



La **GREAM** formation

Groupe de recherche en économie appliquée et théorique

N° 003

"Réfléchir à changer"

Mars 2015

Modèle linéaire général



Massa Coulibaly

Table des matières

Introduction	1
1. Hypothèses	2
1.1. Hypothèses structurelles	2
1.2. Hypothèses stochastiques	2
2. Estimation des paramètres	3
3. Propriétés	6
4. Prévvision	10
5. Coefficients de corrélations	12
5.1. Coefficients de corrélations simples	12
5.2. Le coefficient de corrélation multiple	13
5.3. Les coefficients de corrélation partiels	14
6. Analyse de variance et tests	15
6.1. Test de validité de modèle	15
6.2. Test d'amélioration	16
6.3. Test de valeur particulière des regresseurs	17
6.4. Test de Chow	17
6.5. Test de stabilité des coefficients de régression	18
6.6. Test de restriction sur les paramètres	19
7. Analyse de covariance	21

Introduction

Supposons qu'il existe une relation linéaire entre une variable dépendante Y et $(k-1)$ variables explicatives $X_2, X_3, \dots, X_k$.

Sur ces variables (dépendante et indépendantes), on dispose de n observations.

Soit le modèle: $Y_j = \beta_1 + \beta_2 X_{2j} + \dots + \beta_k X_{kj} + u_j$

qu'on écrira plus simplement:

(1) $Y = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$ où les β_i ($1, 2, k$) et les paramètres de la distribution de u sont inconnus et sont à estimer.

Adoptons les notations matricielles suivantes:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2n} & \dots & X_{kn} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Le modèle (1) peut s'écrire:

$$(2) \quad Y = X\beta + u$$

On suppose que les X sont des données non aléatoires donc que la seule source de variation pour Y provient du terme aléatoire u : $E(X'u) = 0$

Le facteur aléatoire u est introduit dans le modèle pour prendre en compte l'influence de diverses "erreurs":

- les variables omises (nombreuses et évoluant dans des directions différentes, ce qui rend leur effet sur Y imprévisible)
- la forme mathématique du modèle
- la mesure de la variable dépendante
- l'élément erratique inhérent au comportement humain.

1. Hypothèses

Pour l'estimation du vecteur β , nous devons faire certaines hypothèses sur les observations ainsi que sur la distribution de u . Ces hypothèses sont soit structurelles soit stochastiques.

1.1. Hypothèses structurelles

Elles définissent la structure du modèle

H1. (*Absence de colinéarité des variables exogènes*)

La matrice X est de rang $k < n$: $\text{rang}(X) = \text{rang}(X'X) = k$ et $(X'X)^{-1} \exists$ i.e.

- le nombre d'observations est supérieur au nombre de paramètres à estimer
- il n'existe pas de combinaison linéaire des X_i qui soit nulle.

H2. (*Espérance et variance finies*)

Lorsque n tend vers l'infini, la matrice $M_{XX} = \frac{1}{n}(X'X)$ tend vers une matrice finie non singulière M_0 .

1.2. Hypothèses stochastiques

H3. (*Hypothèse fondamentale*)

$E(u) = 0, \forall i = 1, 2, \dots, n$ i.e. approximativement, les valeurs positives et négatives de u ont une somme nulle.

H4. (*Homoscédasticité et absence d'autocorrélation des erreurs*)

$E(uu') = \sigma^2 I_n$ constant

Cette écriture est un moyen condensé de formuler une double hypothèse:

$$E(uu') = \begin{pmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ \vdots & \vdots & & \vdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

Soit: $E(u_i^2) = \sigma^2 \quad \forall i=1,2,\dots,n$ les u_i ont une variance constante, c'est l'homoscédasticité

$E(u_i u_j) = 0$ pour $i \neq j$ les u_i ne sont pas deux à deux corrélés

H5. (*Normalité*)

La loi de u est une loi normale. Cette hypothèse est nécessaire pour conduire les tests statistiques de signification des estimateurs (les tests t , χ^2 et F sont basés sur des distributions normales) et construire les intervalles de confiance.

2. Estimation des paramètres

Pour l'estimation, on admet qu'il n'existe a priori aucune restriction sur la valeur des coefficients β . Seulement dans la pratique, la théorie économique peut indiquer leurs signes ou les limites de leurs valeurs.

Soit: $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$ le vecteur des estimations de β .

Nous avons: $\hat{Y} = X\hat{\beta}$

Le vecteur colonnes des résidus sera: $e = Y - \hat{Y} = Y - X\hat{\beta} \Rightarrow Y = X\hat{\beta} + e$

Par la méthode des moindres carrés ordinaires (OLS), on a:

$$\begin{aligned} \min RSS &= \sum e^2 = e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

où chaque terme est un scalaire et donc égal à sa transposée, e.g.

$$Y'X\hat{\beta} = (Y'X\hat{\beta})' = \hat{\beta}'X'Y$$

Il s'ensuit: $e'e = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$

$$\begin{aligned}\frac{\partial e'e}{\partial \hat{\beta}} = 0 &\Leftrightarrow -2X'Y + 2X'X\hat{\beta} = 0 \\ &\Leftrightarrow X'X\hat{\beta} = X'Y \text{ système d'équations normales.}\end{aligned}$$

En prémultipliant chacun des deux termes par $(X'X)^{-1}$, on obtient:

$$(3) \quad \hat{\beta} = (X'X)^{-1}X'Y$$

Remarque: Alternativement RSS pouvait se mettre sous la forme:

$$\begin{aligned}RSS = e'e &= Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ \text{On aurait: } \frac{\partial e'e}{\partial \hat{\beta}} &= -2Y'X + 2\hat{\beta}'X'X = 0 \Leftrightarrow \hat{\beta}'X'X = Y'X \\ &\Rightarrow \hat{\beta}' = (Y'X)(X'X)^{-1} \text{ ou } \hat{\beta} = ((Y'X)(X'X)^{-1})'\end{aligned}$$

$X'X$ étant une matrice symétrique, $(X'X)^{-1}$ aussi, alors:

$$((X'X)^{-1})' = ((X'X)')^{-1} = (X'X)^{-1} \text{ et alors}$$

$$(3) \quad \hat{\beta} = (X'X)^{-1}X'Y$$

(3) est le résultat fondamental des LSE. Il redonne les résultats des modèles simples.

➤ Ainsi dans le **modèle linéaire simple** à deux variables:
 $Y_j = \beta_1 + \beta_2 X_j + u_j$

Ici: $Y = X\beta + u$ avec:

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$\Rightarrow X'X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} n & \sum X \\ \sum X & \sum X^2 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y \\ \sum XY \end{pmatrix}$$

d'où

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'Y \\ &= \frac{1}{n \sum X^2 - (\sum X)^2} \begin{pmatrix} \sum X^2 \sum Y - \sum X \sum XY \\ n \sum XY - \sum X \sum Y \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \end{aligned}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}_2 \bar{X} \\ \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \end{pmatrix}$$

➤ Dans le modèle à 3 variables X_2, X_3 et Y :

$Y_j = \beta_1 + \beta_2 X_{2j} + \beta_3 X_{3j} + u_j$ s'écrira: $Y = X\beta + u$ avec:

$$X = \begin{pmatrix} 1 & X_{21} & X_{31} \\ 1 & X_{22} & X_{32} \\ \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Les équations normales: $(X'X)\hat{\beta} = X'Y$ s'écriront:

$$(S) \begin{cases} n\hat{\beta}_1 + \hat{\beta}_2 \sum X_2 + \hat{\beta}_3 \sum X_3 = \sum Y \\ \hat{\beta}_1 \sum X_2 + \hat{\beta}_2 \sum X_2^2 + \hat{\beta}_3 \sum X_2 X_3 = \sum X_2 Y \\ \hat{\beta}_1 \sum X_3 + \hat{\beta}_2 \sum X_2 X_3 + \hat{\beta}_3 \sum X_3^2 = \sum X_3 Y \end{cases}$$

En divisant les termes de la première équation par n, on obtient:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$

Les deux dernières équations du système deviennent alors:

$$(S) \begin{cases} \hat{\beta}_2 \left(\sum X_2^2 - \bar{X}_2 \sum X_2 \right) + \hat{\beta}_3 \left(\sum X_2 X_3 - \bar{X}_3 \sum X_2 \right) = \sum X_2 Y - \bar{Y} \sum X_2 \\ \hat{\beta}_2 \left(\sum X_2 X_3 - \bar{X}_2 \sum X_3 \right) + \hat{\beta}_3 \left(\sum X_3^2 - \bar{X}_3 \sum X_3 \right) = \sum X_3 Y - \bar{Y} \sum X_3 \end{cases}$$

Un système à deux équations, deux inconnues:

$$(S) \begin{cases} \hat{\beta}_2 \sum x_2^2 + \hat{\beta}_3 \sum x_2 x_3 = \sum x_2 y \\ \hat{\beta}_2 \sum x_2 x_3 + \hat{\beta}_3 \sum x_3^2 = \sum x_3 y \end{cases} \quad \text{avec}$$

$$x = X - \bar{X} \quad \text{et} \quad y = Y - \bar{Y}$$

On trouve ainsi aisément les estimateurs habituels:

$$\hat{\beta}_1 = a_{1.23}, \quad \hat{\beta}_2 = b_{12.3}, \quad \hat{\beta}_3 = b_{13.2}$$

3. Propriétés

P1. Estimateur linéaire

D'après (3), on voit bien que l'estimateur des moindres carrés est fonction linéaire de Y, les éléments de la matrice $(X'X)^{-1}$ étant des constantes

$$\hat{\beta} = (X'X)^{-1} X'Y = LY$$

P2. Estimateur sans biais

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X'Y = (X'X)^{-1} X'(X\beta + u) \\ &= (X'X)^{-1} (X'X)\beta + (X'X)^{-1} X'u = \beta + (X'X)^{-1} X'u\end{aligned}$$

$$E(\hat{\beta}) = E(\beta) + (X'X)^{-1} X'E(u) \text{ avec } E(u) = 0 \text{ par hypothèse (H3)}$$

$$E(\hat{\beta}) = E(\beta) = \beta, \text{ i.e. } \hat{\beta} \text{ est un estimateur sans biais de } \beta.$$

P3. BLUE (Best Linear Unbiased Estimator)

Soit la matrice de variance-covariance ou matrice de dispersion:

$$\begin{aligned}V(\hat{\beta}) &= E((\hat{\beta} - \beta)(\hat{\beta} - \beta)') \\ &= \begin{pmatrix} E(\hat{\beta}_1 - \beta_1)^2 & \cdots & E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_k - \beta_k) \\ \vdots & \ddots & \vdots \\ E(\hat{\beta}_k - \beta_k)(\hat{\beta}_1 - \beta_1) & \cdots & E(\hat{\beta}_k - \beta_k)^2 \end{pmatrix}\end{aligned}$$

$$\text{où: } E(\hat{\beta}_i - \beta_i)^2 = V(\hat{\beta}_i) \text{ et}$$

$$E(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j) = \text{cov}(\hat{\beta}_i, \hat{\beta}_j)$$

$$V(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(X'X)^{-1} X'uu' X (X'X)^{-1}]$$

$$= (X'X)^{-1} X' E(uu') X (X'X)^{-1}$$

$$= (X'X)^{-1} X' (\sigma_u^2 I_n) X (X'X)^{-1}$$

$$V(\hat{\beta}) = \sigma_u^2 (X'X)^{-1} X' X (X'X)^{-1} = \sigma_u^2 (X'X)^{-1}$$

Pour obtenir la variance de $\hat{\beta}_i$, on multiplie le $i^{\text{ème}}$ terme de la diagonale principale de $(X'X)^{-1}$ par σ_u^2 . La covariance entre $\hat{\beta}_i$ et $\hat{\beta}_j$ s'obtient en multipliant σ_u^2 par le terme de la $i^{\text{ème}}$ ligne, $j^{\text{ème}}$ colonne de $(X'X)^{-1}$.

Pour obtenir l'estimateur de σ_u^2 , on développe RSS

$$\begin{aligned}
RSS &= e'e = (Y - \hat{X}\hat{\beta})'(Y - \hat{X}\hat{\beta}) = (X\beta + u - \hat{X}\hat{\beta})'(X\beta + u - \hat{X}\hat{\beta}) \\
&= (u - X(\hat{\beta} - \beta))'(u - X(\hat{\beta} - \beta)) \\
&= \left[(I_n - X(X'X)^{-1}X')u \right]' \left[(I_n - X(X'X)^{-1}X')u \right] \\
&= (Nu)'(Nu) = u'N'Nu
\end{aligned}$$

où $N = (I_n - X(X'X)^{-1}X')$ est une matrice symétrique et idempotente:

$$N' = N \text{ et } N^2 = N.$$

En plus, N est semi-définie positive

Il s'en suit: $RSS = e'e = u'N'Nu = u'N^2u = u'Nu$

$$RSS = u'(I_n - X(X'X)^{-1}X')u = \sum e^2 \geq 0 \text{ i.e. } N \text{ est semi-définie positive}$$

$$\begin{aligned}
E(e'e) &= E(u'Nu) \\
&= E(u_1 \ u_2 \ \dots \ u_n) \begin{pmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & & \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}
\end{aligned}$$

Puisque par hypothèse (H4), $E(u_i u_j) = 0$ alors:

$$RSS = E(e'e) = \sum_i a_{ii} E(u_i^2) = \sigma_u^2 \sum_i a_{ii}$$

où $\sum a_{ii}$ est justement la trace de la matrice N; on aura donc:

$$\begin{aligned}
E(e'e) &= E(u'Nu) = \sigma_u^2 \text{tr}(N) \\
&= \sigma_u^2 \text{tr}(I_n - X(X'X)^{-1}X') \\
&= \sigma_u^2 (\text{tr}(I_n) - \text{tr}(X(X'X)^{-1}X')) \text{ avec } \text{tr}(AB) = \text{tr}(BA) \\
&= \sigma_u^2 (\text{tr}(I_n) - \text{tr}((X'X)^{-1}X'X)) \\
&= \sigma_u^2 (\text{tr}(I_n) - \text{tr}(I_k)) \\
&= \sigma_u^2 (n - k) = e'e = RSS
\end{aligned}$$

Ainsi, l'estimateur sans biais de σ_u^2 est: $\hat{\sigma}_u^2 = \frac{RSS}{n-k}$

Maintenant, l'estimateur de β est BLUE si de tous les estimateurs linéaires sans biais, il est celui qui a la plus faible variance:

On considère β^* , un autre estimateur linéaire de β et on montre:

$$Var(\hat{\beta}) \leq Var(\beta^*)$$

Ce qui revient à montrer que $V(\beta^*) - V(\hat{\beta})$ est une matrice semi-définie positive.

Rappelons qu'une matrice $A = (a_{ij})_{n,n}$ est semi-définie positive si $\forall$ le vecteur colonne Z , l'on a: $Z'AZ \geq 0$.

$$\begin{aligned} \text{Soit } \beta^* &= ((X'X)^{-1}X' + C)Y \text{ où } C = (c_{ij})_{k,n} \\ &= (X'X)^{-1}X'Y + CY \\ &= (X'X)^{-1}X'(X\beta + u) + C(X\beta + u) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u + CX\beta + Cu \\ &= \beta + CX\beta + ((X'X)^{-1}X' + C)u \\ &= (I_k + CX)\beta + ((X'X)^{-1}X' + C)u \end{aligned}$$

$$E(\beta^*) = \beta \text{ puisque } \beta^* \text{ est sans biais, } \Rightarrow CX = 0$$

$$\begin{aligned} V(\beta^*) &= E((\beta^* - \beta)(\beta^* - \beta)') \\ &= ((X'X)^{-1}X' + C)E(u'u)(X(X'X)^{-1} + C') \\ &= \sigma_u^2 ((X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'C' + CX(X'X)^{-1} + CC') \\ &= \sigma_u^2 ((X'X)^{-1} + (X'X)^{-1}(CX)' + CX(X'X)^{-1} + CC') \\ &= \sigma_u^2 ((X'X)^{-1} + CC') \text{ puisque } CX = 0 \end{aligned}$$

$$\text{Il s'en suit que: } V(\beta^*) - V(\hat{\beta}) = \sigma_u^2 CC'$$

CC' est bien une matrice semi-définie positive puisqu'on ne s'intéresse qu'aux éléments de la diagonale principale, lesquels éléments sont de la forme:

$$\sum_{j=1}^n c_{ij}^2 \quad i=1,2,\dots,k$$

Note Pour que l'estimateur de β soit convergent, il est nécessaire et suffisant que la matrice $(X'X)^{-1}$ tende vers zéro lorsque le nombre n d'observations croît à l'infini i.e. les variables exogènes ne tendent pas à devenir colinéaires (H1).

4. Prévision

Connaissant la relation $Y = X\beta + u$, nous pouvons faire des prévisions sur la valeur de Y associée à des valeurs X données au temps futur θ .

Soit le vecteur: $F = \begin{pmatrix} 1 \\ X_{2\theta} \\ \vdots \\ X_{k\theta} \end{pmatrix}$

1). $\hat{Y}_\theta = F' \hat{\beta}$

$$E(\hat{Y}_\theta) = E(F' \hat{\beta}) = F' E(\hat{\beta}) = F' \beta$$

$$V(\hat{Y}_\theta) = E[(F' \hat{\beta} - F' \beta)(F' \hat{\beta} - F' \beta)'] = E[F'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'F]$$

avec $\hat{\beta} - \beta = (X'X)^{-1} X'u$

$$= F'(X'X)^{-1} X' E(uu') X (X'X)^{-1} F$$

$$= \sigma_u^2 F'(X'X)^{-1} F$$

$$\hat{Y}_\theta = F' \hat{\beta} \mapsto N(F' \beta, \sigma_u^2 F'(X'X)^{-1} F) \quad \text{d'où}$$

$$\frac{F' \hat{\beta} - F' \beta}{SE(\hat{Y}_\theta)} \mapsto t(n-k \text{ ddl}) = \frac{F' \hat{\beta} - F' \beta}{\sqrt{\hat{\sigma}_u^2 F'(X'X)^{-1} F}}$$

d'où la prévision:

$$Y_{\theta|F} = \hat{Y}_{\theta} \pm t_{\varepsilon/2} SE(\hat{Y}_{\theta}) = F' \hat{\beta} \pm t_{\varepsilon/2} \sqrt{\hat{\sigma}_u^2 F' (X' X)^{-1} F}$$

➤ **Observation**

De ce résultat général, il est aisé de retrouver la prévision dans le cas du modèle linéaire simple $Y = \alpha + \beta X + u$

$$(X' X) = \begin{pmatrix} n & \sum X \\ \sum X & \sum X^2 \end{pmatrix}$$

$$\Rightarrow (X' X)^{-1} = \frac{1}{n \sum X^2 - (\sum X)^2} \begin{pmatrix} \sum X^2 & -\sum X \\ -\sum X & n \end{pmatrix}$$

Soit $F' = (1 \quad X_{\theta})$ alors:

$$F'(X' X)^{-1} F = \frac{\sum X^2 - 2X_{\theta} \sum X + nX_{\theta}^2}{n \sum X^2 - (\sum X)^2} = \frac{\sum X^2 - 2X_{\theta} n\bar{X} + nX_{\theta}^2}{n \sum X^2 - (n\bar{X})^2}$$

$$= \frac{\sum X^2 - n\bar{X}^2 + n(X_{\theta}^2 - 2X_{\theta}\bar{X} + \bar{X}^2)}{n(\sum (X - \bar{X})^2)} = \frac{\sum x^2 + n(X_{\theta} - \bar{X})^2}{n \sum x^2}$$

$$\text{Soit: } F'(X' X)^{-1} F = \frac{1}{n} + \frac{(X_{\theta} - \bar{X})^2}{\sum x^2}$$

Soit donc l'intervalle de confiance:

$$E(Y_{\theta|X_{\theta}}) = \hat{\alpha} + \hat{\beta} X_{\theta} \pm t_{\varepsilon/2} \hat{\sigma}_u \sqrt{\frac{1}{n} + \frac{(X_{\theta} - \bar{X})^2}{\sum x^2}}$$

(Modification du modèle) Si nous nous intéressons, comme nous l'avons fait pour le modèle linéaire simple, à l'intervalle de confiance de Y en tant que point de prévision:

$$\hat{Y}_{\theta} = F' \hat{\beta}$$

$$Y_{\theta} = F' \beta + u_{\theta}$$

$$\hat{Y}_\theta - Y_\theta = F'(\hat{\beta} - \beta) - u_\theta = z$$

z est l'erreur de prévision, avec:

$$E(z) = 0$$

puisque $E(\hat{\beta}) = \beta$ et $E(u_\theta) = 0$ par hypothèse

$$V(z) = V(F'(\hat{\beta} - \beta) - u_\theta)$$

Les $\hat{\beta}$ ne dépendent pas des u_i , $i = 1, 2, \dots, n$ avec $n < \theta$; il n'existe donc pas de corrélation entre ces $\hat{\beta}$ et u_θ :

$$\begin{aligned} V(z) &= V(F'(\hat{\beta} - \beta)) + V(u_\theta) = E(F'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'F) + E(u_\theta u_\theta') \\ &= F'V(\hat{\beta})F + \sigma_u^2 \\ &= \sigma_u^2(1 + F'(X'X)^{-1}F) \end{aligned}$$

Par conséquent:

$$\frac{\hat{Y}_\theta - Y_\theta}{SE(z)} \mapsto t(n-k) = \frac{\hat{Y}_\theta - Y_\theta}{\sqrt{\hat{\sigma}_u^2(1 + F'(X'X)^{-1}F)}}$$

L'intervalle de prévision est tel que:

$$Y_\theta = \hat{Y}_\theta \pm t_{\xi/2} SE(z) = F'\hat{\beta} \pm t_{\xi/2} \sqrt{\hat{\sigma}_u^2(1 + F'(X'X)^{-1}F)}$$

5. Coefficients de corrélations

5.1. Coefficients de corrélations simples

Soit la matrice de corrélation, formée des coefficients de corrélation simples:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \vdots & r_{2k} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{pmatrix} \text{ avec } r_{ij} = \begin{cases} r_{ji} \\ 1 & \text{si } i = j \end{cases}$$

Le coefficient de corrélation multiple et les coefficients partiels peuvent être estimés en terme de cofacteurs de la matrice R. Ils peuvent également être calculés directement.

5.2. Le coefficient de corrélation multiple

Soit le coefficient de détermination multiple:

$$R_{1.23\dots k}^2 = \frac{\sum \hat{y}^2}{\sum y^2} = \frac{\hat{y}'\hat{y}}{y'y} = \frac{(x\hat{\beta})'(x\hat{\beta})}{y'y}$$

$$\text{De: } \hat{\beta} = (x'x)^{-1}x'y$$

$$\text{il s'en suit: } (x\hat{\beta})'(x\hat{\beta}) = \hat{\beta}'(x'x)\hat{\beta} = \hat{\beta}'x'y$$

$$(1) \quad R_{1.23\dots k}^2 = \frac{\hat{\beta}'x'y}{y'y}$$

$$\text{ou encore: } R_{1.23\dots k}^2 = 1 - \frac{\sum e^2}{\sum y^2}$$

Le coefficient de détermination multiple traduit le rôle des variables explicatives dans leur ensemble sur l'évolution de la variable dépendante. Il ne traduit pas l'influence des variables exogènes prises séparément sur l'évolution de l'endogène, influence que donnent les regressseurs du modèle.

Le coefficient de détermination ($R_{1.2\dots k}^2$) tel que calculé plus haut ne tient compte ni du nombre n d'observations, ni du nombre (k-1) de variables explicatives dans le modèle. Or, sur un même ensemble d'observations, on peut estimer 2 modèles distincts, le second faisant intervenir un certain nombre de variables explicatives nouvelles en plus de celles rencontrées dans le premier. Le coefficient de corrélation multiple de la seconde régression ne peut qu'augmenter (au pire rester constant) du fait de l'augmentation de la variabilité

expliquée. En d'autres termes, le $R^2_{1,2,\dots,k}$ croît avec le nombre de variables explicatives retenues dans le modèle, indépendamment de leur pouvoir explicatif.

Il y a donc un biais dans l'estimation du coefficient de corrélation multiple dû au fait que l'on ne tient compte ni du nombre d'observations ni du nombre de variables explicatives. Pour y remédier, on a défini le coefficient corrigé, $\bar{R}^2$ /:

$$\bar{R}^2 = 1 - \frac{\sum e^2 / (n-k)}{\sum y^2 / (n-1)} = 1 - \frac{n-1}{n-k} \cdot \frac{\sum e^2}{\sum y^2}$$

$$(2) \quad \bar{R}^2 = 1 - \frac{n-1}{n-k} (1 - R^2)$$

Si $k=1 \rightarrow \bar{R}^2 = R^2$

Si $k > 1 \rightarrow \bar{R}^2 < R^2$ car $\bar{R}^2 - R^2 = \frac{1-R^2}{n-k} (1-k) < 0$

$\bar{R}^2$ peut prendre des valeurs négatives si $R^2 < \frac{k-1}{n-1}$

Si l'accroissement de $\frac{n-1}{n-k}$ compense l'augmentation de R^2 alors $\bar{R}^2$ peut diminuer avec l'introduction de nouvelles variables exogènes dans le modèle.

5.3. Les coefficients de corrélation partiels

Soient les modèles linéaires simples

$$DY(X): Y = \beta X + \alpha + u \text{ et } DX(Y): X = b Y + a + u.$$

Le coefficient de corrélation r est / $r^2 = \hat{\beta} \hat{b}$

Le coefficient de corrélation partiel est défini en les mêmes termes de produit de 2 coefficients de régression. Ainsi, dans un modèle à 3 variables, pour obtenir $r_{12.3}$, il faut supposer avoir fait 2 régressions multiples: $D_Y(X_2, X_3)$ et $D_{X_2}(Y, X_3)$

L'extension à plus de 3 variables est simple puisqu'il s'agit seulement d'ajouter d'autres termes (maintenus constants) dans les 2 régressions.

6. Analyse de variance et tests

Soient les modèles

(m₁) $Y = X\beta + u$ modèle à k variables et

(m₂) $Y = Xb + v$ modèle à m variables parmi les k variables de (m₁).

Dans un tableau dit d'analyse de variance, on regroupe:

- la variabilité non expliquée par les (m-1) exogènes de (m₂) i.e. la variabilité résiduelle associée à (m₂)
- la variabilité non expliquée par les (k-1) exogènes de (m₁)
- la variabilité expliquée par (k-m) exogènes de (m₁).

Après estimation des paramètres, nous cherchons à établir la crédibilité des estimateurs. Il existe 3 types de critères:

- critère économique d'a priori, les postulats de la théorie économique au regard du signe et de la grandeur des paramètres
- critère statistique ou tests de premier ordre, les tests t, F, ..., tests validés si les hypothèses de la régression linéaire sont satisfaites
- critère économétrique ou tests de second ordre, ce sont des tests portant sur des hypothèses stochastiques (examen des résidus e en tant qu'estimation des u) ou non stochastiques (examen des variables explicatives et leur interdépendance) d'identification, de colinéarité et d'agrégation.

6.1. Test de validité de modèle

Le test est de savoir si les variables explicatives ont une influence significative sur la variable dépendante:

l'hypothèse nulle $H_0 : \beta_j = 0 \quad (j = \overline{1, k})$ contre l'hypothèse alternative $H_1 :$

les β_j ne sont pas tous nuls

Le test s'effectue avec le tableau d'analyse de variance:

$$F = \frac{\sum \hat{y}^2 / (k-1)}{\sum e^2 / (n-k)} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1}$$

Si $F > F_\varepsilon(k-1, n-k)$ on dira que la régression est significative ou que tous les β_j ne sont pas nuls.

En général, des valeurs élevées de F suggèrent une relation significative entre Y et les X

6.2. Test d'amélioration

Des modèles (m_1) et (m_2), on déduit le tableau suivant d'analyse de variance:

Source de variabilité	Somme des carrés	ddl	Carrés moyens
$X_2, X_3, \dots, X_m$	$r^2 y' y$	$m-1$	$r^2 y' y / (m-1)$
$X_2, \dots, X_m, X_{m+1}, \dots, X_k$	$R^2 y' y$	$k-1$	$R^2 y' y / (k-1)$
Additionnelle	$(R^2 - r^2) y' y$	$k-m$	$(R^2 - r^2) y' y / (k-m)$
Résiduelle	$(1-R^2) y' y$	$n-k$	$(1-R^2) y' y / (n-k)$
Totale	$y' y$	$n-1$	

Le test d'amélioration due aux regresseurs additionnels:

$$F = \frac{R^2 - r^2}{1-R^2} \cdot \frac{n-k}{k-m}$$

Si $F < F_\varepsilon(k-m, n-k)$ on accepte l'hypothèse nulle i.e. que l'addition des $X_{m+1}, \dots, X_k$ n'améliore pas significativement le modèle (m_2) donc que l'explication de Y peut se limiter aux seules variables $X_2, X_3, \dots, X_m$

Si $F > F_\varepsilon(k-m, n-k)$ alors les variables additionnelles $X_{m+1}, \dots, X_k$ sont d'importantes variables explicatives ou qu'on ne pouvait se contenter du seul modèle (m_2) pour expliquer Y. Dans ce cas on peut en même temps calculer le **gain de précision** qu'on obtient avec le modèle (m_1) par rapport à (m_2):

$$G_{1/2} = \frac{\sum e_2^2 - \sum e^2}{\sum e_2^2} = \frac{(1-r^2) - (1-R^2)}{1-r^2} = \frac{R^2 - r^2}{1-r^2} = r_{1(m+1) \dots k.23 \dots m}^2$$

6.3. Test de valeur particulière des régresseurs

H: $\beta_j = \beta_j^0$ une valeur particulière.

On calcule: $\frac{\hat{\beta}_j - \beta_j^0}{SE(\hat{\beta}_j)} = t$

Si $|t| > t_{\alpha/2}(n-k)$ alors β_j est distinct de β_j^0 .

6.4. Test de Chow

C'est le test d'égalité entre coefficients obtenus à partir d'échantillons différents. Sur les variables Y et X on considère 2 échantillons de taille n_1

respectivement n_2 avec les 2 régressions: $\hat{Y}_1 = X_1 \hat{\beta}_1$
 $\hat{Y}_2 = X_2 \hat{\beta}_2$

Nous voulons tester si ces 2 relations estimées diffèrent significativement auquel cas nous concluons que la relation change d'un échantillon à l'autre: $H_0: \beta_1 = \beta_2$

On applique le test de Chow qui comprend les étapes suivantes:

1) On met ensemble les 2 échantillons et on fait la régression:

$$\hat{Y} = X\hat{\beta} \rightarrow \sum e^2 = y'y - \hat{y}'\hat{y} \text{ à } (n_1 + n_2 - k) \text{ ddl}$$

2) On fait la régression pour chacun des 2 échantillons:

$$\begin{aligned} \hat{Y}_1 &= X_1 \hat{\beta}_1 \rightarrow \sum e_1^2 = y_1'y_1 - \hat{y}_1'\hat{y}_1 \text{ à } (n_1 - k) \text{ ddl} \\ \hat{Y}_2 &= X_2 \hat{\beta}_2 \rightarrow \sum e_2^2 = y_2'y_2 - \hat{y}_2'\hat{y}_2 \text{ à } (n_2 - k) \text{ ddl} \end{aligned}$$

3) On forme la variabilité résiduelle totale:

$$\sum e_*^2 = \sum e_1^2 + \sum e_2^2 \text{ à } (n_1 - k + n_2 - k) = (n_1 + n_2 - 2k) \text{ ddl}$$

4) On calcule: $\sum e^2 - \sum e_*^2 \text{ à } (n_1 + n_2 - k - n_1 - n_2 + 2k) = k \text{ ddl}$

5) On calcule la statistique:

$$F_* = \frac{\left(\sum e^2 - (\sum e_1^2 + \sum e_2^2)\right)/k}{\left(\sum e_1^2 + \sum e_2^2\right)/(n_1 + n_2 - 2k)}$$

Si $F_* > F_\epsilon(k, n_1 + n_2 - 2k)$ on dira que les 2 modèles diffèrent significativement.

6.5. Test de stabilité des coefficients de régression

Quand la taille de l'échantillon augmente:

a) Si le nombre d'observations additionnelles est supérieur au nombre de paramètres dans le modèle, on utilise le test précédent i.e. on utilise les observations additionnelles comme un échantillon à part et on applique le test de Chow.

b) Si par contre le nombre d'observations nouvelles n_2 est inférieur au nombre de paramètres:

1) On fait la régression de l'échantillon augmenté (de n_2):

$$\hat{Y} = X\hat{\beta} \rightarrow \sum e^2 = y'y - \hat{y}'\hat{y} \text{ à } (n_1 + n_2 - k) \text{ ddl}$$

2) On fait la régression sur l'échantillon originel (de taille n_1):

$$\hat{Y}_1 = X_1\hat{\beta}_1 \rightarrow \sum e_1^2 = y_1'y_1 - \hat{y}_1'\hat{y}_1 \text{ à } (n_1 - k) \text{ ddl}$$

3) On calcule la quantité: $\sum e^2 - \sum e_1^2 \text{ à } (n_1 + n_2 - k - n_1 + k) = n_2 \text{ ddl}$

4) On calcule la statistique:

$$F_{**} = \frac{(\sum e^2 - \sum e_1^2)/n_2}{\sum e_1^2/(n_1 - k)} \text{ statistique qui permet de tester: } H_0: \beta_1 = \beta$$

Si $F_{**} > F_{\varepsilon}(n_2, n_1 - k)$ on dira que les coefficients sont instables, leur valeur changeant avec l'extension de la série.

6.6. Test de restriction sur les paramètres

Soit le modèle $Y = X\beta + u$

Supposons que l'on ait un nombre C de restrictions linéaires e.g.

$$\beta_1 = 1$$

$$\beta_2 = \beta_3$$

⋮

et que l'on veuille tester l'hypothèse nulle H_0 : toutes les restrictions sont vraies contre l'hypothèse alternative:

1) On applique les OLS au modèle sans restriction, on obtient:

$$URSS = \sum e^2 \text{ à } (n - k) \text{ ddl}$$

2) On applique les OLS au modèle qui incorpore les C restrictions, on obtient:

$$RRSS = \sum e_r^2 \text{ à } (n - (k - C)) \text{ ddl}$$

3) On calcule la statistique:

$$F_r = \frac{(\sum e_r^2 - \sum e^2)/C}{\sum e^2/(n - k)}$$

Si $F_r > F_{\varepsilon}(C, n - k)$ on dira que le modèle ne satisfait pas toutes les restrictions qu'impose la théorie économique aux régresseurs.

Exemple: On considère sur la période 1970-1995, l'évolution de trois grandeurs économiques:

PIB : le produit intérieur brut

K : la formation brute du capital fixe

L : la population occupée

Ces trois grandeurs sont liées entre elles, au temps t , par une relation de type fonction de production CES:

$$PIB_t = A(\alpha L_t^{-\rho} + \beta K_{t-1}^{-\rho})^{-\frac{1}{\rho}} \text{ avec: } \alpha + \beta = 1 \text{ et } \rho = \frac{1-\sigma}{\sigma}$$

où: A : un paramètre d'échelles
 α, β : des paramètres de répartition
 σ : l'élasticité de substitution

- 1) Pour $\sigma = 1$, que devient la fonction CES, notée CD
- 2) Estimer par la méthode des moindres carrés les paramètres de la fonction CD
- 3) Tester sur CD, l'hypothèse de rendement d'échelle constant

$$PIB = A(\alpha L^{-\rho} + \beta K^{-\rho})^{-\frac{1}{\rho}}$$

1) De $\rho = \frac{1-\sigma}{\sigma}$ si $\sigma = 1 \Rightarrow \rho = 0$

$PIB(\rho=0) = 1^\infty$ une indétermination, qu'on lève en linéarisant la fonction :

$$\begin{aligned} \lim_{\rho \rightarrow 0} \ln(PIB) &= A \lim_{\rho \rightarrow 0} \frac{-\ln(\alpha L^{-\rho} + \beta K^{-\rho})}{\rho} \\ &= A \lim_{\rho \rightarrow 0} \frac{\frac{\alpha L^{-\rho} \ln L + \beta K^{-\rho} \ln K}{\alpha L^{-\rho} + \beta K^{-\rho}}}{1} = A \frac{\alpha \ln L + \beta \ln K}{\alpha + \beta} = A \ln L^\alpha K^\beta \end{aligned}$$

$$\Rightarrow \lim_{\rho \rightarrow 0} (PIB) = A L^\alpha K^\beta$$

Soit la fonction CD : $PIB_t = A L_t^\alpha K_{t-1}^\beta$

2) On linéarise d'abord la fonction CD ci-dessus :

$$\ln PIB_t = \ln A + \alpha \ln L_t + \beta \ln K_{t-1} + u_t$$

$$Y_t = \beta_1 + \alpha X_{2t} + \beta X_{3t-1} + u_t$$

$$Y = X_\beta + u$$

Par la méthode des moindres carrés: $\hat{\beta} = (X'X)^{-1} X'Y$ ou $\hat{b} = (x'x)^{-1} x'y$

$$V(\hat{b}) = \hat{\sigma}_u^2 (x'x)^{-1}$$

$$\hat{\sigma}_u^2 = \frac{\sum e^2}{n-k} = \frac{(1-R^2)y'y}{n-k}$$

3) Rendement d'échelle :

$$F(\lambda L, \lambda K) = \lambda^r F(K, L)$$

$$\text{Sur CD: } A (\lambda L)^\alpha (\lambda K)^\beta = \lambda^{\alpha+\beta} A L^\alpha K^\beta = \lambda^{\alpha+\beta} F(K, L)$$

$r = \alpha + \beta \rightarrow$ rendement d'échelle constant :

$$\alpha + \beta = 1 \Rightarrow \alpha = 1 - \beta$$

$$\text{CD : PIB} = A L_t^{1-\beta} K_{t-1}^\beta e^{u_t}$$

$$Y_t = \beta_1 + (1 - \beta) X_{2t} + \beta X_{3t-1} + U_t$$

$$Y_t = \beta_1 + X_{2t} + \beta(X_{2t} - X_{3t-1}) + U_t$$

Soit le modèle :

$$Y_t - X_{2t} = \beta_1 - \beta(X_{2t} - X_{3t-1}) + U_t$$

Après régression on réécrit :

$$\hat{Y}_t = \hat{\beta}_1 - \hat{\beta}(X_{2t} - X_{3t-1}) + X_{2t} \rightarrow \sum e_r^2 = \text{RRSS à } (n - (k - c)) \text{ ddl c'est à dire à } (n - 2) \text{ ddl}$$

Or sur le modèle sans restriction, on a : $\sum e^2 = \text{URSS à } (n - k) \text{ ddl.}$

$$\text{Soit } F_r = \frac{(\text{RRSS} - \text{URSS})/c}{\text{URSS}/(n - k)}$$

Si $F_r < F_{\varepsilon}(C, n-k)$, le modèle satisfait à l'hypothèse de rendement d'échelle constant. Sinon, l'hypothèse n'est pas satisfaite.

7. Analyse de covariance

Le tableau d'ANOVA est une méthode qui permet de décomposer la variabilité totale de la variable dépendante en variabilité expliquée (par les variables exogènes) et variabilité résiduelle. La variabilité expliquée peut à son tour être décomposée en variabilité associée à n'importe quel sous-ensemble de variables explicatives et variabilité additionnelle.

Supposant à présent un échantillon portant sur la variable Y tel que:

$$Y_{ij} \quad i = \overline{1, p} \quad (\text{la classe à laquelle appartient Y})$$

$$j = \overline{1, m} \quad (\text{le nombre d'observations par classe})$$

e.g.: Y = le revenu de m ménages pris dans p régions du pays

	1	2	...	i	...	p
1	Y_{11}	Y_{12}		Y_{1i}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2i}		Y_{2p}
j	Y_{j1}	Y_{j2}		Y_{ji}		Y_{jp}
m	Y_{m1}	Y_{m2}		Y_{mi}		Y_{mp}
Total				$\sum_j Y_{ij}$		

$$\bar{Y}_i = \frac{\sum_j Y_{ij}}{m} \quad i = 1, p \text{ les moyennes de classe}$$

On se pose la question de savoir si les moyennes d'échantillon indiquent une variabilité significative (de Y) entre les classes.

La variabilité totale de Y peut s'exprimer comme la somme de la variabilité entre classes et de la variabilité à l'intérieur des classes i.e.

$$\sum_{i,j} (Y_{ij} - \bar{Y})^2 = \sum_i m(\bar{Y}_i - \bar{Y})^2 + \sum_{i,j} (Y_{ij} - \bar{Y}_i)^2 \quad \text{avec } \bar{Y} = \frac{1}{mp} \sum_{i,j} Y_{ij}$$

On sait que:

$$\begin{aligned} \sum_j (Y_{ij} - \bar{Y})^2 &= \sum_j \left((Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y}) \right)^2 \\ &= \sum_j (Y_{ij} - \bar{Y}_i)^2 + 2(\bar{Y}_i - \bar{Y}) \underbrace{\sum_j (Y_{ij} - \bar{Y}_i)}_{=0} + m(\bar{Y}_i - \bar{Y})^2 \\ &= \sum_j (Y_{ij} - \bar{Y}_i)^2 + m(\bar{Y}_i - \bar{Y})^2 \\ \Rightarrow \sum_i \sum_j (Y_{ij} - \bar{Y})^2 &= \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 + m \sum_i (\bar{Y}_i - \bar{Y})^2 \\ &= \underbrace{\sum_{i,j} (Y_{ij} - \bar{Y}_i)^2}_{S_2^2} + \underbrace{m \sum_i (\bar{Y}_i - \bar{Y})^2}_{S_1^2} \end{aligned}$$

Le test F de signifiacnce de la variabilité entre moyennes de classe est:

$$F = \frac{\sum m(\bar{Y}_i - \bar{Y})^2 / (p-1)}{\sum_{ij} (Y_{ij} - \bar{Y}_i)^2 / p(m-1)}$$

$p(m-1)$ car il y a p classes et sur chaque classe il y a $(m-1)$ ddl d'où $p(m-1)$ ddl.

Les écarts entre classes proviennent des variabilités explicatives à l'intérieur des classes e.g. les classes de revenu peuvent différer selon la taille des ménages, selon que les femmes gagnent ou non un revenu. Ainsi ces écarts reflètent-ils non seulement l'effet de classe mais aussi l'effet de toute différence liée aux variables explicatives.

L'analyse de covariance consiste essentiellement à corriger statistiquement les effets des variables explicatives entre classes. Pour faire ces corrections statistiques, on suppose que le coefficient de régression de Y sur X est le même quelle que soit la classe. Sinon, dans la pratique, on teste l'hypothèse d'égalité des coefficients de régression de toutes les classes e.g. la fonction de demande a-t-elle le même coefficient pour différents groupes sociaux, différentes régions du pays ou différentes périodes de temps.

L'analyse de covariance est utilisée pour tester:

- les écarts à l'intérieur de classes
- les différences de pentes (coef. de régression) entre classes
- les différences entre classes en ne prenant que les équations totales sans distinction entre pente et terme constant.

Soit un échantillon de $n (= mp)$ observations:

$$Y_{ij} \text{ et } X_{rij} \quad i = \overline{1, p} \quad j = \overline{1, m} \quad \text{et} \quad r = \overline{1, k}$$

$$(1) \quad Y = X\beta + u \quad \text{où:}$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = (X_{ij})_{nk} = \begin{pmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{2n} & \dots & X_{kn} \end{pmatrix}$$

Le modèle (1) fait abstraction du fait que les observations sont groupées par classe (e.g. les ménages du pays par région). En prenant en compte l'existence possible d'un effet de classe, on considère un modèle plus général:

$$(2) Y = D\alpha + X\beta + u \quad \text{où } \alpha = \begin{pmatrix} \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{pmatrix} \quad D = (D_2 \ D_3 \cdots D_p)$$

chaque colonne de D contenant p sous vecteurs à m éléments chacun, D se mettra sous la forme:

$$D = \begin{pmatrix} \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{matrix} \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{matrix} & \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{matrix} & \cdots & \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{matrix} \end{pmatrix} = (d_{ij})_{mp, (p-1)}$$

Le modèle (2) s'écrit pour les différentes classes:

$$\text{Cl. 1} \quad Y_{1j} = \beta_1 + \beta_2 X_{21j} + \cdots + \beta_k X_{k1j} + u_{1j} \quad j = \overline{1, m}$$

$$\text{Cl. 2} \quad Y_{2j} = (\beta_1 + \alpha_2) + \beta_2 X_{22j} + \cdots + \beta_k X_{k2j} + u_{2j}$$

$$\vdots$$

$$\text{Cl. p} \quad Y_{pj} = (\beta_1 + \alpha_p) + \beta_2 X_{2pj} + \cdots + \beta_k X_{kpj} + u_{pj}$$

1) Ainsi les termes constants varient d'une classe à l'autre mais nous avons un même ensemble de coefficient $\beta_2, \dots, \beta_k$

S'il n'y a pas d'effet (différentiel) de classe, alors $\alpha_2 = \alpha_3 = \cdots = \alpha_p = 0$

Nous testons donc l'hypothèse $\alpha = 0$. Ce qui revient à demander si le modèle (2) donne un accroissement significatif de la variabilité expliquée par rapport à

(1) ou ce qui revient au même si la variabilité résiduelle associée à (2) est significativement moindre ou plus petite que celle associée à (1).

2) Les OLS sur (1) $\rightarrow Y = X\hat{\beta} + s$ avec $\hat{\beta} = (X'X)^{-1}X'Y$

$$\begin{aligned} Y'Y &= (X\hat{\beta} + s)'(X\hat{\beta} + s) \\ &= \hat{\beta}'X'X\hat{\beta} + s's + 2\hat{\beta}'X's \end{aligned}$$

où: $X's = X'(Y - X\hat{\beta}) = X'Y - X'X((X'X)^{-1}X'Y) = X'Y - X'Y = 0$

$$\Rightarrow Y'Y = \hat{\beta}'X'Y + s's \quad \text{d'où:}$$

$$(3) \quad s's = Y'Y - \hat{\beta}'X'Y$$

3) Les OLS sur (2) $\rightarrow Y = D\hat{\alpha} + X\hat{\beta} + e$

$$\text{avec } \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} D'D & D'X \\ X'D & X'X \end{pmatrix}^{-1} \begin{pmatrix} D'Y \\ X'Y \end{pmatrix}$$

$$Y'Y = \hat{\alpha}'D'Y + \hat{\beta}'X'Y + e'e \quad \text{d'où:}$$

$$(4) \quad e'e = Y'Y - \hat{\alpha}'D'Y - \hat{\beta}'X'Y$$

4) La réduction de RSS en quittant (1) pour (2) est:

$$(5) \quad s's - e'e = \hat{\alpha}'D'Y + \hat{\beta}'X'Y - \hat{\beta}'X'Y$$

5) La structure du test de $H: \alpha = 0$ est donnée dans le tableau:

Source	Somme des carrés	Ddl	Carrés moyens
X	$s's = Y'Y - \hat{\beta}'X'Y$	mp-k	$s's / mp - k$
X et D	$e'e = Y'Y - \hat{\alpha}'D'Y - \hat{\beta}'X'Y = S_2$	mp-k-(p-1)	$S_2 / (mp - k - p + 1)$
D	$s's - e'e = \hat{\alpha}'D'Y + \hat{\beta}'X'Y - \hat{\beta}'X'Y = S_1$	p-1	$S_1 / (p - 1)$

6) Le test de $H: \alpha = 0$ est donné par la statistique:

$$(T1) \quad F(1) = \frac{S_1 / (p-1)}{S_2 / (mp-k-p+1)}$$

Ce test est basé sur la supposition d'un ensemble commun de pentes $\beta_2, \dots, \beta_k$ sur toutes les classes. Cette supposition reste à tester également.

7) Supposons donc que les coefficients varient selon les classes

- On fait la régression séparément sur chaque classe
Soit la classe i:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im} \end{pmatrix} \quad X_i = \begin{pmatrix} 1 & X_{i21} & \cdots & X_{ik1} \\ 1 & X_{i22} & & X_{ik2} \\ \vdots & & & \vdots \\ 1 & X_{i2m} & \cdots & X_{ikm} \end{pmatrix}$$

Le modèle sur cette classe est:

$$(mi) \quad Y_i = X_i b_i + r_i$$

- On définit une matrice de blocs diagonale:

$$Z = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & X_p \end{pmatrix}$$

ainsi, l'ensemble des p régressions (mi) s'écrira:

$$(m) \quad Y = Zb + r \quad \text{avec } b = (b_{ij})_{pk(1)} \quad \text{et } r = (r_{ij})_{mp(1)}$$

Les OLS sur (m) $\rightarrow$

$$\hat{b} = (Z'Z)^{-1} Z'Y = \begin{pmatrix} (X_1'X_1)^{-1} X_1'Y_1 \\ \vdots \\ (X_p'X_p)^{-1} X_p'Y_p \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

$$Y'Y = \hat{b}'Z'Y + r'r \quad \text{d'où:}$$

$$(6) \quad r'r = Y'Y - \hat{b}'Z'Y$$

- La réduction de RSS en quittant (2) pour (m) est:

$$(7) \quad e'e - r'r = \hat{b}'Z'Y - \hat{\alpha}'D'Y - \hat{\beta}'X'Y = S_3$$

L'analyse complète de covariance devient:

Source	Somme des carrés	ddl	Carrés moyens
X	$s's = Y'Y - \hat{\beta}' X'Y$	mp-k	$s's/mp-k$
X et D	$e'e = Y'Y - \hat{\alpha}' D'Y - \hat{\beta}' X'Y = S_3 + S_4 = S_2$	mp-k-(p-1)	$S_2/(mp-k-p+1)$
D	$s's - e'e = \hat{\alpha}' D'Y + \hat{\beta}' X'Y - \hat{\beta}' X'Y = S_1$	p-1	$S_1/(p-1)$
Z	$r'r = Y'Y + b'Z'Y = S_4$	p(m-k)	$S_4/p(m-k)$
	$e'e - r'r = b'Z'Y - \hat{\alpha}' D'Y - \hat{\beta}' X'Y = S_3$	pk-k-(p-1)	$S_3/pk-k-(p-1)$

Les différents tests utilisés sont:

i) Test de termes constants différents

$$(T_1) \quad F(1) = \frac{S_1/(p-1)}{S_2/(mp-k-p+1)}$$

ii) Test de vecteurs pentes différents

$$(T_2) \quad F(2) = \frac{S_3/(pk-k-p+1)}{S_4/p(m-k)}$$

iii) Test d'homogénéité (effet de classe)

$$(T_3) \quad F(3) = \frac{(S_1 + S_3)/k(p-1)}{S_4/p(m-k)}$$

8). Résumé

1) Les tests de covariance sont basés sur les comparaisons par paires des 3 régressions que sont:

(i). Le modèle général

$$Y = X\beta + u \rightarrow \hat{Y} = X\hat{\beta} \rightarrow X = (X_{ij})_{n,k}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

$$s's = y'y - \hat{\beta}' x' y = (1 - R^2) y'y \text{ à } mp-k \text{ ddl}$$

(ii). Les modèles de groupe

$$Y_i = X_i \beta + u \rightarrow \hat{Y}_i = X_i \hat{\beta} \quad i = \overline{1, p}$$

$$(r'r)_i = (1 - r_i^2) y'_i y_i \quad \text{à } m-k \quad \text{ddl}$$

$$\Rightarrow r'r = \sum_i (r'r)_i \quad \text{à } p(m-k) \quad \text{ddl}$$

(iii). Le modèle général de groupes

$$Y = D\alpha + X\beta + u \rightarrow \hat{Y} = D\hat{\alpha} + X\hat{\beta}$$

$$e'e = Y'Y - \hat{Y}'\hat{Y}$$

$$= Y'Y - (Y'D \quad Y'X) \begin{pmatrix} D'D & D'X \\ X'D & X'X \end{pmatrix}^{-1} \begin{pmatrix} D'Y \\ X'Y \end{pmatrix} \text{à } mp-k-(p-1) \text{ ddl}$$

2) Les tests proprement dits sont basés sur les effets différentiels de ces 3 types de modèles:

(T1). La variabilité des termes constants

On passe du modèle (1) au modèle (2)

$$F_{(1)} = \frac{(s's - e'e)/(p-1)}{e'e/[mp-k-(p-1)]} = \frac{S_1/(p-1)}{S_2/(mp-k-p+1)}$$

(T2). L'homogénéité des pentes

On passe du modèle (2) au modèle (i)

$$F_{(2)} = \frac{(e'e - r'r)/(p-1)(k-1)}{r'r/p(m-k)} = \frac{S_3/(p-1)(k-1)}{S_4/p(m-k)}$$

(T3). L'homogénéité des régresseurs: (termes constants et pentes)

On passe de (1) à (i).

$$F_{(3)} = \frac{(s's - r'r)/k(p-1)}{r'r/p(m-k)} = \frac{(S_1 + S_3)/k(p-1)}{S_4/p(m-k)}$$

Note Dans beaucoup d'applications pratiques, $p = 2$. Alors, tant que le nombre d'observations (m) dans chaque classe excède le nombre de paramètres à estimer (k), les tests de ci-dessus sont équivalents au test de Chow. Sinon (i.e. $m < k$), ils correspondent au test de stabilité des régresseurs.

Exemple25. Faites l'analyse de covariance, après spécification du modèle, sur la répartition suivante de 30 pays en trois groupes selon un ensemble d'indicateurs économiques

Pays	PNB	Pop	%tnPop	%Urbain
Groupe I				
BWA	2230	1277	3.3	25
BFA	270	9016	2.8	15
BDI	210	5492	2.9	5
CMR	910	11524	3.1	40
MLI	200	8460	2.9	24
CAF	410	3008	2.8	47
TCD	180	5680	2.5	32
COG	1000	2276	3.3	41
MOZ	80	15707	2.6	27
TGO	410	3638	3.8	29
Groupe II				
CAN	20360	26522	1.1	77
DNK	22470	5140	0.1	85
FRA	19460	56735	0.5	73
DEU	22350	79452	0.9	85
HKG	11950	5705	0.3	94
ITA	16950	57661	0.2	69
GBR	16030	57411	0.3	89
USA	21970	150372	0.9	75
NLD	17890	14952	0.7	89
NOR	22840	4242	0.4	75
Groupe III				
KOR	5450	42869	1.1	72
MYS	2360	17763	2.3	43
MUS	2240	1075	1.1	41
MEX	2640	81724	1.7	73
MAR	980	25091	2.3	46
NGA	340	96203	2.8	35
THA	1470	56303	1.5	22
POL	1680	38119	0.4	62
ZAF	2420	37959	2.4	49
TUN	1430	8074	2.1	56