LES RESIDUS

PROPRIETES, INTERVALLES DE CONFIANCE ET POINTS ABERRANTS

Note : On va utiliser les propriétés H_0 et H_1 des erreurs.

1 Rappels des propriétés des résidus

Les résidus $e_t = Y_t - \widehat{Y}_t$ sont les écarts entre les valeurs réalisées de la variable endogène et les valeurs estimées par la méthode des MCO.

1.1 Propriétés géométriques

Par hypothèse des MCO le vecteur des résidus $\overrightarrow{e}$ est orthogonal à la variété linéaire engendrée par les variables explicatives du modèle. La variété linéaire H_k est un sous-espace de R^n de dimension k si le modèle a k variables explicatives (y compris la constante). Comme le vecteur des résidus est orthogonal à H_k il appartient au complémentaire de H_k dans R^n noté H_{n-k} ce vecteur ayant n composantes appartient donc à un espace de dimension (n-k) , de plus il étant orthogonal à H_k il est donc orthogonal à tous les vecteurs de H_k et en particulier aux vecteurs des k variables explicatives qui définissent H_k , par conséquent le produit scalaire de ce vecteur avec les vecteurs des variables explicatives est nul.
Le vecteur $\overrightarrow{\widehat{Y}} = X\overrightarrow{\widehat{a}} = X({}^tXX)^{-1}\,{}^tX\overrightarrow{Y} = \underset{(n,n)}{N}\overrightarrow{Y}$ appartient au sous-espace vectoriel H_k car $\overrightarrow{\widehat{Y}} = \widehat{a_1}\overrightarrow{X_1} + ... + \widehat{a_k}\overrightarrow{X_k}$ est une combinaison linéaire des vecteurs des variables explicatives.

1.2 Conséquences de ces propriétés

La matrice N est la matrice de projection sur H_k, en effet on constate que $N = N^2 = {}^tN$, donc bien qu'ayant n lignes et n colonnes elle est de rang inférieur à n , comme H_k est défini par les k vecteurs elle est de rang k.
Le vecteur résidus $\overrightarrow{e} = \overrightarrow{Y} - \overrightarrow{\widehat{Y}} = \overrightarrow{Y} - X\overrightarrow{\widehat{a}} = \overrightarrow{Y} - X({}^tXX)^{-1}\,{}^tX\overrightarrow{Y} = (I_n - X({}^tXX)^{-1}\,{}^tX)\overrightarrow{Y} = M\overrightarrow{Y}$ où la matrice M a aussi n lignes et n colonnes. On remarque que $M = M^2 = {}^tM$, la matrice M est donc une matrice de projection orthogonale sur H_{n-k} et effet $I_n = N + M$ si N est matrice de projection sur H_k sa complémentaire (la somme redonne la matrice identité) est la matrice de projection sur H_{n-k} . Cette matrice ayant n lignes et n colonnes est donc de rang (n-k).
On a $\overrightarrow{e} = (I_n - X({}^tXX)^{-1}\,{}^tX)\overrightarrow{Y} = (I_n - X({}^tXX)^{-1}\,{}^tX)(X\overrightarrow{a} + \overrightarrow{\epsilon}) = X\overrightarrow{a} - X({}^tXX)^{-1}\,{}^tXX\overrightarrow{a} + M\overrightarrow{\epsilon} = M\overrightarrow{\epsilon}$
On en déduit donc que M étant la matrice de projection sur H_{n-k} le vecteur résidu est aussi la projection du vecteur erreur $\overrightarrow{\epsilon}$ sur H_{n-k} .

1.3 Propriétés statistiques des résidus

Etude des propriétés statistiques des résidus dans le cas où les erreurs suivent les **hypothèses classiques** définies dans la partie précédente.

1.3.1 Rappel

Nous avons déjà démontré que si on impose aux erreurs les propriétés classiques alors la matrice de variances-covariances des erreurs s'écrit $V_{\overrightarrow{\epsilon}} = \sigma^2 I_n$ et comme $\overrightarrow{e} = M\overrightarrow{\epsilon}$ alors $V_{\overrightarrow{e}} = \sigma^2 M$. Si de plus on impose aux erreurs de suivre une loi normale nous avons vu que le vecteur des résidus suit une loi normale dégénérée de dimension (n-k). Chaque résidu e_t ne suit pas une loi normale.

1.3.2 Comparaison avec les erreurs

$V_{\overrightarrow{e}} = \sigma^2 M = \sigma^2(I_n - N)$entraîne si on note h_{ij} l'élément générique de la matrice N, que la variance de chaque résidu est $V(e_t) = \sigma^2(1 - h_{tt})$ et $\mathrm{Cov}(e_i, e_j) = \sigma^2(1 - h_{ij})$.

- $V(e_t) = \sigma^2(1 - h_{tt})$ entraîne que les résidus n'ont pas la même variance car $h_{tt} \neq h_{t't'}$ contrairement aux erreurs qui sous les hypothèses de base ont la même variance σ^2. les résidus sont donc en théorie toujours hétéroscédastiques.
- $\mathrm{Cov}(e_i, e_j) = \sigma^2(1 - h_{ij})$ entraîne que les covariances des résidus ne sont en général pas nulles car $h_{ij} \neq 1$ là aussi contrairement aux erreurs qui ont des covariances nulles sous les hypothèses de base des MCO.
- Si les erreurs suivent la loi Normale les résidus ne la suivent pas.

En conclusion les résidus n'ont aucune des propriétés des erreurs qui leur correspondent. Sauf leur espérance qui est nulle dans les deux cas si les variables explicatives sont non aléatoires. En effet on a $E(\overrightarrow{\epsilon}) = \overrightarrow{0}$ par hypothèse, et $E(\overrightarrow{e}) = E(M\overrightarrow{\epsilon}) = ME(\overrightarrow{\epsilon}) = \overrightarrow{0}$. Et c'est pourtant à l'aide des résidus que l'on testera les propriétés des erreurs. Il existent d'autres sortes de résidus qui ont les mêmes propriétés que les erreurs ce sont les résidus récursifs que l'on étudiera dans un autre chapitre.

2 La notion d'intervalle de confiance

2.1 Loi asymptotique des résidus

Si n est grand devant k on aura n-k proche de n et on fera l'approximation que les résidus e_t suivent asymptotiquement une loi $N(0, \sigma^2(1 - h_{tt}))$, soit que $e_t/\sigma\sqrt{(1 - h_{tt})}$ suit asymptotiquement une loi Normale centrée réduite. On estime σ^2 par $s^2 = SRC/(n - k)$ et alors $e_t/s\sqrt{(1 - h_{tt})}$ suit asymptotiquement une loi de Student à (n-k) degrés de liberté, or l'asymptotique de la loi de Student est la loi normale centrée réduite donc

$$\frac{e_t}{s\sqrt{(1 - h_{tt})}} \text{ suit asymptotiquement une loi } N(0, 1)$$

On l'appelle le résidu studentisé

2.2 Intervalle de confiance

Donc au risque α on construit l'intervalle de confiance asymptotique pour chaque résidu

$$-t_\alpha < \frac{e_t}{s\sqrt{(1-h_{tt})}} < t_\alpha$$

On constate que cet intervalle peut aussi s'écrire

$$-t_\alpha\sqrt{(1-h_{tt})} < \frac{e_t}{s} < t_\alpha\sqrt{(1-h_{tt})}$$

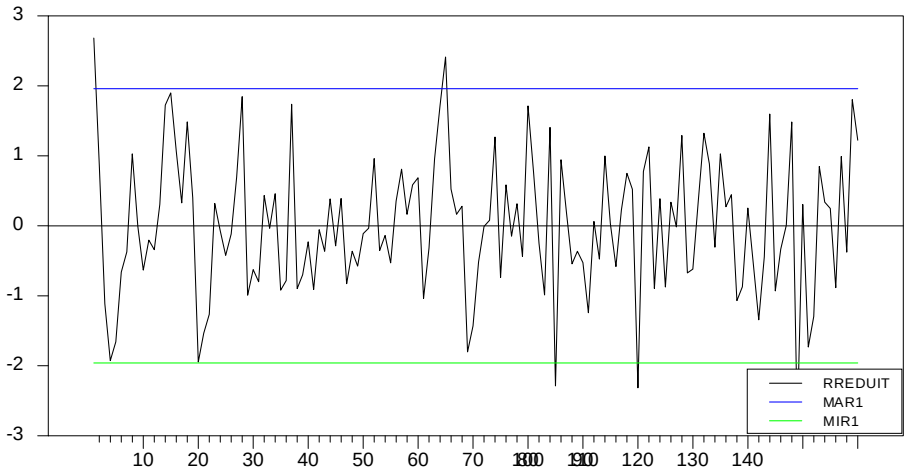
, on construit alors pour chaque résidu un intervalle différent puisqu'il dépend de h_{tt}
Or si n est grand on constate que h_{tt} est proche de 0 (on le calculera dans des exemples), on est donc souvent amené à faire l'approximation suivante

$$-t_\alpha < \frac{e_t}{s} < t_\alpha \quad \text{On appelle } \frac{e_t}{s} \text{ le résidu standardisé}$$

Cet intervalle peut s'écrire $-t_\alpha s < e_t < -t_\alpha s$, dans l'échantillon t_α et s étant fixes on constate ici que l'intervalle est le même pour tous les résidus , on obtient donc ce que l'on appelle une bande de confiance, idem bien sur pour le résidu standardisé au-dessus.

2.3 Sous-programme de construction de l'intervalle de confiance

On va utiliser le sous-programme points_ab.src qui va construire les intervalles de confiance au risque $\alpha = 0.05$. Si vous souhaitez un risque supérieur à 5% il suffit de regarder dans les tableaux fournis par le sous-programme et comparer non plus avec 1.96 mais avec la borne de la loi normale correspondant au risque choisi (voir dans l'exemple).
On travaille sur les données mco.rat et le programme points_aberrants.prg.
end xxx
*** on utilise les mêmes données que dans le chapitre MCO**
*** pas de calendrier car les données ne sont pas des chroniques**
*** le fichier de données est mco.rat**
all 140
open data mco.rat
data(for=rats) /
source points_ab.src
@points_ab Y 1 140
 # constant X1 X2
Le logiciel fournit le graphe des résidus divisés par S et les bornes de l'intervalle 1.96 et -1.96 donc ce qui correspond au résidu standardisé. Normalement les résidus standardisés doivent se trouver dans la bande de confiance -1.96,1.96 si le risque est de 5%
On peut sauvegarder ce graphique en utilisant comme d'habitude SAVE AS et en lui donnant un nom. On peut aussi si
Afin de mieux repérer les points qui sortent éventuellement de l'intervalle de confiance on trouve dans le fichier des sorties les points sortant de cet intervalle.



- On donne un premier tableau utilisant la valeur exacte de la variance des résidus, ce qui donne comme on l'a vu plus haut l'intervalle de confiance

$$-t_\alpha\sqrt{(1-h_{tt})} < \frac{e_t}{s} < t_\alpha\sqrt{(1-h_{tt})}$$

pour le résidu standardisé $\frac{e_t}{s}$

```
liste des points non compris  dans l intervalle de confiance
****************************************************************
 date     -1.96*sqrt(1-htt)   Residus/S     1.96*sqrt(1-htt)
****************************************************************
1             -1.92957        2.68388        1.92957
20            -1.94148       -1.94655        1.94148
65            -1.94095        2.41314        1.94095
85            -1.95183       -2.28827        1.95183
100           -1.92528       -2.31546        1.92528
129           -1.93369       -2.86208        1.93369
```

- Le second tableau correspond à l'approximation h_{tt} proche de 0 et donc à l'intervalle de confiance

$$-t_\alpha < \frac{e_t}{s} < -t_\alpha$$

```
liste des points non compris  dans l intervalle de confiance avec approximation de la variance
****************************************************************
 date          -1.96          Residus/S       1.96
****************************************************************
1             -1.96000        2.68388        1.96000
65            -1.96000        2.41314        1.96000
85            -1.96000       -2.28827        1.96000
100           -1.96000       -2.31546        1.96000
129           -1.96000       -2.86208        1.96000
```

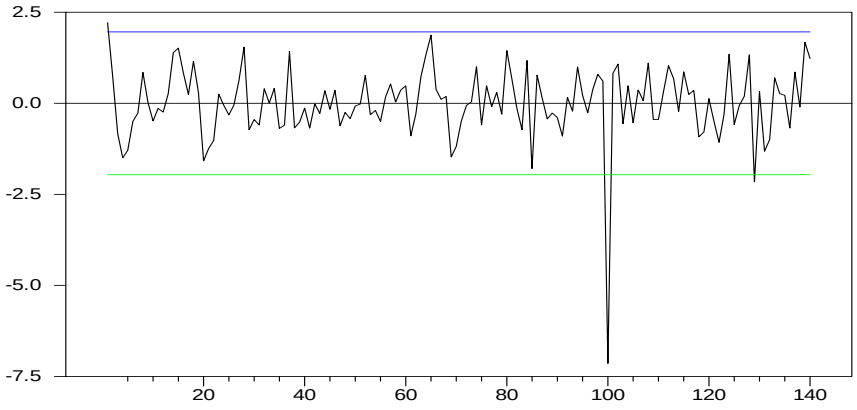
CONCLUSION: dans cet exemple on constate tout d'abord que les valeurs $1.96\sqrt{(1-h_{tt})}$ sont assez proches de 1.96; ensuite on remarque que les résidus sortant de l'intervalle sont les mêmes sauf le numéro 20 qui correspond à une valeur très proche de -1.96.
REMARQUE: si on souhaite prendre un autre risque que 5% , il suffit d'ouvrir le fichier points_ab.src et de changer COM TALPHA = 1.96 par la valeur correspondant au risque souhaité (par exemple 1.64 si vous voulez un intervalle de confiance au risque 10%).

3 La notion de point aberrant

3.1 Un exemple

Comme on vient de le voir à l'aide de l'intervalle de confiance, dans certains modèles des points sortent de l'intervalle de confiance. En économie ce sont des points qui correspondent souvent à des phénomènes arrivés à certaines périodes, qui sont exceptionnels, qui normalement ne devront plus se reproduire et donc dont on ne souhaite pas voir l'impact influer sur les estimations du modèle.

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
lin Y / res
 # constant X1 X2



On obtient le graphe des résidus et l'intervalle de confiance.

```
 on utilise les residus divises par S
liste des points non compris  dans l intervalle de confiance
************************************************************
 date     -talpha*sqrt(1-htt) Residus/S     talpha*sqrt(1-htt)
************************************************************
1             -1.92957        2.21436        1.92957
100           -1.92528       -7.14200        1.92528
129           -1.93369       -2.15517        1.93369

liste des points non compris  dans l intervalle de confiance
avec approximation de la variance
************************************************************
 date          -talpha        Residus/S       talpha
************************************************************
```

```
1             -1.96000        2.21436        1.96000
100           -1.96000       -7.14200        1.96000
129           -1.96000       -2.15517        1.96000
```

On constate un point qui sort très nettement de l'intervalle de confiance pour t=100, les autres sont négligeables par rapport à celui-ci.

```
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
```

3.2 Quand parle-t-on de point aberrant ?

- Un point aberrant est donc tout d'abord un point sortant nettement de l'intervalle de confiance.

- Il faut ensuite que les points aberrants ne soient pas nombreux. Ici on en trouve trois points mais on constate que deux points sont très proches des bornes et on n'en tiendra pas compte. Si dans un modèle ont trouve une grande proportion de points aberrants c'est que le modèle en lui-même est mauvais, il ne fait alors pas traiter ces points mais réfléchir à la construction d'un meilleur modèle. On trouve ce genre de souci par exemple si on essaie de modéliser le cours d'une variable boursière, avec un modèle simple souvent près de la moitié des résidus peuvent être considérés comme aberrants, le modèle est alors à rejeter au profit d'un modèle plus sophistiqué tenant compte de la volatilité de ces séries.

- Il faut aussi et surtout pouvoir expliquer économiquement ce point aberrant. Puisque par définition il correspond à un phénomène que l'on ne souhaite pas prendre en compte dans l'estimation du modèle car ce phénomène n'a aucune raison de se reproduire, il faut donc expliquer pourquoi il apparaît et montrer ainsi que c'est seulement un accident de parcours. La guerre du golfe, une grève importante, une machine hors d'usage dans l'entreprise ... peuvent être considérés comme des phénomènes rares et non pris en compte dans l'estimation. Tout point aberrant doit donc faire l'objet d'une explication. S'il n'y a aucune explication c'est que nous n'avons pas affaire à un point aberrant mais c'est encore ici le modèle qu'il faut critiquer: oubli de variables, modèle trop simpliste ....

3.3 Quelles sont les conséquences d'un point aberrant ?

- La première conséquence est le poids que ce point peut avoir dans l'estimation des coefficients et donc aussi dans les prévisions.

- La seconde est que ces points aberrants ayant des résidus très forts, ils peuvent modifier les tests classiques qui sont basés sur ces résidus. Un exemple classique est celui du test de normalité qui est très sensible à la présence de ces points. Nous en verrons des exemples dans le chapitre 4 partie 1 sur les tests de Normalité.

4 La notion de variable muette

Que faire quand on quelques points aberrants répondant aux définitions ci-dessus?

- L'idée qui vient en premier est de supprimer ce ou ces points, mais elle n'est pas très appliquée car elle provoque des "trous" dans les séries et outre que les économistes n'aiment pas cela dans le cadre des séries chonologiques (en fonction du temps) elles perturbent aussi certains tests statistiques.

- L'autre solution, la plus utilisée est de créer ce que l'on appelle une variable muette (dummy variable) qui a des 0 partout sauf au point aberrant où l'on met la valeur 1. Il faut une variable muette par point aberrant en général, sauf dans des cas très particuliers où plusieurs points aberrants correspondent au même phénomène ; par exemple une grève très importante des postes a provoqué une baisse très importante du niveau d'activité un trimestre (j'ai oublié l'année) puis une hausse identique le trimestre suivant du au rattrapage, il est possible dans des cas très rares comme celui-ci de ne créer qu'une seule variable muette avec 0 partout, sauf 1 le trimestre de la grève et -1 le trimestre de rattrapage suivant.

Application des deux possibilités dans l'exemple vu plus haut.

4.1 On retire le point

On n'a donc plus que n=140-1=139 éléments dans l'échantillon.
Comment programmer le fait que l'on enlève ici t=100 c'est-à-dire que l'on passe du numéro 99 à 101

```
********* on enlève la valeur en t=100
*** pour cela on crée une variable Z égale à Y jusqu' à 99
*** et égale à Y(t+1) jusqu' à 139 idem pour les deux variables explicatives X1 et X2
set Z = Y
set Z1 = X1
set Z2 = X2
do i=100,139,1
com Z(i) = Y(i+1)
com Z1(i) = X1(i+1)
com Z2(i) = X2(i+1)
```

end do i
**** l'échantillon est donc de 139 maintenant**
smpl 1 139
lin Z
 # constant Z1 Z2

```
Linear Regression - Estimation by Least Squares
Dependent Variable Z
Usable Observations    139      Degrees of Freedom   136
Centered R**2     0.999998      R Bar **2   0.999998
Uncentered R**2   1.000000      T x R**2     139.000
Mean of Dependent Variable      691298.00010
Std Error of Dependent Variable 206553.08489
Standard Error of Estimate         286.95323
Sum of Squared Residuals        11198532.934
Regression F(2,136)             35751100.3162
Significance Level of F           0.00000000
Log Likelihood                    -982.36140
Durbin-Watson Statistic             1.699181

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                 849.67115079 101.14639383      8.40041  0.00000000
2.  Z1                         0.80217529   0.00287211    279.29821  0.00000000
3.  Z2                         0.49992775   0.00020642   2421.87131  0.00000000
```

Le résultat est un peu différent du résultat ci-dessus avec les 140 points, mais dans certaines études, l'écart entre les coefficients estimés peut être plus important.

4.2 On construit une variable muette

Suite du programme précédent:
******* autre solution créer une variable muette**
smpl 1 140
set DU100 = t.eq.100
lin Y 1 140 res2
 # constant X1 X2 du100
**** vérification de la variable muette du100**
pri / du100
ATTENTION: bien remettre SMPL 1 140 sinon RATS ne prendrait que de 1 à 139 et en particulier ne définirait DU100 que sur la période 1 139. Il faut toujours bien vérifier que l'on est dans la bonne période c'est-à-dire que le SMPL est bon. C'est une remarque très importante en pratique.
Si on était en données trimestrielles avec un point aberrant en 1974:4 on écrirait **set DU19744 = t.eq.1974:4** , L'INSEE a cette bonne habitude de mettre dans le nom de la variable la date du point aberrant ce qui facilite la lecture du résultat. Vous pouvez vérifier en regardant les résultats de PRI / DU100 que cette variable a des 0 partout sauf pour t=100 où la valeur est 1.

```
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
```

4.3 Aucune différence entre les deux résultats

On constate qu'enlever le point aberrant ou créer une variable muette donne les mêmes résultats. Dans le premier cas on a une taille d'échantillon de 139 dans le second n=140. Cependant les coefficients estimés sont identiques, leurs "t de Student" également. La somme des carrés des résidus est la même. Cette dernière remarque indique que si SCR_{139}=SCR_{140} , comme les coefficients estimés sont identiques les 139 résidus sont aussi identiques cela conduit au fait que le résidu en t=100 est nul (vous pouvez le vérifier en regardant le graphe des résidus RES2), toute l'information sur ce point se retrouve dans la valeur du coefficient de DU100 soit -2700.44, ce point n'influe donc sur aucun élément du modèle estimé. Le $S_{139} = S_{140}$ =286.95323, en effet $\mathrm{S}^2_{139} = \frac{SCR_{139}}{139-3}$ et $\mathrm{S}^2_{140} = \frac{SCR_{140}}{140-4}$ ces deux S^2 sont donc identiques.
On remarque ici que le coefficient de DU100 est très significatif "t de Student"=-9.24; si on trouvait un coefficient non significatif le point ne serait pas un point aberrant cela voudrait dire que l'on s'est trompé dans la lecture des points aberrants.