LES MCO SANS HYPOTHESES SUR LES ERREURS

Note : Dans tous les chapitres vous trouverez les lignes de commande RATS en caractère gras. Pour ce chapitre les corrigés des exercices sont donnés dans le fichier corrige mco.prg.

Dans ce premier chapitre les MCO sont vus comme une technique permettant de trouver des valeurs pour approcher à l'aide l'échantillon les valeurs inconnues des coefficients du modèle. Aucune hypothèse sur les erreurs n'est utile pour cette technique.

1 LES MCO : RESUME

1.1 L'échantillon

Prenons le cas d'une variable Y expliquée par les variables X1 , X2 et un terme constant. La variable Y est dite variable endogène ou expliquée et les variables X1 et X2 sont les variables exogènes ou explicatives; dans la plupart des modèles comme ici on va ajouter un terme constant ce qui conduit à ajouter une variable explicative égale à 1. Le modèle linéaire va s'écrire :

$$Y = a_0 + a_1 X1 + a_2 X2 + erreur$$

Le terme erreur tient compte de tout ce qui n'a pas été expliqué par $a_0 + a_1 X1 + a_2 X2$ en particulier d'éventuelles variables non prises en compte. Afin d'estimer les coefficients inconnus, on doit disposer d'un échantillon de taille n. Pour chaque valeur de t comprise entre 1 et n l'équation linéaire doit être vérifiée. Cela donne un système de n équations :

$$Y_t = a_0 + a_1 X1_t + a_2 X2_t + \epsilon_t \qquad \text{pour } t = 1 \text{ à } n$$

On a donc n équations dans lesquelles on connaît les résultats de l'échantillon soient les Y_t, $X1_t$ et $X2_t$ alors que les a_0, a_1,a_2 et ϵ_t sont inconnus. Les éléments inconnus sont donc les trois coefficients a_i et les n erreurs. Ces n erreurs forment un vecteur $\overrightarrow{\epsilon}$ à n dimensions. De même les n valeurs de Y dans l'échantillon forment le vecteur $\overrightarrow{Y}$, les n valeurs de la variable X1 le vecteur $\overrightarrow{X1}$ et les n valeurs de la variable X2 le vecteur $\overrightarrow{X2}$. Le coefficient de la constante a_0 est toujours égal à 1 , il va donc former une nouvelle variable U telle que $U_t = 1$, on aura ainsi un nouveau vecteur $\overrightarrow{U}$ également de dimension n dont les composantes sont égales à 1. Les n équations du système peuvent maintenant s'écrire sous forme vectorielle:

$$\overrightarrow{Y} = a_0 \overrightarrow{U} + a_1 \overrightarrow{X1} + a_2 \overrightarrow{X2} + \overrightarrow{\epsilon}$$

Passage à la forme matricielle: On va définir une matrice X dont le nombre de colonnes est k nombre de variables (ici k=3 , l'unité, X1 et X2) et le nombre de lignes est la taille n de l'échantillon. Les colonnes de X sont les vecteurs des variables explicatives.

$$\underset{(n,k)}{X} = \begin{pmatrix} 1 & X1_1 & X2_1 \\ 1 & X1_2 & X2_2 \\ .. & .. & .. \\ 1 & X1_t & X2_t \\ .. & .. & .. \\ 1 & X1_n & X2_n \end{pmatrix} = \begin{pmatrix} \overrightarrow{U} & \overrightarrow{X1} & \overrightarrow{X2} \end{pmatrix}$$

Le vecteur des coefficients inconnus $\overrightarrow{a}$ a k composantes.

$$\underset{(k,1)}{\overrightarrow{a}} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}$$

Le système de n équations s'écrit donc pour k variables explicatives (ici k=3)

$$\overrightarrow{Y} = X\overrightarrow{a} + \overrightarrow{\epsilon}$$

1.2 Les MCO

La méthode des moindres carrés consiste à minimiser la somme des carrés des erreurs du modèle (voir les démonstrations dans votre cours). Si on nomme H_k le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les vecteurs des k variables explicatives (ici $\overrightarrow{U}$, $\overrightarrow{X1}$ et $\overrightarrow{X2}$) , le minimum est obtenu pour un vecteur orthogonal à H_k.On notera $\overrightarrow{e}$ ce vecteur orthogonal à H_k donc tel que le produit scalaire avec les vecteurs de X est nul, ce qui se traduit par ${}^tX\overrightarrow{e} = \overrightarrow{0}$, et $\overrightarrow{\widehat{a}}$ le vecteur qui correspond à $\overrightarrow{e}$

$$\begin{aligned} \overrightarrow{Y} &= X\overrightarrow{\widehat{a}} + \overrightarrow{e} \\ \overrightarrow{{}^tXY} &= {}^tXX\overrightarrow{\widehat{a}} + {}^tX\overrightarrow{e} \text{ en multipliant par } {}^tX \end{aligned}$$

Comme ${}^tX\overrightarrow{e} = \overrightarrow{0}$, en multipliant par l'inverse de ${}^tXX$ (s'il existe) on obtient le résultat des MCO :

$$\overrightarrow{\widehat{a}} = ({}^tXX)^{-1}\ {}^tXY$$

NOTE: un sous-espace H_k est engendré par les k variables explicatives du modèle et contient toutes les combinaisons linéaires de ces variables.

HYPOTHESES NECESSAIRES POUR CE RESULTAT. Pour utiliser le résultat des MCO 2 hypothèses de base seulement sont nécessaires,elles sont notées avec l'indice 0

H_0^1 : $n \geq k$

H_0^2 : la matrice X doit être de plein rang k pour que ${}^tXX$ soit inversible

En effet comme X est (n,k), d'après l'hypothèse H_0^1 le rang de X est inférieur ou égal au min(n,k) donc à k. La matrice ${}^tXX$ est (k,k) , de plus elle a le même rang que X donc k d'après H_0^2, elle est donc inversible. Si la seconde hypothèse n'est pas vérifiée X est de rang inférieur à k et il y a une ou plusieurs relations linéaires entre les variables explicatives, les MCO ne peuvent alors s'appliquer (voir l'exemple en fin de ce chapitre pour savoir comment résoudre ce problème).

1.3 LES RESULTATS DE BASE

Dans cette première étape des MCO, les seuls résultats que l'on peut commenter sont les suivants

- La valeur et le signe des coefficients estimés par les MCO
- le R^2 dans le cas d'un modèle avec terme constant. On note SCR la somme des carrés des résidus

$$R^2 = 1 - \frac{SRC}{\sum(Y_t - \overline{Y})^2} = \frac{\text{variance expliquée}}{\text{variance totale}}$$

voir détails dans votre cours.

Ce R^2 fait souvent l'objet d'une attention qu'il ne mérite pas . On sait qu'il mesure le rapport entre la variance de la variable endogène estimée et la variance réelle dans l'échantillon de cette variable endogène. Un bon (proche de 1) R^2 n'est en aucun cas le signe d'un bon modèle comme nous le verrons avec les propriétés statistiques des résultats. Mais dès ce chapitre on peut voir qu'un même modèle , suivant sa présentation peut avoir ou non un bon R^2 (voir dans l'exemple).

ON COMPARE LES R^2 SUR DES MODELES AYANT LA MEME VARIABLE ENDOGENE ET LE MEME ECHANTILLON DONC MEME VALEUR DE n, SEULES LES VARIABLES EXPLICATIVES CHANGENT.

La propriété du R^2 posant problème est que plus on ajoute de variables dans le modèle plus le R^2 augmente car plus la somme des carrés des résidus diminue. Il est évident que l'ajout d'une variable non explicative dans le modèle augmente peu le R^2 , mais il est difficile de savoir jusqu'où aller dans l'ajout de variables. On peut remarquer la propriété suivante

$$Max(R^2) \text{ est identique à } Min(SCR)$$

car $\sum(Y_t - \overline{Y})^2$ reste identique si on change les variables explicatives.

La définition que l'on vient de donner du R^2 est valable seulement dans le cas de modèle ayant un terme constant, cas le plus fréquent en pratique.

- le R^2 dans le cas d'un modèle sans terme constant

$$R^2 = 1 - \frac{SCR}{\sum Y_t^2}$$

- le R^2 bar ou R^2 corrigé. Dans la suite du cours nous verrons que l'estimateur de la variance des erreurs est $S^2 = SCR/(n-k)$ et que l'on peut ainsi en déduire un R^2 beaucoup plus intéressant (voir CORRIGE MCO.PDF)

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

Pour comparer des modèles ayant le même échantillon et la même variable endogène Y avec donc seulement k et SRC qui changent, on prendra le modèle ayant le $\overline{R^2}$ maximum ou ce qui revient au même SCR/(n-k)=S^2 minimum

$$Max(\overline{R^2}) \text{ est identique à } Min(\frac{SCR}{n-k})$$

2 EXEMPLE

Les données sont dans MCO.RAT , le fichier de programme dans MCO.PRG et le **corrigé dans CORRIGE MCO.PDF**
Etude d'une variable Y en fonction tout d'abord d'une variable X1 puis de X1 et X2. Nous sommes ici dans le cas de données non chronologiques (cross-section) , elles ne sont donc pas classées.

all 140
open data mco.rat
data(for=rats) /

aller dans Wizards+show séries windows pour voir la liste des variables. Pour voir l'échantillon

pri / Y X1 X2

2.1 Les MCO dans le modèle $Y = a_0 + a_1 X_1 + \epsilon$

On va construire le modèle linéaire où la variable endogène est Y , les variables exogènes étant l'unité et X1 donc k=2 et n=140

$$\overrightarrow{Y} = X\overrightarrow{a} + \overrightarrow{\epsilon}$$

Dans ce cas particulier :
définir la matrice X
construire la matrice ${}^tXX$ (rats définit seul l'unité en mettant CONSTANT)
cmo(print) ; # constant X1
construire la matrice de corrélation
cmo(corr,print) ; # constant X1
expliquez le résultat trouvé
graphique
gra(header='Y') 1 ; # Y
nuage des points
sca(header='Y en fonction de X1') 1 ; # Y X1

2.2 résultats de la méthode des MCO

lin Y / residus
 # constant X1
définir et commenter les résultats

indiquer comment calculer le vecteur $\overrightarrow{yestime} = \overrightarrow{\widehat{Y}} = \widehat{a_0}\overrightarrow{U} + \widehat{a_1}\overrightarrow{X1}$, estimation de $\overrightarrow{Y}$ par les MCO. Ce vecteur appartient à H_k, il est en effet une combinaison linéaire des vecteurs des variables explicatives (ici le vecteur unité et le vecteur $\overrightarrow{X1}$). Dans RATS ce vecteur est obtenu par la commande PRJ suivie du nom que vous souhaitez donner au vecteur Y estimé.

prj yestime

gra(header='Y et Y estimé') 2 ; #Y ; # yestime

on constate des écarts parfois importants entre les deux. Pour chaque valeur de t l'écart entre Y_t et $yestime_t$ est le résidu noté e_t.

$$Y_t - Yestimé_t = résidu = e_t$$

On note RESIDUS le vecteur de tous ces écarts. Rats construit ce vecteur, on peut le récupérer en mettant un nom de vecteur après LIN Y / .....Ce vecteur est noté dans cet exemple RESIDUS.

Calcul de diverses statistiques sur les résidus

stat residus

pourquoi trouvez-vous une moyenne nulle ?

on reprend le modèle sans la constante, que penser du résultat ?

lin Y / residus1

 # X1

stat residus1

2.3 Quelques remarques :

2.3.1 si on change les unités :

si on divise la variable explicative X1 par 100 notée Z1: pour définir un vecteur on utilise la commande SET suivie du nom que l'on veut donner à ce vecteur cela permet de faire toutes les modifications souhaitées sur les variables à l'intérieur de RATS.

set Z1 = X1/100

lin Y / residus ; # constant Z1

indiquer ce qui a changé dans le modèle estimé

si on divise la variable expliquée par 100 en lui donnant le nom Y100

set Y100 = Y/100

lin Y100 / residus3 ; # constant X1

définir ce qui a changé dans le modèle.

2.3.2 si on centre les variables

on va centrer les variables de base, pour cela on calcule les moyennes mY et mX1 . On les récupère à l'aide la commande STATISTIQUE. on définit une valeur avec COMMPUTE (COM) my=$\overline{Y}$ et mX1 =$\overline{X1}$

stat(noprint) Y ; com my= %mean

sta(noprint) X1 ; com mX1 = %mean

puis on centre. On définit un vecteur (avec SET) pour chaque valeur de t $Yc_t = Y_t - \overline{Y}$ les valeurs centrées de Y puis de X1
set Yc = Y-my
set X1c = X1-mX1
lin Yc / residus4 ; # X1c
comparer les résultats avec le modèle de base
lin Y / residus ; # constant X1
ces deux modèles sont-ils identiques ?
comment recalculer la constante dans le modèle centré ?
quel R2 utilise-t-on dans le modèle centré ?
si on avait mis la constante dans le modèle centré ?
lin Yc / residus4 ; # constant X1c

2.3.3 si on explique X1 en fonction de Y

Que devient le modèle si on explique X1 en fonction de Y ?
lin X1 / residus5 ; # constant Y

2.3.4 l'ordre des données a-t-il une importance ?

si on classe les observations en fonction croissante de X1 va-t-on obtenir les mêmes résultats des MCO ?
Classement: ORDER (vous pouvez voir le détail de toutes les commandes de RATS dans le Help)
modèle avec les données dans l'ordre de base
lin Y / residus ; # constant X1
changement d'ordre : on va ordonner les données par exemple par ordre croissant de X1.Afin de ne pas changer l'ordre dans les données de base on va donner un autre nom aux données classées, on va leur mettre l'indice 0. Il est bien sur évident que les couples ne changent pas.
order(index=ix) X1 / Y
set YO = y(ix)
set X1O = X1(ix)
pri / X1O YO
lin YO / residus2
 # constant X1O
Quelle est la conclusion ?

2.3.5 Ajout de variables explicatives

lin Y / res
 # constant X1 X2
Comparer ce modèle avec celui de la partie I.
VOUS TROUVEREZ LE CORRIGE DANS CORRIGE MCO.PDF.
LA DEFINITION DU R^2 CORRIGE SE TROUVE DANS CE MEME FICHIER.