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

TD n° 5 : Tests

Ce TD concerne les procédure de tests. Dans l'exercice 1, on généralise le test de Student, vu en cours pour tester $b_j = 0$: on le modifie ici pour tester $b_j = 1$. Dans l'exercice 2, on étudie plusieurs modèles sur de vraies données. Les tests sont ici sensés nous aider à choisir l'un des différents modèles proposés.

Exercice 1. Dans le contexte du modèle linéaire général

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

avec les hypothèses usuelles **(H0)-(H5)**, et les notations matricielles habituelles, on souhaite construire un test de l'hypothèse $\mathbf{H}_0 : b_0 = 1$.

1) En utilisant le fait que l'estimateur $\hat{b} = (X'X)^{-1}X'Y$, rappeler la loi de $\hat{b}_0 - 1$ si $\mathbf{H}_0 : b_0 = 1$ est vraie.
2) En utilisant le fait que $X\hat{b}$ est la projection orthogonale de y sur $\{X\beta, \beta \in \mathbb{R}^p\}$, démontrer que $\hat{b}$ est indépendant de
$$\widehat{\sigma^2} = \frac{\|Y - X\hat{b}\|^2}{n - (p+1)}.$$
3) Quelle est la loi de $\widehat{\sigma^2}$?
4) On rappelle que pour tout j, $\hat{\sigma}_j = \sqrt{\widehat{\sigma^2}(X'X)^{-1}_{j,j}}$, quelle est la loi de la statistique
$$t = \frac{\hat{b}_0 - 1}{\hat{\sigma}_0} ?$$
5) En déduire une procédure de test de l'hypothèse $\mathbf{H}_0$.
6) Application numérique : avec $n = 1500$, $p = 12$, on obtient $\hat{b}_0 = 1500$ et $\hat{\sigma}_j = 950$, rejette-t'on $\mathbf{H}_0$ au seuil $\alpha = 0.05$?

Exercice 2. Un économètre souhaite modéliser la part alimentaire de la consommation des ménages constitués d'un couple et de deux enfants, C_i, en fonction de leur revenu, R_i, et d'autres facteurs qui sont la catégorie socio-professionnelle du chef de ménage, CSP_i (1 les agriculteurs, 2 les artisans, commerçants et chefs d'entreprises, 3 les cadres, professions intellectuelles supérieures, 4 les professions intermédiaires, 5 les employés, 6 les ouvriers), de l'âge du chef de famille age_i et du fait que le ménage réside dans la région Ile-de-France ou non, $\mathrm{IDF}_i = 1$ ou $\mathrm{IDF}_i = 0$. Il commence par consulter un certain nombre de manuels d'économie et en tire quatre idées de modèles :

$$\begin{cases} (1) \quad \ln(C_i) = a + b\ln(R_i) + c\ln(R_i)^2 + \alpha_1 \mathrm{CSP}_i + \alpha_2 \mathrm{age}_i + \alpha_3 \mathrm{IDF}_i + \varepsilon_i, \\ (2) \quad \ln(C_i) = a + \frac{b}{R_i} + \alpha_1 \mathrm{CSP}_i + \alpha_2 \mathrm{age}_i + \alpha_3 \mathrm{IDF}_i + \varepsilon_i, \\ (3) \quad C_i = a + b\ln(R_i) + c\ln(R_i)^2 + \alpha_1 \mathrm{CSP}_i + \alpha_2 \mathrm{age}_i + \alpha_3 \mathrm{IDF}_i + \varepsilon_i, \\ (4) \quad C_i = a + \frac{b}{R_i} + \alpha_1 \mathrm{CSP}_i + \alpha_2 \mathrm{age}_i + \alpha_3 \mathrm{IDF}_i + \varepsilon_i. \end{cases}$$

Il dispose d'un échantillon de $n = 1520$ individus (en fait, il s'agit ici de vraies données tirées de l'enquête "budget des familles" de l'INSEE en 1989).

1) L'économètre estime le modèle (1) et obtient :

$$\widehat{\ln(C_i)} = -8.7_{(3.33)} + 2.58_{(0.54)} \ln(R_i) - 0.08_{(0.02)} \ln(R_i)^2 \\ - 0.007_{(0.005)}\text{CSP}_i + 0.009_{(0.001)}\text{age}_i + 0.07_{(0.02)}\text{IDF}_i \text{ et } R^2 = 0.43,$$

la statistique de Fisher $F = 193.379$ et la p-valeur associée est 0.0001. Le modèle est-il globalement significatif ? Le coefficient c est-il significativement différent de 0 ?

2) Il estime ensuite le modèle (2) et obtient :

$$\widehat{\ln(C_i)} = 10.71_{(0.05)} - 79500_{(2872)} \frac{1}{R_i} \\ - 0.01_{(0.005)}\text{CSP}_i + 0.01_{(0.001)}\text{age}_i + 0.11_{(0.02)}\text{IDF}_i \text{ et } R^2 = 0.41,$$

la statistique de Fisher $F = 216.159$ et la p-valeur associée est 0.0001. Le modèle est-il globalement significatif ? Que penser globalement du modèle (2) par rapport au modèle (3) ?

3) De façon à choisir entre les deux modèles, l'économètre propose d'estimer le modèle suivant :

$$(1\&2) \quad \ln(C_i) = a + b \ln(R_i) + c \ln(R_i)^2 + \frac{d}{R_i} + \alpha_1 \text{CSP}_i + \alpha_2 \text{age}_i + \alpha_3 \text{IDF}_i + \varepsilon_i.$$

Il obtient

$$\widehat{\ln(C_i)} = 20.76_{(12.65)} - 1.86_{1.92} \ln(R_i) + 0.08_{(0.07)} \ln(R_i)^2 - 57363_{(23770)} \frac{1}{R_i} \\ - 0.007_{(0.005)}\text{CSP}_i + 0.009_{(0.001)}\text{age}_i + 0.07_{(0.02)}\text{IDF}_i \text{ et } R^2 = 0.43.$$

A priori, quel modèle devrait-on choisir entre (1) et (2) ? Comment pourrait-on rigoureusement tester ce choix ?

4) De la même façon, dans le but de choisir entre les modèles (3) et (4), l'économètre propose le modèle

$$(3\&4) \quad C_i = a + b \ln(R_i) + c \ln(R_i)^2 + \frac{d}{R_i} + \alpha_1 \text{CSP}_i + \alpha_2 \text{age}_i + \alpha_3 \text{IDF}_i + \varepsilon_i.$$

Dans ce modèle, quel(s) test(s) pourraient lui permettre de choisir entre les modèles (3) et (4) ? $R^2 = 0.41$

5) L'économètre compare ensuite les modèles (1&2) avec (3&4) et préfère le modèle (1&2) car son R^2 est plus élevé. Que penser de ce choix ?

6) Que penser de la variable CSP introduite par l'économètre ?