HETEROSCEDASTICITE

Source : D.N.Gujarati, "Basic Econometrics", Third Ed., McGraw Hill, 1995

Le concept

L'une des hypothèses importantes de la régression est le fait que la variance des termes d'erreur est, conditionnellement à la valeur des variables explicatives, constante. Si l'on constate que la variance des termes d'erreur, conditionnellement à l'une des variables explicatives, a tendance à adopter un comportement systématique (par exemple, elle augmente lorsque la variable en question augmente), on parle d'hétéroscédastivité.

L'estimation par les moindres carrés ordinaires (OLS)

En présence d'hétéroscédasticité, les coefficients obtenus à l'aide des moindres carrés ordinaires sont toujours non biaisés. Ils ne sont cependant plus de variance minimum.

L'approche des moindres carrés généralisés (GLS)

Les moindres carrés ordinaires ne tiennent pas compte du fait que la variable explicative peut connaître un niveau de variance très différents selon les valeurs prises par la variable (ou les variables) explicatives.

Exemple : dans un modèle qui chercherait à expliquer le niveau de salaire en fonction du type de travail, il peut se faire que la variance des rémunérations varie très fortement d'un métier à l'autre.

L'approche des moindres carrés généralisées (GLS) permet d'en tenir compte. L'approche suivie est la suivante. Partant du modèle $Y_i = b_1 + b_2 X_i + u_i$ et réécrivant le modèle $Y_i = b_1 X_{0i} + b_2 X_i + u_i$ avec $X_{0i} = 1$ pour chaque i, on peut diviser diviser l'ensemble des termes par la variance des résidus (en supposant qu'elle soit connue et en la notant s_i pour bien indiquer qu'elle n'est pas constante). On obtient alors :

$$\frac{Y_i}{s_i} = b_1 \frac{X_{0i}}{s_i} + b_2 \frac{X_i}{s_i} + \frac{u_i}{s_i}$$

Cette approche revient donc à pondérer chaque observation par l'inverse de sa variance (une observation provenant d'une population dont la variance est plus forte "pèsera" moins fort). En remplaçant les variables originales par $Y_i^* = \frac{Y_i}{s_i}, X_{0i}^* = \frac{X_{0i}}{s_i}, X_i^* = \frac{X_i}{s_i}, u_i^* = \frac{u_i}{s_i}$. La variance des résidus sera à présent constante :

$$\text{var}(u_i^*) = E(u_i^*)^2 = E(\frac{u_i}{s_i})^2 = \frac{1}{s_i^2} E(u_i^2) = \frac{1}{s_i^2} s_i^2 = 1$$

Il suffit à présent d'estimater $Y_i^* = b_1 X^*{}_{0i} + b_2 X^*{}_i + u^*{}_i$ à l'aide de l'approche classique OLS.

L'identification de l'existence de l'hétéroscédasticité

La première approche consiste bien entendu à représenter graphiquement l'évolution les résidus en fonction des différentes variables explicatives (graphique X/Y) et de juger visuellement de la stabilité de la variance.

Au-delà, différents tests statistiques existent. Le coefficient de corrélation de range de Spearman peut par exemple être utilisé. Il se définit comme :

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right]$$

où d_i est la différence entre les rangs assignés pour deux caractéristiques d'un même individu et n est le nombre d'observations. Dans le cas qui nous occupe, la méthode à suivre est la suivante :

1. réaliser la régression et obtenir les résidus;
2. prendre la valeur absolue des résidus;
3. classer les résidus par ordre croissant et définir les rangs;
4. classer les individus par ordre croissant de la variable explicative prise en compte et définir les rangs;
5. calculer le coefficient de corrélation de Spearman en partant des différences entre les rangs par individu;

Pour autant que n > 8 et en posant l'hypothèse que le coefficient de corrélation de Spearman est nul, on peut tester la significativité de r_s à l'aide du test de student suivant

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

avec n-2 degré de liberté.