Rappel:

Modèle de regression multiple

⇒ Hypothèses:

* H_1 : La matrice X est non aléatoire

$\Rightarrow cov(X, \varepsilon) = 0$

* H_2 : X est de plein rang

⇒ les variables explicatives sont linéairement indépendantes ou orthogonales

$\Rightarrow (X'X)^{-1}$ existe

* $H_3 : E(\varepsilon) = 0$

* $H_4 : Var(\varepsilon) = \sigma_\varepsilon^2 . I$

$\rightarrow Var(\varepsilon_t) = \sigma_\varepsilon^2$ (constante $\forall t$)

$\rightarrow cov(\varepsilon_t, \varepsilon_s) = 0$ $(\forall t \neq s)$

Hypothèse d'absence d'autocorrelation des erreurs.

Hypothèse d'homoscédasticité

* $H_5 : \varepsilon \rightsquigarrow \mathcal{N}(0, \sigma_\varepsilon^2 . I)$

⇒ Hypothèse de normalité des erreurs.

CH1 : Problèmes liées aux variables explicatives

I La multicolinéarité :

1) Problèmes

Il y a multicolinéarité lorsque l'hypothèse de l'orthogonalité des var explicatives ou encore de leur indépendance linéaire est relachée (c-à-d lorsque la matrice X n'est plus de plein rang. Ceci implique que la matrice X'X devient singulière et par conséquent son inverse $(X'X)^{-1}$ n'existe pas ce qui rend la méthode MCO complètement défaillante, il n'est donc pas possible devant une telle situation d'estimer les paramètres du modèle.

On distingue 2 types de multicolinéarité la multicolinéarité exacte parfaite et la multicolinéarité imparfaite ou quasi-multicolinéarité.

Dans la pratique c'est plutôt le cas de quasi-multicolinéarité est fréquent En effet, la multicolinéarité imparfaite correspond au cas où la matrice X'X est non singulière mais son det est proche de 0. La conséquence directe est qu'on aura des valeurs très grandes dans la matrice inverse $(X'X)^{-1}$ et ceci implique qu'on aura aussi des valeurs assez grande pour la matrice de variance covariance $\hat{\Sigma}_{\hat{\beta}}$.

estimée de l'estimateur

⇒ les estimateurs deviennent moins précis

⇒ les t de Student $\left(t_c = \left|\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}\right|\right)$ deviennent de plus en plus faibles donc les paramètres deviennent statistiquement non significatifs.

⇒ F (Fisher) très élevé

⇒ R^2 élevé

Rayen Bchya

2) Tests de détection

Les tests de détection de la multicolinéarité

Les plus fréquemment utilisés sont:
- le test de Klein
- le test de Farrar et Glauber

a) Test de Klein:

Le test de Klein n'est pas un test-stat au sens test d'hypothèse mais il s'agit simplement d'un critère de présomption de multicolinéarité

Soit le 1er modèle suivant:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_k x_{kt} + \varepsilon_t$$
$$\forall t = 1, \dots, T$$

Le test de Klein se fait en 3 étapes qui se présentent comme suit:

1- Estimer le modèle de régression et calculer son coeff de détermination R^2

2- calculer la matrice des coefficients de corrélation linéaire prises 2 à 2 entre les variables explicatives telle que:

$$\begin{pmatrix} 1 & \rho_{x_1x_2} & \cdots & \rho_{x_1x_k} \\ \rho_{x_2x_1} & 1 & & \\ & & 1 & \\ \rho_{x_kx_1} & & & 1 \end{pmatrix}$$

3- Comparer le R^2 aux différents coefficients de corrélation.

* Règle de décision

si au moins 1 des coeff de corrélation au carré

$$\rho^2_{x_i x_j} > R^2 \quad i \neq j$$

alors il y a présomption de colinéarité.

b) Test de Farrar et Glauber.

Ce test repose sur les hypothèses suivantes

H_0: les variables explicatives sont orthogonales

contre

H_1: les variables explicatives sont dépendantes

La première étape consiste à calculer le det de la matrice des coeff de corrélation entre les variables explicatives lorsque la valeur du det $D \to 0$ le risque de multicolinéarité est très grand par exemple pour un modèle à 2 variables explicatives si les deux séries sont parfaitement corrélés

$$D = \begin{vmatrix} 1 & \rho_{x_1x_2} \\ \rho_{x_2x_1} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 0$$

$\Rightarrow$ Variables explicatives parfaitement corrélées

Dans le cas opposé si les variables explicatives sont orthogonales

$$D = \begin{vmatrix} 1 & \rho_{x_1x_2} \\ \rho_{x_2x_1} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1$$

2ème étape:

La 2ème étape consiste à effectuer un test de Khi 2 (χ^2) en posant les hypothèses suivantes

H_0: les variables sont orthogonales

$\Rightarrow D = 1$

contre

H_1: les séries sont dépendantes

la statistique utilisée et sa loi de probabilité.

$$\chi^2_c = -\left[T - 1 - \frac{1}{6}(2K+5)\right] \times D \rightarrow \chi^2_{\alpha}\left(\frac{1}{2}K(K-1)\right)$$

avec :

T : taille de l'échantillon

$K = k+1$ le nombre de paramètres à estimer du modèle.

* Règle de décision :

- Si $\chi^2_c \leq \chi^2_\alpha$ alors H_0 est vraie
  => les séries sont orthogonales (indep)
- si $\chi^2_c > \chi^2_\alpha$ alors H_1 est vraie
  => présence de multicolinéarité

3) Les solutions au problème de multicolinéarité :

La solution la plus efficace reste, lors de la spécification du modèle, l'élimination des variables explicatives susceptibles de représenter les mêmes phénomènes et donc d'être corrélées avec elles. Sinon, il existe d'autres techniques permettant d'apporter des remèdes à la multicolinéarité parmi ces techniques on peut citer l'augmentation de la taille de l'échantillon, cette technique n'est efficace que si l'ajout d'observations diffère significativement de celle figurant déjà dans le modèle sinon il y aura reconduction de la multicolinéarité

* 3ème Solution : la "Ridge Regression" : c'est une technique proposée par Hoerl et Kennard (1970) qui consiste à traiter la multicolinéarité mécaniquement d'une façon numérique. Il s'agit de transformer la matrice $(X'X)$ en une matrice $(X'X + cI)$ où c est une constante choisie arbitrairement. Le problème vient du fait que la multicolinéarité se traduit par l'existence dans la matrice X d'une colonne représentant une combinaison linéaire d'autres colonnes ce qui affecte aussi la matrice $(X'X)$, on propose alors de détruire cette comb. linéaire en ajoutant une constante aux éléments de la diagonale de la matrice $(X'X)$. Le principe de la "Ridge Regression" consiste ainsi à définir l'estimateur Ridge suivant :

$$\hat{\beta}_{MCO} = (X'X)^{-1} \cdot X'Y$$

Si multicolinéarité

Estimateur "Ridge"

$$\Rightarrow \hat{\beta}_{Ridge} = (X'X + cI)^{-1} X'Y$$

Il est à noter que l'estimateur "Ridge" peut être estimé en fct de l'estimateur $\hat{\beta}_{MCO}$ telle que :

$$\hat{\beta}_{Ridge} = \left[I + c(X'X)^{-1}\right]^{-1} \cdot \hat{\beta}_{MCO}$$

Par ailleurs, on peut montrer que $\hat{\beta}_{Ridge}$

d'un estimateur biaisé mais convergent (de variance minimale par rapport à l'estimateur des MCO)

Dem Greene (doc)

II - Variables aléatoires :

explicatives : problème d'endogénéité :

Lorsque l'hypothèse H_1 des MCO est relachée, cela implique que les variables explicatives sont dépendantes du terme d'erreur dans ces conditions les estimateurs des MCO ne sont plus convergents et il est nécessaire de recourir à un autre estimateur appelé estimateur des variables instrumentales.

1) Les causes de l'endogénéité :

Le problème d'endogénéité peut être dû à plusieurs causes parmi lesquelles on peut citer :

- existence d'un double sens de causalité entre la variable à expliquer (y) et la/ou les variable(s) explicatives)

Exemple : croissance économique et développement financier

- Omission de variables explicatives dans le modèle

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$$

$$\Rightarrow y_t = \alpha + \beta_1 x_{1t} + u_t$$

- Erreurs de mesure, évoquée dans le cas où on utilise des données d'enquête

2) La méthode des variables instrumentales (VI) :

Soit X la matrice des variables explicatives et Z la matrice des variables instrumentales ou instruments.

Un instrument Z d'une variable X et une variable qui permet de nettoyer X. C'est telle que la covariance entre Z et le terme d'erreur est égale à 0 et Z doit être corrélée avec la variable X

$$cov(X, Z) \neq 0$$

(- Le seul lien) Par ailleurs il faut que le seul lien qui peut exister entre Z et y et la variable X.

Autrement dit l'instrument z ne peut pas être une variable explicative en elle même.

Remarque :

On peut parfois utiliser plusieurs instruments pour une même variable endogène ceci permet d'avoir l'estimation de x_i plus précise (explicative endo)

Dans tous les cas, il faut que le nombre de variables instrumentales soit supérieur ou égal au nombre de variables explicatives

~~L'estimateur des~~

(9) Solution :

. estimateur des variables instrumentales VI donné par la formule suivante :

$$\hat{\beta}_{VI} = (Z'X)^{-1} Z'Y$$

avec :

X = matrice initiale des variables explicatives

Z = la matrice des variables instrumentales.

2) L'estimateur de la matrice de variances-covariances des paramètres estimés et défini tel que :

$$\hat{\Omega}_{\hat{\beta}_{VI}} = \hat{\sigma}_{\varepsilon}^2 (Z'X)^{-1} Z'Z (X'Z)^{-1}$$

Une méthode particulière des variables instrumentales est la doubles moindres carrés Ordinaires (DMCO) (2SLS)

qui revient à appliquer la MCO en deux étapes

Soit le modèle suivant

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 z_{3t} + \varepsilon_t$$

et soit (z_{1t}) un instrument de la variable (x_{3t}). $\Rightarrow$ On suppose que (x_{3t}) est endo

Comme son nom l'indique la méthode des doubles moindres carrés ordinaires se fait en deux étapes :

La 1ère étape consiste à regresser (x_{3t}) sur z_t et les autres variables explicatives du modèles :

$$x_{3t} = a_0 + a_1 z_t + a_2 x_{1t} + a_3 x_{2t-1} + u_t$$

$$\Rightarrow \hat{x}_{3t} = \hat{a}_0 + \hat{a}_1 z_t + \hat{a}_2 x_{1t-1} + \hat{a}_3 x_{2t}$$

$\Rightarrow \hat{x}_{3t}$ valeurs ajustées

2ème étape :

Régression de (y_t) sur $(\hat{x}_{3t})$ et le reste des variables explicatives

$$y_t = \alpha_0 + \alpha_1 x_{1t-1} + \alpha_2 x_{2t-1} + \alpha_3 \hat{x}_{3t} + v_t$$

3) Test d'endogénéité d'Hausman

Le test d'endogénéité d'Hausman permet de détecter une éventuelle corrélation entre le l'erme d'erreur ε_t et une ou plusieurs variables explicatives x_{it}. Dans ce cas on ne peut plus utiliser l'estimateur des MCO qui devient non convergent.

Les hypothèses du test sont les suivantes :

$H_0 : cov(x_{it}, \varepsilon_t) = 0 \Rightarrow$ Pas de pb d'endogénéité

contre

$H_1 : cov(x_{it}, \varepsilon_t) \neq 0$

$\Rightarrow x_{it}$ est endogène

On retient les estimateurs de la MCO

On retient les estimateurs VI

* Statistique du test et loi de probabilité :

5/

$$H = (\hat{\beta}_{VI} - \hat{\beta}_{MCO})' \left[\hat{\Omega}_{\hat{\beta}_{VI}} - \hat{\Omega}_{\hat{\beta}_{MCO}} \right]^{-1} (\hat{\beta}_{VI} - \hat{\beta}_{MCO}) \rightsquigarrow \chi^2(K)$$

- K = nombre de variables explicatives + constante (0 + 1)
= rang de la matrice $\left(\hat{\Omega}_{\hat{\beta}_{VI}} - \hat{\Omega}_{\hat{\beta}_{MCO}} \right)$

* Règle de décision :

- Si $H \leq \chi^2_\alpha$ alors on accepte H_0
  => pas de problème d'endogénéité
  => on retient l'estimation par MCO.
- Si $H > \chi^2_\alpha$ alors on accepte H_1
  => il y a un problème d'endogénéité
  => on doit retenir l'estimation par VI

4] Test sur l'instrument :

* Test de Student :

Ce test permet de vérifier que l'instrument est bien corrélé avec la variable instrumentée, il est appliqué au niveau de la première étape de la méthode des doubles moindres carrés ordinaires

$$x_{3t} = a_0 + a_1 z_t + a_2 x_{1t} + a_3 x_{2t} + u_t$$

* On estime les paramètres
* On effectue le test significativité individuelle pour le paramètre associé à la VI (a_1 ou a_3)
=> Test de Student

* Hypothèses :

$H_0 : a_1 = 0$ contre $H_1 : a_1 \neq 0$

Statistique du test et loi :

Sous H_0 vraie : $\frac{a_1}{\hat{\sigma}_{a_1}} \rightsquigarrow St(T-K)$

* Règle de décision :

- Si $t_c = \left| \frac{a_1}{\hat{\sigma}_{a_1}} \right| \leq t_{\frac{\alpha}{2}}^{T-K}$ alors H_0 est vraie => a_1 est statistiquement non significatif => z_t n'est corrélé avec la variable instrumentée
- Si $t_c > t_{\frac{\alpha}{2}}^{T-K}$ donc H_1 est vraie
  => a_1 est statistiquement significatif
  => l'instrument z_t est corrélé avec x_{3t}

* Test de Sargan :

Ce test permet de vérifier que les instruments utilisés sont exogènes et donc valide. Ce test n'est valable que lorsque l'on utilise au moins deux instruments ou une variable endogène.

* Hypothèses :

$H_0 : cov(Z, \varepsilon) = 0 \Rightarrow Z$ est exogène

Contre

$H_1 : cov(Z, \varepsilon) \neq 0 \Rightarrow Z$ n'est pas

- Statistique du test est lor :

$S = T \cdot R^2$ → $\chi^2(r)$

nb observations totales α

r = degré de liberté

= nombre d'instruments - nombre de variables explicatives endogènes.

R^2 = coefficient de détermination relatif à la régression suivante (modèle)

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t$$

1ère étape :

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t$$

$\Rightarrow \hat{y}_t \Rightarrow \hat{u}_t$

2ème étape :

$$\hat{u}_t = \alpha_0 + \alpha_1 z_{1t} + \alpha_2 z_{2t} + v_t$$

$\Rightarrow R^2$

* Règle de décision :

- si $S < \chi^2_\alpha$ alors H_0 est vraie

$\Rightarrow$ Z est endogène.

- si $S > \chi^2_\alpha$ alors H_1 est vraie

$\Rightarrow$ Z n'est pas exogène