# Subject

Computing the correlation coefficient between two or more variables.

Computing a statistical indicator and sorting the results according to this indicator is a recurring task of the data miner. In this tutorial, we show how to quickly set up the calculation of the linear correlation (1) of an endogenous variable with exogenous variables in order to detect relevant attributes; (2) between exogenous variables in order to detect collinearities.

### Dataset

We use the CARS\_ACCELERATION.XLS dataset: ACCELERATION is the endogenous attribute (acceleration time from 0 to x mph).

## Correlation coefficient

#### Creating a new diagram

The simplest way in order to create a diagram is to load the dataset in the EXCEL spreadsheet. Then, we select the data range and we click on the menu TANAGRA/EXECUTE TANAGRA<sup>1</sup>.

м	icrosoft Ex	cel - cars_	_accel	eration.xls									X
1	Fichier Edition	n <u>A</u> ffichage	Insert	tion Forma <u>t</u> (	utils <u>D</u> onnées	Fe <u>n</u> être ?	Tanagra Sipir	ha				_ 6	9 ×
	🛩 🖬 🖉	) 🖪 🖤	χ 🗈	n 🗈 💅 🖂	) + Ci + 🍓	Σ f* 🕃	Execute T	anagra 🔸	- 2	. G	I 🛛	- 🕭	• *
-	Al	<b>▼</b> =	cylii	nders	1		*		1	14			
	A	В		с	D	E	Г		;	Н		1	
1	cylinders	engine.o	displa I	horesepower	weight	acceleration	n i						
2		8	350	165	3693	11	.5						
3		8	318	150	3436	1	11		/				
4		8	302	140	3449	10	.5						
5		8	429	198	4341	1	LO						
6		8	440	215	4312	8	.5						
7		8	455	225	4425	1	LO	/			_		_
8		8	383	170	3563		10						_
9		8	340	Execute Ta	nagra								_
10		4	400					-					_
+++		6	100	Datas	et range (includin	g the name of I	the attributes -	first row):					
13		4	97	\$44	1:\$F\$241								_
14		4	110	1 41.14					-				
15		4	104				01	]					
16		4	121				UK	Can	cel				
17		8	360										
18		8	307	200	4376	1	.5						
19		8	304	193	4732	18	. 5						
20		4	97	88	2130	14	. 5						
<u>21</u>		4	113	95	2228	1	14						_
22		6	250	100	3329	15	.5						
23		6	232	100	3288	15	.5						
24		8	350	165	4209	]	.2				_		_
25		8	318	150	4096		13				_		
26		0	400	1/0	4/46		12						_
21		0	400	1/5	2408		2				_		-
20		6	250	100	3282	1	15						-
H I	▶ N\ cars	accelerat	ion /	100			•					•	1
Prêt							Somme=	803044.2		NUM			

We check the range selection. If it is right, we click on the OK button. TANAGRA is automatically started and a new stream diagram is ready.

<sup>&</sup>lt;sup>1</sup> The EXCEL add-in TANAGRA.XLA is available since the version 1.4.11. See the tutorial on the web site for the installation of this add-in in your spreadsheet.



#### Correlation between the variable of interest and the others

We want to measure the relation between ACCELERATION and the other variables. We add a component in the diagram, using the short cut into the toolbar. We set ACCELERATION as TARGET and all the others variables as INPUT.



Then we insert the LINEAR CORRELATION component (STATISTICS tab).

We activate the PARAMETERS contextual menu. We state that the results must be sorted according to the absolute value of the correlation coefficient.

💯 TANAGRA 1.4.16 - [Dataset	(tanA6.txt)]			
Tile Diagram Component Wind	dow Help			_ 8 ×
D 🛩 🔲 🗱				
Analysis			Dataset (tanA6.txt)	
🖃 🏢 Dataset (tanA6.txt)			l inear correlation ontions	
🖃 🎇 Define status 1		Databa		
Linear correlation	1 Decembers		Parameters	
1	Parameters		Sort results	
	Execute		Sort by	- I II
	, new		◯ Y attribute name	
		Datasc	◯ X attribute name	
		Compu	Or-value	
		Allocat	⊙  r  - value	
		Data		
		Date	Input list	
		5 attrib	<ul> <li>Target and Input</li> </ul>	
		240 ex	Cross Input	
		_		╴╷┠
Data visualization	Statistics			
Feature construction	eature selection		OK Cancel	Help
PLS	Clustering	•	Spv learning Meta-spv learning	
Spv learning assessment	Scoring		Association	
🛱 Bartlett's test 🛛 🏹	Fisher's test		Group exploration	🕂 Norn
🚰 Brown - Forsythe's test 🛛 👖	Group characte	enzation	⇔Levene's test kore Univariate cont sta	t 📶 One-
<				>

We do not modify the INPUT LIST option here. We validate the parameters specification. By clicking on the VIEW menu, we obtain the following results.

TANAGRA 1.4.16 - [Lin	near correlation 1] Window Help									
) 🖙 🖪 🙀										
Analysis		Line	ar correlatio	n 1						
🖃 🎹 Dataset (tanA6.txt)		Parameters								
😑 🚰 Define status 1	Ci	ross-tab parameters				192				
Linear corre	Sort result	ts yes								
	Sort crite	Sort criterion  r  statistic								
	Input list	Target (Y) and input (X)								
			Results							
	Y	x	r	F <sup>2</sup>	t	Pr(> t )				
	accelerati	ion horesepower	-0.6819	0.4650	-14.3835	0.0000				
	accelerati	ion engine.displacement	-0.5592	0.3127	-10.4070	0.0000				
	accelerati	ion cylinders	-0.5133	0.2635	-9.2280	0.0000				
	accelerati	ion weight	-0.4252	0.1808	-7.2468	0.0000				
		Components								
Data visualization	Statistics	Nonparametric statistic	s Inst	ance select	tion					
Feature construction	Feature selection	Regression	Fac	Factorial analysis						
PLS	Clustering	Spv learning	Meta-spv learning							
Spv learning assessment	Scoring	Association	- Î							
Bartlett's test	👬 Fisher's test	🚺 Group explorat	ion	Linear	correlation	<u> A</u> N				
🔆 Brown - Forsythe's test	👖 Group characterizat	tion 🕁 Levene's test		K More U	Inivariate co	ont stat 📶 O				
<										

HORSEPOWER is the most correlated variable to ACCELERATION. All the correlations are significant at 1% level.

#### Cross-correlations

In order to check the collinearities between the candidates' variables, we want to compute the linear correlations between the INPUT attributes.

We insert again the LINEAR CORRELATION component below DEFINE STATUS 1 into the diagram. We activate the menu PARAMETERS, we always want that the results are sorted, but in addition, we state now that calculation must be applied between the INPUT variables.



Even if the number of INPUT variable is large, the computation is rather quick. The possibility to sort the results according to a statistical criterion is useful.

We obtain the following results.

LINGRA 1.4.16 - LING	window Help	n Zj						
	mildon holp							1.77.1 AT
Analysis				Dar	matare			_
Dataset (tanA6.txt)		Cross-	tab para	meters	dillectris			
🖃 📷 Denne status 1	ation 1	Sort results yes						
Linear correl	ation 2	Sort criterion  r  statistic						
	enecessenten er	Input list	Cross-	-input $(Y \times X)$				
				F	tesults			
		Y		x	r	<b>г</b> ²	t	Pr(> t )
		cylinders		engine.displacement	0.9493	0.9011	46.5716	0.0000
		engine.displacement		weight	0.9320	0.8687	39.6822	0.0000
		engine.displacement		horesepower	0.9078	0.8242	33,3998	0.0000
		cylinders horesepower cylinders		weight	0.9006 0.8112 31.	31.9728	0.0000	
				weight	0.8658	0.7496	26.6940	0.0000
				horesepower	0.8481	0.7194	24.6993	0.0000
			0					
Data visualization	Statist	tes	Nonpa	rametric statistics	Instance	e selection		
Feature construction	Feature se	election		Regression	Factorial analysis			
PLS	Cluster	ring		Spy learning	Meta-spy learning			
5pv learning assessment	Scori	ng		Association				
🛱 Bartlett's test	test 🔠 Grou		Group exploration	K Linear correlation		A Norm		
🚰 Brown - Forsythe's test	🚺 Group cl	haracterizatio	in 🧯	🛱 Levene's test	K.	More Univar	iate cont s	stat 📶 One-v
<				3)				

ACCELERATION does not appear in this calculation. We note that CYLINDERS (number of cylinders) and ENGINE.DISPLACEMENT (engine size) are strongly related. This is not a surprise.

We see also that ENGINE.DISPLACEMENT and HORSEPOWERE are highly correlated.

## Partial correlation

The negative correlation (-0.4252) between ACCELEREATION and WEIGHT seems unusual. It means that the larger is the weight of a car, the faster is acceleration.

This result becomes clear when we see that the WEIGHT is highly correlated to ENGINE.DISPLACEMENT, which is anyway highly correlated to all the variables. We can think that this last variable hides and/or disturbs the correlation between the various variables. We want to proceed again to all analysis by controlling the role of this variable. This is the idea of the partial correlation (see. <a href="http://www2.chass.ncsu.edu/garson/pa765/partialr.htm">http://www2.chass.ncsu.edu/garson/pa765/partialr.htm</a>).

We proceed according to the following steps: (1) we remove to all the variables the effect of the control variable (ENGINE.DISPLACEMENT), by computing the residuals of the linear regression of each variable on this control variable; (2) we perform the above analyses on the residuals corresponding to each variable, where the effect of ENGINE.DISPLACEMENT is removed.

#### Computing the residuals variables

We insert the DEFINE STATUS component in the root of the diagram. We set as TARGET: CYLINDERS, HORSEPOWER, WEIGHT and ACCELERATION; we set as INPUT ENGINE.DISPLACEMENT.



The idea is to remove to the TARGET variables, the effect of the INPUT(s) by computing the residuals of the linear (multiple) regression. We place for this purpose RESIDUAL SCORES component (FEATURE CONSTRUCTION tab). We activate the VIEW menu to obtain the results.



On all regressions, excepted ACCELERATION, the proportion of variance explained is larger than 80%.

### Computation of the partial correlation

In order to compute partial correlations, we add the DEFINE STATUS component into the diagram. We set as TARGET the residuals of ACCELERATION i.e. RS\_ACCELERATION\_1, and as INPUT the other residuals (RS\_CYLINDERS\_1, RS\_HORSEPOWER\_1 et RS-WEIGHT\_1).



Then we insert the LINEAR CORRELATION component. We sort the results according the absolute value of the correlation coefficient.



We obtain the following results.

TANAGRA 1.4.16 - [Lin	ear correlation 3]									
Pile Diagram Component	window Help						*			
Analysis			Line	ar correlatio	n 3					
□ IIII Dataset (tanA6.txt) □ III Dataset (tanA6.txt) □ III Define status 1 □ IIII Linear correl □ IIII Linear correl	ation 1 ation 2	Parameters Cross-tab parameters Sort results yes								
ia - 🏠 Define status 2 ia - 🔅 Residual Scor ia - 🏠 Define sta	res 1 Inpu	Sort criterion Irl statistic Input list Target (Y) and input (X)								
Linear	r correlation 3			Results						
		Y	x	r	r <sup>2</sup>	1	Pr(> t )			
	rs_a	rs_acceleration_1 rs_horesepower_1			0.2512	-8.9359	0.0000			
	rs_a	rs_acceleration_1 rs_weight_1			0.1023	5.2071	0.0000			
	Comp Crea	Computation time : 0 ms. Created at 20/04/2007 11:19:33								
		Cor	mponents							
Data visualization Feature construction PLS Spv learning assessment	Statistics Feature selection Clustering Scoring	Nonparametric statistics stion Regression g Spv learning Association			Instance selection Factorial analysis Meta-spv learning					
🕁 Bartlett's test	🚠 Fisher's test	st 🔠 Group exploration			🖄 Linear correlation 🖉 Norm					
🔆 Brown - Forsythe's test	👖 Group characteriz	racterization 🛱 Levene's test 🛛 🖾 More Univariate cont stat				stat 📶 One-way				
<							>			
1										

HORSEPOWER is again the most correlated variable to ACCELERATION. It seems natural, the larger is the horsepower, and the smaller is the time needed to reach a certain speed. And two cars with the same engine displacement can have different horsepower.

The positive correlation (0.3198) between the ACCELERATION and the WEIGHT is natural. The heavy cars are not very quick. It is well known. The statistical indicator confirms this fact now. The negative original correlation computed above was a spurious correlation.

## Conclusion

In this tutorial, we show how to compute correlations between a large number of variables, and how to sort the results in order to display first the most relevant results.