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[1] (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[2], available on line[3]. 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

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

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 REG Procedure
Model: MODEL1
Dependent Variable: Population

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	71799	35900	4641.72	<.0001
Error	16	123.74557	7.73410		
Corrected Total	18	71923			

Root MSE	2.78102	R-Square	0.9983
Dependent Mean	69.76747	Adj R-Sq	0.9981
Coeff Var	3.98613		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20450	843.47533	24.25	<.0001
Year	1	-22.78061	0.89785	-25.37	<.0001
YearSq	1	0.00635	0.00023877	26.58	<.0001

[1] http://www-stat.stanford.edu/~jtaylo/courses/stats203/notes/diagnostics.pdf

[2] http://v8doc.sas.com/sashtml/stat/chap55/sect33.htm#regprv

[3] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/USPopulation.xls

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

The REG Procedure
Model: MODEL1
Dependent Variable: Population

Output Statistics								
						DFBETAS		
Obs	Residual	RStudent	Hat Diag H	Cov 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

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.

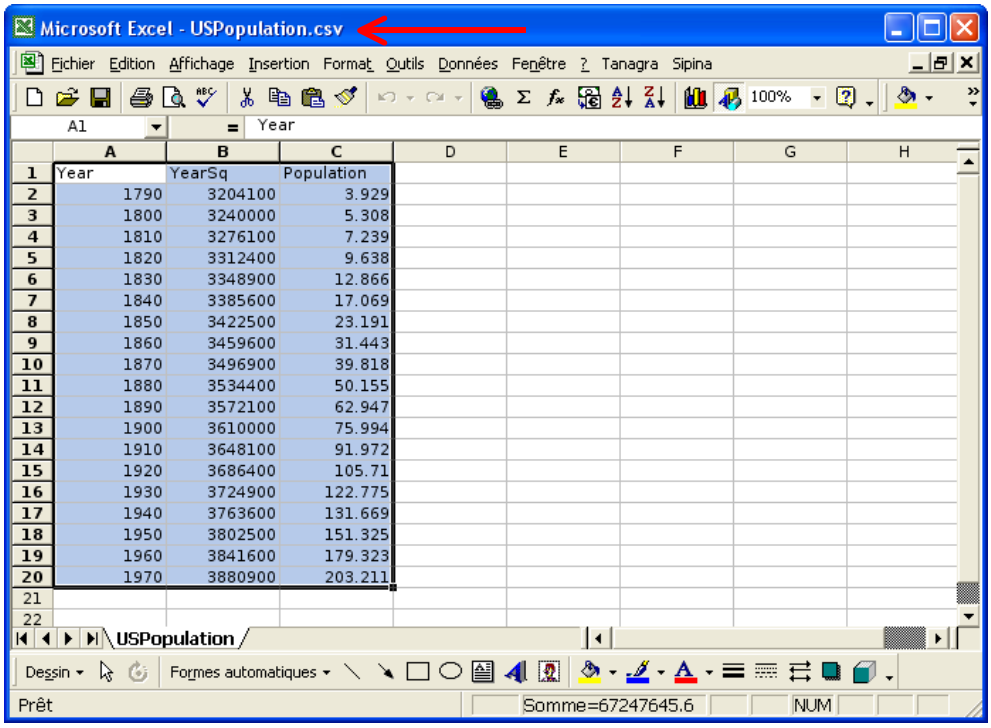
4. Detection under TANAGRA

Creating the diagram and importing the dataset

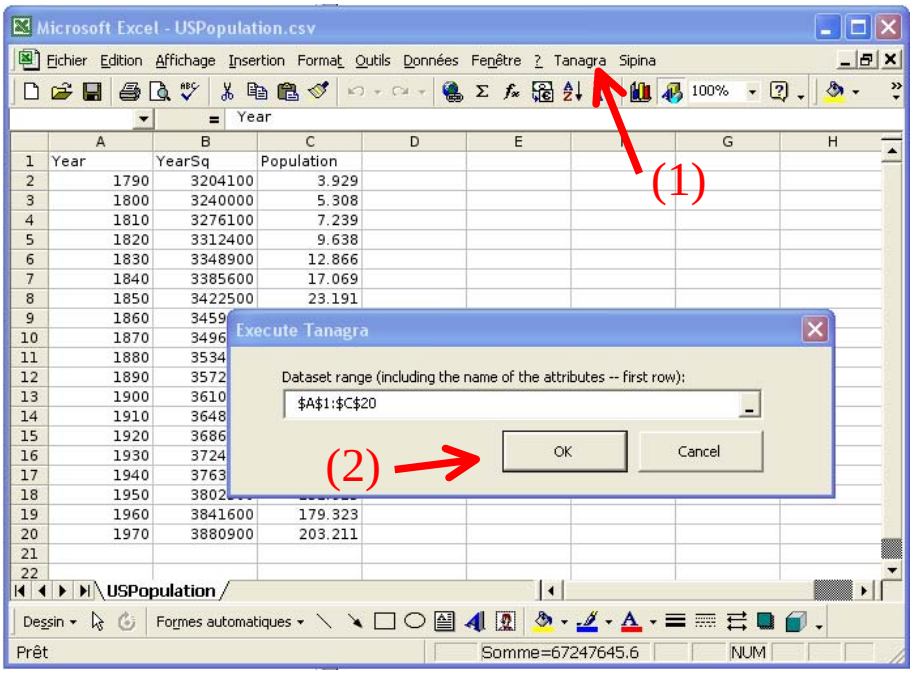
The easiest way to handle the dataset is to open the file into Excel.

We want to launch Tanagra by sending the data using the installed menu with the TANAGRA.XLA add-in[4].

[4] http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html
Tanagra can be launched also from the OOCalc spreadsheet using an add-in. See http://data-mining-tutorials.blogspot.com/2008/10/ooocalc-file-handling-using-add-in.html

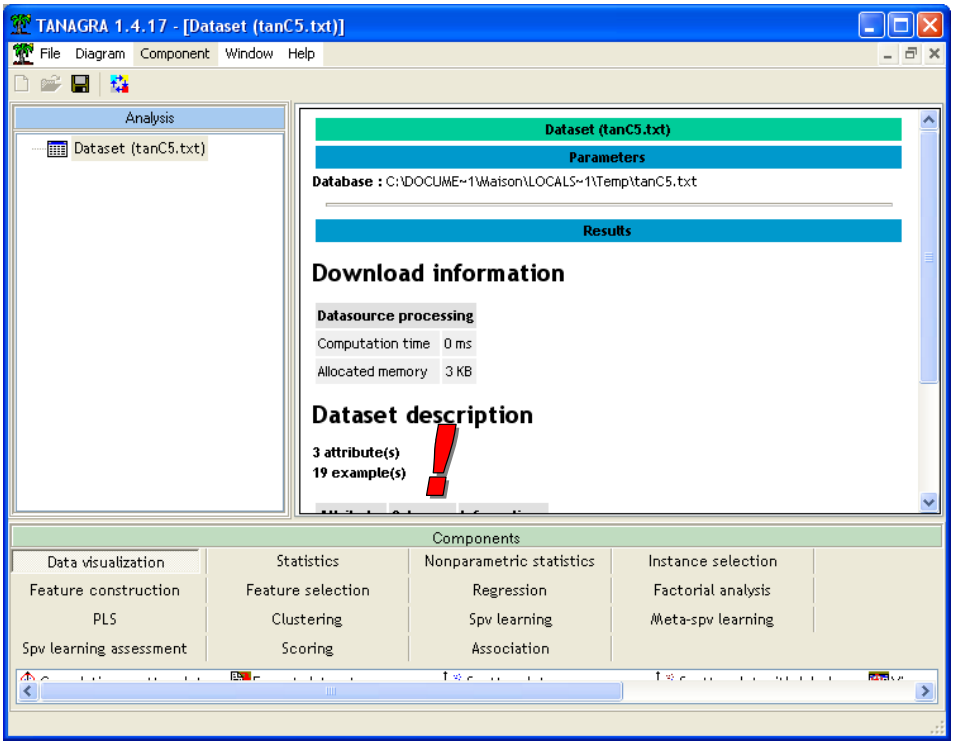


We selected the data range, and then we click on the TANAGRA/EXECUTE TANAGRA.



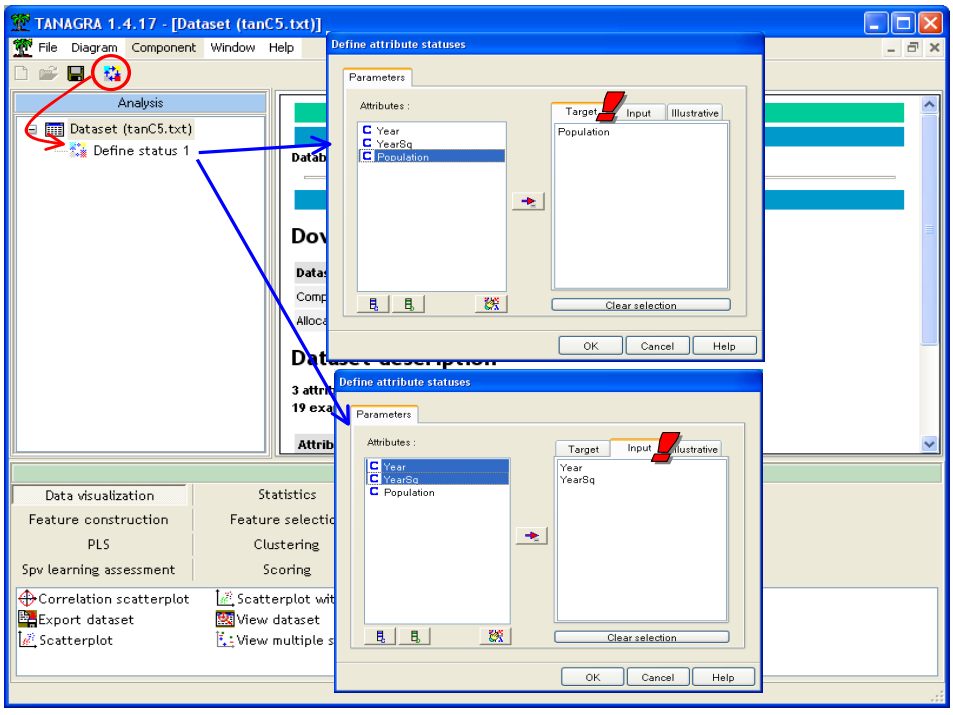
We check the range of selected cells. We validate by clicking on the OK button.

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

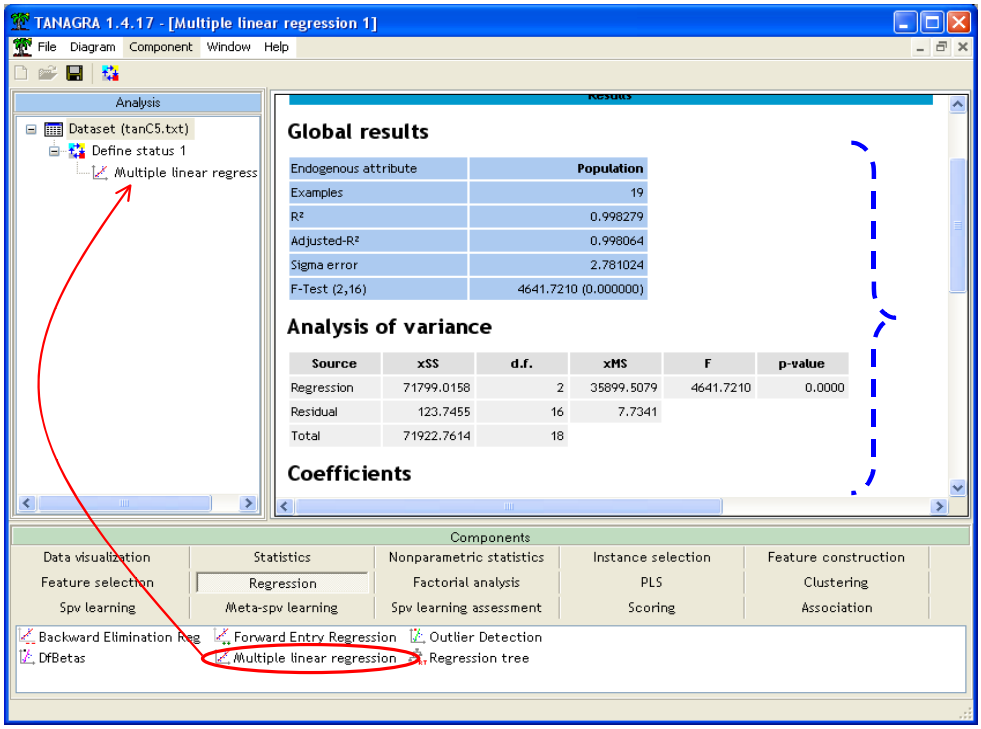


Linear regression

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.



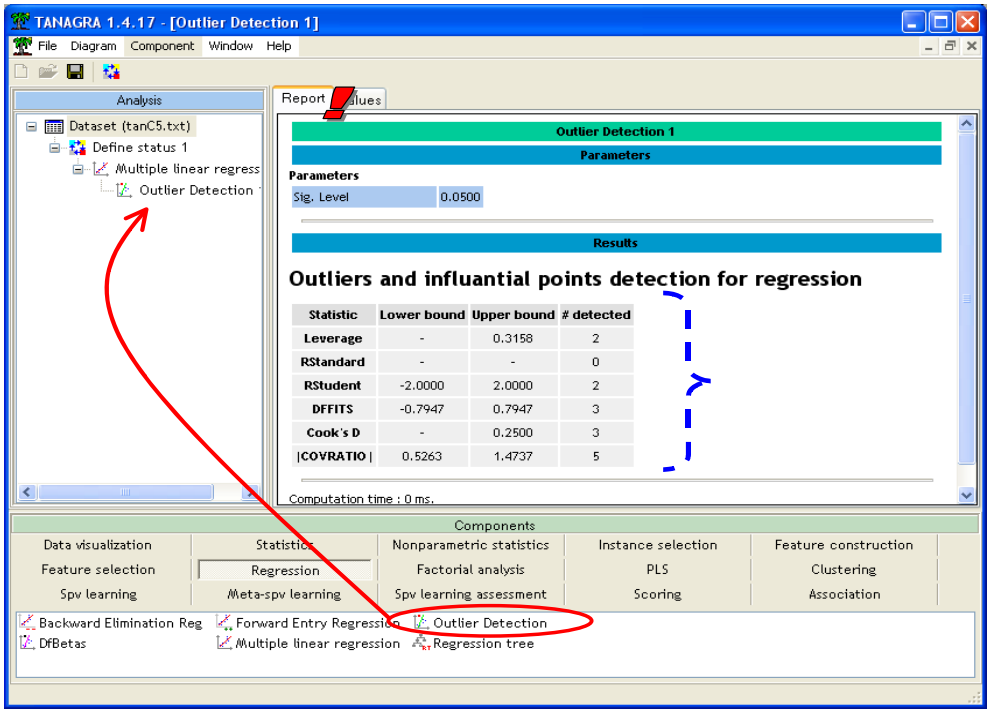
Then we insert the MULTIPLE LINEAR REGRESSION component (REGRESSION tab) into the diagram. We click on the VIEW menu in order to obtain the results.



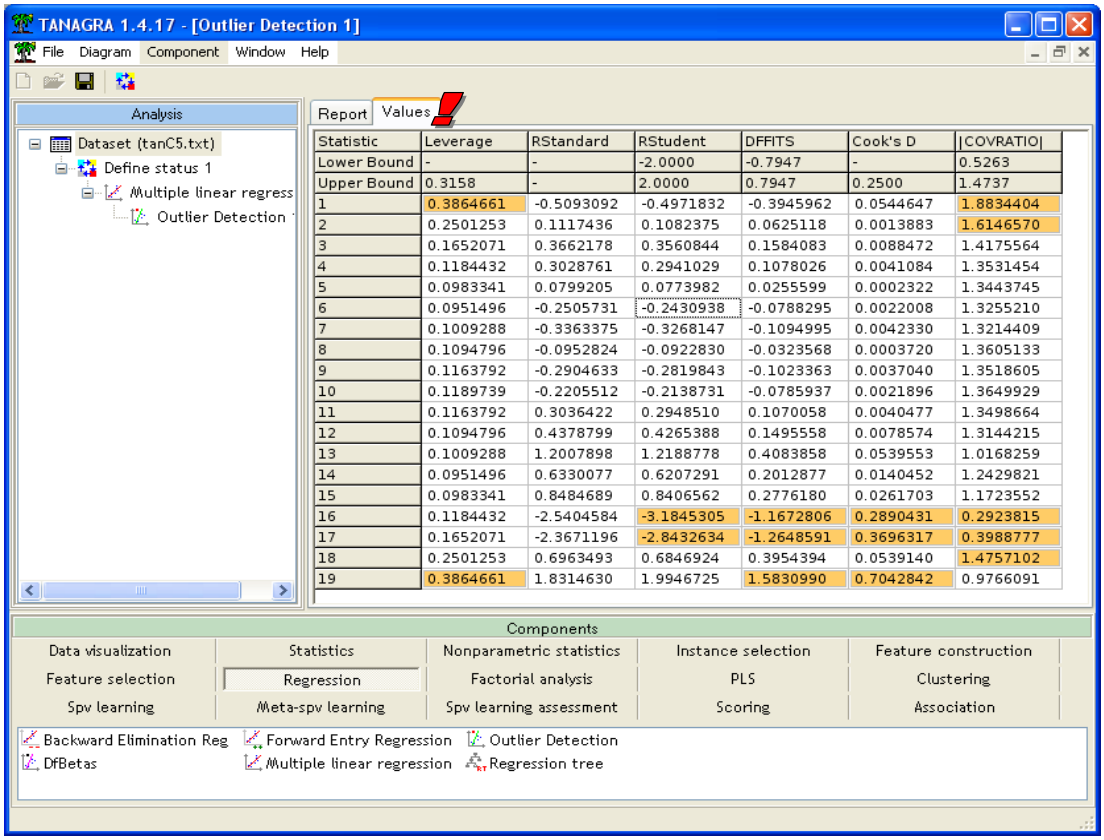
We find the same results than SAS.

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.



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



Statistic	Leverage	RStandard	RStudent	DFFITS	Cook's D	|COVRATIO|
Lower Bound	-	-	-2.0000	-0.7947	-	0.5263
Upper Bound	0.3158	-	2.0000	0.7947	0.2500	1.4737
1	0.3864661	-0.5093092	-0.4971832	-0.3945962	0.0544647	1.8834404
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

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

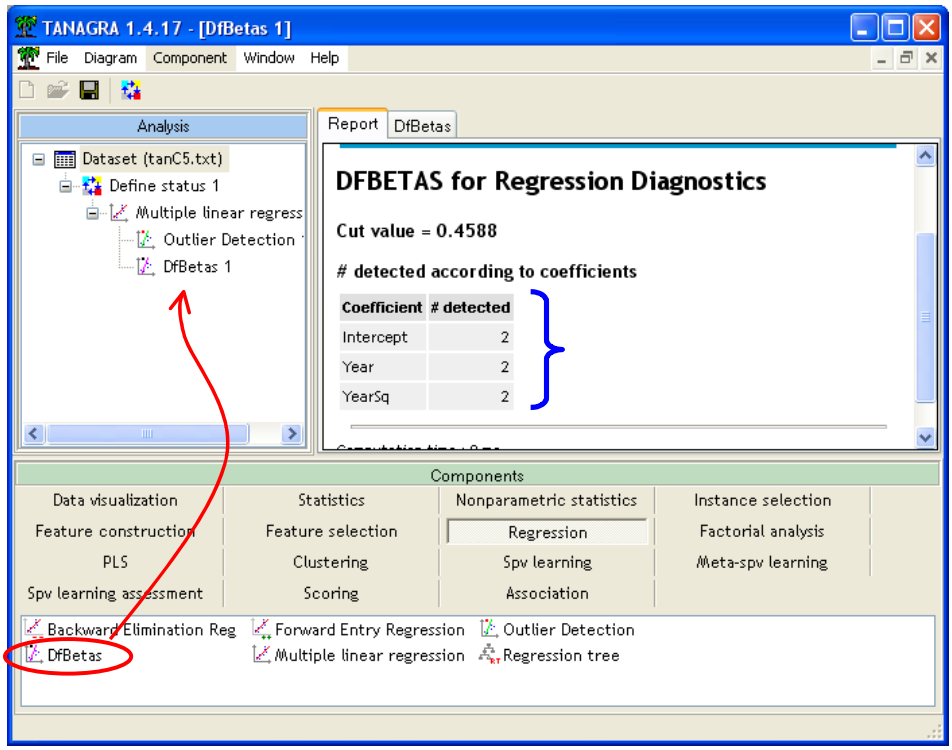
The detailed results can be copied into a spreadsheet, if we want for instance to analyze them deeply. We use the COMPONENT/COPY RESULTS menu when the VALUES tab is activated.

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.

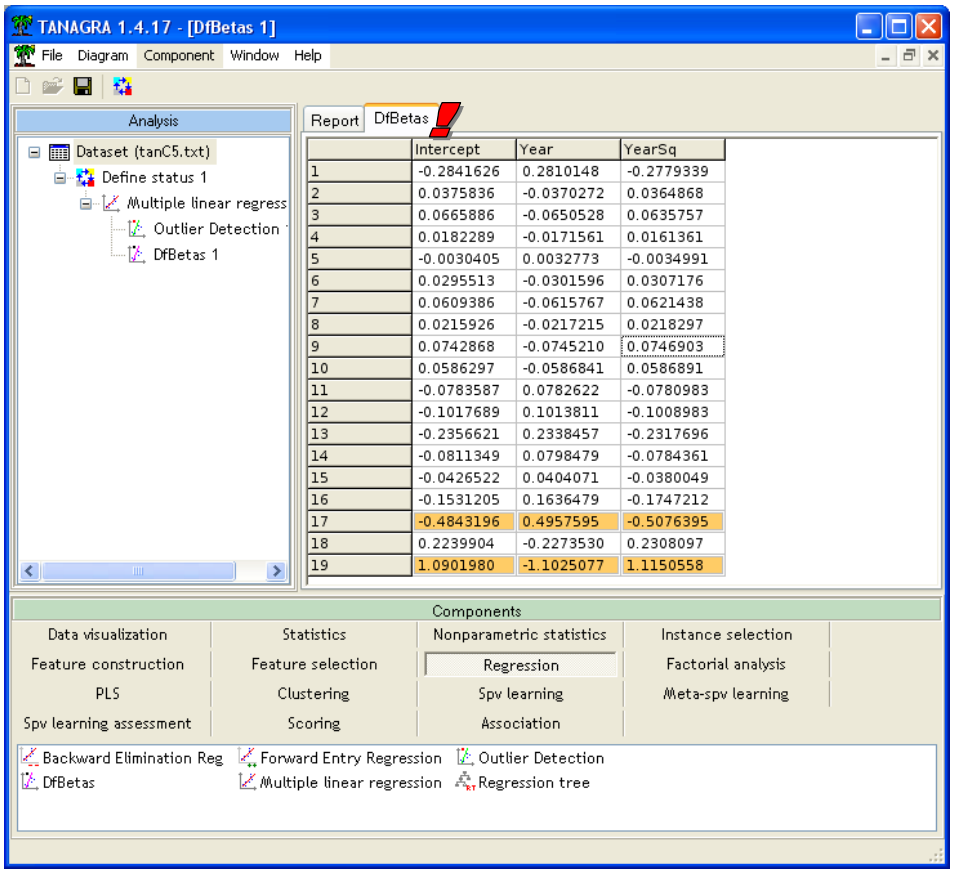
Detection of outliers and influential points -- DFBETAS (2)

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.



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



	Intercept	Year	YearSq
1	-0.2841626	0.2810148	-0.2779339
2	0.0375836	-0.0370272	0.0364868
3	0.0665886	-0.0650528	0.0635757
4	0.0182289	-0.0171561	0.0161361
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

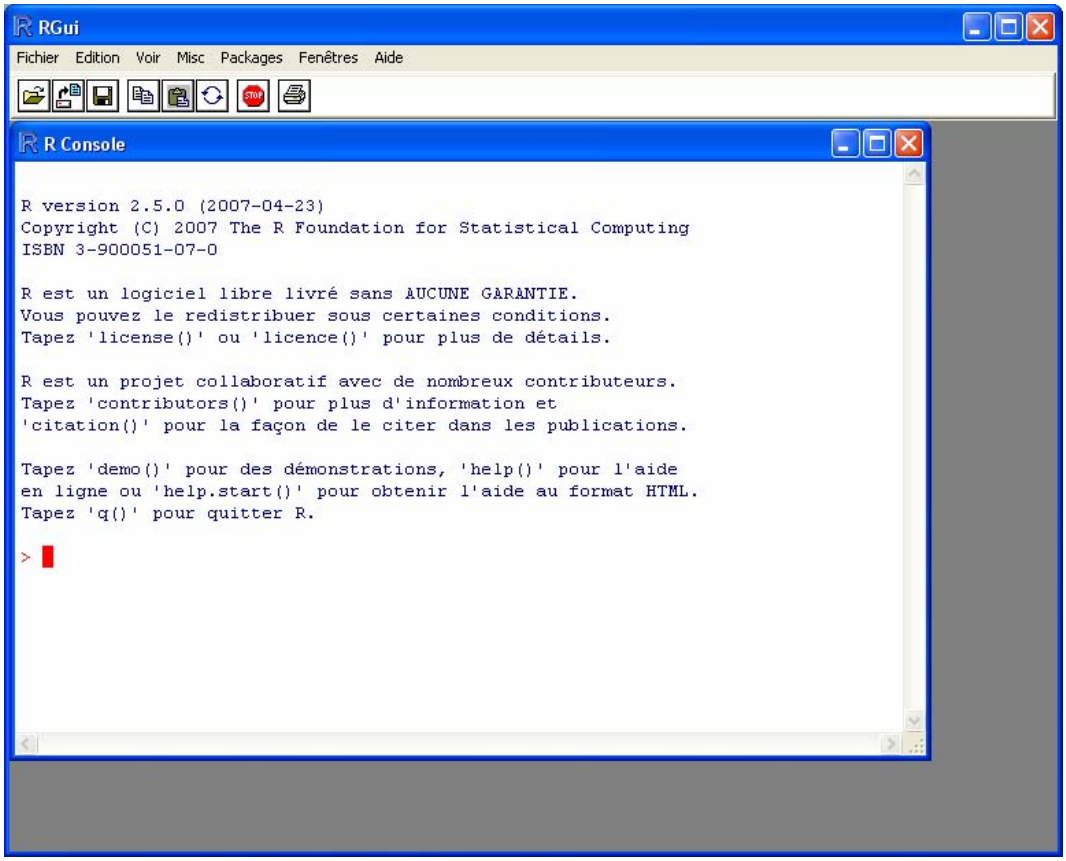
The threshold value is **2/SQRT(n)**. The suspicious values are highlighted. Here also, we can copy the values into a spreadsheet (COMPONENT / COPY RESULTS).

5. Detection under R

The R software is free tool (http://www.r-project.org/). 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 [5].

Launching R

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

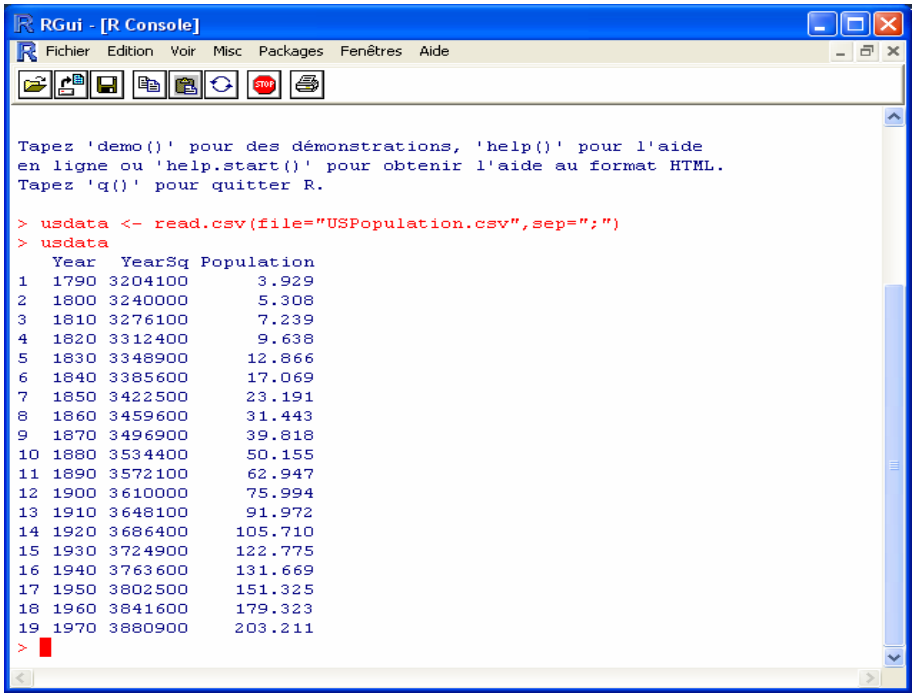


Importing the dataset

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

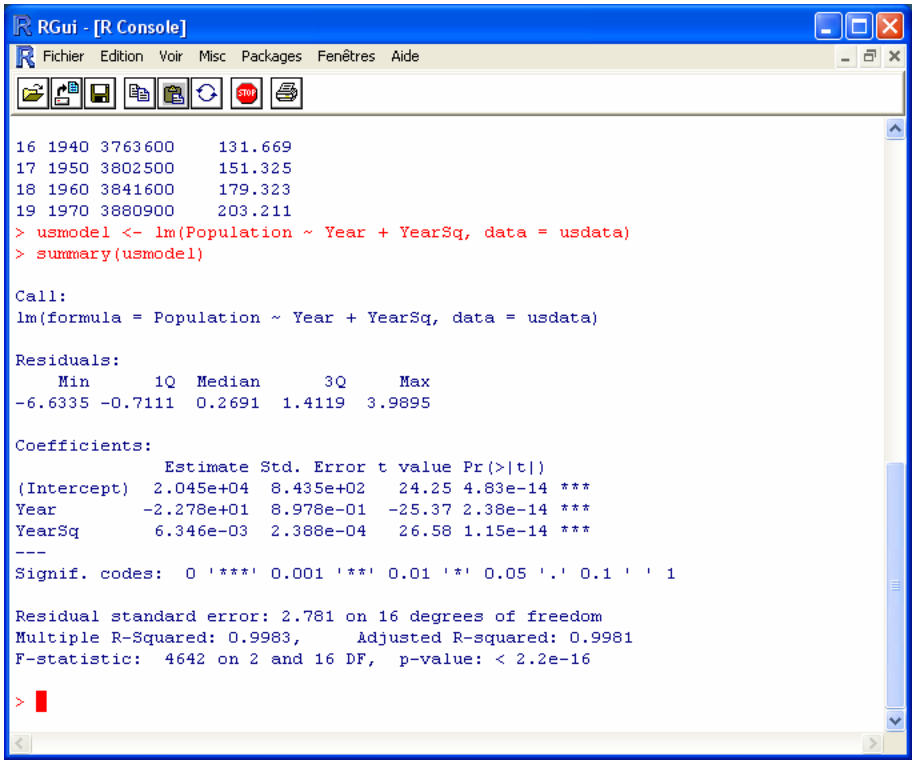
We display the dataset here.

[5] 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.



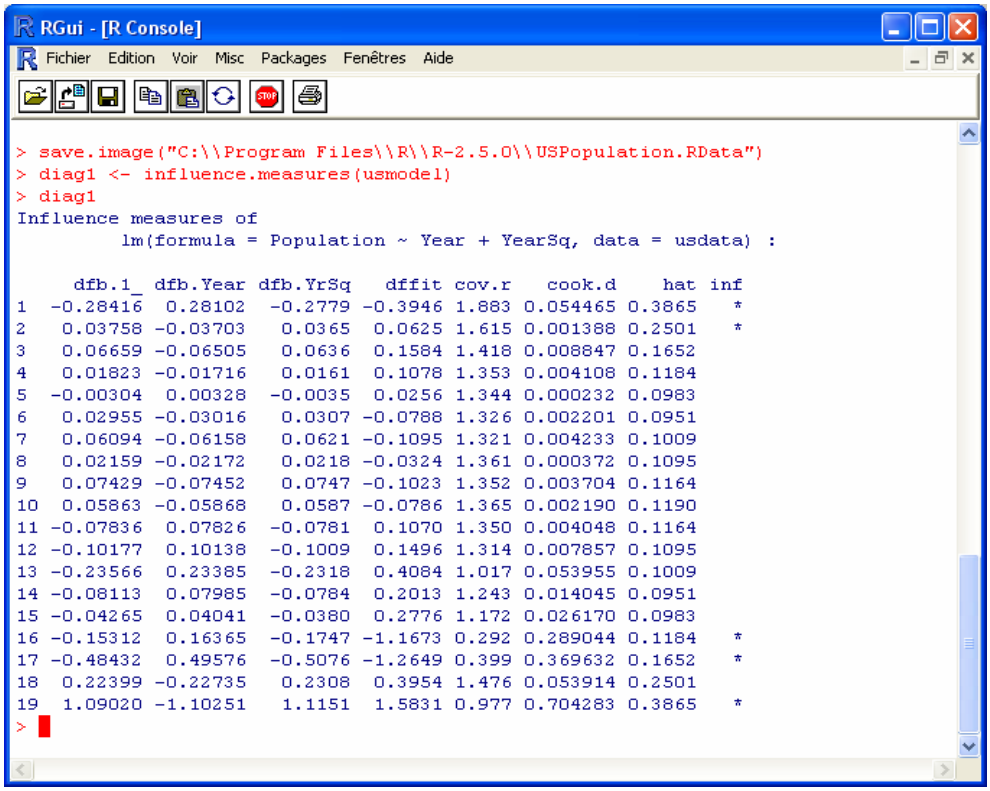
Linear regression

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



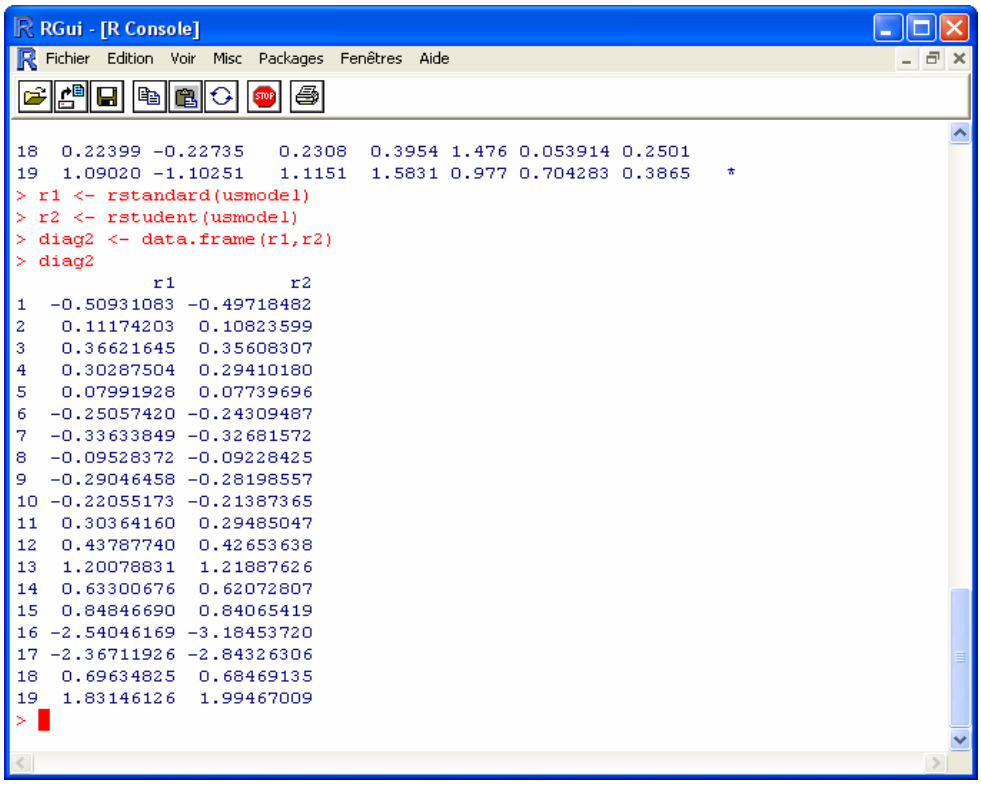
Measures for detecting outliers and influential points

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.



RStandard and RStudent

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



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. http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm).

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