0,000001	Entered
"FS5"		1			0,00610	0,939045	Out
"FS6"		1			0,43979	0,519759	Out
"FS7"		1			0,37655	0,550911	Out
"FS8"		1			0,25836	0,620462	Out
"FS9"		1			0,10272	0,754103	Out
"FS1"	Step Number 4	1	722,9796	0,000000			In
"FS3"		1	571,7847	0,000000			In
"FS4"		1	79,2276	0,000001			In
"FS2"		1			0,06818	0,798824	Out
"FS5"		1			0,04264	0,840182	Out
"FS6"		1			4,04313	0,069510	Out
"FS7"		1			3,30979	0,096167	Out
"FS8"		1			2,09881	0,175314	Out
"FS9"		1			0,75870	0,402344	Out

Effect	Comment (B/Z/P)	P Param.	P Std.Err	P t	P p
Intercept		36,10625	0,309462	116,6743	0,000000
FS1		-8,59379	0,319611	-26,8883	0,000000
FS2	Pooled				
FS3		7,64254	0,319611	23,9120	0,000000
FS4		2,84485	0,319611	8,9010	0,000001
FS5	Pooled				
FS6	Pooled				
FS7	Pooled				
FS8	Pooled				
FS9	Pooled				

retenues
FS1 FS3 FS4

modèle de P sur les Factor Score (FS)

P = 36,10 - 8,59*FS1 + 7,64*FS3 + 2,84*FS4

27 c26	28 Factor 1	29 Factor 2	30 Factor 3	31 Factor 4	32 Factor 5	33 Factor 6	34 Factor 7	35 Factor 8	36 Factor 9	
matrice des	-0,338691	0,105762	0,649400	-0,012103	-0,142553	0,249490	0,221473	-0,538744	-0,174280	T
vecteurs propres	-0,132359	0,339016	-0,001450	0,724723	0,583891	-0,020702	0,011304	-0,028780	0,003324	H
=	0,413622	-0,097999	-0,469273	0,075249	0,018469	-0,016032	0,168848	-0,712893	-0,236874	C
Factor 1 ... Factor 9	0,219512	0,540056	0,087054	-0,361627	0,166331	-0,366240	0,589696	0,110707	-0,002548	TH
	-0,449170	0,086411	-0,287945	-0,189378	0,094467	-0,029220	-0,062038	0,150412	-0,797341	TC
définies par des	-0,252758	-0,517197	-0,054713	0,344885	-0,201039	-0,315494	0,620188	0,148413	-0,008333	HC
combinaisons des	0,405575	-0,074762	0,441687	0,219505	-0,144309	-0,540689	-0,318931	0,047691	-0,411690	TT
variables initiales	-0,025556	0,531633	-0,221360	0,342909	-0,734417	0,070755	0,002471	0,075636	-0,005021	HH
H T C ... CC	0,466601	-0,097142	0,143135	0,132722	0,034777	0,636550	0,290459	0,368523	-0,328882	CC

étape 4 : relation entre les FS et les variables centrées-réduites T H C TH TC HC TT HH CC

FS1 = 0,3387*T - 0,1324*H + 0,4136*C + 0,2195*T*H - 0,4492*T*C - 0,2527*H*C + 0,4058*T*T - 0,0255*H*H + 0,4666*C*C
FS3 = 0,6494*T - 0,0014*H - 0,4693*C + 0,0870*T*H - 0,2879*T*C - 0,0547*H*C + 0,4417*T*T - 0,2213*H*H + 0,1431*C*C
FS4 = - 0,0121*T + 0,7247*H + 0,0752*C - 0,3616*T*H - 0,1894*T*C + 0,3449*H*C + 0,2195*T*T + 0,3429*H*H + 0,1327*C*C

modèle de P sur les variables d'origine centrées-réduites via FS1 FS3 FS4

P = 36,10 - 8,59*(0,3387*T - 0,1324*H + 0,4136*C + 0,2195*T*H - 0,4492*T*C - 0,2527*H*C + 0,4058*T*T - 0,0255*H*H + 0,4666*C*C)
+ 7,64*(0,6494*T - 0,0014*H - 0,4693*C + 0,0870*T*H - 0,2879*T*C - 0,0547*H*C + 0,4417*T*T - 0,2213*H*H + 0,1431*C*C)
+ 2,84*(- 0,0121*T + 0,7247*H + 0,0752*C - 0,3616*T*H - 0,1894*T*C + 0,3449*H*C + 0,2195*T*T + 0,3429*H*H + 0,1327*C*C)

étape 5 : modèle de P sur les variables d'origine centrées-réduites : P = 36,10 - 8,59*FS1 + 7,64*FS3 + 2,84*FS4

P = 36,10 + T* [(-8,59*(0,3387) + 7,64*(0,6494) + 2,84*(- 0,0121)] + H*[(-8,59)*(-0,1324) + 7,64*(-0,0014) + 2,84*(0,7247)]
+ C*[(-8,59)*(0,4136) + 7,64*(-0,4693) + 2,84*(0,0752)] + T*H*[(-8,59)*(0,2195) + 7,64*(0,0870) + 2,84*(-0,3616)]
T*C*[(-8,59)*(-0,4492) + 7,64*(-0,2879) + 2,84*(-0,1894)] + H*C*[(-8,59)*(-0,2527) + 7,64*(-0,0547) + 2,84*(0,3449)]
T*T*[(-8,59)*(0,4058) + 7,64*(0,4417) + 2,84*(0,2195)] + H*H*[(-8,59)*(-0,0255) + 7,64*(-0,2213) + 2,84*(0,3429)]
C*C*[(-8,59)*(0,4666) + 7,64*(0,1431) + 2,84*(0,1327)]

P = 36,10 + 2,02*T + 0,91*H - 6,92*C - 2,25*T*H + 1,12*T*C + 2,73*H*C + 0,51*T*T - 0,50*H*H - 2,54*C*C

Exemple 3 : Acétylène - étude de cas — régression sur les facteurs de l'ACP

étape 5: modèle de P sur les variables d'origine centrées-réduites : P = 36,10 - 8,59*FS1 + 7,64*FS3 + 2,84*FS4

P = 36,10 + T* [(-8,59*(0,3387) + 7,64*(0,6494) + 2,84*(- 0,0121)] + H*[(-8,59)*(-0,1324) + 7,64*(-0,0014) + 2,84*(0,7247)]
+ C*[(-8,59)*(0,4136) + 7,64*(-0,4693) + 2,84*(0,0752)] + T*H*[(-8,59)*(0,2195) + 7,64*(0,0870) + 2,84*(-0,3616)]
T*C*[(-8,59)*(-0,4492) + 7,64*(-0,2879) + 2,84*(-0,1894)] + H*C*[(-8,59)*(-0,2527) + 7,64*(-0,0547) + 2,84*(0,3449)]
T*T*[(-8,59)*(0,4058) + 7,64*(0,4417) + 2,84*(0,2195)] + H*H*[(-8,59)*(-0,0255) + 7,64*(-0,2213) + 2,84*(0,3429)]
C*C*[(-8,59)*(0,4666) + 7,64*(0,1431) + 2,84*(0,1327)]

P = 36,10 + 2,02*T + 0,91*H - 6,92*C - 2,25*T*H + 1,12*T*C + 2,73*H*C + 0,51*T*T - 0,50*H*H - 2,54*C*C

étape 6 : modèle de P sur les variables d'origine : Temp H2 Cont

Relations entre les variables centrées-réduites T H C et les variables d'origine Temp H2 Cont

T = (Temp - 1212,50) / 80,62 H = (H2 - 12,44) / 5,56 C = (Cont - 0,040) / 0,032

T = 0,0124*Temp - 15,06 H = 0,1798*H2 - 2,23 C = 31,25*Cont - 1,25

1 stat	2 Temp	3 H2	4 Cont
MEAN	1212,50	12,44	0,040
SD	80,62	5,66	0,032

Modèle de P en fonction de Temp H2 Cont

P = 36,10 + 2,02*(0,0124*Temp - 15,06) + 0,91*(0,1798*H2 - 2,23) + (-6,92)*(31,25*Cont - 1,25)
+ (-2,25)*(0,0124*Temp - 15,06)*(0,1798*H2 - 2,23) + 1,12*(0,0124*Temp - 15,06)*(31,25*Cont - 1,25) + 2,73*(0,1798*H2 - 2,23)*(31,25*Cont - 1,25)
+ 0,51*(0,0124*Temp - 15,06)^2 + (- 0,50)*(0,1798*H2 - 2,23)^2 + (-2,54)*(31,25*Cont - 1,25)^2

P = c0 + c1*Temp + c2*H2 + c3*Cont + c4*Temp*H2 + c5*TEMP*Cont + c6*H2*Cont + c7*Temp^2 + c8*H2^2 + c9*Cont^2

c0 = 36,10 + (2,02)*(-15,06) + (0,91)*(-2,23) + (-6,92)*(-1,25) + (-2,25)*(-15,06)*(-2,23) + (1,12)*(-15,06)*(-1,25) + (2,73)*(-2,23)*(-1,15)
+ (0,51)*(-15,06)*(-15,06) + (-0,50)*(-2,23)*(-2,23) - (-2,54)*(-1,25)*(-1,25) = 81,97

c1 = (2,02)*(0,0124) + (-2,25)*(0,0124)*(-2,23) + (1,12)*(0,0124)*(-1,25) + (0,51)*2*(0,0124)*(-15,06) = 0,12

c2 = (0,91)*(0,1798) + (-2,25)*(0,1798) + (2,73)*(0,1798) + (2,73)*(0,1798)*(-1,25) + (-0,50)*2*(0,1798)*(-2,23) = 0,04

c3 = (-6,92)*(31,25) + (1,12)*(-15,06)*(31,25) + (2,73)*(31,25)*(-2,23) + (-2,54)*2*(31,25)*(-1,25) = 735,16

c4 = (-2,25)*(0,0124)*(0,1798) = - 0,01

c5 = (1,12)*(0,0124)*(31,25) = 0,43

C6 = (2,73)*(0,1798)*(31,25) = 15,34

c7 = (0,51)*(0,51*(0,0124)*(0,0124) = 0,00004

c8 = (-0,50)*(-0,50)*(0,1798)*(0,1798) = 0,0081

c9 = (-2,54)*(-2,54)*(31,25)*(31,25) = 6300,39

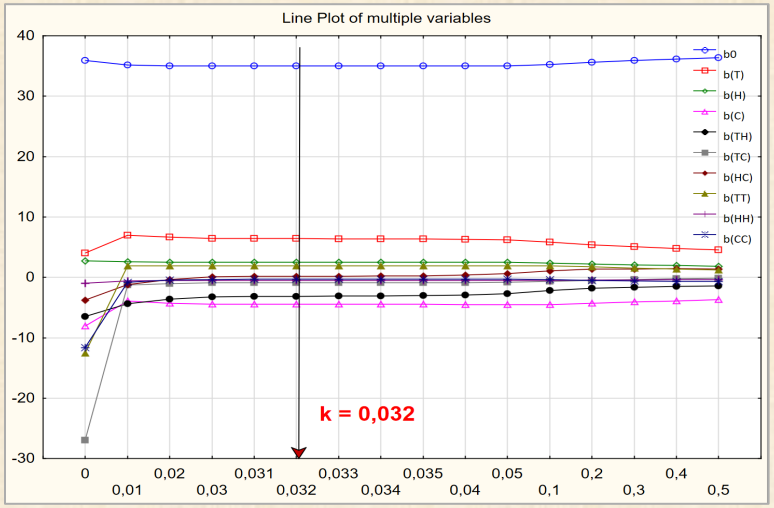
P = 81,97 + 0,12*Temp + 0,04*H2 + 735,16*Cont
-0,01*Temp*H2 + 0,43*Temp*Cont + 15,34*H2*Cont
+ 0,00004*Temp^2 + 0,0081*H2^2 + 6300,39*Cont^2

	P Observed	P Predictd
1	49,0	49,3
2	50,2	49,3
3	50,5	49,2
4	48,5	48,7
5	47,5	47,9
6	44,5	45,9
7	28,0	29,6
8	31,5	31,4
9	34,5	34,4
10	35,0	36,2
11	38,0	36,3
12	38,5	37,3
13	15,0	15,1
14	17,0	15,1
15	20,5	21,7
16	29,5	30,4
C	31,84	
D	42,14	
G	31,07	
H	44,59	

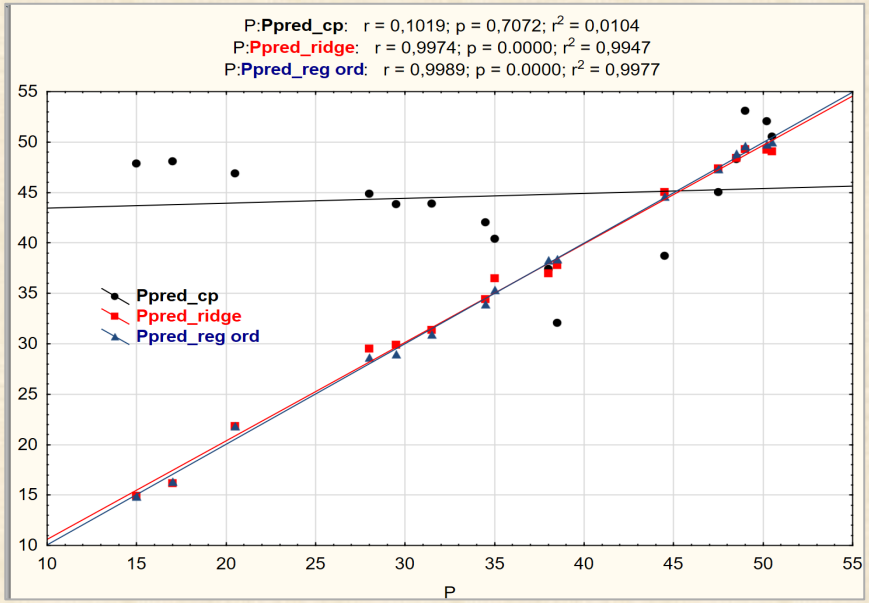
Exemple 3 : Acétylène - étude de cas

Régression Ridge - OLS - ACP

1 k	2 b0	3 b(T)	4 b(H)	5 b(C)	6 b(TH)	7 b(TC)	8 b(HC)	9 b(TT)	10 b(HH)	11 b(CC)
0,000	35,90	4,00	2,78	-8,04	-6,46	-26,98	-3,77	-12,52	-0,97	-11,59
0,010	35,16	7,01	2,58	-3,89	-4,39	-1,25	-1,18	1,89	-0,58	-0,80
0,020	35,03	6,63	2,54	-4,29	-3,63	-1,01	-0,35	1,87	-0,54	-0,43
0,030	35,01	6,44	2,52	-4,43	-3,20	-0,91	0,11	1,87	-0,52	-0,32
0,031	35,02	6,43	2,52	-4,44	-3,17	-0,90	0,14	1,87	-0,52	-0,31
0,032	35,02	6,41	2,52	-4,44	-3,13	-0,90	0,17	1,87	-0,52	-0,31
0,033	35,02	6,40	2,52	-4,45	-3,10	-0,89	0,21	1,87	-0,52	-0,30
0,034	35,02	6,39	2,51	-4,46	-3,07	-0,88	0,24	1,87	-0,52	-0,30
0,035	35,02	6,37	2,51	-4,46	-3,04	-0,88	0,27	1,87	-0,51	-0,30
0,040	35,03	6,31	2,50	-4,49	-2,92	-0,85	0,40	1,88	-0,51	-0,28
0,050	35,06	6,21	2,48	-4,51	-2,71	-0,81	0,60	1,88	-0,51	-0,28
0,100	35,24	5,86	2,39	-4,48	-2,18	-0,66	1,08	1,87	-0,49	-0,35
0,200	35,61	5,41	2,23	-4,28	-1,79	-0,48	1,35	1,74	-0,45	-0,48
0,300	35,92	5,07	2,09	-4,08	-1,61	-0,34	1,41	1,57	-0,42	-0,57
0,400	36,18	4,78	1,97	-3,89	-1,50	-0,23	1,42	1,41	-0,40	-0,63
0,500	36,40	4,53	1,86	-3,72	-1,41	-0,13	1,41	1,26	-0,37	-0,67



26 essai	27 ID2	28 Temp	29 H2	30 Cont	31 new	32 T	33 H	34 C	35 P	36 Ppred_cp	37 Ppred_ridge	38 Ppred_reg ord
1		1300	7,5	0,0120	recodage	1,0853	-0,8731	-0,8949	49,0	53,11	49,30	49,61
2		1300	9,0	0,0120	de	1,0853	-0,6082	-0,8949	50,2	52,04	49,22	49,77
3	J	1300	11,0	0,0115	Temp	1,0853	-0,2550	-0,9107	50,5	50,57	49,09	50,01
4		1300	13,5	0,0130	H2	1,0853	0,1865	-0,8633	48,5	48,33	48,42	48,85
5	I	1300	17,0	0,0135	Cont	1,0853	0,8047	-0,8475	47,5	45,06	47,39	47,33
6		1300	23,0	0,0120	en	1,0853	1,8644	-0,8949	44,5	38,72	45,05	44,63
7		1200	5,3	0,0400	valeurs	-0,1550	-1,2617	-0,0099	28,0	44,85	29,50	28,65
8		1200	7,5	0,0380	centrées	-0,1550	-0,8731	-0,0731	31,5	43,89	31,37	30,91
9		1200	11,0	0,0320	réduites	-0,1550	-0,2550	-0,2627	34,5	42,04	34,39	33,91
10		1200	13,5	0,0260	T	-0,1550	0,1865	-0,4524	35,0	40,38	36,48	35,34
11	F	1200	17,0	0,0340	H	-0,1550	0,8047	-0,1995	38,0	37,43	36,97	38,30
12	E	1200	23,0	0,0410	C	-0,1550	1,8644	0,0217	38,5	32,08	37,78	38,40
13		1100	5,3	0,0840		-1,3954	-1,2617	1,3808	15,0	47,89	14,89	14,84
14	A	1100	7,5	0,0980		-1,3954	-0,8731	1,8233	17,0	48,09	16,17	16,33
15	B	1100	11,0	0,0920		-1,3954	-0,2550	1,6337	20,5	46,89	21,81	21,85
16		1100	17,0	0,0860		-1,3954	0,8047	1,4440	29,5	43,85	29,87	28,98
extrapol	C	1100	23,0	0,0115		-1,3954	1,8643	-0,9008	?	31,84	43,10	-5,23
extrapol	D	1200	7,5	0,0980		-0,1550	-0,8732	1,8331	?	42,14	21,84	-9,07
extrapol	G	1200	23,0	0,0115		-0,1550	1,8643	-0,9008	?	31,07	41,21	39,04
extrapol	H	1300	7,5	0,0980		1,0853	-0,8732	1,8331	?	44,59	33,31	-72,90



Conclusion générale

- Méthode Ridge : meilleure
- Régression ordinaire bien évaluée mais seulement avec les 16 observations
  extrapolation : très mauvaise

Extensions du modèle de régression multiple RM

- **Régression Multiple (RM):** $Y = \beta_o + \beta_1 X_1 + \ldots + \beta_p X_p$ / **X et Y continues**
- **Extensions**
  - **AD : Analyse Discriminante : Y est une variable catégorique**
    prédiction d'appartenance à un groupe avec X
  - **AC : Analyse Canonique : 2 groupes de variables continues X et Y**
    prédiction des facteurs de Y par les facteurs de X
  - **ACP : Analyse en Composantes Principales :**
    prédiction de Y par les facteurs de X

propriétés en commun de RM – AD – AC – ACP

1. **les facteurs sont extraits des matrices Y'Y et X'X**
2. **nombre de fonctions de prédictions ≤ min (dim X, dim Y) = (p, m)**

Régression PLS : régression Partial Least Square

- **étend la RM sans imposer les restrictions 1 et 2 ci-haut**
- **fonctions de prédictions: facteurs extraits de la matrice Y'XX'Y**
- **nombre fonctions de prédictions ≥ max (dimX, dimY)**
- **peut s'employer si : nombre observations < nombre variables**
- **proche de la régression sur composantes principales**
- **composantes PLS optimisées pour être prédictives de Y**

Régression PLS (Projection on Latent Structure) : Partial Least Square

RAPPEL : régression sur composantes principales (p. 20)

$Y = XB + E$ $\quad Y : n \times m \quad X : n \times p \quad B : p \times m \quad E : n \times m$

n : nombre de cas (observations)

m : nombre de variables Y $\quad$ p : nombre de variables X

B : coefficients de régression $\quad$ E : terme d'erreur (bruit)

$T = XW$ scores factoriels de X $\quad$ W: matrice de poids

$Y = TQ + E$ $\quad$ Q: matrice des coefficients de régression (loadings)

$B = WQ$ $\quad Y = XB + E = XWQ + E$

Régression PLS

déterminer W, T, Q telles $T = XW$ et $B = WQ$

W : matrice p x c poids de X vérifie $T = XW$

vecteurs (colonnes) de poids (loadings) pour X

W maximise la covariance entre les réponses Y et T

T : matrice n x c = scores factoriels de X

régression de Y sur X et sur T :

$Y = XB + E = XWQ + E = TQ + E$

Q : poids (loadings) de Y sur T

régression de X sur P : $X = TP + F$

P : matrice p x c de poids (loadings) de X sur T

Régression PLS

déterminer W, T, Q telles T = XW et B = WQ

W : matrice p x c poids de X

vecteurs (colonnes) de poids (loadings) pour X

W maximise la covariance entre les réponses Y et T

T : matrice n x c = scores factoriels de X

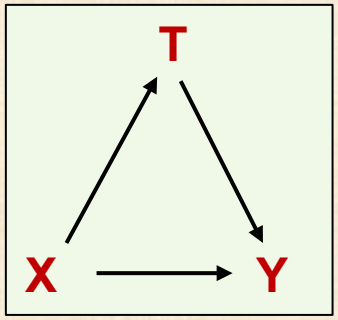
régression de Y sur X et sur T :

$$Y = XB + E = XWQ + E = TQ + E$$

Q : poids (loadings) pour Y sur T

régression de X sur P : $X = TP + F$

P : matrice p x c de poids (loadings) pour X sur T



<u>modèle de régression double sur X et Y</u>

déterminer un petit nombre de composants PLS
(facteurs ou variables latentes non observées directement)
qui expliquent une

- grande partie de la variabilité des prédicteurs X
- grande partie de la variabilité des variables de réponse Y

<u>bon modèle</u> nombre restreint de composants latents
et forte corrélation entre les X Y

2 algorithmes (NIPALS, SIMPLS) : choisir des composantes (facteurs) successifs orthogonaux qui maximisent la covariance entre les X et Y

ALGORITHMES (extrait de la documentation Statistica)

calculs avec les X et Y en centrées réduites : moyennes = 0 écart types = 1

NIPALS: Nonlinear Iterative Partial Least Squares

Pour h = 1,…, c (nombre composantes) $A_0=X'Y$, $M_0 = X'X$, $C_0 = I$

1. calculer q_h , le vecteur propre dominant (plus grande valeur propre) de $A_h'A_h$
2. $w_h = G_hA_hq_h$, $w_h = w_h / ||w_h||$ mettre w_h dans W comme une colonne
3. $p_h = M_hw_h$, $c_h = w_h'M_hw_h$, $p_h = p_h /c_h$ mettre p_h dans P comme une colonne
4. $q_h = A_h'w_h / c_h$ mettre q_h dans Q comme une colonne
5. $A_{h+1} = A_h - c_hp_hq_h'$ et $B_{h+1} = M_h - c_hp_hp_h'$
6. $C_{h+1} =C_h - w_hp_h'$

T = XW = scores factoriels B = WQ = coefficients régression PLS

SIMPLS : Simple Iterative Multiple Partial Least Square

Pour h =1,…, c $A_0 = X'Y$, $M_0 = X'X$, $C_0 = I$, et c donné

1. calculer q_h, le vecteur propre dominant de $A_h'A_h$
2. $w_h = A_hq_h$, $c_h = w_h'M_hw_h$, $w_h = w_h/sqrt(c_h)$, mettre w_h dans W comme colonne
3. $p_h = M_hw_h$, mettre p_h dans P comme colonne
4. $q_h = A_h'w_h$, mettre q_h dans Q comme colonne
5. $v_h = C_hp_h$, $v_h = v_h / ||v_h||$
6. $C_{h+1} = C_h - v_hv_h'$ et $M_{h+1} = M_h - p_hp_h'$
7. $A_{h+1} = C_hA_h$

T = XW = scores factoriels B = WQ = coefficients régression PLS

si un seul Y : les algorithmes sont équivalents

Partial Least Squares (Ref Statistica Electronic Manual)

Partial Least Squares (PLS) (also known as *Projection to Latent Structure*) is a popular method for modeling industrial applications. It was developed by Wold in the 1960s as an economic technique, but soon its usefulness was recognized by many areas of science and applications including *Multivariate Statistical Process Control (MSPC)* in general and chemical engineering in particular.

It many ways, *PLS* can be regarded as a substitute for the method of multiple regression, especially when the number of predictor variables is large. In such cases, there is seldom enough data to construct a reliable model that can be used for predicting the dependent data Y from the predictor variables X. Instead, we get a model that can perfectly fit the training data while performing poorly on unseen samples. This problem is known as over-fitting (Bishop 1995). *PLS* alleviates this problem by adopting the "often correct" assumption that, although there might be a large number of predictor variables, the data may actually be much simpler and can be modeled with the aid of just a handful of components (also known as latent). This in fact is the same technique used by *PCA* for representing the X variables with the aid of a lesser number of principal components.

The idea is to construct a set of components that accounts for as much as possible variation in the data while also modeling the *Y* variables well. The technique works by extracting a set of components that transforms the original data X to a set of *t*-scores T (as in *PCA*). Similarly, Y is used to define another set of components known the *u*-scores U. The *t*-scores are then used to predict the *u*-scores, which in turn are used to predict the response variables Y. This multistage process is hidden, and all we see as the outcome is that for a set of predictor variables *X*, *PLS* predicts a set of relating responses *Y*. Thus, *PLS* works just as any other regression model.

where Q is the *q*-loadings and F is the matrix of residuals that models the noise. The above equations represent the so-called outer relations of the *PLS* model. In addition, there is also an inner relation that results from the interdependency of *Y* variables on *X*. As we shall see later, both the outer and inner relations can be built using the *NIPALS* algorithm (see *NIPALS Algorithm for PLS*, below).

Just as all regression models, it is the aim of *PLS* to minimize the residuals F as much as possible while also having a representation of the relationship between *X* and *Y* that generalizes well for unseen (validation) data. Because of the inner relationship, the principal components used for representing *X* and *Y* cannot be calculated separately. In other words, when building *PLS* models, the outer relationships should know about each other so they cooperate in building the model as a whole. This means that the outer relations cannot be separately handled but rather treated as parts of the same problem. This condition is fulfilled by rewriting the second outer equation as:

$$Y = TBQ^T + F$$

were B is the matrix of coefficients which relates the matrix of *t*-scores T to the matrix of *u*-scores U. In essence, TB defines an inner regression problem that relates the *t*-scores to the *u*-scores. In other words, T is used to regress U. In that sense, the elements of B are merely the regression coefficients.

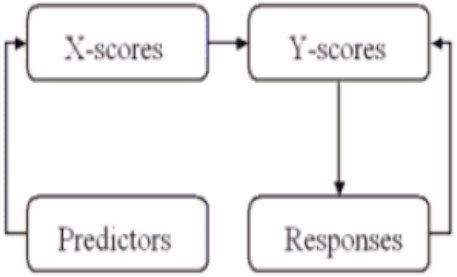


Figure 9. A schematic representation of *PLS* analysis.

Just as before, both X and Y are matrices of data:

$$\mathbf{X} = \begin{pmatrix} x_{11} \Lambda & x_{1M} \\ \mathrm{M} & \\ x_{N1} \Lambda & x_{NM} \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_{11} \Lambda & y_{1D} \\ \mathrm{M} & \\ x_{N1} \Lambda & x_{ND} \end{pmatrix}$$

And it is assumed there is a relationship of the form:

$$(y_{11} \Lambda \; y_{1D}) = f(x_{11} \Lambda \; x_{1M}) + \text{noise}$$

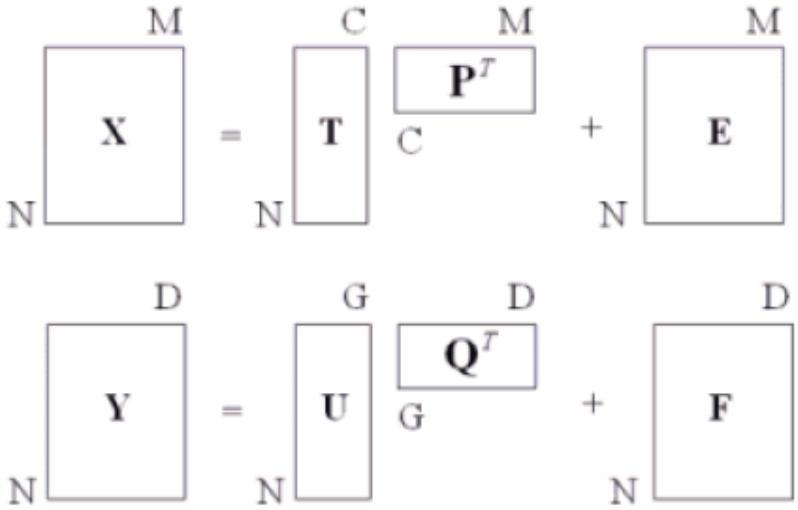


Figure 10. An algebraic representation of a PLS model.

Just as the *X* variables, the R^2Y can also be defined for the *Y* variables from the fraction of the explained variance:

$$R^2 Y = 1 - \frac{\sum_{j=1}^{D} (y_{ij} - \hat{y}_{ij})^2}{\sum_{j=1}^{D} y_{ij}^2}$$

Similarly, you can detect outliers using distance-to-model defined for *Y*,

$$D\text{-}To\text{-}Model = \sqrt{\frac{\sum_{j=1}^{D} (y_{ij} - \hat{y}_{ij})^2}{D}}$$

where *D* is the number of *Y* variables.

Statistical Details for the Partial Least Squares Platform (PLS)

(Ref JMP Documentation)

This section contains statistical details about some of the methods used in the Partial Least Squares platform. See Hoskuldsson (1988), Garthwaite (1994), or Cox and Gaudard (2013).

- "Statistical Details for Partial Least Squares"
- "Statistical Details for the van der Voet T2 Test"
- "Statistical Details for the T2 Plot"
- "Statistical Details for Confidence Ellipses for X Score Scatterplot Matrix"
- "Statistical Details for Prediction and Confidence Limits"
- "Statistical Details for Standardized Scores and Loadings"
- "Statistical Details for PLS Discriminant Analysis"

Statistical Details for Partial Least Squares

Partial least squares fits linear models based on linear combinations, called factors, of the explanatory variables (Xs). These factors are obtained in a way that attempts to maximize the covariance between the Xs and the response or responses (Ys). In this way, PLS exploits the correlations between the Xs and the Ys to reveal underlying latent structures. The factors address the combined goals of explaining response variation and predictor variation. Partial least squares is particularly useful when you have more X variables than observations or when the X variables are highly correlated.

NIPALS

The NIPALS method works by extracting one factor at a time. Let $X = X_0$ be the centered and scaled matrix of predictors and $Y = Y_0$ the centered and scaled matrix of response values. The PLS method starts with a linear combination $t = X_0 w$ of the predictors, where t is called a *score vector* and w is its associated *weight vector*. The PLS method predicts both X_0 and Y_0 by regression on t:

$$\hat{X}_0 = tp', \text{ where } p' = (t't)^{-1}t'X_0$$

$$\hat{Y}_0 = tc', \text{ where } c' = (t't)^{-1}t'Y_0$$

The vectors p and c are called the *X*- and *Y*-*loadings*, respectively.

The specific linear combination $t = X_0 w$ is the one that has maximum covariance $t'u$ with some response linear combination $u = Y_0 q$. Another characterization is that the *X*- and *Y*-weights, w and q, are proportional to the first left and right singular vectors of the covariance matrix $X_0'Y_0$. Or, equivalently, the first eigenvectors of $X_0'Y_0Y_0'X_0$ and $Y_0'X_0X_0'Y_0$ respectively.

This accounts for how the first PLS factor is extracted. The second factor is extracted in the same way by replacing X_0 and Y_0 with the *X*- and *Y*-residuals from the first factor:

$$X_1 = X_0 - \hat{X}_0$$

$$Y_1 = Y_0 - \hat{Y}_0$$

These residuals are also called the *deflated* *X* and *Y* blocks. The process of extracting a score vector and deflating the data matrices is repeated for as many extracted factors as desired.

SIMPLS

The SIMPLS algorithm was developed to optimize a statistical criterion: it finds score vectors that maximize the covariance between linear combinations of Xs and Ys, subject to the requirement that the X-scores are orthogonal. Unlike NIPALS, where the matrices X_0 and Y_0 are deflated, SIMPLS deflates the cross-product matrix, $X_0'Y_0$.

In the case of a single Y variable, these two algorithms are equivalent. However, for multivariate Y, the models differ. SIMPLS was suggested by De Jong (1993).

Statistical Details for the van der Voet T^2 Test

In the Partial Least Squares platform, the van der Voet T^2 test helps determine whether a model with a specified number of extracted factors differs significantly from a proposed optimum model. The test is a randomization test based on the null hypothesis that the squared residuals for both models have the same distribution. Intuitively, one can think of the null hypothesis as stating that both models have the same predictive ability.

To obtain the van der Voet T^2 statistic given in the Cross Validation report, the calculation below is performed on each validation set. In the case of a single validation set, the result is the reported value. In the case of Leave-One-Out and KFold validation, the results for each validation set are averaged.

Denote by $R_{i,jk}$ the jth predicted residual for response k for the model with i extracted factors. Denote by $R_{opt,jk}$ is the corresponding quantity for the model based on the proposed optimum number of factors, *opt*. The test statistic is based on the following differences:

$$D_{i,jk} = R^2_{i,jk} - R^2_{opt,jk}$$

Suppose that there are K responses. Consider the following notation:

$$\mathbf{d}_{i,j} = (D_{i,j1}, D_{i,j2}, \ldots, D_{i,jK})'$$

$$\mathbf{d}_{i,.} = \sum_j \mathbf{d}_{i,j}$$

$$\mathbf{S_i} = \sum_j \mathbf{d}_{i,j}\mathbf{d}_{i,j}'$$

The van der Voet statistic for i extracted factors is defined as follows:

$$C_i = \mathbf{d}_{i,.}'\mathbf{S_i}^{-1}\mathbf{d}_{i,.}$$

The significance level is obtained by comparing C_i with the distribution of values that results from randomly exchanging $R^2_{i,jk}$ and $R^2_{opt,jk}$. A Monte Carlo sample of such values is simulated and the significance level is approximated as the proportion of simulated critical values that are greater than or equal to C_i.

Statistical Details for the T^2 Plot

In the Partial Least Squares platform, the T^2 value for the i^{th} observation is computed as follows:

$$T_i^2 = (n-1)\sum_{j=1}^{p}\left(t_{ij}^2 / \sum_{k=1}^{n} t_{kj}^2\right)$$

where t_{ij} = X score for the i^{th} row and j^{th} extracted factor, p = number of extracted factors, and n = number of observations used to train the model. If validation is not used, n = total number of observations.

The control limit for the T^2 Plot is computed as follows:

$((n\text{-}1)^2/n)$*BetaQuantile(0.95, p/2, $(n\text{-}p\text{-}1)/2$)

where p = number of extracted factors, and n = number of observations used to train the model. If validation is not used, n = total number of observations. See Tracy et al. (1992).

Statistical Details for Prediction and Confidence Limits

This section describes the calculation of standard errors of prediction and confidence limits in the Partial Least Squares platform. Let X denote the matrix of predictors and Y the matrix of response values, which might be centered and scaled based on your selections in the launch window. Assume that the components of Y are independent and normally distributed with a common variance σ^2.

Hoskuldsson (1988) observes that the PLS model for Y in terms of scores is formally similar to a multiple linear regression model. He uses this similarity to derive an approximate formula for the variance of a predicted value. See also Umetrics (1995). However, Denham (1997) points out that any value predicted by PLS is a non-linear function of the Ys. He suggests bootstrap and cross validation techniques for obtaining prediction intervals. The PLS platform uses the normality-based approach described in Umetrics (1995).

Denote the matrix whose columns are the scores by T and consider a new observation on X, x_0. The predictive model for Y is obtained by regressing Y on T. Denote the score vector associated with x_0 by t_0.

Let a denote the number of factors. Define s^2 to be the sum of squares of residuals divided by $df = n - a - 1$ if the data are centered and $df = n - a$ if the data are not centered. The value of s^2 is an estimate of σ^2.

Standard Error of Prediction Formula

The standard error of the predicted mean at x_0 is estimated by the following:

$$SE(\bar{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + t_0(T'T)^{-1}t_0'\right)}$$

Mean Confidence Limit Formula

Let $t_{0.975, df}$ denote the 0.975 quantile of a t distribution with degrees of freedom $df = n - a - 1$ if the data are centered and $df = n - a$ if the data are not centered.

The 95% confidence interval for the mean is computed as follows:

$$\bar{Y}_{x_0} \pm t_{0.975, df} SE(\bar{Y}_{x_0})$$

Indiv Confidence Limit Formula

The standard error of a predicted individual response at x_0 is estimated by the following:

$$SE(\hat{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + 1 + t_0(T'T)^{-1}t_0'\right)}$$

Let $t_{0.975, df}$ denote the 0.975 quantile of a t distribution with degrees of freedom $df = n - a - 1$ if the data are centered and $df = n - a$ if the data are not centered.

The 95% prediction interval for an individual response is computed as follows:

$$\bar{Y}_{x_0} \pm t_{0.975, df} SE(\hat{Y}_{x_0})$$

Statistical Details for Standardized Scores and Loadings

This section describes the calculations of the standardized scores and loadings in the Partial Least Squares platform.

Consider the following notation:

- n_{tr} is the number of observations in the training set
- m is the number of effects in X
- k is the number of responses in Y
- $VarX_i$ is the percent variation in X explained by the i^{th} factor
- $VarY_i$ is the percent variation in Y explained by the i^{th} factor
- $\mathbf{XScore}_i$ is the vector of X scores for the i^{th} factor
- $\mathbf{YScore}_i$ is the vector of Y scores for the i^{th} factor
- $\mathbf{XLoad}_i$ is the vector of X loadings for the i^{th} factor
- $\mathbf{YLoad}_i$ is the vector of Y loadings for the i^{th} factor

Standardized Scores

The vector of i^{th} Standardized X Scores is defined as follows:

$$\frac{\mathbf{XScore}_i}{(n_{tr}-1)\sqrt{mVarX_i/n_{tr}}}$$

The vector of i^{th} Standardized Y Scores is defined as follows:

$$\frac{\mathbf{YScore}_i}{(n_{tr}-1)\sqrt{kVarY_i/n_{tr}}}$$

Standardized Loadings

The vector of i^{th} Standardized X Loadings is defined as follows:

$$\mathbf{XLoad}_i\sqrt{mVarX_i}$$

The vector of i^{th} Standardized Y Loadings is defined as follows:

$$\mathbf{YLoad}_i\sqrt{kVarY_i}$$

Pour des variables catégoriques Y

On peut faire une analyse Partial Least Square Discriminant (PLS-DA) en utilisant Partial Least Square personality dans le Fit Model platform.

La variable Y utilise le codage disjonctif complet.

You can perform a Partial Least Squares Discriminant Analysis (PLS-DA) by using the Partial Least Squares personality in the Fit Model platform. When a categorical variable is entered as Y in the launch window, it is coded using indicator coding. If there are k levels, each level is represented by an indicator variable with the value 1 for rows in that level and 0 otherwise. The resulting k indicator variables are treated as continuous and the PLS analysis proceeds as it would with continuous Ys.

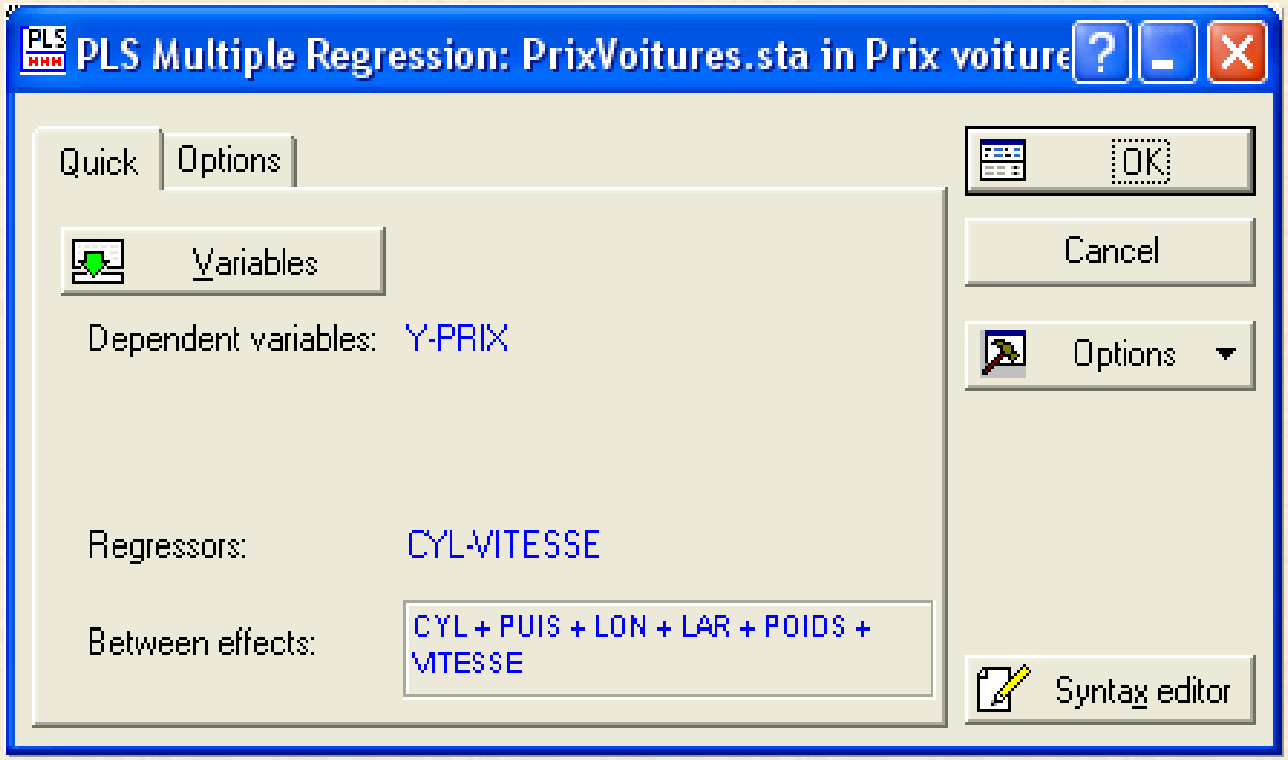
Avantages de la régression PLS

- peut traiter le cas: nombre de variables > nombre d'observations
- première composante PLS toujours plus corrélée avec Y que la première composante principale de X

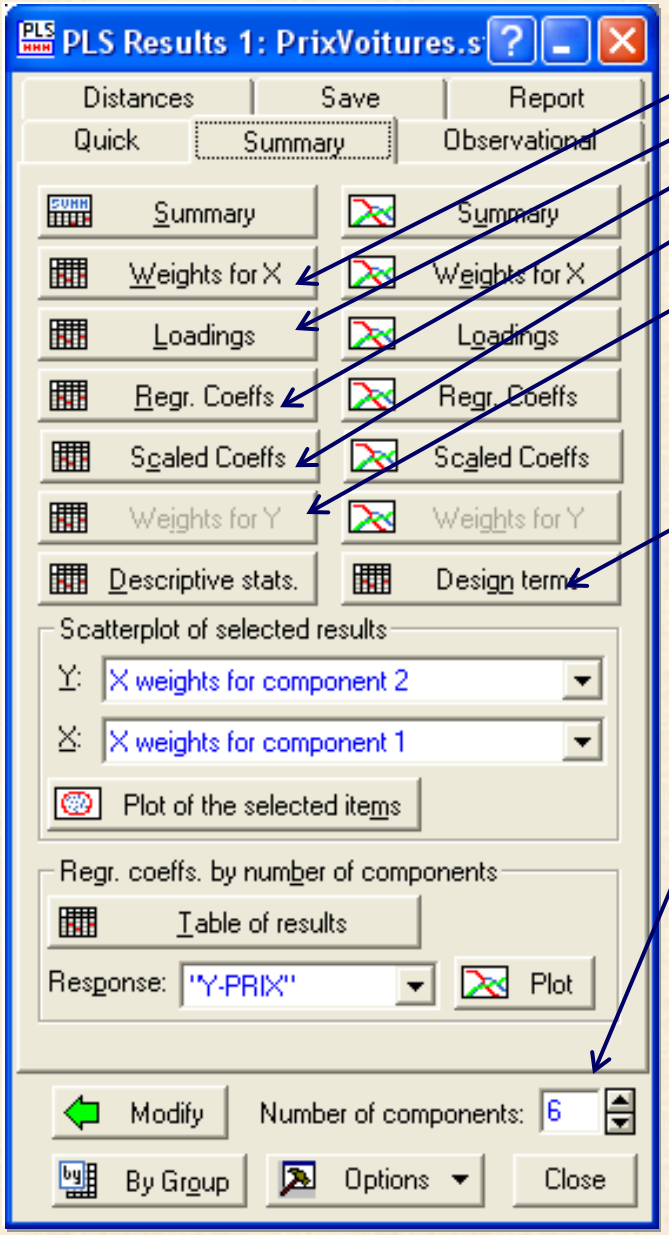
PLS avec STATISTICA Advanced Linear / Nonlinear …. General Partial Least Square Models

Exemple 4:

prix des voitures



Régression PLS : Partial Least Square



W : Weights for X

P : Loadings

B : Regr. Coeffs – variables dans leurs unités

Scaled Coeffs (beta) (var centrées-éduites)

Q : Weights for Y – seulement si plusieurs Y

identification du modèle employé

colonnes de la matrice design utile avec des variables catégoriques pour montrer comment les modalités d'une variable catégorique seront codées avec des variables à valeurs -1 0 1

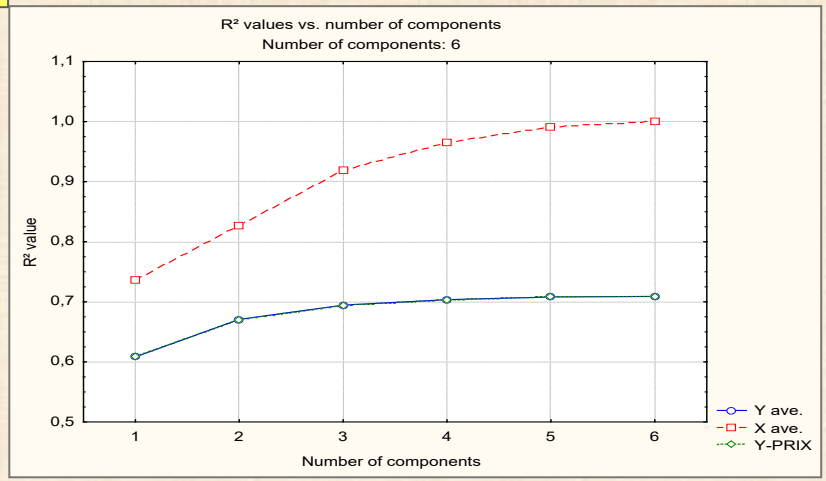
nombre de composants : ici, 6 est le nombre maximal car il y a 6 variables continues;

en général, le nombre de composants utiles sera plus petit

Régression PLS

data = voitures

NOM	CYL	PUIS	LON	LAR	POIDS	VITESSE
ALFASUD-TI-1350	1350	79	393	161	870	165
AUDI-100-L	1588	85	468	177	1110	160
SIMCA-1307-GLS	1294	68	424	168	1050	152
CITROEN-GS-CLUB	1222	59	412	161	930	151
FIAT-132-1600GLS	1585	98	439	164	1105	165
LANCIA-BETA-1300	1297	82	429	169	1080	160
PEUGEOT-504	1796	79	449	169	1160	154



Summary of PLS Responses: Y-PRIX

C nombre composants	Increase R² of Y	Average R² of Y cumul	Increase R² of X	Average R² of X cumul
Comp 1	0,608	0,608 min	0,736	0,736
Comp 2	0,062	0,671	0,090	0,827
Comp 3	0,024	0,695	0,093	0,919
Comp 4	0,009	0,704	0,047	0,966
Comp 5	0,005	0,708	0,025	0,991
Comp 6	0,001	0,709 max	0,009	1,000

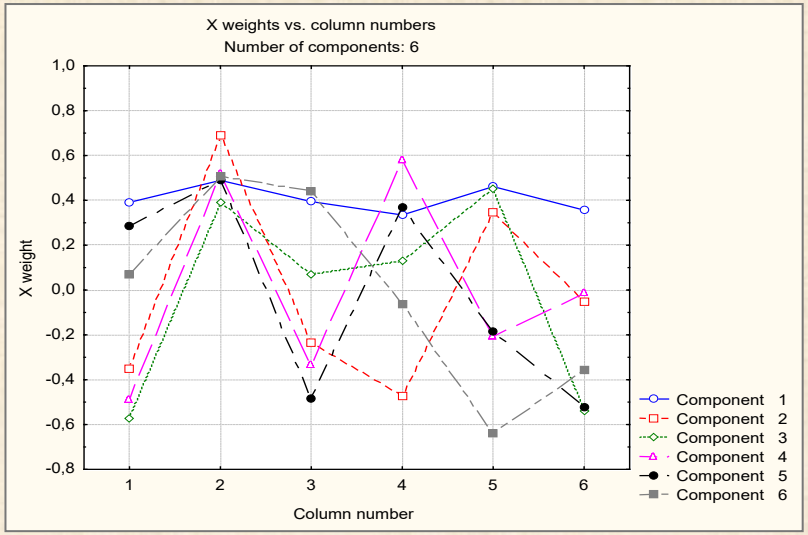
combien de composantes retenir ?

critères: c petit et R^2 de Y (cumul) max

Cet exemple: modèle à 2 composantes semble un bon choix R^2 de Y = 0,67

Predictor weights Responses: Y-PRIX

	CYL	PUIS	LON	LAR	POIDS	VITESSE
Compo 1	0,391	0,489	0,394	0,335	0,461	0,356
Compo 2	-0,350	0,689	-0,235	-0,472	0,349	-0,054
Compo 3	-0,575	0,390	0,068	0,131	0,453	-0,539
Compo 4	-0,489	0,518	-0,335	0,581	-0,207	-0,014
Compo 5	0,285	0,490	-0,481	0,370	-0,186	-0,524
Compo 6	0,069	0,507	0,442	-0,065	-0,641	-0,357

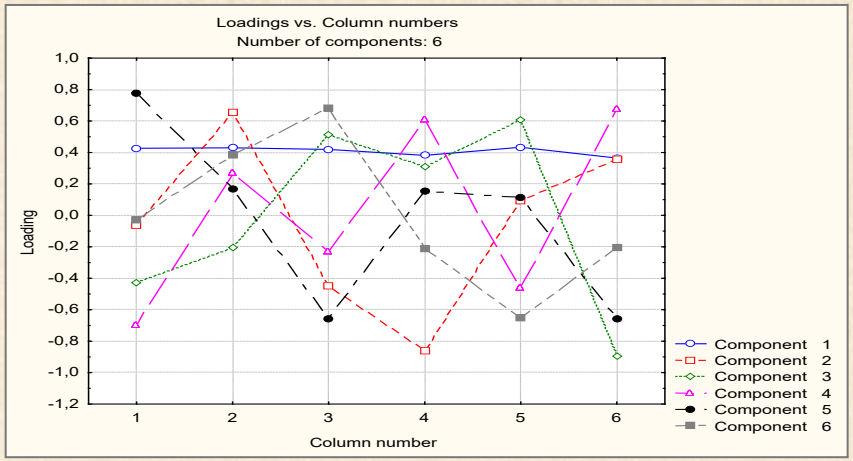


Régression PLS

modèle à 6 composants

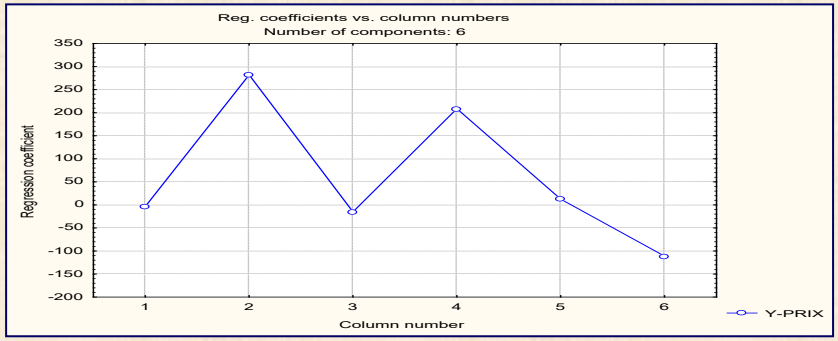
X loadings Responses: Y-PRIX

	CYL	PUIS	LON	LAR	POIDS	VITES
Comp 1	0,427	0,430	0,419	0,381	0,433	0,364
Comp 2	-0,064	0,658	-0,451	-0,859	0,091	0,355
Comp 3	-0,430	-0,208	0,513	0,313	0,609	-0,893
Comp 4	-0,702	0,270	-0,230	0,611	-0,460	0,676
Comp 5	0,780	0,168	-0,655	0,155	0,116	-0,657
Comp 6	-0,031	0,385	0,683	-0,214	-0,653	-0,203



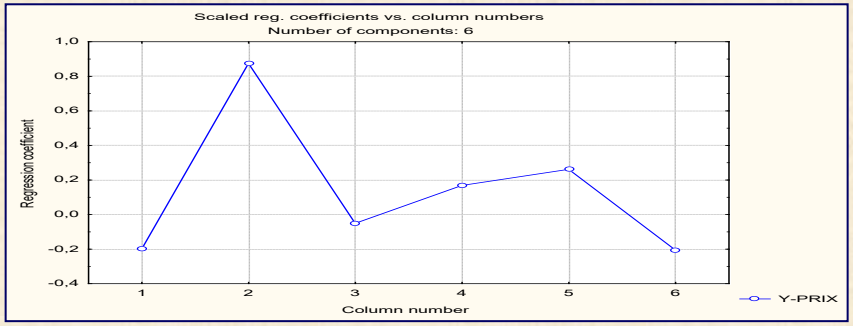
PLS regression coeff. Responses: Y-PRIX

	Interc.	CYL	PUIS	LON	LAR	POIDS	VITESSE
Y-PRIX	-8239,4	-3,5	282,2	-15,0	208,7	12,6	-111,1



PLS scaled regression Responses: Y-PRIX

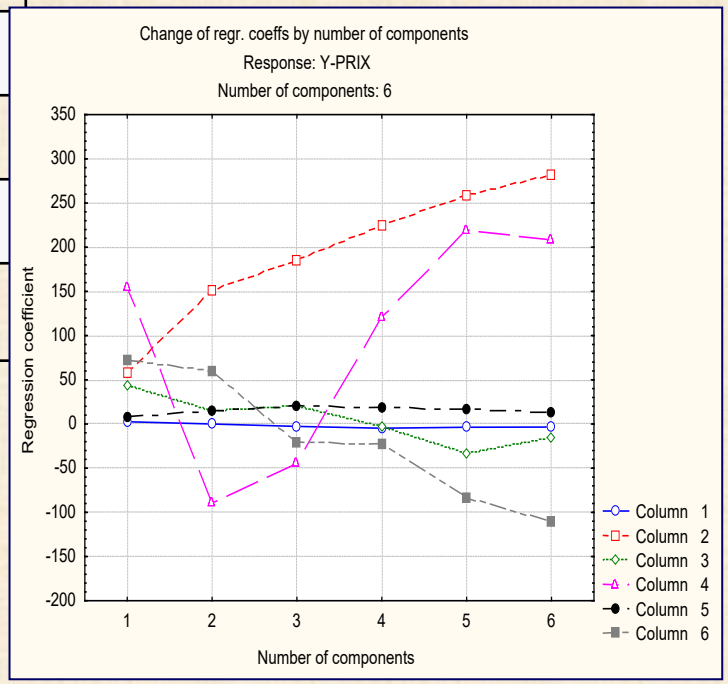
	CYL	PUIS	LON	LAR	POIDS	VITESSE
Y-PRIX	-0,199	0,875	-0,051	0,169	0,262	-0,205



Régression PLS

Change of regr. Coeffs Responses: Y-PRIX						
column	1 Comp Y-PRIX	2 Comp Y-PRIX	3 Comp Y-PRIX	4 Comp Y-PRIX	5 Comp Y-PRIX	6 Comp Y-PRIX
CYL	2,56	-0,00	-2,77	-4,76	-3,68	-3,5
PUIS	58,81	151,56	186,03	224,64	258,78	282,2
LON	43,69	14,54	20,05	-2,94	-33,84	-15,0
LAR	154,34	-89,24	-44,79	121,25	220,13	208,7
POIDS	8,25	15,24	21,20	18,90	16,98	12,6
VTES	71,89	59,72	-20,33	-22,14	-83,43	-111,1

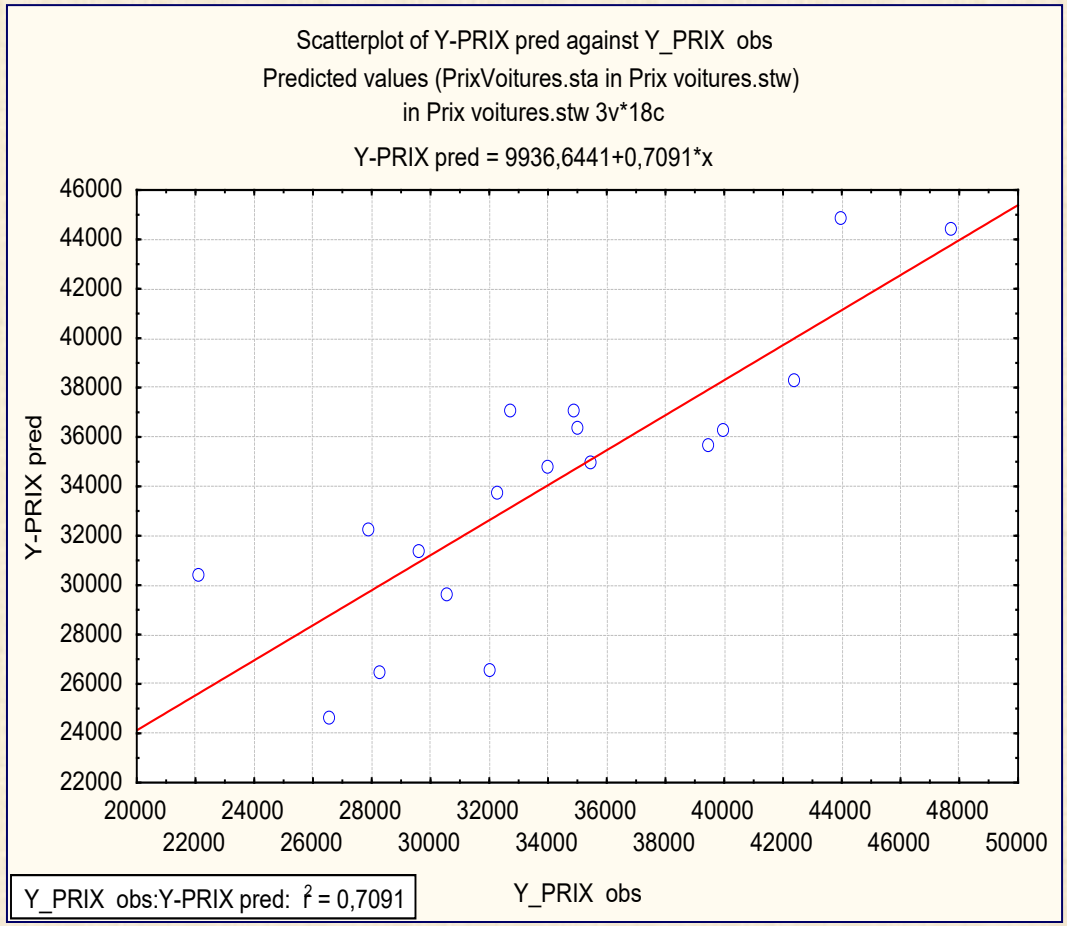
modèle à 6 composants



Régression PLS

modèle à 6 composants - $R^2 = 0,71$

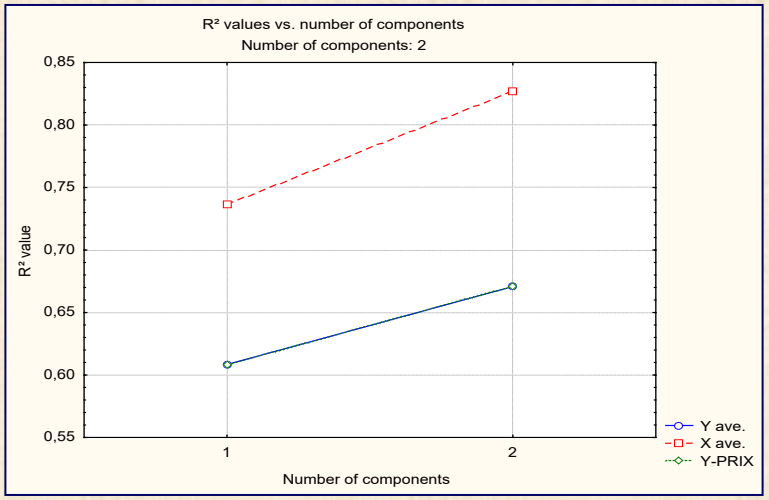
Predicted values	Responses: Y-PRIX		
	Y-PRIX pred	Y_PRIX obs	écart
ALFASUD-	29616	30570	954
AUDI-100	36260	39990	3730
SIMCA-13	31411	29600	-1811
CITROEN-	26446	28250	1804
FIAT-132	37043	34900	-2143
LANCIA-B	34973	35480	507
PEUGEOT-	33749	32300	-1449
RENAULT-	26580	32000	5420
RENAULT-	44446	47700	3254
TOYOTA-C	24650	26540	1890
ALFETTA-	38270	42395	4125
PRINCESS	34830	33990	-840
DATSUN-2	44872	43980	-892
TAUNUS-2	36343	35010	-1333
RANCHO	35638	39450	3812
MAZDA-92	32233	27900	-4333
OPEL-REK	37103	32700	-4403
LADA-130	30390	22100	-8290



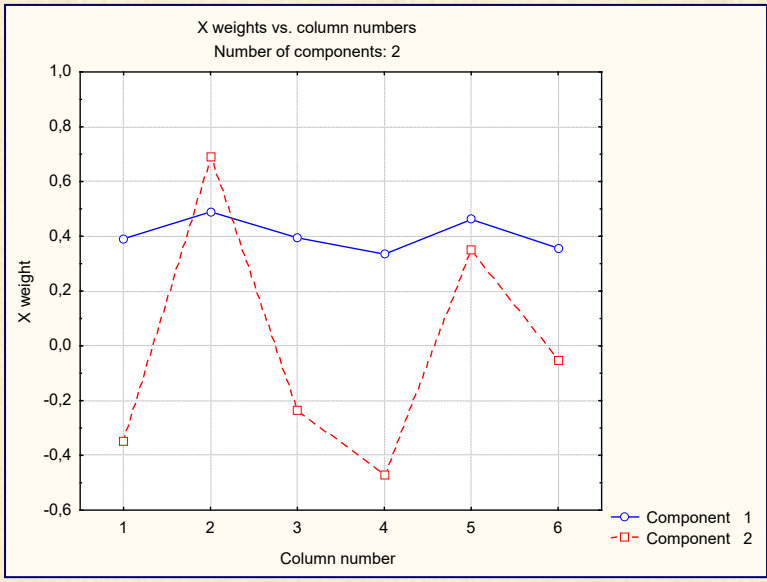
Régression PLS

modèle à 2 composants - $R^2 = 0,67$

Summary of PLS Responses: Y-PRIX				
comp	Increase R^2 of Y	Average R^2 of Y	Increase R^2 of X	Average R^2 of X
Comp 1	0,608	0,608	0,736	0,736
Comp 2	0,062	0,671	0,090	0,827



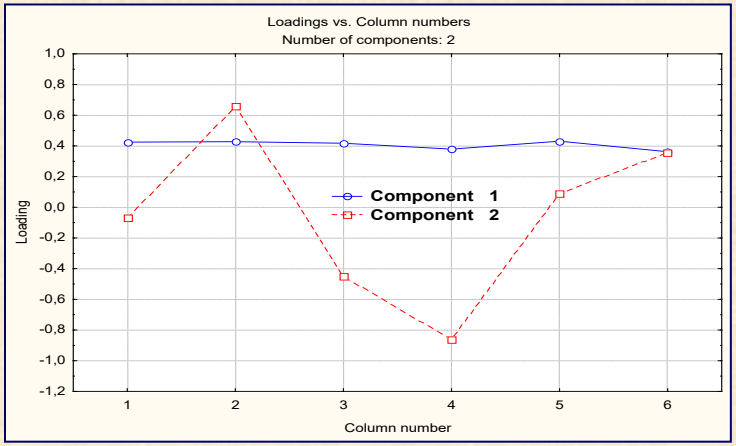
Predictor weights Responses: Y-PRIX						
comp	CYL	PUIS	LON	LAR	POIDS	VITESSE
Compo 1	0,391	0,489	0,394	0,335	0,461	0,356
Compo 2	-0,350	0,689	-0,235	-0,472	0,349	-0,054



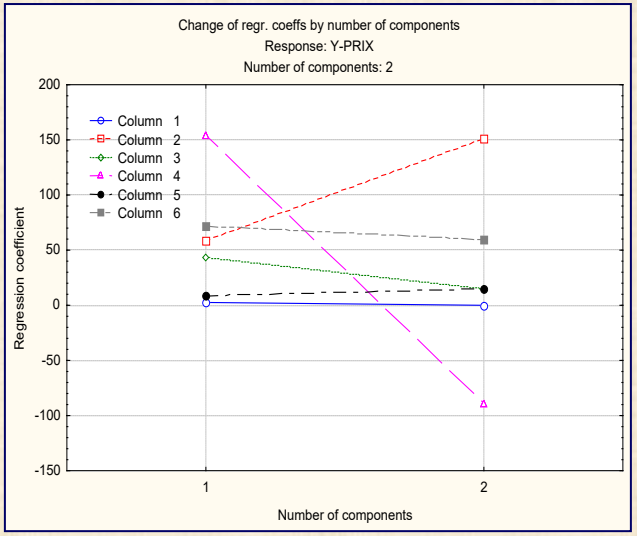
Régression PLS

modèle à 2 composants - $R^2 = 0,67$

X loadings Responses: Y-PRIX						
c	CYL	PUIS	LON	LAR	POIDS	VITESSE
Comp 1	0,427	0,430	0,419	0,381	0,433	0,364
Comp 2	-0,064	0,658	-0,451	-0,859	0,091	0,355



PLS regression coefficients							
	Interc.	CYL	PUIS	LON	LAR	POIDS	VITESSE
Y-PRIX	4013,63	-0,0011	151,56	14,54	- 89,24	15,24	59,72

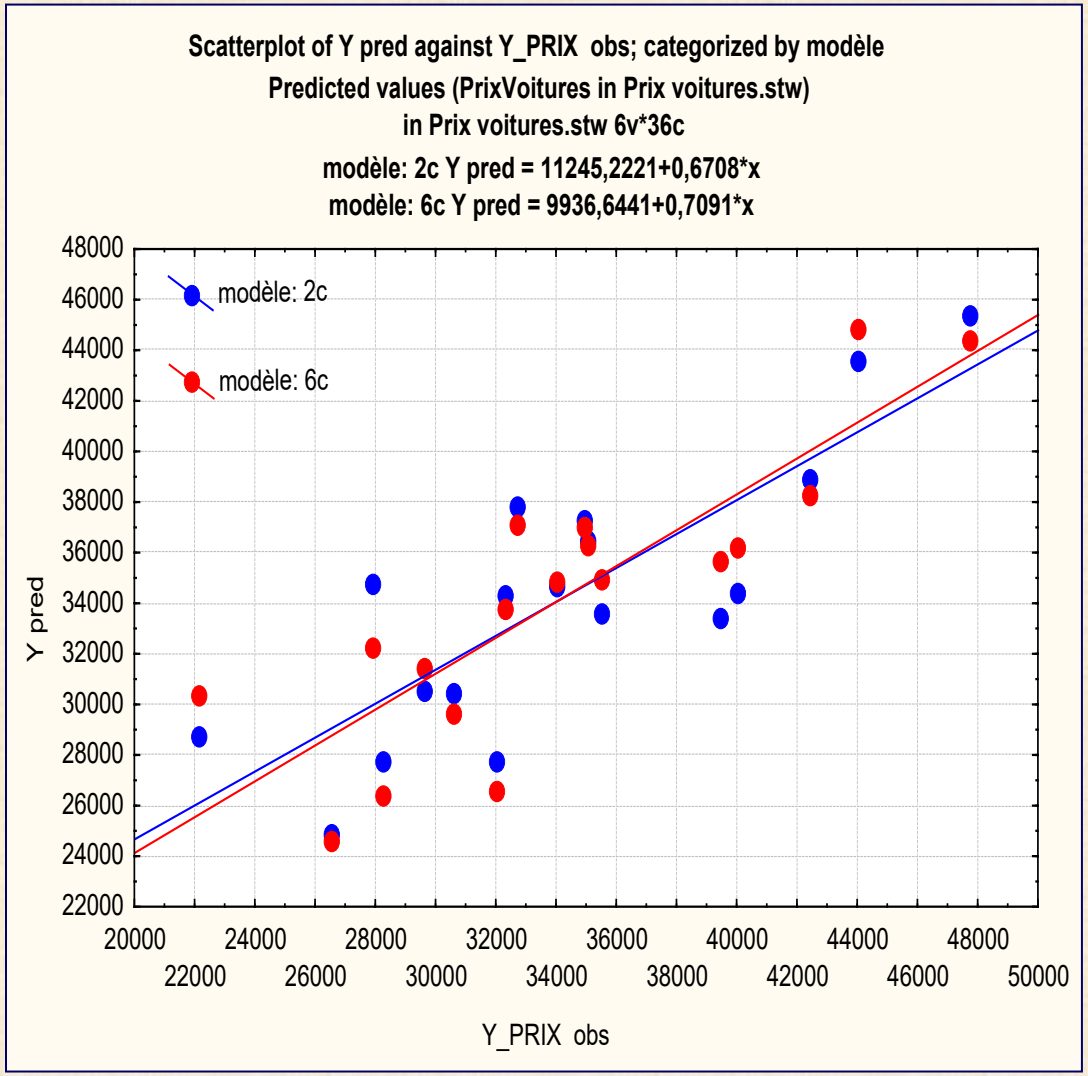


PLS scaled regression coefficients Responses						
	CYL	PUIS	LON	LAR	POIDS	VITESS
Y-PRIX	-0,00006	0,4699	0,0489	-0,0722	0,3176	0,1103

Régression PLS

modèle à 2 composants et modèle à 6 composants

Predicted values Responses: Y-PRIX			
voiture	Y_PRIX obs	Y_PRIX pred 2 comps	Y-PRIX pred 6 comps
ALFASUD-	30570	30444	29616
AUDI-100	39990	34375	36260
SIMCA-13	29600	30570	31411
CITROEN-	28250	27768	26446
FIAT-132	34900	37306	37043
LANCIA-B	35480	33610	34973
PEUGEOT-	32300	34307	33749
RENAULT-	32000	27719	26580
RENAULT-	47700	45410	44446
TOYOTA-C	26540	24920	24650
ALFETTA-	42395	38903	38270
PRINCESS	33990	34674	34830
DATSUN-2	43980	43612	44872
TAUNUS-2	35010	36494	36343
RANCHO	39450	33395	35638
MAZDA-92	27900	34805	32233
OPEL-REK	32700	37803	37103
LADA-130	22100	28739	30390



Exemple 5 : Étude pollution mer Baltique - Lindberg, Person, Wold (1983)

16 échantillons d'eau de mer EM1,..., EM16

variables Y : 3 indicateurs de pollution Y1 Y2 Y3

Y1 = LS = quantité de Lignin Sulfonate (pollution industrie papetière)

Y2 = HA = quantité de Humic Acids (produits naturels forêt)

Y3 = DT = quantité de DétergenT (agent blanchiment)

variables X : intensités émissions spectrales à

27 fréquences différentes : X1-X27

Régression PLS

cas avec nombre d'observations (n=16) inférieur au nombre de variables (p=30)
3 Y et 27 X

nom	LS	HA	DT	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
EM1	3,011	0,000	0,00	2766	2610	3306	3630	3600	3438	3213	3051	2907	2844	2796	2787	2760
EM2	0,000	0,401	0,00	1492	1419	1369	1158	958	887	905	929	920	887	800	710	617
EM3	0,000	0,000	90,63	2450	2379	2400	2055	1689	1355	1109	908	750	673	644	640	630
EM4	1,482	0,158	40,00	2751	2883	3492	3570	3282	2937	2634	2370	2187	2070	2007	1974	1950
EM5	1,116	0,410	30,45	2652	2691	3225	3285	3033	2784							
EM6	3,397	0,303	50,82	3993	4722	6147	6720	6531	5970							
EM7	2,428	0,298	70,59	4032	4350	5430	5763	5490	4974							
EM8	4,024	0,115	89,39	4530	5190	6910	7580	7510	6930							
EM9	2,275	0,504	81,75	4077	4410	5460	5857	5607	5097							
EM10	0,959	0,145	101,10	3450	3432	3969	4020	3678	3237							
EM11	3,190	0,253	120,00	4989	5301	6807	7425	7155	6525							
EM12	4,132	0,569	117,70	5340	5790	7590	8390	8310	7670							
EM13	2,160	0,436	27,59	3162	3477	4365	4650	4470	4107							
EM14	3,094	0,247	61,71	4380	4695	6018	6510	6342	5760							
EM15	1,604	0,286	108,80	4587	4200	5040	5289	4965	4449							
EM16	3,162	0,701	60,00	4017	4725	6090	6570	6354	5895							

suite

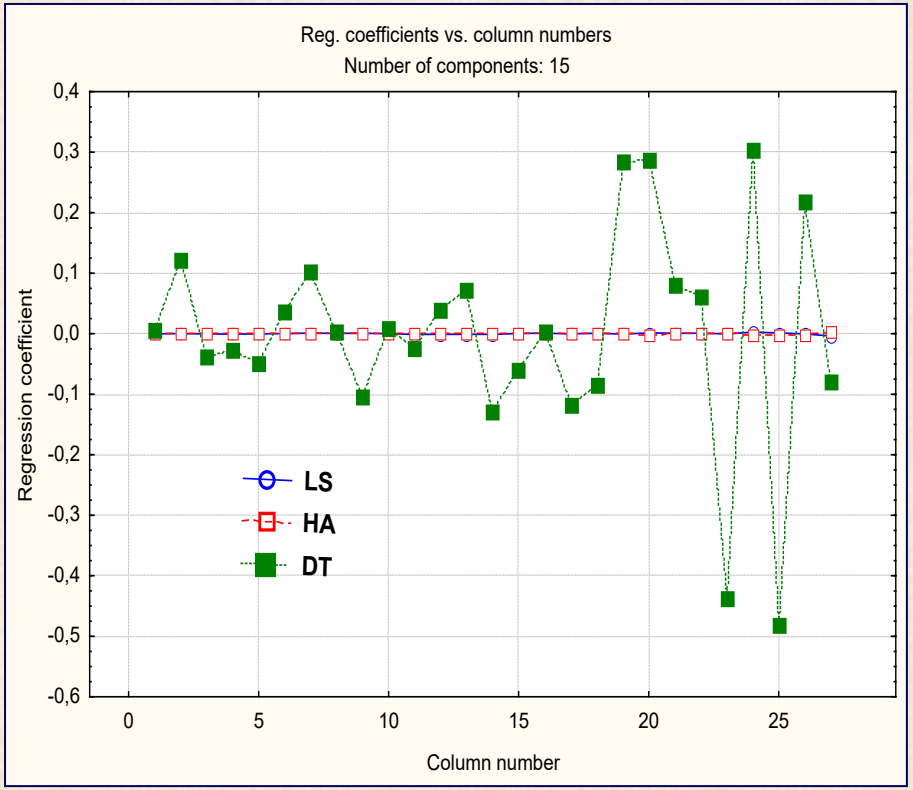
nom	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27
EM1	2754	2670	2520	2310	2100	1917	1755	1602	1467	1353	1260	1167	1101	1017
EM2	535	451	368	296	241	190	157	128	106	89	70	65	56	50
EM3	618	571	512	440	368	305	247	196	156	120	98	80	61	50
EM4	1890	1824	1680	1527	1350	1206	1080	984	888	810	732	669	630	582
EM5	1917	1800	1650	1464	1299	1140	1020	909	810	726	657	594	549	507
EM6	3864	3663	3390	3090	2787	2481	2241	2028	1830	1680	1533	1440	1314	1227
EM7	3147	3000	2772	2490	2220	1980	1779	1599	1440	1320	1200	1119	1032	957
EM8	4210	4000	3770	3420	3060	2760	2490	2230	2060	1860	1700	1590	1490	1380
EM9	3330	3192	2910	2610	2325	2064	1830	1638	1476	1350	1236	1122	1044	963
EM10	1890	1776	1635	1452	1278	1128	981	867	753	663	600	552	507	468
EM11	3972	3777	3531	3168	2835	2517	2244	2004	1809	1620	1470	1359	1266	1167
EM12	4900	4700	4390	3970	3540	3170	2810	2490	2240	2060	1870	1700	1590	1470
EM13	2838	2694	2490	2253	2013	1788	1599	1431	1305	1194	1077	990	927	855
EM14	3567	3438	3171	2880	2571	2280	2046	1857	1680	1548	1413	1314	1200	1119
EM15	2670	2529	2328	2088	1851	1641	1431	1284	1134	1020	918	840	756	714
EM16	4110	3915	3600	3240	2913	2598	2325	2088	1917	1734	1587	1452	1356	1257

Régression PLS

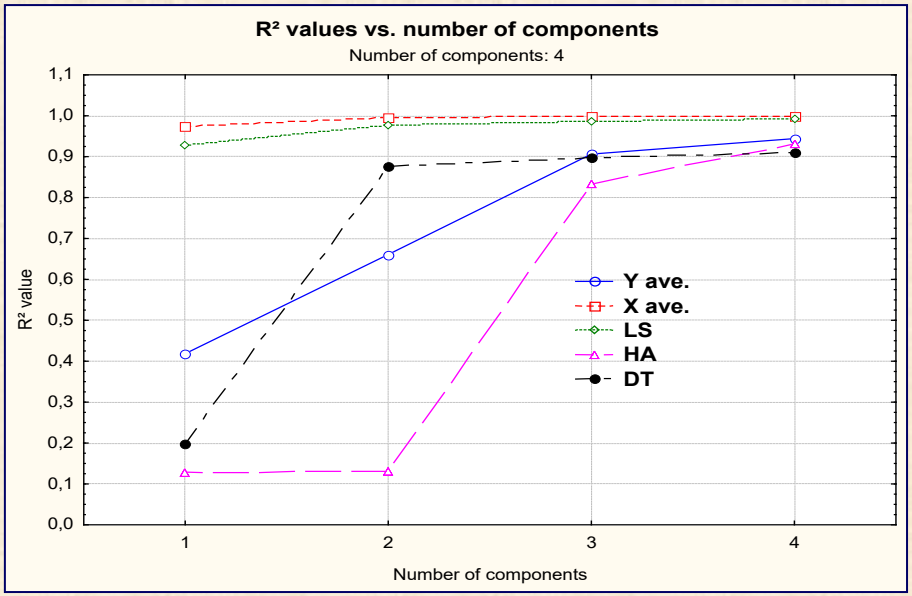
Summary of PLS Responses: LS HA DT							
	Increase - R^2 of Y	Average - R^2 of Y	Increase - R^2 of X	Average - R^2 of X	R^2 for - LS	R^2 for - HA	R^2 for - DT
Comp 1	0,419	0,419	0,975	0,975	0,930	0,130	0,198
Comp 2	0,242	0,662	0,022	0,996	0,976	0,130	0,878
Comp 3	0,245	0,907	0,002	0,998	0,987	0,834	0,900
Comp 4	0,038	0,945	0,001	0,999	0,993	0,932	0,909
Comp 5	0,010	0,955	0,000	1,000	0,997	0,955	0,913
Comp 6	0,023	0,978	0,000	1,000	0,998	0,973	0,962
Comp 7	0,012	0,989	0,000	1,000	0,998	0,977	0,993
Comp 8	0,005	0,994	0,000	1,000	0,998	0,992	0,993
Comp 9	0,001	0,996	0,000	1,000	0,999	0,995	0,993
Comp 10	0,001	0,997	0,000	1,000	0,999	0,996	0,995
Comp 11	0,002	0,998	0,000	1,000	0,999	0,996	1,000
Comp 12	0,001	1,000	0,000	1,000	1,000	0,999	1,000
Comp 13	0,000	1,000	0,000	1,000	1,000	1,000	1,000
Comp 14	0,000	1,000	0,000	1,000	1,000	1,000	1,000
Comp 15	0,000	1,000	0,000	1,000	1,000	1,000	1,000

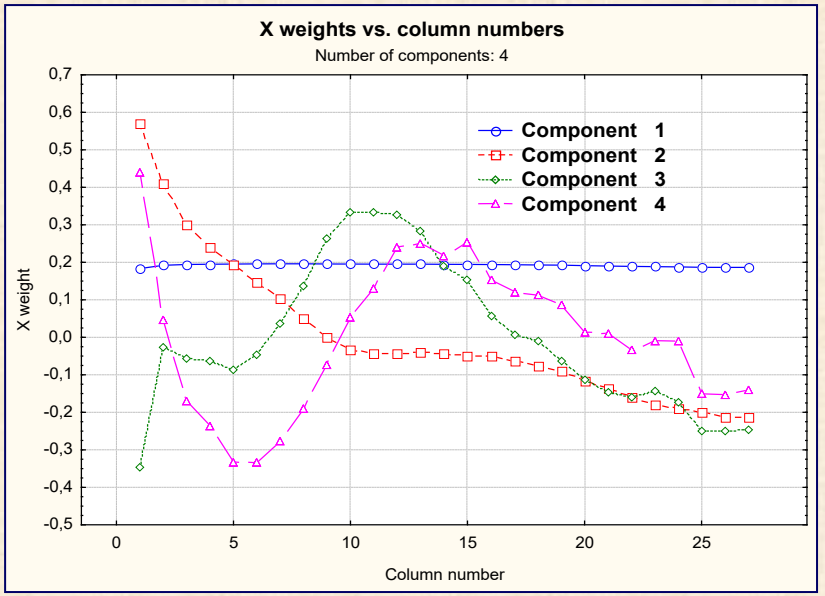
combien de composants retenir?

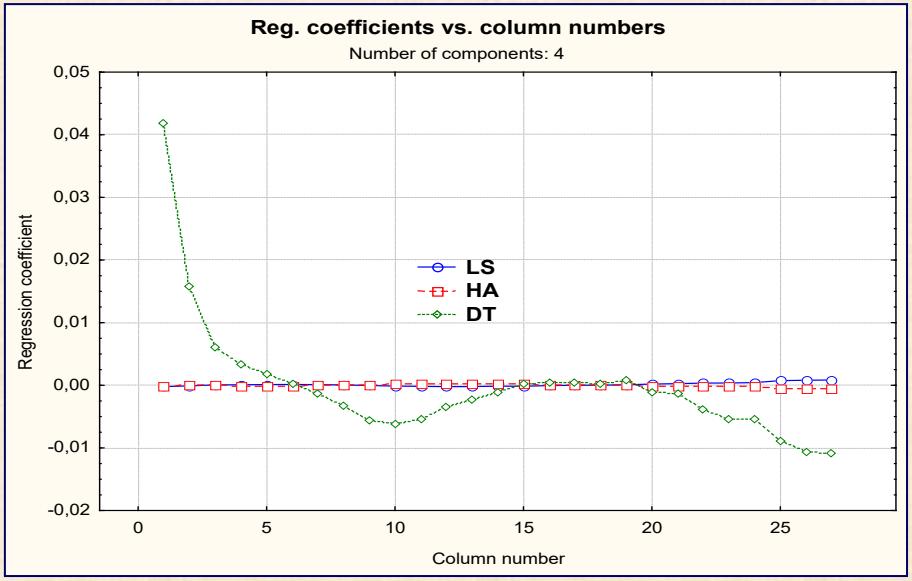
4 composants semblent OK

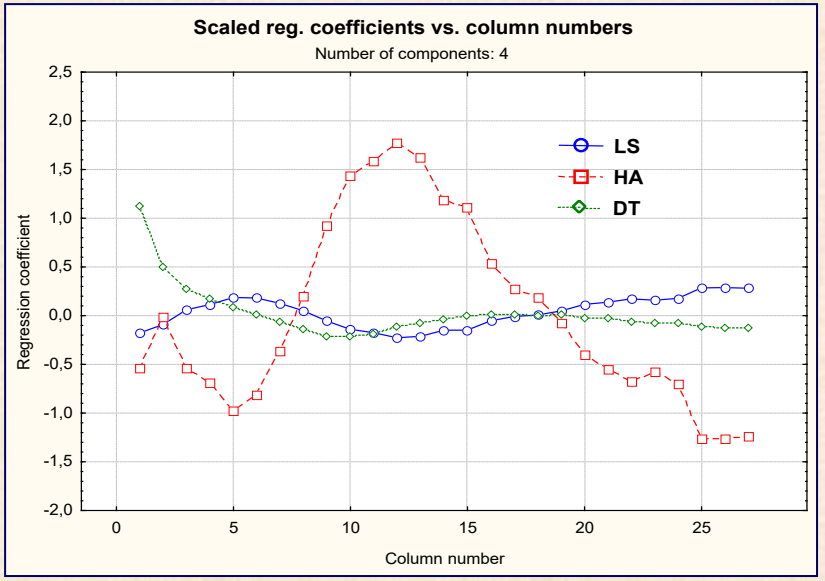


Régression PLS









Régression PLS

Predicted values Responses: LS HA DT						
	LS pred	LS obs	HA pred	HA obs	DT pred	DT obs
EM1	2,90	3,01	0,06	0,00	-6,19	0,00
EM2	-0,16	0,00	0,44	0,40	4,82	0,00
EM3	-0,06	0,00	-0,01	0,00	73,19	90,63
EM4	1,57	1,48	0,09	0,16	41,69	40,00
EM5	1,15	1,12	0,44	0,41	33,61	30,45
EM6	3,46	3,40	0,29	0,30	59,47	50,82
EM7	2,52	2,43	0,24	0,30	81,31	70,59
EM8	4,00	4,02	0,15	0,12	77,55	89,39
EM9	2,39	2,27	0,51	0,50	78,17	81,75
EM10	1,04	0,96	0,18	0,14	89,89	101,10
EM11	3,10	3,19	0,35	0,25	115,17	120,00
EM12	3,90	4,13	0,62	0,57	103,41	117,70
EM13	2,24	2,16	0,41	0,44	36,00	27,59
EM14	3,07	3,09	0,20	0,25	86,42	61,71
EM15	1,59	1,60	0,26	0,29	123,84	108,80
EM16	3,31	3,16	0,61	0,70	52,20	60,00

Prédictions erronées

valeurs observées = 0

difficultés reliées au seuil de détection du processus de mesurage

DT = 0 ou DT = 0,00 ??

Régression PLS avec JMP Pro
data = Postate Cancer.jmp 97 observations

11 variables

- ID
- poids
- age
- PSA
- volume
- hyperplasia
- invasion
- GleasonScore
- Y_degré
- Train_Test
- train_test

5 premières observations

	ID	poids	age	PSA	volume	hyperplasia	invasion	GleasonScore	Y_degré	Train_Test	train_test
1	1	15,959	50	0,651	0,5599	0,000	non	6	0,0000	train	1
2	2	27,660	58	0,852	0,3716	0,000	non	7	0,0000	train	1
3	3	14,732	74	0,852	0,6005	0,000	non	7	0,0000	train	1
4	4	26,576	58	0,852	0,3012	0,000	non	6	0,0000	train	1
5	5	30,877	62	1,448	2,1170	0,000	non	6	0,0000	train	1

analyses : 12 méthodes différentes

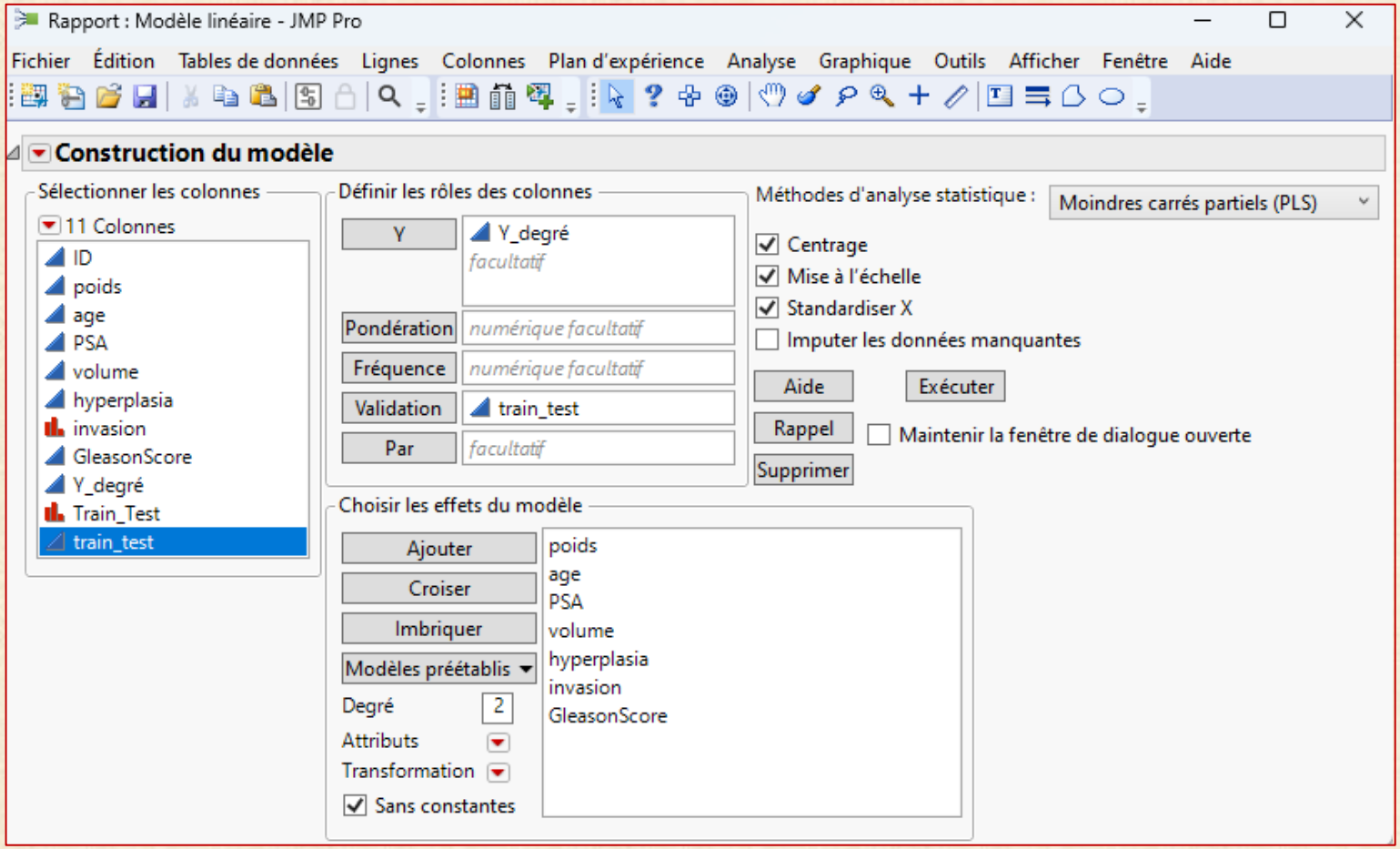
- Criblage du modèle de Y_degré
- Régression généralisée pour Y_degré
- Partition pour Y_degré
- Bootstrap forest pour Y_degré
- Boosted tree pour Y_degré
- Fit Model
- Fit Model-neurones
- Moindres carrés partiels
- Réseaux de neurones
- Ajustement pas-à-pas pour Y_degré
- Régression généralisée pour Y_degré 2
- Moindres carrés partiels 2

En particulier, par la méthode moindres carrés partiels (PLS)

Régression PLS avec JMP Pro
data = Postate Cancer.jmp 97 observations

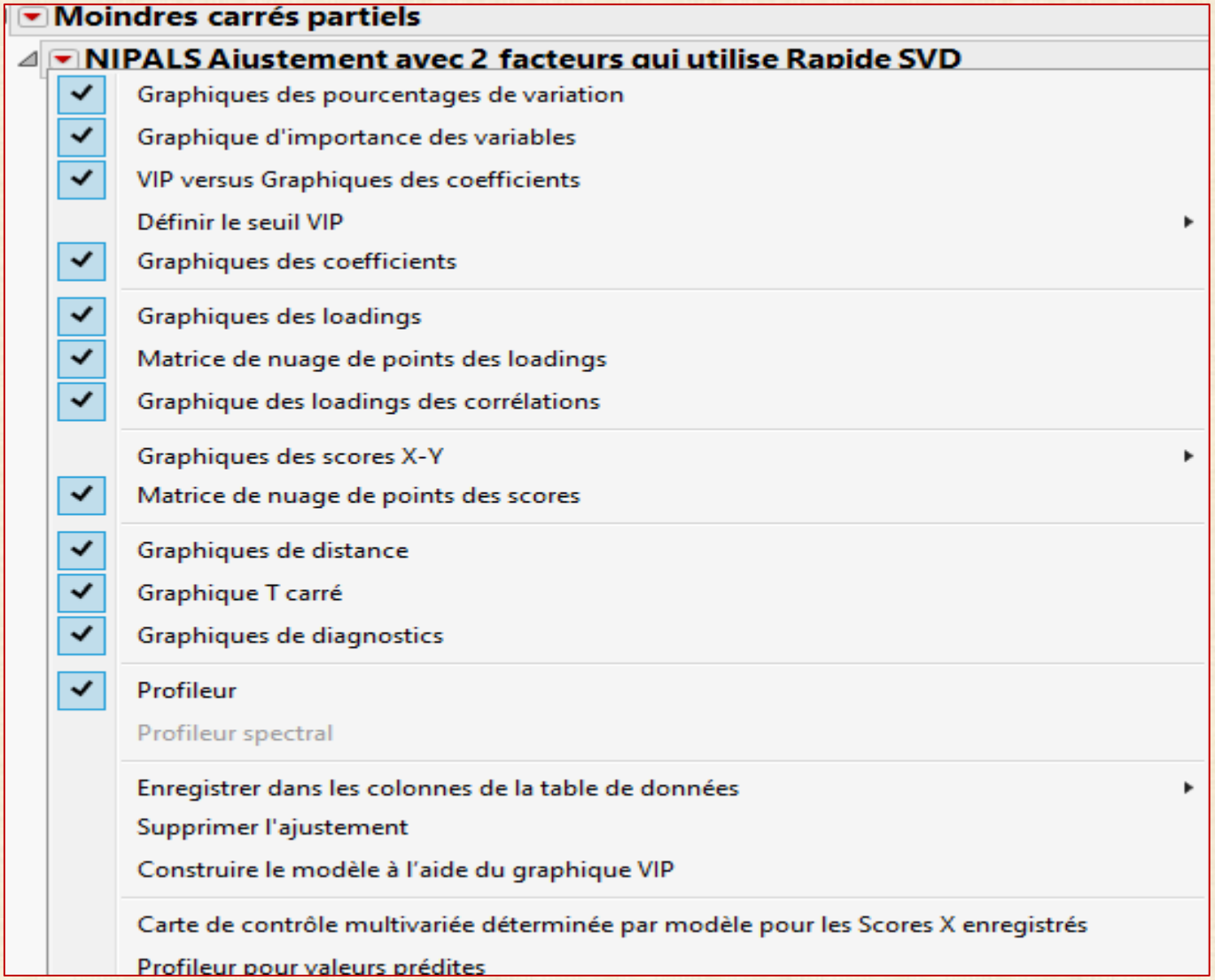
Mise en œuvre de l'analyse PLS classique (Y une réponse continue)
Modèle développé avec les 80 premières lignes (train-test = 1)

Boite de dialogue pour la mise en œuvre avec moindres carrés partiels.



Régression PLS avec JMP Pro
data = Postate Cancer.jmp 97 observations

Toutes les sorties de la procédure PLS; ….il y en a beaucoup …



Régression PLS avec JMP Pro
data = Postate Cancer.jmp 97 observations

Moindres carrés partiels

Réponse Y : Y_degré
Colonne de validation : train_test

Lancement du modèle

Spécification de la méthode

- (•) NIPALS
- ( ) SIMPLS

Spécification des facteurs

Nombre de facteurs (1 à max=8) : 2

Lancer

Résumé des comparaisons de modèles

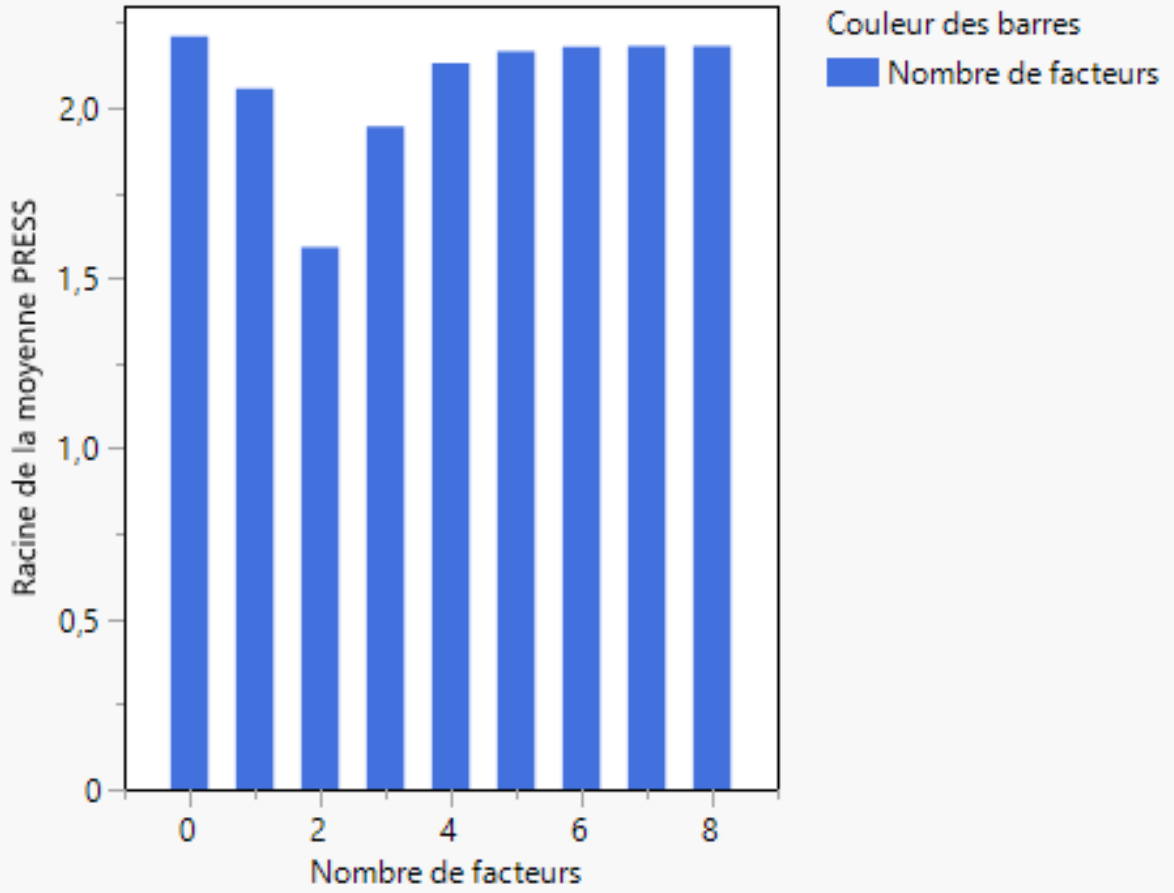
Méthode	SVD	Nombre de lignes	Nombre de facteurs	Pourcentage de variation expliquée pour X cumulé	Pourcentage de variation expliquée pour Y cumulé	Nombre de VIP > 0.8
NIPALS	Rapide	80	2	56,579185	68,285572	4

Validation croisée avec colonne de validation et méthode = NIPALS qui utilise Rapide SVD

Nombre de facteurs	Racine de la moyenne PRESS	T^2 de Van der Voet	Prob. > van der Voet T^2	Q^2	Cumulé(e) Q^2	D^2X	Cumulé(e) D^2X	D^2Y	Cumulé(e) D^2Y
0	2,210130	1,388769	0,2670	-0,587350	-0,587350	0,000000	0,000000	0,000000	0,000000
1	2,055514	1,680663	0,1240	-0,373023	-0,373023	0,387480	0,387480	0,641114	0,641114
2	1,590132	0,000000	1,0000	0,178319	-0,128186	0,178312	0,565792	0,041742	0,682856
3	1,946046	0,854258	0,4800	-0,230674	-0,388429	0,109539	0,675331	0,016528	0,699383
4	2,130364	1,091836	0,4480	-0,474839	-1,047709	0,100139	0,775469	0,003297	0,702681
5	2,165554	1,122205	0,4380	-0,523964	-2,120634	0,077919	0,853388	0,000430	0,703110
6	2,179005	1,137909	0,4130	-0,542956	-3,815001	0,077280	0,930668	0,000014	0,703124
7	2,179858	1,138946	0,4290	-0,544163	-6,435148	0,069332	1,000000	0,000000	0,703124
8	2,179858	1,138946	0,4480	-0,544163	-10,48108	0,000000	1,000000	0,000000	0,703124

Régression PLS avec JMP Pro
data = Postate Cancer.jmp 80 observations

Graphique de la racine de la moyenne PRESS

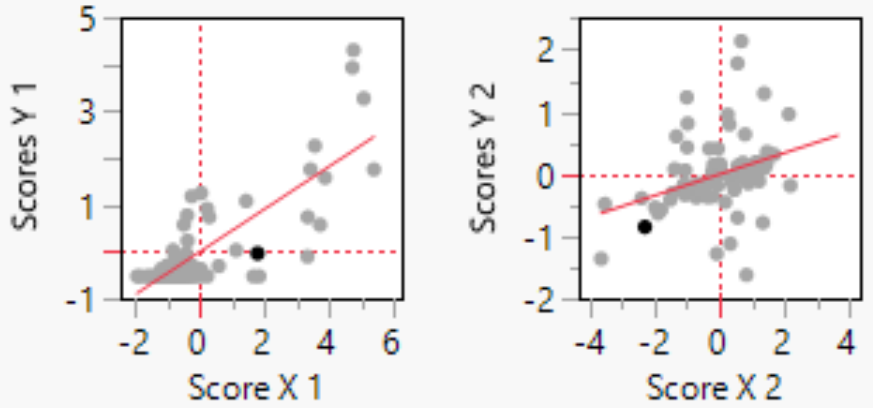


Remarque : la racine minimale de la moyenne PRESS est 1,59013 et le nombre de facteurs minimisant est 2.

Régression PLS avec JMP Pro
data = Postate Cancer.jmp 80 observations

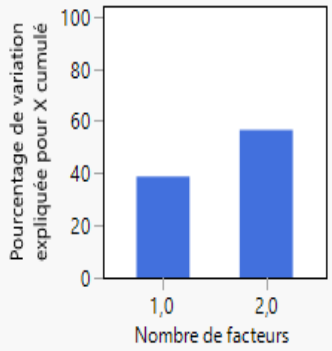
NIPALS Ajustement avec 2 facteurs qui utilise Rapide SVD

Graphiques des scores X-Y

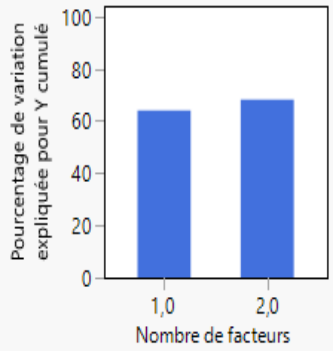


Pourcentage de variation expliquée

Nombre de facteurs	Effets de X	X cumulé
1	38,7480	38,748
2	17,8312	56,579



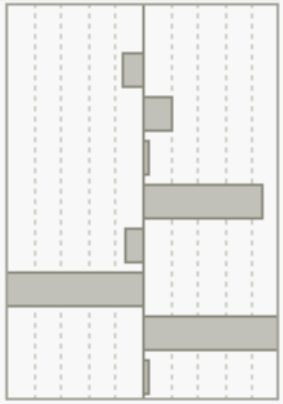
Réponses Y	Y cumulé
64,1114	64,1114
4,1742	68,2856



Régression PLS avec JMP Pro
data = Postate Cancer.jmp 80 observations

Coefficients du modèle pour les données centrées et réduites

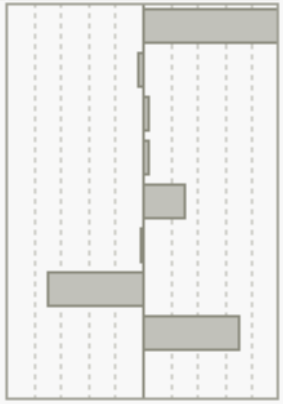
Terme	Y_degré
Constante	0,0000
poids	-0,0450
age	0,0625
PSA	0,0125
volume	0,2694
hyperplasia	-0,0386
invasion[non]	-0,3137
invasion[oui]	0,3137
GleasonScore	0,0102



Remarque: les variables X ont été standardisées.

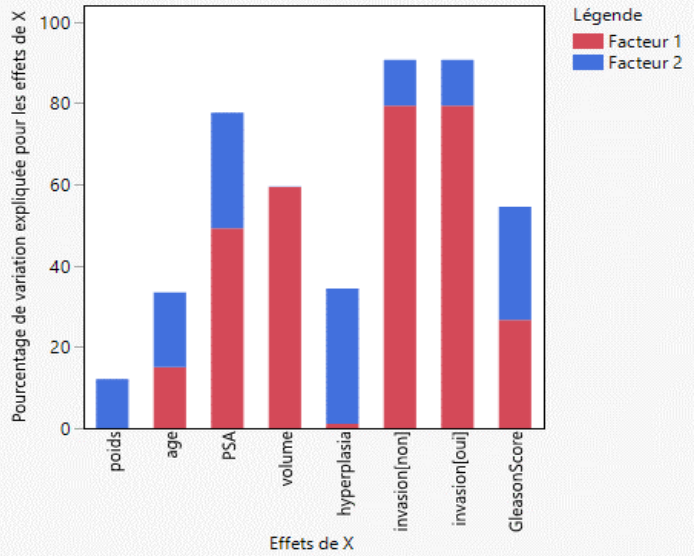
Coefficients du modèle pour les données originales

Terme	Y_degré
Constante	4,0168
poids	-0,1229
age	0,1873
PSA	0,1890
volume	1,2082
hyperplasia	-0,1085
invasion[non]	-2,7917
invasion[oui]	2,7917
GleasonScore	0,0328

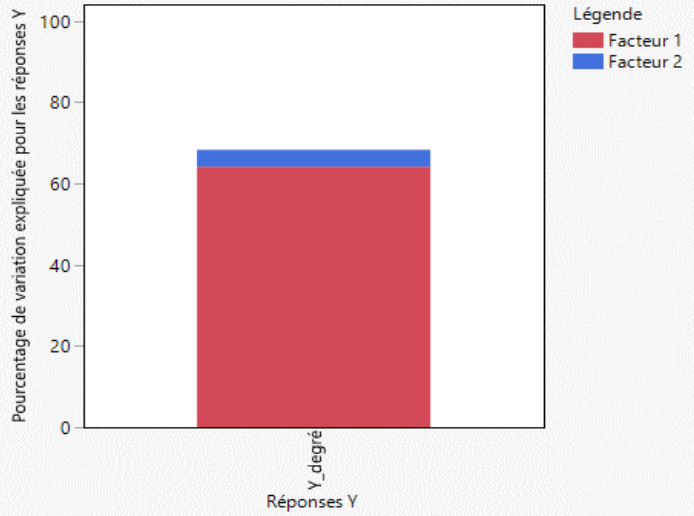


Remarque: les variables X ont été standardisées.

Pourcentage de variation expliquée pour les effets de X



Pourcentage de variation expliquée pour les réponses Y



Régression PLS avec JMP Pro
data = Postate Cancer.jmp 80 observations

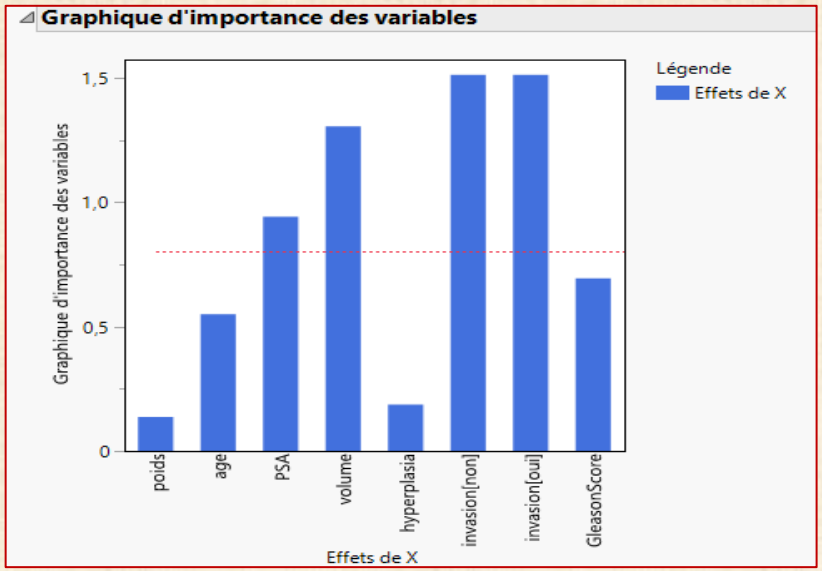
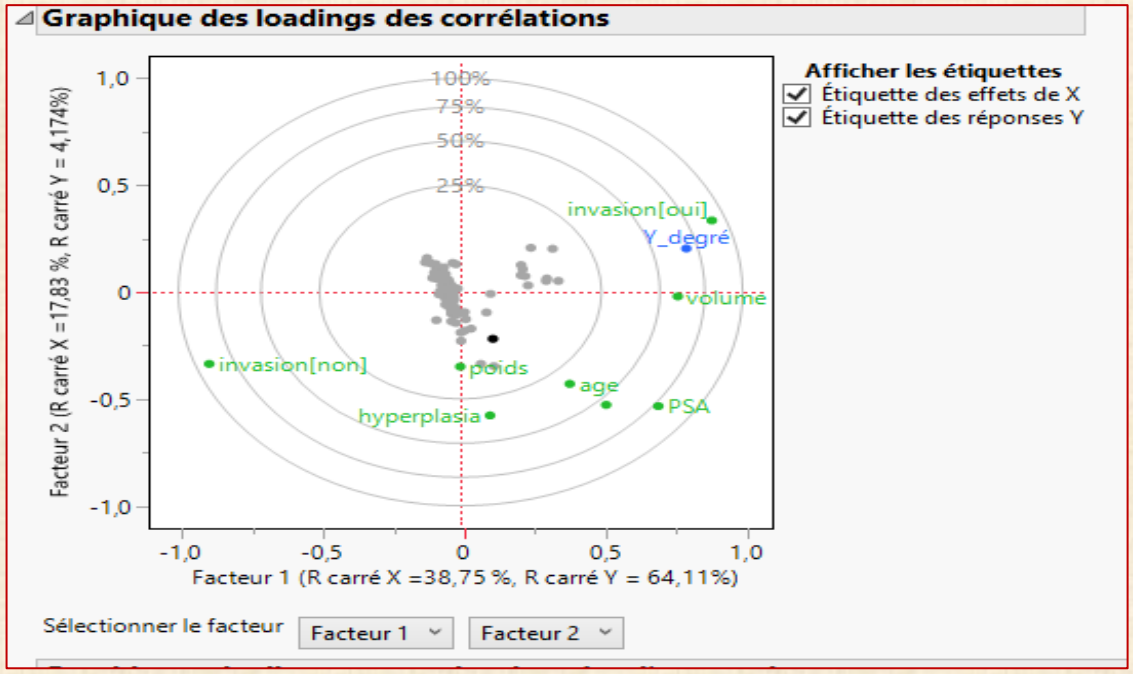


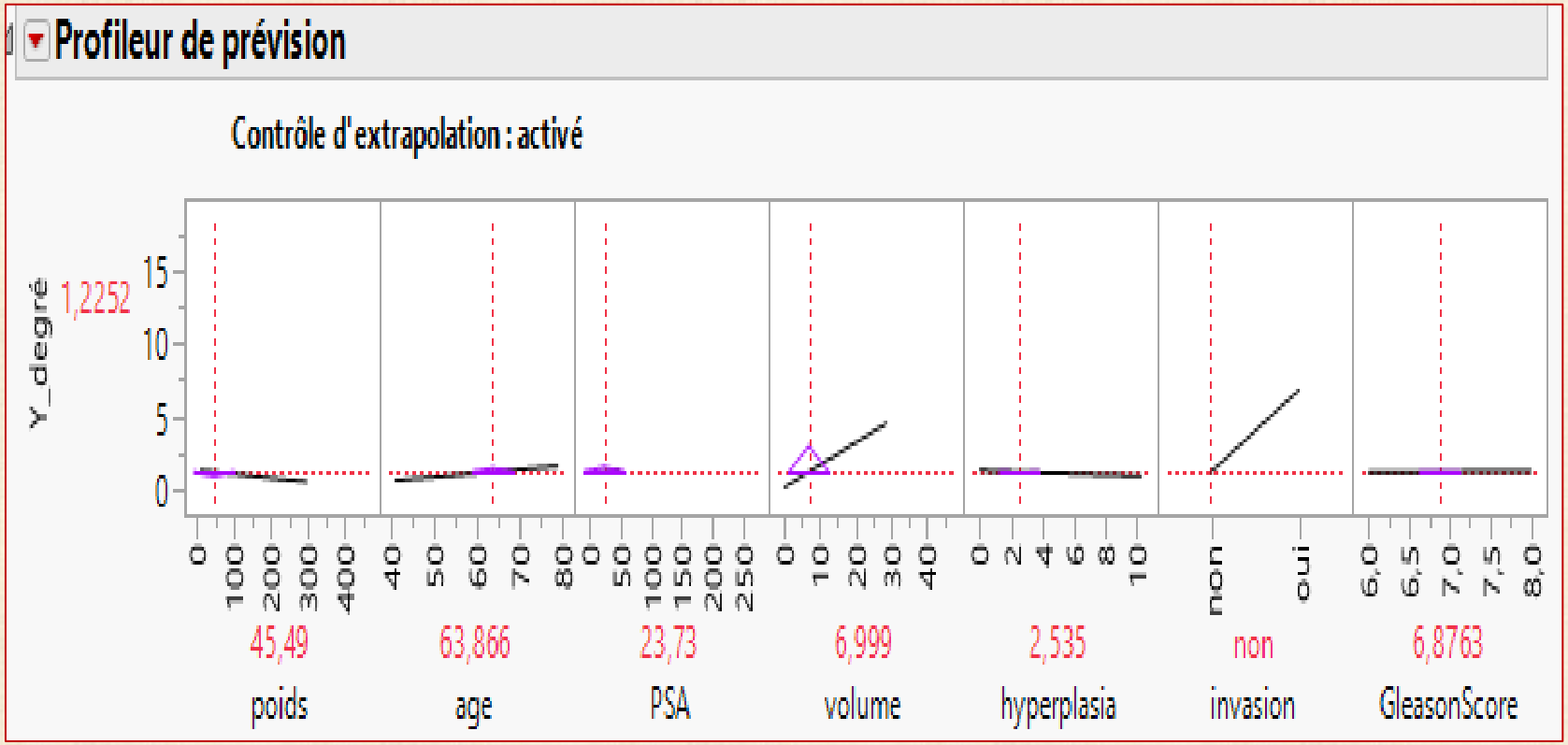
Tableau des valeurs d'importance des variables

X	Graphique d'importance des variables
poids	0,1376
age	0,5495
PSA	0,9421
volume	1,3044
hyperplasia	0,1863
invasion[non]	1,5123
invasion[oui]	1,5123
GleasonScore	0,6939

VIP versus Coefficients pour données centrées et réduites



Régression PLS avec JMP Pro
data = Postate Cancer.jmp 80 observations

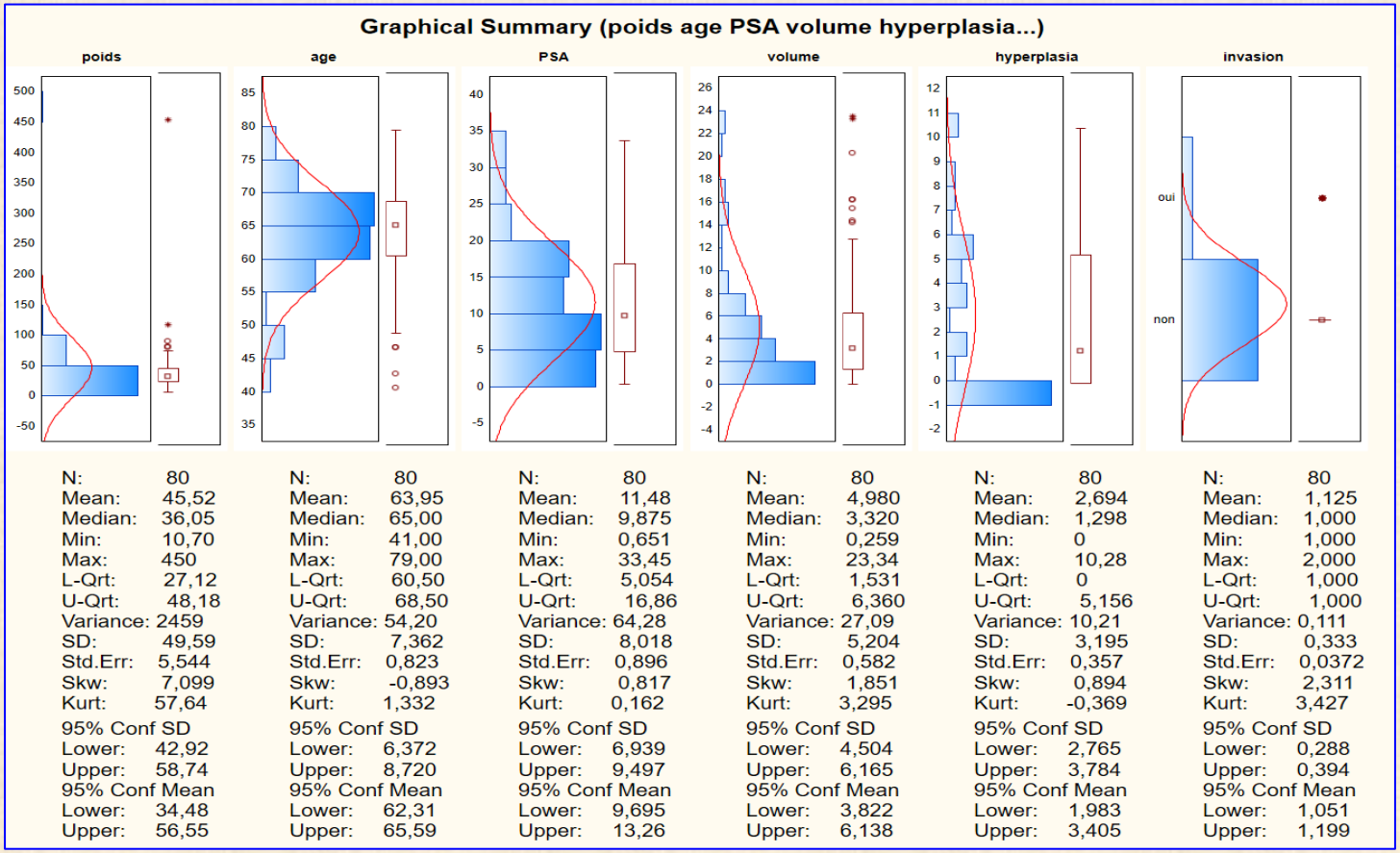


Régression PLS avec Statistica
data = Postate Cancer.sta n = 80 obs. parmi 97

DESCRIPTION DES DONNÉES

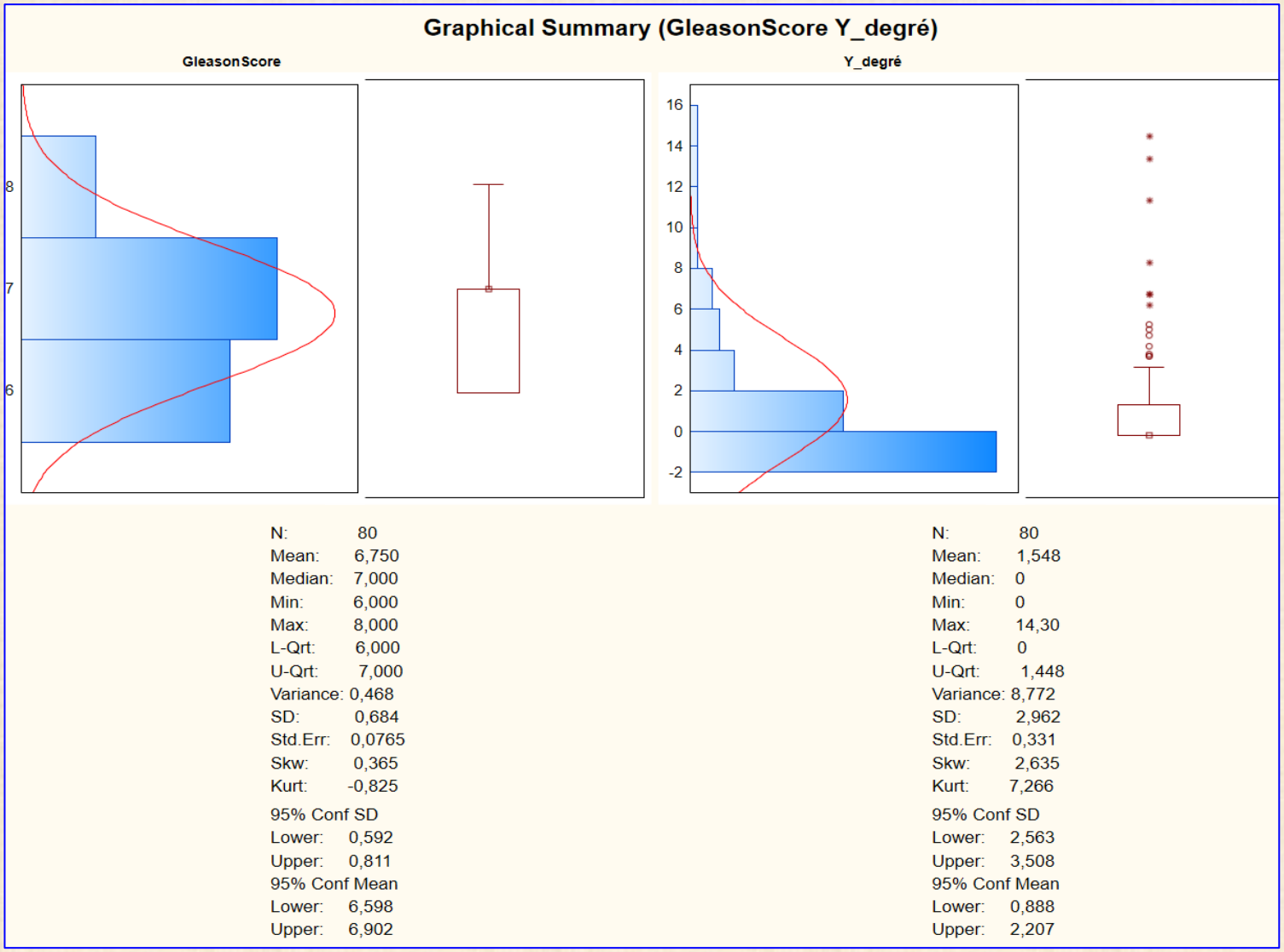
Subset of CancerProstate.sta
Variables: 1-11 Include condition: v11=1 n=80 obs.

	1 ID	2 poids	3 age	4 PSA	5 volume	6 hyperplasia	7 invasion	8 GleasonScore	9 Y_degré	10 Train_Test	11 train_test
1	1	15,959	50	0,651	0,5599	0	non	6	0	train	1
2	2	27,66	58	0,852	0,3716	0	non	7	0	train	1
3	3	14,732	74	0,852	0,6005	0	non	7	0	train	1



	poids	age	PSA	volume	hyperplasia	invasion
N:	80	80	80	80	80	80
Mean:	45,52	63,95	11,48	4,980	2,694	1,125
Median:	36,05	65,00	9,875	3,320	1,298	1,000
Min:	10,70	41,00	0,651	0,259	0	1,000
Max:	450	79,00	33,45	23,34	10,28	2,000
L-Qrt:	27,12	60,50	5,054	1,531	0	1,000
U-Qrt:	48,18	68,50	16,86	6,360	5,156	1,000
Variance:	2459	54,20	64,28	27,09	10,21	0,111
SD:	49,59	7,362	8,018	5,204	3,195	0,333
Std.Err:	5,544	0,823	0,896	0,582	0,357	0,0372
Skw:	7,099	-0,893	0,817	1,851	0,894	2,311
Kurt:	57,64	1,332	0,162	3,295	-0,369	3,427
95% Conf SD						
Lower:	42,92	6,372	6,939	4,504	2,765	0,288
Upper:	58,74	8,720	9,497	6,165	3,784	0,394
95% Conf Mean						
Lower:	34,48	62,31	9,695	3,822	1,983	1,051
Upper:	56,55	65,59	13,26	6,138	3,405	1,199

Régression PLS avec Statistica
data = Postate Cancer.sta 80 observations
DESCRIPTION DES DONNÉES



Régression PLS avec Statistica
data = Postate Cancer.sta n = 80 obs. parmi 97

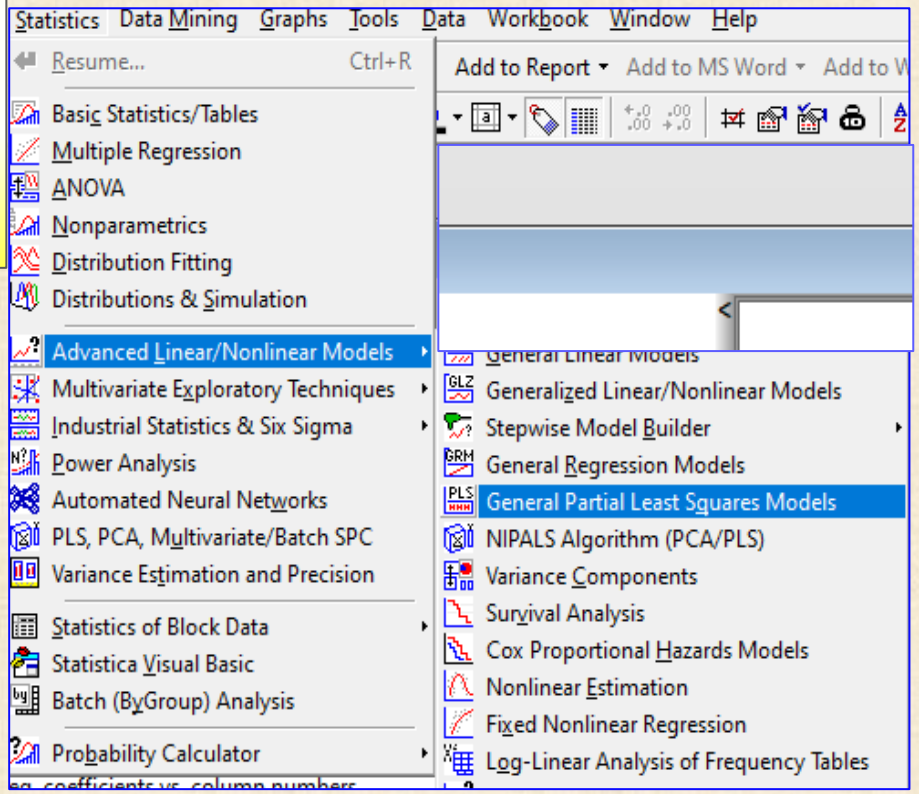
Subset of CancerProstate.sta
Variables: 1-11 Include condition: v11=1 n=80 obs.

	1 ID	2 poids	3 age	4 PSA	5 volume	6 hyperplasia	7 invasion	8 GleasonScore	9 Y_degré	10 Train_Test	11 train_test
1	1	15,959	50	0,651	0,5599	0	non	6	0	train	1
2	2	27,66	58	0,852	0,3716	0	non	7	0	train	1
3	3	14,732	74	0,852	0,6005	0	non	7	0	train	1

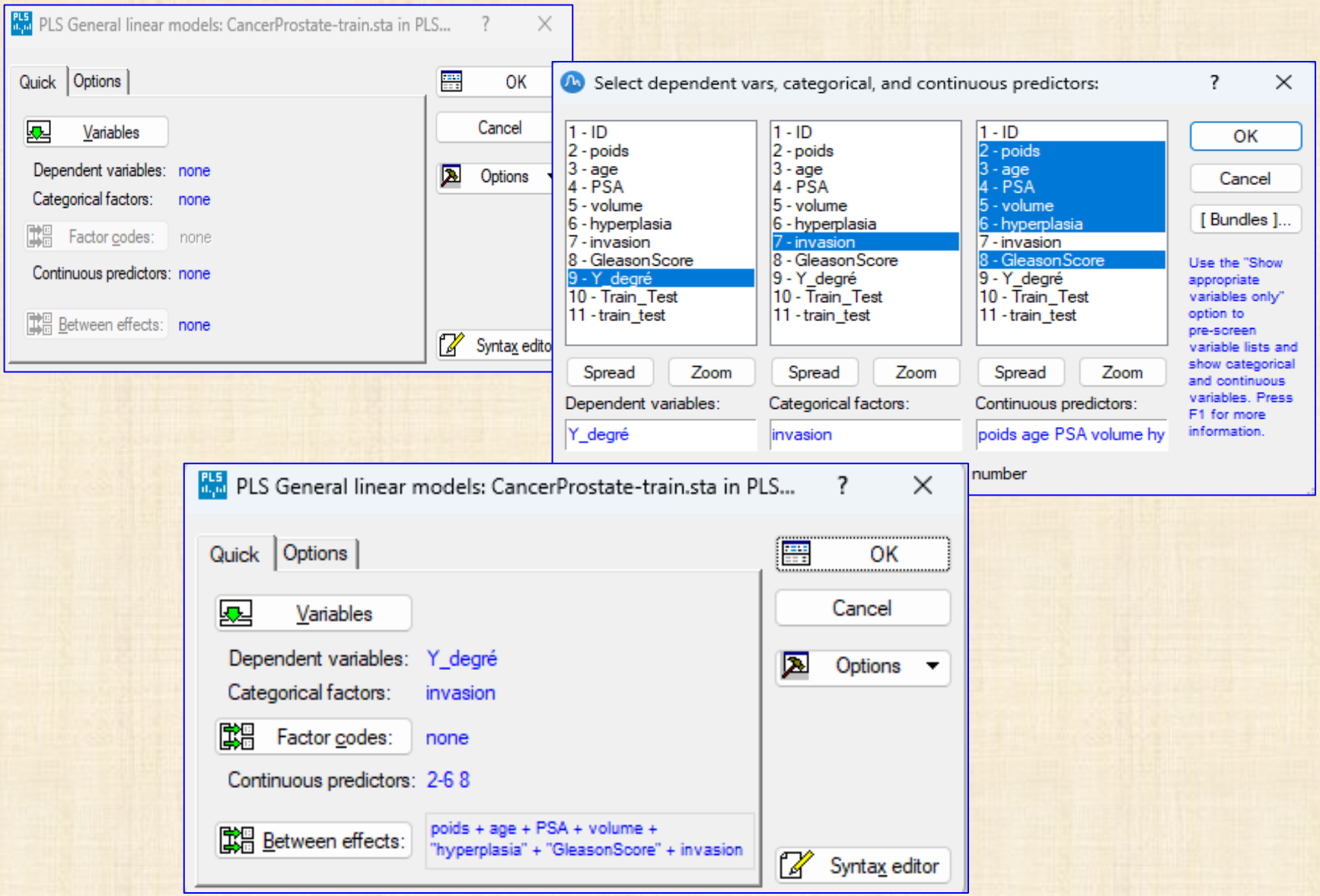
PLS avec STATISTICA

Advanced Linear / Nonlinear

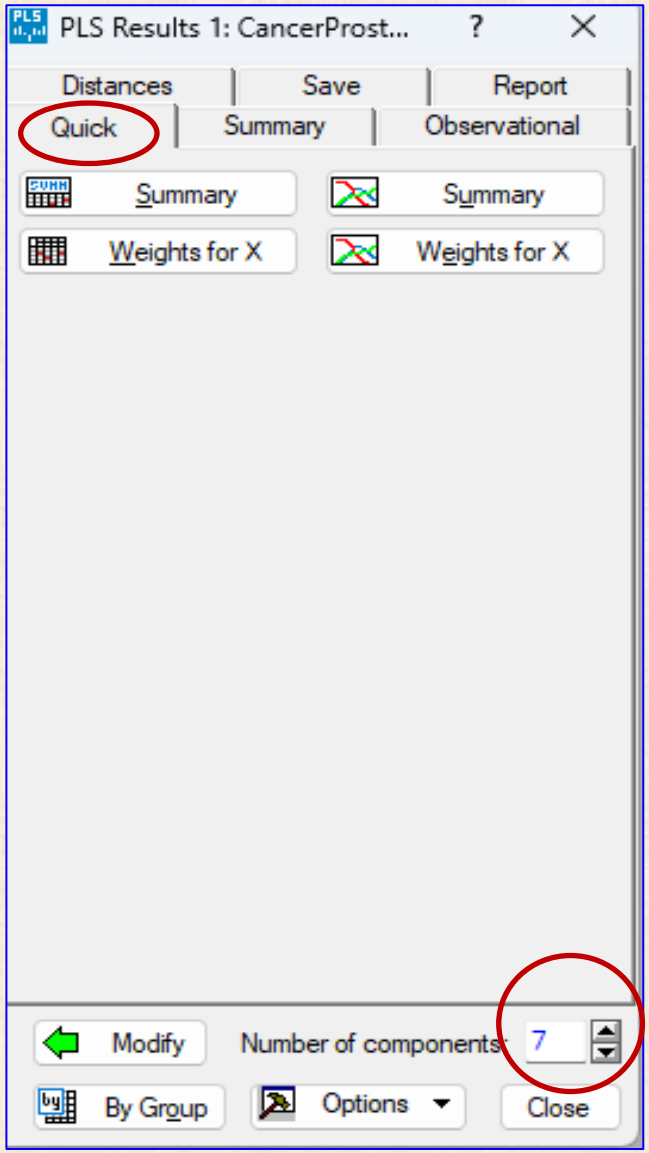
… General Partial Least Square Models

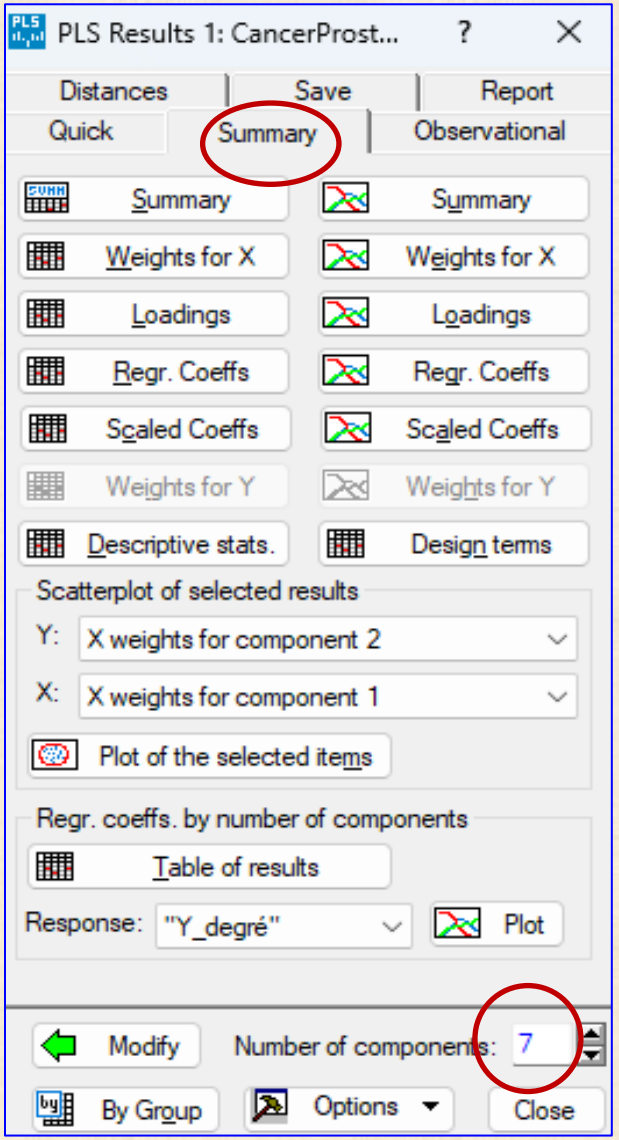


Régression PLS avec Statistica
data = Postate Cancer.sta 80 observations

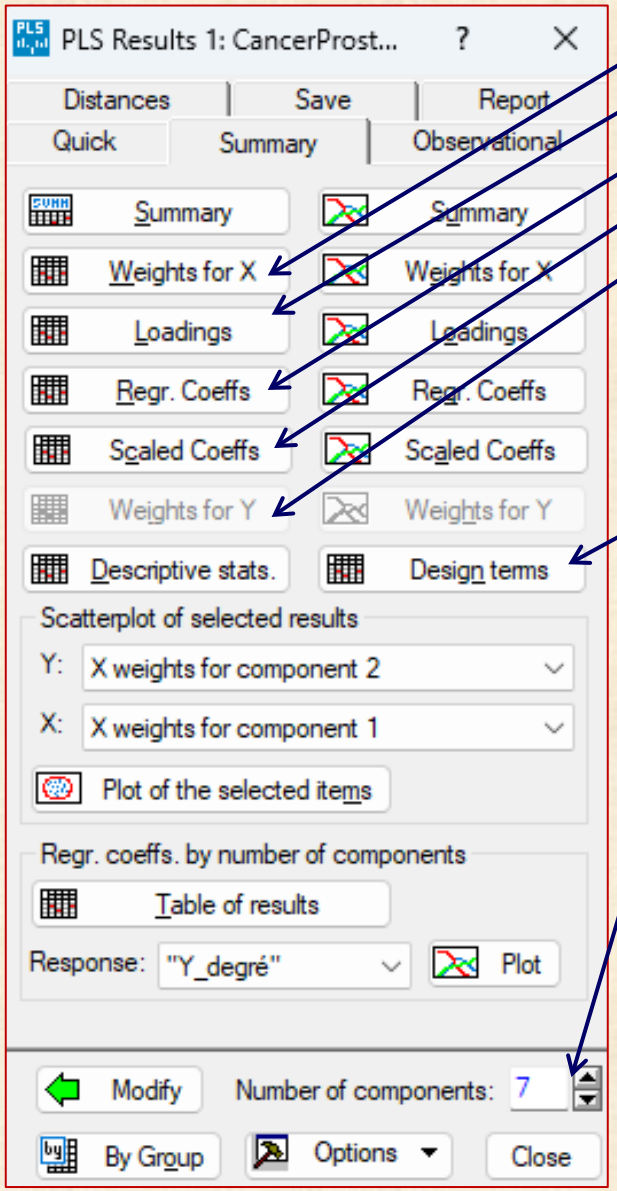


Régression PLS avec Statistica
data = Postate Cancer.sta 80 observations





Régression PLS : Partial Least Square



W : Weights for X

P : Loadings

B : Regr. Coeffs – variables dans leurs unités

Scaled Coeffs (beta) (var centrées-éduites)

Q : Weights for Y – seulement si plusieurs Y

identification du modèle employé

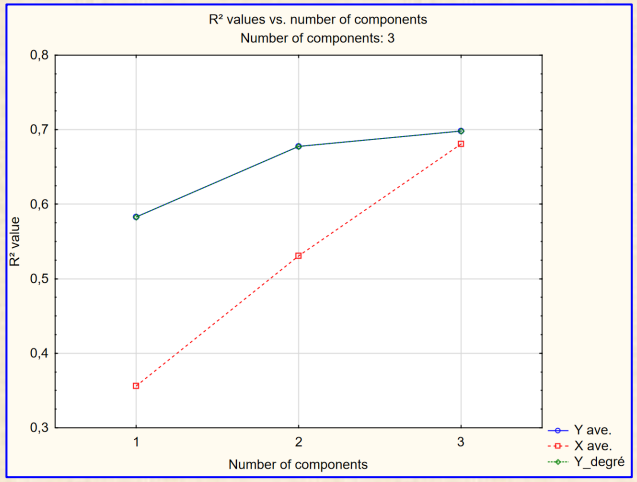
colonnes de la matrice design utile avec des variables catégoriques pour montrer comment les modalités d'une variable catégorique seront codées avec des variables à valeurs -1 0 1

nombre de composants : ici, 7 est le nombre maximal car il y a 7 variables continues;

en général, le nombre de composants utiles sera plus petit

Régression PLS avec Statistica
data = Postate Cancer.sta 80 observations

	Increase R^2 of Y	Average R^2 of Y	Increase R^2 of X	Average R^2 of X
Comp 1	0,582735	0,582735	0,356161	0,356161
Comp 2	0,094923	0,677658	0,174886	0,531048
Comp 3	0,020700	0,698359	0,149946	0,680994
Comp 4	0,004318	0,702677	0,083021	0,764015
Comp 5	0,000438	0,703115	0,085723	0,849738
Comp 6	0,000009	0,703124	0,074833	0,924571
Comp 7	0,000000	0,703124	0,075429	1,000000



Predictor weights (CancerProstate-train in PLS examples)
Responses: Y_degré
Options: NO-INTERCEPT AUTOSCALE

	poids	age	PSA	volume	hyperplasia	GleasonScore	invasion 1
Compo 1	-0,031652	0,233846	0,353134	0,566136	0,024549	0,260767	-0,656124
Compo 2	-0,146244	-0,180575	-0,451944	0,211327	-0,261173	-0,404603	-0,682040
Compo 3	0,335165	0,284103	-0,562333	0,079317	0,502090	-0,266250	-0,401817
Compo 4	-0,254649	-0,009790	-0,728234	0,298448	0,284131	0,484311	0,019336
Compo 5	0,387539	-0,703508	-0,183129	0,462265	0,263338	0,173820	0,090024
Compo 6	-0,417757	-0,049997	0,118623	0,534176	0,386006	-0,512035	0,335237
Compo 7	0,340129	0,386586	-0,196274	0,519482	-0,505348	-0,137858	0,389996

PLS scaled regression coefficients (CancerProstate-train in PLS examples)
Responses: Y_degré
Options: NO-INTERCEPT AUTOSCALE

	poids	age	PSA	volume	hyperplasia	GleasonScore	invasion 1
Y_degré	-0,022352	0,085028	-0,133702	0,407901	0,048566	0,009038	-0,603448

Régression PLS avec Statistica
Y_degré : observées VS prédites