Econométrie

Répétition 9

1. Matière : Problèmes de spécification et variables explicatives binaires

- Utilisation des variables explicatives binaires.
- Test de la forme fonctionnelle.
- Test d'hétéroscédasticité et utilisation de $\hat{V}_{HC}(\hat{\beta})$.
- Test d'auto-corrélation et utilisation de $\hat{V}_{HAC}(\hat{\beta})$.
- Test de normalité.

2. Exercices

1- On considère les données du fichier R9_consommation.txt, où rev = le revenu par habitant et $cons$ = la consommation par habitant au cours des années 1929-1970.

a- Estimez par MCO les paramètres du modèle :

$$cons_t = \beta_1 + \beta_2 rev_t + e_t$$

Examinez graphiquement les valeurs observées $cons_t$, prédites $\hat{cons}_t$, les résidus $\hat{e}_t$ en fonction du temps. Pouvez-vous identifier la période de guerre ?

b- Estimez par MCO les paramètres du modèle :

$$cons_t = \beta_1 + \beta_2 G_t + \beta_3 rev_t + e_t\,,$$

où $G_t = 1$ si $t = 1941,...,1945$, 0 sinon. Réexaminez $cons_t$, $\hat{cons}_t$ et $\hat{e}_t$ en fonction du temps, et commentez.

c- Créez une variable binaire $P_t = 1 - G_t$ et réestimez le modèle en remplaçant G_t par P_t. Comparez les résultats obtenus. Sont-ils cohérents ?

d- Estimez par MCO les paramètres du modèle :

$$cons_t = \beta_1 + \beta_2 G_t + \beta_3 rev_t + \beta_4 rev_t G_t + e_t$$

i- Réexaminez $cons_t$, $\hat{cons}_t$ et $\hat{e}_t$ en fonction du temps, et commentez.
ii- Testez la significativité de β_2 et de β_4 par un F-test joint au seuil de 5%. Commentez.

2- On considère les données individuelles reprises dans le fichier R9_salaire.xlsx, dont la définition des variables est : Sal = le salaire horaire en \$, $Educ$ = le nombre d'années d'étude, $Expe$ = le nombre d'années d'expérience professionnelle, $Sexe = 1$ si c'est une femme et 0 sinon, $Syndic = 1$ si syndiqué et 0 sinon.

a- On envisage le modèle :

$$Sal_i = \beta_1 + \beta_2 Hom_i + \beta_3 Syndic_i + \beta_4 \left(Hom_i \times Syndic_i \right) + e_i \,,$$

où la variable $Hom_i = 1$ si l'individu i est un homme, 0 sinon.

i- Quelle est la signification des paramètres de ce modèle ?
ii- Estimez par MCO les paramètres de ce modèle.
iii- Calculez un intervalle de prévision à 95% pour (a) le salaire d'une femme non-syndiquée et (b) le salaire d'un homme syndiqué. Sur quelle hypothèse implicite la validité de vos calculs repose-t-elle ?
iv- Peut-on affirmer que les salaires moyens des travailleurs syndiqués sont différents de ceux des travailleurs non-syndiqués ? Justifiez votre réponse à l'aide d'un test.
v- Peut-on affirmer que les salaires moyens des hommes sont différents de ceux des femmes ? Justifiez votre réponse à l'aide d'un test.
vi- Peut-on affirmer que l'écart entre les salaires moyens des hommes et des femmes est différent selon que l'on envisage les travailleurs syndiqués ou non-syndiqués ? Justifiez votre réponse à l'aide d'un test.
vii- Peut-on affirmer que l'écart entre les salaires moyens des travailleurs syndiqués et non-syndiqués est différent selon que l'on envisage la population des hommes ou des femmes ? Justifiez votre réponse à l'aide d'un test.

b- On considère à présent le modèle :

$$\begin{aligned} Sal_i &= \beta_1 + \beta_2 Hom_i + \beta_3 Syndic_i + \beta_4 \left(Hom_i \times Syndic_i \right) \\ &\quad + \beta_5 Educ_i + \beta_6 Expe_i + e_i \end{aligned}$$

i- Estimez par MCO les paramètres de ce modèle et comparez-les aux paramètres estimés du modèle précédent. Commentez.
ii- Calculez un intervalle de prévision pour (a) le salaire d'une femme non-syndiquée ayant poursuivi 12 années d'études et ayant 5 années d'expérience professionnelle, et (b) le salaire d'un homme syndiqué ayant de même poursuivi 12 années d'études et ayant 5 années d'expérience professionnelle. Sur quelle hypothèse implicite la validité de vos calculs repose-t-elle ?
iii- Les résultats d'estimation de ce modèle suggèrent-ils l'existence d'une discrimination salariale entre hommes et femmes ? Justifiez votre réponse à l'aide d'un test. Aurait-on raisonnablement pu conclure à l'existence ou non d'une telle discrimination sur base du

modèle précédent ?

iv- Les résultats d'estimation de ce modèle suggèrent-ils l'existence d'une discrimination salariale entre travailleurs syndiqués et non-syndiqués ? Justifiez votre réponse à l'aide d'un test. Aurait-on raisonnablement pu conclure à l'existence ou non d'une telle discrimination sur base du modèle précédent ?

3- On considère le modèle :

$$\begin{aligned} Sal_i &= \beta_1 + \beta_2(Educ_i - 12) + \beta_3(Expe_i - 15) + \beta_4(Educ_i - 12)^2 \\ &\quad +\beta_5(Expe_i - 15)^2 + \beta_6((Educ_i - 12)(Expe_i - 15)) + e_i\,, \end{aligned} \tag{1}$$

où : Sal_i = le salaire horaire de l'individu i
$Educ_i$ = le nombre d'années d'étude de l'individu i
$Expe_i$ = le nombre d'années d'expérience professionnelle de l'individu i

pour un ensemble d'observations américaines (la variable Sal est le salaire horaire mesuré en \$) contenu dans le fichier R8_salaire.txt.

a- Testez la forme fonctionnelle du modèle :

i- en considérant l'ajout des variables $(Educ_i - 12)^3$, $(Expe_i - 15)^3$, $(Educ_i - 12)^2(Expe_i - 15)$ et $(Educ_i - 12)(Expe_i - 15)^2$.

ii- au travers d'un test RESET, en considérant l'ajout des variables $\hat{Sal}_i^2$ et $\hat{Sal}_i^3$.

b- Testez l'hypothèse d'homoscédasticité en considérant comme variables de la régression auxiliaire les variables du modèle (1). Effectuez ce test :

i- en estimant le modèle auxiliaire, puis en réalisant un F-test de la significativité de la régression dans son ensemble (nullité de tous les paramètres, sauf l'intercept).

ii- en estimant le modèle auxiliaire, puis en réalisant un test basé sur la statistique de test $LM_H = n \times R^2$ (utiliser le fonction 'Save - T*R-squared de GRETL).

iii- sur base du l'estimation du modèle (1), en utilisant la fonction 'Tests - Heteroskedasticity - Koenker' de GRETL[1].

c- Testez l'hypothèse d'homoscédasticité en considérant comme variables de la régression auxiliaire les valeurs prédites $\hat{Sal}_i$ et $\hat{Sal}_i^2$ du modèle (1). Effectuez ce test :

i- en estimant le modèle auxiliaire, puis en réalisant un F-test de la significativité de la régression dans son ensemble (nullité de tous les paramètres, sauf l'intercept).

ii- en estimant le modèle auxiliaire, puis en réalisant un test basé sur la statistique de test $LM_H = n \times R^2$.(utiliser le fonction 'Save - T*R-squared de GRETL).

d- Testez l'hypothèse de normalité de ce modèle en utilisant la fonction 'Tests - Normality of Residuals' de GRETL.

[1] Il ne faut pas utiliser la fonction 'Tests - Heteroskedasticity - Breush-Pagan', car celle-ci effectue un test qui n'est valable que sous l'hypothèse de normalité.

e- Estimez le modèle log-lin alternatif :

$$\begin{aligned}\ln Sal_i &= \beta_1 + \beta_2(Educ_i - 12) + \beta_3(Expe_i - 15) + \beta_4(Educ_i - 12)^2 \\ &\quad + \beta_5(Expe_i - 15)^2 + \beta_6((Educ_i - 12)(Expe_i - 15)) + e_i\end{aligned} \tag{2}$$

i- Testez la forme fonctionnelle de ce modèle à l'aide d'un test RESET, en considérant l'ajout des variables $\hat{Sal}_i^2$ et $\hat{Sal}_i^3$.

ii- Testez l'hypothèse d'homoscédasticité de ce modèle en utilisant la fonction 'Tests - Heteroskedasticity - Koenker' de GRETL.

iii- Testez l'hypothèse de normalité de ce modèle en utilisant la fonction 'Tests - Normality of Residuals' de GRETL.

f- La validité de l'hypothèse d'homoscédasticité du modèle (2) étant plausible, mais pas certaine, réestimez le modèle avec l'option 'Robust Standard Errors - HC0' et comparez les résultats à ceux obtenus précédemment. Ensuite :

i- calculez un intervalle de confiance à 99% pour β_2.

ii- testez au seuil de 5%, à l'aide d'un t-test, $\mathrm{H}_0 :\ \beta_2 \leq 0,06$ contre $\mathrm{H}_1 :\ \beta_2 > 0,06$. Quelle est la P-valeur de ce test ?

iii- testez, à l'aide d'un F-test, Testez $\mathrm{H}_0 :\ \beta_4 = \beta_5 = \beta_6 = 0$ contre $\mathrm{H}_1 :\ \beta_4 \neq 0$ et/ou $\beta_5 \neq 0$ et/ou $\beta_6 \neq 0$. Effectuez ce test (1) en utilisant les fonctions de calcul matriciel de GRETL, et (2) en utilisant la fonction 'Tests - Omit Variables - Wald test' de GRETL.

4- On considère le modèle :

$$\ln(Sup_t) = \beta_1 + \beta_2 \ln(Prix_t) + e_t\,, \tag{3}$$

où : $Sup_t =$ la superficie plantée de canne à sucre au cours de l'année t dans une région du Bangladesh (en milliers d'hectares)

$Prix_t =$ un index du prix (relatif) de la canne à sucre au cours de l'année t

pour un ensemble d'observations annuelles contenues dans le fichier R9_superficie.txt.

a- Quelle est la signification du paramètre β_2 ? Quel est le signe attendu de ce paramètre ?

b- Estimez par MCO les paramètres du modèle, et examinez graphiquement les résidus $\hat{e}_t$ en fonction du temps et de la variable $\ln(Prix_t)$. Ces graphiques suggèrent-ils une mauvaise spécification de la forme fonctionnelle ? une violation de l'hypothèse d'homoscédasticité ? une violation de l'hypothèse de non-corrélation ?

c- Testez la forme fonctionnelle du modèle en considérant :

i- l'ajout de la variable $\ln(Prix_t)^2$.

ii- l'ajout de la variable prédite $\hat{Sup}_t^2$.

d- Testez l'hypothèse d'homoscédasticité du modèle en utilisant la fonction 'Tests - Heteroskedasticity - Koenker' de GRETL.

e- Testez l'hypothèse de non-corrélation en considérant comme variables de la régression auxiliaire, outre les variables du modèle (3), les résidus retardés $\hat{e}_{t-1}$. Effectuez ce test :

i- en estimant le modèle auxiliaire, puis en réalisant un F-test de la significativité du paramètre des résidus retardés $\hat{e}_{t-1}$.

ii- en estimant le modèle auxiliaire, puis en réalisant un test basé sur la statistique de test $LM_A = T \times R^2$ (utiliser le fonction 'Save - T*R-squared de GRETL).

iii- sur base du l'estimation du modèle (3), en utilisant la fonction 'Tests - Autocorrelation - Lag Order : 1' de GRETL[2].

f- Réestimez le modèle avec l'option 'Robust Standard Errors - HAC' et comparez les résultats à ceux obtenus au point b.

5- On considère le modèle :

$$Prix_i = \beta_1 + \beta_2 sup_i + \beta_3 Chamb_i + \beta_4 D_{V_i} + \beta_5 D_{G_i} + \beta_6 D_{V_i} D_{G_i} + e_i \,,$$

où : $Prix_i =$ le prix de vente de la maison i (en millions de Frs)
$sup_i =$ la superficie habitable de la maison i (en m^2)
$Chamb_i =$ le nombre de chambres de la maison i
$D_{V_i} = 1$ si la maison i est située en milieu urbain, 0 sinon
$D_{G_i} =$ 1 si la maison i possède un garage, 0 sinon

L'estimation préliminaire de ce modèle sur un échantillon de maisons vendues en région liégeoise en 1999, suivie d'un test d'hétéroscédasticité de Breush-Pagan, ayant montré que l'hypothèse d'homoscédasticité n'était pas satisfaite, le modèle a été réestimé avec l'option 'Robust Standard Error - HC0' de GRETL, qui a donné les résultats suivants :

Model 1:
OLS, using observations 1-273
Dependent variable: PRIX
Heteroskedasticity-robust standard errors, variant HC0

	coefficient	std. error	t-ratio	p-value
const	-0.249	0.137	-1.817	0.0346 **
SUP	0.025	0.004	6.250	2.71e-09 ***
CHAMB	0.158	0.027	5.851	2.15e-08 ***
D_V	0.215	0.079	2.721	0.0033 ***
D_G	0.173	0.068	2.544	0.0055 ***
$D_V D_G$	0.086	0.032	2.687	0.0036 ***

a- Quel est le prix moyen estimé de vente au m^2 d'une maison de 140 m^2, comprenant 2 chambres et 1 garage, et située en zone rurale ?

b- Peut-on considérer que la présence d'une chambre supplémentaire accroît la valeur vénale moyenne d'une maison d'au moins 175 000 frs ?

c- Testez l'hypothèse nulle que la présence d'un garage n'accroît pas plus la valeur vénale moyenne d'une maison située en milieu urbain que celle d'une maison située en zone rurale. Que concluez-vous ?

[2] L'utilisation de la fonction 'Tests - Autocorrelation' de GRETL donne des résultats légèrement différents de ceux obtenus ci-avant car plutôt que d'estimer la régression auxiliaire en omettant la première observation, GRETL utilise ici toutes les observations, en attribuant une valeur nulle au résidu retardé $\hat{e}_0$ de la première observation.