Chapitre 5 Multiple-3

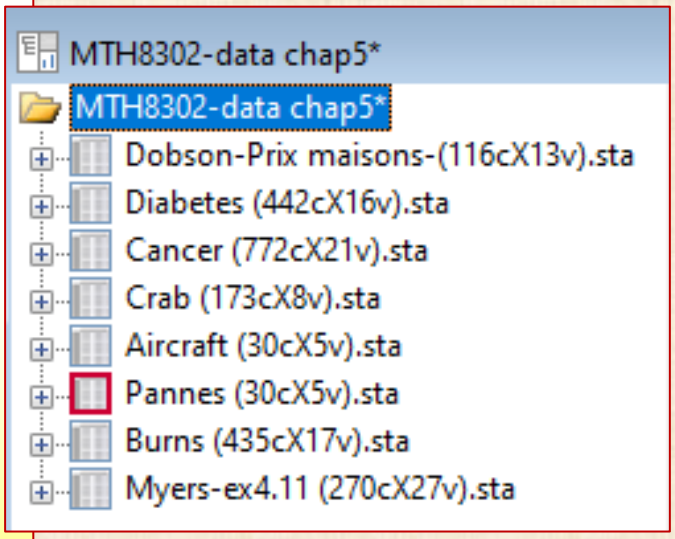
distribution réponse Y

- non symétrique (ex. gamma)
- binaire (multinomiale avec 2 modalités)
- Poisson (valeurs entières : 0, 1, 2, ..)
- multinomiale (k modalités $k \geq 3$)

cas 1 : modalités non ordonnées

cas 2 : modalités ordonnées

EXEMPLES



codage des variables catégoriques

Exemple de données avec variables continues et variables catégoriques

Dobson vol 1 p.322-323 Dobson vol 1 p.322-323

prix de maison en vente sur le marché résidentiel /
X1_SUP : superficie (pi.car) - variable continue / valeurs de 1000 à 1500
X2_AG : agence immobiliaire : A ou B - variable catégorique à 2 modalités
Z : codage de X2_AG : Z = 1 si X2_AG = A / Z = -1 si X2_AG = B
X3_PER : période de l'année ; variable catégorique avec 3 modalités
jan-fev-mar-avr = 1 / jui-aou-sep-oct = 2 / ma-ju-nov-dec = 3
Z1 Z2 Z3 : codage disjonctif complet de X3_PER
Z1 = 1 si X3_PER = jan-fev-mar-avr / = 0 sinon
Z2 = 1 si X3_PER = jui-aou-sep-oct / = 0 sinon
Z3 = 1 si X3_PER = ma-ju-nov-dec / =0 sinon
Z4, Z5 : codage è effet de X3_PER
Z4 : 1 vs 3 / Z2 : 2 vs 3
Y_PR = prix_demandé
/

Dobson-Prix maisons.ca

1 ID	2 X1_SUP	3 X2_AG	4 Z	5 X3_PER	6 info	7 Z1	8 Z2	9 Z3	10 info	11 Z4	12 Z5	13 Y_PR
1	1165	A	1	jan-fev-mar-avr	Z1 Z2 Z3	1	0	0	Z4 Z5	1	0	395,0
2	1170	B	-1	jan-fev-mar-avr		1	0	0		1	0	425,0
3	1160	B	-1	jan-fev-mar-avr	codage	1	0	0	codage	1	0	455,0
4	1306	B	-1	jan-fev-mar-avr	disjonctif	1	0	0	à effet	1	0	539,0
5	1120	A	1	jan-fev-mar-avr	complet	1	0	0	de X3_PER	1	0	415,0
6	1040	B	-1	jan-fev-mar-avr	de X3_PER	1	0	0	modalité de	1	0	425,0
7	1130	B	-1	jan-fev-mar-avr		1	0	0	référence	1	0	427,5
8	1232	A	1	jan-fev-mar-avr		1	0	0	= 3	1	0	449,0
9	1364	A	1	jan-fev-mar-avr		1	0	0		1	0	488,0
10	1260	B	-1	jan-fev-mar-avr		1	0	0		1	0	472,0
110	1152	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	515,0
111	1001	A	1	ma-ju-nov-dec		0	0	1		-1	-1	439,0
112	1227	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	509,0
113	1308	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	529,0
114	1113	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	487,0
115	1345	A	1	ma-ju-nov-dec		0	0	1		-1	-1	525,0
116	1392	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	529,0

variables catégoriques : 3 types de codage
codage 1 : disjonctif complet
codage 2 : disjonctif complet restreint
codage 3 : codage à effet

codage disjonctif complet avec variables indicatrices

X3_PER : catégorique 3 modalités m1 m2 m3
m1=jan-fev-mar-avr / m2=jui-aou-sep-oct
m3=ma-ju-nov-dec
codage avec Z1 Z2 Z3 à valeurs 1 ou 0
selon modalité m1 m2 m3
contrainte : Z1 + Z2 + Z3 = 1
conséquence : multi colinéarité

codage disjonctif complet restreint

X3_PER : catégorique avec 3 modalités
m1=jan-fev-mar-avr / m2=jui-aou-sep-oct
m3=ma-ju-nov-dec
codage avec 2 indicatrices Z1 Z2
selon modalité m1 m2 seulement
m3 représentée par Z1 = 0 Z2 = 0
pas de contrainte entre Z1 et Z2
conséquence : représente effet général dans modèle statistique

Codage à effet (recommandé) utilisé par Statistica

U variable catégorique modalités m1 m2 … mk
utilisation de (k-1) variables, disons Z, à valeurs 1 / 0 / -1
m_k modalité de référence pour comparaison aux autres

$Z_1 = 1$ si $U = m_1$
$= 0$ si $U = m_2,\ldots,m_{k-1}$
$= -1$ si $U = m_k$
$Z_2 = 1$ si $U = m_2$
$= 0$ si $U = m_1,\ldots,m_{k-1}$
$= -1$ si $U = m_k$
……
$Z_{k-1} = 1$ si $U = m_{k-1}$
$= 0$ si $U = m_1,\ldots,m_{k-2}$
$= -1$ si $U = m_k$

Exemple X3_PER catégorique 3 modalités codée avec Z4 et Z5 seulement

Exemple X2_AG : catégorique 2 modalités A B codée avec Z = 1 ou -1 selon modalité A ou B
cas particulier codage à effet

Modèles avec variables continues et variables catégoriques

DIFFÉRENTS MODÈLES pour mesurer l'influence des 3 variables sur Y_prix

3 variables explicatives X : X1_SUP X2_AG X3_PER influence (?) sur Y_prix

M1 : Y vs SUP (1) $Y\_prix = \beta_0 + \beta_1 * X1\_SUP$ X2_AG et X3_PER pas tenu en compte

M2 : bottom-up 2 modèles selon les 2 valeurs de X2_AG X3_PER pas tenu en compte

(2a) $Y\_prix\_A = \beta_{0A} + \beta_{1A} * X1\_SUP$ pour X2_AG = A

(2b) $Y\_prix\_B = \beta_{0B} + \beta_{1B} * X1\_SUP$ pour X2_AG = B

l'influence variable AG : représentée par les 4 coefficients des 2 modèles

test : si $\beta_{0A} = \beta_{0B}$ et $\beta_{1A} = \beta_{1B}$ alors X2_AG pas d'influence sur Y_prix

M3 : top-down modèle avec X2_AG remplacée par Z X3_PER pas tenu en compte

(3) $Y\_prix = \beta_0 + \beta_1 * X1\_SUP + \beta_2 * Z$

l'influence variable AG : représentée par β_2

test : si $\beta_2 = 0$ alors X2_AG n'influence pas sur Y_prix

Question quelle approche bottom-up ou top-down préférable ? **Réponse : ???**

M4 : top down modèle avec 3 facteurs : X1_SUP X2_AG X 3_PER

X2_AG remplacée par Z et X3_PER remplacée par Z4 Z5 (codage à effet)

(4) $Y\_prix = \beta_0 + \beta_1 * X1\_SUP + \beta_2 * Z + \beta_3 * Z4 + \beta_4 Z * 5$

M5 : général (5) $Y\_prix = \beta_0 + \beta_1 * X1\_SUP + \beta_2 * Z + \beta_3 * Z4 + \beta_4 * Z5 + \beta_5 * X1\_SUP * Z + \beta_6 * X1\_SUP * Z4 + \beta_7 * X1\_SUP * Z5 + \beta_8 * Z * Z4 + \beta_9 * Z * Z5$

contient 5 interactions entre les 4 variables X2_SUP Z Z4 Z5

Remarques

1 : ne pas mettre des interactions entre les variables de codage représentant les modalités d'une variable catégorique

2 : ne pas mettre d'effet quadratique pour les variables de codage représentant les modalités d'une variable catégorique

3 : M4 (eq. 4) est appelé un modèle d'analyse de covariance : pas d'interaction entre un facteur continu X2_SUP et des facteurs catégoriques représentés par les variables Z Z4 Z5

4 : en présence de variables catégoriques et continues, l'intérêt de de savoir si les facteurs catégoriques sont significatifs enlevant les effets des variables continues sur la réponse Y

Modèles avec variables continues et variables catégoriques

M6 modèle global avec le codage disjonctif complet Z1 Z2 Z3 (variable X3_PER)

M6 (6) $Y\_prix = \beta_0 + \beta_1 * X1\_SUP + \beta_2 * Z + \beta_3 * Z1 + \beta_4 * Z2 + \beta_5 * Z3 +$
$+ \beta_6 * X1\_SUP * Z + \beta_7 * X1\_SUP * Z1 + \beta_8 * X1\_SUP * Z2 + \beta_9 * X1\_SUP * Z3$
$+ \beta_{10} * Z * Z1 + \beta_{11} * Z * Z2 + \beta_{12} * Z * Z3$

matrice design

X1-SUP	Z	Z1	Z2	Z3
x1	1	1	0	0
x2	-1	0	0	0
x3	1	0	0	1

Z1 + Z2 + Z3 = 1
dépendance linéaire

M7 modèle global avec le codage disjonctif complet restreint avec Z1 Z2 seulement
quand Z3 = 1 alors Z1 = Z2 = 0 l'équation (6) devient

(7) $Y\_prix = \beta_0 + \beta_1 * X1\_SUP + \beta_2 * Z + \beta_5 + \beta_6 * X1\_SUP * Z + \beta_9 * X1\_SUP + \beta_{12} * Z$

$= (\beta_0 + \beta_5) + (\beta_1 + \beta_9) * X1\_SUP + (\beta_2 + \beta_{12}) * Z + \beta_6 * X1\_SUP * Z$

l'effet général β_0 est modifié par β_5 qui est l'effet de la modalité m3

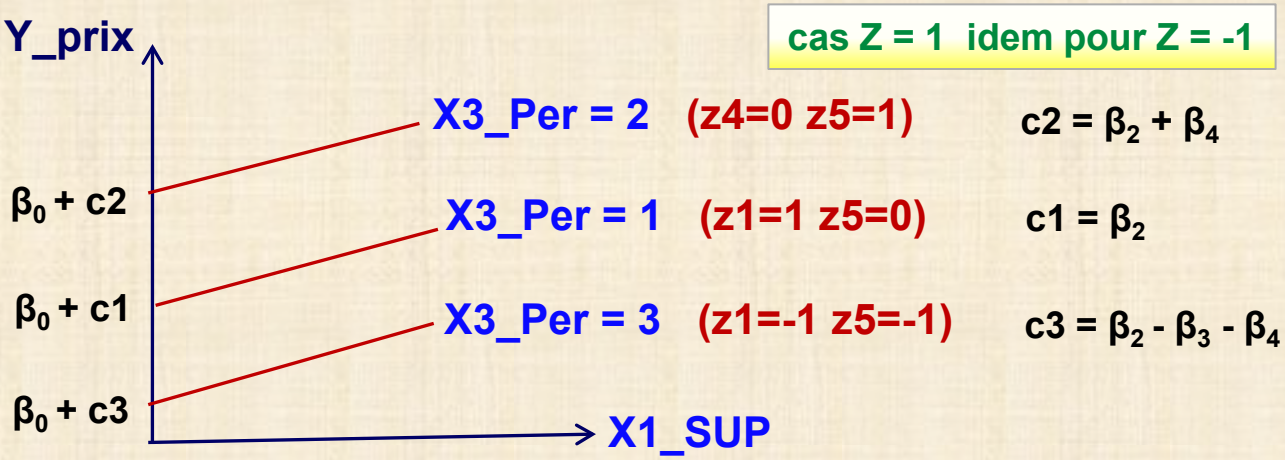
cette situation n'est pas présente avec le codage à effet
conclusion : codage a effet est préférable au codage disjonctif complet restreint

Modèle d'analyse de covariance : pas d'interaction entre les variables continues et les variables catégoriques

M4 : (4) $Y\_prix = \beta_0 + \beta_1 * X1\_SUP + \beta_2 * Z + \beta_3 * Z4 + \beta_4 * Z5$

cas Z = 1 idem pour Z = -1

droites parallèles : cordonnées à l'origine possiblement distinctes



$c2 = \beta_2 + \beta_4$

$c1 = \beta_2$

$c3 = \beta_2 - \beta_3 - \beta_4$

droites parallèles: permet de tester l'influence de la variable catégorique X3_Per via Z4 et Z5 indépendamment de la valeur de la variable continue X1_SUP

Test de H0 : c1 = c2 = c3 = 0 ($\beta_2 = \beta_3 = \beta_4 = 0$) si rejetée alors X3_Per influence Y_Prix

général : comparaison des différents modèles se fait par l'intermédiaire du test F

Test : si $F = [SS_{MC} - SS_{MR} / 3] / [SSE_{MC} / (n - 5)] > F_{3,\, n-5,\, \alpha}$

si on ne rejette pas H_0 modèle réduit (MR) est OK

si on rejette H_0 modèle complet (MC) est OK

Si interaction : variable S (superficie) et variables catégoriques Z4, Z5
ajout S*Z4 S*Z5 dans le modèle

Modèles avec pentes distinctes

interaction entre la variable continue (covariable = S)
et la variable catégorique X3_PER

MC : Modèle Complet

$$Y = \beta_0 + \beta_1 * S + \beta_2 * Z4 + \beta_3 * Z5 + \beta_4 * S*Z4 + \beta_5 * S*Z5$$

Sous modèles: MR1 MR2 MR3

MR1 : $Y = \beta_0 + \beta_1 * S + \beta_2 * Z4 + \beta_3 * Z5$

H_0: $\beta_4 = \beta_5 = 0$ pentes identiques - interceptes distincts

MR2 : $Y = \beta_0 + \beta_1 * S + \beta_4 * S*Z4 + \beta_5 * S*Z5$

H_0: $\beta_2 = \beta_3 = 0$ pentes distinctes - interceptes identiques

MR3 : $Y = \beta_0 + \beta_1 * Z$

H_0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ modèle identique 3 périodes

Utilisation de tests F pour choisir le modèle approprié

MODÈLES : variables continues + variables catégoriques

3 Variables : X1_SUP / X2_AG / X1_PER — MODÈLES

Dobson vol 1 p.322-323 Dobson vol 1 p.322-323

prix de maison en vente sur le marché résidentiel /
X1_SUP : superficie (pi.car) - variable continue / valeurs de 1000 à 1500
X2_AG : agence immobiliaire : A ou B - variable catégorique à 2 modalités
Z : codage de X2_AG : Z = 1 si X2_AG = A / Z = -1 si X2_AG = B
X3_PER : période de l'année ; variable catégorique avec 3 modalités
jan-fev-mar-avr = 1 / jui-aou-sep-oct = 2 / ma-ju-nov-dec = 3
Z1 Z2 Z3 : codage disjonctif complet de X3_PER
Z1 = 1 si X3_PER = jan-fev-mar-avr / = 0 sinon
Z2 = 1 si X3_PER = jui-aou-sep-oct / = 0 sinon
Z3 = 1 si X3_PER = ma-ju-nov-dec / =0 sinon
Z4, Z5 : codage è effet de X3_PER
Z4 : 1 vs 3 / Z2 : 2 vs 3
Y_PR = prix_demandé
/

1 ID	2 X1_SUP	3 X2_AG	4 Z	5 X3_PER	6 info	7 Z1	8 Z2	9 Z3	10 info	11 Z4	12 Z5	13 Y_PR
1	1165	A	1	jan-fev-mar-avr	Z1 Z2 Z3	1	0	0	Z4 Z5	1	0	395,0
2	1170	B	-1	jan-fev-mar-avr		1	0	0		1	0	425,0
3	1160	B	-1	jan-fev-mar-avr	codage	1	0	0	codage	1	0	455,0
4	1306	B	-1	jan-fev-mar-avr	disjonctif	1	0	0	à effet	1	0	539,0
5	1120	A	1	jan-fev-mar-avr	complet	1	0	0	de X3_PER	1	0	415,0
6	1040	B	-1	jan-fev-mar-avr	de X3_PER	1	0	0	modalité de	1	0	425,0
7	1130	B	-1	jan-fev-mar-avr		1	0	0	référence	1	0	427,5
8	1232	A	1	jan-fev-mar-avr		1	0	0	= 3	1	0	449,0
9	1364	A	1	jan-fev-mar-avr		1	0	0		1	0	488,0
10	1260	B	-1	jan-fev-mar-avr		1	0	0		1	0	472,0

. . . . . .

110	1152	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	515,0
111	1001	A	1	ma-ju-nov-dec		0	0	1		-1	-1	439,0
112	1227	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	509,0
113	1308	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	529,0
114	1113	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	487,0
115	1345	A	1	ma-ju-nov-dec		0	0	1		-1	-1	525,0
116	1392	B	-1	ma-ju-nov-dec		0	0	1		-1	-1	529,0

M1 Y_prix vs X1_SUP

selon agence immobilière (X2_AG)
X3_PER pas tenu en compte
Q : droites distinctes selon agence X2_AG?

M2 Y_prix vs X1_AG + X2_SUP

modèle d'analyse de covariance (ANCOVA)
pas d'interaction entre X1 et X2
conséquence : pentes égales
X3_PER pas tenu en compte

M3 Y_prix vs X1_AG + X2_SUP + X1*X2

modèle avec interaction
conséquence : pentes distinctes
X3_PER pas tenu en compte

M4 : Y_prix vs X1_SUP + X2_AG + X3_PER + X1*X2 + X1*X3 + X2*X3

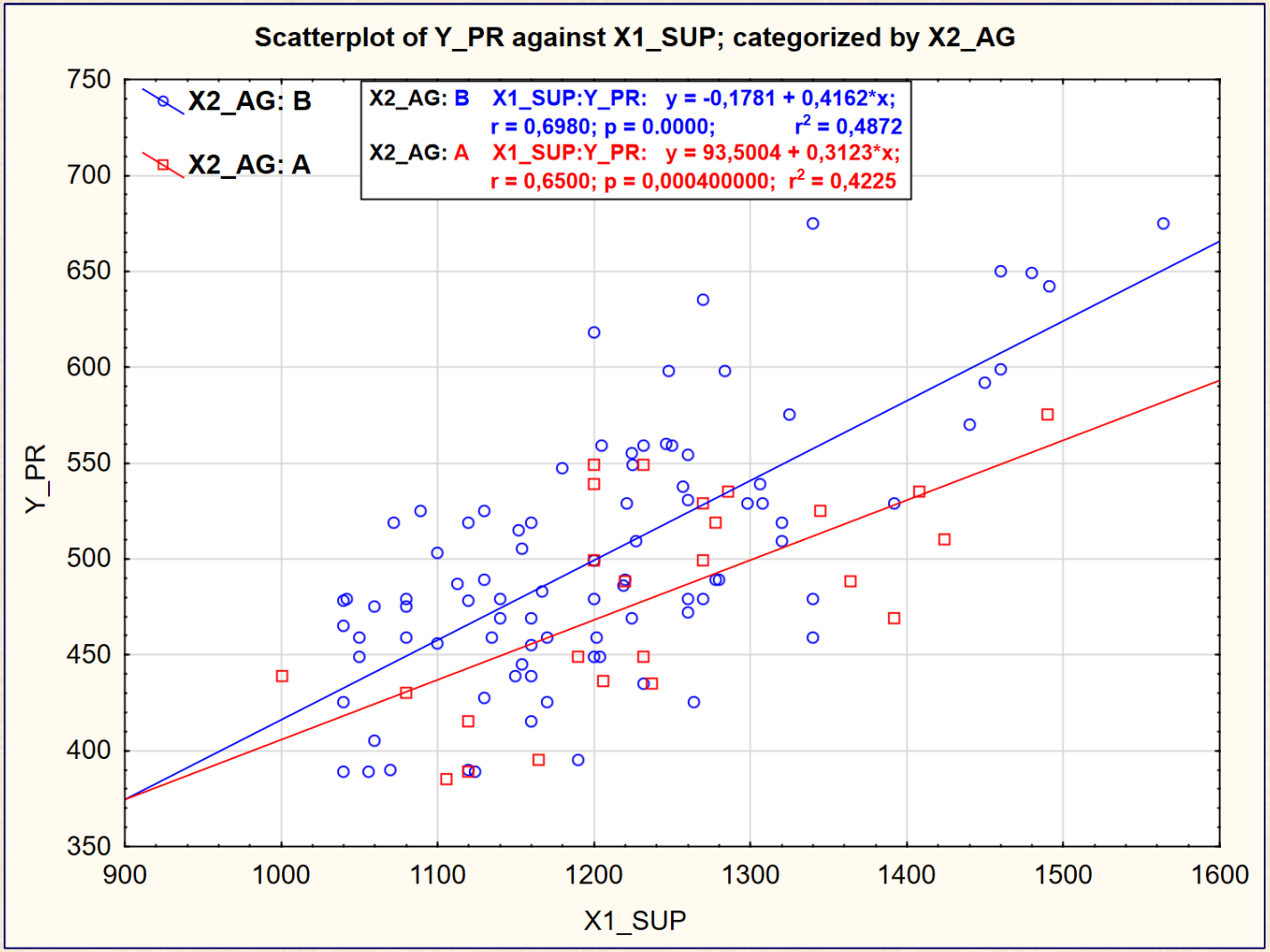
modèle avec interactions d'ordre 2
conséquence : pentes distinctes

M1 Y_prix vs X1_SUP

selon agence immobilière (X2_AG)

X3_PER pas tenu en compte

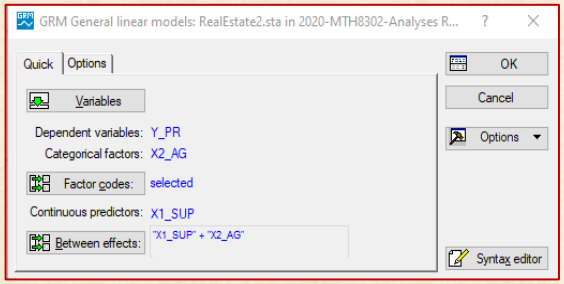
Q : droites distinctes selon agence X2_AG?



M2 Y_prix vs X1_AG + X2_SUP

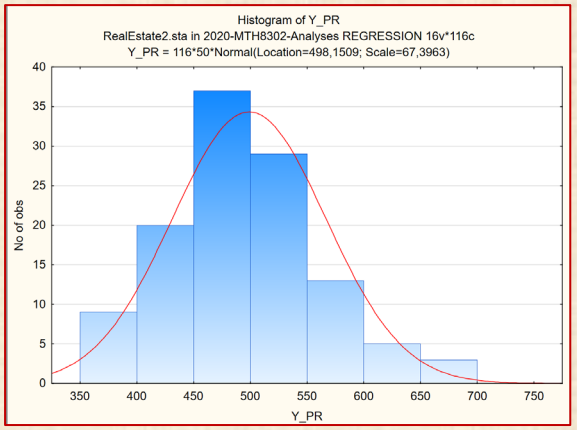
modèle d'analyse de covariance (ANCOVA)
pas d'interaction entre X1 et X2
conséquence : pentes égales
X3_PER pas tenu en compte

avec GRM

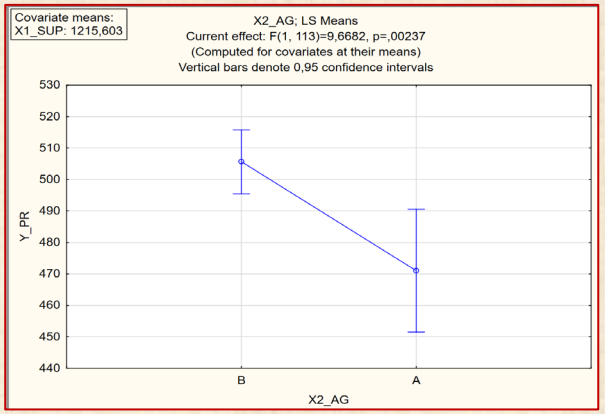


Label	Column	Variable	Level of Variable	versus Level
	Column Labels (Labels for the columns of the design matrix X			
Intercept	1			
X1_SUP	2	X1_SUP		
X2_AG	3	X2_AG	B	A

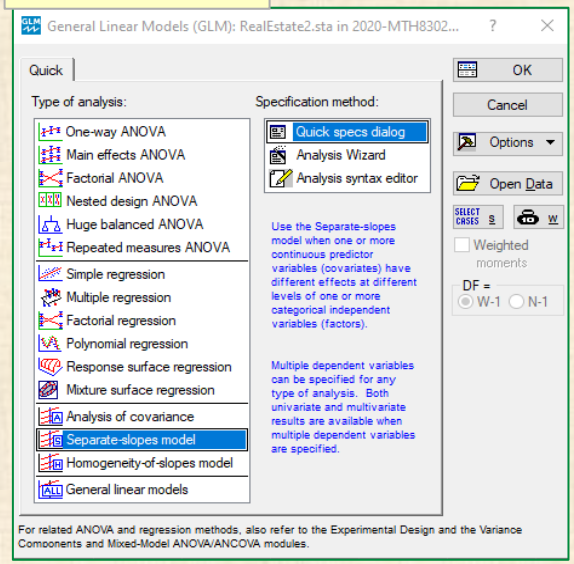
Effect	Degr. of Freedom	Y_PR SS	Y_PR MS	Y_PR F	Y_PR p
Intercept	1	76,0	76,0	0,0317	0,859014
X1_SUP	1	242421,9	242421,9	101,1653	0,000000
X2_AG	1	23167,8	23167,8	9,6682	0,002373
Error	113	270781,4	2396,3		
Total	115	522360,1			

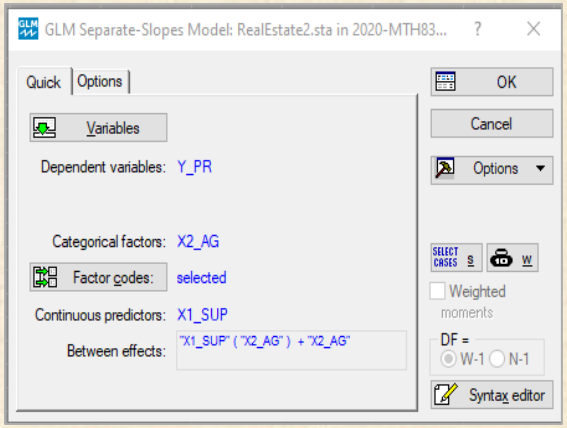


Effect	Level of Effect	Column	Y_PR Param.	Y_PR Std.Err	Y_PR t	Y_PR p
Intercept		1	8,61276	48,37665	0,17804	0,859014
X1_SUP		2	0,39461	0,03923	10,05809	0,000000
X2_AG	B	3	17,30205	5,56449	3,10937	0,002373



avec GLM

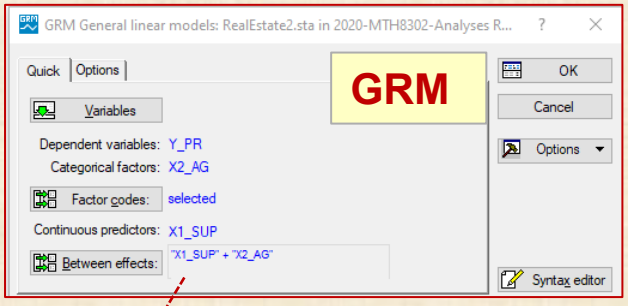




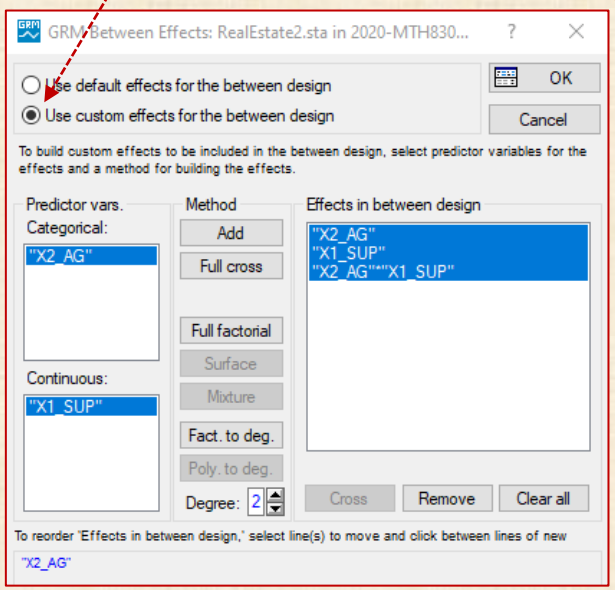
GLM résultats identiques à GRM

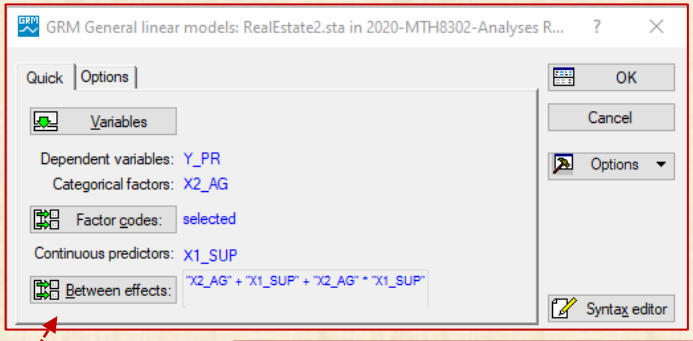
M3 Y_prix vs X1_AG + X2_SUP + X1*X2

modèle avec interaction
conséquence : pentes distinctes
X3_PER pas tenu en compte



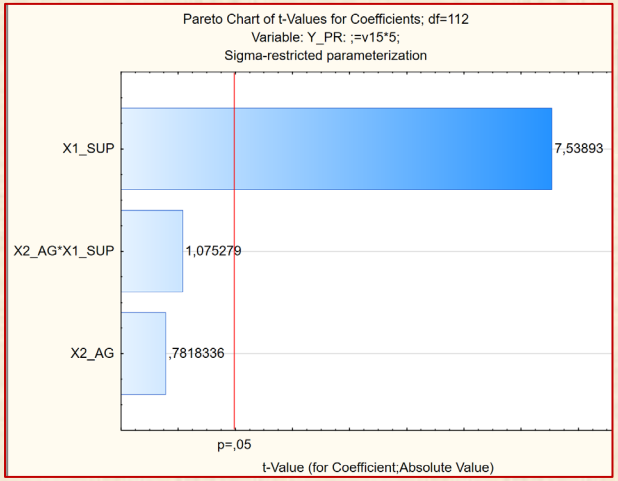
spécification modèle





Effect	Level of Effect	Column	Y_PR Param.	Y_PR Std.Err	Y_PR t	Y_PR p
Intercept		1	46,6612	59,90948	0,778861	0,437704
X2_AG	B	2	-46,8392	59,90948	-0,781834	0,435962
X1_SUP		3	0,3643	0,04832	7,538930	0,000000
X2_AG*X1_SUP	1	4	0,0520	0,04832	1,075279	0,284561

Effect	Degr. of Freedom	Y_PR SS	Y_PR MS	Y_PR F	Y_PR p
Intercept	1	1451,6	1451,6	0,60662	0,437704
X2_AG	1	1462,7	1462,7	0,61126	0,435962
X1_SUP	1	136006,6	136006,6	56,83546	0,000000
X2_AG*X1_SUP	1	2766,8	2766,8	1,15622	0,284561
Error	112	268014,6	2393,0		
Total	115	522360,1			

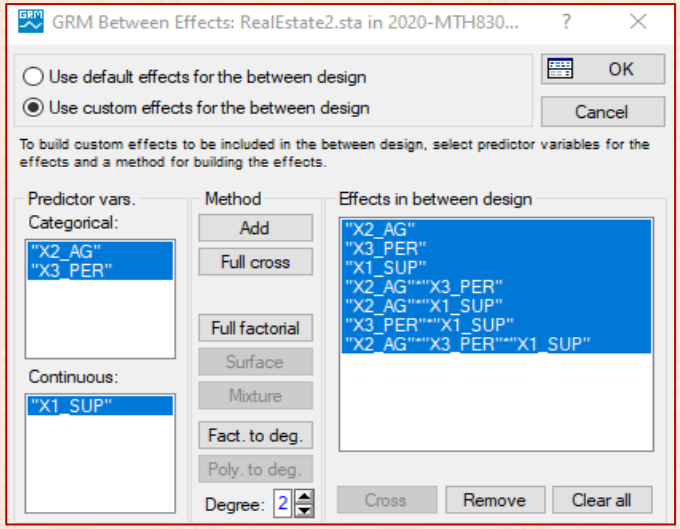


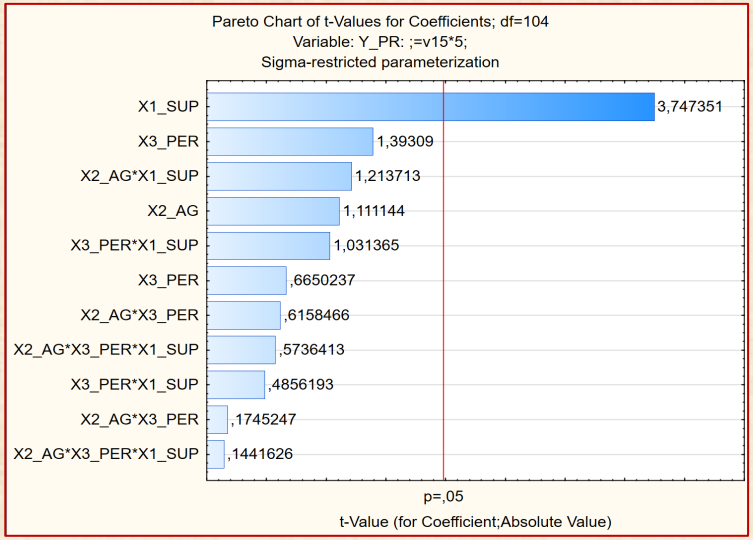
GLM : résultats identiques à GRM

M4 : Y_prix vs X1_SUP + X2_AG + X3_PER + X1*X2 + X1*X3 + X2*X3

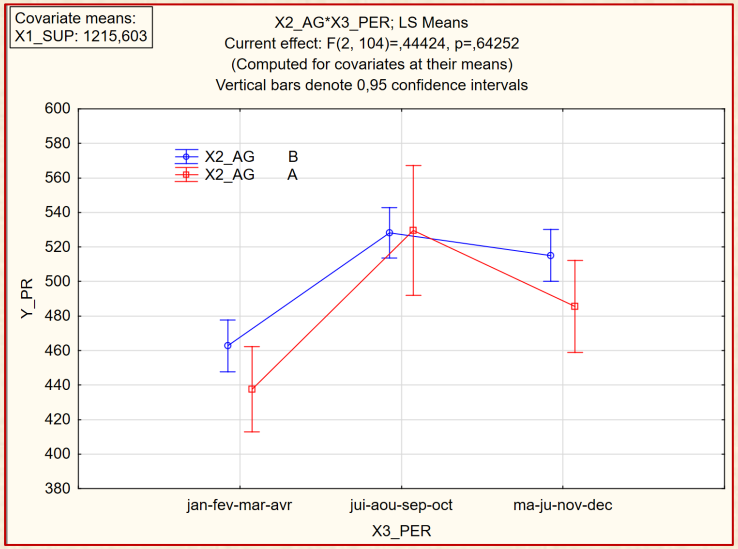
modèle avec interactions d'ordre 2
conséquence : pentes distinctes

Analyse avec GRM





Effect	Degr. of Freedom	Y_PR SS	Y_PR MS	Y_PR F	Y_PR p
Intercept	1	2008,5	2008,54	1,30482	0,255958
X2_AG	1	1900,5	1900,52	1,23464	0,269068
X3_PER	2	4768,8	2384,40	1,54899	0,217325
X1_SUP	1	21616,2	21616,22	14,04264	0,000294
X2_AG*X3_PER	2	1367,7	683,84	0,44424	0,642520
X2_AG*X1_SUP	1	2267,6	2267,58	1,47310	0,227607
X3_PER*X1_SUP	2	2638,5	1319,27	0,85705	0,427390
X2_AG*X3_PER*X1_SUP	2	1258,4	629,19	0,40875	0,665546
Error	104	160090,0	1539,33		
Total	115	522360,1			



Régressions pénalisées

méthode de vraisemblance pénalisée pour ajuster un modèle de régression classique ou logistique.
Pénalisation effectuée sur une grille de valeurs d'un paramètre d'ajustement λ
Résultat : tend à réduire les coefficients du modele vers 0. (comme ridge) – utile en multicolinéarité.

FOB : Fonction OBjective : à minimiser

$$FOB = (1 / N) \sum l(\beta, x, y, w) + P(\beta, \alpha, \lambda)$$

$$= 1 / N) \sum (1/2) (y - y)^2 + P(\beta, \alpha, \lambda)$$

α paramètre d'ajustement
β paramètres du modèle
λ paramètre de mélange variant entre 0 et 1
N nombre d'observations

$$l(\beta, x, y, w) = \sum (1/2) (y - \hat{y})^2$$ fonction de vraisemblance

régression Ridge	$\alpha = 0$	$P = \lambda \|\|\beta\|\|_2^2 / 2$	norme quadratique
régression Lasso	$\alpha = 1$	$P = \lambda \|\|\beta\|\|_1$	norme valeur absolue
régression Elasticnet	$\alpha = 0,1$	$P = \lambda [0,9*\|\|\beta\|\|_2^2/2 + 0,1*\|\|\beta\|\|_1]$	

LASSO est une méthode d'estimation de régression généralisée qui applique une pénalité L1 (valeur absolue) à la vraisemblance lors de l'estimation des paramètres.

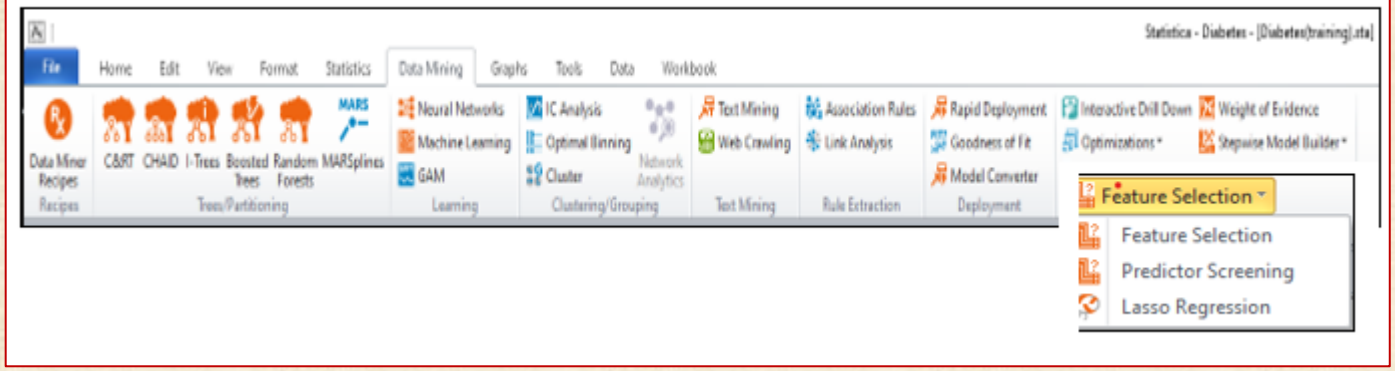
LASSO fait une sélection des variables (régresseurs X) et tend à choisir un modèle plus parcimonieux en présence de variables corrélées.

LASSO ne peut pas sélectionner plus de régresseurs que d'observations lorsque que le nombre de régresseurs est plus grand que le nombre d'observations.

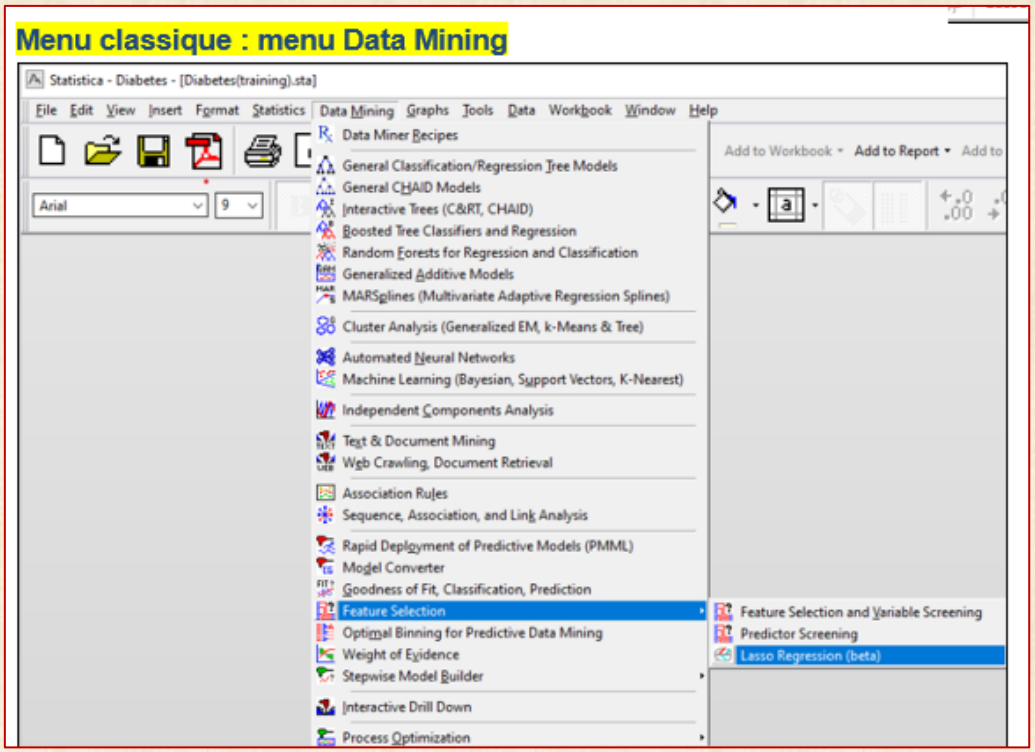
La version adaptative de LASSO pénalise moins les variables actives que les variables inactives et se rapproche asymptotiquement d'un modèle incluant uniquement des régresseurs actifs.

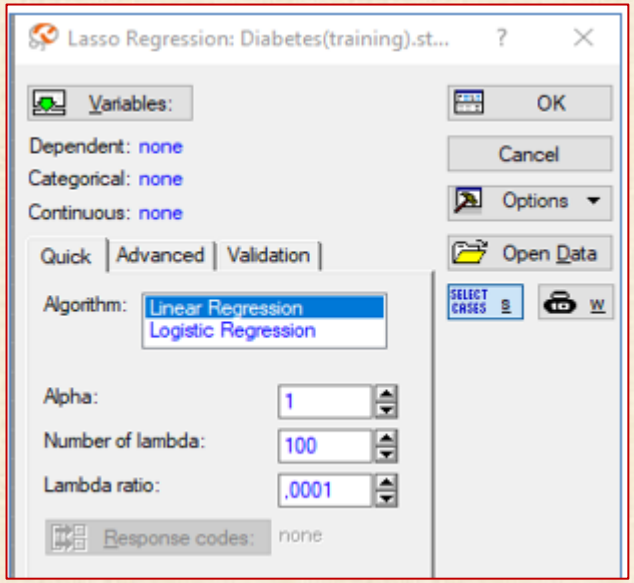
Régression pénalisée avec STATISTICA

Barre menu DataMining



Menu classique : menu Data Mining





Régression pénalisée avec STATISTICA

EXEMPLE DE DONNÉES diabetes.sta diabetes.jmp

Efron, B., Hastie, T., Johnstone, J., and Tibshirani, R. (2004). Least Angle Regression
Annals of Statistics (with discussion), vol. 32, pp. 407-499.

DATA Diabetes n = 442 observations X 13 variables : 10 X explicatives
X = Age Gender BMI BP Total Cholesterol LDL HDL TCH LGT Glucose
Y : 3 réponses Y1 Y2 Y3 (continue, binaire, ordinale)

Y1_continue is a quantitative measure of disease progression one year after baseline. (25 à 346)
Y2_binaire et Y3_ordinale sont des recodages de Y1_continue
Y2_binaire = Low si Y_continue = 200 ou moins / = High si Y_continue 201 ou plus
Y3_ordinale = Low si Y_continue = 150 ou moins
= Medium si Y_continue comprise entre 151 et 200
= High si Y_continue = 201 ou plus
observations séparées en 2 groupes : Training (309 obs.) (v16=1) Validation (133 obs.) (v16=2)

DATA

	1 ID	2 Age	3 Gender	4 BMI	5 BP	6 Total Cholesterol	7 LDL	8 HDL	9 TCH	10 LTG	11 Glucose	12 Y1_continue	13 Y2_binaire	14 Y3_ordinale	15 Validation	16 Validation2
1	1	59	2	32,1	101	157	93,2	38	4,00	4,8598	87	151	Low	Medium	Training	1
2	2	48	1	21,6	87	183	103,2	70	3,00	3,8918	69	75	Low	Low	Validation	2
3	3	72	2	30,5	93	156	93,6	41	4,00	4,6728	85	141	Low	Low	Training	1
4	4	24	1	25,3	84	198	131,4	40	5,00	4,8903	89	206	High	High	Training	1
5	5	50	1	23,0	101	192	125,4	52	4,00	4,2905	80	135	Low	Low	Training	1
.....																
437	437	33	1	19,5	80	171	85,4	75	2,00	3,9703	80	48	Low	Low	Training	1
438	438	60	2	28,2	112	185	113,8	42	4,00	4,9836	93	178	Low	Medium	Training	1
439	439	47	2	24,9	75	225	166,0	42	5,00	4,4427	102	104	Low	Low	Validation	2
440	440	60	2	24,9	99,67	162	106,6	43	3,77	4,1271	95	132	Low	Low	Validation	2
441	441	36	1	30,0	95	201	125,2	42	4,79	5,1299	85	220	High	High	Validation	2
442	442	36	1	19,6	71	250	133,2	97	3,00	4,5951	92	57	Low	Low	Validation	2

Régression pénalisée avec STATISTICA

Analyse des corrélations

Correlations (Diabetes.(validation).sta in Diabetes.stw)
Marked correlations are significant at p < ,05000
N=442 (Casewise deletion of missing data)

Variable	Age	Gender	BMI	BP	**Total Cholesterol**	LDL	HDL	TCH	LTG	Glucose	Y1_continue
Age	1,000000	0,173737	0,185085	0,335428	0,260061	0,219243	-0,075181	0,203841	0,270774	0,301731	0,187889
Gender	0,173737	1,000000	0,088161	0,241010	0,035277	0,142637	-0,379090	0,332115	0,149916	0,208133	0,043062
BMI	0,185085	0,088161	1,000000	0,395411	0,249777	0,261170	-0,366811	0,413807	0,446157	0,388680	0,586450
BP	0,335428	0,241010	0,395411	1,000000	0,242464	0,185548	-0,178762	0,257650	0,393480	0,390430	0,441482
Total Cholesterol	0,260061	0,035277	0,249777	0,242464	1,000000	0,896663	0,051519	0,542207	0,515503	0,325717	0,212022
LDL	0,219243	0,142637	0,261170	0,185548	**0,896663**	1,000000	-0,196455	0,659817	0,318357	0,290600	0,174054
HDL	-0,075181	-0,379090	-0,366811	-0,178762	0,051519	-0,196455	1,000000	-0,738493	-0,398577	-0,273697	-0,394789
TCH	0,203841	0,332115	0,413807	0,257650	0,542207	0,659817	-0,738493	1,000000	0,617859	0,417212	0,430453
LTG	0,270774	0,149916	0,446157	0,393480	0,515503	0,318357	-0,398577	0,617859	1,000000	0,464669	0,565883
Glucose	0,301731	0,208133	0,388680	0,390430	0,325717	0,290600	-0,273697	0,417212	0,464669	1,000000	0,382483
Y1_continue	0,187889	0,043062	0,586450	0,441482	0,212022	0,174054	-0,394789	0,430453	0,565883	0,382483	1,000000

Analyse de la multi colinéarité

Eigenvalues of correlation matrix, and related statistics (Diabetes.(validation).sta in Diab
Active variables only

Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %	**IC =4,024211/Eigenvalue;**
1	4,024211	40,24211	4,02421	40,2421	**1,00**
2	1,492320	14,92320	5,51653	55,1653	**2,70**
3	1,205966	12,05966	6,72250	67,2250	**3,34**
4	0,955476	9,55476	7,67797	76,7797	**4,21**
5	0,662181	6,62181	8,34015	83,4015	**6,08**
6	0,602717	6,02717	8,94287	89,4287	**6,68**
7	0,536566	5,36566	9,47944	94,7944	**7,50**
8	0,433682	4,33682	9,91312	99,1312	**9,28**
9	0,078320	0,78320	9,99144	99,9144	**51,38**
10	0,008561	0,08561	10,00000	100,0000	**470,08**

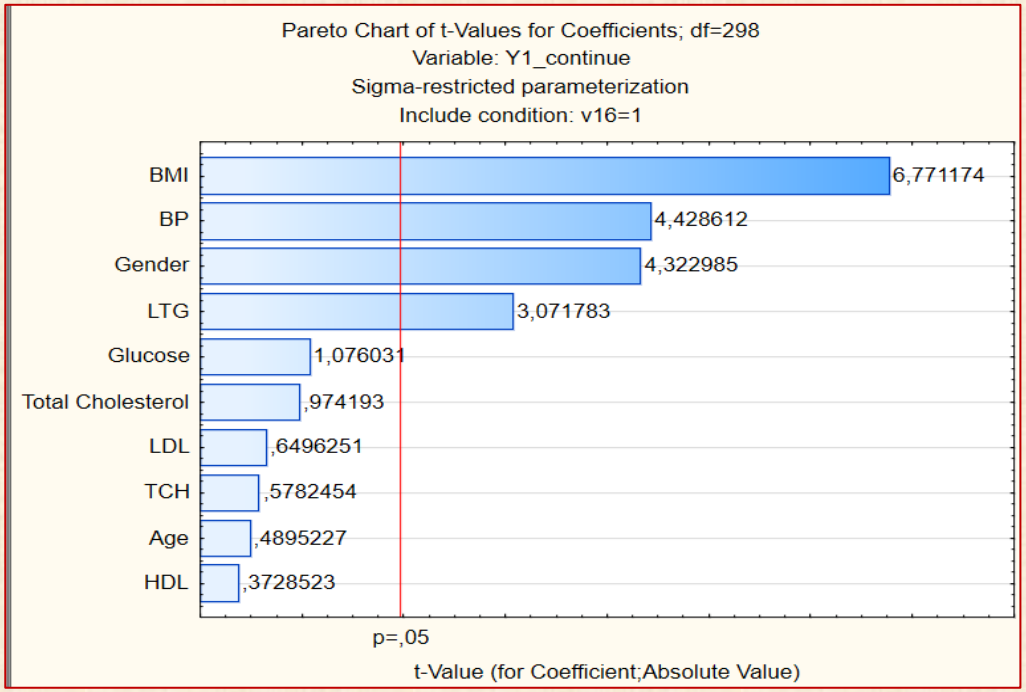
Diagnostic : multi colinéarité présente

Régression pénalisée avec STATISTICA

Modèle de régression ordinaire sur l'ensemble d'entrainement

Parameter Estimates (Diabetes(training).sta in Diabetes.stw)
Sigma-restricted parameterization
Include condition: v16=1

Effect	Level of Effect	Column	Y1_continue Param.	Y1_continue Std.Err	Y1_continue t	Y1_continue p
Intercept		1	-322,238	76,87505	-4,19171	0,000037
Age		2	-0,122	0,24962	-0,48952	0,624832
BMI		3	5,728	0,84595	6,77117	0,000000
BP		4	1,181	0,26665	4,42861	0,000013
Total Cholesterol		5	-0,666	0,68403	-0,97419	0,330751
LDL		6	0,422	0,65012	0,64963	0,516435
HDL		7	-0,342	0,91855	-0,37285	0,709523
LTG		8	57,017	18,56156	3,07178	0,002324
Glucose		9	0,358	0,33290	1,07603	0,282784
TCH		10	4,021	6,95344	0,57825	0,563535
Gender	1	11	14,728	3,40689	4,32299	0,000021



remarque

modèle pas satisfaisant

pourquoi ?

Régression pénalisée avec STATISTICA

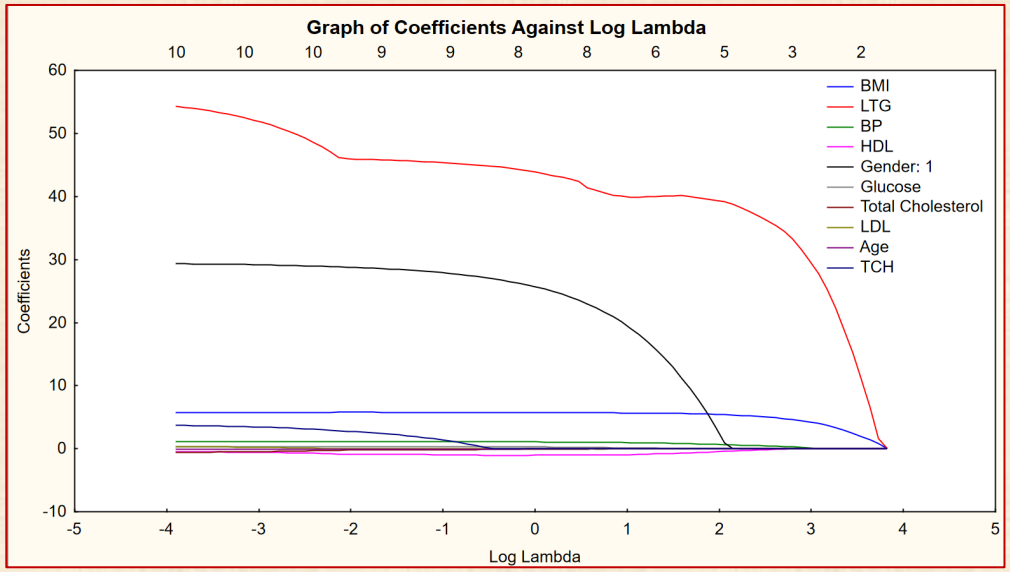
Modèle de régression LASSO sur l'ensemble d'entrainement
Matrice des coefficients

Y1_continue - Coefficient Matrix (Diabetes(training).sta in Diabetes.stw)
Linear Regression

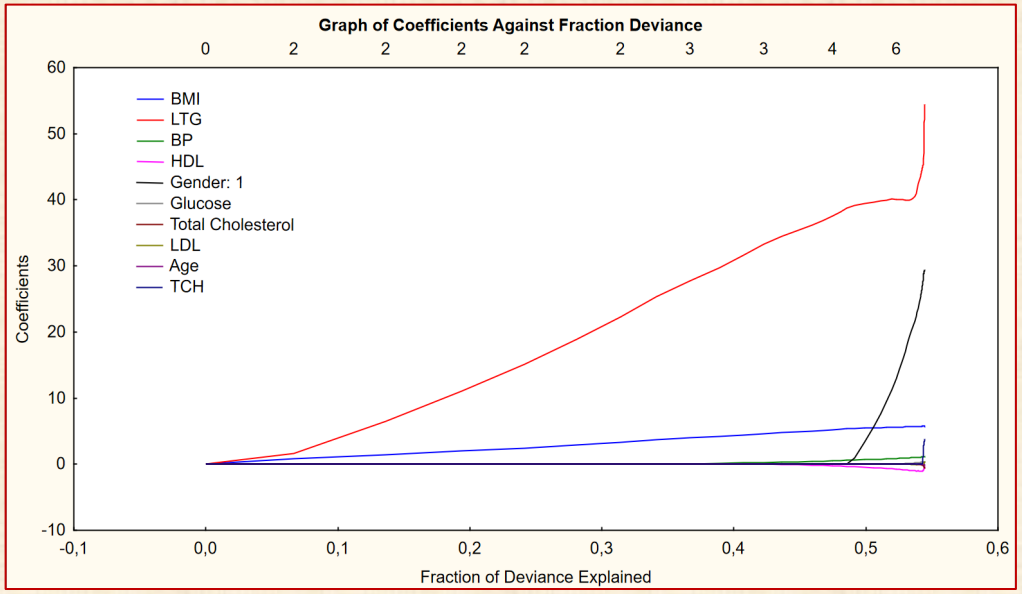
Include condition: v16=1

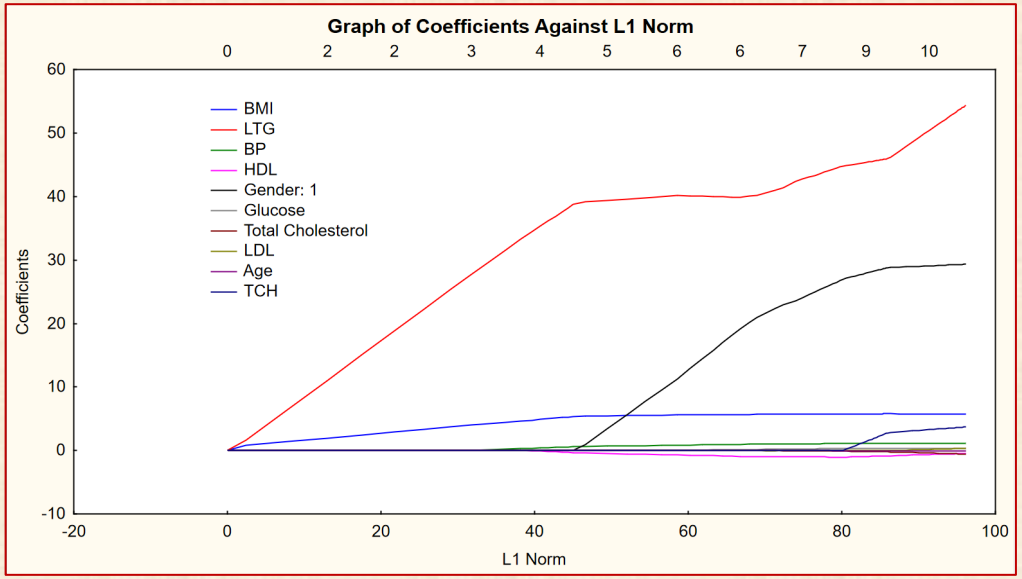
	Lambda	Df	Intercept	Age	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose	Gender 1	Gender 2
1	45,77	0	152,16											
2	41,70	2	122,52		0,85						1,61			
3	38,00	2	84,21		1,44						6,54			
4	34,62	2	49,31		1,98						11,04			
5	31,55	2	17,51		2,47						15,14			
6	28,75	2	-11,47		2,92						18,87			
7	26,19	2	-37,88		3,33						22,27			
8	23,86	2	-61,94		3,70						25,37			
9	21,74	3	-85,91		3,98	0,06					27,78			
10	19,81	3	-108,70		4,22	0,14					29,77			
11	18,05	3	-129,46		4,43	0,21					31,60			
12	16,45	3	-148,37		4,63	0,27					33,26			
13	14,99	4	-162,02		4,78	0,33			-0,04		34,50			
14	13,66	4	-171,28		4,90	0,39			-0,10		35,38			
15	12,44	4	-179,73		5,01	0,44			-0,16		36,19			
16	11,34	4	-187,43		5,11	0,48			-0,21		36,92			
17	10,33	4	-194,45		5,20	0,53			-0,26		37,59			
18	9,41	4	-200,84		5,29	0,56			-0,30		38,20			
19	8,58	4	-206,66		5,36	0,60			-0,35		38,76			
20	7,81	5	-211,83		5,42	0,64			-0,40		39,18		0,94	

79	0,03	10	-322,45	-0,12	5,75	1,18	-0,52	0,28	-0,51	3,60	53,33	0,36	29,27	
80	0,03	10	-323,36	-0,12	5,75	1,18	-0,53	0,29	-0,50	3,63	53,57	0,36	29,29	
continue - Parameter Estimates (Diabetes(training).sta in Diabetes.stw)}							-0,53	0,30	-0,49	3,65	53,78	0,36	29,30	
82	0,02	10	-324,97	-0,12	5,74	1,18	-0,54	0,31	-0,48	3,67	53,97	0,36	29,31	
83	0,02	10	-325,66	-0,12	5,74	1,18	-0,55	0,31	-0,48	3,69	54,15	0,36	29,32	
84	0,02	10	-326,34	-0,12	5,74	1,18	-0,56	0,32	-0,47	3,71	54,32	0,36	29,34	



Régression pénalisée avec STATISTICA





Y1_continue - Parameter Estimates
Linear Regression
Model Lambda = 0,020279; %Dev = 0,544347
Include condition: v16=1

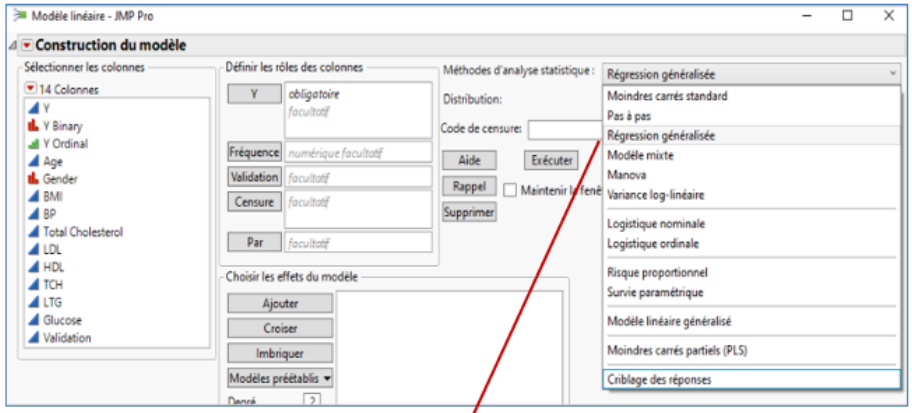
1 effect	2 Estimates	Régression ordinaire Y1_continue Param.
Intercept	-326,3379	-322,238
Age	-0,1178	-0,122
BMI	5,7430	5,728
BP	1,1771	1,181
Total Cholesterol	-0,5560	-0,666
LDL	0,3204	0,422
HDL	-0,4683	-0,342
TCH	3,7081	57,017
LTG	54,3212	0,358
Glucose	0,3596	4,021
Gender 1	29,3353	14,728

comparaison modèles
ordinaire, LASSO, RIDGE, ...

sur l'ensemble validation

Régression pénalisée avec JMP

Modèle linéaire ..... Régression généralisée



JMP PRO Overview of the Generalized Regression Personality

The Generalized Regression personality features regularized, or penalized, regression techniques. Such techniques attempt to fit better models by shrinking the model coefficients toward zero. The resulting estimates are biased. This increase in bias can result in decreased prediction variance, thus lowering overall prediction error compared to non-penalized models. Two of these techniques, the Elastic Net and the Lasso, include variable selection as part of the modeling procedure.

Modeling techniques such as the Elastic Net and the Lasso are particularly useful for large data sets, where collinearity is typically a problem. In addition, modern data sets often include more variables than observations. This situation is sometimes referred to as the $p > n$ problem, where n is the number of observations and p is the number of predictors. Such data sets require variable selection if traditional modeling techniques are to be used.

The Elastic Net and Lasso can also be used for small data sets with little correlation, including designed experiments. They can be used to build predictive models or to select variables for model reduction or for future study.

The personality provides the following classes of modeling techniques:

- Maximum Likelihood
- Step-Based Estimation
- Penalized Regression

The Elastic Net and Lasso are relatively recent techniques (Tibshirani 1996; Zou and Hastie 2005). Both techniques penalize the size of the model coefficients, resulting in a continuous shrinkage. The amount of shrinkage is determined by a *tuning parameter*. An optimal level of shrinkage is determined by one of several validation methods. Both techniques have the ability to shrink coefficients to zero. In this way, variable selection is built into the modeling procedure. The Elastic Net model subsumes both the Lasso and ridge regression as special cases. See "Statistical Details for Estimation Methods" on page 333.

NOTATION

- $\sum_{j=1}^{p} |\beta_j|$ is the l_1 penalty
- $\sum_{j=1}^{p} \beta_j^2$ is the l_2 penalty
- λ is the tuning parameter
- α is a parameter that determines the mix of the l_1 and l_2 penalties
- N is the number of rows
- p is the number of variables

RIDGE

An l_2 penalty is applied to the regression coefficients during ridge regression. Ridge regression coefficient estimates are given by the following:

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

where $\sum_{j=1}^{p}\beta_j^2$ is the l_2 penalty, λ is the tuning parameter, N is the number of rows, and p is the number of variables.

JMP PRO Dantzig Selector

An l_∞ penalty is applied to the regression coefficients during Dantzig Selector. Coefficient estimates for the Dantzig Selector satisfy the following criterion:

$$\min_{\beta}\left\|X^T(y - X\beta)\right\|_\infty \quad \text{subject to } \|\beta\|_1 \le t$$

where $\|v\|_\infty$ denotes the l_∞ norm, which is the maximum absolute value of the components of the vector v.

JMP PRO Lasso Regression

An l_1 penalty is applied to the regression coefficients during Lasso. Coefficient estimates for the Lasso are given by the following:

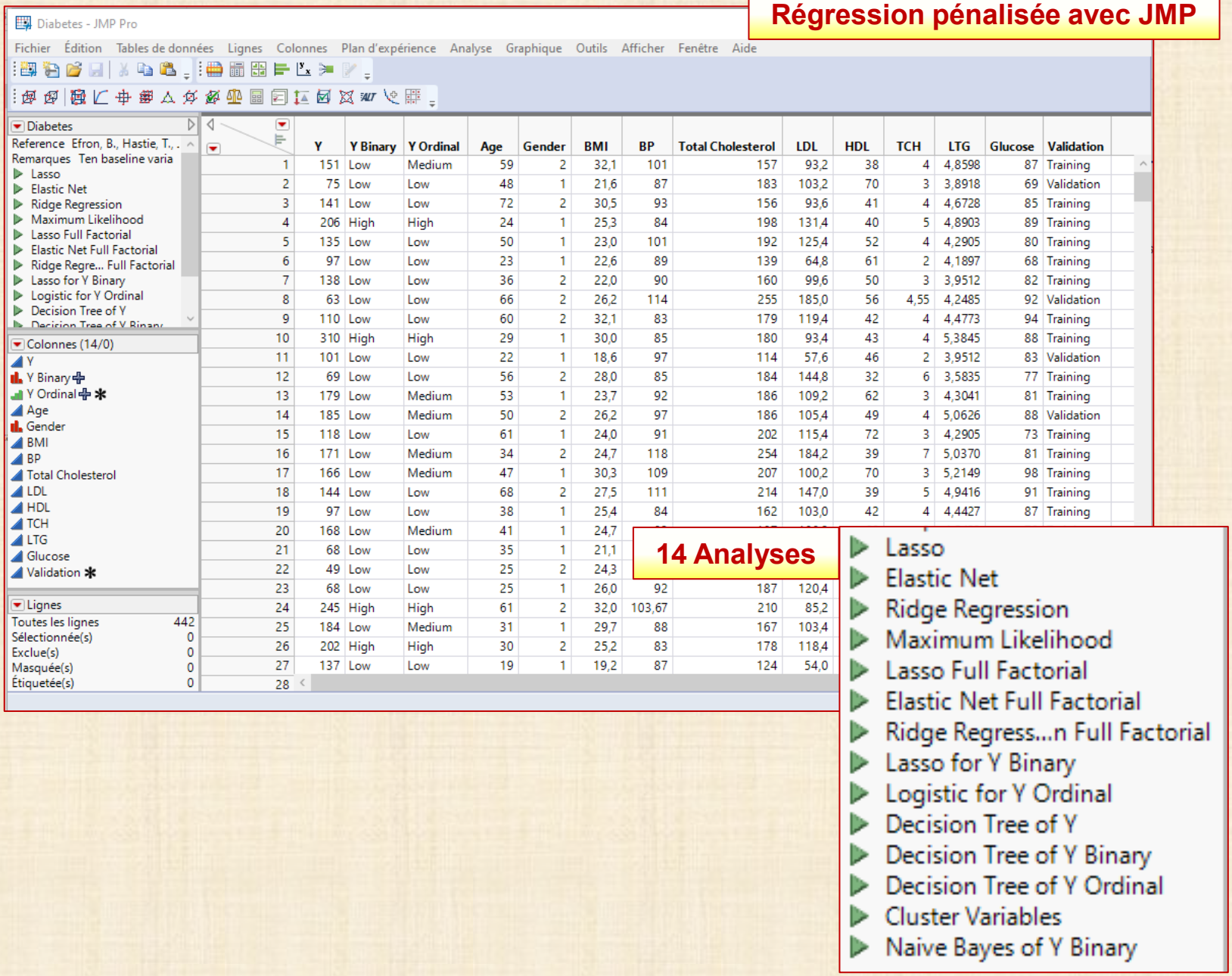
$$\hat{\beta}^{lasso} = \text{argmin}_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N} -LogLikelihood(\beta;y_i) + \lambda\sum_{j=1}^{p}|\beta_j|\right\},$$

JMP PRO Elastic Net

The Elastic Net combines both l_1 and l_2 penalties. Coefficient estimates for the Elastic Net are given by the following:

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

Régression pénalisée avec JMP



	Y	Y Binary	Y Ordinal	Age	Gender	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose	Validation
1	151	Low	Medium	59	2	32,1	101	157	93,2	38	4	4,8598	87	Training
2	75	Low	Low	48	1	21,6	87	183	103,2	70	3	3,8918	69	Validation
3	141	Low	Low	72	2	30,5	93	156	93,6	41	4	4,6728	85	Training
4	206	High	High	24	1	25,3	84	198	131,4	40	5	4,8903	89	Training
5	135	Low	Low	50	1	23,0	101	192	125,4	52	4	4,2905	80	Training
6	97	Low	Low	23	1	22,6	89	139	64,8	61	2	4,1897	68	Training
7	138	Low	Low	36	2	22,0	90	160	99,6	50	3	3,9512	82	Training
8	63	Low	Low	66	2	26,2	114	255	185,0	56	4,55	4,2485	92	Validation
9	110	Low	Low	60	2	32,1	83	179	119,4	42	4	4,4773	94	Training
10	310	High	High	29	1	30,0	85	180	93,4	43	4	5,3845	88	Training
11	101	Low	Low	22	1	18,6	97	114	57,6	46	2	3,9512	83	Validation
12	69	Low	Low	56	2	28,0	85	184	144,8	32	6	3,5835	77	Training
13	179	Low	Medium	53	1	23,7	92	186	109,2	62	3	4,3041	81	Training
14	185	Low	Medium	50	2	26,2	97	186	105,4	49	4	5,0626	88	Validation
15	118	Low	Low	61	1	24,0	91	202	115,4	72	3	4,2905	73	Training
16	171	Low	Medium	34	2	24,7	118	254	184,2	39	7	5,0370	81	Training
17	166	Low	Medium	47	1	30,3	109	207	100,2	70	3	5,2149	98	Training
18	144	Low	Low	68	2	27,5	111	214	147,0	39	5	4,9416	91	Training
19	97	Low	Low	38	1	25,4	84	162	103,0	42	4	4,4427	87	Training
20	168	Low	Medium	41	1	24,7								
21	68	Low	Low	35	1	21,1								
22	49	Low	Low	25	2	24,3								
23	68	Low	Low	25	1	26,0	92	187	120,4					
24	245	High	High	61	2	32,0	103,67	210	85,2					
25	184	Low	Medium	31	1	29,7	88	167	103,4					
26	202	High	High	30	2	25,2	83	178	118,4					
27	137	Low	Low	19	1	19,2	87	124	54,0					
28														

14 Analyses

- Lasso
- Elastic Net
- Ridge Regression
- Maximum Likelihood
- Lasso Full Factorial
- Elastic Net Full Factorial
- Ridge Regress...n Full Factorial
- Lasso for Y Binary
- Logistic for Y Ordinal
- Decision Tree of Y
- Decision Tree of Y Binary
- Decision Tree of Y Ordinal
- Cluster Variables
- Naive Bayes of Y Binary

Régression pénalisée avec JMP

Régression ORDINAIRE : data = diabetes.jmp

Moindres carrés standard avec Colonne de validation

Résumé du modèle

Réponse	Y
Distribution	Normale
Méthode d'estimation	Moindres carrés standard
Méthode de validation	Colonne de validation
Moyenne Lien du modèle	Unité
Échelle Lien du modèle	Unité

Mesure	Apprentissage	Validation
Nombre de lignes	309	133
Somme des fréquences	309	133
-Log-vraisemblance	1659,3426	729,09775
Nombre de paramètres	12	12
BIC	3387,4853	1516,8797
AICc	3343,7392	1484,7955
R carré	0,544387	0,4290597
R carré ajusté	0,529098	.
Racine de l'erreur quadratique moyenne (RMSE)	51,992643	57,516988

Estimation des paramètres pour les régresseurs d'origine

Terme	Estimation	Erreur standard	Khi deux de Wald	Prob. > Khi-deux	Limite de confiance inférieure (pour 95% de confiance)	Limite de confiance supérieure (pour 95% de confiance)
Constante	-336,9656	76,984501	19,158635	<,0001*	-487,8525	-186,0788
Age	-0,122192	0,2496154	0,2396325	0,6245	-0,61143	0,3670448
Gender[1-2]	29,455858	6,8137767	18,688204	<,0001*	16,101101	42,810615
BMI	5,7280953	0,8459531	45,848797	<,0001*	4,0700578	7,3861328
BP	1,1808708	0,2666458	19,612605	<,0001*	0,6582546	1,703487
Total Cholesterol	-0,666377	0,6840299	0,949052	0,3300	-2,007051	0,6742968
LDL	0,4223324	0,6501172	0,4220127	0,5159	-0,851874	1,6965386
HDL	-0,342483	0,9185475	0,1390188	0,7093	-2,142803	1,4578375
TCH	4,020793	6,9534367	0,3343678	0,5631	-9,607692	17,649279
LTG	57,017088	18,561559	9,4358535	0,0021*	20,637102	93,397074
Glucose	0,3582082	0,3328978	1,1578418	0,2819	-0,294259	1,0106758

Tests des effets

Source	Nparm	Degrés de liberté	Somme des carrés	Rapport F	Prob. > F
BMI	1	1	128515,04	45,848797	<,0001*
BP	1	1	54974,5	19,612605	<,0001*
Gender	1	1	52383,386	18,688204	<,0001*
LTG	1	1	26448,875	9,4358535	0,0023*
Glucose	1	1	3245,4522	1,1578418	0,2828
Total Cholesterol	1	1	2660,2107	0,949052	0,3308
LDL	1	1	1182,9097	0,4220127	0,5164
TCH	1	1	937,23918	0,3343678	0,5635
Age	1	1	671,69432	0,2396325	0,6248
HDL	1	1	389,67243	0,1390188	0,7095

Régression pénalisée avec JMP

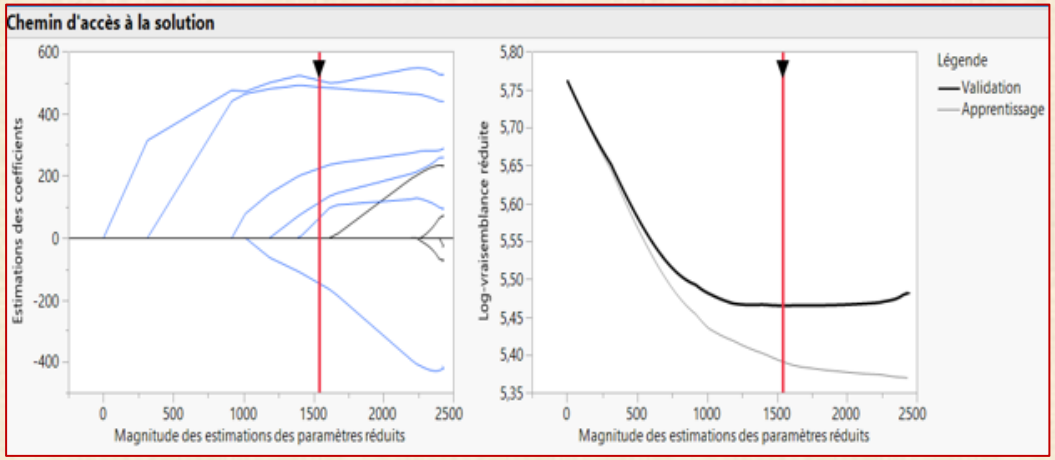
Régression LASSO

Lasso avec Retenue Validation

Résumé du modèle

Réponse	Y
Distribution	Normale
Méthode d'estimation	Lasso
Méthode de validation	Retenue
Moyenne Lien du modèle	Unité
Échelle Lien du modèle	Unité

Mesure	Apprentissage	Validation
Nombre de lignes	309	133
Somme des fréquences	309	133
-Log-vraisemblance	1674,1142	719,15872
Nombre de paramètres	9	9
BIC	3399,8285	1482,3306
AICc	3366,8304	1457,7808
R carré	0,5301083	0,4334666
Racine de l'erreur quadratique moyenne (RMSE)	54,538497	53,955255
Pénalité lambda	74,863327	.



Estimation des paramètres pour les régresseurs d'origine

Terme	Estimation	Erreur standard	Khi deux de Wald	Prob. > Khi-deux	Limit
Constante	-329,1998	27,480478	143,50618	<,0001*	
Age	0	0	0	1,0000	
Gender[1-2]	12,894716	6,7198929	3,6821279	0,0550	
BMI	6,339605	0,8632354	53,934404	<,0001*	
BP	0,9171642	0,2529557	13,146354	0,0003*	
Total Cholesterol	-0,235943	0,1141011	4,2759776	0,0387*	
LDL	0	0	0	1,0000	
HDL	0	0	0	1,0000	
TCH	2,7654017	3,8000222	0,5295954	0,4668	
LTG	55,179547	8,4026575	43,124388	<,0001*	
Glucose	0	0	0	1,0000	

Tests des effets

Source	Nparm	Degrés de liberté	Khi deux de Wald	Prob. > Khi-deux	
BMI	1	1	53,934404	<,0001*	
LTG	1	1	43,124388	<,0001*	
BP	1	1	13,146354	0,0003*	
Total Cholesterol	1	1	4,2759776	0,0387*	
Gender	1	1	3,6821279	0,0550	
TCH	1	1	0,5295954	0,4668	
Age	1	0	0	1,0000	Supprimé
LDL	1	0	0	1,0000	Supprimé
HDL	1	0	0	1,0000	Supprimé
Glucose	1	0	0	1,0000	Supprimé

Univariate Tests of Significance for Y1_continue (Diabetes(traini
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 52,9
Include condition: v16=1

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	49250,2	1	49250,2	17,57041	0,000037
Age	671,7	1	671,7	0,23963	0,624832
BMI	128515,0	1	128515,0	45,84880	0,000000
BP	54974,5	1	54974,5	19,61260	0,000013
Total Cholesterol	2660,2	1	2660,2	0,94905	0,330751
LDL	1182,9	1	1182,9	0,42201	0,516435
HDL	389,7	1	389,7	0,13902	0,709523
LTG	26448,9	1	26448,9	9,43585	0,002324
Glucose	3245,5	1	3245,5	1,15784	0,282784
TCH	937,2	1	937,2	0,33437	0,563535
Gender	52383,4	1	52383,4	18,68820	0,000021
Error	835299,6	298	2803,0		

MODÈLES : linéaires - non linéaires - linéarisables

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

Modèle linéaire dans les paramètres à estimer

$$y_i = f(x_{i1}, x_{i2},\ldots, X_{id} ; \beta_0, \beta_1,\ldots, \beta_k) + \varepsilon_i \qquad i = 1, 2,\ldots, n$$

est linéaire dans les paramètres β_j si

$$f(x_{i1}, x_{i2},\ldots, X_{id} ; \beta_0, \beta_1,\ldots, \beta_1) = \Sigma\, \beta_j\, g_j(x_{i1}, x_{i2},\ldots, x_{id}) \qquad (1)$$

$g_j(x_{i1}, x_{i2},\ldots, X_{id})$ fonctions connues sans paramètres inconnus

Si l'équation (1) n'est pas vérifiée, le modèle est *non-linéaire* dans les β

autre vérification de (1) : calcul les dérivées par rapport aux β

donne un système d'équations linéaires dans les β

Remarque

β_j peuvent être remplacés par de nouveaux paramètres $\gamma_k = h_k(\beta_j)$ dans l'équation (1)

$$f(x_{i1}, x_{i2},\ldots, X_{id} ; \beta_0, \beta_1,\ldots, \beta_1) = \Sigma\, h_k(\beta_j)\, g_j(x_{i1}, x_{i2},\ldots, X_{id}) = \Sigma\, \gamma_k\, g_j(x_{i1}, x_{i2},\ldots, x_{id}) \qquad (2)$$

système (2) est linéaire dans les paramètres γ_k nombre de γ_k = nombre de β_j

modèle non-linéaire

intrinsèquement linéaires (« linéarisables »)

par transformation sur Y et /ou X

exemple : fonction logistique avec 2 variables

$$y_i = \exp(\beta_0+\beta_1*x_{i1}+\beta_2*x_{i2})/[1+\exp(\beta_0+\beta_1*x_{i1}+\beta_2*x_{i2})]$$

$$\ln[y_i / (1-y_i)] = \beta_0 + \beta_1*x_i + + \beta_2*x_{i2}$$

intrinsèquement non-linéaire

exemple $y_i = \beta_0 + \beta_1(1 - \exp(\beta_2*x_i) + \varepsilon_i$

Modèles linéaires généralisés: generalized linear model
Utilisation de GLZ de Statistica

régression ordinaire classique : Y continue et normale

$$y_i = f(x_{i1}, x_{i2},\ldots, X_{id} ; \beta_0, \beta_1,\ldots, \beta_k) + \varepsilon_i \quad \varepsilon \sim N(0, \sigma^2) \quad i = 1, 2,\ldots,n$$

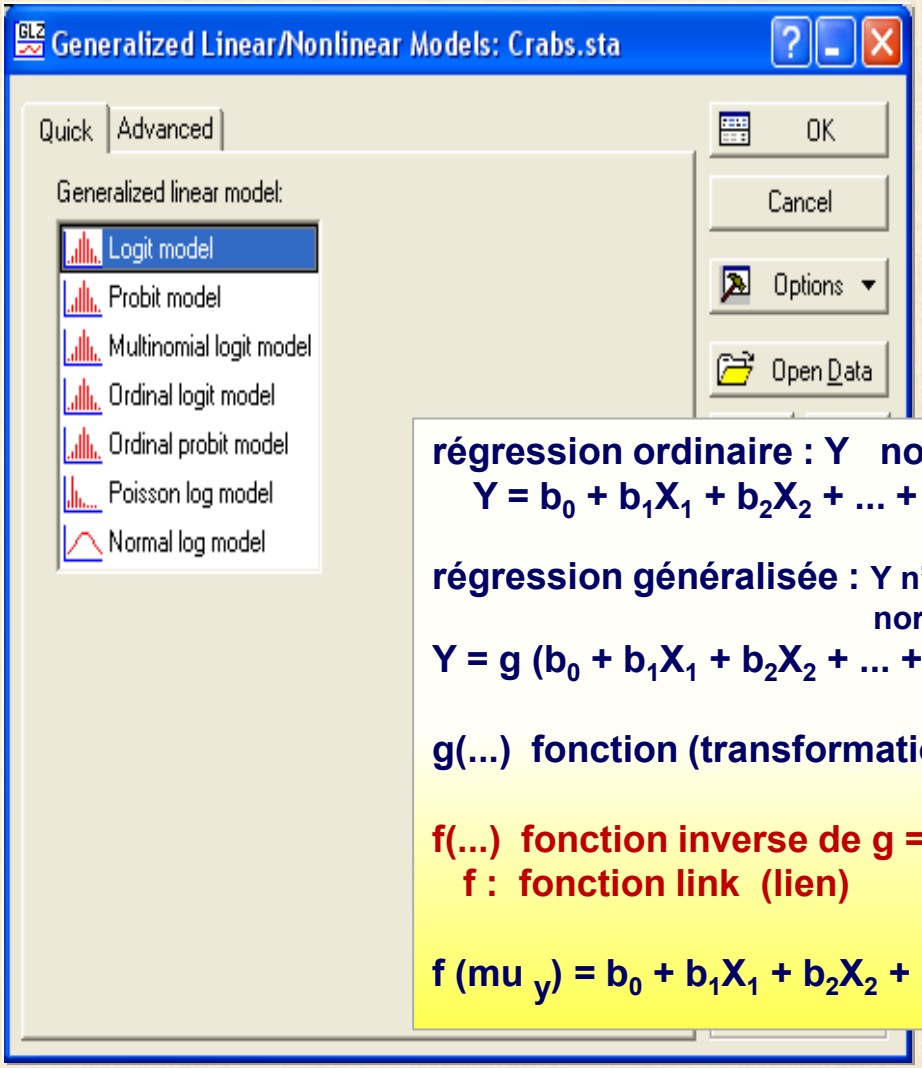
Y continue et normale $Y \sim N(f(x_{i1}, x_{i2},\ldots, X_{id} ; \beta_0, \beta_1,\ldots, \beta_k) ; \sigma^2)$

Autres cas de Y

y = 0 ou 1	: régression logistique
y = variable catégorique k modalités	: régression multinomiale linéaire généralisée
y = comptage 0, 1, 2,…	: régression type Poisson linéaire généralisée
y suit loi non normale	: exemple loi gamma linéaire généralisée

Utilisation de *STATISTICA* : régression linéaire généralisée

Statistics…. Advanced Linear/Nonlinear Models…
… Generalized Linear/Nonlinear Models (GLZ)



régression ordinaire : Y normale

$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e$

régression généralisée : Y n'est pas normale

$Y = g\,(b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k) + e$

g(...) fonction (transformation)

f(...) fonction inverse de g = link = g^{-1}
f : fonction link (lien)

$f(mu_y) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

Y	f (fonction link)	
normale	f(z) = z	identité
binomiale	f(z) = ln(z/(1-z))	logit
Poissson	f(z) = ln(z)	
	f(z) = invnorm(z)	probit
…….	f(z) = ln(-ln(1-z))	
	$f(z) = z^a$	
Multinomiale	……f(z) = …	

Estimation des paramètres et tests

paramètres sont estimés par la méthode de vraisemblance maximale:
nécessite la résolution d'équations non linéaires par méthode itérative

- si on peut linéariser le modèle et résoudre le modèle linéaire équivalent:
  mais pas recommandé en général – présence de zéros…
- estimations, écarts types d'estimation, intervalles confiance.
  le test de $H_0 : \beta = 0$ s'appelle la statistique de Wald
- on peut introduire plusieurs variables explicatives continues et des variables catégoriques et des produits (interactions)
  cas du modèle logistique

$$\pi_i (x_1,..,x_p) = \exp(\beta_0+\beta_1*x_{i1}+\beta_2*x_{i2}+ \dots) / [1+\exp(\beta_0+\beta_1*x_{i1}+\beta_2*x_{i2} \dots)]$$

$$\ln[\pi_i / (1- \pi_i)] = \beta_0 + \beta_1*x_i + + \beta_2*x_{i2} \dots + \beta_p*x_{ip}$$

- on peut aussi avoir une variable de réponse avec 3 ou plus catégories:
  régression multinomiale

régression logistique avec plusieurs facteurs continus et catégoriques et variable réponse Y (0,1) avec comptage n

Exemple 1 – voir chapitre REG SIMPLE

Étude de survie cancer de 772 sujets
data = cancer.sta

tableau des données : tableau d'effectifs (nombre) croisant
3 facteurs X (ville, âge, type tumeur) et Y_Survie (oui, non)
tableau initial de 764 lignes (patients) provenant de
3 villes (X1) + 3 catégories âge (X2) + 4 catégories de tumeurs (X3) donnant 3x3x4 = 36 sous-groupes

1. ville (X1) : Tokyo, Boston MA, Glamorgan
2. cat âge (X2) : 50 et moins, 50 à 69, 70 et plus
3. type tumeurs (X3) : 4 types
   MIN_MAL ……Minimum inflammation, malignant
   MIN_BEGN ….Mimimum inflammation, benign
   GRT_MAL ……Greater inflammation, malignant
   GRT_BEGN …Greater inflammation, benign

Y_survie (oui ou non) n = nombre dans chaque catégorie

id	ville	age	type tumeur	Survie Y	nombre n
1	TOKYO	50 et moins	MIN_MAL	non	9
2	TOKYO	50 et moins	MIN_BEGN	non	7
3	TOKYO	50 et moins	GRT_MAL	non	4
4	BOSTON	50 et moins	GRT_BEGN	non	3
5	TOKYO	50 et moins	MIN_MAL	oui	26
.	.	.	.	.	.
72	GLAMORGN	70 et plus	GRT_BEGN	oui	1

Exemple 2 : crabs satellite - data = crabs.sta

crab satellites
by female's color, spine condition, width, and weight

4 facteurs X : 2 catégoriques (X1, X2) + 2 continus (X3, X4)

1. Color (X1) : Color of the crab avec 4 categories
2. Spine (X2) Spine condition for the crab avec 3 categories)
3. Width (X3) : Carapace width of the female crab (cm)
   CATWIDTH : recodage de WIDTH en 8 categories 22,72, 23,75,…28,75
4. Weight (X4) : Weight of the crab (kg)

Y1 Satellts
number of satellites; i.e. the number of male crabs attached to the female's nest, in addition to the single male crab attached to each nest.

Y2 : indicator variable
= 0 if Y1 = 0
= 1 if Y1 > 0

Purpose of the study : determine the factors that predict whether or not additional satellites are attached to a female horseshoe crab's nest. search for a logit model with the minimum number of factors to predict the binary Y.

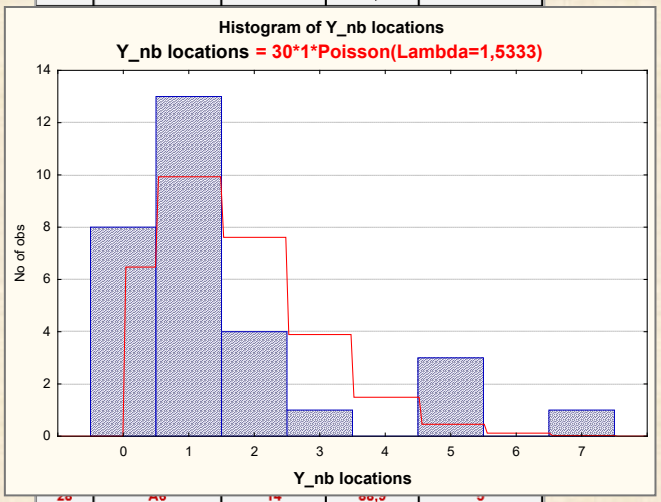
id	COLOR	SPINE	WIDTH	WEIGHT	CATWIDTH	SATELLTS Y1	Y2
1	medium	bothworn	28,3	3,05	28,75	8	1
2	darkmed	bothworn	22,5	1,55	22,75	0	0
3	lightmed	bothgood	26,0	2,30	25,75	9	1
.	.	.	.	.	.	.	.
173	medium	oneworn	24,5	2,00	24,75	0	0

Exemple 3 : hélicoptères

Y variable comptage : 0, 1, 2,…
Y ~ Poisson

Aircraft damage - Montgomery 4ed. p 450
Y_ nb locations dommage (0, 1, 2,..)
X1_type hélicoptère
A4=Douglas Skyhawk
A6=Grumman Intruder
X2_load (tons)
X3_exp (months) n = 30

id	x1_helicop	x2_load	x3_exp	Y_nb locations
1	A4	4	91,5	0
2	A4	4	84,0	1
3	A4	4	76,5	0
4	A4	5	69,0	0
5	A4	5	61,5	0
6	A4	5	80,0	0
7	A4	6	72,5	1
8	A4	6	65,0	0
9	A4	6	57,5	0
29	A6	14	73,7	5
30	A6	14	57,8	7



Exemple 4 Plan expérimental 3**3 = 27 essais

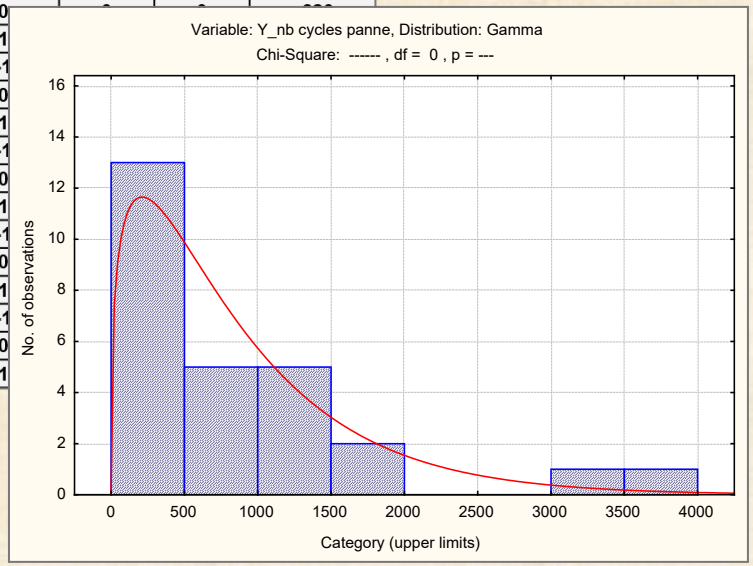
3 facteurs X1 X2 X3 à 3 modalités
réponse : Y_nb cycles_panne

Montgomery 4 th ed p. 458
Données plan expérimental complet
3 facteurs à 3 modalités.

ID	x1_long	x2_amp	x3_load	Y_nb cycles panne
1	-1	-1	-1	674
2	0	-1	-1	1414
3	1	-1	-1	3636
4	-1	0	-1	338
5	0	0	-1	1022
6	1	0	-1	1568
7	-1	1	-1	170
8	0	1	-1	442
9	1	1	-1	1140
10	-1	-1	0	370
11	0	-1	0	1198
12	1	-1	0	3184
13	-1	0	0	266
14	0			
15	1			
16	-1			
17	0			
18	1			
19	-1			
20	0			
21	1			
22	-1			
23	0			
24	1			
25	-1			
26	0			
27	1			

Y non normale

Y ~ Gamma



Exemple 2: crab satellites analyse

crab satellites

by female's color, spine condition, width, and weight

4 facteurs X : 2 catégoriques (X1, X2) + 2 continus (X3, X4)

1. Color (X1) : Color of the crab avec 4 categories
2. Spine (X2) Spine condition for the crab avec 3 categories)
3. Width (X3) : Carapace width of the female crab (cm)
   CATWIDTH : recodage de WIDTH en 8 categories 22,72, 23,75,...28,75
4. Weight (X4) : Weight of the crab (kg)

Y1 Satellts
number of satellites; i.e. the number of male crabs attached to the female's nest, in addition to the single male crab attached to each nest.

Y2 : indicator variable
= 0 if Y1 = 0
= 1 if Y1 > 0

Purpose of the study : determine the factors that predict whether or not additional satellites are attached to a female horseshoe crab's nest. search for a logit model with the minimum number of factors to predict the binary Y.

id	COLOR	SPINE	WIDTH	WEIGHT	CATWIDTH	SATELLTS Y1	Y2
1	medium	bothworn	28,3	3,05	28,75	8	1
2	darkmed	bothworn	22,5	1,55	22,75	0	0
3	lightmed	bothgood	26,0	2,30	25,75	9	1
.	.	.	.	.	.	.	.
173	medium	oneworn	24,5	2,00	24,75	0	0

Modèle 1: logit Y = f(SPINE , WIDTH, WEIGHT)

	Col	Variable	Level of Variable	Versus Level
Intercep	1			.
WIDTH	2	WIDTH		
WEIGHT	3	WEIGHT		
SPINE	4	SPINE	bothgood	bothworn
SPINE	5	SPINE	oneworn	bothworn

Y - Parameter estimates
Distribution : BINOMIAL, Link function: LOGIT
Modeled probability that Y = 0

	Level of - Effect	Column	Estimate	Standard - Error	Wald - Stat.	p	Lower CL - 95, %	Upper CL - 95, %
Intercept		1	9,210285	3,612313	6,500931	0,010782	2,13028	16,29029
WIDTH		2	-0,297166	0,186462	2,539911	0,111001	-0,66262	0,06829
WEIGHT		3	-0,857189	0,677029	1,603021	0,205476	-2,	
SPINE	bothgood	4	-0,004547	0,339528	0,000179	0,989316	-0,	
SPINE	oneworn	5	0,085980	0,409746	0,044031	0,833795	-0,71711	0,88907

WIDTH WEIGHT SPINE non significatifs

Modèle 2: logit Y = f(SPINE , WIDTH)

Y - Parameter estimates
Distribution : BINOMIAL, Link function:
LOGIT Modeled probability that Y = 0

	Level of - Effect	Column	Estimate	Standard - Error	Wald - Stat.	p
Intercept		1	12,32830	2,691917	20,97407	0,000005
WIDTH		2	-0,49531	0,104796	22,33902	0,000002
SPINE	bothgood	3	0,00069	0,338837	0,00000	0,998379
SPINE	oneworn	4	0,04359	0,408801	0,01137	0,915088

SPINE pas significatif

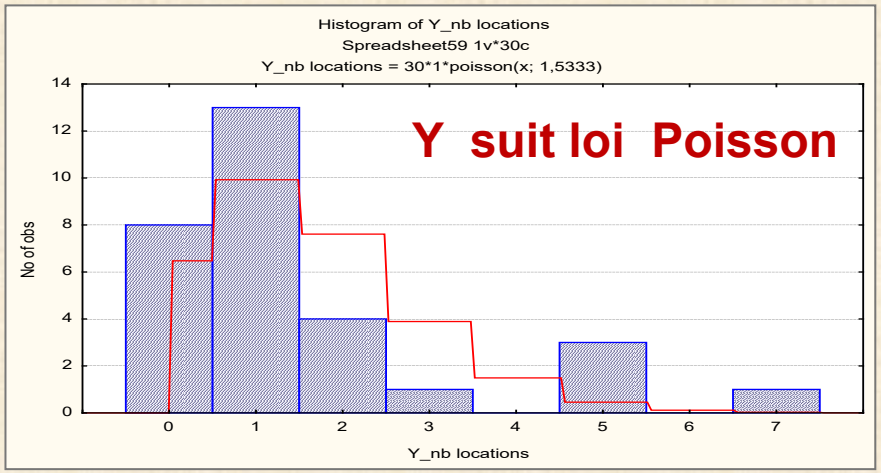
Modèle 3: logit Y = f(WIDTH) final

Classification of cases

Observed	Predicted 0	Predicted 1	Percent - correct
0	27	35	43,54
1	16	95	85,59

Exemple 3 : Y ~ Poisson Y variable comptage : 0, 1, 2,… Aircraft damage

id	x1_helicop	x2_load	x3_exp	Y_nb locations
1	A4	4	91,5	0
2	A4	4	84,0	1
3	A4	4	76,5	0
4	A4	5	69,0	0
5	A4	5	61,5	0
6	A4	5	80,0	0
7	A4	6	72,5	1
8	A4	6	65,0	0
9	A4	6	57,5	0
10	A4	7	50,0	2
11	A4	7	103,0	1
12	A4	7	95,5	1
13	A4	8	88,0	1
14	A4	8	80,5	1
15	A4	8	73,0	2
16	A6	7	116,1	3
17	A6	7	100,6	
18	A6	7	85,0	
19	A6	10	69,4	
20	A6	10	53,9	
21	A6	10	112,3	
22	A6	12	96,7	
23	A6	12	81,1	
24	A6	12	65,6	
25	A6	8	50,0	
26	A6	8	120,0	
27	A6	8	104,4	
28	A6	14	88,9	
29	A6	14	73,7	
30	A6	14	57,8	



Y_nb locations - Parameter estimates
Distribution : POISSON Link function: LOG

	Level of - Effect	Estimate	Standard - Error	Wald - Stat.	p	Lower CL - 95, %	Upper CL - 95, %
Intercept		-0,1216	0,9712	0,0157	0,9003	-2,025	1,7818
x2_load		0,1654	0,0675	5,9989	0,0143	0,033	0,2978
x3_exp		-0,0135	0,0083	2,6666	0,1025	-0,030	0,0027
x1_helicop	A4	-0,2844	0,2522	1,2717	0,2594	-0,779	0,2099
Scale		1,0000	0,0000				

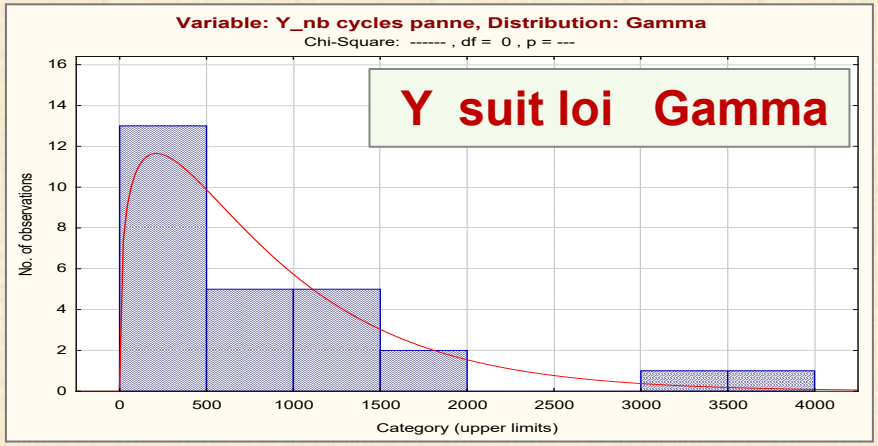
Exemple 4 Plan expérimental 3**3

3 facteurs X1 X2 X3 à 3 modalités - 27 essais

Y : nb cycles panne

Montgomery 4 th ed p. 458
Données plan expérimental complet
3 facteurs à 3 modalités.

ID	x1_long	x2_amp	x3_load	Y_nb cycles panne
1	-1	-1	-1	674
2	0	-1	-1	1414
3	1	-1	-1	3636
4	-1	0	-1	338
5	0	0	-1	1022
6	1	0	-1	1568
7	-1	1	-1	170
8	0	1	-1	442
9	1	1	-1	1140
10	-1	-1	0	370
11	0	-1	0	1198
12	1	-1	0	3184
13	-1	0	0	266
14	0	0	0	620
15	1	0	0	1070
16	-1	1	0	118
17	0	1	0	332
18	1	1	0	884
19	-1	-1	1	292
20	0	-1	1	634
21	1	-1	1	2000
22	-1	0	1	210
23	0	0	1	438
24	1	0	1	566
25	-1	1	1	90
26	0	1	1	220
27	1	1	1	360



Y_nb cycles panne - Parameter estimates (data-ex14-9 (Worsted Yarn) in Generalized examples.stw) Distribution : GAMMA Link function: LOG

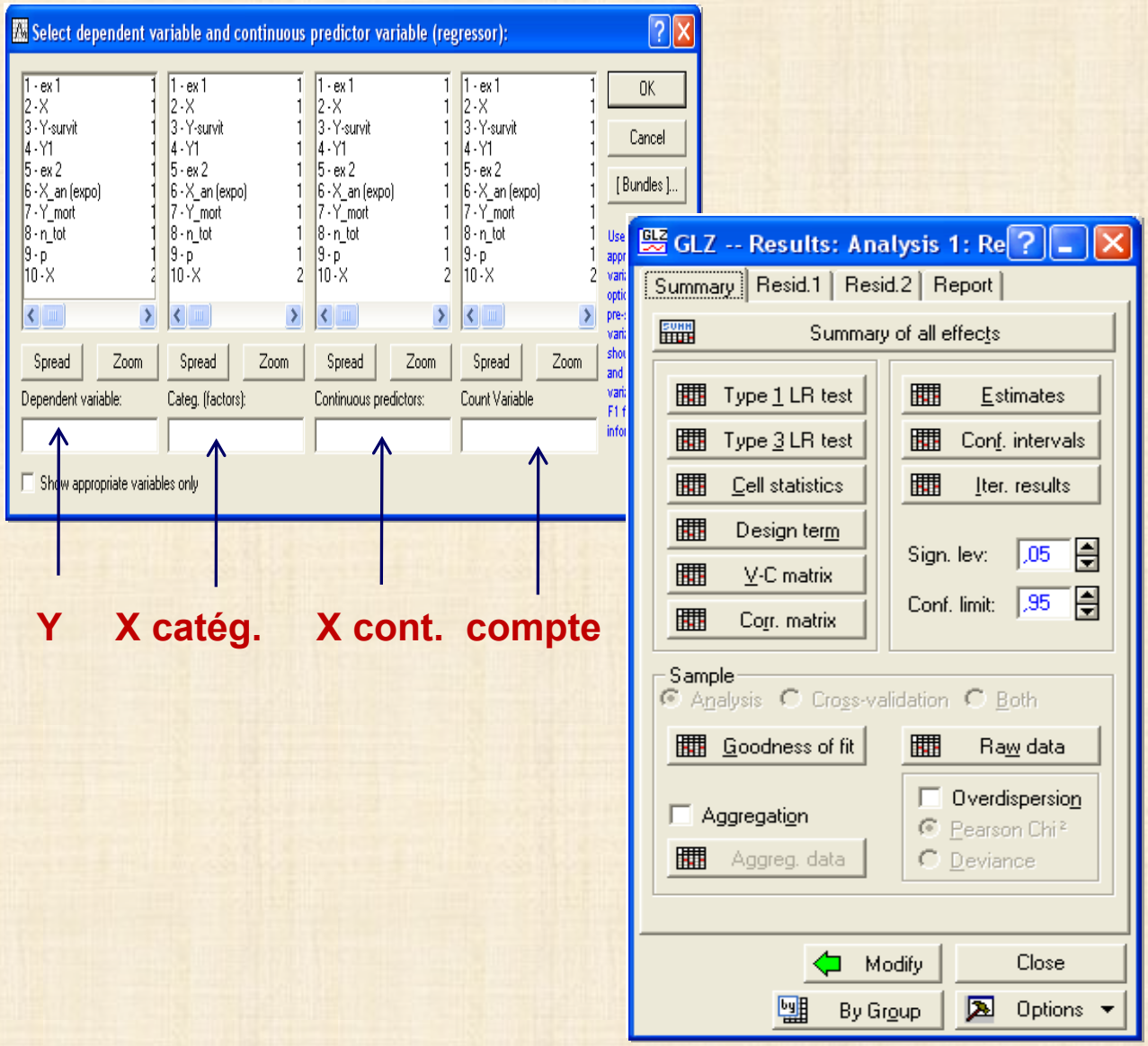
	Column	Estimate	Standard - Error	Wald - Stat.	p	Lower CL - 95, %	Upper CL - 95, %
Intercept	1	6,3489	0,03241	38373,04	0,000000	6,28539	6,41244
x1_long	2	0,8425	0,03969	450,49	0,000000	0,76471	0,92031
x2_amp	3	-0,6313	0,03969	252,95	0,000000	-0,70912	-0,55352
x3_load	4	-0,3851	0,03969	94,14	0,000000	-0,46293	-0,30733
Scale		35,2585	9,55111	13,63	0,000223		

facteurs X1 X2 X3 sont significatifs

Exemple 5: brulures

détails : chapitre régression simple

Données : burns.sta			
id	X	y	nombre
1	1,35	survit	13
2	1,35	décès	0
3	1,60	survit	19
4	1,60	décès	0
5	1,75	survit	67
6	1,75	décès	2
7	1,85	survit	45
8	1,85	décès	5
9	1,95	survit	71
10	1,95	décès	8
11	2,05	survit	50
12	2,05	décès	20
13	2,15	survit	35
14	2,15	décès	31
15	2,25	survit	7
16	2,25	décès	49
17	2,35	survit	1
18	2,35	décès	12



Exemple 5: brulures

détails : chapitre régression simple

y - Parameter estimates BINOMIAL, Link function: LOGIT Modeled probability that y = survit								
	Level of Effec	Col	Estimate	Standard Error	Wald - Stat.	p	CL -95%	CL +95%
Intercept		1	22,71	2,2661	100,42	0,0000	18,27	27,15
X		2	-10,66	1,0826	97,00	0,0000	-12,78	-8,54

y - Statistics of goodness of fit Distribution : BINOMIAL, Link function: LOGIT Modeled probability that y = survit			
	Df	Stat.	Stat/Df
Deviance	433	335,23	0,7742
Scaled Deviance	433	335,23	0,7742
Pearson Chi²	433	468,79	1,0826
Scaled P. Chi²	433	468,79	1,0826
Loglikelihood		-167,62	

Classification of cases			
	Predicted survit	Predicted décès	Percent correct
survit	265	43	86,0
décès	35	92	72,4

y - Predicted Values (Regression_logist.sta in Logistique.stw) Distribution : BINOMIAL, Link function: LOGIT Modeled probability that y = survit						
id	Response '1' = survit	Pred. - Value	LINEAR - Pred.	Standard - Error	Lower CL 95, %	Upper CL 95, %
1	1	1,000	8,31	0,812	0,999	1,000
3	1	0,996	5,65	0,547	0,990	0,999
5	1	0,983	4,05	0,392	0,964	0,992
6	0	0,983	4,05	0,392	0,964	0,992
7	1	0,952	2,98	0,293	0,918	0,972
8	0	0,952	2,98	0,293	0,918	0,972
9	1	0,872	1,92	0,203	0,820	0,910
10	0	0,872	1,92	0,203	0,820	0,910
11	1	0,701	0,85	0,143	0,639	0,756
12	0	0,701	0,85	0,143	0,639	0,756
13	1	0,446	-0,22	0,151	0,375	0,520
14	0	0,446	-0,22	0,151	0,375	0,520
15	1	0,217	-1,28	0,221	0,153	0,300
16	0	0,217	-1,28	0,221	0,153	0,300
17	1	0,087	-2,35	0,313	0,049	0,150
18	0	0,087	-2,35	0,313	0,049	0,150

survit à x = 1,5 versus à x = 2,5

$\exp(-\beta) = \exp(10,66) \approx 43\,000$ à 1

Exemple 6: enquête satisfaction client - data : Myers-ex4.11.sta

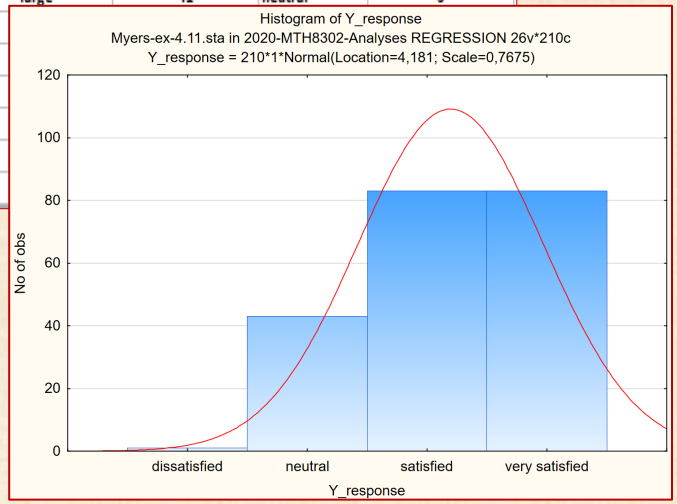
Exemple 4.11 - livre Myers - Montgomery-Robinson - Generalized Linear Model, 2nd ed, Wiley (2011)

Exemple de régression logistique multinomiale avec une réponse multinomiale avec modalités ordonnées: Y = very dissatified (=1) / dissatisfied (=2) / neutral (=3) / satisfied (=4) / very satisfied (=5)

(a) j'ai créé n = 210 données individuelles correspondant au tableau 4.11 de Myers (colonnes 2 à 5)

(b) j'ai ajouté une variable continue, X3_continuous (valeurs entre 0 et 1000) selon distribution uniforme. Cette variable n'est aucunement liée à la réponse Y et ne devrait pas avoir un effet significatif. But : je voulais une variable continue en plus des 2 variables catégoriques X1 et X2 pour avoir un cas mixte de 2 variables catégoriques et 1 variable continue

1 ID	2 X1_repeat_customer	3 X2_size_account	4 X3_continuous	5 Y_response	6 Y_reponse_num
1	yes	small	189	very satisfied	5
2	no	large	628	satisfied	4
3	no	large	943	satisfied	4
4	yes	large	566	very satisfied	5
5	no	large	494	very satisfied	5
6	yes	small	155	satisfied	4
7	no	small	867	neutral	3
8	yes	small	945	neutral	3
...					
201	yes	small	678	satisfied	4
202	no	small	485	neutral	3
203	no	large	702	satisfied	4
204	no	large	41	neutral	3
205	no				
206	no				
207	yes				
208	yes				
209	yes				
210	yes				



régression multinomiale : un peu de théorie

Y : variable output avec m+1 modalités nominales non ordonnées représentées conventionnellement par 0, 1, 2,,…, m

0 = modalité de référence (convention)

attention : Statistica utilise la dernière lue (m)

Modèle

X variables explicatives continues ou catégoriques

si X catégorique alors codage à effet

estimation des β par vraisemblance maximale

logit →

$$P(y_i = 0) = \frac{1}{1 + \sum_{j=1}^{m} \exp\left[\mathbf{x}'_i \boldsymbol{\beta}^{(j)}\right]}$$

$$P(y_i = 1) = \frac{\exp\left[\mathbf{x}'_i \boldsymbol{\beta}^{(1)}\right]}{1 + \sum_{j=1}^{m} \exp\left[\mathbf{x}'_i \boldsymbol{\beta}^{(j)}\right]}$$

$$\vdots \qquad\qquad \vdots$$

$$P(y_i = m) = \frac{\exp\left[\mathbf{x}'_i \boldsymbol{\beta}^{(m)}\right]}{1 + \sum_{j=1}^{m} \exp\left[\mathbf{x}'_i \boldsymbol{\beta}^{(j)}\right]}$$

$$\ln \frac{p(y_i = 1)}{p(y_i = 0)} = \mathbf{x}'_i \boldsymbol{\beta}^{(1)}$$

$$\ln \frac{p(y_i = 2)}{p(y_i = 0)} = \mathbf{x}'_i \boldsymbol{\beta}^{(2)}$$

$$\vdots$$

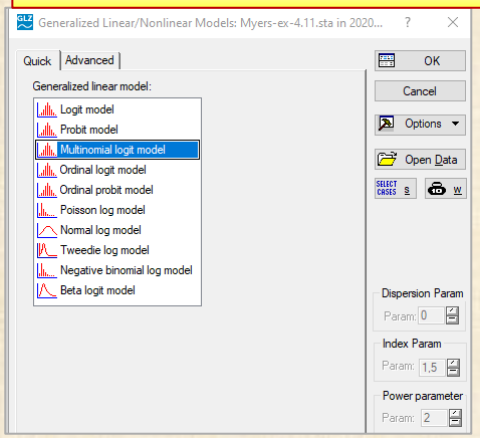
$$\ln \frac{p(y_i = m)}{p(y_i = 0)} = \mathbf{x}'_i \boldsymbol{\beta}^{(m)}$$

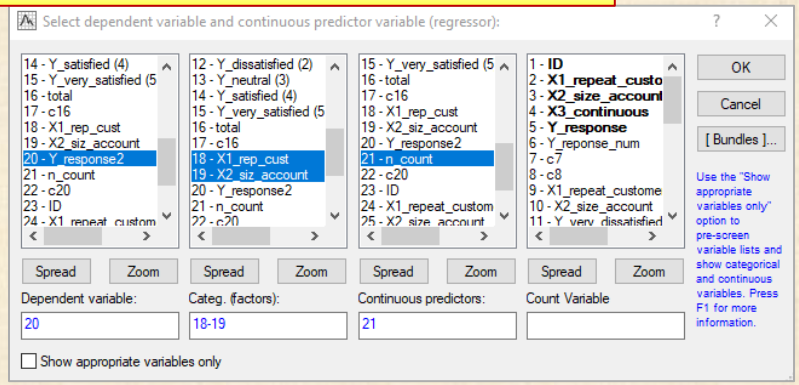
Exemple 6: enquête satisfaction client - data = Myers-ex4.11.sta - analyse avec STATITICA

Tableau croisé des données sans perte d'information

18 X1_rep_cust	19 X2_siz_account	20 Y_response2	21 n_count
yes	small	very dissastisfied	0
yes	small	dissatisfied	0
yes	small	neutral	7
yes	small	satisfied	10
yes	small	very satisfied	30
yes	large	very dissastisfied	0
yes	large	dissatisfied	0
yes	large	neutral	5
yes	large	satisfied	13
yes	large	very satisfied	25
no	small	very dissastisfied	0
no	small	dissatisfied	1
no	small	neutral	15
no	small	satisfied	33
no	small	very satisfied	15
no	large	very dissastisfied	0
no	large	dissatisfied	0
no	large	neutral	16
no	large	satisfied	28
no	large	very satisfied	12

analyse : comme si les modalités n'étaient pas ordonnées





Label	Column	Variable	Level of Variable	versus Level
Intercept	1			
X1_rep_cust	2	X1_rep_cust	yes	no
X2_siz_account	3	X2_siz_account	small	large

Effect	Degr. of Freedom	Wald Stat.	p
Intercept	4	128,2722	0,000000
X1_rep_cust	1	26,9481	0,000000
X2_siz_account	1	0,2624	0,608502

Effect	Level of Effect	Column	Estimate	Standard Error	Wald Stat.
Intercept 1		1	-22,4709	4435,896	0,00003
Intercept 2		2	-5,6691	1,006	31,76444
Intercept 3		3	-1,6059	0,186	74,29837
Intercept 4		4	0,3619	0,150	5,84503
X1_rep_cust	yes	5	-0,7298	0,141	26,94806
X2_siz_account	small	6	-0,0677	0,132	0,26236
Scale			1,0000	0,000	

	Df	Stat.	Stat/Df
Deviance	46	426,66	9,28
Scaled Deviance	46	426,66	9,28
Pearson Chi²	46	594,55	12,92
Scaled P. Chi²	46	594,55	12,92
AIC		438,66	
AICC		452,66	
BIC		442,05	
Loglikelihood		-213,33	

Exemple 6: enquête satisfaction client - data = Myers-ex4.11.sta - analyse avec STATISTICA

analyse : avec modalités ordonnées

Théorie logit des probabilités cumulées

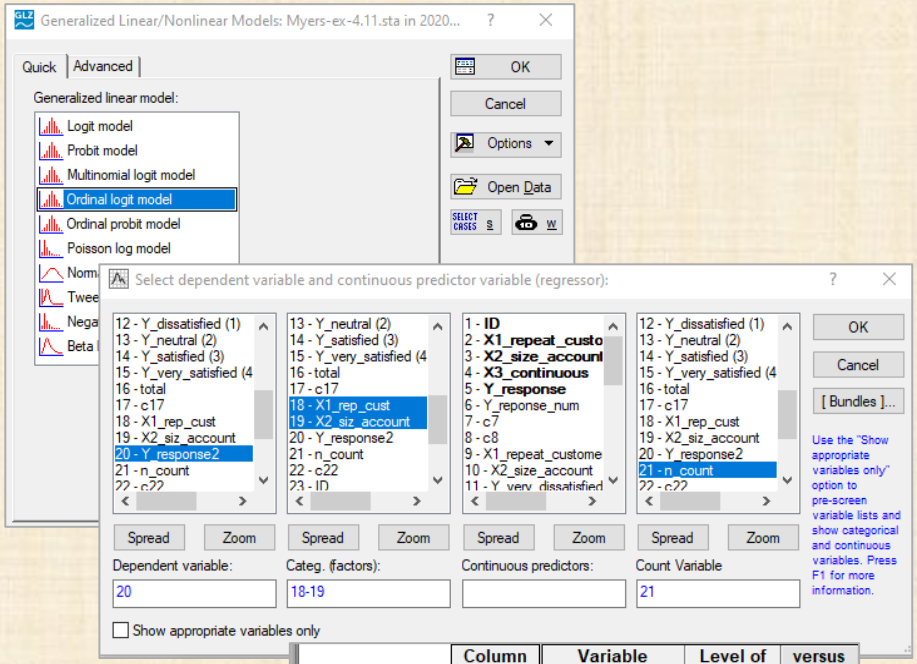
A second case involving a multilevel categorical response is an **ordinal** response. For example, customer satisfaction may be measured on a scale as not satisfied, indifferent, somewhat satisfied, and very satisfied. These outcomes would be coded as 1, 2, 3, and 4 respectively. The usual approach for modeling this type of response data is to use logits of cumulative probabilities:

$$\ln\frac{P(y_i \le k)}{1-P(y_i \le k)} = \alpha_k + \mathbf{x}_i'\boldsymbol{\beta},\ k = 1,\ldots,m-1 \qquad (4.45)$$

The cumulative probabilities are then

$$P(y_i \le k) = \frac{\exp(\alpha_k + \mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\alpha_k + \mathbf{x}_i'\boldsymbol{\beta})},\ k = 1,\ldots,m-1 \qquad (4.46)$$

The cumulative logit models presented above are also known as *proportional odds* models since the only difference in the models from one response category to the next is the intercepts. The effects of the regresssors in $\mathbf{x}_i$ are assumed to be the same for all response categories.



Label	Column	Variable	Level of Variable	versus Level
Intercept	1			
X1_rep_cust	2	X1_rep_cust	yes	no
X2_siz_account	3	X2_siz_account	small	large

Effect	Level of Effect	Column	Estimate	Standard Error	Wald Stat.
Intercept 1		1	-22,4709	4435,896	0,00003
Intercept 2		2	-5,6691	1,006	31,76444
Intercept 3		3	-1,6059	0,186	74,29837
Intercept 4		4	0,3619	0,150	5,84503
X1_rep_cust	yes	5	-0,7298	0,141	26,94806
X2_siz_account	small	6	-0,0677	0,132	0,26236
Scale			1,0000	0,000	

Effect	Degr. of Freedom	Log-Likelihd	Chi-Square	p
Intercept	4	-227,621		
X1_rep_cust	1	-213,461	28,31995	0,000000
X2_siz_account	1	-213,330	0,26318	0,607945

	Df	Stat.	Stat/Df
Deviance	46	426,66	9,28
Scaled Deviance	46	426,66	9,28
Pearson Chi²	46	594,55	12,92
Scaled P. Chi²	46	594,55	12,92
AIC		438,66	
AICC		452,66	
BIC		442,05	
Loglikelihood		-213,33	

	Predicted: very dissastisfied	Predicted: dissatisfied	Predicted: neutral	Predicted: satisfied	Predicted: very satisfied	Percent correct
Observed: very dissastisfied	0	0	0	0	0	0
Observed: dissatisfied	0	0	0	1	0	0
Observed: neutral	0	0	0	0	0	0
Observed: satisfied	0	0	0	0	0	0
Observed: very satisfied	0	0	0	119	90	43