TESTS DE NORMALITE

Dans le chapitre précédent on a vu les propriétés nécessaires sur les erreurs pour que les coefficients des MCO soient les "meilleurs". Dans la pratique bien sur ce ne sera pas toujours le cas. Dans ce chapitre nous allons étudier les tests classiques qui seront systématiquement faits afin de voir si les erreurs ont les bonnes propriétés et si donc les MCO sont utilisables. Nous verrons également bien sur les conséquences d'une hypothèse non vérifiée et les premiers éléments pour régler le problème.

1 Tests le Normalité des erreurs: Théorie

Hypothèses du test :

$$\begin{array}{lcl} H_0 & : & \text{les erreurs suivent une loi Normale} \\ H_1 & : & \text{les erreurs ne suivent pas une loi Normale} \end{array}$$

Sous l'hypothèse H1 la loi des erreurs est donc inconnue.
On caractérise la loi normale N(m,σ^2) par le fait

- qu'elle est symétrique $\Rightarrow$ son moment centré d'ordre 3 est nul $\quad \mu_3 = 0$
- que son moment centré d'ordre 4 est $\mu_4 = 3\mu_2^2 = 3(\sigma^2)^2 \Longrightarrow$ sa Kurtosys K=$\mu_4/\sigma^4 = 3$

les hypothèses du test peuvent alors s'écrire

$$\begin{array}{lcl} H_o & : & \mu_3 = 0 \text{ et } \mu_4 = 3\sigma^4 \\ H_1 & : & \text{l'une au moins de ces deux propriétés n'est pas vérifiée} \end{array}$$

le moment σ^2 est estimé à l'aide des résidus par $s^2 = \sum e_t^2/(n-k)$
le moment $\mu_3 = E\left[(\epsilon - E(\varepsilon))^3\right] = E\left[\epsilon^3\right]$ est estimé à l'aide des résidus par $\widehat{\mu_3} = \sum e_t^3/n$
le moment $\mu_4 = E\left[(\epsilon - E(\varepsilon))^4\right] = E\left[\epsilon^4\right]$ est estimé à l'aide des résidus par $\widehat{\mu_4} = \sum e_t^4/n$

- on définit le coefficient de symétrie (skewness)

$$\alpha_3 = \frac{\mu_3}{\sigma^3} \text{ estimé par } \widehat{\alpha_3} = \frac{\sum e_t^3/n}{s^3}$$

qui suit asymptotiquement une loi N(0,3!/n) sous l'hypothèse H_0

- on définit le coefficient d'aplatissement (Kurtosis)

$$\alpha_4 = \frac{\mu_4}{\sigma^4} \text{ estimé par } \widehat{\alpha_4} = \frac{\sum e_t^4/n}{s^4}$$

qui suit asymptotiquement une loi N(3,4!/n) sous l'hypothèse H_0

Les hypothèses du test peuvent donc s'écrire

$$\begin{array}{lll} H_o & : & \alpha_3 = 0 \ \text{ et } \ \alpha_4 = 3 \\ H_1 & : & \text{l'une au moins de ces deux propriétés n'est pas vérifiée} \end{array}$$

Il existe deux tests de normalité l'un testant séparément les deux parties de H_0 , l'autre global

1.1 tests de Skewness et Kurtosis

1.1.1 test de Skewness

$$\begin{array}{lll} H_{01} : \mu_3 & = & 0 \Rightarrow \ \alpha_3 = 0 \\ H_{11} : \mu_3 & \neq & 0 \Rightarrow \ \alpha_3 \neq 0 \end{array}$$

sous l'hypothèse H_0 l'estimateur $\widehat{\alpha_3}$ de α_3 suit asymptotiquement une loi N(0 , 3!/n) où (3!=6), sa variable centrée et réduite t

$$t = \sqrt{n/6}(\widehat{\alpha_3}) \text{ suit asymptotiquement une N(0,1)}$$

si $-1.96 < t_{estimation} < 1.96$ on décide H_{01} sinon H_{11}
si on décide H_{11} le test est terminé car la loi n'étant pas symétrique elle ne peut être normale.
si la décision est H_{01} on passe au test suivant .

1.1.2 test de Kurtosis

$$\begin{array}{lll} H_{02} : \mu_4 & = & 3\sigma^4 \Rightarrow \ \alpha_4 = 3 \\ H_{12} : \mu_4 & \neq & 3\sigma^4 \Rightarrow \ \alpha_4 \neq 3 \end{array}$$

sous l'hypothèse H_0 l'estimateur $\widehat{\alpha_4}$ de α_4 suit asymptotiquement une loi N(3 , 4!/n) ou (4!=24)

$$t = \sqrt{n/24}(\widehat{\alpha_4} - 3) \text{ suit asymptotiquement une N(0,1)}$$

si $-1.96 < t_{estimation} < 1.96$ on décide H_{02} sinon H_{12}
si la décision est H_{02} , on a donc les deux propriétés vérifiées , on décide alors normalité des erreurs
si on décide H_{12} le test est terminé car la loi n'étant pas un coefficient d'aplatissement égal à 3 nous ne sommes pas dans le cadre de la loi normale.

REMARQUE : les logiciels donnent les estimations $\widehat{\alpha_3}$ et $\widehat{\alpha_4}$, mais il faut remarquer que certains donnent directement les valeurs centrées soit $(\widehat{\alpha_3} - 0)$ et $(\widehat{\alpha_4} - 3)$, on parle alors de Kurtosis centrée. Il est donc prudent de bien lire le mode d'utilisation du test. RATS donne la Kurtosis centrée.
lin y / residus
 # constant X Z
stat residus (donne skewness et kurtosis centrée , les versions 5 et 6 donnent aussi le tests suivant de Jarque et Berra)

1.2 Test Global de Jarque et Berra

Il teste donc globalement

$$\begin{aligned} H_o &: \mu_3 = 0 \text{ et } \mu_4 = 3\sigma^4 \\ H_1 &: \text{l'une au moins de ces deux propriétés n'est pas vérifiée} \end{aligned}$$

sous l'hypothèse H_0 vraie la variable aléatoire S somme des carrés des deux précédents résultats centrés réduits suit donc le loi du χ^2 à deux degrés de liberté.

$$S = \frac{n}{6}\widehat{\alpha_3}^2 + \frac{n}{24}(\widehat{\alpha_4} - 3)^2$$

Si $S_{estimation}$ est inférieur à la borne du χ^2_2 on décide H_0 sinon on décide H_1 .

2 APPLICATION

Le test de Normalité s'effectue à l'aide de la commande STATISTIQUE dans RATS.
On effectue la régression puis on récupère les résidus notés par exemple Res

2.1 Exemple 1 : Normalite1.prg

On reprend l'exemple du chapitre précédent MCO. Le programme est dans le fichier normalite1.prg

2.1.1 On explique Y en fonction de X1 seulement.

all 140
open data mco.rat
data(for=rats) /
smpl 1 140
lin Y / residus1
 # constant X1

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   138
Centered R**2     0.915883      R Bar **2   0.915274
Uncentered R**2   0.993195      T x R**2     139.047
Mean of Dependent Variable      692268.63700
```

```
Std Error of Dependent Variable 206128.93876
Standard Error of Estimate       59999.61073
Sum of Squared Residuals         4.96794e+11
Regression F(1,138)                1502.5742
Significance Level of F           0.00000000
Log Likelihood                   -1737.93725
Durbin-Watson Statistic             0.120471

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                -68443.58769  20269.23173     -3.37672  0.00095323
2.  X1                          7.40009      0.19091     38.76305  0.00000000
```

stat residus1 1 140

```
Statistics on Series RESIDUS1
Observations                  140
Sample Mean              0.000000      Variance            3574054343.543488
Standard Error       59783.395216      of Sample Mean      5052.619083
t-Statistic (Mean=0)     0.000000      Signif Level        1.000000
Skewness                 0.590044      Signif Level (Sk=0) 0.004808
Kurtosis (excess)        0.226655      Signif Level (Ku=0) 0.593468
Jarque-Bera              8.423217      Signif Level (JB=0) 0.014823
```

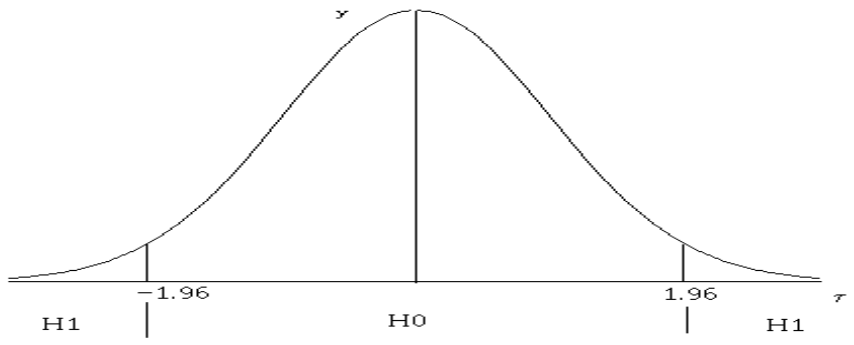
- Résultats du test de Normalité en deux parties

  On teste tout d'abord si la loi des erreurs en symétrique, test de Skewness

$$\begin{aligned} H_{01}: \mu_3 &= 0 \Rightarrow \alpha_3 = 0 \\ H_{11}: \mu_3 &\neq 0 \Rightarrow \alpha_3 \neq 0 \end{aligned}$$

  Les résultats de RATS montrent que $\widehat{\alpha_3} = \frac{\sum e_i^3/n}{s^3}$ =0.590044 . Sous l'hypothèse H_0 $t = \sqrt{n/6}(\widehat{\alpha_3})$ suit asymptotiquement une N(0,1)

  Règle de décision:



  Dans cet exemple $t = \widehat{\alpha_3}\sqrt{n/6} = 0.590044\sqrt{\frac{140}{6}} = 2.85 > 1.96$ on décide donc H1 $\alpha_3 \neq 0$ la loi des erreurs n'est donc pas symétrique.

  On peut également voir le résultat sans calcul en utilisant la P-value : le logiciel donne la P-value ou niveau de significativité =Prob($|t| > 2.85$)=0.004808 valeur nettement

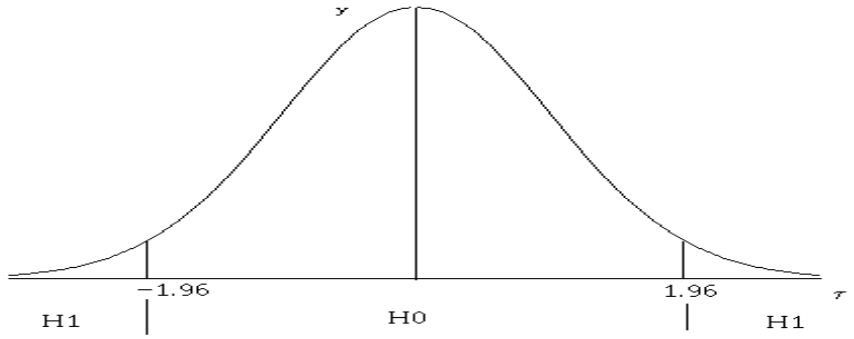
inférieure à 0.05 , on décide donc H1. Il est donc inutile d'aller plus loin dans le test, puisque la loi n'est pas symétrique, ce ne peut pas être une loi Normale. Pour l'exemple nous allons cependant poursuivre.

On regarde maintenant si la Kurtosis est égale à 3.

$$\begin{aligned} H_{02}: \mu_4 &= 3\sigma^4 \Rightarrow \alpha_4 = 3 \\ H_{12}: \mu_4 &\neq 3\sigma^4 \Rightarrow \alpha_4 \neq 3 \end{aligned}$$

Les résultats de RATS montrent que $\widehat{\alpha_4} - 3 = \frac{\sum e^4/n}{s^4} - 3 = 0.226655$. ATTENTION RATS donne l'excès de Kurtosis c'est-à-dire l'écart à 3. Sous l'hypothèse H_0 $t = (\widehat{\alpha_4} - 3)\sqrt{n/24}$suit asymptotiquement une N(0,1)

Règle de décision:



Dans cet exemple $t = (\widehat{\alpha_4} - 3)\sqrt{n/24} = 0.226655\sqrt{\frac{140}{24}} = 0.547$ compris dans l'intervalle $[-1.96, +1.96]$ on décide donc H0 $\alpha_4 = 3$ la loi des erreurs a une Kurtosis égale à 3.

On peut également voir le résultat sans calcul en utilisant la P-value : le logiciel donne la P-value ou niveau de significativité = 0.590698 valeur nettement supérieure à 0.05 , on décide donc H0. Conclusion générale: comme la loi n'est pas symétrique ce ne peut être une loi Normale.

- test global de Jarque et Berra

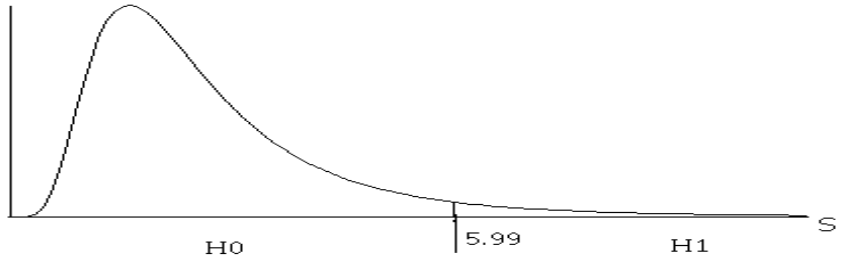
  Il teste donc globalement

$$\begin{aligned} H_o &: \mu_3 = 0 \text{ et } \mu_4 = 3\sigma^4 \text{ soit } \alpha_3 = 0 \text{ et } \alpha_4 = 3 \\ H_1 &: \text{l'une au moins de ces deux propriétés n'est pas vérifiée} \end{aligned}$$

sous l'hypothèse H_0 vraie la variable aléatoire S somme des carrés des deux précédents résultats centrés réduits suit donc le loi du χ^2 à deux degrés de liberté.

$$S = \frac{n}{6}\widehat{\alpha_3}^2 + \frac{n}{24}(\widehat{\alpha_4} - 3)^2$$

Règle de décision



Le résultat calculé par RATS est S=8.423217 nettement supérieur à la borne 5.99, on décide donc H1.

On peut aussi utiliser le niveau de significativité : il est de 0.014823 nettement inférieur à 5%, on décide donc H1 la loi n'est pas une loi Normale.

2.1.2 On explique Y en fonction de X1 et de X2

smpl 1 140
lin Y / residus2
 # constant X1 X2
en notant residus2 les résidus de cette nouvelle régression

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   137
Centered R**2     0.999998      R Bar **2   0.999998
Uncentered R**2   1.000000      T x R**2     140.000
Mean of Dependent Variable      692268.63700
Std Error of Dependent Variable 206128.93876
Standard Error of Estimate         291.88445
Sum of Squared Residuals        11671924.919
Regression F(2,137)             34660910.1766
Significance Level of F           0.00000000
Log Likelihood                   -991.82521
Durbin-Watson Statistic             1.753973

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                834.92234985 102.69414057      8.13018  0.00000000
2.  X1                        0.80325350   0.00288544    278.38191  0.00000000
3.  X2                        0.49984476   0.00020700   2414.74856  0.00000000
```

stat residus2 1 140

```
Statistics on Series RESIDUS2
Observations                140
Sample Mean            0.000000      Variance            83970.682869
Standard Error       289.776954      of Sample Mean      24.490623
t-Statistic (Mean=0)   0.000000      Signif Level        1.000000
Skewness               0.000120      Signif Level (Sk=0) 0.999541
Kurtosis (excess)      0.143768      Signif Level (Ku=0) 0.734908
Jarque-Bera            0.120572      Signif Level (JB=0) 0.941495
```

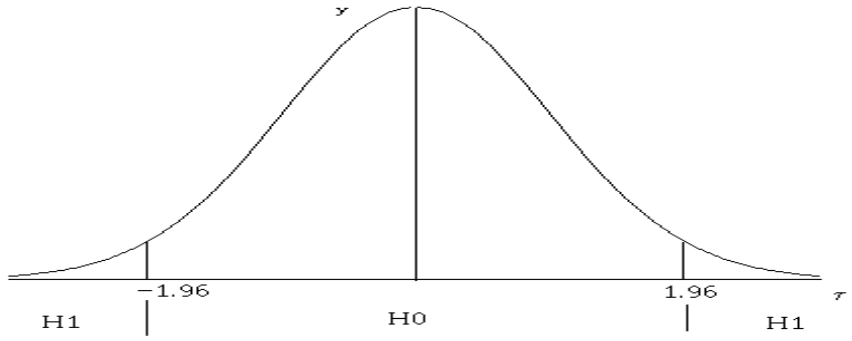
- Résultats du test de Normalité en deux parties

  On teste tout d'abord si la loi des erreurs est symétrique, test de Skewness

$$\begin{aligned} H_{01} : \mu_3 &= 0 \Rightarrow \alpha_3 = 0 \\ H_{11} : \mu_3 &\neq 0 \Rightarrow \alpha_3 \neq 0 \end{aligned}$$

  Les résultats de RATS montrent que $\widehat{\alpha_3} = \frac{\sum e_t^3/n}{s^3} = 0.590345$. Sous l'hypothèse H_0 $t = \sqrt{n/6}(\widehat{\alpha_3})$ suit asymptotiquement une N(0,1)

  Règle de décision:



  Dans cet exemple $t = \widehat{\alpha_3}\sqrt{n/6} = 0.000120\sqrt{\frac{140}{6}} = 0.000581.96$ compris dans l'intervalle $[-1.96, +1.96]$on décide donc H0 $\alpha_3 = 0$ la loi des erreurs est donc symétrique.

  On peut également voir le résultat sans calcul en utilisant la P-value : le logiciel donne la P-value ou niveau de significativité = 0.999541 valeur nettement supérieure à 0.05 , on décide donc H0. La première partie du test indique donc que la loi est symétrique. On peut donc poursuivre l'étude.

  On regarde maintenant si la Kurtosis est égale à 3.

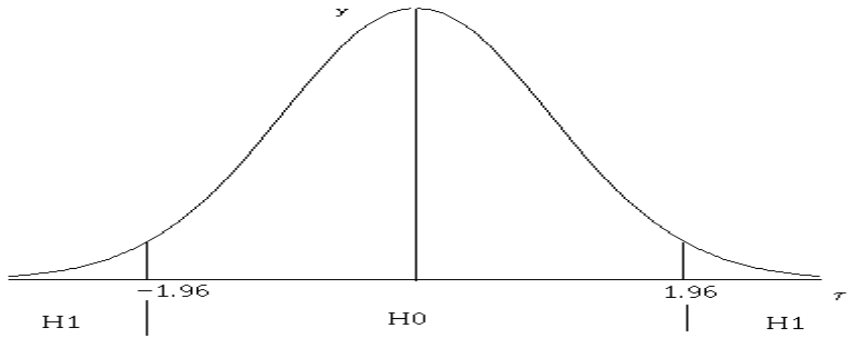
$$\begin{aligned} H_{02} : \mu_4 &= 3\sigma^4 \Rightarrow \alpha_4 = 3 \\ H_{12} : \mu_4 &\neq 3\sigma^4 \Rightarrow \alpha_4 \neq 3 \end{aligned}$$

  Les résultats de RATS montrent que $\widehat{\alpha_4} - 3 = \frac{\sum e^4/n}{s^4} - 3 =$0.143768. ATTENTION RATS donne l'excès de Kurtosis c'est à dire l'écart à 3. Sous l'hypothèse H_0 $t = (\widehat{\alpha_4} - 3)\sqrt{n/24}$suit asymptotiquement une N(0,1)

  Règle de décision:

  Dans cet exemple $t = \sqrt{n/24}(\widehat{\alpha_4} - 3) = 0.143768\sqrt{\frac{140}{24}} = 0.347$ compris dans l'intervalle $[-1.96, +1.96]$ on décide donc H0 $\alpha_4 = 3$ la loi des erreurs a une Kurtosis égale à 3.

  On peut également voir le résultat sans calcul en utilisant la P-value : le logiciel donne la P-value ou niveau de significativité = 0.734908 valeur nettement supérieure à 0.05 , on décide donc H0. Conclusion générale: la loi est symétrique et possède une Kurtosis de 3 c'est donc une loi Normale.



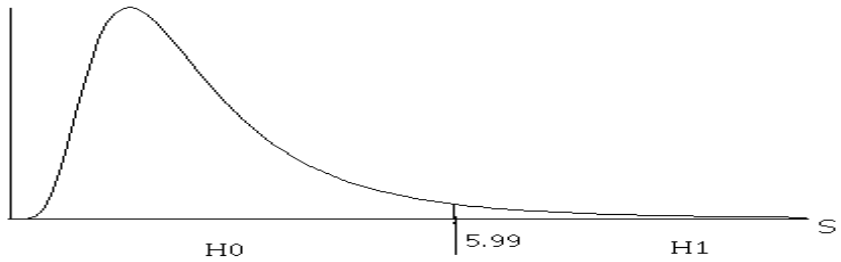
- test global de Jarque et Berra

Il teste donc globalement

$$\begin{aligned} H_o &: \mu_3 = 0 \text{ et } \mu_4 = 3\sigma^4 \text{ soit } \alpha_3 = 0 \text{ et } \alpha_4 = 3 \\ H_1 &: \text{l'une au moins de ces deux propriétés n'est pas vérifiée} \end{aligned}$$

sous l'hypothèse H_0 vraie la variable aléatoire S somme des carrés des deux précédents résultats centrés réduits suit donc le loi du χ^2_2 à deux degrés de liberté.

$$S = \frac{n}{6}\widehat{\alpha_3}^2 + \frac{n}{24}(\widehat{\alpha_4} - 3)^2$$



Le résultat calculé par RATS est S=0.120572 nettement inférieur à la borne 5.99, on décide donc H1.
On peut aussi utiliser le niveau de significativité : il est de 0.941495 nettement supérieur à 5%, on décide donc H0 la loi est une loi Normale.

3 Conséquence de points aberrants dans le modèle

Pour illustrer ce cas assez fréquent en pratique dont nous avons déjà parlé dans le chapitre 2 partie sur les résidus, nous allons reprendre l'exemple residu.prg
end xxx
*** pas de calendrier car les données ne sont pas des chroniques**
*** le fichier de données est residu.rat**

all 140
open data residu.rat
data(for=rats) /
smpl 1 140
source points_ab.src
@points_ab Y 1 140
 # constant X1 X2
****** Estimation du modèle de base**
smpl 1 140
lin Y / res
 # constant X1 X2
stat res

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   137
Centered R**2     0.999997      R Bar **2   0.999997
Uncentered R**2   1.000000      T x R**2     140.000
Mean of Dependent Variable      692254.35129
Std Error of Dependent Variable 206119.59008
Standard Error of Estimate         364.83023
Sum of Squared Residuals        18234850.449
Regression F(2,137)             22184025.9745
Significance Level of F           0.00000000
Log Likelihood                   -1023.05559
Durbin-Watson Statistic             1.935119

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                792.80956360 128.35876356      6.17651  0.00000001
2.  X1                        0.80633215   0.00360655    223.57462  0.00000000
3.  X2                        0.49960780   0.00025873   1931.01707  0.00000000

Statistics on Series RES
Observations                 140
Sample Mean             0.000000      Variance            131185.974452
Standard Error        362.196044      of Sample Mean      30.611153
t-Statistic (Mean=0)    0.000000      Signif Level        1.000000
Skewness               -2.583469      Signif Level (Sk=0) 0.000000
Kurtosis (excess)      18.304580      Signif Level (Ku=0) 0.000000
Jarque-Bera          2110.236793      Signif Level (JB=0) 0.000000
```

Le test de symétrie indique un niveau de significativité ns=0.0000<0.05 on décide donc H1 la loi n'est pas symétrique
Le test de la Kurtosis indique lui aussi un NS=0.000 on décide donc H1 la kurtosis n'est pas égale à 3
Le test de Jarque et Bera donne le même résultat, la loi n'est donc pas une loi normale.
Avant de sangloter sur la perte de Normalité il faut toujours regarder la présence d'éventuels points aberrants.
********* après traitement du point aberrant en t=100**
set DU100 = t.eq.100
lin Y 1 140 res2
 # constant X1 X2 du100
stat res2

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   136
Centered R**2     0.999998      R Bar **2   0.999998
Uncentered R**2   1.000000      T x R**2     140.000
Mean of Dependent Variable      692254.35129
Std Error of Dependent Variable 206119.59008
Standard Error of Estimate         286.95323
Sum of Squared Residuals        11198532.934
Regression F(3,136)             23906116.9066
Significance Level of F           0.00000000
Log Likelihood                    -988.92696
Durbin-Watson Statistic             1.703769

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                  849.671151   101.146394      8.40041  0.00000000
2.  X1                          0.802175     0.002872    279.29821  0.00000000
3.  X2                          0.499928     0.000206   2421.87131  0.00000000
4.  DU100                   -2700.442894   292.128141     -9.24404  0.00000000

Statistics on Series RES2
Observations                140
Sample Mean            0.000000      Variance            80564.985136
Standard Error       283.839717      of Sample Mean      23.988834
t-Statistic (Mean=0)   0.000000      Signif Level        1.000000
Skewness               0.040280      Signif Level (Sk=0) 0.847362
Kurtosis (excess)      0.191060      Signif Level (Ku=0) 0.652721
Jarque-Bera            0.250798      Signif Level (JB=0) 0.882145
```

Le test de symétrie indique un niveau de significativité ns=0.847>0.05 on décide donc H0 la loi est symétrique
Le test de la Kurtosis indique lui aussi un NS=0.6527 on décide donc H0 la kurtosis est à 3
Le test de Jarque et Bera donne le même résultat, la loi est donc bien une loi normale.
Ces résultats sont assez fréquents en pratique voilà pourquoi il **faut dans une étude toujours commencer par la recherche d'éventuels points aberrants**.