# 1. Subject

### Detecting outliers and influential points for regression analysis.

The analysis of outliers and influential points is an important step of the regression diagnostics. The goal is to detect (1) the points which are very different to the others (outliers) i.e. they seem do not belong to the analyzed population; or (2) the points that if they are removed (influential points), leads us to a different model. The distinction between these kinds of points is not always obvious.

In this tutorial, we implement several indicators for the analysis of outliers and influential points. To avoid confusion about the definitions of indicators<sup>1</sup> (some indicators are calculated differently from one tool to another), we compare our results with state-of-the-art tool such as SAS and R. In a first step, we give the results described into the SAS documentation. In a second step, we describe the process and the results under Tanagra and R. In conclusion, we note that these tools give the same results.

## 2. Dataset

The dataset comes from the SAS documentation<sup>2</sup>, available on line<sup>3</sup>. The goal is to predict US population size (USPopulation) from the Year (Year) and the squared Year (YearSq). We will mainly focus on the implementation of calculations and comparison of results in this tutorial.

## 3. SAS results

Year	YearSq	Population
1790	3204100	3.929
1800	3240000	5.308
1810	3276100	7.239
1820	3312400	9.638
1830	3348900	12.866
1840	3385600	17.069
1850	3422500	23.191
1860	3459600	31.443
1870	3496900	39.818
1880	3534400	50.155
1890	3572100	62.947
1900	3610000	75.994
1910	3648100	91.972
1920	3686400	105.71
1930	3724900	122.775
1940	3763600	131.669
1950	3802500	151.325
1960	3841600	179.323
1970	3880900	203.211



The used dataset and the results provided by the regression under SAS are the following.

<sup>&</sup>lt;sup>1</sup> <u>http://www-stat.stanford.edu/~jtaylo/courses/stats203/notes/diagnostics.pdf</u>

<sup>&</sup>lt;sup>2</sup> <u>http://v8doc.sas.com/sashtml/stat/chap55/sect33.htm#regprv</u>

<sup>&</sup>lt;sup>3</sup> <u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/USPopulation.xls</u>

	Model: MODEL1 Dependent Variable: Population									
	Output Statistics									
			C	FBETAS						
Obs	Residual	RStudent	H	Ratio	DFFITS	Intercept	Year	YearSq		
1	-1.1094	-0.4972	0.3865	1.8834	-0.3946	-0.2842	0.2810	-0.2779		
2	0.2691	0.1082	0.2501	1.6147	0.0625	0.0376	-0.0370	0.0365		
3	0.9305	0.3561	0.1652	1.4176	0.1584	0.0666	-0.0651	0.0636		
4	0.7908	0.2941	0.1184	1.3531	0.1078	0.0182	-0.0172	0.0161		
5	0.2110	0.0774	0.0983	1.3444	0.0256	-0.0030	0.0033	-0.0035		
6	-0.6629	-0.2431	0.0951	1.3255	-0.0788	0.0296	-0.0302	0.0307		
7	-0.8869	-0.3268	0.1009	1.3214	-0.1095	0.0609	-0.0616	0.0621		
8	-0.2501	-0.0923	0.1095	1.3605	-0.0324	0.0216	-0.0217	0.0218		
9	-0.7593	-0.2820	0.1164	1.3519	-0.1023	0.0743	-0.0745	0.0747		
10	-0.5757	-0.2139	0.1190	1.3650	-0.0786	0.0586	-0.0587	0.0587		
11	0.7938	0.2949	0.1164	1.3499	0.1070	-0.0784	0.0783	-0.0781		
12	1.1492	0.4265	0.1095	1.3144	0.1496	-0.1018	0.1014	-0.1009		
13	3.1664	1.2189	0.1009	1.0168	0.4084	-0.2357	0.2338	-0.2318		
14	1.6746	0.6207	0.0951	1.2430	0.2013	-0.0811	0.0798	-0.0784		
15	2.2406	0.8407	0.0983	1.1724	0.2776	-0.0427	0.0404	-0.0380		
16	-6.6335	-3.1845	0.1184	0.2924	-1.1673	-0.1531	0.1636	-0.1747		
17	-6.0147	-2.8433	0.1652	0.3989	-1.2649	-0.4843	0.4958	-0.5076		
18	1.6770	0.6847	0.2501	1.4757	0.3954	0.2240	-0.2274	0.2308		
19	3.9895	1.9947	0.3865	0.9766	1.5831	1.0902	-1.1025	1.1151		

The procedure for the detection of the outliers and influential points supplies the following table.

The REG Procedure

We observe from the left to the right: (1) the observation number; (2) the residual (deviation between the observed value and the predicted value); (3) the studentized residual; (4) the leverage; (5) the COVRATIO; (6) the DFFITS; (7) the DFBETAS for the coefficients, including the intercept.

## 4. Detection under TANAGRA

### Creating the diagram and importing the dataset

The easiest way to handle the dataset is to open the file into Excel.

We want to launch Tanagra by sending the data using the installed menu with the TANAGRA.XLA add-in<sup>4</sup>.

<sup>&</sup>lt;sup>4</sup> <u>http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html</u>

Tanagra can be launched also from the OOCalc spreadsheet using an add-in. See <u>http://data-mining-tutorials.blogspot.com/2008/10/000calc-file-handling-using-add-in.html</u>

M 🖾	licrosof	it Exce	l - USPop	oulati	on.csv 🧲										
	<u>Fi</u> chier (	<u>E</u> dition	<u>A</u> ffichage	Inser	tion Forma <u>t (</u>	<u>)</u> utils <u>D</u> onr	nées	Fe <u>n</u> être	<u>?</u> T	anagra	Sipina				8 ×
l D	🚔 🔲	4	a #59	X 🗈	n 🙉 🛷 🖂	) + CH +	0	$\Sigma f_{x}$	<b>R</b>	Al Z.		<u>7</u> 100%	- 6	ก_ 💩 -	»
				Yea	r			· ·		<b>-</b> , -,		v		<b>a</b> • ]	•
	A	<u>ــــــــــــــــــــــــــــــــــــ</u>	B	<u> </u>	с	D		E			F	G		Н	
1	Year	-	YearSq		Population						-				
Z		1790	3204	100	3.929										
3		1800	3240	0000	5.308										_
4	]	1810	3276	5100	7.239										
5		1820	3312	2400	9.638										
6		1830	3348	3900	12.866										
7		1840	3385	5600	17.069										_
8		1850	3422	2500	23.191					_					_
9		1860	3459	9600	31.443										_
10		1870	3496	5900	39.818										_
11		1880	3534	1400	50.155					_		_			_
12		1890	3572	2100	62.947					_					_
13		1900	3610	2100	/5.994					_					_
14		1910	2694	400	91.972										_
16		1920	372/	1900	122 775					_					_
17		1940	3763	3600	131 669					_					_
18		1950	3802	2500	151.325										_
19	1	1960	3841	600	179.323										_
20		1970	3880	0900	203.211										
21															
22															
<b>H</b> 4	► ► ►	USPo	pulation ,	/					•						•
	isin + <table-cell></table-cell>	6	Formes au	tomati	ques 🗸 🔪 🍾			4 🙎	<u></u>	- 🚄 -	· <u>A</u> -	= = =	‡ 🖬	<b>-</b>	
Prêt								Somn	ne=6	72476	15.6	N	IUM [		

We selected the data range, and then we click on the TANAGRA/EXECUTE TANAGRA.

N 🔊	licrosoft Exce	l - USPopulat						_	
	Fichier Edition	Affichage Inse	rtion Forma <u>t</u> (	utils <u>D</u> onnées	Fenêtre ? "	Tanagra Sipina	э		_ 8 ×
			h 📾 🛷 🔽		T E	A1	<b>a</b> 100%	- 2 8	, »
1 LL					, ∠ <i>]*</i> <u>⊌€</u>	2+	100 /0	·	<u>-</u>
_	<u> </u>	= 10	ar	-	-				
	A	В	С	D	E		G	H	
1	Year	YearSq	Population				(1)		
2	1/90	3204100	3.929	-			(1)		
3	1800	3240000	5.308				<b>`</b>		
4	1810	3276100	7.239						
5	1820	3312400	9.638						
6	1830	3348900	12.866						
/	1840	3385600	17.069						
8	1850	3422500	23.191						
9	1860	3459 FX	ecute Tanagr					X	
10	18/0	3496		×					ŝ.
11	1880	3534		(* 1. K. 11	e., , ,				
12	1890	3572	Dataset ranç	ge (including the	name or the at	tributes first	row):		
13	1900	3610	\$A\$1:\$C\$	20					
14	1910	3648			3 <u></u>			-	
15	1920	3666		<ul> <li>•</li> </ul>		or	Cancel		
10	1950	3724	- C2				cancor		
1/	1940	3/63	(4	· · ·					¥
10	1950	2841600	170 222						
20	1960	2880000	202 211						
20	1970	3880900	205.211		-	-			
22									-
14	ISPO	nulation /	1		1		1		
		policically						<b>.</b>	
Des	asin + 🗟 🕝	Formes automat	tiques 🔹 🔪 🍾		4 🖳 🖄	• 🛃 • 🗛 ·	• <b>=</b> = £	Ξ∎ 🗊 ₊	
Prêt	t				Somme=6	57247645.6	1	NUM MUK	

We check the range of selected cells. We validate by clicking on the OK button.

Tanagra is automatically launched. We check that we have 19 instances and 3 variables.

🌋 TANAGRA 1.4.17 - [Da	TANAGRA 1.4.17 - [Dataset (tanC5.txt)]								
Tile Diagram Component	t Window H	lelp			_ 8 ×				
🗅 📽 🖬 🔯									
Analysis			Dataset (t	anC5.txt)	<u>^</u>				
			Parameters						
		Database : C:\DOCUME~1\Waison\LOCALS~1\Temp\tanC5.txt							
Results									
Download information									
		Datacourses a	recording						
	Datasource processing								
		Allocated mem	orv 3KB						
		nilood cod nioni	ory one		_				
		Dataset	description						
		3 attribute(s)							
		19 example(s)							
			· · · · ·		~				
			Components						
Data visualization	Sta	tistics	Nonparametric statistics	Instance selection					
Feature construction	Feature	e selection	Regression	Factorial analysis					
PLS	Clu	stering	Spv learning	Meta-spv learning					
Spv learning assessment	So	oring	Association						

#### Linear regression

We want to predict the values of POPULATION using YEAR and YEARSQ. We specify the type of the variables using the DEFINE STATUS component. We set POPULATION as TARGET; YEAR and YEARSQ as INPUT.



Then we insert the MULTIPLE LINEAR REGRESSION component (REGRESSION tab) into the diagram. We click on the VIEW menu in order to obtain the results.

File Diagram Component	Window H	elp							- 7
) 📽 🔛  🍇									
Analysis					Results				
Dataset (tanC5.txt) Dataset (tanC5.txt)		Global r	esults						
🖳 🔀 Multiple lines	ar regress	Endogenous a	ttribute		Population			1	
1		Examples		19				1.1	
		R²			0.998279			÷	
Adjusted-R <sup>2</sup>									
Sigma error					2.781024				
F-Test (2,16)				4641.7210	0.000000)			1	
		Analysis	of variance	•				7	
		Source	xSS	d.f.	×MS	F	p-value		
		Regression	71799.0158	2	35899.5079	4641.7210	0.0000		
		Residual	123.7455	16	7.7341			1.1	
		Total	71922.7614	18				- i - i -	
$\mathbf{A}$		Cooffici	ante					÷	
		coerrick	ents						
		<		Ш		J			>
			Comp	onents					
Data visualization	Sta	tistics	Nonparametric	statistics	Instance sel	ection	Feature const	ruction	
Feature selection	Reg	ression	Factorial an	alysis	PLS		Clusteri	ng	
Spv learning	Meta-sp	w learning	Spv learning ass	sessment	Scoring	2	Associati	ion	
🖞 Backward Elimination Reg	g 🛛 🕌 Forwa	rd Entry Regres	sion 🛛 🖄 Outlier D	etection					
L DfBetas	📿 Multip	le linear regres	sion 📩 Regressio	on tree					

We find the same results than SAS.

#### Detection of outliers and influential points (1)

We insert now the diagnostic component OUTLIER DETECTION (REGRESSION tab) into the diagram. We click on the VIEW menu. There are two tabs into the visualization window. REPORT summarizes the number of suspicious observations identified according to the indicators.

💯 TANAGRA 1.4.17 - [Outlier Det	tection 1]						
Tile Diagram Component Window	Help						_ 8 ×
🗅 📽 🔲 🗱							
Analysis	Report <b>A</b> lue	s					
Dataset (tanC5.txt)				Outling Datas	tion 1		~
🖃 🎎 Define status 1				Paramete	0011 1 re		
🛓 📝 Multiple linear regres	S Parameters			r ar anicce	.,		_
🦾 🔀 Outlier Detection	Sig. Level	0.05	00				
1							_
				Results			
	Outliers	and influ	uantial no	inte dat	ection for	regression	
	Outtiers		iantiat po	units dei		regression	=
	Statistic	Lower bound	Upper bound	# detected	- <b>1</b>		
	Leverage	-	0.3158	2			
	RStandard	-	-	0			
	RStudent	-2.0000	2.0000	2	~		
	DFFITS	-0.7947	0.7947	3	- <b>1</b>		
	Cook's D	-	0.2500	3			
	COVRATIO	0.5263	1.4737	5			
	Computation ti	me:0ms.					~
		G	omponents				
Data visualization	Statistic	Nonparame	tric statistics	Instan	ce selection	Feature construction	
Feature selection	Regression	Factoria	al analysis		PLS	Clustering	
Spv learning Met	a-spv learning	Spv learnin	g assessment	S	coring	Association	
🛃 Backward Elimination Reg 🛛 🐇 Fo	rward Entry Regres	si 🖉 🖟 Outli	er Detection	>			
🖾 DfBetas 🖾 Mu	ultiple linear regres	sion 🧟 Regre	ession tree				
L							

VALUES tab gives the computed values for each individual. The suspicious values are underlined.

💯 TANAGRA 1.4.17 - [01	utlier Detec	tion 1]							
Tile Diagram Component	t Window	Help						_ 8 ×	
🗅 💕 🖪  🎇									
Autoria Values									
Anaiysis		Report Value	° <b>4</b>	,		,		, ,	
🖃 🏢 Dataset (tanC5.txt)		Statistic	Leverage	RStandard	RStudent	DFFITS	Cook's D	COVRATIO	
📄 🚰 Define status 1		Lower Bound	-	-	-2.0000	-0.7947	-	0.5263	
📥 📈 Multiple line	ear regress	Upper Bound	0.3158	-	2.0000	0.7947	0.2500	1.4737	
 ™_Outlier I	Detection :	1	0.3864661	-0.5093092	-0.4971832	-0.3945962	0.0544647	1.8834404	
	Deceetion	2	0.2501253	0.1117436	0.1082375	0.0625118	0.0013883	1.6146570	
		3	0.1652071	0.3662178	0.3560844	0.1584083	0.0088472	1.4175564	
		4	0.1184432	0.3028761	0.2941029	0.1078026	0.0041084	1.3531454	
		5	0.0983341	0.0799205	0.0773982	0.0255599	0.0002322	1.3443745	
		6	0.0951496	-0.2505731	-0.2430938	-0.0788295	0.0022008	1.3255210	
		7	0.1009288	-0.3363375	-0.3268147	-0.1094995	0.0042330	1.3214409	
		8	0.1094796	-0.0952824	-0.0922830	-0.0323568	0.0003720	1.3605133	
		9	0.1163792	-0.2904633	-0.2819843	-0.1023363	0.0037040	1.3518605	
		10	0.1189739	-0.2205512	-0.2138731	-0.0785937	0.0021896	1.3649929	
		11	0.1163792	0.3036422	0.2948510	0.1070058	0.0040477	1.3498664	
		12	0.1094796	0.4378799	0.4265388	0.1495558	0.0078574	1.3144215	
		13	0.1009288	1.2007898	1.2188778	0.4083858	0.0539553	1.0168259	
		14	0.0951496	0.6330077	0.6207291	0.2012877	0.0140452	1.2429821	
		15	0.0983341	0.8484689	0.8406562	0.2776180	0.0261703	1.1723552	
		16	0.1184432	-2.5404584	-3.1845305	-1.1672806	0.2890431	0.2923815	
		17	0.1652071	-2.3671196	-2.8432634	-1.2648591	0.3696317	0.3988777	
		18	0.2501253	0.6963493	0.6846924	0.3954394	0.0539140	1.4757102	
		19	0.3864661	1.8314630	1.9946725	1.5830990	0.7042842	0.9766091	
			Co	mponents					
Data visualization	St	atistics	Nonparame	tric statistics	Instance	selection	Feature c	onstruction	
Feature selection	Re	gression	Factori	al analysis	P	LS	Clust	tering	
Spv learning	Meta-s	spv learning	Spv learnin	g assessment	Sec	oring	Asso	ciation	
Z Backward Elimination Re	ee 🔀 Forw	ard Entry Regres	ssion 🚺 Outli	er Detection					
Deckward Lummation Reg     Image Linky Regression       DfBetas     Multiple linear regression									
L									

The choice of the cut values is very important. It enables to highlight the suspicious individuals. We show below the cut values reported into to the literature: n is the dataset size, p is the number of estimated parameters (number of descriptors + 1 for a regression with intercept).

Indicator	Cut value
Leverage	2*p/n
RStandard	-
RStudent	2
DFFITS	2*SQRT(p/n)
Cook's D	4/(n-p)
COVRATIO-1	3*p/n

The detailed results can be copied into a spreadsheet, if we want for instance to analyze them deeply. We use the COMPONENT/COPY RESULTS menu when the VALUES tab is activated.

By comparing our results with SAS, we observe that we obtain the same results. We will note below that R gives also the same values for all the indicators.

### Detection of outliers and influential points -- DFBETAS (2)

The second tool is used to evaluate the influence of each observation on each estimated coefficient. We use a second component for these calculations because it does not seem desirable to expand tabs into the visualization window, especially since the results are different: OUTLIER DETECTION evaluates the overall influence of each observation; DFBETAS evaluates the influence of each observation on each estimated coefficient. We add the DFBETAS component (REGRESSION tab) under the linear regression. We click on the VIEW menu in order to obtain the results.

💯 TANAGRA 1.4.17 - [Dff	Betas 1]								
Tile Diagram Component	Window Help			- 8 ×					
D 📽 🖪 🔤									
Analysis	Report DfBeta	38							
Dataset (tanC5.txt)	ar regress totaction : Cut value = (	o for Regression Di	agnostics						
ji DfBetas 1	# detected a	ccording to coefficients							
		≡							
	Year	2							
	YearSq	2							
<				~					
	C	components							
Data visualization	Statistics	Nonparametric statistics	Instance selection						
Feature construction	Feature selection	Regression	Factorial analysis						
PLS	Clustering	Spv learning	Meta-spv learning						
Spv learning assessment	Scoring	Association							
Kackwary Elimination Reg       Kackwary Elimination Reg         DrBetas       Multiple linear regression         Kackwary Elimination Reg       Kackwary Elimination         Multiple linear regression       Kackwary Regression tree									
				.:					

Again, two tabs are available: the summary into the first, the detailed results into the second.

💯 TANAGRA 1.4.17 - [Di	🂯 TANAGRA 1.4.17 - [DfBetas 1]									
🏆 File Diagram Componen	t Window	Help					- 8 ×			
D 📽 🖪   👫										
Analysis		Report DfBet	Report DfBetas							
🖃 🏢 Dataset (tanC5.txt)			Intercept	Year	YearSq					
🚊 🔂 Define status 1		1	-0.2841626	0.2810148	-0.2779339					
🔲 🕹 Multiple lip	Aultiple linear regress			-0.0370272	0.0364868					
		3	0.0665886	-0.0650528	0.0635757					
Uuther	Detection	4	0.0182289	-0.0171561	0.0161361					
🔤 📝 DfBetas	1	5	-0.0030405	0.0032773	-0.0034991					
		6	0.0295513	-0.0301596	0.0307176					
		7	0.0609386	-0.0615767	0.0621438					
		8	0.0215926	-0.0217215	0.0218297					
		9	0.0742868	-0.0745210	0.0746903					
		10	0.0586297	-0.0586841	0.0586891					
		11	-0.0783587	0.0782622	-0.0780983					
		12	-0.1017689	0.1013811	-0.1008983					
		13	-0.2356621	0.2338457	-0.2317696					
		14	-0.0811349	0.0798479	-0.0784361					
		15	-0.0426522	0.0404071	-0.0380049					
		16	-0.1531205	0.1636479	-0.1747212					
		17	-0.4843196	0.4957595	-0.5076395					
		18	0.2239904	-0.2273530	0.2308097					
<		19	1.0901980	-1.1025077	1.1150558					
		J.,	Componen	ts						
Data visualization	St	atistics	Nonparame	tric statistics	Instance	selection				
Feature construction	Featu	re selection	Reg	ression	Factoria	l analysis				
PLS	Cl	ustering	Spv l	earning	Meta-sp	/ learning				
Spv learning assessment	S	coring	Asso	ciation						
🖉 Backward Elimination Ro 🖄 DfBetas	〈Backward Elimination Reg  人 Forward Entry Regression  注 Outlier Detection 注 DfBetas									

The threshold value is **2/SQRT(n)**. The suspicious values are highlighted. Here also, we can copy the values into a spreadsheet (COMPONENT / COPY RESULTS).

## 5. Detection under R

The R software is free tool (<u>http://www.r-project.org/</u>). It has an excellent reputation among statisticians, quite justified in light of his many qualities. Its library of methods is impressive. The other reason for which we use R is that the techniques we study in this tutorial are implemented as standard by leading experts. The calculations implemented in R are often used as reference <sup>5</sup>.

### Launching R

We use the 2.5.0 version in this tutorial. We launch the tool.



#### Importing the dataset

We use the following commands in order to import the data file (USPopulation.CSV).

We display the dataset here.

<sup>&</sup>lt;sup>5</sup> I confess to being a true novice with regard to R. The manipulations described in this tutorial are simplified to the extreme. Perhaps they are not optimal. My goal was to obtain results comparable to those of SAS and TANAGRA to verify the accuracy of the calculations.

R RGui - [R Console]	
R Fichier Edition Voir Misc Packages Fenêtres Aide	_ 8 ×
	~
Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide	
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.	
Tapez 'q()' pour quitter R.	
N wedere ( weed weedfile HUCDevelopies week erself.)	
<pre>&gt; usdata &lt;- read.csv(file="05Population.csv",sep=";") &gt; usdata</pre>	
Year YearSg Population	
1 1790 3204100 3.929	_
2 1800 3240000 5.308	
3 1810 3276100 7.239	
4 1820 3312400 9.638	
5 1830 3348900 12.866	
6 1840 3385600 17.069	
7 1850 3422500 23.191	
8 1860 3459600 31.443	
9 1870 3496900 39.818	
10 1880 3534400 50.155	-
11 1890 3572100 62.947	-
12 1900 3610000 75.994	
13 1910 3648100 91.972	
14 1920 3686400 105.710	
15 1930 3724900 122.775	
16 1940 3763600 131.669	
17 1950 3802500 151.325	
18 1960 3841600 179.323	
	~
<u>&lt;</u>	>

#### Linear regression

The Im() command performs a linear regression. We set the model in the "usmodel" object.

ℝ RGui - [R Console]	
R Fichier Edition Voir Misc Packages Fenêtres Aide	_ 8 ×
2 C C C C C C C C C C C C C C C C C C C	
	^
16 1940 3763600 131.669	_
18 1960 3841600 179.323	
19 1970 3880900 203.211	
<pre>&gt; usmodel &lt;- lm(Population ~ Year + YearSq, data = usdata) &gt; usmound(usmodel)</pre>	
> summary(usmodel)	
Call:	
lm(formula = Population ~ Year + YearSq, data = usdata)	
Residuals:	
Min 1Q Median 3Q Max	
-6.6335 -0.7111 0.2691 1.4119 3.9895	
Coefficients:	
Estimate Std. Error t value Pr(> t )	
(Intercept) 2.045e+04 8.435e+02 24.25 4.83e-14 ***	
Year -2.278e+01 8.978e-01 -25.37 2.38e-14 ***	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	=
Desidual standard error: 2 781 on 16 degrees of freedom	
Multiple R-Squared: 0.9983, Adjusted R-squared: 0.9981	
F-statistic: 4642 on 2 and 16 DF, p-value: < 2.2e-16	
2 <b>•</b>	*
<	>

#### Measures for detecting outliers and influential points

We obtain the measures for detection influential points by using the **influence.measures()** command. We display the obtained values below. The command computes, from the left to the right: DFBETAS, DFFITS, COVRATIO, Cook's distance, leverage.

🥂 RGui - [R Console]						
R Fichier Edition Voir Misc Packages Fenêtres Aide	- 8 ×					
A seve image/#Ci\\Dreamen Files\\D\\D 2 E O\\USDenulation Dote#\						
<pre>&gt; save.image("C:\\rrogram Flee(\k\k\k-2.3.0\\05F0pulation.kbata") &gt; diad[&lt;- influence.measures(usmodel)</pre>						
> diag1						
Influence measures of						
lm(formula = Population ~ Year + YearSq, data = usdata) :						
dfb.1_dfb.Year dfb.YrSq dffit cov.r cook.d hat inf						
1 -0.28416 0.28102 -0.2779 -0.3946 1.883 0.054465 0.3865 *						
7 0.06094 -0.06158 0.0621 -0.1095 1.321 0.004233 0.1009						
8 0.02159 -0.02172 0.0218 -0.0324 1.361 0.000372 0.1095						
9 0.07429 -0.07452 0.0747 -0.1023 1.352 0.003704 0.1164						
10 0.05863 -0.05868 0.0587 -0.0786 1.365 0.002190 0.1190						
11 -0.07836 0.07826 -0.0781 0.1070 1.350 0.004048 0.1164						
12 -0.10177 0.10138 -0.1009 0.1496 1.314 0.007857 0.1095	_					
13 -0.23566 0.23385 -0.2318 0.4084 1.017 0.053955 0.1009						
14 -0.08113 0.07985 -0.0784 0.2013 1.243 0.014045 0.0951						
15 -0.04265 0.04041 -0.0380 0.2776 1.172 0.026170 0.0983						
16 -0.15312 0.16365 -0.1747 -1.1673 0.292 0.289044 0.1184 *						
17 -0.48432 0.49576 -0.5076 -1.2649 0.399 0.369632 0.1652 *						
18 0.22399 -0.22735 0.2308 0.3954 1.476 0.053914 0.2501						
19 1.09020 -1.10251 1.1151 1.5831 0.977 0.704283 0.3865 *						
	~					
<u>&lt;</u>						

#### RStandard and RStudent

To obtain the standardized residuals (RSTANDARD) and the studentized residuals (RSTUDENT), we perform additional operations.

🥂 RGui - [R Console]	
R Fichier Edition Voir Misc Packages Fenêtres Aide	- 8 ×
18 0.22399 -0.22735 0.2308 0.3954 1.476 0.053914 0.2501	<u> </u>
19 1.09020 -1.10251 1.1151 1.5831 0.977 0.704283 0.3865 *	
> r1 <- rstandard(usmodel)	
> r2 <- rstudent(usmodel)	
> diag2 <= data.frame(f1,f2)	
r1 r2	
1 -0.50931083 -0.49718482	
2 0.11174203 0.10823599	
3 0.36621645 0.35608307	
4 0.30287504 0.29410180 5 0.07991928 0.07739696	
6 -0.25057420 -0.24309487	
7 -0.33633849 -0.32681572	
8 -0.09528372 -0.09228425	
9 -0.29046458 -0.28198557	
10 -0.22055173 -0.21387365	
12 0.43787740 0.42653638	
13 1.20078831 1.21887626	
14 0.63300676 0.62072807	
15 0.84846690 0.84065419	
16 -2.54046169 -3.18453720	_
18 0.69634825 0.68469135	
19 1.83146126 1.99467009	
>	
	×
	7

All the computed values are the same than those of SAS and Tanagra.

# 6. Conclusion

The analysis of the residuals is a key step in the linear regression diagnostic process. There are the detection of outliers and influential that we described in this tutorial. But other procedures are also important. There are the simple graphical representation of the residuals; the checking of the normality of residuals that we can implement with intuitive graphics such as QQ-plots; we can check the heteroscedasticity and autocorrelation (when the data are longitudinal), etc.

We use the threshold values in order to highlight the suspicious individuals according to the indicators. But these thresholds are not irrevocable. In the most of situations, it is more convenient to sorting the data according to the indicators in order to highlight the individuals which differ strongly from the others. We can use also very simplistic approaches in order to detect the remarkable values (e.g. <u>http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm</u>).

In the table below, we sort the dataset according the RSTUDENT. We observe that the instance n°16 and n°17 are problematic. But we note also that the instance n°19 is suspicious, it needs deeper analysis.

Statistic	Leverage	RStandard	RStudent	DFFITS	Cook's D	COVRATIO
16	0.11844316	-2.54045844	-3.1845305	-1.16728055	0.2890431	0.29238147
17	0.16520713	-2.36711955	-2.84326339	-1.26485908	0.36963168	0.39887768
1	0.38646615	-0.50930923	-0.49718323	-0.39459622	0.05446466	1.88344038
7	0.1009288	-0.33633751	-0.32681474	-0.10949951	0.00423302	1.32144094
9	0.11637919	-0.2904633	-0.28198433	-0.10233629	0.003704	1.35186052
6	0.09514964	-0.25057307	-0.24309376	-0.07882955	0.00220079	1.32552099
10	0.11897392	-0.22055119	-0.21387313	-0.0785937	0.00218958	1.36499286
8	0.10947959	-0.09528238	-0.09228296	-0.03235684	0.00037204	1.36051333
5	0.09833407	0.07992052	0.07739815	0.02555994	0.0002322	1.34437454
2	0.25012529	0.11174358	0.10823749	0.06251182	0.00138833	1.61465704
4	0.11844316	0.30287614	0.29410288	0.10780257	0.00410836	1.35314536
11	0.11637919	0.30364218	0.29485103	0.10700581	0.00404774	1.34986639
3	0.16520713	0.36621782	0.35608441	0.15840831	0.00884724	1.41755641
12	0.10947959	0.43787986	0.4265388	0.14955577	0.00785738	1.31442153
14	0.09514964	0.63300771	0.62072909	0.20128772	0.01404517	1.24298215
18	0.25012529	0.69634926	0.68469238	0.39543939	0.05391404	1.47571015
15	0.09833407	0.8484689	0.84065622	0.27761799	0.02617032	1.17235518
13	0.1009288	1.20078981	1.21887779	0.40838584	0.05395525	1.01682591
19	0.38646615	1.83146298	1.99467254	1.58309901	0.70428419	0.97660911

Median	0.10823749
Q1	-0.26253905
Q3	0.523633945
IQR	0.78617299
Lower Outer Fence	-2.62105802
Lower Inner Fence	-1.44179853
Upper Inner Fence	1.70289343
Upper Outer Fence	2.882152915