Objectif

Dans ce didacticiel, nous montrons l'utilisation du composant **FORWARD ENTRY REGRESSION** qui réalise une régression en sélectionnant les meilleures variables. L'approche s'appuie sur une stratégie de sélection progressive fondée sur le calcul des coefficients de corrélation partielle.

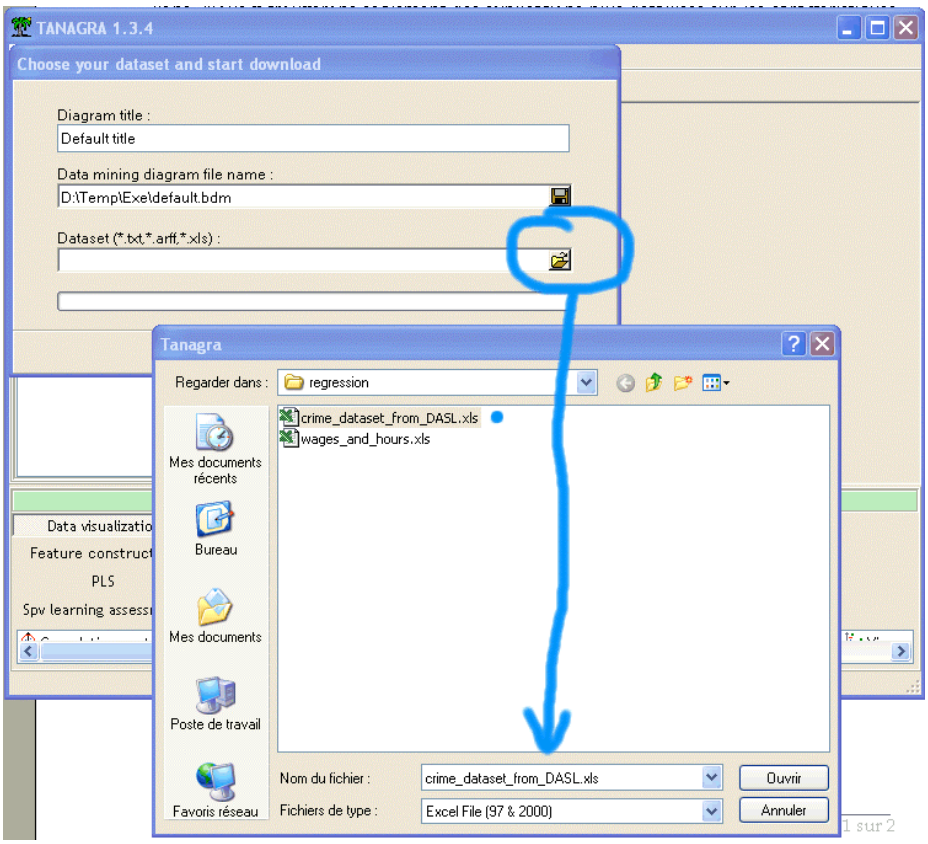
Fichier

Nous utilisons le fichier CRIME_DATASET_FROM_DASL.XL, il décrit différentes caractéristiques de 47 états des USA en 1960. L'objectif de la régression est d'expliquer le taux de criminalité à partir d'une série d'indicateurs socio-économiques : taux de chômage, niveau d'éducation, revenu moyen, budget de la police etc. Les données sont disponibles en ligne. Vous y trouverez également des explications plus détaillées sur les variables utilisées (http://lib.stat.cmu.edu/DASL/Stories/USCrime.html).

Sélection progressive dans la régression

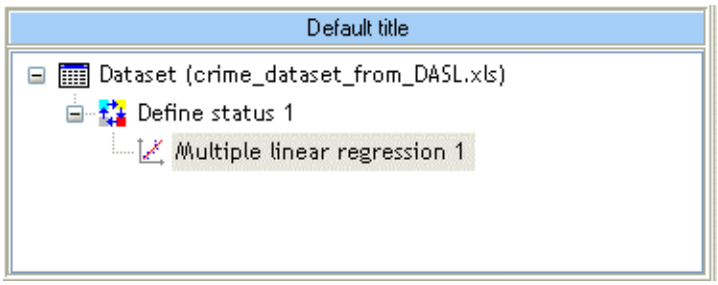
Charger le fichier de données

La première étape consiste à importer les données dans TANAGRA. Nous activons le menu FILE/NEW pour créer un nouveau diagramme.



Régression linéaire multiple sur toutes les variables

Nous essayons de réaliser une régression linéaire multiple directement sur l'ensemble des exogènes candidats. Pour ce faire, nous ajoutons un composant DEFINE STATUS dans le diagramme, nous plaçons CRIME RATE en TARGET, et toutes les autres variables en INPUT. Nous insérons à la suite le composant LINEAR MULTIPLE REGRESSION.



Les résultats sont les suivants.

Global results

Endogenous attribute	**CrimeRate**
Examples	47
R^2	0.769236
Adjusted-R^2	0.678329
Sigma error	21.935649
F-Test (13,33)	8.4618 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	52930.5756	13	4071.5827	8.4618	0.0000
Residual	15878.6992	33	481.1727		
Total	68809.2747	46			

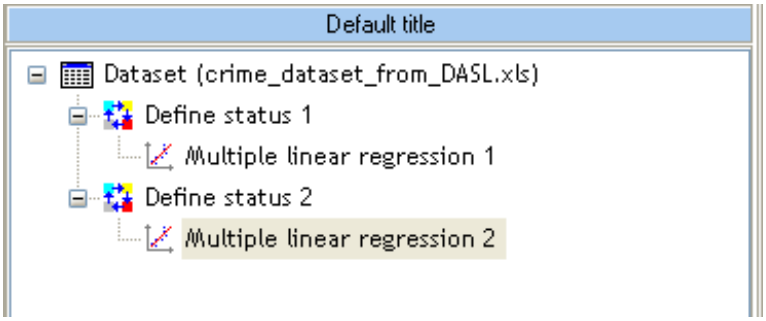
Coefficients

Attribute	Coef.	std	t(33)	p-value
Constant	-691.837589	155.887910	-4.438045	0.000096
Male14-24	1.039810	0.422708	2.459875	0.019306
Southern	-8.308312	14.911587	-0.557172	0.581170
Education	1.801601	0.649650	2.773186	0.009060
Expend60	1.607818	1.058667	1.518720	0.138357
Expend59	-0.667258	1.148773	-0.580844	0.565292
Labor	-0.041031	0.153477	-0.267344	0.790868
Male	0.164795	0.209932	0.784993	0.438057
PopSize	-0.041277	0.129516	-0.318701	0.751962
NonWhite	0.007175	0.063867	0.112338	0.911236
Unemp14-24	-0.601675	0.437154	-1.376345	0.177983
Unemp35-39	1.792263	0.856111	2.093493	0.044069
FamIncome	0.137358	0.105830	1.297913	0.203316
IncUnderMed	0.792933	0.235085	3.372959	0.001913

Apparemment, la régression semble de bonne qualité puisqu'elle est globalement très hautement significative : le niveau de signification réel est inférieur à 0.001. Nous constatons que 77% de la variance de l'endogène est expliquée par la régression, 4 variables se démarquent (pour un seuil de signification nominal de 5%) : MALE14-24, EDUCATION, UNEMP35-39, INCUNDERMED.

Régression sur les variables significatives

Un premier réflexe qui semble de bon sens consiste à recommencer la régression en ne conservant que ce sous-ensemble de variables. Nous ajoutons de nouveau un composant DEFINE STATUS à la racine du diagramme, et nous réalisons une nouvelle régression avec les variables ci-dessus.



Les résultats sont particulièrement décevants !

Global results

Endogenous attribute	**CrimeRate**
Examples	47
R^2	0.229784
Adjusted-R^2	0.156430
Sigma error	35.522631
F-Test (4,42)	3.1325 (0.024221)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	15811.2675	4	3952.8169	3.1325	0.0242
Residual	52998.0072	42	1261.8573		
Total	68809.2747	46			

Coefficients

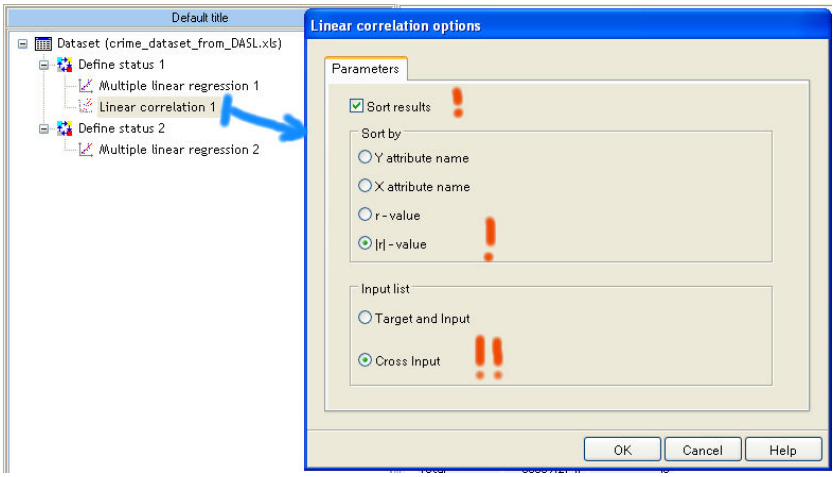
Attribute	Coef.	std	t(42)	p-value
Constant	-349.158324	155.025802	-2.252259	0.029592
Male14-24	0.767473	0.587023	1.307399	0.198189
Education	2.299542	0.789091	2.914165	0.005695
Unemp35-39	1.736663	0.706527	2.458027	0.018178
IncUnderMed	0.161780	0.227497	0.711130	0.480934

Nous seulement la qualité globale de la régression a fortement chuté (y compris pour le R^2 corrigé qui tient compte de la complexité du modèle), mais en plus le rôle des variables n'est

plus du tout cohérent avec la première régression : seules EDUCATION et UNEMP35-39 semblent se démarquer finalement.

Corrélations croisées entre les exogènes

Suspectant un problème de colinéarité, nous décidons de vérifier le degré de colinéarité entre les variables exogènes. Nous plaçons donc le composant LINEAR CORRELATION de la palette STATISTICS dans le diagramme (au même niveau que la première régression), et nous le paramétrons de la manière suivante.



La corrélation entre chaque exogène est donc calculée, et les valeurs sont rangées de manière décroissante suivant la valeur absolue du coefficient -- ou selon le carré du coefficient de corrélation, ce qui revient à la même chose. Nous n'affichons que les premiers résultats ici.

Linear correlation 1

Parameters

Cross-tab parameters	
Sort results	yes
Sort criterion	|r| statistic
Input list	Cross-input (Y x X)

Results

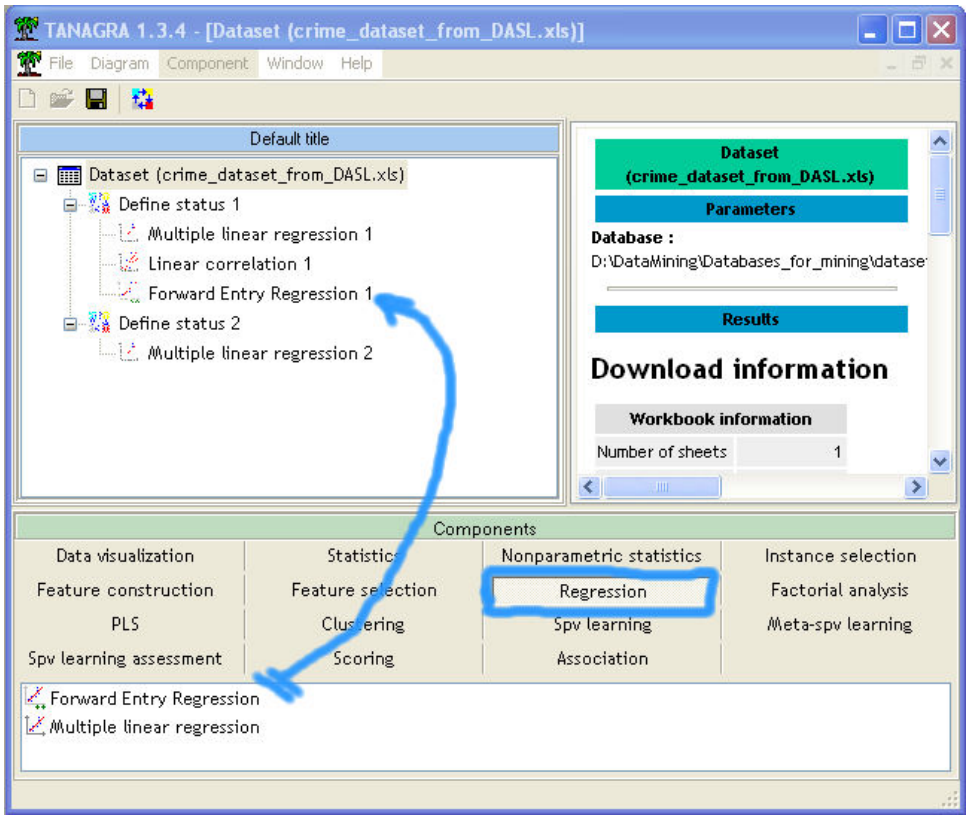
Y	X	r	r²	t	Pr(>|t|)
Expend60	Expend59	0.9936	0.9872	58.9449	0.0000
FamIncome	IncUnderMed	-0.8840	0.7815	-12.6848	0.0000
Expend59	FamIncome	0.7943	0.6309	8.7694	0.0000
Expend60	FamIncome	0.7872	0.6197	8.5636	0.0000
Education	IncUnderMed	-0.7687	0.5908	-8.0610	0.0000
Southern	NonWhite	0.7671	0.5884	8.0213	0.0000
Unemp14-24	Unemp35-39	0.7459	0.5564	7.5129	0.0000
Southern	IncUnderMed	0.7372	0.5434	7.3186	0.0000
Education	FamIncome	0.7360	0.5417	7.2930	0.0000
Southern	Education	-0.7027	0.4938	-6.6261	0.0000
NonWhite	IncUnderMed	0.6773	0.4588	6.1759	0.0000

Effectivement, les variables exogènes sont très liées entre elles, certains carrés du coefficient de corrélation dépassant même le coefficient de détermination R^2 de la régression initiale.

Sélection progressive des variables

Parmi les différentes approches possibles pour résoudre ce problème de colinéarité, la sélection de variable a le mérite de l'automaticité, l'utilisateur n'a pas besoin d'intervenir dans le processus de calcul. La tentation est forte de mettre le nombre maximal de variables candidates et laisser la machine trouver la « bonne » solution. Attention, car étant totalement mécanique, cette méthode peut proposer des solutions qui n'ont aucun sens pour le praticien, en introduisant par exemple des variables totalement farfelues dues un artefact statistique.

Le composant FORWARD ENTRY REGRESSION de TANAGRA effectue une « sélection progressive des variables » : elle consiste à partir d'un ensemble vide de variables, d'en rajouter au fur et à mesure tant que l'adjonction introduit une amélioration significative du modèle. La méthode programmée s'appuie sur le calcul des corrélations partielles.



Les résultats sont nettement plus encourageants que pour la régression précédente.

Global results

Endogenous attribute	**CrimeRate**
Examples	47
R^2	0.729635
Adjusted-R^2	0.696663
Sigma error	21.301348
F-Test (5,41)	22.1293 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	50205.6311	5	10041.1262	22.1293	0.0000
Residual	18603.6437	41	453.7474		
Total	68809.2747	46			

Coefficients

Attribute	Coef.	std	t(41)	p-value
Constant	-524.374333	95.115565	-5.513023	0.000002
Expend60	1.233122	0.141635	8.706359	0.000000
IncUnderMed	0.634926	0.146846	4.323752	0.000096
Education	2.030773	0.474189	4.282623	0.000109
Male14-24	1.019822	0.353203	2.887356	0.006175
Unemp35-39	0.913608	0.434092	2.104642	0.041496

Parmi les 13 variables exogènes candidates, 5 ont finalement été retenues. Nous constatons que le R^2 est tout à fait comparable à celui de la première régression, le R^2 ajusté qui tient compte de la complexité du modèle est même meilleur : des variables non-pertinentes ont bien été éliminées.

En descendant plus bas dans la fenêtre de résultats, nous disposons du détail des calculs lors du processus de sélection.

Forward Selection Process

partial corr. F (p-value)	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
d.f.	45	44	43	42	41	40
r(Y,Xj*/Xj1,Xj2...)	Expend60 : 0.6876	IncUnderMed : 0.4516	Education : 0.4509	Male14-24 : 0.3226	Unemp35-39 : 0.3123	-
R^2	0.4728	0.5803	0.6656	0.7004	0.7296	-
Male14-24	-0.0895 0.36 (0.5498)	0.4123 9.01 (0.0044)	0.2505 2.88 (0.0970)	0.3226 4.88 (0.0327)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)
Southern	-0.0906 0.37 (0.5446)	0.2458 2.83 (0.0997)	-0.1081 0.51 (0.4797)	0.0424 0.08 (0.7847)	-0.0393 0.06 (0.8023)	-0.0489 0.10 (0.7586)
Education	0.3228 5.24 (0.0269)	-0.0145 0.01 (0.9236)	0.4509 10.97 (0.0019)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)
Expend60	0.6876 40.36 (0.0000)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)
Expend59	0.6667 36.01 (0.0000)	-0.2007 1.85 (0.1810)	-0.1031 0.46 (0.5005)	-0.1360 0.79 (0.3786)	-0.1484 0.92 (0.3423)	-0.1285 0.67 (0.4172)
Labor	0.1889 1.66 (0.2036)	0.1461 0.96 (0.3325)	0.3004 4.26 (0.0450)	0.0562 0.13 (0.7173)	0.0381 0.06 (0.8085)	0.1501 0.92 (0.3428)
Male	0.2139 2.16 (0.1488)	0.2628 3.26 (0.0777)	0.3967 8.03 (0.0070)	0.2255 2.25 (0.1410)	0.1900 1.54 (0.2224)	0.1135 0.52 (0.4743)
PopSize	0.3375 5.78 (0.0204)	-0.0395 0.07 (0.7943)	-0.2125 2.03 (0.1611)	-0.1307 0.73 (0.3977)	-0.0627 0.16 (0.6896)	-0.0734 0.22 (0.6440)
NonWhite	0.0326 0.05 (0.8278)	0.2531 3.01 (0.0896)	-0.1123 0.55 (0.4625)	0.0428 0.08 (0.7828)	-0.0894 0.33 (0.5685)	-0.0988 0.39 (0.5335)
Unemp14-24	-0.0505 0.11 (0.7362)	-0.0282 0.03 (0.8526)	0.0283 0.03 (0.8537)	0.0566 0.13 (0.7153)	0.1570 1.04 (0.3146)	-0.1861 1.44 (0.2380)
Unemp35-39	0.1773 1.46 (0.2331)	0.0701 0.22 (0.6432)	-0.0094 0.00 (0.9513)	0.1643 1.17 (0.2865)	0.3123 4.43 (0.0415)	0.0000 0.00 (0.0000)
FamIncome	0.4413 10.88 (0.0019)	-0.2233 2.31 (0.1358)	0.2722 3.44 (0.0705)	0.1859 1.50 (0.2269)	0.2815 3.53 (0.0674)	0.2595 2.89 (0.0970)
IncUnderMed	-0.1790 1.49 (0.2286)	0.4516 11.27 (0.0016)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)	0.0000 0.00 (0.0000)

A la première étape, EXPEND60 est la variable la plus corrélée avec l'endogène (r = 0.6876). Le test de signification peut être réalisé avec un t de Student, nous préférons prendre le carré de sa valeur ce qui revient à un F de Fisher à [1 ; n-2] degrés de liberté ($t^2 = F = 40.36$), il est très hautement significatif.

Nous sélectionnons donc cette première variable, puis nous calculons la corrélation entre l'endogène et les exogènes candidates restantes en enlevant l'information apportée par EXPEND60 : c'est la notion de corrélation partielle, nous retrouvons les valeurs dans la colonne « STEP 2 » de notre tableau. Dans notre cas, la variable INCUNDERMED se démarque, $r_{crime,incUnderMed/\exp and60} = 0.4516$; le F de Fisher à [1 ; n-3 (!)] degrés de liberté est égal à 11.27, il est hautement significatif (inférieur à 1%).

Nous continuons le processus jusqu'à ce que nous ne puissions plus introduire de nouvelle variable. Ce qui nous amène à sélectionner 5 variables au final.

Remarque : L'utilisation de la loi de Fisher telle quelle est sujette à caution dans ce processus. En effet si le calcul du degré de liberté est approprié, il faut noter que la variable a été retenue à l'issue d'un processus de sélection. Le F de la variable à tester correspond en fait à un F(max). A l'instar de ce qui se fait dans les processus de comparaisons multiples, il apparaît nécessaire de corriger le niveau de signification empirique obtenu. A défaut de solution totalement satisfaisante à cet égard, la plupart des logiciels commerciaux donnent le choix entre (1) l'introduction d'un niveau de signification nominal à comparer avec la p-value à chaque étape (comme nous le faisons dans TANAGRA) et (2) l'introduction d'une valeur de coupure que l'on compare directement avec le F calculé (F-to-enter dans certains logiciels – en le fixant à 4, nous « approchons » un niveau de signification de 5%). Dans la pratique, les résultats que produisent ces deux techniques diffèrent peu.

Revenons à notre exemple, à titre de comparaison, voici la solution proposée sur notre site de référence -- http://lib.stat.cmu.edu/DASL/Stories/USCrime.html.

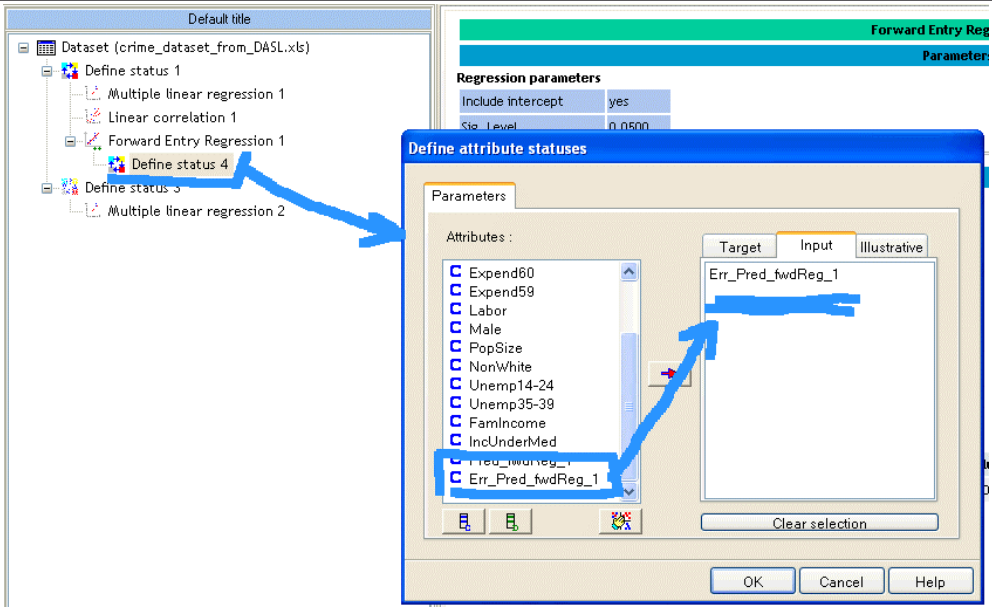
Dependent variable is: **R**
No Selector
48 total cases of which 1 is missing
R squared = 73.0% R squared (adjusted) = 69.7%
s = 21.30 with 47 - 6 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	50205.6	5	10041.1	22.1
Residual	18603.6	41	453.747	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-524.374	95.12	-5.51	≤ 0.0001
Age	1.01982	0.3532	2.89	0.0062
Ed	2.03077	0.4742	4.28	0.0001
U2	0.913608	0.4341	2.10	0.0415
X	0.634926	0.1468	4.32	≤ 0.0001
Ex0	1.23312	0.1416	8.71	≤ 0.0001

Normalité des résidus

Enfin, dernière étape, il est possible de tester la normalité des résidus. Le composant de régression produit automatiquement deux nouvelles variables : la prédiction et le résidu. Nous allons donc placer un nouveau composant DEFINE STATUS pour sélectionner en INPUT la variable ERR_PRED_LMREG_2.



Puis nous ajoutons le composant NORMALITY TEST. Nous constatons que l'hypothèse de normalité des résidus est compatible avec les résidus observés : à un niveau de signification de 5%, tous les tests sont cohérents.