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.

Activités

Attention : ceci est la version corrigée de l'activité.

Exercice

On cherche à étudier la relation entre le taux de mortalité (TM), les dépenses de santé (DS), le pourcentage de plus de 65 ans (POP) et la densité des médecins (DM) au cours des 26 dernières années. La spécification que nous cherchons à estimer a donc la forme linéaire suivante :

$$TM_t = \beta_0 + \beta_1 DS_t + \beta_2 POP_t + \beta_3 DM_t + u_t$$

Les résultats de l'estimation sont donnés dans la matrice et l'équation suivante :

$$\widehat{TM}_t = 21 - 0{,}24\, DS_t - 0{,}68\, POP_t - 2{,}7\, DM_t \qquad \text{(A)}$$

$$(X'X)^{-1} = \begin{pmatrix} 48 & -3 & 1 & -0{,}1 \\ -3 & 0{,}05 & -0{,}01 & -2 \\ 1 & -0{,}01 & 3 & -0{,}02 \\ -0{,}1 & -2 & -0{,}02 & 4 \end{pmatrix}$$

Consigne

PARTIE 1 :

1. Existe-t-il une influence d'au moins un des facteurs ?

Calculez les valeurs manquantes des tableaux suivants afin de répondre à cette question :

Variables	$\widehat{\beta}_j$	$\widehat{\sigma}(\widehat{\beta}_j)$	$t_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}(\widehat{\beta}_j)}$	Prob > |t|
Constante	21	…	…	…
DS	-0.34	…	…	…
POP	-0.68	…	…	…
DM	-3.7	…	…	…

Nombre d'observations	…	SCT	…
SCR	…	R^2	…
Variance estimée du terme d'erreur	0.34	$\bar{R}^2$	…
Ecart-type de la variable dépendante	0.98	F	…
SCE	…	Prob > F	…

2. Peut-on affirmer que le coefficient de la densité des médecins est dix fois plus élevé que celui des dépenses de santé ?

3. Nous cherchons à savoir si l'ajout des variables explicatives POP et DM améliore significativement la qualité de l'estimation par rapport à DS seul.

Pour cela, on régresse le taux de mortalité uniquement sur la variable DS. La spécification à estimer est donc :

$$\boldsymbol{TM_t = \beta_0 + \beta_1 DS_t + v_t}$$

Le nouveau modèle obtenu est :

$$\widehat{TM_t} = 25 - 0,65 DS_t \qquad \text{(B)}$$

avec $\widehat{\sigma}_v = 0.676$ et $R^2 = 0.56$

Quel modèle est le plus adapté ?

PARTIE 2 : Un économiste suggère qu'il y a eu un changement structurel en 1998. Nous vérifions cette dernière théorie en effectuant les régressions sur deux sous-périodes, la première de 1985 à 1997, la seconde de 1998 à 2010.

Les résultats sont présentés dans les tableaux ci-dessous :

1. Compléter les cases vides des tableaux suivants :

Variables	$\widehat{\beta}_j$	$\widehat{\sigma}(\widehat{\beta}_j)$	$t_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}(\widehat{\beta}_j)}$	Prob > |t|
Constante	22	1.76	…	…
DS	-0.31	0.04	…	…
POP	-0.82	0.23	…	…
DM	-1.76	0.59	…	…

1985-1997

Variables	$\widehat{\beta}_j$	$\widehat{\sigma}(\widehat{\beta}_j)$	$t_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}(\widehat{\beta}_j)}$	Prob > |t|
Constante	19	4.48	…	…
DS	-0.27	0.25	…	…
POP	-0.57	0.98	…	…
DM	-3.54	1.02	…	…

1998-2010

Période	1985-1997	1998-2010
Nombre d'observations	…	…
SCR	…	…
Variance estimée du terme d'erreur	0.18	0.31
Ecart-type de la variable dépendante	0.7	0.8
SCE	…	…
SCT	…	…
$\boldsymbol{R^2}$	…	…
$\boldsymbol{\bar{R}^2}$	…	…
F	…	…
Prob > F	…	…

2. Le modèle à trois variables explicatives (modèle (A)) est-il stable sur l'ensemble de la période ou est-il préférable de procéder à deux estimations, l'une allant de 1985 à 1997, l'autre de la période 1998 à 2010 ?

Correction

PARTIE 1 :

1. Pour remplir le premier tableau il faut calculer les écart-types estimés des coefficients estimés.

On sait que :

$\hat{\sigma}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2.x^{jj}}$ où $\hat{\sigma}^2$ est l'estimateur de la variance de la perturbation et x^{jj} le j[ème] élément de la diagonale principale de la matrice $(X'X)^{-1}$. On nous donne $\hat{\sigma}^2 = 0.2$ ainsi que la matrice $(X'X)^{-1}$. Une fois les écart-types estimés calculés, les statistiques de Student (t_j) se déterminent immédiatement. Dans la dernière colonne du premier tableau, on demande de calculer les *p-values* associées à chaque test de Student. La loi de Student ayant ici 22 degrés de liberté, on rappelle que la *p-value* α_j associée à un test de statistique t_j est définie par :

$$\alpha_j = Prob\{\lceil T_{22} \rceil > |t_j|\}$$

Pour calculer ces p-values on utilisera la table statistique de la loi de Student ou la **fonction LOI.STUDENT.BILATERALE** du tableur Excel.

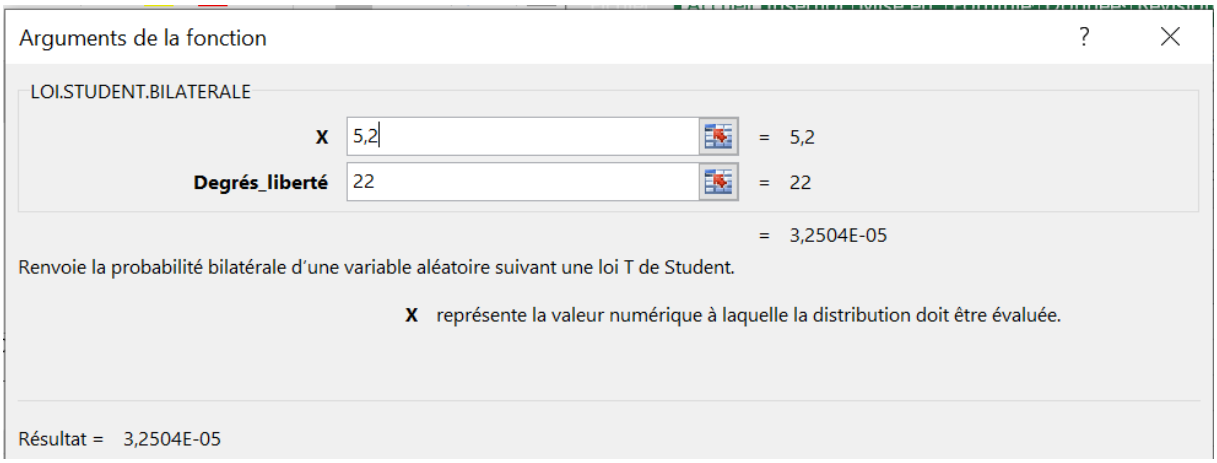
Détaillons les calculs pour la première ligne du premier tableau :

$$\hat{\sigma}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2.x^{11}} = \sqrt{0.34 * 48} = 4.039$$

$$t_0 = \frac{\hat{\beta}_0}{\hat{\sigma}(\hat{\beta}_0)} = \frac{21}{4.039} = 5.2$$

$$\alpha_0 = Prob\{\lceil T_{22} \rceil > 5.2\} = 0.000$$

Ci-dessous la fenêtre de calcul de la p-value avec la fonction Excel :



Les autres lignes du premier tableau se remplissent de la même façon. On obtient :

Variables	$\hat{\beta}_j$	$\hat{\sigma}(\hat{\beta}_j)$	$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$	Prob > |t|
Constante	21	4.039	5.2	0.000
DS	-0.34	0.13	-2.61	0.016
POP	-0.68	1.01	-0.673	0.508
DM	-3.7	1.166	-3.17	0.004

On peut d'ores et déjà voir que les seules variables explicatives influant significativement la variable dépendante sont les dépenses de santé (DS) et la densité des médecins (DM). Ces deux variables jouent négativement sur le taux de mortalité. Lorsque ces deux variables augmentent, le taux de mortalité diminue.

Le second tableau se remplit aussi facilement que le premier.

Le nombre d'observations est donné dans l'énoncé et est égal à 26. Il faut maintenant calculer les quantités SCR, SCE et SCT. Ces trois termes sont les composants de l'équation d'analyse de la variance. SCT (Somme des Carrés Totale) correspond, au facteur $\frac{1}{n}$ près, à la variance empirique de la variable dépendante :

$$SCT = \sum_{t=1}^{n}(y_t - \bar{y})^2 = n * Var(y) = n * \sigma_y^2 = 26 * (0.98)^2 = 24.9704$$

Pour déterminer la quantité SCR (Somme des Carrés des Résidus estimés) on utilise la valeur fournie de $\hat{\sigma}^2$. En effet on sait que $\hat{\sigma}^2 = \frac{SCR}{n-k}$ et donc $SCR = (n-k)\hat{\sigma}^2 = 22 * 0.34 = 7.48$.

Enfin, la quantité SCE (Somme des Carrés Expliquée) est obtenue par différence, compte tenu de l'équation d'analyse de la variance : $SCE = SCT - SCR = 24.9704 - 7.48 = 17.4904$.

On peut maintenant calculer le coefficient de détermination (R^2) et le coefficient de détermination ajusté ($\bar{R}^2$). On a :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = \frac{17.4904}{24.9704} = 0.7$$

Le modèle (A) explique donc 69.9% de la variance de la variable dépendante. Concernant le coefficient ajusté, on sait que :

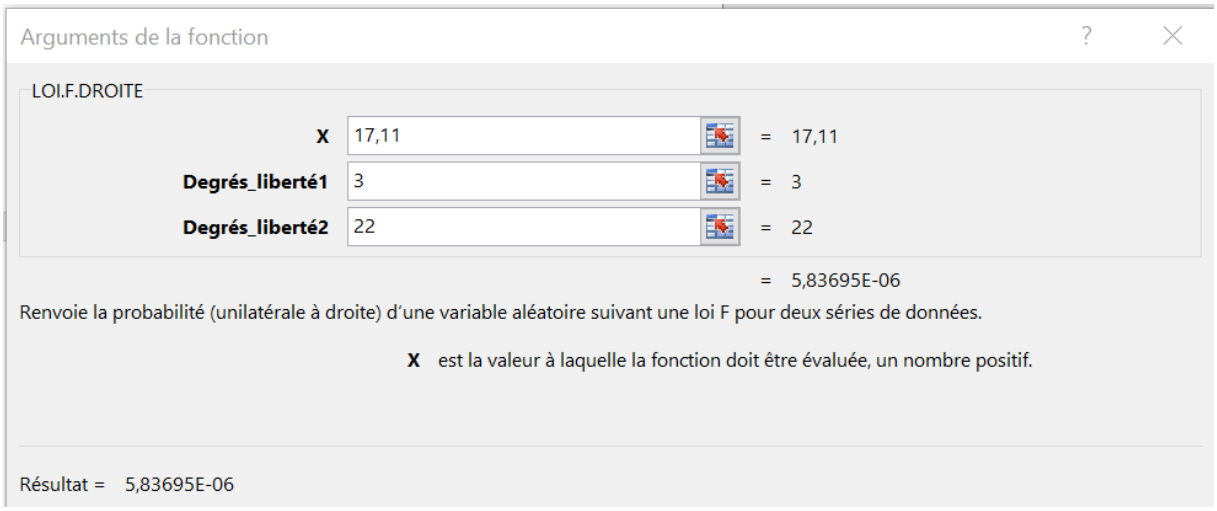
$$\bar{R}^2 = 1 - \frac{\frac{\sum_{t=1}^{n}(\hat{u}_t)^2}{n-k}}{\frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{n-1}} = 1 - \frac{\frac{SCR}{n-k}}{\frac{SCT}{n-1}} = 1 - \frac{\frac{7.48}{22}}{\frac{24.9704}{25}} = 0{,}66$$

Les deux dernières cases du tableau concernent le test de significativité globale de la régression. Il faut calculer la statistique de Fisher, puis la *p-value* associée à ce test. On aura :

$$f = \frac{R_{nc}^2}{1 - R_{nc}^2} \cdot \frac{n-k}{k-1} = \frac{0.7}{1-0.7} * \frac{22}{3} = 17.11$$

Sous l'hypothèse nulle du test, la statistique *f* suit une loi de Fisher de degrés de liberté 3 et 22. Pour déterminer la *p-value* associée à ce test, on utilise une table statistique de la loi de Fisher ou, comme dans le cas des tests de Student, la fonction adéquate du tableur Excel. Cette **fonction se nomme : LOI.F.DROITE**.

La fenêtre de calcul de la p-value avec la fonction Excel est :



La *p-value* du test de significativité globale de l'estimation est nulle. On rejette donc H_0 et on conclut que le modèle est globalement significatif.

L'ensemble de ces résultats permet de remplir complètement le second tableau :

Nombre d'observations	26	SCT	24.97
SCR	7.48	R^2	0.7
Variance estimée du terme d'erreur	0.34	$\bar{R}^2$	0.66
Ecart-type de la variable dépendante	0.98	F	17.11
SCE	17.49	Prob > F	0.000

2. On demande dans cette question si le coefficient de la densité des médecins est dix fois plus élevé que celui des dépenses de santé.

Cette affirmation est vraie si l'on observe les coefficients estimés associés aux deux variables explicatives puisque le coefficient estimé de la variable DS est égal à -0.34 alors que celui de la variable DM est égal à -3.7, soit plus de 10 fois plus.

Par contre, ces informations ne permettent pas de savoir si ce rapport de 1 à 10 se retrouve au niveau des coefficients β_j.

Pour répondre à la question posée nous devons tester l'hypothèse :

$H_0 : \beta_3 = 10\beta_1 \Leftrightarrow \beta_3 - 10\beta_1 = 0$ en utilisant la méthode d'estimation sous contrainte présentée au cours de ce chapitre. La question suivante va nous donner l'occasion d'appliquer cette méthode.

3. On doit tester le modèle contraint (B) contre le modèle non contraint (A).

Le modèle contraint correspond au modèle initial dans lequel on a intégré les deux contraintes :

$$\beta_2 = \beta_3 = 0.$$

Pour prendre une décision, la méthode consiste à calculer la statistique *f*, soit à partir de son expression en fonction des sommes des carrés des résidus estimés (21), soit à partir de la relation (22) dépendant des coefficients de détermination des estimations contrainte et non-contrainte. Le plus simple est sans doute d'utiliser cette dernière expression puisque nous connaissons les deux coefficients de détermination. On a en effet calculé $R^2_{nc} = 0.699$ et l'énoncé nous donne $R^2_c = 0.56$. Le test de ces deux contraintes linéaires sur les coefficients de l'équation (A) se présente donc sous la forme suivante :

$$\begin{cases} H_0 : \quad TM_t = \beta_0 + \beta_1 DS_t + v_t \\ contre\ \bar{H} \ : TM_t = \beta_0 + \beta_1 DS_t + \beta_2 POP_t + \beta_3 DM_t + u_t \end{cases}$$

Si à l'issue du test, on rejette l'hypothèse nulle, on considérera que les deux contraintes imposées ne sont pas pertinentes et donc que le meilleur modèle est le modèle non contraint (A). Calculons la statistique *f* :

$$f = \frac{(R^2_{nc} - R^2_c)}{1 - R^2_{nc}} . \frac{n-k}{r} = \frac{0.7 - 0.56}{1 - 0.7} * \frac{22}{2} = 5.13$$

On peut vérifier que l'expression de *f* en fonction des sommes des carrés des résidus (21) donne le même résultat. Il faut pour cela calculer la somme des carrés des résidus estimés du modèle contraint :

$$SCR_c = \big(n - (k-r)\big).\hat{\sigma}^2_v = 24 * 0.676^2 = 10.97$$

On aura donc :

$$f = \frac{\frac{SCR_c - SCR_{nc}}{dl_c - dl_{nc}}}{\frac{SCR_{nc}}{dl_{nc}}} = \frac{\frac{10.97 - 7.48}{2}}{\frac{7.48}{22}} = 5.13$$

Le test est basé sur le résultat suivant :

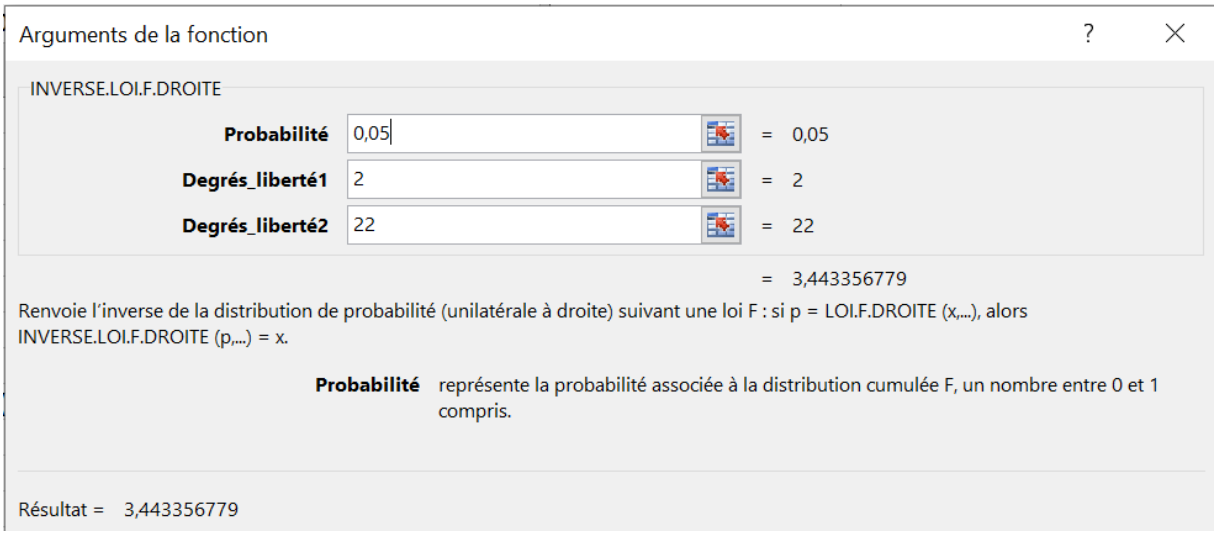
Si H₀ est vraie alors : $f \rightarrow F_{(2,22)}$

Pour un risque de première espèce égal à 5%, la valeur critique de la loi de Fisher est :

$$f^*_{0.05} = 3.44$$

Pour trouver cette valeur critique, on peut utiliser la table de la loi de Fisher ou la **fonction INVERSE.LOI.F.DROITE** du tableur Excel.

La fenêtre de calcul de la valeur critique est donnée ci-dessous :



La valeur calculée de la statistique *f* est supérieure à la valeur critique, ce qui nous conduit à rejeter l'hypothèse nulle. Les contraintes imposées ne sont pas pertinentes et le meilleur modèle est le modèle non contraint (A).

PARTIE 2 :

1. Les tableaux se remplissent en utilisant la même méthode que dans la première partie de l'exercice.

Les résultats sont les suivants :

1985-1997

Variables	$\hat{\beta}_j$	$\hat{\sigma}(\hat{\beta}_j)$	$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$	Prob > |t|
Constante	22	1.76	22.5	0.000
DS	-0.31	0.04	-7.75	0.000
POP	-0.82	0.23	-3.57	0.006
DM	-1.76	0.59	-2.98	0.015

1998-2010

Variables	$\hat{\beta}_j$	$\hat{\sigma}(\hat{\beta}_j)$	$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$	Prob > |t|
Constante	19	4.48	4.24	0.002
DS	-0.27	0.25	-1,08	0.308
POP	-0.57	0.98	-0.58	0.576
DM	-3.54	1.02	-3.47	0.007

Période	1985-1997	1998-2010
Nombre d'observations	13	13
SCR	1.62	2.79
Variance estimée du terme d'erreur	0.18	0.31
Ecart-type de la variable dépendante	0.7	0.8
SCE	4.75	5.53
SCT	6.37	8.32
R^2	0.746	0.665
$\bar{R}^2$	0.661	0.553
F	8.81	5.96
Prob > F	0.005	0.016

2. Pour tester la stabilité temporelle du modèle (A), nous allons utiliser un test de Chow.

Avant ceci, comme il l'a été précisé dans la partie dédiée à ce test, il convient de vérifier que l'on ne se trouve pas dans une situation d'hétéroscédasticité entre blocs. Ce test est une adaptation de la méthode générale présentée par Goldfeld et Quandt.

Test d'homoscédasticité des perturbations :

Les résultats des estimations relatives aux deux périodes permettent de connaître les variances estimées des perturbations pour chacun des deux groupes. On aura :

Période 1 (1985-1997) : $\hat{\sigma}_1^2 = 0.18$

Période 2 (1998-2010) : $\hat{\sigma}_2^2 = 0.31$

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

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

Le test est basé sur le résultat suivant :

Si H0 est vraie alors : $\varphi_2 = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \rightarrow F_{(9,9)}$

Avec un seuil de 5%, il suffit de comparer la valeur calculée de φ_2, soit $\varphi_2 = \frac{0.31}{0.18} = 1.72$ et la valeur critique de la loi de Fisher correspondante, soit $f_{0,05}^* = 3.18$. Puisque $\varphi_2 < f_{0,05}^*$, on ne rejette pas H0 et on considère donc que la variance de la perturbation est la même dans les deux sous-échantillons.

Test de stabilité temporelle :

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

$$\begin{cases} H_0 : & TM_t = \beta_0 + \beta_1 DS_t + \beta_2 POP_t + \beta_3 DM_t + u_t \quad \forall t \in \{1985 - 2010\} \\ contre\ \bar{H} : & \begin{cases} TM_t = \gamma_0 + \gamma_1 DS_t + \gamma_2 POP_t + \gamma_3 DM_t + u_t \quad \forall t \in \{1985 - 1997\} \\ TM_t = \mu_0 + \mu_1 DS_t + \mu_2 POP_t + \mu_3 DM_t + u_t \quad \forall t \in \{1998 - 2010\} \end{cases} \end{cases}$$

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 : $\begin{cases} SCR_c = 7.48 \text{ et } dl_c = 22 \\ SCR_1 = 1.62 \text{ et } dl_1 = 9 \\ SCR_2 = 2.79 \text{ et } dl_2 = 9 \end{cases}$, soit :

$$f = \frac{\frac{SCR_c - (SCR_1 + SCR_2)}{k}}{\frac{(SCR_1 + SCR_2)}{n - 2k}} = \frac{\frac{7.48 - (1.62 + 2.79)}{4}}{\frac{(1.62 + 2.79)}{18}} = \frac{0.7675}{0.245} = 3.13$$

Le test est basé sur le résultat suivant :

Si H₀ est vraie alors : $f \rightarrow F_{(4,18)}$

En notant $f^*_{0,05}$ la valeur critique de cette loi de Fisher pour un risque de première espèce de 5%, la table fournit $f^*_{0,05} = 2.93$.

La valeur calculée de *f* est supérieure à la valeur critique ce qui nous conduit à rejeter l'hypothèse nulle de stabilité temporelle de la régression. Il est donc préférable de procéder à deux estimations, l'une allant de 1985 à 1997, l'autre de 1998 à 2010.

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/).