Subject

In this tutorial, we show how to use the **FORWARD ENTRY REGRESSION** component: it performs a multiple linear regression with a forward variable selection based on partial correlation.

Dataset

We use CRIME_DATASET_FROM_DASL.XL from the DASL website¹. It contains various characteristics of 47 states of USA. We want to explain the criminality from unemployment, education level, ...

Forward Selection for Regression

Import the dataset

We build a new diagram and import the dataset with the FILE / NEW menu.

TANAGRA 1.3.4							_ 🗆 🗙
Choose your datas	et and start dov	vnload					
Diagram title : Default title Data mining di D:\Temp\Exelu Dataset (*.bxt,*.			C				
	T						
	Tanagra					? ×	
	Regarder dans :	regression		🕑 G 💋	i 🖻 🛄 -		
	Mes documents récents	Crime_dataset_from いいのでは、 のでは、 ので、 ので、 ので、 ので、 ので、 ので、 ので、 ので、 ので、 ので					
Data visualizatio Feature construct							
PLS Spy learning assessi	Mes documents						17 . c.e
	Poste de travail		V	/			
		Nom du fichier :	crime_dataset_from_DA	ASL.xls	~	Ouvrir	
	Favoris réseau	Fichiers de type :	Excel File (97 & 2000)		~	Annuler	1 sur 2

R.R.

¹ <u>http://lib.stat.cmu.edu/DASL/Stories/USCrime.html</u>

Linear multiple regression

First, we want to perform a regression with the whole variables. We add a DEFINE STATUS component in the diagram and set CRIME RATE as TARGET, and the other variables as INPUT. We add the LINEAR MULTIPLE REGRESSION component.

	Default title
🖃 🏢 Data	aset (crime_dataset_from_DASL.xls)
ė 🔁	Define status 1
	🛃 Multiple linear regression 1

We obtain the following results.

Endogenous att	tribute		CrimeRate		
Examples			47		
R²			0.769236	•	
Adjusted-R ^z			0.678329	•	
Sigma error			21.935649		
F-Test (13,33)		8,461	8 (0.000000)	•	
Analysis (of varianc	e			
Source	x\$\$	d.f.	xMS	F	p-value
Regression	52930.5756	13	4071.5823	7 8.4618	0.0000
Residual	15878.6992	33	481,1723	7	
Total	68809.2747	46			
Coefficie	nts				
Coefficie Attrii		Coef.	std	t(33)	p-value
		Coef. -691.837589	std 155,887910	t(33) -4.438045	p-value 0.000096
Attril				-4.438045	-
Attril Constant		-691.837589	155.887910	-4.438045 2.459875	0.000096
Attril Constant Male14-24		-691.837589 1.039810	155.887910 0.422708	-4.438045 2.459875	0.000096 0.019306 0.581170
Attril Constant Male14-24 Southern		-691.837589 1.039810 -8.308312	155.887910 0.422708 14.911587	-4.438045 2.459875 -0.557172 2.773186	0.000096 0.019306 0.581170
Attril Constant Male14-24 Southern Education		-691.837589 1.039810 -8.308312 1.801601	155.887910 0.422708 14.911587 0.649650	-4.438045 2.459875 -0.557172 2.773186 1.518720	0.000096 0.019306 0.581170 0.009060
Attril Constant Male14-24 Southern Education Expend60		-691.837589 1.039810 -8.308312 1.801601 1.607818	155.887910 0.422708 14.911587 0.649650 1.058667	-4.438045 2.459875 -0.557172 2.773186 1.518720 -0.580844	0.000096 0.019306 0.581170 0.009060 0.138357
Attril Constant Male14-24 Southern Education Expend60 Expend59		-691.837589 1.039810 -8.308312 1.801601 1.607818 -0.667258	155.887910 0.422708 14.911587 0.649650 1.058667 1.148773	-4.438045 2.459875 -0.557172 2.773186 1.518720 -0.580844	0.000096 0.019306 0.581170 0.009060 0.138357 0.565292
Attril Constant Male14-24 Southern Education Expend60 Expend59 Labor		-691.837589 1.039810 -8.308312 1.801601 1.607818 -0.667258 -0.041031	155.887910 0.422708 14.911587 0.649650 1.058667 1.148773 0.153477	-4.438045 2.459875 -0.557172 2.773186 1.518720 -0.580844 -0.267344 0.784993	0.000096 0.019306 0.581170 0.009060 0.138357 0.565292 0.790868
Attril Constant Male14-24 Southern Education Expend60 Expend59 Labor Male		-691.837589 1.039810 -8.308312 1.801601 1.607818 -0.667258 -0.041031 0.164795	155.887910 0.422708 14.911587 0.649650 1.058667 1.148773 0.153477 0.209932	-4.438045 2.459875 -0.557172 2.773186 1.518720 -0.580844 -0.267344 0.784993 -0.318701	0.000096 0.019306 0.581170 0.009060 0.138357 0.565292 0.790868 0.438057
Attril Constant Male14-24 Southern Education Expend60 Expend59 Labor Male PopSize		-691.837589 1.039810 -8.308312 1.801601 1.607818 -0.667258 -0.041031 0.164795 -0.041277	155.887910 0.422708 14.911587 0.649650 1.058667 1.148773 0.153477 0.209932 0.129516	-4.438045 2.459875 -0.557172 2.773186 1.518720 -0.580844 -0.267344 0.784993 -0.318701 0.112338	0.000096 0.019306 0.581170 0.009060 0.138357 0.565292 0.790868 0.438057 0.751962

0.137358

0.792933

0.105830

0.235085

1.297913

3.372959

0.203316

0.001913 🙍

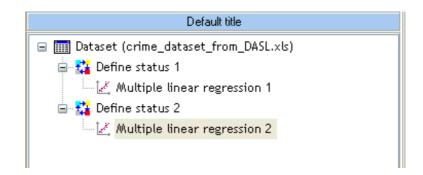
Famincome

IncUnderMed

The results seem encouraging -- $R^2 = 0.77$ -- and 4 variables are significant -- p-value lower than 0.05: MALE14-24, EDUCATION, UNEMP35-39, and INCUNDERMED.

Regression with the significant exogenous variables

We want to perform a new regression with only the significant variables. We add a DEFINE STATUS component and set as INPUT the significant variables above, TARGET is always CRIMERATE. We add a new REGRESSION component.



The results are particularly disappointing!

Global re	sults						
Endogenous att	ribute		CrimeRate				
Examples			47				
R²			0.229784	•			
Adjusted-R ²			0.156430	•			
Sigma error			35,522631				
F-Test (4,42)		3,13	25 (0.024221)				
Analysis of variance							
Source	x\$\$ 15811.2675	u 4	xMS 3952.816	F 9 3.1325	p-value 0.0242		
Regression Residual	52998.0072	42			0.0242		
Total	68809.2747	42		3			
TULA	00007.2747	40					
Coefficie	nts						
Attrib	ute	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		

Explained variance is significantly lower and only two variables seems relevant: EDUCATION andUNEMP35-39. These results are not at all coherent with what we saw previously.

Correlation between the exogenous variables

We suspect a problem of colinearity between the exogenous variables. In order to check that, we add the LINEAR CORREALTION in the diagram and we set the following parameters.

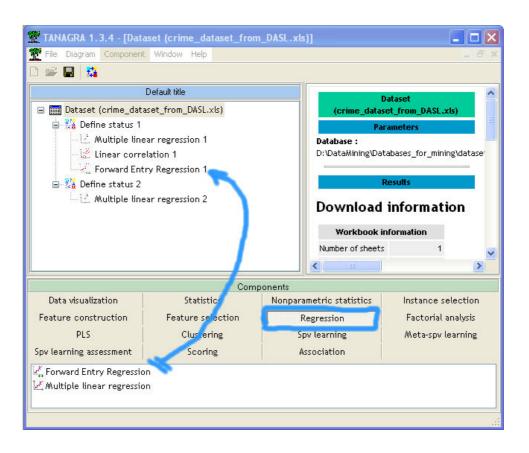
Default title	Linear correlation options
Dataset (crime_dataset_from_DASL.xls) Define status 1 // Multiple linear regression 1 // Iniear correlation 1 // Define status 2 // Multiple linear regression 2	Parameters Sort results Sort by Y attribute name A attribute name r - value Input list Target and Input Cross Input OK Cancel Help

We see that several variables are highly correlated. In some cases, the square of the correlation coefficient is higher than the R^2 of the regression.

					Linear corr
					Parame
Cross-ta	b parameters				
Sort results	yes				
Sort criterion	r statistic				
Input list	$Cross-input\;(Y\timesX)$				
					Resu
Y	х	г	г ²	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

Forward selection for regression

There are various solutions for colinearity problem in the regression. The FORWARD ENTRY REGRESSION performs a forward selection of variables using the partial correlation measurements.



The quality of the regression with 5 variables is close to the first regression with the whole dataset: irrelevant variables are rejected.

Global res	sults				
Endogenous attr	ribute		CrimeRate		
Examples			47		
R²			0.729635		
Adjusted-R ²			0.696663		
Sigma error			21.301348		
F-Test (5,41)		22,129	3 (0.000000)		
Analysis o	of varianc	e			
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			
Coefficie	nts				
Attrib	ute	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

Below, TANAGRA shows the detailed results and the steps of the computation.

Forward Se	lection Pro	cess				
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	-
R²	0.4728	0.5803	0.6656	0.7004	0.7296	-
Male14-24	-0.0895	0.4123	0.2505	0.3226	0.0000	0.0000
	0.36 (0.5498)	9.01 (0.0044)	2.88 (0.0970)	4.88 (0.0327)	0.00 (0.0000)	0.00 (0.0000)
Southern	-0.0906	0.2458	-0.1081	0.0424	-0.0393	-0.0489
	0.37 (0.5446)	2.83 (0.0997)	0.51 (0.4797)	0.08 (0.7847)	0.06 (0.8023)	0.10 (0.7586)
Education	0.3228	-0.0145	0.4509	0.0000	0.0000	0.0000
	5.24 (0.0269)	0.01 (0.9236)	10.97 (0.0019)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)
Expend60	0.6876	0.0000	0.0000	0.0000	0.0000	0.0000
	40.36 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)
Expend59	0.6667	-0.2007	-0.1031	-0.1360	-0.1484	-0.1285
	36.01 (0.0000)	1.85 (0.1810)	0.46 (0.5005)	0.79 (0.3786)	0.92 (0.3423)	0.67 (0.4172)
Labor	0.1889	0.1461	0.3004	0.0562	0.0381	0.1501
	1.66 (0.2036)	0.96 (0.3325)	4.26 (0.0450)	0.13 (0.7173)	0.06 (0.8085)	0.92 (0.3428)
Male	0.2139	0.2628	0.3967	0.2255	0.1900	0.1135
	2.16 (0.1488)	3.26 (0.0777)	8.03 (0.0070)	2.25 (0.1410)	1.54 (0.2224)	0.52 (0.4743)
PopSize	0.3375	-0.0395	-0.2125	-0.1307	-0.0627	-0.0734
	5.78 (0.0204)	0.07 (0.7943)	2.03 (0.1611)	0.73 (0.3977)	0.16 (0.6896)	0.22 (0.6440)
NonWhite	0.0326	0.2531	-0.1123	0.0428	-0.0894	-0.0988
	0.05 (0.8278)	3.01 (0.0896)	0.55 (0.4625)	0.08 (0.7828)	0.33 (0.5685)	0.39 (0.5335)
Unemp14-24	-0.0505	-0.0282	0.0283	0.0566	0.1570	-0.1861
	0.11 (0.7362)	0.03 (0.8526)	0.03 (0.8537)	0.13 (0.7153)	1.04 (0.3146)	1.44 (0.2380)
Unemp35-39	0.1773	0.0701	-0.0094	0.1643	0.3123	0.0000
	1.46 (0.2331)	0.22 (0.6432)	0.00 (0.9513)	1.17 (0.2865)	4.43 (0.0415)	0.00 (0.0000)
Famincome	0.4413	-0.2233	0.2722	0.1859	0.2815	0.2595
	10.88 (0.0019)	2.31 (0.1358)	3.44 (0.0705)	1.50 (0.2269)	3.53 (0.0674)	2.89 (0.0970)
IncUnderMed	-0.1790	0.4516	0.0000	0.0000	0.0000	0.0000
	1.49 (0.2286)	11.27 (0.0016)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)

Tutorial

Forward Selection for Regression

At the first step, EXPEND60 is the most correlated with the endogenous variable (r = 0.6876). With the t of Student (the Fisher's F is the square of the Student's t), we see that this correlation is significant (at the significance level of 0.05).

At the second step, we compute the correlation between the endogenous the resulting variables, by removing the information given by EXPEND60: this is a partial correlation. We see that INCUNDERMED is the most correlated with the endogenous and it is highly significant (the degree of freedom of the test has been modified!).

We continue the process until we cannot introduce anymore a new variable: we obtain 5 relevant variables.

On the DASL website, the authors propose the same regression - 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 Regression	Sum of Sq 50205.6	uares df 5	Mean Squ 100			
Residual	18603.6	41	453			
Variable Constant Age Ed U2 X Ex0	Coefficient -524.374 1.01982 2.03077 0.913608 0.634926 1.23312	s.e. of Coe 95.12 0.3532 0.4742 0.4341 0.1468 0.1416	ff t-ratio -5.51 2.89 4.28 2.10 4.32 8.71	prob ≤ 0.0001 0.0062 0.0415 ≤ 0.0001 ≤ 0.0001		

Normality test

The regression component produces two variables: the prediction and the residuals. In order to check the validity of the regression, it is possible to test the normality of the residuals.

We add a new DEFINE STATUS in the diagram and set the variable ERR_PRED_LMREG_2 as INPUT.

Tutorial Forward Selection for Regression

Default title	Forward Entry Reg
Dataset (crime_dataset_from_DASL.xls)	Parameter
Dataset (crime_dataset_from_DASL.xls) Define status 1 Multiple linear regression 1 Define status 4 Define status 3 L: Multiple linear regression 2	Parameters Include intercept yes Sie level 0.0500 Define attribute statuses Parameters Attributes : Target Input Illustrative Err_Pred_fwdReg_1 Expend60 E Expend59 E Labor C Male C PopSize C NonWhite Unemp14-24 Unemp15-39 E FamIncome IncUnderMed C Treo_worneg E Err_Pred_fwdReg_1
	E E Clear selection
	OK Cancel Help

We add the NORMALITY TEST component. At the significance level of 5%, we see that the observed residuals are compatible with the assumption of normality.

Default title	Rormality Test 1						
Dataset (crime_dataset_from_DASL.xls)							
□ ↓ Define status 1 □ ↓ Multiple linear regression 1				Normality Test 1			
- Linear correlation 1	Parameters						
E V Forward Entry Regression 1	Attributes : 1 Examples : 47						
🚽 🎇 Define status 4							
Normality Test 1				Results			
□ ♣ Define status o □ ↓ Multiple linear regression 2	Attribute	Mu ; Sigma	Shapiro-Wilk (p-value)	Lilliefors D = max[D-,D+] (p-value)	Anderson-Darling (p-value)	d'Agostino (p-value)	
1	Err_Pred_fwdReg_1	0.0000; 20.1104	0.971494 (0.3017)	0.0819 = max[0.0819,0.0811] (p ≻= 0.20)	0.428719	1.0000 ^2 + 1.7925 ^2 = 4.2131 (0. <i>1217</i>)	
	Computation time : 0 Created at 03/08/20						