
Principales entorses aux MCO : l'endogénéité

Avril 2020

1 Exercice théorique

Considérons le modèle multiple homoscédastique

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

1. Expliquer comment une forte colinéarité des variables x_{1t} et x_{2t} diminue la précision avec laquelle les coefficients de régression β_1 et β_2 sont estimés.
2. On ajuste ce modèle à un échantillon de 100 individus, pour lesquels on trouve une corrélation empirique entre x_{1t} et x_{2t} de $\hat{\rho}_{23} = 0,9$.
Testez la multicolinéarité de ces variables au niveau 5% à l'aide du test de Farrar-Glauber.
Quelle est votre conclusion ?
3. Expliquer pourquoi des coefficients β_1 et β_2 non significatifs (sur base des tests t) ne conduisent pas nécessairement à conclure que, pris ensemble, x_{1t} et x_{2t} n'influencent pas y_t (sur base du test F).

2 Applications

2.1 Application 1 : non-normalité et points aberrants

Nous cherchons à expliquer la nocivité des cigarettes à partir de leur teneur en divers composants. À travers cet exercice, nous allons passer en revue différents tests sur les résidus, notamment dans le but de détecter d'éventuels problèmes de non-normalité et de points aberrants.

*Q1 : Importez le fichier **cigarettes.txt**. Réalisez les nuages de points en croisant deux à deux les variables.*

*Q2 : Régressez **CD** sur les autres variables de la base de données.*

- Récupérez les résidus de la régression. Calculez sa moyenne. Que constatez-vous ?
- Construisez le nuage de points avec en abscisse **CD** et en ordonnée le résidu. Y a-t-il des éléments saillants dans le graphique ?
- Réalisez la droite de Henry pour vérifier la compatibilité des résidus avec l'hypothèse de normalité (commande **qqnorm**). Que constatez-vous ?

- Calculez alors la statistique de Jarque-Bera. Est-ce que les résidus sont compatibles avec l'hypothèse de normalité ?

Dans la suite, nous nous concentrons sur la détection des points aberrants

Q3 : Calculez le résidu “studentisé” de la régression, ainsi que son seuil critique pour un risque de 10%. Quelles sont les marques de cigarette atypiques au sens de ce seuil ?

Q4 : Construisez le nuage de points avec en abscisse la variable CO et en ordonnée le résidu “studentisé”. Insérez dans le graphique les lignes matérialisant les seuils critiques. Faites apparaître nommément les cigarettes atypiques. Calculez le levier de chaque observation. Quels sont les points atypiques au sens du levier ?

Q5 : Créez un nouveau data frame excluant les observations atypiques au sens du résidu “studentisé” OU du levier. De combien d’observations dispose-t-on maintenant? Réalisez de nouveau la régression CO vs les autres variables à partir de ce nouvel ensemble de données. Quelle est la valeur du R^2 maintenant ?

2.2 Application 2 : modèles à équations simultanées et endogénéité

On s’intéresse à la relation entre inflation et rendements des marchés des actions. Selon la théorie économique, le taux d’inflation intervient dans la valorisation des actions à travers le facteur d’escompte, mais dans le même temps, l’évolution des marchés financiers peut stimuler la demande et donc l’inflation. Il est clair qu’il existe une relation simultanée entre ces 2 grandeurs.

Nous voulons estimer le système d’équations simultanées suivant :

$$\begin{cases} \text{inflation}_t &= \alpha_0 + \alpha_1 \text{returns}_t + \alpha_2 \text{dcredit}_t + \alpha_3 \text{dprod}_t + \alpha_4 \text{dmoney}_t + \varepsilon_{1t} \\ \text{returns}_t &= \beta_0 + \beta_1 \text{dprod}_t + \beta_2 \text{dspread}_t + \beta_3 \text{inflation}_t + \beta_4 \text{rterm}_t + \varepsilon_{2t} \end{cases}$$

où *inflation* est le taux d’inflation américain, *returns* est le rendement logarithmique de l’indice S&P500, *dcredit* est la variation mensuelle des crédits à la consommation, *dprod* est la variation mensuelle de l’indice de la production industrielle, *dmoney* la variation mensuelle de la masse monétaire, *dspread* la variation mensuelle de l’écart de crédit (entre les obligations BAA et AAA) et *rterm* la variation mensuelle entre les bons à 10 ans et 3 mois du Trésor américain.

Etant donné la présence de *returns* dans l’équation de *inflation* et vice-versa, la méthodologie des **2SLS** (Two-stage least squares) est appropriée pour répondre au problème d’endogénéité. Cette méthode nécessite l’utilisation d’**instruments** qui sont représentés ici par la liste des autres variables.

*Q1 : Importez le fichier **macro.xls**. Chargez le package **AER**. Effectuez les 2 régressions correspondant aux 2 équations à l’aide de la commande **ivreg**. Pour spécifier les instruments dans la même régression, utiliser le séparateur | suivi de la liste des instruments. Comme avec **lm**, les variables sont séparées par +.*

Q2 : Effectuez un test de Hausman pour vérifier l’endogénéité des 2 variables.