Module 4 : vérification des hypothèses d'application de la régression et robustesse du modèle

Dans les paragraphes précédents on a supposé que les hypothèses d'application de la régression étaient vérifiées ce qui permet de montrer les propriétés remarquables (BLUE) des estimateurs, de construire des tests des paramètres et du coefficient de détermination et enfin d'élaborer des intervalles de confiance prévisionnels. L'importance de ces hypothèses étant manifeste, il est indispensable de les vérifier pour contrôler la qualité statistique et donc opérationnelle du modèle de régression.

L'hypothèse d'indépendance de la variable explicative est une hypothèse ad hoc. Il en est de même dans ce cours de celle concernant le sens de causalité entre deux variables ainsi que l'absence de tendances communes pouvant conduire à une « spurious régression » (régression factice, c'est-à-dire une régression qui semble de bonne qualité à cause d'une tendance semblable entre les deux variables (r^2 élevé) mais qui dans la réalité n'est qu'une covariation).

En définitive, ce sont les hypothèses sur l'aléa qui font l'objet de ce paragraphe. Rappelons que l'aléa est une succession temporelle (pour le modèle choisi ici) de variables aléatoires centrées, homoscédastiques, non autocorrélées et obéissant à une loi normale. Cet aléa est inconnu. L'hypothèse fondamentale sur laquelle repose le modèle de régression c'est que le résidu du modèle (connu) $e_t = Y_t - \hat{Y}_t$ est un échantillon de cette famille de variables aléatoires. De ce fait, si le résidu vérifie, à partir de ses caractéristiques, les propriétés de l'aléa, on dira qu'il est issu de la famille des variables aléatoires. On utilise ainsi la moyenne, la variance, l'autocorrélation, et l'histogramme des résidus pour vérifier les hypothèses d'application du modèle de régression (unités 1, 2, 3 et 4).

Il est enfin possible de vérifier si le modèle estimé est valide dans diverses circonstances : c'est la robustesse (unité 5)

1 L'hypothèse de nullité de l'espérance mathématique de l'erreur $E[\varepsilon_t] = 0$

On veut tester $E[\varepsilon_t] = 0$ On utilise la moyenne des résidus $\bar{e} = \frac{1}{n}\sum_t e_t$ pour vérifier cette hypothèse.

On sait que : $\bar{e} \equiv N\left(m, \sigma_e / \sqrt{n}\right)$ soit $\frac{\bar{e} - m}{\sigma_e}\sqrt{n} \equiv N(0,1)$

On construit alors le test de signification : $H_0 : m = 0$ contre $H_1 : m \neq 0$

Si $\frac{|\bar{e} - 0|}{\sigma_e}\sqrt{n} < 1{,}96$ (le quantile à 95% de la loi normale centrée réduite) alors l'hypothèse H_0 est vérifiée.

Cette hypothèse ne joue pas un rôle important dans la régression puisqu'on sait que $e_t = y_t - \hat{y}_t$ et donc par construction $\bar{e} = 0$. Il s'agit donc d'une hypothèse ad hoc et l'utilité de ce test ne se justifie que dans d'autres applications (séries temporelles par exemple)

2 L'hypothèse de non autocorrélation des erreurs $E[\varepsilon_t \varepsilon_{t'}] = 0$

On va tester $E[\varepsilon_t \varepsilon_{t'}] = 0 \qquad \forall t, \forall t', t \neq t'$

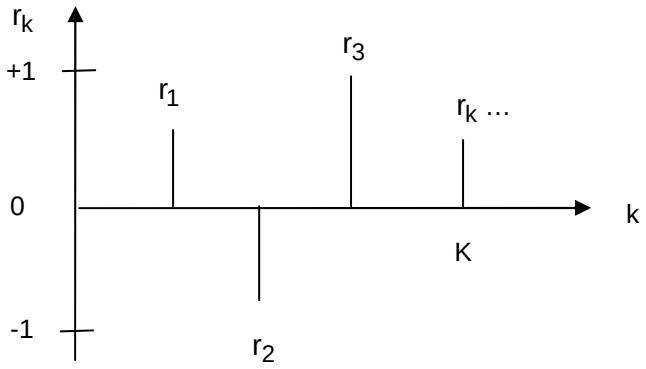
2.1 Détection de l'autocorrélation

L'autocorrélation concerne les résidus : $e_t = Y_t - \hat{Y}_t$. Il y a autocorrélation toutes les fois où on peut trouver un coefficient de corrélation linéaire significativement différent de 0 entre la chronique des résidus et cette même chronique décalée d'un ou de plusieurs pas de temps.

Si on note k le pas de temps du décalage temporel $\left(k \in N^*\right)$ et r_k le coefficient de corrélation linéaire simple correspondant, on peut construire la Fonction d'AutoCorrélation des résidus :

$$k \in N^* \xrightarrow{FAC} r_k \in [-1,+1]$$

dont la représentation graphique est le corrélogramme :



K est le décalage maximal pour lequel r_k a un sens statistique (le nombre de points permettant le calcul de r_k). En général $\frac{n}{6} \le K \le \frac{n}{3}$

Si les résidus sont une bonne représentation de l'aléa, ils doivent vérifier l'hypothèse de non autocorrélation ; cela signifie que toutes les autocorrélations successives doivent être non significativement différentes de 0.

2.2 Principales causes de l'autocorrélation

Plusieurs raisons peuvent être la cause d'une autocorrélation. On peut citer :

- Les variables de départ ne vérifient pas l'hypothèse de stationnarité, c'est-à-dire qu'elles peuvent contenir des tendances déterministes (trend linéaire) ou stochastiques (promenade aléatoire) communes, ce qui est générateur d'une régression factice pour laquelle le r^2 est proche de 1 avec une autocorrélation importante du résidu.

- Les variables de départ étaient saisonnières et elles ont été mal désaisonnalisées.

- Les variables contiennent des phénomènes exceptionnels (grèves,…) qui sont mal expliqués par le modèle.

- Les variables de départ possédaient des « non informations » qui ont été corrigées par extrapolation linéaire…

2.3 Les effets de l'autocorrélation des erreurs

Considérons le modèle sous la forme :

$$y_t = \beta x_t + \varepsilon_t$$

Et supposons que :

ε_t obéisse à un processus autorégressif d'ordre 1, c'est-à-dire qu'il existe entre ε_t et ε_{t-1} un modèle de régression linéaire :

$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$ avec

$|\rho| < 1$ (qui assure la stabilité du modèle) et $\begin{cases} E[\eta_t] = 0 \\ V[\eta_t] = \sigma_\eta^2 \\ Cov[\eta_t \eta_{t'}] = 0 \end{cases}$

On sait que :

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \sum w_t y_t \quad \text{avec } w_t = \frac{x_t}{\sum x_t^2}$$

$$\hat{\beta} = \sum w_t (Y_t - \overline{Y}) = \sum w_t Y_t - \overline{Y} \underbrace{\sum w_t}_{=0}$$

D'où

$$\begin{aligned} \hat{\beta} &= \sum w_t (\alpha + \beta X_t + \varepsilon_t) \\ &= \alpha \underbrace{\sum w_t}_{0} + \beta \underbrace{\sum w_t X_t}_{=1} + \sum w_t \varepsilon_t \\ &= \beta + \sum w_t \varepsilon_t \end{aligned}$$

D'où

$$E[\hat{\beta}] = \beta + \underbrace{\sum w_t}_{=0} E[\varepsilon_t]$$

- L'estimateur reste sans biais quelque soit $E[\varepsilon_t]$.

On peut vérifier cependant que :

$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$ s'écrit :

$$\begin{aligned} \varepsilon_t &= \rho(\rho\varepsilon_{t-2} + \eta_{t-1}) + \eta_t \\ &= \rho^2 \varepsilon_{t-2} + \rho\eta_{t-1} + \eta_t \\ &\cdots \text{etc} \cdots \end{aligned}$$

$$\begin{aligned} \varepsilon_t &= \eta_t + \rho\eta_{t-1} + \rho^2 \eta_{t-2} + \cdots \\ &= \sum_{\theta=0}^{+\infty} \rho^\theta \eta_{t-\theta} \end{aligned}$$

$$E[\varepsilon_t] = \sum_{\theta=0}^{+\infty} \rho^\theta \underbrace{E[\eta_{t-\theta}]}_{=0}$$

- La variance de $\hat{\beta}$ s'écrit :

$$V[\hat{\beta}] = \sigma_\varepsilon^2 \,/\, \sum x_t^2 = E[\hat{\beta} - E[\hat{\beta}]]^2$$

Comme $V[\varepsilon_t] = E\left[\varepsilon_t - \underbrace{E[\varepsilon_t]}_{=0}\right]^2 = E[\varepsilon_t^2]$

On a :

$$\varepsilon_t = \eta_t + \rho\eta_{t-1} + \rho^2\eta_{t-2} + \cdots$$

$$\varepsilon_t^2 = \eta_t^2 + \rho^2\eta_{t-1}^2 + \cdots + 2\rho\eta_t\eta_{t-1} + \cdots$$

D'où

$$E\left[\varepsilon_t^2\right] = E\left[\eta_t^2\right] + \rho^2 E\left[\eta_{t-1}^2\right] + \cdots + 2\rho\underbrace{E\left[\eta_t\eta_{t-1}\right]}_{=0} + \cdots$$

$$= \sigma_\eta^2 + \rho^2\sigma_\eta^2 + \ldots + 0$$

$$= \sigma_\eta^2\left(1 + \rho^2 + \cdots\right)$$

Or : $|\rho| < 1$ d'où :

$$E\left[\varepsilon_t^2\right] = \frac{1}{1-\rho^2}\sigma_\eta^2 = \sigma_\varepsilon^2$$

De ce fait :

$$V\left[\hat{\beta}\right] = \frac{\sigma_\eta^2}{1-\rho^2}\frac{1}{\sum_t x_t^2}$$

En conclusion, lorsqu'il y a autocorrélation des erreurs (ici d'ordre 1)

- les estimateurs restent sans biais.

- les variances d'échantillon des coefficients de régression ne sont plus minimales : plus l'autocorrélation est forte $(|\rho| \to 1)$ plus la quantité $\frac{1}{1-\rho^2}$ est grande et donc plus forte est la sous-estimation de la variance de $\hat{\beta}$.

La méthode des MCO n'est donc pas, dans ce cas, la meilleure des méthodes pour estimer le modèle. Elle sous-estime les variances vraies dans le cas d'une autocorrélation positive par exemple, ce qui a pour conséquence une surestimation de la prévision de l'estimation. Dans le cas d'une prévision, on n'aura plus des valeurs de la variable endogène les meilleures possibles.

L'autocorrélation remet en cause l'estimation du MLGS par les MCO ; on doit disposer de tests permettant de la détecter.

2.4 Tests d'autocorrélation des résidus

Le plus utilisé est le test de Durbin-Watson. Ces auteurs proposent la statistique suivante :

$$DW = \frac{\sum_{t=1}^{n-1}(e_{t+1} - e_t)^2}{\sum_{t=1}^{n} e_t^2}$$

Pour n grand :

$\sum_{t=1}^{n-1} e_{t+1}^2 \approx \sum_{t=1}^{n-1} e_t^2 \approx \sum_{t=1}^{n} e_t^2$; on peut alors approximer DW par la quantité : $DW = 2(1-\hat{\rho})$

avec $\hat{\rho} = \dfrac{\sum_{t} e_{t+1} e_t}{\sum_{t=1}^{n} e_t^2}$

$\hat{\rho}$ est l'estimation, par les MCO, du modèle $e_{t+1} = \rho e_t + \varepsilon_t$ avec $|\rho| < 1$

- si $\hat{\rho} \to 0$, absence de corrélation dans les résidus $\Rightarrow DW \to 2$;

- si $\hat{\rho} \to 1$, forte autocorrélation positive dans les résidus $\Rightarrow DW \to 0$

- si $\hat{\rho} \to -1$, forte autocorrélation négative dans les résidus $\Rightarrow DW \to 4$

Durbin-Watson ont montré que la statistique DW dépendait de deux valeurs d_1 et d_2, indépendantes de X_t ; ce sont des variables aléatoires fonction de ε_t. Ils en ont tabulé les valeurs pour n (nombre d'observations), K (nombre de variables exogènes) données et deux seuils 5% et 10%. Le test se déroule de la façon suivante :

- On calcule DW_c (avec e_t et la formule non simplifiée) ;

- On place le résultat trouvé dans le tableau suivant :

0	d_1	d_2	2	$4-d_2$	$4-d_1$	4
Autocorrélation >0	Doute	indépendance		Doute	Autocorrélation <0	

Ce test présente l'inconvénient de ne pouvoir déceler que les autocorrélations d'ordre 1. On peut remédier à ce problème en utilisant les résultats de la FAC (Fonction d'autocorrélation). Chaque autocorrélation peut être testée par un test classique de signification de Student :

$H_0 : \rho_k = 0 \quad H_1 : \rho_k \neq 0$

$t_c = \dfrac{|r_k|}{\sqrt{1-r_k^2}}\sqrt{n-2} < T_{lu}(n-2)$ on est sous l'hypothèse H_0.

3 L'hypothèse d'homoscédasticité des erreurs $E\left[\varepsilon_t^2\right] = \sigma_\varepsilon^2$

On va tester $E\left[\varepsilon_t^2\right] = \sigma_\varepsilon^2 \qquad \forall t$

3.1 Définition

L'homoscédasticité peut être considérée comme un cas particulier de la non autocorrélation $E[\varepsilon_t \varepsilon_{t'}] = 0$; lorsque t = t' alors :

$$Cov[\varepsilon_t, \varepsilon_{t'}] = Cov[\varepsilon_t, \varepsilon_t] = E[[\varepsilon_t - E[\varepsilon_t]][\varepsilon_t - E[\varepsilon_t]]] = E\left[\varepsilon_t^2\right] = \sigma_\varepsilon^2$$

Il y a hétéroscédasticité lorsque la variance des variables aléatoires qui composent ε_t sont différentes au cours du temps. Les conséquences de l'hétéroscédasticité sont, par construction, identiques à celles de l'autocorrélation :

- les estimateurs des paramètres restent sans biais.

- les estimateurs des paramètres ne sont plus de variance minimale.

Il faut donc, comme pour l'autocorrélation, détecter une présence possible d'hétéroscédasticité en utilisant le résidu e_t, seule information disponible concernant ε_t.

3.2 Tests d'hétéroscédasticité

➔ *Test paramétrique de Goldfeld-Quandt*

Il s'applique toutes les fois où l'écart type de l'erreur du modèle s'accroît proportionnellement avec la variable explicative du modèle.

Ecriture de cette hypothèse :

$$\sqrt{E\left[\varepsilon_t^2\right]} = aX_t \Leftrightarrow E\left[\varepsilon_t^2\right] = a^2X_t^2$$

Principe du test :

On ordonne les observations des variables Y_t et X_t en fonction des valeurs croissantes de X_t.

On néglige les observations centrales de l'échantillon. On appelle m le nombre de ces observations négligées.

Comme m dépend de n, on prend pour n = 30, m = 8 et pour n = 60, m = 16, etc.

On obtient deux sous échantillons, l'un correspond aux faibles valeurs de X_t (premier échantillon), l'autre aux fortes valeurs (deuxième échantillon). On applique les MCO sur les $\frac{n-m}{2}$ observations faibles et sur les $\frac{n-m}{2}$ observations fortes. (Il faut que les deux échantillons soient suffisamment importants).

On appelle SCR_1 la somme des carrés des résidus du premier échantillon, SCR_2 la somme des carrés des résidus du second échantillon. On démontre alors que :

$$\frac{SCR_2}{SCR_1} \equiv F_{1-p}\left(\frac{n-m-4}{2};\frac{n-m-4}{2}\right)$$

Les hypothèses du test sont :

H_0 : homoscédasticité H_1 : hétéroscédasticité ($SCR_2 > SCR_1$)

Règle de décision :

$$\begin{cases} \text{si } \frac{SCR_2}{SCR_1} < F_{1-\rho} \Rightarrow H_0 \text{ acceptée au rique de } 1^{\text{ère}} \text{ espèce p} \rightarrow \text{homoscédasticité} \\ \text{si } \frac{SCR_2}{SCR_1} \geq F_{1-\rho} \Rightarrow H_0 \text{ rejetée au rique de } 1^{\text{ère}} \text{ espèce p} \rightarrow \text{hétéroscédasticité} \end{cases}$$

➔ *Test de Glejser*

Ce test propose de régresser la valeur absolue des résidus de la régression avec la variable explicative X_t. On considère des fonctions simples du type, (selon l'hypothèse précédente) :

$|e_t| = a_0 + a_1X_t + \eta_t$ avec η_t =aléa vérifiant les hypothèses de base

$$|e_t| = a_0 + \frac{a_1}{X_t} + \eta_t$$

$$|e_t| = a_0 + a_1\sqrt{X_t} + \eta_t$$

$$|e_t| = a_0 + \frac{a_1}{\sqrt{X_t}} + \eta_t$$

L'hypothèse d'homoscédasticité est vérifiée si le paramètre a_1 n'est pas significativement différent de zéro.

D'où le test :

$H_0 : a_1 = 0$ (homoscédasticité) $H_1 : a_1 \neq 0$ (hétéroscédasticité)

On applique alors la méthode des MCO aux différents modèles proposés par Glejser :

$$t_c = \frac{\hat{a}_1}{\sigma_{\hat{a}_1}} \equiv T(n-2)$$

Si $t_c < T_{lu}(n-2) \Rightarrow H_0$ acceptée au risque de 1ère espèce $p \rightarrow$ homoscédasticité

➔ *Test Arch - LM*

Il s'agit d'un test de conception différente utilisé principalement pour les séries temporelles. Les modèles AutoRégressifs Conditionnellement Hétéroscédastique (ARCH) ont été introduits par Engle en 1982 pour modéliser la volatilité des cours boursiers. Un représentant de ce modèle est associé au test du Multiplicateur de Lagrange (test du χ^2) pour vérifier l'hypothèse d'homoscédasticité du résidu e_t (qui est une série chronologique).

Déroulement du test :

- On considère le modèle suivant, appelé modèle autorégressif de retard p, sur le carré des résidus :

$$e_t^2 = \phi_0 + \phi_1 e_{t-1}^2 + \cdots + \phi_p e_{t-p}^2 + \eta_t$$

- On estime le modèle par la méthode des MCO (il s'agit d'un modèle à plusieurs variables qui sera étudié ultérieurement).

$$\hat{e}_t^2 = \hat{\phi}_0 + \hat{\phi}_1 e_{t-1}^2 + \cdots + \hat{\phi}_p e_{t-p}^2$$

- On calcule la statistique :

nR^2 avec R^2 le coefficient de détermination du modèle

n le nombre d'observations

- On démontre que : $nR^2 \equiv \chi^2(p)$

Sous l'hypothèse H_0 du test, les coefficients du modèle ne sont pas significativement différents de zéro (ils sont donc significativement égal à zéro) : $\phi_1 = \cdots \phi_p = 0$

De ce fait : $e_t^2 = \phi_0 + \varepsilon_t$ et $V[e_t] = \frac{1}{n}\sum e_t^2 = \frac{1}{n}\sum(\phi_0 + \varepsilon_t) = \bar{\varepsilon} + \phi_0 = \phi_0$, il y a homoscédasticité

D'où le test :

$H_0 : \phi_1 = \cdots \phi_p = 0$ (homoscédasticité) H_1 : un au moins des coefficients ≠ de 0 (hétéroscédasticité)

Si $nR^2 < \chi^2(p)$ on est sous l'hypothèse H_0 donc homoscédasticité.

p le nombre de retard est choisi successivement dans N*.

➔ *Test de White*

On effectue une régression entre le carré du résidu et une ou plusieurs variables explicatives en niveau et au carré (ici, on considère une seule variable explicative puisque l'on se place dans le cas du modèle linéaire général simple à 2 variables), c'est-à-dire :

$$e_t^2 = a_0 + a_1 X_{1t} + b_1 X_{1t}^2 + \eta_t$$

Si l'un de ces coefficients de régression (a_1 ou b_1) est significativement différent de 0, on accepte l'hypothèse d'hétéroscédasticité. Deux manières pour effectuer le test :

1) On effectue un test de Fisher : $H_0 : a_1 = b_1 = a_0 = 0$

On construit le Fisher calculé suivant :

$F_c = \frac{R^2}{1-R^2}\frac{n-k}{k-1}$ où k représente le nombre total de paramètres estimés (ici, k=3)

$F_c \equiv F(k-1, n-k)$

Règle de décision :

Si $F_c < F_{1-p}(k-1, n-k)$ alors H_o acceptée au risque de 1ère espèce p $\Rightarrow$ homoscédasticité

Si $F_c \geq F_{1-p}(k-1, n-k)$ alors H_o rejetée au risque de 1ère espèce p $\Rightarrow$ hétéroscédasticité

2) Soit on recourt à la statistique $LM \equiv \chi^2(p = K)$

K étant le nombre de variable explicatives, ici K=2

$LM = nR^2 > \chi^2(p)$ H_0 rejetée au risque de 1ère espèce p $\Rightarrow$ hétéroscédasticité

$LM = nR^2 < \chi^2(p)$ H_0 acceptée au risque de 1ère espèce p $\Rightarrow$ homoscédasticité

4. L'hypothèse de normalité des erreurs : $\varepsilon_t \equiv N(0, \sigma_\varepsilon)$

On veut tester $\varepsilon_t \equiv N(0, \sigma_\varepsilon)$

Cette hypothèse est indispensable pour calculer les estimateurs du maximum de vraisemblance mais aussi et surtout pour réaliser nombre de tests concernant les caractéristiques du modèle de régression (test de Student des paramètres, test de Fisher du r² etc …). On utilise dans la pratique le test de Jarque et Béra fondé sur la notion de skewness (asymétrie) et du Kurtosis (aplatissement). Néanmoins il est toujours possible de recourir aux tests standards d'ajustement d'une loi normale à une distribution empirique (comme le test du χ^2).

Principales étapes du test de Jarque et Béra :

On construit l'histogramme du résidu e_t en découpant l'étendu du résidu en classes d'amplitudes égales. On calcule alors après avoir affecté à chacune des classes le nombre de fois que le résidu se répète :

Le coefficient du skewness : $\beta_1^{1/2} = \frac{\mu_3}{\sigma^3}$ où μ_3 est le moment centré d'ordre 3 de la distribution.

Le coefficient du Kurtosis $\beta_2 = \frac{\mu_4}{\sigma^4}$ où μ_4 est le moment centré d'ordre 4 de la distribution.

On démontre que :

$$\beta_1^{1/2} \equiv N\left(0, \sqrt{\frac{6}{n}}\right) \text{ et } \beta_2 \equiv N\left(3, \sqrt{\frac{24}{n}}\right)$$

Remarque : il est donc possible de réaliser un test de symétrie et d'aplatissement en utilisant les lois normales centrées réduites. :

$$\nu_{1=} \frac{\beta_1^{1/2} - 0}{\sqrt{\frac{6}{n}}} \equiv N(0,1) \text{ et } \nu_{2=} \frac{\beta_2 - 3}{\sqrt{\frac{24}{n}}} \equiv N(0,1)$$

Les tests de symétrie et d'aplatissement normal se font ainsi :

- H_0 : aplatissement normal

si $\left|\frac{\beta_1^{1/2} - 0}{\sqrt{\frac{6}{n}}}\right| < 1{,}96$ (le quantile à 95% de la loi normale centrée réduite) alors H_0 est acceptée au risque de 5% donc aplatissement normal.

si $\left|\frac{\beta_1^{1/2} - 0}{\sqrt{\frac{6}{n}}}\right| \geq 1{,}96$ alors H_0 est rejetée au risque de 5%.

- H_0 : symétrie normale

Si $\left|\frac{\beta_2 - 3}{\sqrt{\frac{24}{n}}}\right| < 1{,}96$ alors H_0 est acceptée au risque de 5% donc symétrie normale.

Si $\left|\frac{\beta_2 - 3}{\sqrt{\frac{24}{n}}}\right| \geq 1{,}96$ alors H_0 est rejetée au risque de 5%.

Pour vérifier l'hypothèse de normalité, il faut à la fois l'aplatissement normal et la symétrie normale.

De ce fait la statistique (due à Jarque Béra) notée JB s'écrit :

$JB = \frac{n}{6}\beta_1 + \frac{n}{24}(\beta_2 - 3)^2$ Elle obéït à un $\chi^2(2)$ (somme de deux lois normales au carré). Le test se déroule de la façon suivante :

- Hypothèse : H_0 : la distribution obéit à une loi normale H_1 : la distribution n'obéit pas à une loi normale
- Calcul de JB
- Si JB < $\chi^2(2)$ (égal à 5,99 au seuil $\alpha = 0{,}05$) on est sous l'hypothèse H_0 de normalité.

5. La robustesse du modèle

Un modèle est dit robuste lorsqu'il est valide dans des circonstances différentes.

Exemple : l'estimation de la fonction de consommation pendant la première moitié du XXè siècle est-elle restée identique à celle de la deuxième moitié ?

La relation prix – récolte de vin est-elle restée identique après l'introduction de la viticulture dans le marché commun en 1970 ?

Dans ces exemples, appelés exemples de robustesses structurelles, l'étude porte sur des époques de temps consécutives, mais elle peut concerner des périodes qui se chevauchent. Cette robustesse peut aussi être liée à des problèmes d'homogénéité spatiale. La robustesse concerne aussi le sens de la causalité de la relation économique.

Dans ce cours, on dira qu'un modèle est robuste, si quels que soient les sous-ensembles constitués à partir d'observations consécutives sur la période [1,n], les estimateurs du même modèle sur chacun de ces sous-ensembles sont :

- valides (test des paramètres, R^2, résidus…)
- stables : les paramètres estimés ne sont pas significativement différents entre eux, et différents de $\hat{\alpha}$ et $\hat{\beta}$.

Cette définition amène sur le plan statistique à comparer les estimations des paramètres entre eux et les qualités de la régression entre elles. Trois tests de stabilité sont présentés

5.1 Tests de comparaison de deux coefficients de corrélation

Ils permettent de s'assurer que les relations sont bien de type linéaire et qu'elles ne sont pas globalement différentes.

Supposons que, sur la période [1,n], on construise 2 sous périodes de cardinal n_1 et n_2. Soient r_1 et r_2, les coefficients de corrélation linéaire des deux sous périodes échantillons.

On montre qu'un coefficient de corrélation ne suit pas une distribution d'expression simple autour de son espérance mathématique : la distribution est fortement asymétrique pour les valeurs éloignées de zéro. Ainsi pour comparer deux coefficients de corrélation on peut utiliser au préalable la transformation non linéaire de FISHER :

$z = \text{Argth}\,\rho = \frac{1}{2}\text{Log}\frac{1+\rho}{1-\rho}$ avec Argth : fonction Argument tangente hyperbolique et Log le logarithme népérien :

$$\begin{cases} E(z) = \text{Argth}\,\rho \quad \text{avec} \quad E[r] = \rho \\ V(z) = s^2(z) = \dfrac{1}{n-3} \end{cases}$$

Si on note alors : $\begin{cases} z_1 = \text{Argth}\,\rho_1 \\ z_2 = \text{Argth}\,\rho_2 \end{cases}$

La différence $d = z_1 - z_2$ a pour caractéristique :

$$E(d) = E(z_1) - E(z_2) = 0$$

$$V(d) = V(z_1) + V(Z_2)$$

$$= \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

La valeur estimée de d est $\hat{d}$. Elle est égale à $\hat{d} = z_1' - z_2'$

Avec $\begin{cases} z_1' = \text{Argth}\, r_1 = \frac{1}{2}\text{Log}\left(\frac{1+r_1}{1-r_1}\right) \\ z_2 = \text{Argth}\, r_2 = \frac{1}{2}\text{Log}\left(\frac{1+r_2}{1-r_2}\right) \end{cases}$

On teste alors l'hypothèse $H_0 : d = 0 \qquad H_1 : d \neq 0$

Sous l'hypothèse H_0 on a : $t_c = \left|\frac{\hat{d}}{s(\hat{d})}\right| \equiv N(0,1)$ avec $s(\hat{d}) = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$

La règle de décision est alors la suivante :

$\begin{cases} \text{si } t_c \geq 1{,}96 \Rightarrow H_0 \text{ rejetée au rique de 5\%} \\ \text{si } t_c < 1{,}96 \Rightarrow H_0 \text{ acceptée au rique de 5\%} \end{cases}$

Si H_0 acceptée, la différence entre les deux coefficients r_1 et r_2 n'est pas significativement différente de 0.

5.2 Tests de comparaison de deux coefficients de régression

Soient $\hat{\beta}^1$ et $\hat{\beta}^2$, les deux coefficients de régression estimés sur les deux sous-ensembles de cardinal n_1 et n_2.Considérons l'hypothèse H_0 : les deux coefficients ne sont pas significativement différents. Si cette hypothèse est vraie alors $\hat{d} = \hat{\beta}^1 - \hat{\beta}^2$ n'est pas significativement différente de zéro. En effet, le caractère non biaisé de $\hat{\beta}^1$ et $\hat{\beta}^2$ permet d'écrire que :

$E[\hat{\beta}^1] = E[\hat{\beta}^2] = \beta$ D'où : $E[\hat{\beta}^1] - E[\hat{\beta}^2] = E[\hat{\beta}^1 - \hat{\beta}^2] = 0$

De plus comme $\hat{\beta}^1$ et $\hat{\beta}^2$ sont deux variables aléatoires indépendantes on a :

$s^2[\hat{d}] = s^2[\hat{\beta}^1] + s^2[\hat{\beta}^2]$ D'où :

$$t_c = \frac{\hat{d}}{s[\hat{d}]} \equiv T_{1-\alpha}(n_1 + n_2 - 4)$$

D'où le test : $H_0 : d = \beta^1 - \beta^2 = 0 \qquad H_1 : d \neq 0$

Et la règle de décision :

Si $t_c = \frac{|\hat{d}|}{s[\hat{d}]} < T_{1-\alpha}(n_1 + n_2 - 4)$ on est sous l'hypothèse H_0 et donc les deux coefficients ne sont pas significativement différents.

5.3 Tests de stabilité du modèle : test de Chow

Ce test est une présentation différente du test de comparaison de deux coefficients de régression.

Soit SCR_0 : la somme des carrés des résidus du modèle sur toute la période et SCR_1 et SCR_2 la somme des carrés des résidus sur chacune des deux sous périodes.

On teste $H_0 : \beta^1 = \beta^2 \qquad H_1 : \beta^1 \neq \beta^2$

Sous H_0 : $F_c = \dfrac{(n-4)SCR_0 - (SCR_1 + SCR_2)}{SCR_1 + SCR_2} \equiv F_{1-p}(2, n-4)$

Règle de décision :

Si $F_c \geq F_{1-p}(2, n-4)$ H_0 rejetée au risque de 1ère espèce p

Si $F_c < F_{1-p}(2, n-4)$ H_0 acceptée au risque de 1ère espèce p donc les deux coefficients sont significativement différents.

5.4 Une étude simple de la robustesse : les régressions roulantes

La régression roulante consiste à régresser le modèle sur un nombre suffisant (n_1) d'observations en début de période puis de réitérer l'estimation en rajoutant une observation à n_1 jusqu'en fin de période. (en accroissant le nombre d'observations sur l'axe du temps). Pour chacune des régressions on mémorise une ou plusieurs caractéristiques (t de Student, r^2, DWc…) que l'on représente graphiquement. La stabilité au cours du temps de ces caractéristiques est une indication de la robustesse du modèle.

Remarque : il est possible d'utiliser les régressions roulantes de l'instant 1 vers l'instant n (régression Forward) ou au contraire de n vers 1 (régression backward)

Bibliographie :

Régis BOURBONNAIS (2009) : Econométrie - Dunod - 7ième édition

J JOHNSTON, J DINARDO (1999) : Méthodes économétriques – Economica - 4ième édition