Université Paris 7 **Économétrie**
Licence 3, 2009-2010 **Pierre Alquier - Joseph Salmon**

TD n° 7 : Hétéroscédasticité

Dans ce TD, on étudie différents cas d'hétéroscédasticité. Le premier exercice est un cas d'école, on suppose que l'on connait la variance des différentes observations et on vérifie que l'estimateur des moindres carrés généralisés est optimal. Dans le second, on suppose que $\mathrm{Var}(\varepsilon_i)$ est une fonction croissante de l'une des variables de la régression, dans ce cas, on peut facilement fabriquer un nouveau test d'hétéroscédasticité, connu dans la littérature sous le nom de **test de Goldfeld-Quandt**.

Exercice 1. On observe deux variables aléatoires y_1 et y_2 telles que

$$y_i = b + \xi_i$$

avec $\xi_1 \sim \mathcal{N}(0, \sigma^2)$, $\xi_2 \sim \mathcal{N}(0, 3\sigma^2)$ et ξ_1 et ξ_2 sont indépendantes. On est donc dans le cadre d'un modèle linéaire extrêmement simple (il n'y a pas de variables explicatives). Le but de cet exercice est de fabriquer un estimateur de b qui soit le plus performant possible.

1) Réécrire le modèle sous forme matricielle.
2) Montrer que l'estimateur des moindres carrés de b s'écrit
$$\hat{b} = \frac{y_1 + y_2}{2}.$$
3) Vérifier que $\hat{b}$ est sans biais et calculer $\mathrm{Var}(\hat{b})$.
4) On se propose maintenant d'étudier la famille d'estimateurs
$$\hat{b}_\lambda = \lambda y_1 + (1-\lambda) y_2$$
pour tout $\lambda \in [0,1]$. Vérifier que $\hat{b}_\lambda$ est sans biais et calculer $\mathrm{Var}(\hat{b}_\lambda)$.
5) Déterminer λ_0 tel que
$$\mathrm{Var}(\hat{b}_{\lambda_0}) = \inf_{\lambda \in [0,1]} \mathrm{Var}(\hat{b}_\lambda).$$
6) A partir de la forme matricielle du modèle, calculer l'estimateur des moindres carrés généralisés $\hat{b}_{\mathrm{MCG}}$ d'après la formule du cours et vérifier que $\hat{b}_{\mathrm{MCG}} = \hat{b}_{\lambda_0}$. Commenter.

Exercice 2. On suppose qu'on a un modèle linéaire

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{j,i} + \varepsilon_i$$

vérifiant les hypothèses usuelles sauf l'homoscédasticité généralisée par $\mathrm{Var}(\varepsilon_i) = f(x_{1,i})$ où f est une fonction croissante :

- **(H0)** les valeurs des x_i sont supposées déterministes (non-aléatoires) ;
- **(H1)** les ε_i sont aléatoires avec $\mathbb{E}(\varepsilon_i) = 0$;
- **(H2)** les ε_i sont indépendants les uns des autres ;
- **(H3')** pour tout i, $\mathrm{Var}(\varepsilon_i) = f(x_{1,i})$ avec f fonction croissante ;
- **(H4)** les ε_i sont gaussiens ;
- **(H5)** la matrice $X'X$ est inversible.

TAB. 1 – Données pour l'Exercice 2, réordonnées.

i	x_i	y_i
1	16	20
2	18	24
3	23	28
4	24	22
5	26	32
6	28	32
7	29	28
8	31	36
9	32	41
10	34	41

En particulier, si f est une fonction constante, on est dans le cadre homoscédastique, mais si f n'est pas constante, on est dans un cadre particulier d'hétéroscédasticité. On souhaite tester l'hypothèse $\mathbf{H}_0$: "f est constante égale à σ^2" contre l'hypothèse $\mathbf{H}_1$: "f est strictement croissante".

Par commodité, on va supposer que $x_{1,1} \leq ... \leq x_{1,n}$, autrement dit, les gens qui ont fait l'enquête ont pris le temps de trier les observations par valeurs croissantes de la variable x_1. On coupe l'échantillon en deux parties, les individus i avec $i \in \{1, ..., \ell\}$ et ceux avec $i \in \{\ell+1, ..., n\}$. On introduit les matrices suivantes :

$$y_{(1)} = \begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix}, X_{(1)} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,\ell} & \dots & x_{p,\ell} \end{pmatrix}, \varepsilon_{(1)} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_\ell \end{pmatrix},$$

$$y_{(2)} = \begin{pmatrix} y_{\ell+1} \\ \vdots \\ y_n \end{pmatrix}, X_{(2)} = \begin{pmatrix} 1 & x_{1,\ell+1} & \dots & x_{p,\ell+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \dots & x_{p,n} \end{pmatrix} \text{ et } \varepsilon_{(2)} = \begin{pmatrix} \varepsilon_{\ell+1} \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On estime séparément les deux modèles $y_{(1)} = X_{(1)}b + \varepsilon_{(1)}$ et $y_{(2)} = X_{(2)}b + \varepsilon_{(2)}$. On récupère dans chaque modèle la quantité SCR, notée respectivement $\text{SCR}_{(1)}$ et $\text{SCR}_{(1)}$.

1) Sous $\mathbf{H}_0$, quelle est la loi de $\text{SCR}_{(1)}/\sigma^2$? Et la loi de $\text{SCR}_{(2)}$?
2) En déduire la loi de la statistique

$$\text{GQ} = \frac{\left(\frac{\text{SCR}_{(2)}}{(n-\ell)-(p+1)}\right)}{\left(\frac{\text{SCR}_{(1)}}{\ell-(p+1)}\right)}$$

si l'hypothèse $\mathbf{H}_0$ est vraie.
3) Que se passe-t'il si $\mathbf{H}_1$ est vraie ?
4) En déduire une procédure de test de l'hypothèse $\mathbf{H}_0$ contre l'hypothèse $\mathbf{H}_1$.
5) On traite maintenant une application numérique, on revient aux données de l'Exercice 2 du TD 2 : on souhaitait modéliser y le rendement de maïs (en quintal) d'une parcelle de terrain en fonction de x la quantité d'engrais utilisée (en kilos), et on avait les données complètes dans la Table 1. On prend $\ell = 5$, discuter ce choix.
6) On estime les deux modèles et on obtient $\text{SCR}_{(1)} = 12.891$ et $\text{SCR}_{(2)} = 9.459$. Calculer la statistique GQ.
7) Un logiciel statistique nous fournit la p-valeur correspondant à cette quantité : 0.40, que peut-on clonclure ?