Chapitre 1 : Le modèle de régression simple

Dans le cadre de l'économétrie, un modèle consiste en une présentation formalisée d'un phénomène sous forme d'équations dont les variables sont des grandeurs économiques. Le modèle est l'outil que le modélisateur utilise lorsqu'il cherche à comprendre et à expliquer des phénomènes.

Dans le modèle de régression simple, une variable endogène est expliquée par une variable exogène[1].

I. Présentation du modèle

A. Exemple introductif

Selon la théorie keynésienne, la consommation des manages est représentée par la relation : C = f (Y) dont la forme est supposée linéaire $C = a_0 + a_1 Y$ dont la série d'observation $(C_t , Y_t)/t \in (1, T)$ est disponible. Le modèle qui est déterministe[2] n'est pas un modèle économétrique. Pour construire la relation qui existe entre le revenu Y et la consommation C , il est supposé que les (C_t , Y_t) correspondent à un échantillon de ménages caractéristique d'une population de ménages beaucoup plus vaste ; c.à.d que (y_t , x_t) est un échantillon dans une distribution à deux dimension, c.à.d que pour une valeur du revenu x déterminée, on observera des valeurs différents de la variable y , correspondante chacune à une observation particulière.

Soit la fonction de consommation keynésienne [3] : $C = a_0 + a_1 Y$

où : C = consommation, Y = revenu, a_1= propension marginale à consommer,

a_0 = consommation autonome ou incompressible.

La variable consommation est appelée « variable à expliquer » ou « variable endogène ». La variable revenu est appelée « variable explicative » ou «

[1]Bourbonnais Régis « économétrie : cours et exercices corrigés», 9eme édition dunod, Paris, 2015.

[2] Les modèles déterministes sont basés sur une loi connue ou hypothétique de la physique, des mathématiques ou d'une quelconque autre discipline, de sorte que des valeurs d'input données produisent toujours le même résultat. Par contre, le modèle stochastique accepte une certaine distribution de probabilité associée à des inputs donnés, dans les processus au sein du modèle et donc dans l'output, de sorte que le même input peut amener à différentes valeur d'output.

[3] Bourbonnais Régis « économétrie : cours et exercices corrigés», 9eme édition dunod, Paris, 2015

variable exogène » (c'est le revenu qui explique la consommation). a_1 et a_0 sont les paramètres du modèle ou encore les coefficients de régression.

Deux types de spécifications peuvent être distingués :

- Les modèles en série temporelle, les variables représentent des phénomènes observés à intervalles de temps réguliers, par exemple la consommation et le revenu annuel sur 20 ans pour un pays donné. Le modèle s'écrit alors :$C_t = a_0+a_1Y_t$ t=1,...,20 où :C_t=consommation au temps t, Y_t = revenu au temps t.
- Les modèles en coupe instantanée, les variables représentent des phénomènes observés au même instant mais concernant plusieurs individus, par exemple la consommation et le revenu observés sur un échantillon de 20 pays. Le modèle s'écrit alors :$C_i = a_0+a_1Y_i$ i =1,...,20 où : C_i=consommation du pays *i* pour une année donnée, Y_i= revenu du pays *i* pour une année donnée.

B. Rôle du terme aléatoire

Le modèle tel qu'il vient d'être spécifié n'est qu'une caricature de la réalité. En effet ne retenir que le revenu pour expliquer la consommation est insuffisant ; il existe une multitude d'autres facteurs susceptibles d'expliquer la consommation. C'est pourquoi il est ajouté un terme (ε_t) qui synthétise l'ensemble de ces informations non explicitées dans le modèle : $C_t = a_0 + a_1Y_t + \varepsilon_t$ si le modèle est spécifié en série temporelle où ε_t représente l'erreur de spécification du modèle, c'est-à-dire l'ensemble des phénomènes explicatifs de la consommation non liés au revenu. Le terme ε_t mesure la différence entre les valeurs réellement observées de C_t et les valeurs qui auraient été observées si la relation spécifiée avait été rigoureusement exacte. Le terme ε_t regroupe donc trois erreurs :

- une erreur de spécification, c'est-à-dire le fait que la seule variable explicative n'est pas suffisante pour rendre compte de la totalité du phénomène expliqué ;
- une erreur de mesure, les données ne représentent pas exactement le phénomène
- une erreur de fluctuation d'échantillonnage, d'un échantillon à l'autre les observations, et donc les estimations, sont légèrement différentes.

Les valeurs vraies a_0 et a_1 ne sont pas connues mais seulement les deux séries d'observations C_t et R_t. Les estimateurs de a_0 et a_1 , notés respectivement $\widehat{a}_0$ et , $\widehat{a}_1$sont des variables aléatoires, qui suivent les mêmes lois de probabilité, celle de ε_t, puisqu'ils sont fonctions de la variable aléatoire ε_t. Les caractéristiques de moyenne et d'écart type de ces coefficients permettent de construire des tests de validité du modèle estimé.

II. Estimation des paramètres

A. Modèle et hypothèses

Soit le modèle suivant : $y_t = a_0 + a_1 x_t + \varepsilon_t$ pour $t = 1, \ldots\ldots, n$ avec :

y_t = variable à expliquer au temps t;
x_t = variable explicative au temps t;
a_0, a_1 = paramètres du modèle ;
ε_t = erreur de spécification (différence entre le modèle vrai et le modèle spécifié), cette erreur est inconnue et restera inconnue ;
n= nombre d'observations.

Hypothèses

• H1 : le modèle est linéaire en x_t (ou en n'importe quelle transformation de x_t).

• H2 : les valeurs x_t sont observées sans erreur (x_t non aléatoire).

• H3 : $E(\varepsilon_t) = 0$, l'espérance mathématique de l'erreur est nulle : en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle. Cette hypothèse signifie que les facteurs secondaires n'ont pas un effet systématique jouant à la hausse ou à la baisse sur la variable y .

• H4 : $E(\varepsilon t^2) = \sigma_\varepsilon^2$, la variance de l'erreur est constante[4]: le risque de l'amplitude de l'erreur est le même quelle que soit la période. Cette hypothèse, souvent appelée la propriété de l'homoscédasticité, traduit l'idée que l'amplitude de la variabilité de l'aléa provenant des facteurs secondaires est invariante à travers les individus ou à travers le temps.[5]

• H5 : $E(\varepsilon_t \varepsilon_{t'}) = 0$ si $t \neq t'$, **les erreurs sont non corrélées** (ou encore indépendantes) : une erreur à l'instant t n'a pas d'influence sur les erreurs suivantes.

• H6 : $Cov(x_t, \varepsilon_t) = 0$, l'erreur est indépendante de la variable explicative.

B. Formulation des estimateurs

En traçant un graphique (1) des couples de données liant le revenu et la consommation observée, nous obtenons un nuage de points que nous pouvons ajuster à l'aide d'une droite. Le principe de base de la méthode des MCO est de choisir parmi toutes les droites possibles celle qui minimise l'écart entre les réalisations de la variable expliquée et les valeurs prévus par le modèle estimé. Mais pour éviter la compensation entre les écarts négatifs et positifs, la minimisation porte sur les erreurs quadratiques comptées parallèlement à l'axe de la variable expliqué.[6] On cherche une droite, d'équation : $y_t = \hat{a}_0 + \hat{a}_1 x_t$ qui

[4] Dans le cas où cette hypothèse n'est pas vérifiée, on parle alors de modèle hétéroscédastique.
[5] Sami Mestiri (2021) Le modèle de régression linéaire simple, Faculté des sciences économiques et de gestion de Mahdia, p.8.
[6] Sami Mestiri (2021), *Op.cit.*, p.9.

approche « au mieux » les données. On l'appelle droite des moindres carrées de y en x ou droite de régression de y en x.

L'estimateur des coefficients a_0 et a_1 est obtenu donc en minimisant la distance au carré entre chaque observation et la droite, d'où le nom d'estimateur des moindres carrés ordinaires (MCO). La résolution analytique est la suivante :

$$Min \sum_{t=1}^{n} \varepsilon_t^{\,2} = Min \sum_{t=1}^{n} (y_t - a_0 - a_1 x_t)^2 = Min\ S$$

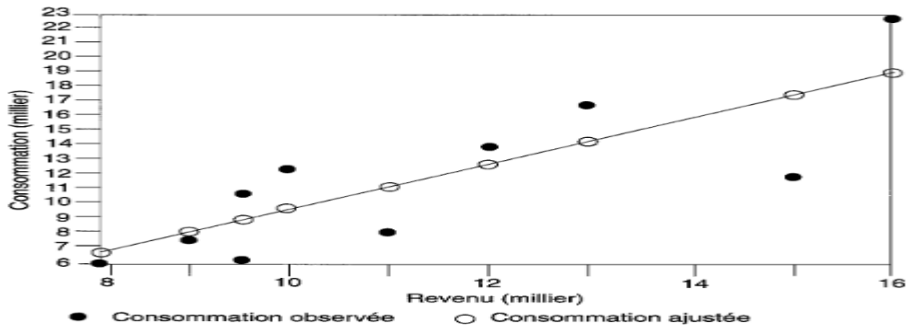
En opérant par dérivation par rapport à a_0 et a_1 afin de trouver le minimum[7] de cette fonction, on obtient les résultats suivants :

$$\frac{\partial S}{\partial a_0} = -2 \sum_t (y_t - \hat{a}_0 - \hat{a}_1 x_t) = 0$$

$$\text{Et } \frac{\partial S}{\partial a_1} = -2 \sum_t x_t (y_t - \hat{a}_0 - \hat{a}_1 x_t) = 0$$

Graphiquement, il s'agira de tracer une droite linéaire qui minimisera l'écart (mis au carré) entre chaque couple (X,Y) et son point correspondant sur la droite de régression.

Graphique 1 – Ajustement d'un nuage de points par une droite



Source : Bourbonnais Régis, « économétrie : cours et exercices corrigés», 9eme édition dunod, Paris, 2015

Sommant par rapport à t, il vient : $\sum_t x_t\ y_t - \hat{a}_0 \sum_t x_t - \hat{a}_1 \sum_t x_t^{\,2} = 0$

$\sum_t y_t - n\,\hat{a}_0 - \hat{a}_1 \sum_t x_t = 0$

Qui est sont appelées les équations normales et qui impliquent que :

$$\hat{a}_1 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}$$

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}$$

C. Les différentes écritures du modèle : erreur et résidu

[7] Les conditions du deuxième ordre sont considérées comme vérifiées car la fonction est convexe.

Le modèle de régression simple peut s'écrire sous deux formes selon qu'il s'agit du modèle théorique spécifié par l'économiste ou du modèle estimé à partir d'un échantillon.

• Modèle théorique spécifié par l'économiste avec ε_t l'erreur inconnue :

$$y_t = a_0 + a_1 x_t + \varepsilon_t$$

• Modèle estimé à partir d'un échantillon d'observations :

$$y_t = \hat{a}_0 + \hat{a}_1 x_t + e_t = \hat{y}_t + e_t \qquad e_t = \text{résidu}$$

Le résidu observé et est donc la différence entre les valeurs observées de la variable à expliquer et les valeurs ajustées à l'aide des estimations des coefficients du modèle ; ou encore : $\hat{y}_t = \hat{a}_0 + \hat{a}_1 x_t$

III. Conséquences des hypothèses : construction des tests

A. Hypothèse de normalité des erreurs

Cette hypothèse n'est pas indispensable afin d'obtenir des estimateurs convergents mais elle permet de construire des tests statistiques concernant la validité du modèle estimé.

B. Conséquences de l'hypothèse de normalité des erreurs

L'estimateur de la variance de l'erreur (σ_ε^2) noté $\hat{\sigma}_\varepsilon^2$ est donc égal à : $\hat{\sigma}^2{}_\varepsilon = \frac{1}{(n-2)} \sum_t e_t^2$

Les estimateurs empiriques de la variance de chacun des coefficients

$$\hat{\sigma}^2{}_{\hat{a}_1} = \frac{\hat{\sigma}^2{}_\varepsilon}{\sum_{t=1}^{n} (x_t - \bar{x})^2} \quad ; \quad \hat{\sigma}^2{}_{\hat{a}_0} = \hat{\sigma}^2{}_\varepsilon \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{t=1}^{n} (x_t - \bar{x})^2}\right)$$

L'hypothèse de normalité des erreurs implique que : $\frac{\hat{a}_1 - a_1}{\sigma_{\hat{a}_1}}$ et $\frac{\hat{a}_0 - a_0}{\sigma_{\hat{a}_0}}$ suivent une loi normale centrée réduite N (0, 1).

C. Test bilatéral

Soit à tester, à un seuil de 5 %, l'hypothèse H0 :$a_1 = 0$ contre l'hypothèseH1 :$a_1 \neq 0$.

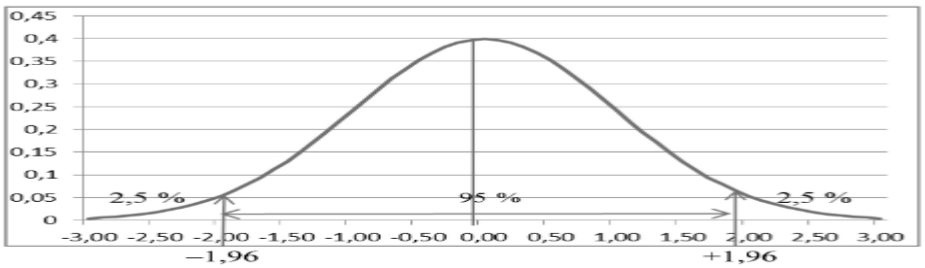
Nous savons que $\frac{\hat{a}_1 - a_1}{\sigma_{\hat{a}_1}}$ suit une loi de student à (n - 2) degrés de liberté.

Sous H0 ($a_1 = 0$) le ratio appelé ratio de Student $\frac{\hat{a}_1 - 0}{\sigma_{\hat{a}_1}}$ suit donc une loi de Student à (n– 2) degrés de liberté[8]. Le test d'hypothèses bilatéral consiste donc à

[8] La notion de degré de liberté correspond au nombre de valeurs restant réellement à disposition après une procédure d'estimation statistique. Si un échantillon comprend 10 observations etqu'on dispose en plus de la moyenne de cet échantillon, on ne peut choisir librement les valeurs que pour 9 de ces observations, la dixième

comparer le ratio de Student empirique $t^* = \frac{|\hat{a}_1|}{\hat{\sigma}_{\hat{a}_1}}$ à la valeur du t de Student lue dans la table à n– 2 degrés de liberté et pour un seuil de probabilité égal à 5 %, soit si n–2 > 30, $t^{0,05}_{\infty}$ =1,96. Si t*> $t^{0,05}_{\infty}$ =1,96, nous rejetons l'hypothèse H0 (graphique 2), le coefficient théorique et inconnu $a_1$est significativement différent de 0.

Graphique 2 – Test bilatéral à 5 %



IV. Équation et tableau d'analyse de la variance

A. Équation d'analyse de la variance

L'équation fondamentale d'analyse de la variance :

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t e_t^2$$

SCT = SCE + SCR

La variabilité totale (SCT) est égale à la variabilité expliquée (SCE) + la variabilité des résidus (SCR).

Cette équation permet de juger de la qualité de l'ajustement d'un modèle.

$$R^2 = \frac{\sum_t (\hat{y}_t - \bar{y})^2}{\sum_t (y_t - \bar{y})^2} = 1 - \frac{\sum_t e_t^2}{\sum_t (y_t - \bar{y})^2}$$

R^2 est appelé le coefficient de détermination

B. Tableau d'analyse de la variance

Le tableau présente l'analyse de la variance pour un modèle de régression simple.

---

se déduisant de la valeur de la moyenne. Dans le cas présent, le modèle de régression simple, le nombre de degrés de liberté est donc de n−2 car nous avons estimé deux paramètres $a_0$et a_1.

Tableau : Analyse de la variance par une régression simple.

Source de variation	**Sommes des carrées**	**Degrés de liberté**	**Carées moyen**
x	$SCE = \sum_t (\hat{y}_t - \bar{y})^2$	1	SCE/1
Résidu	$SCR = \sum_t e_t^2$	n -2	SCR/n-2
Total	$SCT = \sum_t (y_t - \bar{y})^2$	n-1	

Les degrés de liberté correspondent au nombre de valeurs que nous pouvons choisir arbitrairement (par exemple, pour la variabilité totale, connaissant n−1 valeurs, nous pourrons en déduire la n-ième, puisque nous connaissons la moyenne $\bar{y}$.

Le test H0: a_1=0 est équivalent au test d'hypothèse[9]H0: SCE= 0 (la variable explicative x_t ne contribue pas à l'explication du modèle).

Soit le test d'hypothèses H0: SCE = 0 contre l'hypothèse H1:SCE≠ 0.

La statistique[10]de ce test est donnée par : $F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{\sum_t (\hat{y}_t - \bar{y})^2}{1}}{\frac{\sum_t e_t^2}{n-2}}$

Ou encore : $F^* \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{\sum_t (\hat{y}_t - \bar{y})^2}{1}}{\frac{\sum_t e_t^2}{n-2}} = \frac{\frac{R^2}{1}}{\frac{1-R^2}{n-2}}$

F^* suit une statistique de Fisher à 1 et n−2 degrés de liberté. Si $F^* > F^{\alpha}_{1\,;\,n-2}$ nous rejetons au seuil α l'hypothèse H_0 d'égalité des variances, la variable x_t est significative ; dans le cas contraire, nous acceptons l'hypothèse d'égalité des variances, la variable x_t n'est pas explicative de la variable y_t.

[9] Cela n'est vrai que dans le cas du modèle de régression simple

[10] Nous comparons la somme des carrés expliqués SCE à la somme des carrés des résidus SCR qui est représentative de la somme des carrés théoriquement la plus faible.