

Objectif

Détecter les points aberrants et influents de la régression linéaire multiple.

La validation est une étape clé de la modélisation. S'agissant de la régression linéaire multiple, parmi les multiples évaluations à mettre en place figure l'analyse des résidus, plus particulièrement la détection des points aberrants et influents.

La distinction entre points aberrants et points influents n'est pas toujours comprise. Pour simplifier, nous dirons que les points aberrants correspondent à des observations hors normes, c.-à-d. ne correspondant pas à la population étudiée. Il peut s'agir d'un comportement très particulier, par exemple un senior qui se présente à l'épreuve du bac¹. Ces points peuvent fausser les résultats de la régression.

Les points influents correspondent aux observations qui pèsent significativement, voire exagérément, sur les calculs. Ils déterminent, dans une proportion à évaluer, les résultats de la régression. Il convient de les considérer avec précaution au moins pour interpréter correctement les coefficients obtenus. Dans certains cas, on peut considérer qu'une observation pèse trop sur la régression au point d'altérer les paramètres estimés.

Concernant la détection des points aberrants et influents dans la régression, il existe un très grand nombre de documents en ligne, de qualité assez inégale malheureusement. Pour notre part, nous en conseillerons deux qui, tout en étant simples et accessibles, donnent avec précision les principaux repères à retenir : en français (<http://ifr69.vjf.inserm.fr/~webifr/ppt/outilsdiag.ppt>, d'Alice Gueguen) et en anglais (<http://www-stat.stanford.edu/~jtaylo/courses/stats203/notes/diagnostics.pdf>, de Jonathan Taylor). Il est important de bien comprendre le sens à donner aux indicateurs proposés, et ils sont nombreux, pour interpréter correctement les résultats.

Enfin, pour éviter les confusions autour des définitions des indicateurs (la définition des résidus standardisés, studentisés, internes ou externes, est par exemple assez fluctuante d'un logiciel à l'autre), nous cadrerons nos calculs par rapport à deux logiciels reconnus dans le monde scientifique, SAS et le logiciel gratuit R. Nous procéderons de la manière suivante : dans un premier temps, nous affichons les données et les résultats issus de la documentation de SAS, ils nous serviront de référence ; puis nous décrivons la procédure sous TANAGRA ; enfin, nous décrivons la démarche à suivre sous R. Fort heureusement (*ouf ! Je me voyais très mal remettre en question ces logiciels*), les résultats coïncident.

Données

Les données sont issues de la documentation de SAS, disponible en ligne (<http://v8doc.sas.com/sashtml/stat/chap55/sect33.htm#regprv>). L'objectif est d'expliquer la population US (USPopulation) à partir de l'année (Year) et du carré de l'année (YearSq). Nous mettrons essentiellement l'accent sur la mise en œuvre des calculs et la comparaison de résultats dans ce didacticiel.

Résultats de SAS

Les données et les résultats de SAS sont les suivants :

¹ « Aberrant » est peut être un peu péjoratif d'ailleurs, « atypique » serait plus *politiquement correct*.

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

La procédure de calcul des points influents fournit le tableau suivant :

*The REG Procedure
Model: MODEL1
Dependent Variable: Population*

Output Statistics								
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS		
						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

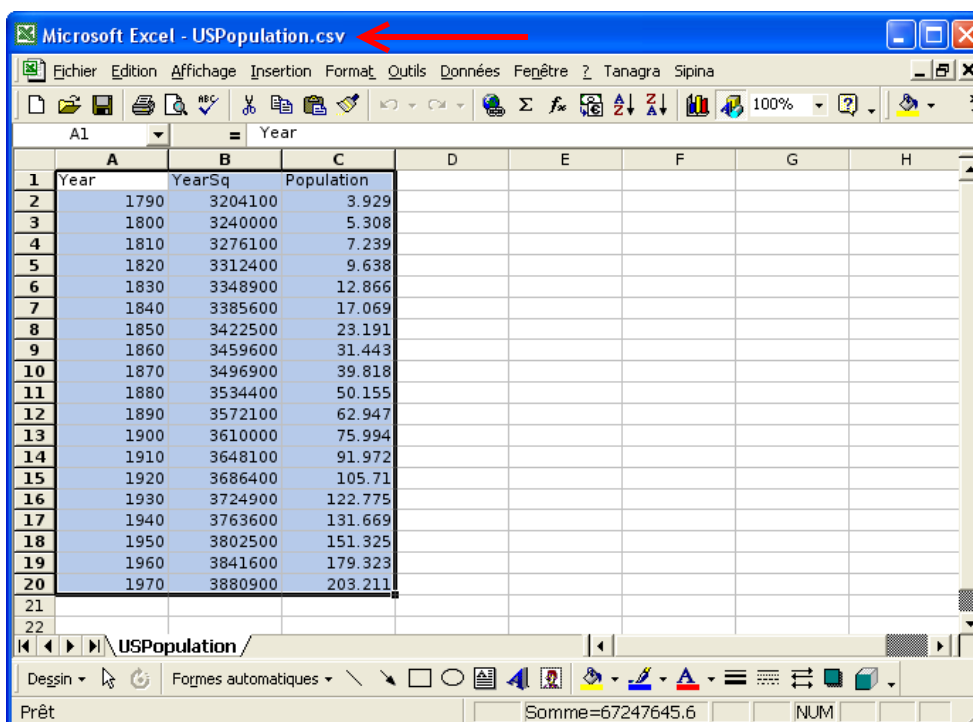
Nous reconnaissons dans l'ordre :

1. Le numéro des observations ;
2. Les résidus (écart entre valeur observée et valeur prédite) ;
3. Les résidus studentisés (*résidus studentisés externes* dans certains logiciels) ;
4. Les leverage (extrait de la hat-matrix H) ;
5. Les COVRATIO ;
6. Les DFFITS ;
7. Et les DFBETAS calculés sur les coefficients, y compris la constante (Intercept).

Détection des points atypiques et influents avec TANAGRA

Création du diagramme et chargement des données

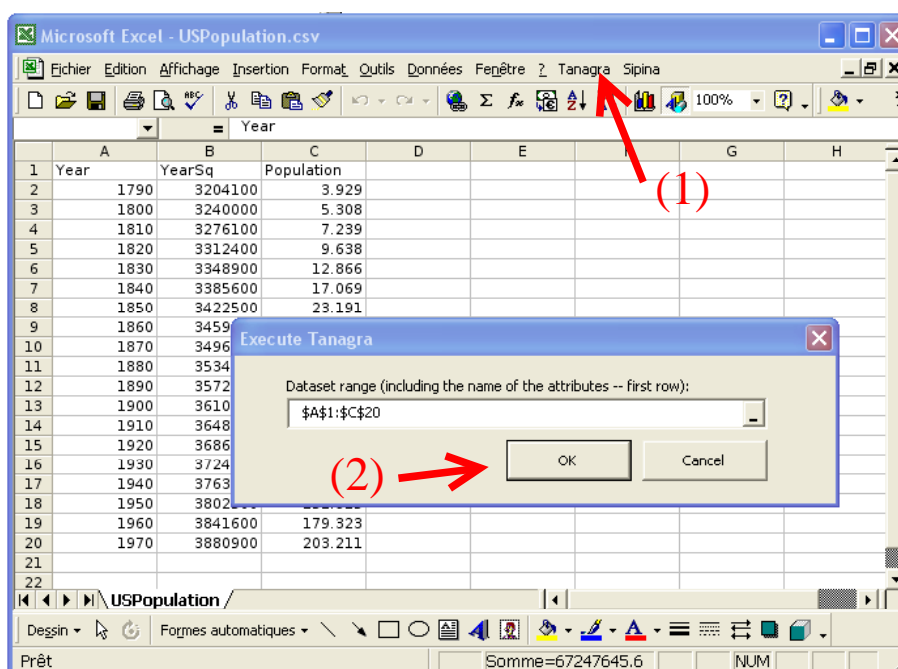
Le fichier USPopulation.CSV est au format CSV, format reconnu par R. TANAGRA ne le prend pas en charge directement. Le plus simple, pour nous, est de charger les données dans le tableur EXCEL².



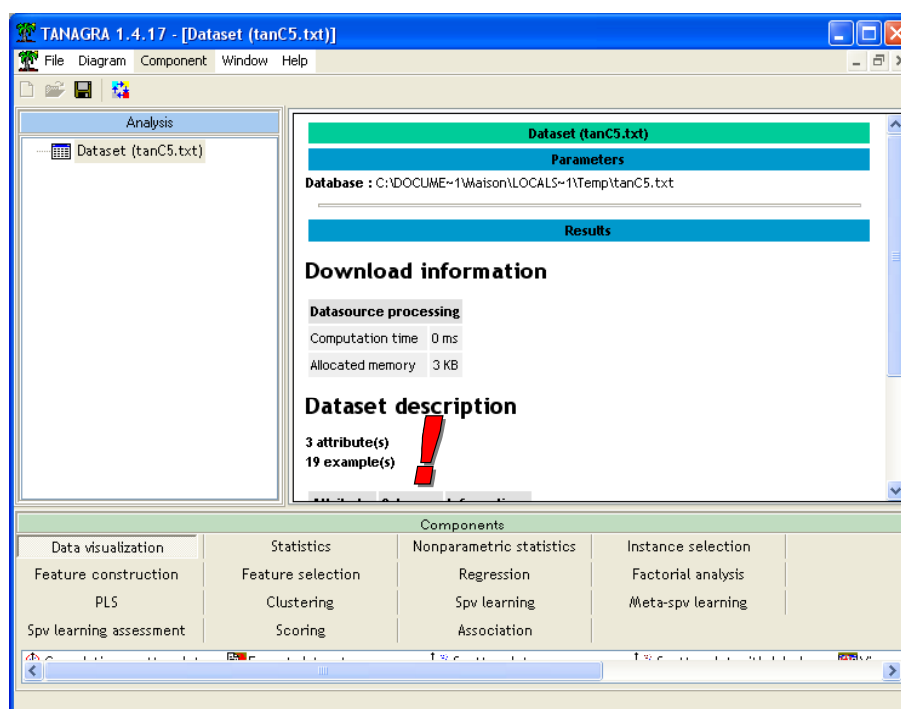
	A	B	C	D	E	F	G	H
1	Year	YearSq	Population					
2	1790	3204100	3.929					
3	1800	3240000	5.308					
4	1810	3276100	7.239					
5	1820	3312400	9.638					
6	1830	3348900	12.866					
7	1840	3385600	17.069					
8	1850	3422500	23.191					
9	1860	3459600	31.443					
10	1870	3496900	39.818					
11	1880	3534400	50.155					
12	1890	3572100	62.947					
13	1900	3610000	75.994					
14	1910	3648100	91.972					
15	1920	3686400	105.71					
16	1930	3724900	122.775					
17	1940	3763600	131.669					
18	1950	3802500	151.325					
19	1960	3841600	179.323					
20	1970	3880900	203.211					

Puis nous activons le menu TANAGRA/EXECUTE TANAGRA³ après avoir sélectionné les données.

² Lancer EXCEL, puis activer le menu Fichier/Ouvrir. Si vous tentez d'accéder au fichier en double-cliquant dans l'explorateur Windows, le tableur ne reconnaîtra pas le séparateur « ; » utilisé.



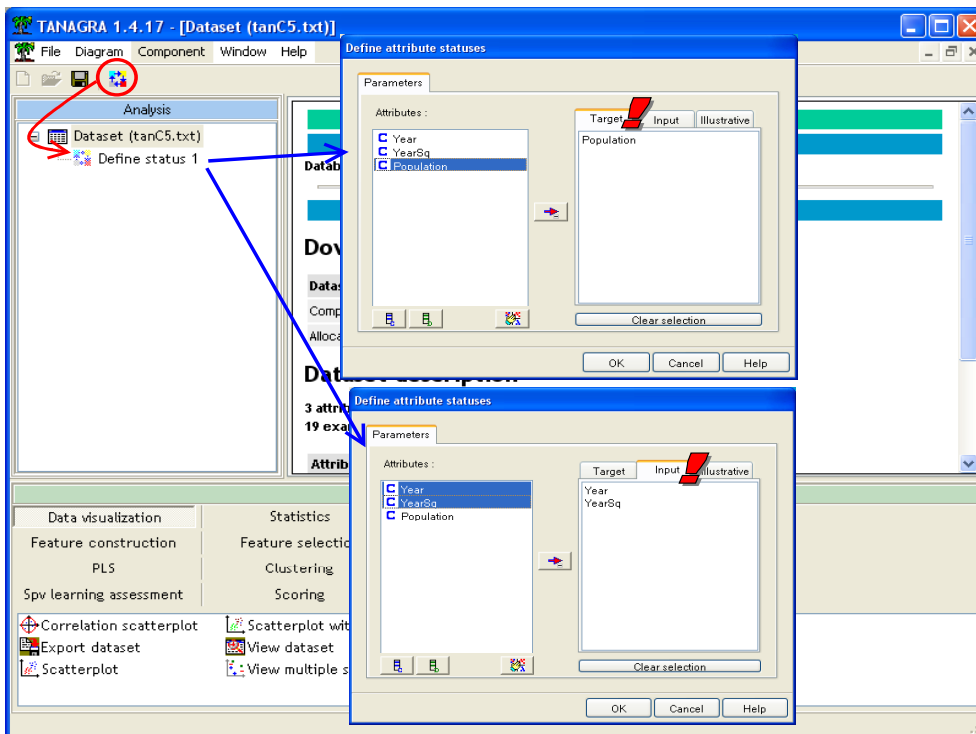
TANAGRA est automatiquement lancé. Nous vérifions que 19 observations et 3 variables sont disponibles.



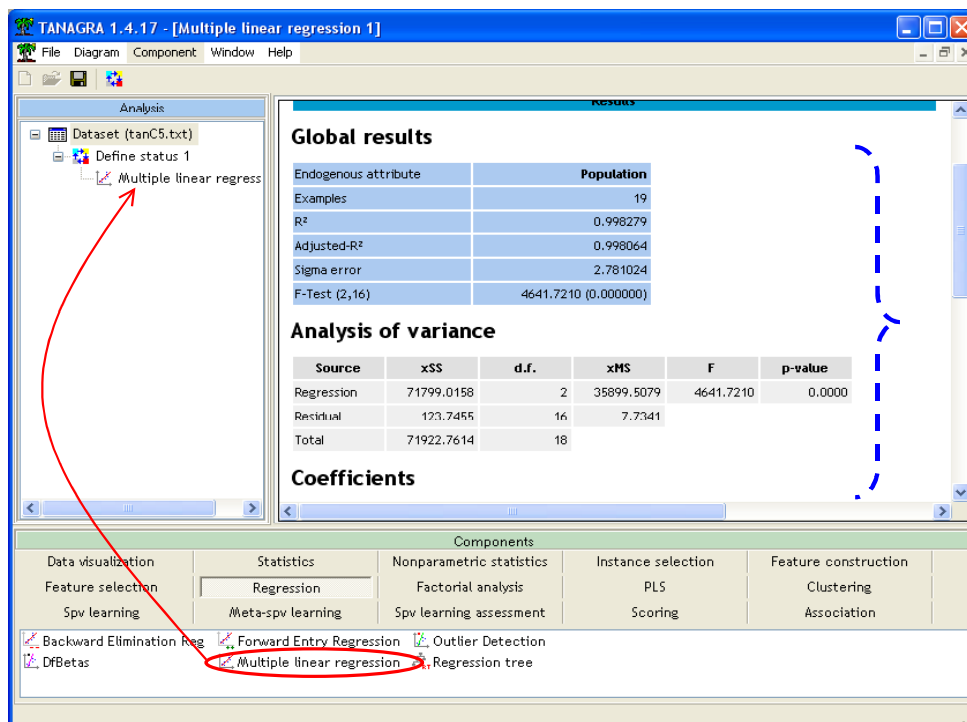
³ Ce menu est disponible avec l'installation de la macro complémentaire TANAGRA.XLA livrée en standard avec le logiciel. Voir les didacticiels sur le site pour une description détaillée de la procédure. Cette macro est disponible depuis la version 1.4.11 de TANAGRA.

Régression linéaire multiple

Il faut définir le problème à traiter. Pour cela, nous insérons le composant DEFINE STATUS dans le diagramme en passant par le raccourci disponible dans la barre d'outils. Nous plaçons en TARGET la variable POPULATION, et en INPUT les variables YEAR et YEARSQ.



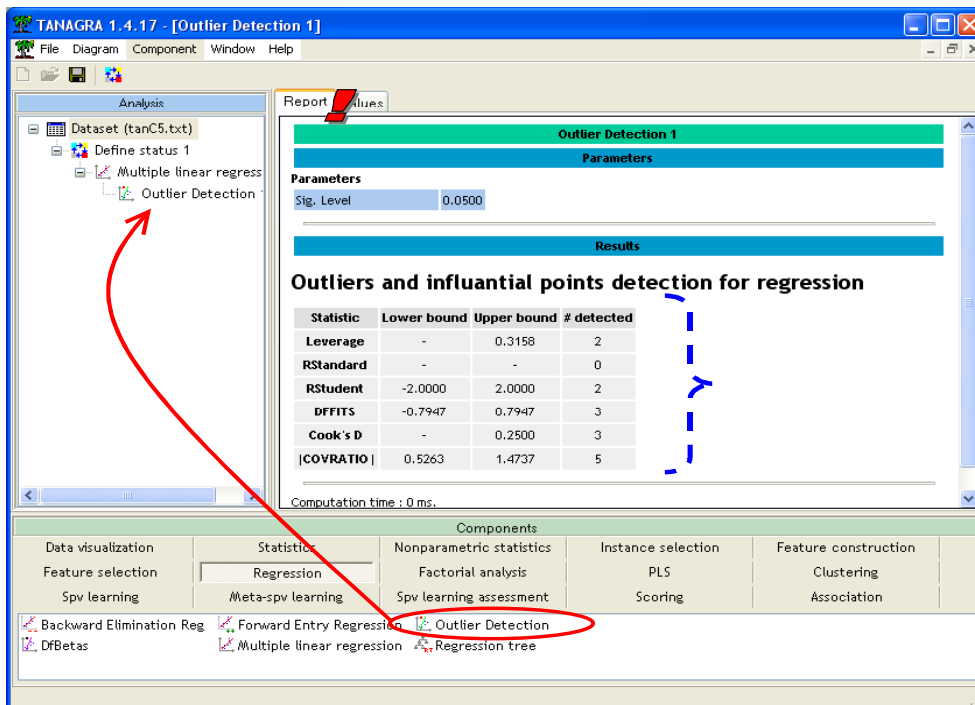
Puis nous insérons le composant MULTIPLE LINEAR REGRESSION (onglet REGRESSION). En activant le menu contextuel VIEW, les résultats s'affichent.



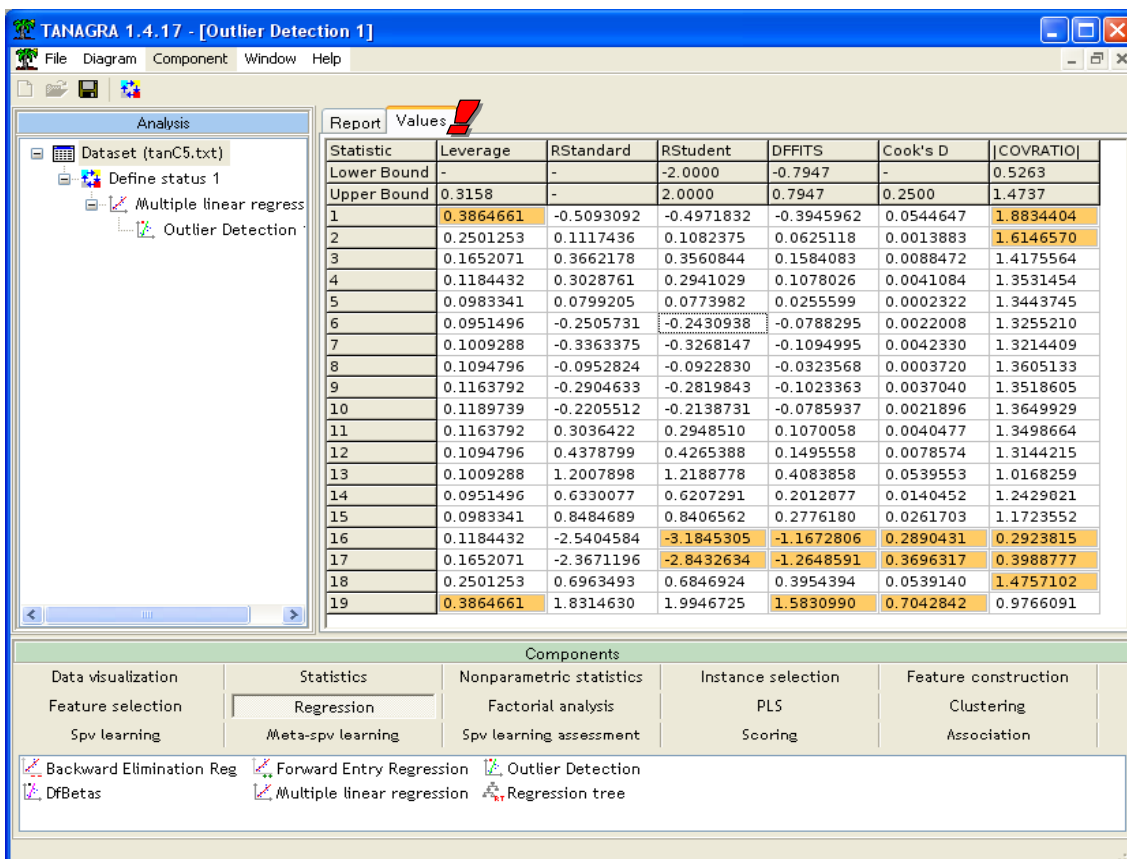
Nous retrouvons bien les résultats de SAS, avec un R² de 0.9983. Tous les paramètres concordent.

Détection des points atypiques et influents (1)

Nous insérons maintenant l'outil de diagnostic OUTLIER DETECTION (onglet REGRESSION). Nous activons le menu VIEW. Deux onglets sont disponibles dans la fenêtre de résultats. Le premier (REPORT) récapitule le nombre d'observations suspectes identifiées selon les indicateurs.



Le second (VALUES) fournit le détail, en surlignant les valeurs à considérer avec attention.



Le choix des valeurs de coupure est un élément important, elles permettent d'identifier les observations marquantes. Nous détaillons dans le tableau ci-dessous les seuils pour chaque indicateur : n étant le nombre d'observations, p le nombre de coefficients estimés (= nombre de variables + 1 si nous effectuons une régression avec constante).

Indicateur	Seuil
Leverage	$2*p/n$
RStandard	-
RStudent	2
DFFITS	$2*SQRT(p/n)$
Cook's D	$4/(n-p)$
COVRATIO-1	$3*p/n$

Les résultats détaillés peuvent être copiés dans un tableur pour des analyses approfondies (cf. conclusion). Il faut actionner le menu COMPONENT/COPY RESULTS lorsque l'onglet VALUES est actif.

En comparant nos résultats avec ceux de SAS, nous constatons que les formules implémentées sont bien les mêmes. Nous verrons qu'il en sera de même avec le logiciel R.

Détection des points atypiques et influents -- DFBETAS (2)

Ce second outil permet d'évaluer le rôle de chaque observation sur chaque coefficient estimé. L'intérêt de proposer un second composant pour ces calculs est purement ergonomique, il ne paraissait pas souhaitable de multiplier les onglets dans une fenêtre de visualisation, d'autant plus que les résultats sont de nature un peu différente : OUTLIER DETECTION évalue l'influence des observations, globalement ; DFBETAS évalue l'action des observations sur chaque coefficient.

Nous insérons le composant DFBETAS (onglet REGRESSION) à la suite de la régression multiple. Nous actions le menu VIEW.

The screenshot shows the TANAGRA 1.4.17 software interface. The main window is titled "TANAGRA 1.4.17 - [DfBetas 1]". The interface is divided into several sections:

- Analysis:** A tree view on the left shows the workflow: Dataset (tanC5.txt) > Define status 1 > Multiple linear regress > Outlier Detection > DfBetas 1. A red arrow points from the "DfBetas 1" component to the "Components" section below.
- Report DfBetas:** A window on the right displays the results for "DFBETAS for Regression Diagnostics". It shows a "Cut value = 0.4588" and a table of "# detected according to coefficients":

Coefficient	# detected
Intercept	2
Year	2
YearSq	2

 A blue bracket groups the three rows of the table.
- Components:** A grid of components is shown below the report. The "Regression" component is selected. Other components include Data visualization, Statistics, Nonparametric statistics, Instance selection, Feature construction, Feature selection, Regression, Factorial analysis, PLS, Clustering, Spv learning, and Meta-spv learning. At the bottom, a row of icons includes "Backward Elimination Reg", "Forward Entry Regression", "Outlier Detection", "DfBetas", "Multiple linear regression", and "Regression tree". The "DfBetas" icon is circled in red.

De nouveau, deux onglets sont disponibles : un récapitulatif dans le premier onglet, les résultats détaillés dans le second.

The screenshot shows the TANAGRA 1.4.17 software interface. The main window displays a report for 'DfBetas' analysis. The report table is as follows:

	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

Below the report, there is a 'Components' section with a grid of options:

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	

At the bottom, there is a list of available analysis methods:

- Backward Elimination Reg
- Forward Entry Regression
- Outlier Detection
- DfBetas
- Multiple linear regression
- Regression tree

La valeur de coupure est unique ici, elle est égale à $2/\sqrt{n}$. Les interférences suspectes sont surlignées.

Ici également, il est possible de copier ces valeurs dans un tableur (COMPONENT / COPY RESULTS).

Détection des points atypiques et influents avec R

Le logiciel R est un logiciel libre. Il dispose d'une excellente réputation auprès des statisticiens, tout à fait justifiée d'ailleurs au regard de ses multiples qualités. Sa bibliothèque de calcul est impressionnante. De plus, R étant ouvert, tout un chacun peut l'enrichir assez facilement en programmant des nouvelles méthodes, il faut néanmoins s'investir un peu pour appréhender correctement son langage de script.

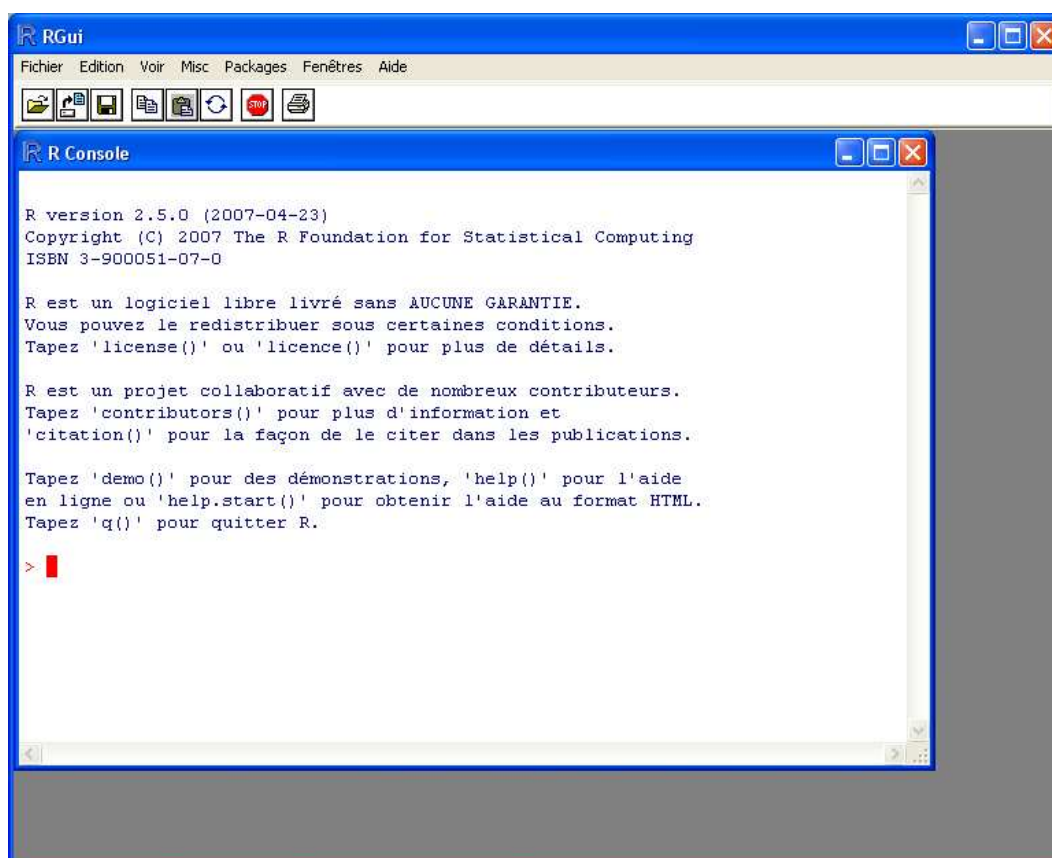
C'est peut-être le seul écueil de ce logiciel. Certes, la ligne de commande associée à un langage cohérent procure d'immenses possibilités, mais elle constitue également une barrière à l'entrée pour les néophytes habitués aux logiciels pilotés par menu ou, à l'extrême limite de complexité, fonctionnant par diagramme. Ce didacticiel est l'occasion de découvrir un peu ce logiciel, surtout pour les « data miner » informaticiens, qui viennent de l'apprentissage automatique et de la reconnaissance de formes, et qui sont plus habitués à des références telles que WEKA ou YALE.

L'autre raison qui nous fait introduire R est que les techniques que nous étudions dans ce didacticiel sont implémentées en standard par les plus grands experts du domaine. Les calculs implémentés dans R servent souvent de référence.

Démarrer R

Le logiciel est disponible sur le site du projet (<http://www.r-project.org/>). L'installation ne pose pas de problèmes particuliers, il faut suivre la procédure avec un peu d'attention. Nous avons installé la version 2.5.0.

Nous démarrons le logiciel en double-cliquant sur le raccourci du menu DEMARRER. La fenêtre de R apparaît⁴.

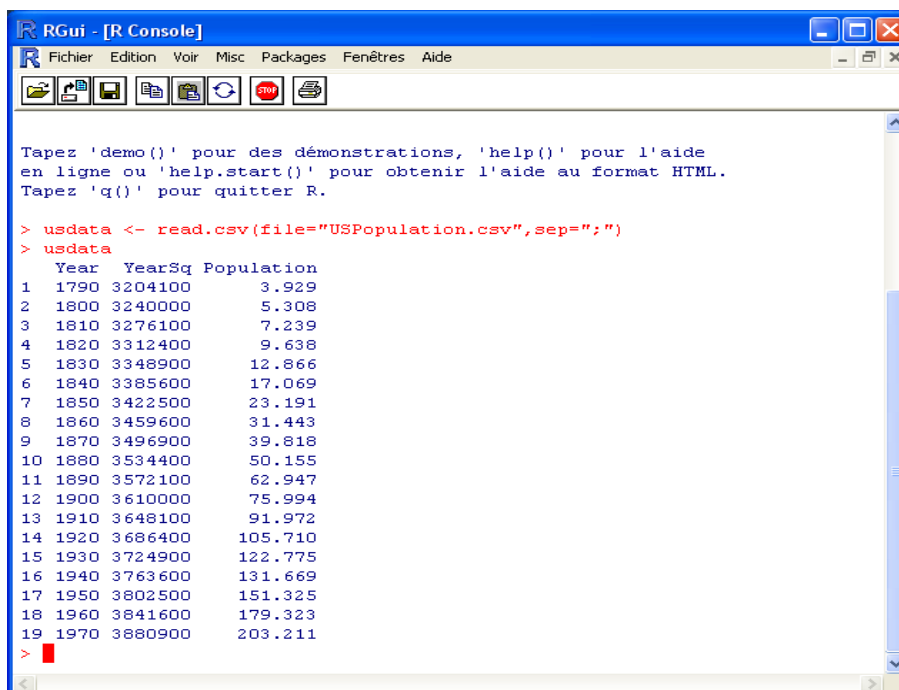


Charger les données

Le plus simple est copier les données USPopulation.CSV dans le répertoire de R (C:\Program Files\R\R-2.5.0).

Des commandes spécifiques permettent de charger les données dans un objet « usdata » et de les afficher.

⁴ Je confesse être un véritable néophyte s'agissant de R. Les manipulations décrites dans ce didacticiel sont donc simplifiées à l'extrême. Il se peut, très vraisemblablement, qu'elles ne soient pas optimales. Mon objectif était de produire des résultats comparables à ceux de SAS et TANAGRA pour vérifier l'exactitude de mes calculs.



```

RGui - [R Console]
Fichier Edition Voir Misc Packages Fenêtres Aide

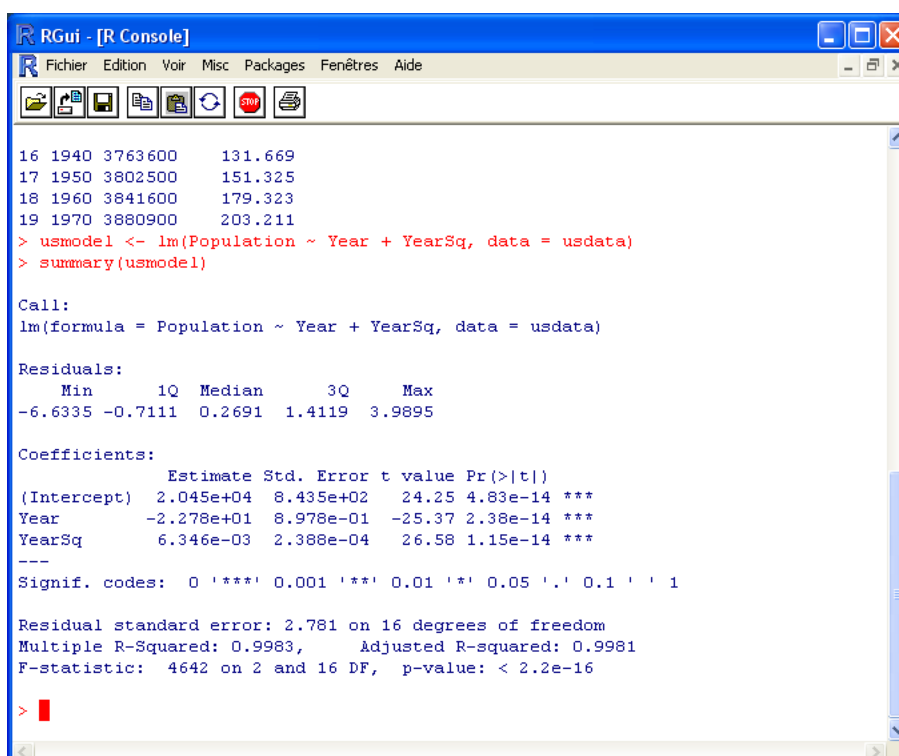
Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

> usdata <- read.csv(file="USPopulation.csv",sep=";")
> usdata
  Year  YearSq 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
19 1970 3880900  203.211
>

```

Régression linéaire multiple

La commande **lm** permet de lancer une régression linéaire multiple. Le modèle est conservé dans un objet nommé « usmodel ».



```

RGui - [R Console]
Fichier Edition Voir Misc Packages Fenêtres Aide

16 1940 3763600    131.669
17 1950 3802500    151.325
18 1960 3841600    179.323
19 1970 3880900    203.211
> usmodel <- lm(Population ~ Year + YearSq, data = usdata)
> 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 ***
YearSq        6.346e-03  2.388e-04   26.58 1.15e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual 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

>

```

Mesures des points influents

Les points influents sont calculés automatiquement et sont disponibles. Il suffit de les récupérer dans un objet pour les afficher.

```

> save.image("C:\\Program Files\\R\\R-2.5.0\\USPopulation.RData")
> diag1 <- influence.measures(usmodel)
> 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 *
2  0.03758 -0.03703  0.0365  0.0625 1.615 0.001388 0.2501 *
3  0.06659 -0.06505  0.0636  0.1584 1.418 0.008847 0.1652
4  0.01823 -0.01716  0.0161  0.1078 1.353 0.004108 0.1184
5 -0.00304  0.00328  -0.0035  0.0256 1.344 0.000232 0.0983
6  0.02955 -0.03016  0.0307 -0.0788 1.326 0.002201 0.0951
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 *
>

```

R affiche, de gauche à droite : les DFBETAS (constante, Year, YearSq), les DFFITS, les COVRATIO, les distances de Cook, et le leverage.

RStandard et RStudent

Pour l'affichage des résidus standardisés et studentisés, il faut quelques manipulations supplémentaires.

```

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 *
> r1 <- rstandard(usmodel)
> r2 <- rstudent(usmodel)
> diag2 <- data.frame(r1,r2)
> diag2
      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
11  0.30364160  0.29485047
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
17 -2.36711926 -2.84326306
18  0.69634825  0.68469135
19  1.83146126  1.99467009
>

```

Toutes les valeurs obtenues sont en adéquation avec ceux de SAS et TANAGRA.

Conclusion

L'analyse des résidus est une étape clé de la validation de la régression linéaire multiple. Il y a la détection des points aberrants et influents que nous avons initié dans ce didacticiel, mais il ne faut pas non plus oublier les autres vérifications. Ils passent par des graphiques « simples » des résidus (résidu vs. Endogène, résidu vs. Les exogènes, etc.) ; la vérification de la normalité des résidus que nous pouvons mettre en œuvre avec des graphiques intuitifs tels que les QQ-plots (tant que ce n'est pas la praline...) ; les vérifications de l'hétéroscédasticité et de l'auto-corrélation (lorsque les données sont longitudinales)....

Enfin, si nous proposons dans TANAGRA les valeurs seuils les plus souvent référencées dans la littérature, il ne faut pas en faire une vérité irrévocable. A mon sens, il faut en réalité s'inquiéter lorsqu'une observation prend une valeur très différente des autres sur un indicateur. Plutôt que de voir si elle dépasse ou pas une valeur de coupure, le plus simple est de les trier et de voir dans quelle mesure elle s'éloigne des autres en utilisant des concepts simples telles que les intervalles interquartiles (cf. <http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>) pour évaluer l'importance de l'écart.

Dans notre exemple ci-dessus, après avoir copié les valeurs dans EXCEL, nous avons trié le tableau selon le RSTUDENT. Nous avons alors calculé les indicateurs simples proposés sur le site de NIST (Lower Outer/Inner Fence ; Upper Inner/Outer Fence). Nous constatons que les observations n°16 et n°17 sont effectivement problématiques, nous découvrons en revanche que l'observation n°19 est aussi à considérer avec beaucoup d'attention. La simple comparaison du RSTUDENT avec la valeur seuil 2 n'aurait pas permis de la détecter, surtout si le nombre d'observations est très élevé.

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	0.52363395
IQR	0.78617299
Lower Outer Fence	-2.62105802
Lower Inner Fence	-1.44179853
Upper Inner Fence	1.70289343
Upper Outer Fence	2.88215292