Introduction à l'économétrie

Tests usuels sur le modèle linéaire multiple

Ce cours vous est proposé par Olivier Baron, Maître de conférences, Université de Bordeaux et par AUNEGe, l'Université Numérique en Économie Gestion.

Illustrations et compléments

Un exemple d'application commentée

Considérons une fonction de production Cobb-Douglas de la forme :

$$\boldsymbol{Q_i = A.e^{\gamma.t}.L_t^{\alpha}.K_t^{\beta}} \qquad \textbf{(A)}$$

où Q_t désigne le volume de la production à la date *t*, L_t et K_t les services du travail et du capital physique utilisés à cette date. Le terme $e^{\gamma.t}$ formalise l'effet du progrès technique.

L'équation (A) doit être transformée si l'on souhaite utiliser la méthode des Moindres Carrés Ordinaire pour l'estimer. En effet, le second membre de l'équation (A) est une fonction non linéaire des paramètres α, β, ,γ et A. La méthode des MCO n'est donc pas applicable dans ce cas.

Pour pouvoir utiliser cette méthode d'estimation, il faut linéariser l'équation (A) en prenant, par exemple, le logarithme népérien de chacun des deux membres. On obtient ainsi :

$$\ln Q_t = \ln\left(A.e^{\gamma.t}.L_t^{\alpha}.K_t^{\beta}\right) = \ln A + \ln e^{\gamma.t} + \ln L_t^{\alpha} + \ln K_t^{\beta}$$

$$\Leftrightarrow \quad \ln Q_t = \ln A + \gamma.t + \alpha.\ln L_t + \beta.\ln K_t$$

En posant $q_t = \ln Q_t$, $l_t = \ln L_t$ et $k_t = \ln K_t$, l'équation ci-dessus s'écrit finalement :

$$q_t = \ln A + \gamma.t + \alpha.\ l_t + \beta.k_t$$

On obtient alors une spécification linéaire en fonction des paramètres qui peut être estimée par la méthode des MCO. Pourtant, on préfèrera par la suite, estimer cette équation en taux de croissance à partir de l'équation suivante :

$$\dot{q}_t = \gamma + \alpha.\ \dot{l}_t + \beta.\dot{k}_t + \omega_t \qquad \text{(B)}$$

où ω_t est un terme aléatoire. Les taux de croissance des différentes variables $\dot{q}_t$, $\dot{l}_t$ et $\dot{k}_t$ sont définis comme les différences premières des logarithmes des variables initiales ($\dot{q}_t = \ln Q_t - \ln Q_{t-1} = q_t - q_{t-1}$, etc ... ...).

Pour obtenir l'équation (B), il suffit d'écrire l'équation (A) à la date $t-1$. On a :

$$q_{t-1} = \ln A + \gamma.(t-1) + \alpha.\ l_{t-1} + \beta.k_{t-1}$$

et donc :

$$q_t - q_{t-1} = \dot{q}_t = \ln A + \gamma.t + \alpha.\ l_t + \beta.k_t - (\ln A + \gamma.(t-1) + \alpha.\ l_{t-1} + \beta.k_{t-1})$$

$$\Leftrightarrow \quad \dot{q}_t = \gamma + \alpha.(l_t - l_{t-1}) + \beta.(k_t - k_{t-1})$$

$$\Leftrightarrow \quad \dot{q}_t = \gamma + \alpha.\dot{l}_t + \beta.\dot{k}_t$$

Bien entendu, pour arriver à la spécification finale, il faut rajouter une perturbation aléatoire, notée ω_t, qui rend compte de l'ensemble des facteurs explicatifs de la production non inclus dans la liste des variables explicatives de ce modèle. On peut noter que compte tenu de la forme log-log de cette spécification, les paramètres α et β s'interprètent comme les élasticités de l'output relativement aux deux facteurs de production, le travail et le capital.

On dispose pour estimer cette équation de données trimestrielles agrégées pour l'économie française allant du deuxième trimestre 1962 au troisième trimestre 1996, soit 136 observations sur la variable dépendante et les deux variables explicatives.

L'application des MCO au modèle (B) conduit aux résultats suivants :

$$\left| \begin{array}{l} \widehat{\dot{q}_t} = 0.003 + 0.558\,\dot{l}_t + 0.393\dot{k}_t \\ R^2 = 0.33\,;SCR = 0.00664 \end{array} \right.$$

Supposons que l'on donne partiellement l'expression de la matrice $(X'X)^{-1}$, soit :

$$(X'X)^{-1} = \begin{pmatrix} 0.02 & . & . \\ . & 480.5 & . \\ . & . & 118.58 \end{pmatrix}$$

On peut alors en déduire les valeurs des écart-types estimés des coefficients estimés.

On sait que l'écart-type estimé d'un coefficient estimé $\hat{\beta}_j$ est donné par :

$\hat{\sigma}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2.x^{jj}}$ où $\hat{\sigma}^2$ est l'estimateur de la variance de la perturbation et x^{jj} le j[ième] élément de la diagonale principale de la matrice $(X'X)^{-1}$. Pour estimer la variance de la perturbation, on utilise $\hat{\sigma}^2 = \frac{SCR}{n-k}$ avec $SCR = 0.00664$, $n = 136$ et $k = 3$.

On obtient $\hat{\sigma}^2 = \frac{0.00664}{133} = 0.00005$. On en déduit les écart-types estimés des coefficients estimés :

$$\begin{cases} \hat{\sigma}(\hat{\gamma}) = \sqrt{0.00005 * 0.02} = 0.001 \\ \hat{\sigma}(\hat{\alpha}) = \sqrt{0.00005 * 480.5} = 0.155 \\ \hat{\sigma}(\hat{\beta}) = \sqrt{0.00005 * 118.58} = 0.077 \end{cases}$$

Il nous reste à **tester la significativité des 3 coefficients apparaissant dans le modèle (B).**

Pour tester la significativité d'un coefficient β_j, c'est-à-dire mettre en œuvre le test :

$$H_0 : \beta_j = 0$$
$$contre\ \overline{H} : \beta_j \neq 0$$

on utilise le résultat suivant :

$$\textbf{Si } \mathbf{H_0} \textbf{ est vraie} \text{ alors: } t_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \rightarrow T_{n-k}$$

Les calculs des statistiques t_j sont immédiats et peuvent être regroupés dans le tableau ci-dessous :

Variable	$\hat{\beta}_j$	$\hat{\sigma}(\hat{\beta}_j)$	$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$	*p-value* : α_j
Constante	0.003	0.001	3	0.0027
$\dot{l}_t$	0.558	0.155	3.6	0.0003
$\dot{k}_t$	0.393	0.077	5.1	0.0000

Pour prendre une décision, il faut comparer les valeurs calculées de la statistique avec la valeur critique de la loi de Student à 133 degrés de liberté.

La table fournit le résultat suivant : $Prob\{|T_{133}| > 1{,}98\} = 0.05$

Les valeurs calculées sont toutes supérieures à la valeur critique (1.98) et donc on rejette l'hypothèse nulle dans les 3 cas. On en déduit que les 3 coefficients de l'équation (2) sont significativement différents de zéro au seuil de 5%.

La dernière colonne du tableau fournit les *p-values* associées aux trois tests précédents soit :

$$\alpha_j = Prob\{|T_{133}| > t_j\}$$

Les trois *p-values* sont inférieures au seuil fixé de 5%, ce qui confirme le rejet de l'hypothèse nulle dans chaque cas.

On peut maintenant **chercher à retrouver la valeur prise par la statistique *f*** permettant de réaliser le **test de significativité globale de la régression**.

Le test de significativité globale de la régression consiste à tester l'hypothèse selon laquelle tous les coefficients sont nuls sauf la constante, soit dans le cas présent :

$$H_0 : \alpha = \beta = 0$$
$$contre\ \bar{H} : \text{ au moins un des coefficients est non nul}$$

On a vu que la statistique permettant de prendre une décision est égale à $f = \frac{R^2}{1-R^2}.\frac{n-k}{r}$, soit ici : $f = \frac{0.33}{1-0.33}.\frac{133}{2} = 32.75$

Le test est basé sur le résultat suivant :

Si H₀ est vraie alors : $f = \frac{R^2}{1-R^2}.\frac{n-k}{r} \rightarrow F_{(k-1,n-k)}$

Pour une loi de Fisher $F_{(2,133)}$, on trouve dans une table statistique la valeur critique associée à un risque de première espèce égal à 5% : $f^*_{0,05} = 3.06$. Puisque $32.75 \gg 3.06$, $f \gg f^*_{0,05}$, et l'hypothèse nulle est donc rejetée. Au moins un des coefficients de pente est non nul. Ce résultat est cohérent avec les conclusions des tests de Student réalisés précédemment.

Posons-nous à présent la question suivante : **compte tenu de la valeur estimée du paramètre de rendements d'échelle $(\alpha + \beta)$[1], que peut-on dire sur la nature des rendements d'échelle ?**

Pour construire un estimateur de $(\alpha + \beta)$ il suffit de remarquer que :

$$E(\hat{\alpha} + \hat{\beta}) = E(\hat{\alpha}) + E(\hat{\beta}) = \alpha + \beta$$

$\widehat{(\alpha + \beta)} = \hat{\alpha} + \hat{\beta}$ est donc un estimateur sans biais de $\alpha + \beta$. Même si l'estimation du paramètre de rendements d'échelle est inférieure à 1 $(0.558 + 0.393 = 0.951)$, tout en lui étant proche, on ne peut rien dire sur la nature des rendements d'échelle, compte tenu des variances affectant les estimations des paramètres α et β et donc de celle de $(\alpha + \beta)$. Un test de Student pourrait permettre de statuer sur cette question mais nécessite la connaissance de $Var(\hat{\alpha} + \hat{\beta})$, grandeur

[1] Dans une fonction de production Cobb-Douglas du type $Q_i = A.e^{\gamma.t}.L_t^{\alpha}.K_t^{\beta}$, le paramètre de rendements d'échelle est égal à la somme des exposants $\alpha + \beta$. Selon que ce paramètre est inférieur, égal ou supérieur à 1, les rendements d'échelle seront dits décroissants, constants ou croissants.

non calculable ici car dépendant de la covariance entre les variables aléatoires $\hat{\alpha}$ et $\hat{\beta}$. Les données de l'exercice ne permettent pas de calculer cette covariance.

On s'intéresse maintenant plus rigoureusement à la question de la nature des rendements d'échelle. On rappelle que la propriété de rendements d'échelle constants s'écrit : $\alpha + \beta = 1$.

En notant $\underset{(3,1)}{b}$ le vecteur des paramètres à estimer du modèle (B), écrivons cette contrainte sous la forme $C.b = q$, où C et q sont deux matrices à expliciter.

La contrainte $\alpha + \beta = 1$ peut s'écrire de la façon suivante :

$$(0 \quad 1 \quad 1).\begin{pmatrix}\gamma\\ \alpha\\ \beta\end{pmatrix} = 1$$

où $\underset{(1,3)}{C} = (0 \quad 1 \quad 1)$, $\underset{(3,1)}{b} = \begin{pmatrix}\gamma\\ \alpha\\ \beta\end{pmatrix}$ et $\underset{(1,1)}{q} = 1$.

Montrons à présent comment l'intégration de cette contrainte au modèle (B) permet de le réécrire sous la forme :

$$\dot{q}_t - \dot{l}_t = \gamma + (1-\alpha).(\dot{k}_t - \dot{l}_t) + \nu_t$$

Pour obtenir cette dernière écriture, il suffit de partir du modèle (B) et de substituer au paramètre β son expression, qui, sous la contrainte envisagée vaut, $(1-\alpha)$. On peut noter que les perturbations des deux modèles sont notées différemment. Ceci vise à se prémunir contre l'éventualité de la non pertinence de la contrainte $\alpha + \beta = 1$. On obtient :

$$\dot{q}_t = \gamma + \alpha.\dot{l}_t + \beta.\dot{k}_t + \omega_t$$

$$\Leftrightarrow \quad \dot{q}_t = \gamma + \alpha.\dot{l}_t + (1-\alpha).\dot{k}_t + \nu_t$$

$$\Leftrightarrow \quad \dot{q}_t - \dot{l}_t = \gamma + (\alpha-1).\dot{l}_t + (1-\alpha).\dot{k}_t + \nu_t$$

$$\Leftrightarrow \quad \dot{q}_t - \dot{l}_t = \gamma + (1-\alpha).(\dot{k}_t - \dot{l}_t) + \nu_t \qquad \text{(C)}$$

Il nous reste à **donner une interprétation économique des variables figurant dans le modèle (C)**.

La variable dépendante du modèle (C) peut être transformée de la façon suivante :

$$\dot{q}_t - \dot{l}_t = q_t - q_{t-1} - (l_t - l_{t-1}) = \ln Q_t - \ln Q_{t-1} - (\ln L_t - \ln L_{t-1})$$

$$\Leftrightarrow \quad \dot{q}_t - \dot{l}_t = \ln\frac{Q_t}{Q_{t-1}} - \ln\frac{L_t}{L_{t-1}} = \ln\frac{\frac{Q_t}{Q_{t-1}}}{\frac{L_t}{L_{t-1}}} = \ln(\frac{Q_t}{Q_{t-1}}.\frac{L_{t-1}}{L_t})$$

$$\Leftrightarrow \quad \dot{q}_t - \dot{l}_t = \ln\left(\frac{Q_t}{L_t}.\frac{L_{t-1}}{Q_{t-1}}\right) = \ln\frac{\frac{Q_t}{L_t}}{\frac{Q_{t-1}}{L_{t-1}}} = \ln\frac{Q_t}{L_t} - \ln\frac{Q_{t-1}}{L_{t-1}}$$

$$\Leftrightarrow \quad \dot{q}_t - \dot{l}_t = \left(\frac{q}{l}\right)_t - \left(\frac{q}{l}\right)_{t-1} = \left(\frac{\dot{q}}{l}\right)_t$$

La variable $\dot{q}_t - \dot{l}_t$ s'interprète donc comme le taux de croissance de la production par tête. On montrerait de même que $\dot{k}_t - \dot{l}_t$ peut s'interpréter comme le taux de croissance du capital par tête.

L'estimation du modèle (C) à partir des mêmes données que précédemment aboutit aux résultats suivants (les écart-types estimés des coefficients estimés sont donnés entre parenthèses sous les coefficients estimés) :

$$\left|\begin{array}{c} \widehat{\dot{q}_t - \dot{l}_t} = \underbrace{0.003}_{(0.001)} + \underbrace{0.395}_{(0.077)}\left(\dot{k}_t - \dot{l}_t\right) \\ R^2 = 0.329\,;SCR = 0.00665 \end{array}\right.$$

Construisons la statistique du test de la contrainte de constance des rendements d'échelle et déterminons la décision à prendre pour un seuil de 5%.

On teste ici le modèle contraint (C) contre le modèle non contraint (B). Pour calculer la valeur de la statistique permettant de prendre une décision, on utilise indifféremment l'une des expressions (21) ou (22) données au cours de ce chapitre. Cette statistique mesure l'accroissement de la variation résiduelle qui résulte de l'imposition de la contrainte. Nous rejetterons celle-ci si cet accroissement est jugé « trop grand ». Choisissons par exemple l'expression (21) qui fait intervenir les sommes des carrés des résidus estimés et les degrés de liberté des régressions contrainte et non contrainte. On a, dans le cas présent :

$$\begin{cases} SCR_{nc} = 0.00664 & SCR_c = 0.00665 \\ dl_{nc} = n - k = 133 & dl_c = n - (k - r) = 134 \end{cases}$$

Sous cette forme, la statistique s'écrit : $f = \frac{\frac{SCR_c - SCR_{nc}}{dl_c - dl_{nc}}}{\frac{SCR_{nc}}{dl_{nc}}} = \frac{0.00001}{0.00664} * 133 = 0.2$

Le test est basé sur le résultat suivant :

Si H₀ est vraie (si la contrainte est pertinente) alors : $f = \frac{\frac{SCR_c - SCR_{nc}}{dl_c - dl_{nc}}}{\frac{SCR_{nc}}{dl_{nc}}} \rightarrow F_{(1,133)}$

Pour un risque de première espèce égal à 5%, la valeur critique $f^*_{0.05}$ de la loi de Fisher ci-dessus est donnée par :

$$Prob\left\{F_{(1,133)} > f^*_{0.05}\right\} = 0.05$$

La table de la loi de Fisher nous donne $f^*_{0.05} = 3.91$

Décision : Puisque la valeur calculée de la statistique est inférieure à la valeur critique, on ne rejette pas la contrainte étudiée. On considère donc que, pour cette fonction de production Cobb-Douglas, les rendements d'échelle peuvent être considérés comme constants.

On dispose maintenant de données désagrégées sur un échantillon de 168 entreprises françaises appartenant à 2 secteurs de production (Parachimie-Pharmacie et Textile-Habillement) permettant d'estimer la même fonction de production Cobb-Douglas spécifiée en taux de croissance de la forme :

$$\dot{q}_i = \gamma + \alpha.\dot{l}_i + \beta.\dot{k}_i + u_i$$

Sur l'échantillon total $(n = 168)$, on obtient les résultats suivants (les écart-types estimés des coefficients estimés sont donnés entre parenthèses sous les coefficients estimés) :

$$\left|\begin{array}{l} \widehat{\dot{q}_i} = \underbrace{0.03}_{(0.021)} + \underbrace{0.658}_{(0.395)}\ \dot{l}_i + \underbrace{0.263}_{(0.172)}\dot{k}_i \\ R^2 = 0.12\ ; SCR = 2.954 \end{array}\right.$$

Sur 61 entreprises du secteur Parachimie-Pharmacie (secteur T12), on obtient :

$$\left|\begin{array}{l} \widehat{\dot{q}_i} = -\underbrace{0.021}_{(0.016)} + \underbrace{0.532}_{(0.290)}\ \dot{l}_i + \underbrace{0.233}_{(0.181)}\dot{k}_i \\ R^2 = 0.15\ ; SCR = 0.9377 \end{array}\right.$$

Sur 107 entreprises du secteur Textile-Habillement (secteur T18), on obtient :

$$\left|\begin{array}{l} \widehat{\dot{q}_i} = \underbrace{0.049}_{(0.012)} + \underbrace{0.730}_{(0.163)}\ \dot{l}_i + \underbrace{0.196}_{(0.102)}\dot{k}_i \\ R^2 = 0.19\ ; SCR = 1.4223 \end{array}\right.$$

La question posée est la suivante : peut-on considérer que les coefficients de ces deux dernières estimations sont significativement différents ?

Les résultats suggèrent que l'élasticité production–emploi (α) est plus élevée dans le secteur T18 que dans le secteur T12. En revanche, l'élasticité production–capital (β) serait plus élevée dans le secteur T12. Mais compte tenu des variances affectant les estimations, un tel commentaire n'est pas légitime. Il convient alors d'examiner si les différences observées sont significatives ou si l'on peut considérer que les coefficients de la fonction de production sont identiques dans les deux secteurs. Dans ce cas, on conclura que la productivité des facteurs est la même dans les deux secteurs. On peut répondre à ce questionnement en utilisant la méthodologie du test de Chow.

Pour appliquer ce test, on calcule la statistique donnée par l'expression (21). Pour cela, il suffit de connaître les valeurs des sommes des carrés des résidus estimés et des degrés de liberté des modèle contraint et non contraint.

Ici, l'estimation contrainte consiste à estimer la fonction de production en empilant les observations relatives aux 168 entreprises des deux secteurs. On trouve :

$$SCR_c = 2.954 \text{ et } dl_c = 165$$

Estimer le modèle non contraint revient à estimer le modèle secteur par secteur. On en déduit :

$$SCR_{nc} = SCR_1 + SCR_2 = 0.9377 + 1.4223 = 2.36 \text{ et } dl_{nc} = dl_1 + dl_2 = 58 + 104 = 162$$

Mais, comme on l'a vu, **la bonne démarche consiste à commencer par tester l'homoscédasticité des perturbations entre les deux secteurs**.

Remarque

Avant de procéder à la mise en œuvre du test de Goldfeld et Quandt et du test de Chow, on peut remarquer que l'estimation sur le secteur T12 fournit une bonne illustration des divergences entre un test de Fisher et des tests de significativité réalisés coefficient par coefficient. En effet, pour cette estimation, aucun paramètre n'est significatif au seuil de 5% (les *t* de Student valent respectivement -1.31, 1.83 et 1.29 alors que la valeur critique de la loi au seuil de 5% est égale à $t^*_{0,05} = 2.00$). Par contre, pour le test de significativité globale, $f = \frac{R^2}{1-R^2}.\frac{n-k}{r} = 5.12$ alors que la valeur critique de la loi de Fisher considérée $\left(F_{(2,58)}\right)$ vaut $f^*_{0,05} = 3.156$.

- **Test d'homoscédasticité des perturbations**

Les résultats des estimations relatives aux secteurs T12 et T18 permettent d'estimer les variances des perturbations pour chacun des deux groupes. On aura :

Secteur T12 : $\hat{\sigma}_1^2 = \frac{SCR_1}{n_1-k} = \frac{0.9377}{58} = 0.0162$

Secteur T18 : $\hat{\sigma}_2^2 = \frac{SCR_2}{n_2-k} = \frac{1.4223}{104} = 0.0137$

Puisque nous observons $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$, on procède au test unilatéral suivant :

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$contre\ \bar{H} : \sigma_1^2 > \sigma_2^2$$

Le test est basé sur le résultat suivant :

Si H₀ est vraie alors : $\varphi_1 = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \rightarrow F_{(58,104)}$

Avec un seuil de 5%, il suffit de comparer la valeur calculée de φ_1, soit $\varphi_1 = \frac{0.0162}{0.0137} = 1.1825$ et la valeur critique de la loi de Fisher correspondante, soit $f_{0,05}^* = 1.45$. Puisque $\varphi_1 < f_{0,05}^*$, on ne rejette pas H_0 et on considère donc que la variance de la perturbation est la même dans les deux sous-échantillons.

- **Test d'homogénéité des comportements**

Compte tenu du résultat précédent, le test de Chow appliqué à cet exemple peut s'écrire de la façon suivante :

$$\begin{cases} H_0 : & \dot{q}_i = \gamma + \alpha.\dot{l}_i + \beta.\dot{k}_i + u_i \quad \forall i \in \{T12 \cup T18\} \\ contre\ \bar{H} & : \begin{cases} \dot{q}_i = \gamma_1 + \alpha_1.\dot{l}_i + \beta_1.\dot{k}_i + u_i \quad \forall i \in \{T12\} \\ \dot{q}_i = \gamma_2 + \alpha_2.\dot{l}_i + \beta_2.\dot{k}_i + u_i \quad \forall i \in \{T18\} \end{cases} \end{cases}$$

Ici encore, on teste le modèle contraint contre le modèle non contraint. Pour prendre une décision, on calcule la statistique *f* à l'aide de l'expression (21)

avec : $\begin{cases} SCR_c = 2.954 \text{ et } dl_c = 165 \\ SCR_1 = 0.9377 \text{ et } dl_1 = 58 \\ SCR_2 = 1.4223 \text{ et } dl_2 = 104 \end{cases}$ soit :

$$f = \frac{\frac{SCR_c - (SCR_1 + SCR_2)}{k}}{\frac{(SCR_1 + SCR_2)}{n - 2k}} = \frac{\frac{2.954 - (0.9377 + 1.4223)}{3}}{\frac{(0.9377 + 1.4223)}{162}} = \frac{0.198}{0.0146} = 13.56$$

Le test est basé sur le résultat suivant :

Si H₀ est vraie alors : $f \rightarrow F_{(3,162)}$

En notant $f_{0,05}^*$ la valeur critique de cette loi de Fisher pour un risque de première espèce de 5%, la table fournit $f_{0,05}^* = 2.66$.

La valeur calculée est largement supérieure à la valeur critique, ce qui nous conduit à rejeter l'hypothèse nulle de ce test. Les élasticités de la production relativement aux facteurs de production sont significativement différentes dans les secteurs de la pharmacie et de l'habillement.

Interprétation des coefficients d'une régression linéaire

Pour des raisons pédagogiques, utilisons une application de la régression linéaire par MCO afin de présenter la façon d'interpréter les coefficients d'un modèle. Considérons l'exemple classique dans lequel nous voulons estimer, entre autres, l'impact du nombre d'années d'études d'un individu sur son salaire. Nous disposons de données en coupe transversale avec *n* observations individuelles sur les variables suivantes :

W_i : le salaire de l'individu *i*

Ed_i : le nombre d'années d'études de l'individu *i*

$Anci_i$: l'ancienneté dans l'emploi de l'individu *i*

Hom_i : variable indicatrice prenant la valeur 1 si l'individu *i* est un homme, 0 sinon.

Le modèle niveau-niveau

Considérons le modèle suivant estimé par Moindres Carrés Ordinaires :

$$W_i = \beta_0 + \beta_1.Ed_i + \beta_2.Anci_i + \beta_3.Anci_i^2 + \beta_4.Hom_i + \varepsilon_i \qquad \forall i \in \{1;n\} \qquad (1)$$

avec ε_i le terme d'erreur.

Dans l'équation (1), le coefficient β_1 s'interprète comme l'effet marginal d'une année supplémentaire d'études sur le salaire. Cela correspond à la variation de β_1 unités du salaire individuel induite par la variation d'une unité du niveau d'études, *toutes choses égales par ailleurs*, c'est-à-dire en considérant que l'ancienneté et le genre de l'individu sont fixés et constants. Formellement, il s'agit de la dérivée partielle du salaire relativement à la variable d'éducation. En effet :

$$\frac{\partial W_i}{\partial Ed_i} = \frac{\partial\left(\beta_0 + \beta_1.Ed_i + \beta_2.Anci_i + \beta_3.Anci_i^2 + \beta_4.Hom_i + \varepsilon_i\right)}{\partial Ed_i} = \beta_1$$

Dans l'équation (1), le coefficient β_4, associé à la variable binaire Hom_i va mesurer l'influence sur le salaire, du fait d'être un homme relativement à une femme, *toutes choses égales par ailleurs*, c'est-à-dire en considérant que l'ancienneté et le nombre d'années d'études de l'individu sont fixés et constants. Si par exemple, β_4 est positif, cela signifie que le fait d'être un homme, procure un bonus salarial relativement à la situation d'une femme ayant un niveau d'études et une ancienneté identique. Dans un modèle linéaire, on utilisera des variables indicatrices (ou binaires ou encore dichotomiques) pour chercher à quantifier l'impact, sur la variable dépendante, de la possession, par un individu, d'une caractéristique particulière, relativement à ceux ou celles

qui ne la possèdent pas. Pour étudier l'influence d'une variable qualitative multinomiale sur une variable dépendante quantitative, on créera autant de variables indicatrices qu'il y a de modalités pour la variable qualitative. La procédure consiste à incorporer comme variables explicatives du modèle, toutes les variables indicatrices, **moins une**. Les coefficients des variables incorporées s'interprètent relativement à la catégorie qui a été laissée de côté.

Dans l'équation (1), la variable d'ancienneté a été incorporée sous forme quadratique. C'est une procédure assez courante pour rendre compte d'effets non linéaires dans un modèle linéaire. On peut imaginer que l'ancienneté joue positivement sur le salaire d'un individu. L'effet marginal d'une année supplémentaire d'ancienneté dans l'emploi serait donc positif. Si le modèle estimé est le suivant :

$$W_i = \beta_0 + \beta_1.Ed_i + \beta_2.Anci_i + \beta_3.Hom_i + \varepsilon_i \qquad \forall i \in \{1;n\}$$

l'effet marginal sera mesuré par β_2 et sera constant, quel que soit le nombre d'années d'ancienneté dans l'emploi d'un individu. Une année d'ancienneté supplémentaire procure, toutes choses égales par ailleurs, un gain salarial constant, donné par β_2. Or il semblerait logique d'imaginer que cet effet marginal devrait, tout en restant positif, être décroissant avec l'ancienneté de l'individu. Pour pouvoir modéliser cette non linéarité dans le modèle linéaire multiple, on incorpore l'ancienneté au carré en tant que variable explicative du salaire, en plus de la variable en niveau, comme dans la relation (1). On s'attend à observer $\beta_2 > 0$ et $\beta_3 < 0$. Ainsi l'effet marginal de l'ancienneté sur le salaire est :

$$\frac{\partial W_i}{\partial Anci_i} = \frac{\partial\left(\beta_0 + \beta_1.Ed_i + \beta_2.Anci_i + \beta_3.Anci_i^2 + \beta_4.Hom_i + \varepsilon_i\right)}{\partial Anci_i} = \beta_2 + 2\beta_3.Anci_i$$

Lorsque β_3 est négatif, l'effet marginal décroit avec le niveau d'ancienneté atteint par un individu.

Le modèle Log-Log

Considérons le même modèle que précédemment mais dans lequel la variable dépendante W_i et la variable explicative Ed_i sont exprimées en logarithmes :

$$\ln W_i = \beta_0 + \beta_1.\ln Ed_i + \beta_2.Anci_i + \beta_3.Hom_i + \varepsilon_i \qquad \forall i \in \{1;n\} \qquad (2)$$

avec ε_i le terme d'erreur.

Pour savoir comment interpréter le coefficient β_1, il suffit d'étudier comme précédemment la dérivée partielle de W_i par rapport à Ed_i. Pour ceci, nous pouvons réécrire l'équation (2) en réalisant une transformation exponentielle :

$$W_i = e^{\beta_0+\beta_1.\ln Ed_i+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}$$

$$\Leftrightarrow \quad W_i = e^{\beta_1.\ln Ed_i}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}$$

$$\Leftrightarrow \quad W_i = Ed_i^{\beta_1}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}$$

Nous pouvons maintenant dériver W_i par rapport à Ed_i :

$$\frac{\partial W_i}{\partial Ed_i} = \frac{\partial\left(Ed_i^{\beta_1}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}\right)}{\partial Ed_i}$$

$$\Leftrightarrow \quad \frac{\partial W_i}{\partial Ed_i} = \beta_1.Ed_i^{\beta_1-1}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}$$

$$\Leftrightarrow \quad \frac{\partial W_i}{\partial Ed_i} = \beta_1\frac{Ed_i^{\beta_1}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}}{Ed_i}$$

$$\Leftrightarrow \quad \frac{\partial W_i}{\partial Ed_i} = \beta_1\frac{W_i}{Ed_i}$$

En isolant β_1, on obtient :

$$\beta_1 = \frac{\frac{\partial W_i}{\partial Ed_i}}{\frac{W_i}{Ed_i}} = \frac{\partial W_i}{W_i}.\frac{Ed_i}{\partial Ed_i} = \frac{\frac{\partial W_i}{W_i}}{\frac{\partial Ed_i}{Ed_i}}$$

On reconnait bien ici une élasticité. Elle s'interprète comme le changement de β_1% du salaire individuel induit par une variation du nombre d'années d'études de 1%.

Le modèle Log-niveau

Considérons le même modèle avec maintenant la variable dépendante W_i en logarithme :

$$\ln W_i = \beta_0 + \beta_1.Ed_i + \beta_2.Anci_i + \beta_3.Hom_i + \varepsilon_i \qquad \forall i \in \{1;n\} \qquad (3)$$

avec ε_i le terme d'erreur.

De la même manière que dans le cas précédent, on réalise une transformation exponentielle du modèle (3) :

$$W_i = e^{\beta_0+\beta_1.Ed_i+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}$$

$$\Leftrightarrow \quad W_i = e^{\beta_1.Ed_i}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}$$

Calculons la dérivée partielle de W_i par rapport à Ed_i :

$$\frac{\partial W_i}{\partial Ed_i} = \frac{\partial\left(e^{\beta_1.Ed_i}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}\right)}{\partial Ed_i}$$

$$\Leftrightarrow \quad \frac{\partial W_i}{\partial Ed_i} = \beta_1.e^{\beta_1.Ed_i}.e^{\beta_0+\beta_2.Anci_i+\beta_3.Hom_i+\varepsilon_i}$$

$$\Leftrightarrow \quad \frac{\partial W_i}{\partial Ed_i} = \beta_1.W_i$$

On peut donc isoler β_1 :

$$\beta_1 = \frac{\partial W_i}{\partial Ed_i.W_i}$$

En multipliant par 100 cette dernière équation, on obtient :

$$100 * \beta_1 = \frac{100 * \frac{\partial W_i}{W_i}}{\partial Ed_i} = \frac{\%\Delta W_i}{\partial Ed_i}$$

Ainsi, on peut interpréter $100 * \beta_1$ comme le changement en pourcentage du salaire lorsque le nombre d'années d'études augmente d'une unité, toutes choses égales par ailleurs. En d'autres termes, lorsque le nombre d'années d'études augmente d'une unité, le salaire augmente de $(100 * \beta_1)\%$, à ancienneté et genre fixés.

Le modèle niveau-Log

Considérons enfin le cas où, cette fois, c'est la variable dépendante qui est en niveau et la variable explicative qui est en logarithme :

$$W_i = \beta_0 + \beta_1.\ln Ed_i + \beta_2.Anci_i + \beta_3.Hom_i + \varepsilon_i \qquad \forall i \in \{1;n\} \qquad (4)$$

avec ε_i le terme d'erreur.

Calculons la dérivée partielle de W_i par rapport à Ed_i :

$$\frac{\partial W_i}{\partial Ed_i} = \frac{\beta_1}{Ed_i}$$

$$\Leftrightarrow \quad \beta_1 = \frac{\partial W_i}{\frac{\partial Ed_i}{Ed_i}}$$

Divisons par 100 les deux membres de cette dernière équation :

$$\frac{\beta_1}{100} = \frac{\partial W_i}{\frac{100 * \partial Ed_i}{Ed_i}} = \frac{\partial W_i}{\%\Delta Ed_i}$$

Cette fois-ci, $\frac{\beta_1}{100}$ s'interprète comme le changement en unités du salaire induit par une augmentation de 1% du nombre d'années d'études, toutes choses égales par ailleurs. En d'autres termes, lorsque le nombre d'années d'études augmente de 1%, le salaire correspondant augmente de $\frac{\beta_1}{100}$ unités, à ancienneté et genre fixés.

L'ensemble de ces résultats peuvent être regroupés dans le tableau récapitulatif suivant :

Type de modèle	**Variable dépendante**	**Variable explicative**	**Interprétation du coefficient β_1**
Niveau-Niveau	y	X	$\Delta y = \beta_1 \Delta x$
Niveau-Log	y	Log(x)	$\Delta y = \left(\frac{\beta_1}{100}\right) \% \, \Delta x$
Log-Niveau	Log(y)	x	$\% \, \Delta y = (100\beta_1) \Delta x$
Log-Log	Log(y)	Log(x)	$\% \, \Delta y = \beta_1 \% \, \Delta x$

Stabilité temporelle d'une estimation : recherche de la date de rupture

On dispose de données annuelles concernant la consommation de viande de poulet aux Etats-Unis pour la période allant de 1960 à 1999, soit 40 observations. Dans le but d'expliquer les déterminants de la demande de ce type de viande, nous disposons aussi d'informations concernant le prix du poulet, le revenu disponible et le prix du bœuf sur la même période. Les données sont regroupées dans le tableau ci-dessous :

Group: UNTITLED Workfile: CHICKEN::Chick6\

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transp

obs	D_POULET	P_BOEUF	P_POULET	Y_DISP
1960	23.52000	20.40000	12.20000	20.22000
1961	25.95000	20.20000	10.10000	20.78000
1962	25.94000	21.30000	10.20000	21.71000
1963	27.22000	19.90000	10.00000	22.46000
1964	27.82000	18.00000	9.200000	24.10000
1965	29.77000	19.90000	8.900000	25.63000
1966	32.08000	22.20000	9.700000	27.34000
1967	32.62000	22.30000	7.900000	28.95000
1968	32.88000	23.40000	8.200000	31.14000
1969	34.90000	26.20000	9.700000	33.24000
1970	36.88000	27.10000	9.100000	35.87000
1971	36.74000	29.00000	7.700000	38.60000
1972	38.49000	33.50000	9.000000	41.40000
1973	37.01000	42.80000	15.10000	46.16000
1974	36.93000	35.60000	9.700000	50.10000
1975	36.70000	32.30000	9.900000	54.98000
1976	39.84000	33.70000	12.90000	59.72000
1977	40.71000	34.50000	12.00000	65.17000
1978	43.10000	48.50000	12.40000	72.24000
1979	46.64000	66.10000	13.90000	79.67000
1980	46.91000	62.40000	11.00000	88.22000
1981	48.45000	58.60000	11.10000	97.65000
1982	49.52000	56.70000	10.30000	104.2600
1983	50.83000	55.50000	12.70000	111.3100
1984	52.83000	57.30000	15.90000	123.1900
1985	54.81000	53.70000	14.80000	130.3700
1986	56.47000	52.60000	12.50000	136.4900
1987	60.27000	61.10000	11.00000	142.4100
1988	62.28000	66.60000	9.200000	152.9700
1989	66.17000	69.50000	14.90000	162.5700
1990	69.08000	74.60000	9.300000	171.3100
1991	72.12000	72.70000	7.100000	176.0900
1992	75.38000	71.30000	8.600000	184.9400
1993	77.14000	72.60000	10.00000	188.7200
1994	78.61000	66.70000	7.400000	195.5500
1995	78.23000	61.80000	6.500000	202.8700
1996	81.42000	58.70000	6.600000	210.9100
1997	83.67000	63.10000	7.700000	219.4000
1998	83.89000	59.60000	8.100000	231.6100
1999	88.87000	63.40000	7.100000	239.6800

Nous allons chercher à estimer la spécification linéaire suivante :

$$D_{Poulet_t} = \beta_0 + \beta_1 . P_{Poulet_t} + \beta_2 . P_{boeuf_t} + \beta_3 . R_{disp_t} + \varepsilon_t \qquad (1)$$

La demande de poulet dépend :

- Du prix du poulet : on attend ici une influence négative du prix sur la demande ($\beta_1 < 0$) ;
- Du prix du bœuf : on attend ici une influence positive du prix sur la demande si le poulet et le bœuf sont des biens substituables ($\beta_2 > 0$) ;
- Du revenu disponible : on attend ici une influence positive du revenu sur la demande ($\beta_3 > 0$).

L'estimation par MCO de la relation (1) fournit les résultats suivants (estimation réalisée avec le logiciel Eviews) :

Equation: EQCONT Workfile: CHICKEN-AUNÈGE::Chick6\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: D_POULET
Method: Least Squares
Date: 07/18/23 Time: 11:37
Sample: 1960 1999
Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	27.59394	1.584458	17.41539	0.0000
P_POULET	-0.607160	0.157120	-3.864300	0.0004
P_BOEUF	0.092188	0.039883	2.311452	0.0266
Y_DISP	0.244860	0.011095	22.06862	0.0000

R-squared	0.990391	Mean dependent var	50.56725
Adjusted R-squared	0.989590	S.D. dependent var	19.53879
S.E. of regression	1.993549	Akaike info criterion	4.312350
Sum squared resid	143.0726	Schwarz criterion	4.481238
Log likelihood	-82.24700	Hannan-Quinn criter.	4.373414
F-statistic	1236.776	Durbin-Watson stat	0.897776
Prob(F-statistic)	0.000000		

Comme nous pouvons le voir, les résultats de cette estimation sont bons et conformes à ce que l'on pouvait attendre. Le modèle explique environ 99% de la variance de la variable dépendante ($R^2 = 0.99$) et la statistique de Fisher prend une valeur très élevée ($F \approx 1237$). Les coefficients sont tous significatifs au seuil de 5% (les *p-values* sont toutes inférieures à 0.05) et les signes sont cohérents avec ce que la théorie économique prédisait. De plus, on lit les informations suivantes :

$$SCR = 143.07 \text{ et } \hat{\sigma} = 1.9935$$

On vérifie bien que $\hat{\sigma} = \sqrt{\frac{SCR}{n-k}} = \sqrt{\frac{143.07}{40-4}} = \sqrt{3.9742} = 1.9935$

Une fois ces constatations faites, on s'interroge maintenant sur la stabilité temporelle de cette relation. On sait que pour statuer sur ce point, on peut utiliser un test de Chow. Il reste néanmoins un problème à résoudre : lorsque on dispose de données en coupes transversales, les groupes sont parfaitement déterminés *ex ante*. On connaît en effet la répartition de l'échantillon initial par groupes de pays, par secteurs d'activité, par classes d'âges, par genre, etc.

Quand on travaille sur des séries chronologiques, comme c'est le cas ici, les groupes correspondent à un découpage en périodes de l'échantillon initial. Or il y a de nombreuses façons de découper la période 1960-1999 en deux sous périodes. Dans la mesure où notre modèle comporte 4 paramètres à estimer, on doit obligatoirement disposer d'au moins 5 observations pour réaliser les estimations non contraintes, ce qui conduit à envisager une date de rupture comprise entre 1964 et 1994, soit 31 dates possibles pour former les deux groupes nécessaires à la mise en œuvre d'un test de Chow.

Choisissons par exemple la date de 1983 pour scinder l'échantillon total en deux. Les résultats des régressions sur les deux sous-périodes (1960-1982 et 1983-1999) sont donnés ci-dessous :

Equation: EQUA1 Workfile: CHICKEN-AUNÈGE::Chick6\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: D_POULET
Method: Least Squares
Date: 07/18/23 Time: 11:48
Sample: 1960 1982
Included observations: 23

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	26.71063	2.738750	9.752855	0.0000
P_POULET	-0.568712	0.311115	-1.827981	0.0833
P_BOEUF	0.182667	0.101228	1.804511	0.0870
Y_DISP	0.194002	0.054628	3.551293	0.0021

R-squared	0.922225	Mean dependent var	36.11391
Adjusted R-squared	0.909945	S.D. dependent var	7.524126
S.E. of regression	2.257931	Akaike info criterion	4.623546
Sum squared resid	96.86678	Schwarz criterion	4.821023
Log likelihood	-49.17077	Hannan-Quinn criter.	4.673211
F-statistic	75.09814	Durbin-Watson stat	0.492583
Prob(F-statistic)	0.000000		

Equation: EQUA2 Workfile: CHICKEN-AUNÈGE::Chick6\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: D_POULET
Method: Least Squares
Date: 07/18/23 Time: 11:48
Sample: 1983 1999
Included observations: 17

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12.84721	5.242387	2.450641	0.0292
P_POULET	-0.235820	0.175794	-1.341456	0.2027
P_BOEUF	0.152610	0.050101	3.046021	0.0094
Y_DISP	0.284753	0.013845	20.56747	0.0000

R-squared	0.990777	Mean dependent var	70.12176
Adjusted R-squared	0.988648	S.D. dependent var	12.08875
S.E. of regression	1.287978	Akaike info criterion	3.546349
Sum squared resid	21.56555	Schwarz criterion	3.742399
Log likelihood	-26.14397	Hannan-Quinn criter.	3.565837
F-statistic	465.5009	Durbin-Watson stat	1.593414
Prob(F-statistic)	0.000000		

On en retire les informations suivantes :

$$\begin{cases} SCR_1 = 96.87 & dl_1 = 23 - 4 = 19 \\ SCR_2 = 21.56 & dl_2 = 17 - 4 = 13 \end{cases}$$

On peut à présent construire la statistique permettant de prendre une décision quant à l'existence d'une rupture en 1982 au niveau de la fonction de demande de poulet aux Etats-Unis.

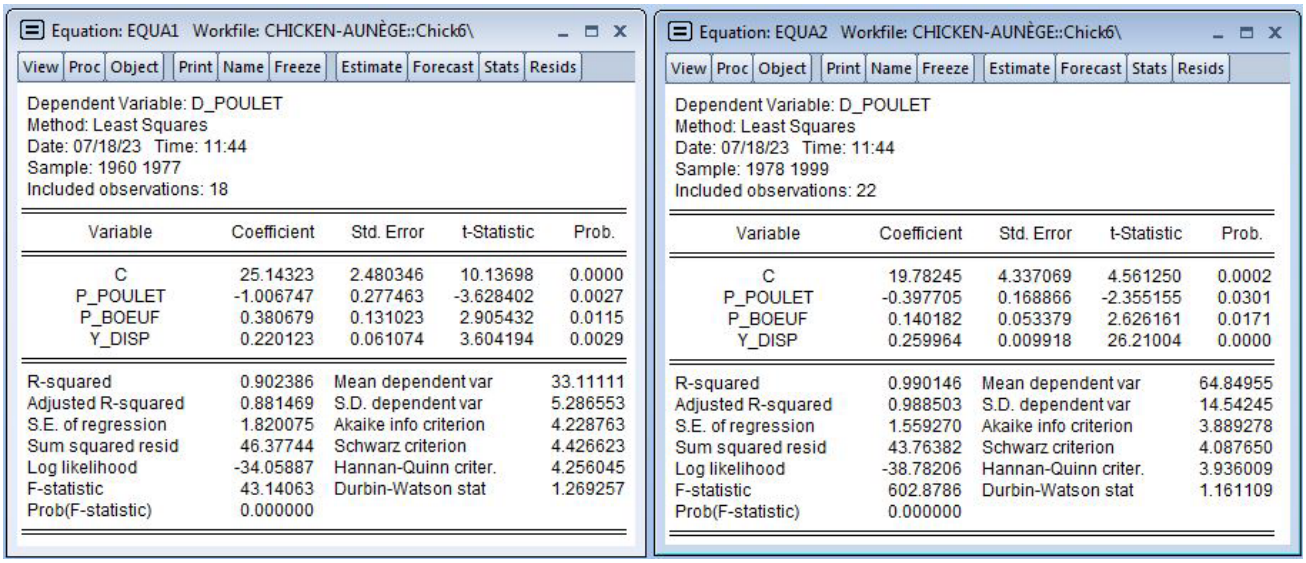
$$f = \frac{\frac{SCR_c - (SCR_1 + SCR_2)}{k}}{\frac{(SCR_1 + SCR_2)}{n - 2k}} = \frac{\frac{143.07 - (96.87 + 21.56)}{4}}{\frac{96.87 + 21.56}{32}} = 1.66$$

Il reste à comparer la valeur calculée de la statistique *f* avec la valeur critique d'une loi de Fisher à 4 et 32 degrés de liberté.

Pour un risque de première espèce égal à 5%, la table nous donne : $f^*_{0,05} = 2.67$

Puisque $f < f^*_{0,05}$, on ne rejette pas H_0 et on conclut donc que les coefficients de l'équation sont stables entre les deux sous-périodes.

Choisissons maintenant une autre date de rupture, 1978 par exemple. On peut à nouveau estimer le modèle (1) sur les deux sous-périodes (1960-1977 et 1978-1999). Les résultats sont donnés ci-dessous :



Equation: EQUA1 Workfile: CHICKEN-AUNÈGE::Chick6\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: D_POULET
Method: Least Squares
Date: 07/18/23 Time: 11:44
Sample: 1960 1977
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	25.14323	2.480346	10.13698	0.0000
P_POULET	-1.006747	0.277463	-3.628402	0.0027
P_BOEUF	0.380679	0.131023	2.905432	0.0115
Y_DISP	0.220123	0.061074	3.604194	0.0029

R-squared	0.902386	Mean dependent var	33.11111
Adjusted R-squared	0.881469	S.D. dependent var	5.286553
S.E. of regression	1.820075	Akaike info criterion	4.228763
Sum squared resid	46.37744	Schwarz criterion	4.426623
Log likelihood	-34.05887	Hannan-Quinn criter.	4.256045
F-statistic	43.14063	Durbin-Watson stat	1.269257
Prob(F-statistic)	0.000000		

Equation: EQUA2 Workfile: CHICKEN-AUNÈGE::Chick6\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: D_POULET
Method: Least Squares
Date: 07/18/23 Time: 11:44
Sample: 1978 1999
Included observations: 22

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	19.78245	4.337069	4.561250	0.0002
P_POULET	-0.397705	0.168866	-2.355155	0.0301
P_BOEUF	0.140182	0.053379	2.626161	0.0171
Y_DISP	0.259964	0.009918	26.21004	0.0000

R-squared	0.990146	Mean dependent var	64.84955
Adjusted R-squared	0.988503	S.D. dependent var	14.54245
S.E. of regression	1.559270	Akaike info criterion	3.889278
Sum squared resid	43.76382	Schwarz criterion	4.087650
Log likelihood	-38.78206	Hannan-Quinn criter.	3.936009
F-statistic	602.8786	Durbin-Watson stat	1.161109
Prob(F-statistic)	0.000000		

On en retire les informations suivantes :

$$\begin{cases} SCR_1 = 46.38 & dl_1 = 18 - 4 = 14 \\ SCR_2 = 43.76 & dl_2 = 22 - 4 = 18 \end{cases}$$

Recalculons la statistique *f* :

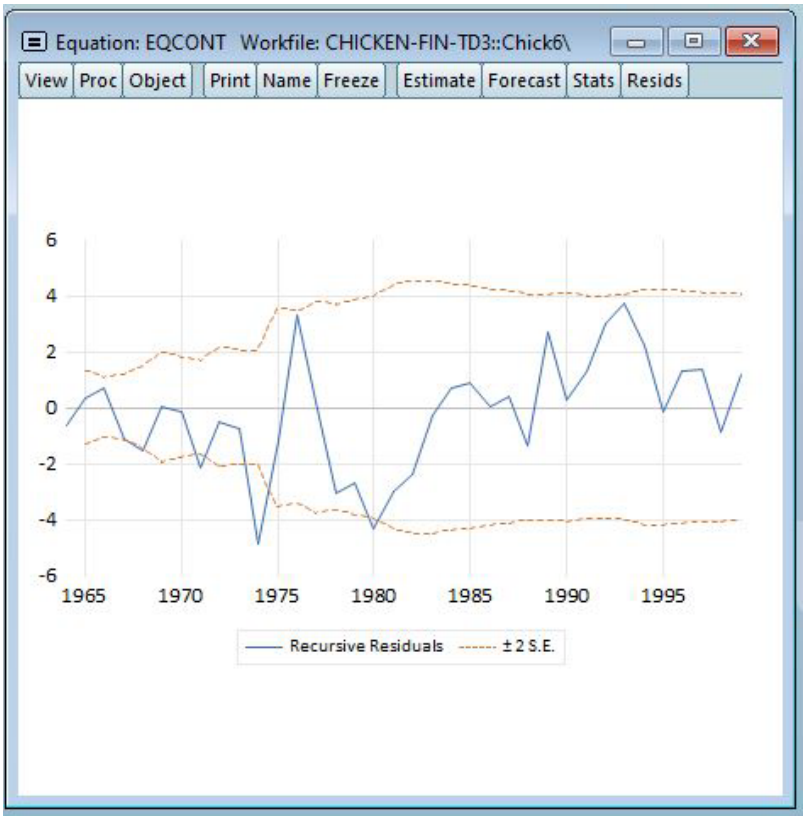
$$f = \frac{\frac{SCR_c - (SCR_1 + SCR_2)}{k}}{\frac{(SCR_1 + SCR_2)}{n - 2k}} = \frac{\frac{143.07 - (46.38 + 43.76)}{4}}{\frac{46.38 + 43.76}{32}} = 4.7$$

La valeur critique de la loi de Fisher vaut toujours $f^*_{0,05} = 2.67$ et donc puisque $f > f^*_{0,05}$, on rejette l'hypothèse nulle et on conclut que les coefficients se sont significativement modifiés entre les deux sous périodes.

On mesure, avec cet exemple, le caractère peu convaincant de la conclusion du test de Chow qui est intimement liée au choix de la date de rupture devant scinder l'échantillon initial en deux sous-échantillons.

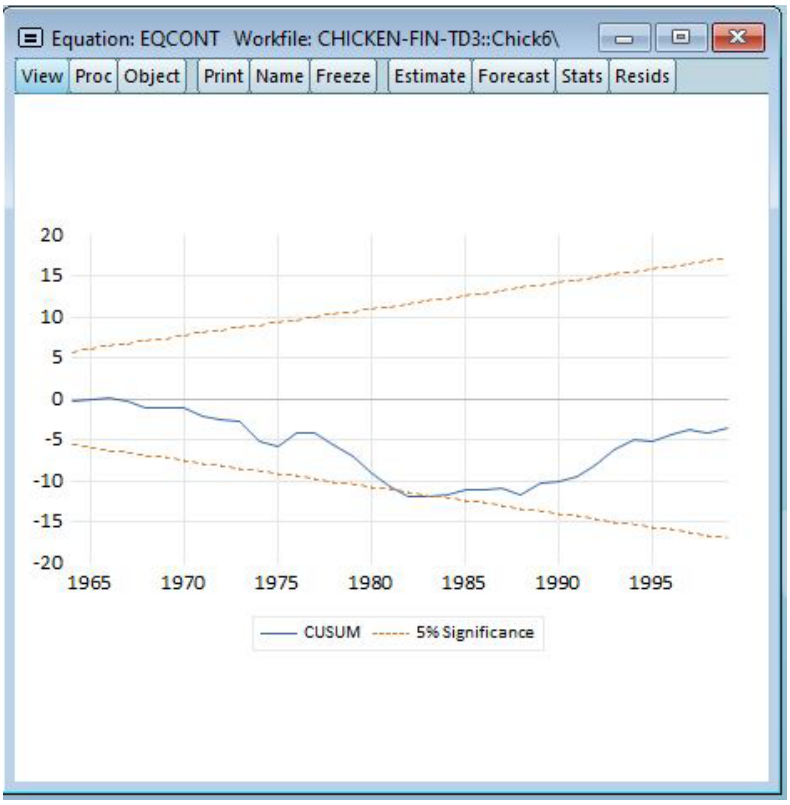
Plusieurs tests ont ainsi été développés pour tenter de déterminer une éventuelle date de rupture dans le cas des séries temporelles[2]. Ces tests utilisent la méthode récursive qui consiste à estimer le modèle en prenant au départ $(k + 1)$ observations puis en rajoutant, avant chaque nouvelle estimation, une observation supplémentaire jusqu'à arriver à l'échantillon complet. Pour chaque estimation, on calcule les résidus estimés, la somme des résidus estimés (CUSUM), la somme des carrés des résidus estimés (CUSUMSQ). Si le modèle est stable, ces grandeurs doivent rester à l'intérieur d'un certain intervalle de confiance, représenté sur les graphiques par deux courbes pointillées rouges.

Concernant notre échantillon de données, voici le graphique des résidus récursifs :

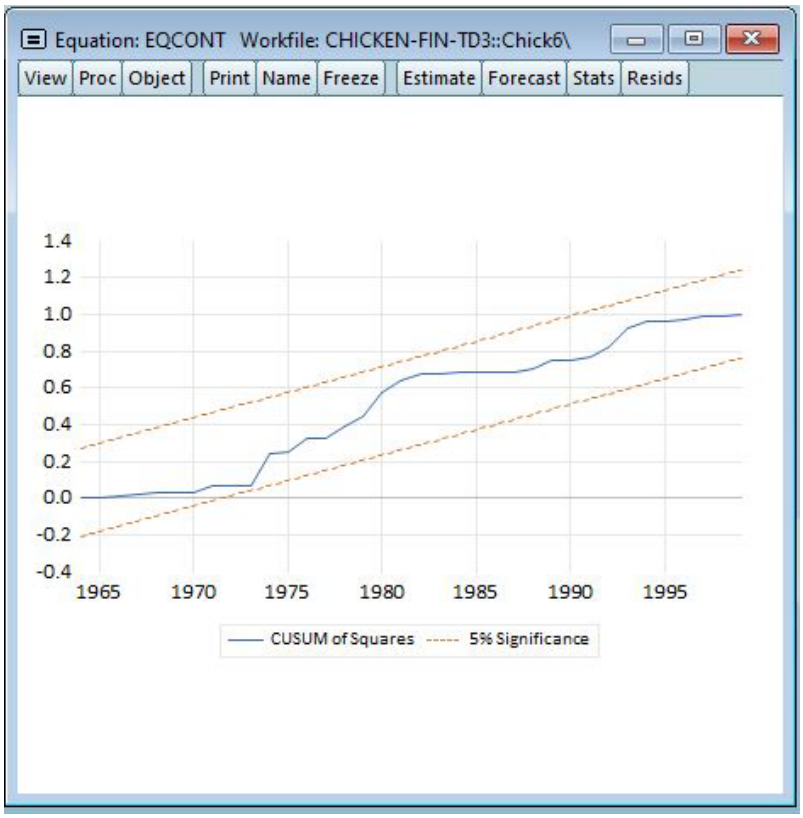


[2] Pour une présentation détaillée de ces tests voir W.H. Green, Econometric Analysis, pp 355-359.

Voici le graphique de la somme des résidus récursifs :

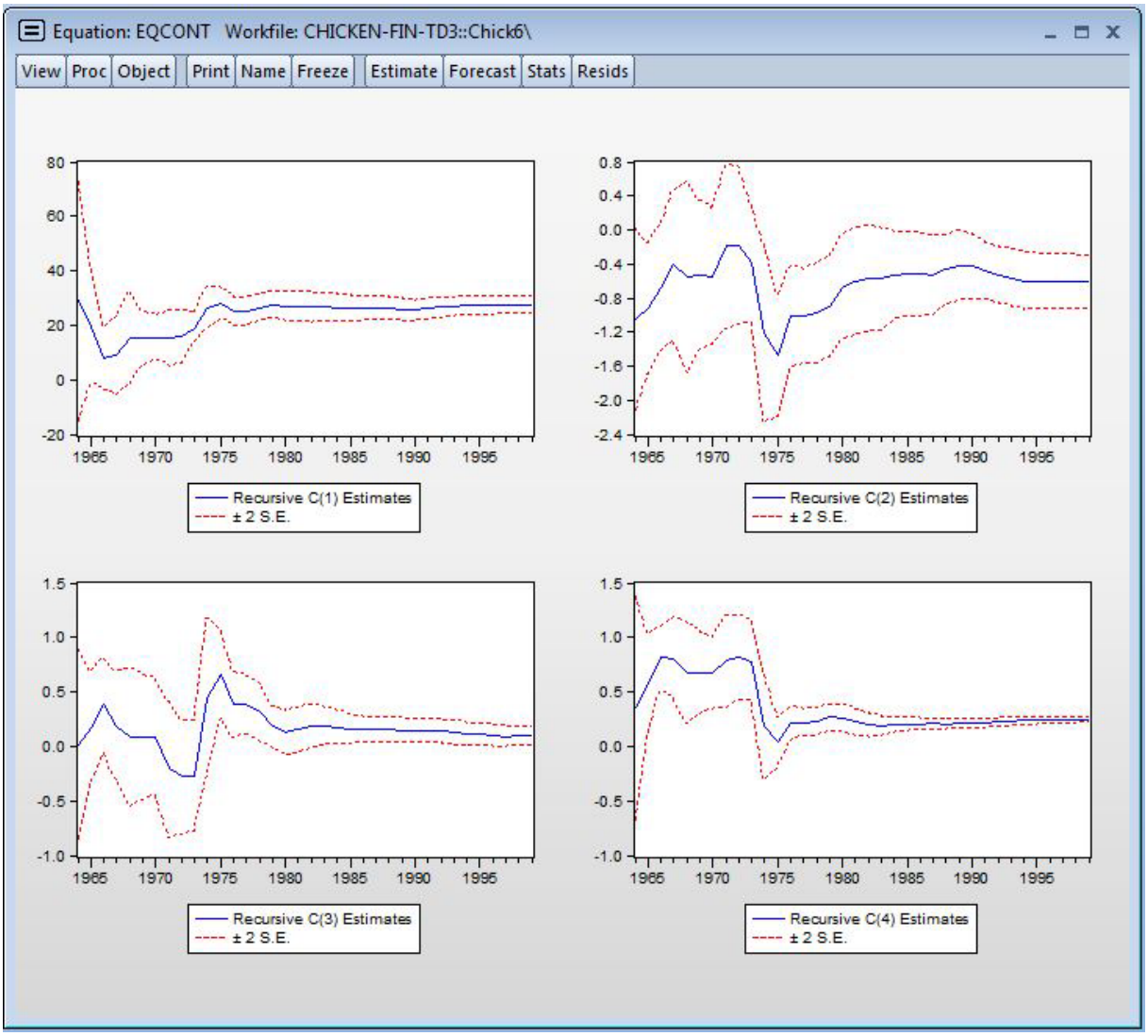


Voici le graphique de la somme des carrés des résidus récursifs :



Enfin, on peut aussi observer l'évolution des paramètres estimés à mesure que l'on rajoute une à une les observations. Si, pour une certaine date, on observe une forte variation de ces coefficients, cela peut nous renseigner sur une éventuelle date de rupture.

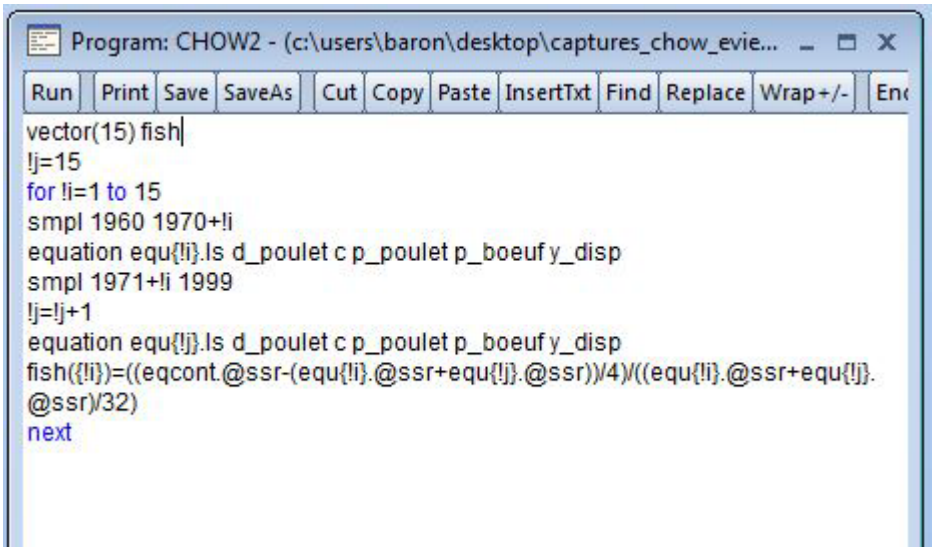
Ci-dessous, les graphiques correspondant aux 4 coefficients estimés du modèle (1) :



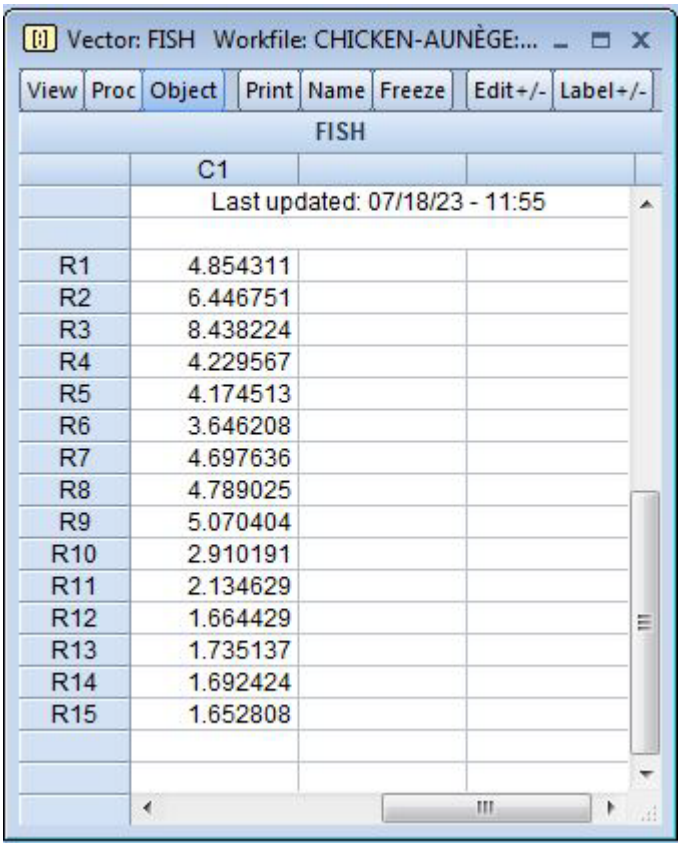
Plusieurs de ces tests suggèrent une date de rupture située entre les années 1973 et 1982.

Pour trouver cette date, nous allons programmer le calcul de la statistique *f* en faisant varier la date de rupture potentielle entre 1971 et 1986. Pour chacune de ces dates, on estime le modèle sur les deux sous-périodes et on calcule la statistique *f*. La date pour laquelle la statistique *f* est la plus élevée sera retenue comme date de rupture. En effet, puisque la valeur critique ne change pas quelle que soit la date choisie, plus grande est valeur calculée de *f*, plus les différences entre les deux sous-échantillons seront significatives.

Le programme permettant de réaliser cette suite de calculs est donné ci-dessous pour information. Les valeurs calculées de la statistique sont stockées dans un vecteur nommé fish.

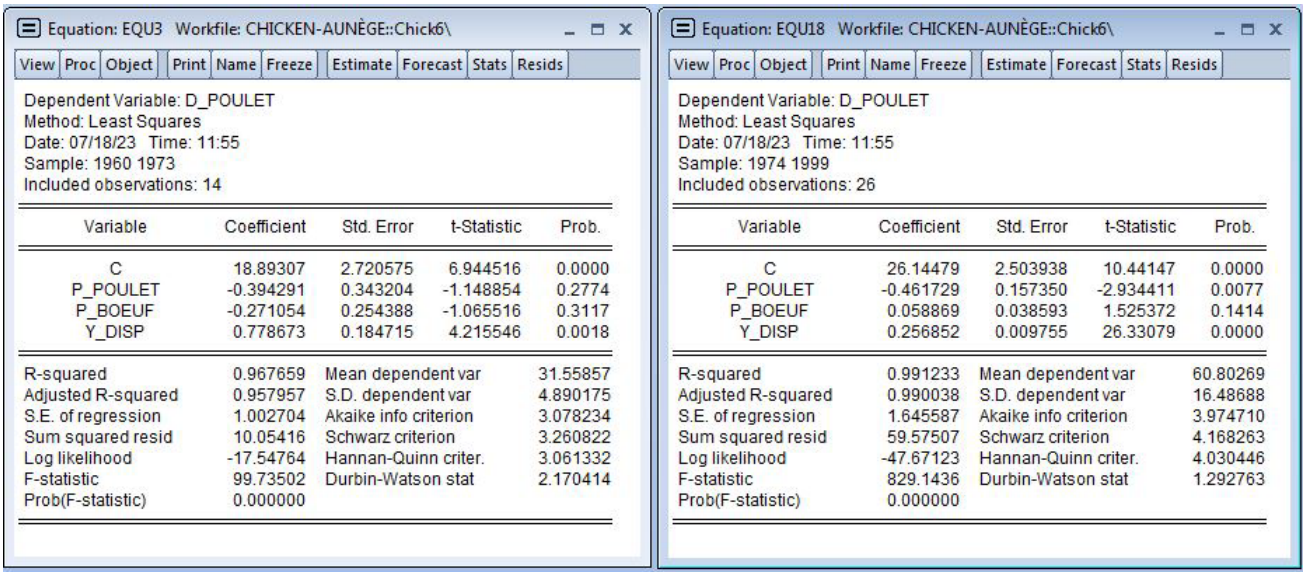


Le vecteur contenant les différentes valeurs de la statistique *f* est :



En examinant les valeurs numériques composant ce vecteur, on retrouve les deux valeurs calculées initialement en choisissant la date de rupture au hasard : 1.66 pour une date de 1983 (ligne 12 du vecteur Fish) et 4.7 pour une date de 1978 (ligne 7 du vecteur Fish).

On voit immédiatement que la statistique du test de Chow prend sa plus grande valeur pour la troisième date de rupture, soit 1974. C'est donc cette date que nous retiendrons pour estimer les deux parties du modèle non contraint. Les deux estimations sont données ci-dessous :



Equation: EQU3 Workfile: CHICKEN-AUNÈGE::Chick6\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: D_POULET
Method: Least Squares
Date: 07/18/23 Time: 11:55
Sample: 1960 1973
Included observations: 14

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	18.89307	2.720575	6.944516	0.0000
P_POULET	-0.394291	0.343204	-1.148854	0.2774
P_BOEUF	-0.271054	0.254388	-1.065516	0.3117
Y_DISP	0.778673	0.184715	4.215546	0.0018

R-squared	0.967659	Mean dependent var	31.55857
Adjusted R-squared	0.957957	S.D. dependent var	4.890175
S.E. of regression	1.002704	Akaike info criterion	3.078234
Sum squared resid	10.05416	Schwarz criterion	3.260822
Log likelihood	-17.54764	Hannan-Quinn criter.	3.061332
F-statistic	99.73502	Durbin-Watson stat	2.170414
Prob(F-statistic)	0.000000		

Equation: EQU18 Workfile: CHICKEN-AUNÈGE::Chick6\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: D_POULET
Method: Least Squares
Date: 07/18/23 Time: 11:55
Sample: 1974 1999
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	26.14479	2.503938	10.44147	0.0000
P_POULET	-0.461729	0.157350	-2.934411	0.0077
P_BOEUF	0.058869	0.038593	1.525372	0.1414
Y_DISP	0.256852	0.009755	26.33079	0.0000

R-squared	0.991233	Mean dependent var	60.80269
Adjusted R-squared	0.990038	S.D. dependent var	16.48688
S.E. of regression	1.645587	Akaike info criterion	3.974710
Sum squared resid	59.57507	Schwarz criterion	4.168263
Log likelihood	-47.67123	Hannan-Quinn criter.	4.030446
F-statistic	829.1436	Durbin-Watson stat	1.292763
Prob(F-statistic)	0.000000		

On en retire les informations suivantes :

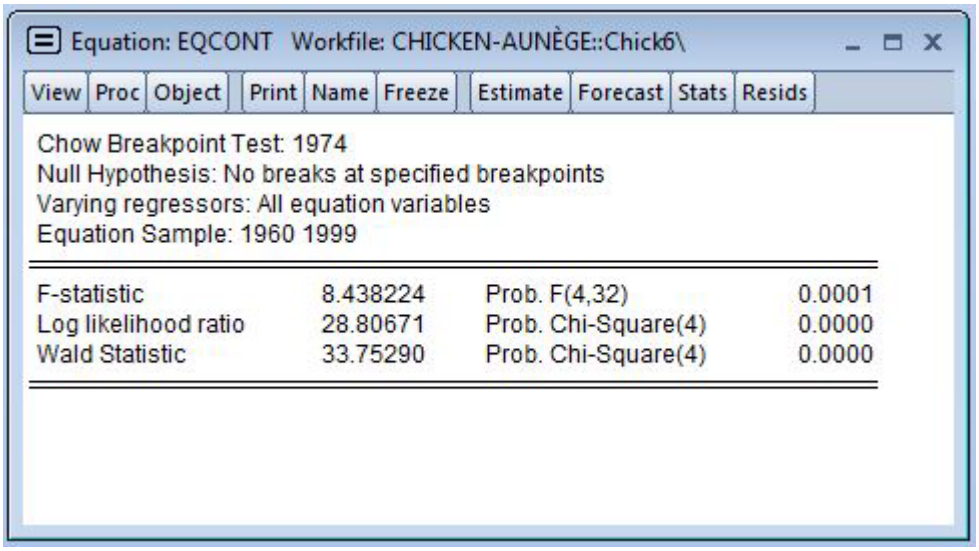
$$\begin{cases} SCR_1 = 10.05 & dl_1 = 14 - 4 = 10 \\ SCR_2 = 59.58 & dl_2 = 26 - 4 = 22 \end{cases}$$

On peut maintenant retrouver la valeur de la statistique *f* :

$$f = \frac{\frac{SCR_c - (SCR_1 + SCR_2)}{k}}{\frac{(SCR_1 + SCR_2)}{n - 2k}} = \frac{\frac{143.07 - (10.05 + 59.58)}{4}}{\frac{10.05 + 59.58}{32}} = 8.44$$

On rejette bien entendu l'hypothèse nulle de stabilité des coefficients de l'équation (1).

Connaissant la date rupture, le logiciel permet bien entendu de réaliser automatiquement ce test. Les résultats sont donnés ci-dessous :



On voit que le commentaire initial effectué à partir de l'estimation sur la totalité de la période, est fortement remis en question lorsqu'on estime le modèle sur les deux sous périodes. Sur la première période (1960-1973) la seule variable influant positivement sur la demande de poulet est le revenu disponible. Les prix du poulet et du bœuf n'ont pas d'influence significative sur la demande (les *p-values* des tests de Student sont respectivement de 0.28 et 0.31, soit largement supérieures au seuil habituel de 5%).

Sur la seconde période (1974-1999) le revenu disponible est toujours une variable explicative significative de la demande mais avec une intensité bien moindre (le coefficient estimé correspondant est divisé par 3) et le prix du poulet retrouve l'influence négative qu'on avait imaginé. Par contre, le prix du bœuf n'a toujours pas d'influence significative sur la demande, ce qui remet en cause l'hypothèse de substituabilité entre ces deux aliments.

Références

Comment citer ce cours ?

Introduction à l'économétrie, Olivier Baron, AUNEGe (http://aunege.fr), CC – BY NC ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).