CORRIGE DE MCO.PRG

C. BABUSIAUX 10/03/2007

Les données ne sont pas des séries chronologiques. Elles sont numérotées de 1 à 140.
Si la numérotation n'a pas de sens économique (numérotation au fur et à mesure de l'arrivée des données par exemple) alors les données ne sont pas classées.
Modèle Y=a+bX1+erreur
La matrice X a n=140 lignes et deux colonnes, la première n'est composée que de 1 et la seconde contient la série X1.

Matrice ${}^tXX$

```
Cross-Product Matrix
             Constant           X1
Constant              140        14391681
X1                14391681   1578209728081
```

Matrice de corrélation entre Y et les variables explicatives autres que la constante car les variables sont centrées pour le calcul des corrélations.

```
Correlation Matrix
            Y              X1
Y  1.000000000000 0.957017820178
X1 0.957017820178 1.000000000000
```

la corrélation entre Y et Y est de 1 ; de même que la corrélation entre X1 et lui-même.
La corrélation entre Y et X1 est de .957 ; elle est bien sur égale à la corrélation entre X1 et Y.

Résultats des MCO

VOIR LES COMMENTAIRES SUR LES SORTIES DANS LE FICHIER COMMENTAIRES LINREG.PDF

```
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
```

On constate que n=140 k=2 (deux variables explicatives l'unité et X1) , le nombre de degrés de liberté n-k=138
Le R2 centré = 0.9158 , le R2 corrigé ou R2 bar = 0.915274
Le R2 non centré n'est pas utilisable ici car on a un modèle avec terme constant.
Le TxR2 est la taille de l'échantillon multipliée par le R2 non centré (utilisé dans certains tests d'autocorrélation).
La moyenne de la variable endogène Y est de 692268.63700
L'écart-type de cette variable endogène est 206128.93876
Le SEE noté S est tel que S^2=Somme des carrés des résidus/(n-k) = SRC/(n-k)
Le S = $\sqrt{SCR/(n-k)}$ = 59999.61073
La somme des carrés des résidus est SCR=4.96794e+11
Les éléments suivants seront étudiés par la suite.
Le R2 est bon , le S semble important mais il dépend des unités de Y , il n'aura dans de signification que par rapport à un autre modèle ayant même taille et même Y.
L'estimation de a0 est -68443.58769 , celle de a1 = 7.40009
Pour chaque t la valeur estimée de Y est donc Yestimé(t) = -68443.58769+7.40009*X1(t)
Pour chaque t l'écart entre Y et sa valeur estimée est le résidu(t) = Y(t)-Yestimé(t)

Statistiques sur le vecteur RESIDUS des résidus du modèle :

```
Statistics on Series RESIDUS
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
```

Nous constatons que la moyenne des résidus est nulle, c'est le cas dans tous les modèles qui ont un terme constant. Les autres résultats seront étudiés par la suite.
Si on reprend le modèle sans terme constant :

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   139
Centered R**2     0.908933      R Bar **2   0.908933
Uncentered R**2   0.992632      T x R**2     138.969
Mean of Dependent Variable      692268.63700
Std Error of Dependent Variable 206128.93876
Standard Error of Estimate       62204.18679
Sum of Squared Residuals         5.37841e+11
Log Likelihood                   -1743.49446
Durbin-Watson Statistic             0.091286

   Variable                    Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  X1                      6.7759518981 0.0495150542    136.84630  0.00000000

Statistics on Series RESIDUS1
Observations                 140
Sample Mean            -4283.778634      Variance            3850878075.125581
Standard Error         62055.443558      of Sample Mean      5244.642215
t-Statistic (Mean=0)     -0.816791      Signif Level        0.415445
```

```
Skewness                0.815594      Signif Level (Sk=0) 0.000097
Kurtosis (excess)       0.243664      Signif Level (Ku=0) 0.566051
Jarque-Bera            15.867521      Signif Level (JB=0) 0.000358
```

On constate que le coefficient estimé de X1 est différent de celui du modèle avec terme constant, ici seul le R2 non centré peut être utilisé. On remarque aussi que la moyenne des nouveaux résidus n'est pas nulle mais égale à -4283.7786

SI ON CHANGE LES UNITES

- Z1=X1/100

```
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.  Z1                         740.00884     19.09057     38.76305  0.00000000
```

On constate que tous les résultats sont identiques au modèle de base avec X1 au lieu de Z1. la seule différence est la valeur du coefficient de Z1 qui est multiplié par 100 .
Yestimé sera le même. Démonstration :
Modèle de base : Y = a0+a1*X1+ erreur
Changement en multipliant et divisant par 100 : Y = a_0+a_1*100*X1/100+erreur. Le modèle reste le même. Si on pose X1/100=Z1, on obtient Y = a_0+a_1*100*Z1+erreur, on remarque bien ainsi que le coefficient de Z1 est a1 multiplié par 100. Conclusion, la valeur des coefficients dépend des unités des variables explicatives, en conséquence on ne peut dire simplement en regardant sa valeur si un coefficient est petit ou non, puisqu'il suffit de diviser les unités de la variable correspondante par 100 pour que ce coefficient soit multiplié par 100 et paraisse alors plus grand, il faut pour décider si un coefficient peut être considéré comme nul faire des tests.

- On divise Y par 100 Y100=Y/100

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y100
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      6922.6863700
```

```
Std Error of Dependent Variable 2061.2893876
Standard Error of Estimate       599.9961073
Sum of Squared Residuals        49679355.375
Regression F(1,138)                1502.5742
Significance Level of F           0.00000000
Log Likelihood                   -1093.21343
Durbin-Watson Statistic             0.120471

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                -684.4358769  202.6923173     -3.37672  0.00095323
2.  X1                         0.0740009    0.0019091     38.76305  0.00000000
```

Modèle de base : $Y = a_0 + a_1 * X1 + \text{erreur}$
Si on divise Y par 100 tout le modèle est divisé par 100. Il devient :
$Y/100 = a_0/100 + (a_1/100) * X1 + \text{erreur}/100$
On constate que l'estimation de a_0 est divisée par 100 comme celle de a_1
La moyenne de l'endogène divisée par 100 comme son écart-type (car les erreurs sont aussi divisées par 100). Les résidus étant divisés aussi par 100 on constate que le nouveau S est divisé par 100 tandis que la SCR est divisée par 10000.

SI ON CENTRE LE MODELE

```
Linear Regression - Estimation by Least Squares
Dependent Variable YC
Usable Observations    140      Degrees of Freedom   139
Centered R**2     0.915883      R Bar **2   0.915883
Uncentered R**2   0.915883      T x R**2     128.224
Mean of Dependent Variable           -0.0000
Std Error of Dependent Variable  206128.9388
Standard Error of Estimate        59783.3952
Sum of Squared Residuals         4.96794e+11
Log Likelihood                   -1737.93725
Durbin-Watson Statistic             0.120471

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  XC                      7.4000883884 0.1902177721     38.90324  0.00000000
```

On retrouve la même estimation de a_1.

```
Linear Regression - Estimation by Least Squares
Dependent Variable YC
Usable Observations    140      Degrees of Freedom   138
Centered R**2     0.915883      R Bar **2   0.915274
Uncentered R**2   0.915883      T x R**2     128.224
Mean of Dependent Variable           -0.0000
Std Error of Dependent Variable  206128.9388
Standard Error of Estimate        59999.6107
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                   0.0000000 5070.8926293  1.61652e-14  1.00000000
2.  XC                         7.4000884    0.1909057     38.76305  0.00000000
```

Si on ajoute une constante à un modèle centré on remarque que le coefficient de la constante est nul puisque l'influence de la constante est déjà prise en compte en centrant.
Démonstration :
Pour chaque t Y(t) = a_0+a_1*X1(t)+erreurs(t) la moyenne Ym = a_0+a_1*Xm+erreursm
Pour chaque t Y(t)-Ym = a_0+a_1*X1(t)+erreurs-(a_0+a_1*Xc+erreursc)
Soit Y(t)-Ym = a_1*(X1(t)-Xm)+erreurs-erreursm . La constante disparaît, mais elle est toujours présente via la transformation.
Le R2 est le même que dans le modèle non centré de base , de plus ici le R2 centré et non centré donne les mêmes résultats. On peut retrouver cette constante estimée par

$$\widehat{a_0} = Y_m - \widehat{a_1} X_m$$

SI ON EXPLIQUE X1 EN FONCTION DE Y

```
Linear Regression - Estimation by Least Squares
Dependent Variable X1
Usable Observations    140      Degrees of Freedom   138
Centered R**2     0.915883      R Bar **2   0.915274
Uncentered R**2   0.994735      T x R**2     139.263
Mean of Dependent Variable      102797.72143
Std Error of Dependent Variable  26657.66370
Standard Error of Estimate        7759.46092
Sum of Squared Residuals        8308874266.1
Regression F(1,138)                1502.5742
Significance Level of F           0.00000000
Log Likelihood                   -1451.57772
Durbin-Watson Statistic             0.134271

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                17118.048748  2305.577127      7.42463  0.00000000
2.  Y                           0.123767     0.003193     38.76305  0.00000000
```

On constate dans le seul cas comme ici d'une variable explicative que les R2 sont identiques.
SI ON CHANGE L'ORDRE DES DONNEES
La fonction ORDER classe les couples X1,Y , on choisit d'effectuer le classement en fonction croissante de X1. Afin de mieux voir la transformation et ne pas modifier l'ordre initial des données, il est préférable de donner un autre nom aux séries ordonnées. Dans cet exemple on nomme YO et X1O les séries Y et X1 ordonnées.

```
Linear Regression - Estimation by Least Squares
Dependent Variable YO
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             1.695235

   Variable                     Coeff       Std Error      T-Stat     Signif
```

```
*****************************************************************************
1.  Constant                -68443.58769  20269.23173     -3.37672  0.00095323
2.  X1O                          7.40009      0.19091     38.76305  0.00000000
```

Tous les résultats des MCO sont identiques (à l'exception du Durbin-Watson qui tient compte de l'ordre)

SI ON AJOUTE UNE VARIABLE

La matrice X des variables explicatives a maintenant 140 lignes et 3 colonnes (la colonne unité, la colonne de la série X1 et celle de la série X2)

```
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
```

On constate beaucoup de changements par rapport au modèle contenant seulement X1. Tout d'abord les coefficients estimés sont très différents : le coefficient de la constante passe de -68443.587 à 834.92 , celui de X1 passe de 7.4 à 0.803 , un grand changement qui est du à un problème de colinéarité entre X1 et X2. Ce problème pose de gros soucis en économie car on constate ainsi qu'il est très difficile d'interpréter économiquement la valeur même des coefficients puisqu'ils sont susceptibles de modifications par l'ajout de variables.
ON PEUT COMPARER CES DEUX MODELES CAR ILS ONT LE MEME ECHANTILLON ET LA MEME VARIABLE ENDOGENE Y. Sinon la comparaison est impossible (sauf pour des modèles en LOG, voir le théorème de BOX-COX. dans un chapitre suivant).
Le R2 a augmenté, c'est normal car on sait que si on ajoute des variables, mécaniquement le R2 va augmenter. Il augmente car la SCR diminue, en effet si on ajoute une variable elle va mécaniquement expliquer un peu de Y (sauf si elle est orthogonale) et donc faire baisser les résidus donc la somme de leurs carrés. Maximiser le R2 est identique à minimiser la SCR. C'est pour cela que l'on ne regarde pas le R2 mais le R2 bar ou R2 corrigé car il tient compte du nombre de variables explicatives.
Définition du R^2 corrigé ou $\overline{R^2}$

$$\overline{R^2} = 1 - \frac{SCR/(n-k)}{\sum(Y_t - \overline{Y})^2/(n-1)}$$

Comme n est fixé (même échantillon) on constate que MAX R^2 corrigé est identique à MINIMISER SCR/(n-k)=S^2 car le dénominateur reste inchangé si Y et n sont les mêmes lors de changement de variables explicatives. C'est souvent ce dernier critère qui est utilisé.

Si k augmente alors SCR diminue, mais (n-k) diminue également, donc en pratique il faut regarder le rapport des deux pour voir comment il évolue, le reste de la formule ne changeant pas. Par exemple, si en a un premier modèle avec SRC_1 et k variables et si on ajoute 1 variable, alors on obtient SCR_2 qui est automatiquement inférieur à SCR_1 et (n-k-1) qui est inférieur à (n-k). A quelle condition a-t-on $\mathrm{S}_2^2 = SCR_2/(n-k-1) < S_1^2 = SCR_1/(n-k)$? Il faut donc :

$$\frac{SCR_2}{SCR_1} < \frac{n-k-1}{n-k}$$

On notera 1 le modèle qui est expliqué par X1 seul et 2 celui qui est expliqué par X1 et X2.
SCR1 = 4.96794e+11 SCR2 = 11671924.919 on constate la diminution des résidus due à une variable supplémentaire.
k le nombre de variables explicatives passe de 2 à 3 et le nombre de degrés de libertés de 138 à 137
$\mathrm{R}_1^2 = 0.915883$ $\mathrm{R}_2^2 = 0.999998$ il y a augmentation automatique du R2 mais on constate ici que cette augmentation est forte, le bon sens fait penser à une amélioration du modèle.
$\overline{R_1^2} = 0.915274$ $\overline{R_2^2} = 0.999998$ le modèle 2 est donc meilleur que le 1
On utile souvent le critère du S qui est rappelons-le identique à celui du R2 bar.
$\mathrm{S}_1 = 59999.61$ $\mathrm{S}_2 = 291.88$, le S_2 est très nettement plus petit que S_1, le modèle 2 est donc nettement meilleur.