

Rappels : Econométrie

Cem Ertur*

8 septembre 2014

— Version 2.0 —

1 Moindres Carrés Ordinaires : Modèle de régression simple

Considérons le modèle de régression simple où y est la variable expliquée et x la variable explicative et supposons que nous disposons de n observations sur le couple (x_i, y_i) . On peut écrire pour l'observation i :

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, \dots, n$$

1.1 Les hypothèses

Les hypothèses du modèle de régression simple sont les suivantes :

Sur la nature de la relation entre les variables :

- H_1 : forme fonctionnelle linéaire stable

Sur la nature de la variable explicative :

- H_{2a} : la variable explicative est non stochastique (ce n'est pas une variable aléatoire, elle est mesurée sans erreurs).
- H_{2b} : il existe au moins 2 valeurs différentes pour les $x_i \implies \sum_i (x_i - \bar{x})^2 > 0$. C'est une condition technique d'identification qui assure l'existence de la solution des MCO dans le cadre de la régression simple : elle est toujours satisfaite.
- H_{2c} : $\lim_{n \rightarrow \infty} \sum_i (x_i - \bar{x})^2 = q > 0$. C'est une condition technique restrictive permettant d'obtenir facilement des résultats asymptotiques importants. Elle assure que la variance de la variable explicative est finie.

Sur les erreurs :

- H_{3a} : $E(\varepsilon_i) = 0 \quad \forall i$ espérance mathématique nulle
- H_{3b} : $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2 \quad \forall i$ homoscedasticité
- H_{3c} : $cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \quad \forall i \neq j$ non corrélation des erreurs

On peut renforcer les hypothèses précédentes en les remplaçant par l'hypothèse suivante où les erreurs sont de plus indépendantes :

*Laboratoire d'Economie d'Orléans (UMR 6221 CNRS), UFR Droit Economie Gestion, Université d'Orléans. E-mail : cem.ertur@univ-orleans.fr, site internet : <http://certur.free.fr/>.

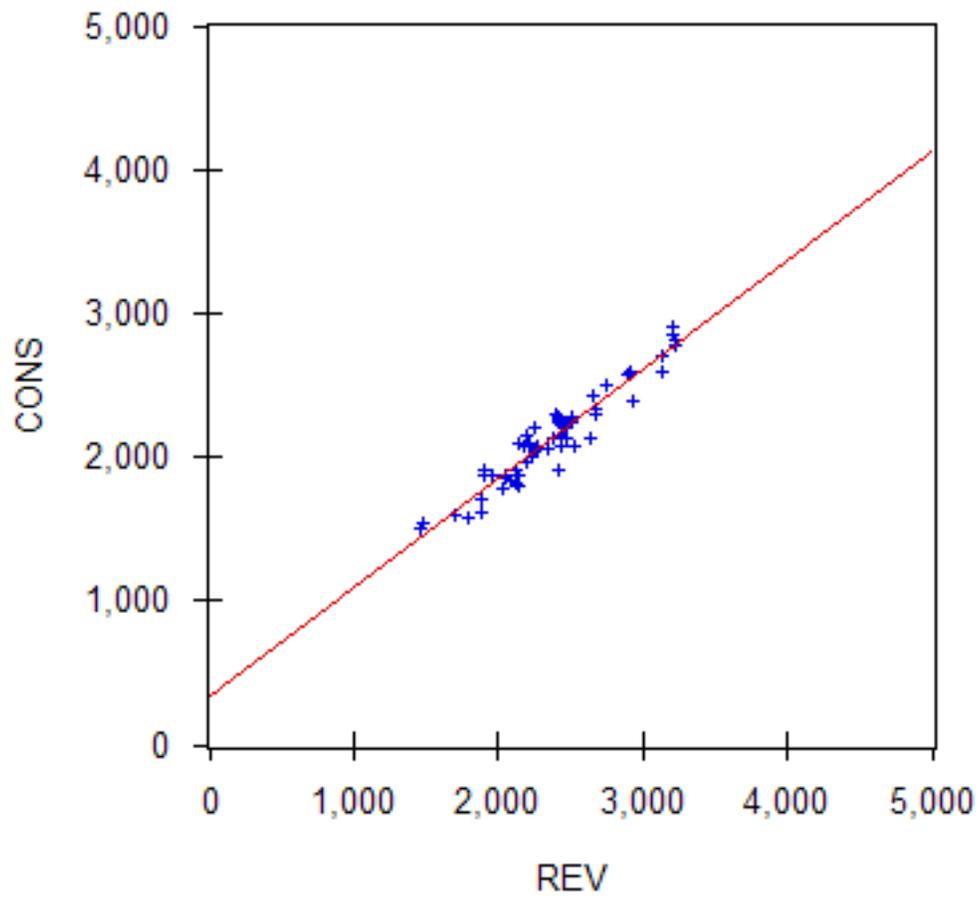


FIGURE 1 – Fonction de consommation keynésienne

- $H_{3d} : \varepsilon_i \sim i.i.d.(0, \sigma^2) \quad \forall i$ les erreurs sont identiquement, indépendamment distribuées

De plus, on peut supposer la normalité de la distribution des erreurs :

- $H_{3e} : \varepsilon_i \sim N(0, \sigma^2) \quad \forall i$ les erreurs sont distribués suivant la loi Normale

Sur la nature de la relation entre la variable explicative et les erreurs :

- $H_4 : cov(x_i, \varepsilon_j) = E(x_i \varepsilon_j) = 0 \quad \forall i, \forall j$ non corrélation entre la variable expliquée et le terme d'erreur (ou hypothèse plus forte d'indépendance).

Remarque 1 Sur l'hypothèse H_{2a} : cette hypothèse est très restrictive. Elle correspond à une situation expérimentale où on choisit les valeurs de x_i et on observe y_i . Cela serait le cas par exemple dans une expérience agricole où y_i est le rendement et x_i la concentration en engrais. Elle facilite cependant l'analyse dans la mesure où la statistique élémentaire suffit pour obtenir des résultats. Toutefois en sciences sociales, il est clair qu'on sera rarement dans une situation expérimentale et par ailleurs peu de modèles justifient la présence de variables explicatives non aléatoires. Par ailleurs, elle introduit une asymétrie injustifiable dans la relation, dans la mesure où la variable expliquée est aléatoire alors que la variable explicative est supposée ne pas l'être. Pour l'exemple de la fonction de consommation, la consommation agrégée serait aléatoire mais pas le revenu agrégé. Notons cependant que cette hypothèse n'est pas indispensable et on verra qu'elle pourra être relâchée pratiquement sans coût. L'hypothèse crucial est alors l'hypothèse H_4 d'absence de corrélation entre x et ε .

On peut donner une autre interprétation de cette hypothèse en disant qu'on effectue l'analyse conditionnellement à l'échantillon observé. Ceci revient à supposer que le modèle de régression et ces hypothèses s'appliquent aux valeurs particulières observées de x . Le modèle de régression simple est alors interprété comme un modèle conditionnel où chaque valeur de y_i est considérée comme la réalisation d'une variable aléatoire d'espérance conditionnelle et de variance conditionnelle :

$$E(y_i/x_i) = \alpha + \beta x_i \quad \text{et} \quad V(y_i/x_i) = \sigma^2$$

Remarque 2 Sur l'hypothèse H_{2c} : cette hypothèse est en fait trop restrictive dans la mesure où elle exclut un cas très simple qui est celui de la régression sur une tendance : dans ce cas, en effet la variance de la variable explicative n'est pas bornée. Mais cette hypothèse peut être relâchée sans nuire aux résultats asymptotiques (conditions de Grenander, cf. Greene p.295).

Remarque 3 Sur l'hypothèse H_{3a} : cette hypothèse est relativement naturelle puisque ε est interprétée comme la somme d'un grand nombre d'effets positifs et négatifs. Même s'il existe une raison de postuler que l'espérance mathématique de l'erreur est différente de zéro, on peut l'intégrer dans la composante systématique du modèle de régression, laissant seulement dans le terme d'erreur la composante inconnue de ε .

Supposons en effet que $E(\varepsilon_i) = \mu, \forall i$ alors $\alpha + \beta x + \varepsilon = (\alpha + \mu) + \beta x + (\varepsilon - \mu) = \alpha' + \beta x + \varepsilon'$ et $E(\varepsilon') = 0$ comme dans le modèle standard. Il faut toutefois noter qu'alors l'estimateur des MCO est un estimateur biaisé de la constante initiale, le biais étant égal à l'espérance mathématique de ε . Mais ceci ne pose pas de problème particulier. Des problèmes peuvent survenir lorsque l'espérance mathématique de ε n'est pas constante :

$E(\varepsilon_i) = \mu_i$ dans le cas d'un modèle à variable dépendante qualitative ou dans le cas de l'estimation d'une frontière de production (cf. A Guide to Econometrics, chap. 7, P. Kennedy).

Remarque 4 Sur l'hypothèse H_{3b} : c'est l'hypothèse d'homoscédasticité, la variance de ε_i est constante pour tout i . C'est surtout pour les modèles en coupe transversale que cette hypothèse risque de ne pas être appropriée. Considérons en effet un modèle établissant par exemple une relation entre le profit des entreprises d'un secteur industriel et la taille de ces entreprises. Même en tenant compte de la taille, mesurée en CA, les profits des grandes entreprises présentera une variation plus importante que ceux des petites entreprises. Cette hypothèse ne sera alors pas appropriée. De même, si l'on considère la fonction de consommation sur données en coupe transversale, on trouvera de l'hétéroscédasticité même après avoir tenu compte du revenu et de la taille des ménages. Cette hypothèse risque là aussi de ne pas être appropriée.

Remarque 5 Sur l'hypothèse H_{3c} : non auto-corrélation des erreurs. Cette hypothèse ne pose généralement pas de problème sur données en coupe transversale, les données provenant d'enquêtes sur des individus tirés aléatoirement dans la population, l'hypothèse d'indépendance semble naturelle. Il faut par contre s'en inquiéter dans les modèles sur données localisées en coupe transversale et sur données temporelles. Il se peut que les ε_i soient corrélés d'une observation à l'autre, c'est-à-dire que chaque erreur ait tendance à être suivie d'une erreur de même signe, reflétant une sorte d'inertie, dans le premier cas on parle d'autocorrélation spatiale, dans le deuxième d'autocorrélation temporelle. Il existe de nombreuses méthodes permettant de traiter l'autocorrélation des erreurs, nous en verrons quelques unes, de même que les conséquences du non respect de cette hypothèse. L'hypothèse H_{3d} remplace l'absence de corrélation par l'hypothèse plus forte d'indépendance des erreurs. L'hypothèse H_{3d} implique les hypothèses précédentes et en particulier H_{3c} .

Remarque 6 Sur l'hypothèse H_{3e} : l'hypothèse H_{3e} est plus forte que H_{3d} et implique l'indépendance des erreurs, on écrit alors pour résumer les hypothèses H_3 : $\varepsilon_i \sim i.i.dN(0, \sigma^2)$. L'hypothèse de normalité est nécessaire pour calculer la distribution exacte des estimateurs et des statistiques utilisées dans les tests. Mais il faut noter que la majorité des résultats des MCO ne requiert pas cette hypothèse de Normalité qu'on pourra relâcher sans coût trop important grâce au théorème central-limit. Par contre elle est nécessaire pour l'utilisation de la méthode du maximum de vraisemblance.

Remarque 7 Etant donné les hypothèses sur les erreurs et pour x_i non stochastique, on peut calculer l'espérance mathématique, la variance et la covariance de y_i :

$$\begin{aligned} E(y_i) &= \alpha + \beta x_i \quad \forall i \\ V(y_i) &= V(\varepsilon_i) = \sigma^2 \quad \forall i \\ cov(y_i, y_j) &= E \left[\underbrace{(y_i - E(y_i))}_{\varepsilon_i} \underbrace{(y_j - E(y_j))}_{\varepsilon_j} \right] = E(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j \end{aligned}$$

De même l'hypothèse de normalité sur les erreurs va impliquer la normalité des y_i : $y_i \sim N(\alpha + \beta x_i, \sigma^2)$. Pour x_i stochastique, tous ces moments de même que la distribution des y_i sont conditionnés par les x_i .

Remarque 8 Sur l'hypothèse H_4 : il est évident que cette hypothèse est redondante quand la variable explicative est non stochastique. En effet dans ce cas $cov(x_i, \varepsilon_j) = E(x_i \varepsilon_j) = x_i E(\varepsilon_j) = 0$ pour tout i et j étant donné que $E(\varepsilon_i) = 0$ par H_{3a} . Elle n'est à considérer que dans le cas où on relâche l'hypothèse H_{2a} comme souligné précédemment.

1.2 Les estimateurs des MCO

Sous cet ensemble d'hypothèses, l'objectif est maintenant d'estimer les paramètres inconnus de ce modèle de régression simple : α, β et σ^2 en supposant que toutes les hypothèses précédentes sont satisfaites, sur la base d'un échantillon de taille n :

$$y_i = \alpha + \beta x_i + \varepsilon_i = E(y_i/x_i) + \varepsilon_i \quad i = 1, \dots, n$$

Il faut distinguer les paramètres inconnus α et β , des paramètres estimés $\hat{\alpha}$ et $\hat{\beta}$. Pour la population la régression s'écrit $E(y_i/x_i) = \alpha + \beta x_i$, alors que l'estimation de $E(y_i/x_i)$ sera notée :

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

L'erreur concernant l'observation i s'écrit :

$$\varepsilon_i = y_i - \alpha - \beta x_i$$

tandis que l'erreur estimée s'écrit :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

on a donc :

$$y_i = \alpha + \beta x_i + \varepsilon_i = \hat{\alpha} + \hat{\beta} x_i + \hat{\varepsilon}_i$$

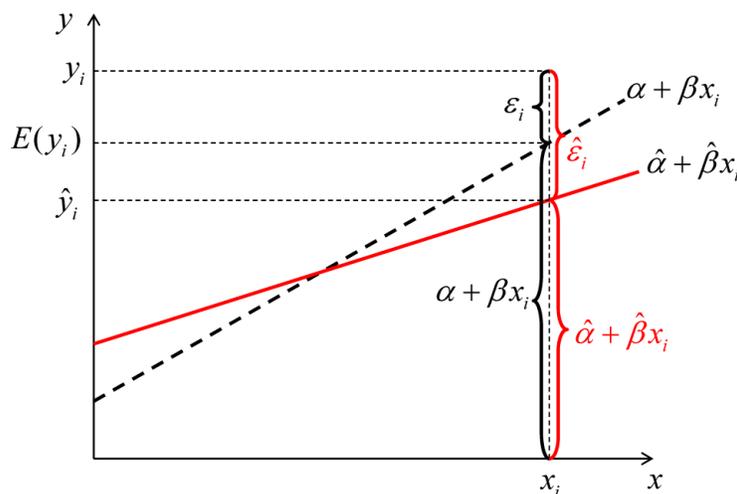


FIGURE 2 – Modèle de régression simple

Les estimateurs de α et β sont obtenus en appliquant la méthode des Moindres Carrés qui consiste à minimiser la somme des carrés des erreurs :

$$\min_{\alpha, \beta} SS = \min_{\alpha, \beta} \sum_{i=1}^n \varepsilon_i^2 = \min_{\alpha, \beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

Les conditions du premier ordre s'écrivent :

$$\begin{cases} \frac{\partial SS}{\partial \alpha} = -2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \implies \sum_{i=1}^n \hat{\varepsilon}_i = 0 \\ \frac{\partial SS}{\partial \beta} = -2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0 \implies \sum_{i=1}^n \hat{\varepsilon}_i x_i = 0 \end{cases}$$

La première équation implique la nullité de la somme des erreurs estimées tandis que la seconde équation implique la nullité de la somme pondérée par les x_i des erreurs estimées. Ces équations peuvent s'écrire de la manière suivante pour donner les équations normales :

$$\begin{cases} \sum_{i=1}^n y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^n x_i \\ \sum_{i=1}^n y_i x_i = \hat{\alpha} \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2 \end{cases}$$

On divise les deux membres de la première équation par n , on obtient :

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

où \bar{y} et \bar{x} sont respectivement les moyennes arithmétiques de y_i et x_i . La droite des moindres carrés ordinaires passe donc par le point moyen de coordonnées (\bar{x}, \bar{y}) . On tire α de l'équation précédente et on l'introduit dans la seconde équation pour avoir la solution pour β en notant que $\sum_i^n x_i = n\bar{x}$.

$$\begin{aligned} \sum_{i=1}^n y_i x_i &= (\bar{y} - \hat{\beta}\bar{x}) \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n y_i x_i &= \bar{y} \sum_{i=1}^n x_i - \hat{\beta}\bar{x} \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n y_i x_i - n\bar{y}\bar{x} &= \hat{\beta} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \\ \hat{\beta} &= \frac{\sum_{i=1}^n y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \end{aligned}$$

La solution est alors :

$$\hat{\beta} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{V(x)} \quad \text{et} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Remarque 9 On n'a pas besoin d'étudier les conditions du second ordre car la convexité de la fonction objectif SS nous garantit qu'il s'agit d'un minimum global strict.

Définissons de nouvelles notations. Sachant que :

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i - \bar{x})y_i - \sum_{i=1}^n (x_i - \bar{x})\bar{y} = \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y} \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} \\ \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (y_i - \bar{y})x_i - \sum_{i=1}^n (y_i - \bar{y})\bar{x} = \sum_{i=1}^n (y_i - \bar{y})x_i - \bar{x} \underbrace{\sum_{i=1}^n (y_i - \bar{y})}_{=0}\end{aligned}$$

par conséquent, définissons m_{xy} comme :

$$m_{xy} = \sum_i x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n (y_i - \bar{y})x_i$$

De même, définissons m_{xx} et m_{yy} :

$$\begin{aligned}m_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i \\ m_{yy} &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \bar{y})y_i\end{aligned}$$

Dans cette notation, les estimateurs des MCO de α et β s'écrivent :

$$\hat{\beta} = \frac{m_{xy}}{m_{xx}} \quad \text{et} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Remarque 10 Condition nécessaire et suffisante pour l'existence de la solution des MCO : il faut avoir au moins 2 valeurs différentes pour les x_i :

$$m_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$$

En effet, $\sum_{i=1}^n (x_i - \bar{x})^2 = 0 \Leftrightarrow (x_i - \bar{x})^2 = 0, \forall i \Leftrightarrow x_i = \bar{x}, \forall i$

Remarque 11 L'essentiel est que la relation soit linéaire dans les paramètres α et β , pas nécessairement dans les variables. Nous pouvons donc appliquer les MCO sur des formes fonctionnelles non linéaires dans les variables qui par une transformation appropriée deviennent linéaires dans les paramètres.

1.3 Transformations de formes fonctionnelles non linéaires

1.3.1 Schéma exponentiel : taux de variation constant

Soit les fonctions exponentielles suivantes :

$$\begin{aligned}y &= ba^x & \text{ou} & & y &= e^{ax+b} \\ y &= be^{x \ln a}\end{aligned}$$

Calculons la dérivée de y par rapport à x et les taux de variation instantanés de y par rapport à x :

$$\begin{aligned} \frac{\partial y}{\partial x} &= (\ln a)be^{x \ln a} & \text{ou} & \quad \frac{\partial y}{\partial x} = ae^{ax+b} \\ \frac{\partial \ln y}{\partial x} = \frac{\partial y}{\partial x} \cdot \frac{1}{y} &= \ln a = \text{cte} & \text{ou} & \quad \frac{\partial y}{\partial x} \cdot \frac{1}{y} = a = \text{cte} \end{aligned}$$

Ce schéma convient très souvent pour décrire l'évolution d'une variable en fonction du temps sous l'hypothèse d'un taux de variation constant dans le temps : par exemple, $y = e^{at+b}$ où $t = 1, 2, \dots$ représente le temps.

On effectue une transformation logarithmique :

$$\ln y = \ln b + x \ln a \quad \text{ou} \quad \ln y = ax + b$$

On pose $z = \ln y$, $\alpha = \ln a$, $\beta = \ln b$ et on obtient $z = \alpha x + \beta$ ou $z = ax + b$. On estime alors le modèle linéarisé par les MCO. On retrouve ensuite les coefficients estimés initiaux \hat{a} et \hat{b} par la transformation inverse.

1.3.2 Schéma puissance : élasticité constante

Soit les fonctions puissances suivantes :

$$\begin{aligned} y &= bx^a & \text{ou} & \quad y = be^{a \ln x} \\ y &= e^{a \ln x + b} \end{aligned}$$

Calculons la dérivée de y par rapport à x et l'élasticité de y par rapport à x :

$$\begin{aligned} \frac{\partial y}{\partial x} &= abx^{a-1} = \frac{ay}{x} \\ \frac{\partial y}{\partial x} \cdot \frac{x}{y} &= a = \text{cte} \end{aligned}$$

Ce schéma convient très souvent pour représenter des phénomènes pour lesquels l'élasticité de y par rapport à x est constante. On effectue une transformation logarithmique :

$$\ln y = \ln b + a \ln x$$

On pose $\tilde{y} = \ln y$, $\tilde{x} = \ln x$ et $\beta = \ln b$ et on obtient $\tilde{y} = a\tilde{x} + \beta$. On estime alors le modèle linéarisé par les MCO. On retrouve ensuite le coefficient estimé initial \hat{b} par la transformation inverse.

1.4 Propriétés des estimateurs des MCO en échantillon de taille finie

P1 Les estimateurs des MCO sont linéaires par rapport à y_i

$$\begin{aligned} \hat{\beta} &= \frac{m_{xy}}{m_{xx}} = \frac{1}{m_{xx}} \sum_i (x_i - \bar{x}) y_i = \sum_i \frac{(x_i - \bar{x})}{m_{xx}} y_i = \sum_i w_i y_i & \text{avec } w_i &= \frac{(x_i - \bar{x})}{m_{xx}} \\ \hat{\alpha} &= \bar{y} - \hat{\beta} \bar{x} = \bar{y} - \bar{x} \sum_i w_i y_i = \sum_i \left(\frac{1}{n} - \bar{x} w_i \right) y_i = \sum_i v_i y_i & \text{avec } v_i &= \frac{1}{n} - \bar{x} w_i \end{aligned}$$

Propriétés des poids de la forme linéaire :

- $\sum_i w_i = 0$, car $\sum_i (x_i - \bar{x}) = 0$;
- $\sum_i w_i x_i = 1$, car $\sum_i (x_i - \bar{x}) x_i = m_{xx}$;
- $\sum_i w_i^2 = \frac{1}{m_{xx}}$, car $\sum_i \left(\frac{x_i - \bar{x}}{m_{xx}}\right)^2 = \frac{m_{xx}^2}{m_{xx}}$;
- $\sum_i v_i = 1$, car $\sum_i \left(\frac{1}{n} - \bar{x} w_i\right) = 1 - \bar{x} \underbrace{\sum_i w_i}_{=0}$;
- $\sum_i v_i x_i = 0$, car $\sum_i \left(\frac{1}{n} - \bar{x} w_i\right) x_i = \bar{x} - \bar{x} \underbrace{\sum_i w_i x_i}_{=1}$;
- $\sum_i v_i^2 = \frac{1}{n} + \frac{\bar{x}^2}{m_{xx}}$, car $\sum_i \left(\frac{1}{n} - \bar{x} w_i\right)^2 = \frac{1}{n} + \bar{x}^2 \underbrace{\sum_i w_i^2}_{=1/m_{xx}} - \frac{2}{n} \bar{x} \underbrace{\sum_i w_i}_{=0} = \frac{1}{n} + \frac{\bar{x}^2}{m_{xx}}$;
- $\sum_i w_i v_i = \frac{-\bar{x}}{m_{xx}}$, car $\sum_i w_i \left(\frac{1}{n} - \bar{x} w_i\right) = \frac{1}{n} \underbrace{\sum_i w_i}_{=0} - \bar{x} \underbrace{\sum_i w_i^2}_{=1/m_{xx}}$

P2 Les estimateurs des MCO sont centrés

$$E(\hat{\beta}) = E\left(\sum_i w_i y_i\right) = \sum_i w_i E(y_i) = \sum_i w_i (\alpha + \beta x_i) = \alpha \underbrace{\sum_i w_i}_{=0} + \beta \underbrace{\sum_i w_i x_i}_{=1} = \beta$$

$$E(\hat{\alpha}) = E\left(\sum_i v_i y_i\right) = \sum_i v_i E(y_i) = \sum_i v_i (\alpha + \beta x_i) = \alpha \underbrace{\sum_i v_i}_{=1} + \beta \underbrace{\sum_i v_i x_i}_{=0} = \alpha$$

P3 Calcul des vraies variances en termes du paramètre σ^2 inconnu :

$$V(\hat{\beta}) = V\left(\sum_i w_i y_i\right)_{y_i \text{ non corrélés}} = \sum_i V(w_i y_i) = \sum_i w_i^2 \underbrace{V(y_i)}_{\sigma^2} = \sigma^2 \sum_i w_i^2 = \frac{\sigma^2}{m_{xx}}$$

$$V(\hat{\alpha}) = V\left(\sum_i v_i y_i\right)_{y_i \text{ non corrélés}} = \sum_i V(v_i y_i) = \sum_i v_i^2 \underbrace{V(y_i)}_{\sigma^2} = \sigma^2 \sum_i v_i^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{m_{xx}}\right)$$

$$\text{cov}(\hat{\beta}, \hat{\alpha}) = \text{cov}\left(\sum_i w_i y_i, \sum_i v_i y_i\right) = \sigma^2 \sum_i w_i v_i = -\sigma^2 \frac{\bar{x}}{m_{xx}}$$

P4 Les estimateurs des MCO sont BLUE (théorème de Gauss-Markov)

a) Dérivation directe de l'estimateur BLUE

Définissons un estimateur linéaire :

$$\beta^* = \sum_i c_i y_i$$

Calculons son espérance mathématique et sa variance :

$$E(\beta^*) = E\left(\sum_i c_i y_i\right) = \sum_i c_i E(y_i) = \sum_i c_i (\alpha + \beta x_i) = \alpha \sum_i c_i + \beta \sum_i c_i x_i$$

$$V(\beta^*) = V\left(\sum_i c_i y_i\right) = \sum_i c_i^2 V(y_i) = \sigma^2 \sum_i c_i^2$$

Imposons à cet estimateur linéaire d'être centré :

$$E(\beta^*) = \beta \implies \alpha \sum_i c_i + \beta \sum_i c_i x_i = \beta \quad \forall \alpha, \beta \implies \sum_i c_i = 0 \text{ et } \sum_i c_i x_i = 1$$

Les contraintes pour que β^* soit centré sont donc :

- (1) $\sum_i c_i = 0$
- (2) $\sum_i c_i x_i = 1$

Cherchons l'estimateur à variance minimale parmi les estimateurs linéaires centrés :
Le problème à résoudre est donc :

$$\min_{c_i} \sum_i V(\beta^*) = \min_{c_i} \sigma^2 \sum_i c_i^2 = \min_{c_i} \sum_i c_i^2$$

$$s.c. \begin{cases} \sum_i c_i = 0 \\ \sum_i c_i x_i = 1 \end{cases}$$

On construit le Lagrangien :

$$\mathbf{L}(c_i, \lambda_1, \lambda_2) = \sum_i c_i^2 + 2\lambda_1 \left(\sum_i c_i\right) + 2\lambda_2 \left(1 - \sum_i c_i x_i\right)$$

Les conditions du premier ordre s'écrivent :

$$\frac{\partial \mathbf{L}}{\partial c_i} = 2c_i + 2\lambda_1 - 2\lambda_2 x_i = 0 \quad i = 1, \dots, n \quad (1)$$

$$\frac{\partial \mathbf{L}}{\partial \lambda_1} = 2 \sum_i c_i = 0 \quad (2)$$

$$\frac{\partial \mathbf{L}}{\partial \lambda_2} = 2 \left(1 - \sum_i c_i x_i\right) = 0 \quad (3)$$

$$c_i = \lambda_2 x_i - \lambda_1 \quad i = 1, \dots, n$$

en sommant (1) sur i , on obtient :

$$\sum_i c_i = \lambda_2 \sum_i x_i - n\lambda_1 = 0$$

par ailleurs en multipliant (1) par x_i , on obtient :

$$x_i c_i = \lambda_2 x_i^2 - \lambda_1 x_i \quad i = 1, \dots, n$$

en sommant sur i , on obtient :

$$\sum_i x_i c_i = \lambda_2 \sum_i x_i^2 - \lambda_1 \sum_i x_i = 1$$

on obtient donc le système : $\begin{cases} \sum_i c_i = \lambda_2 \sum_i x_i - n\lambda_1 = 0 \\ \lambda_1 = \lambda_2 \bar{x} \\ \sum_i x_i c_i = \lambda_2 \sum_i x_i^2 - \lambda_1 \sum_i x_i = 1 \end{cases}$

on obtient : $\lambda_1 = \lambda_2 \bar{x}$, on remplace dans la deuxième équation :

$$\lambda_2 \sum_i x_i^2 - \lambda_2 \bar{x} \sum_i x_i \quad \Longrightarrow \quad \lambda_2 = \frac{1}{\sum_i x_i^2 - \bar{x} \sum_i x_i} = \frac{1}{\sum_i x_i^2 - n\bar{x}^2} = \frac{1}{m_{xx}}$$

Enfin :

$$c_i = \lambda_2 x_i - \lambda_1 = \lambda_2 x_i - \lambda_2 \bar{x} = \lambda_2 (x_i - \bar{x}) = \frac{x_i - \bar{x}}{m_{xx}} = w_i$$

Donc :

$$\beta^* = \sum_i w_i y_i = \hat{\beta}$$

Donc l'estimateur des moindres carrés est BLUE.

b) Montrons que $V(\hat{\beta}) < V(\tilde{\beta})$ pour $\tilde{\beta}$ tout autre estimateur linéaire centré de β

$$\tilde{\beta} = \sum_i d_i y_i$$

avec :

$$E(\tilde{\beta}) = \alpha \sum_i d_i + \beta \sum_i d_i x_i$$

$$V(\tilde{\beta}) = \sigma^2 \sum_i d_i^2$$

Les contraintes pour que $\tilde{\beta}$ soit centré sont donc (1) $\sum_i d_i = 0$ et (2) $\sum_i d_i x_i = 1$, d_i est arbitraire pourvu que (1) et (2) soient satisfaits.

Sans perte de généralité, on peut écrire :

$$d_i = w_i + e_i$$

avec w_i les poids des moindres carrés et e_i arbitraire.

On en déduit :

$$\sum_i d_i = 0 \quad \Longrightarrow \quad \underbrace{\sum_i w_i}_{=0} + \sum_i e_i = 0 \quad \Longrightarrow \quad \sum_i e_i = 0$$

$$\sum_i d_i x_i = 1 \quad \Longrightarrow \quad \underbrace{\sum_i w_i x_i}_{=1} + \sum_i e_i x_i = 1 \quad \Longrightarrow \quad \sum_i e_i x_i = 0$$

On calcule la variance de $\tilde{\beta}$:

$$\begin{aligned}
 V(\tilde{\beta}) &= \sigma^2 \sum_i d_i^2 \\
 &= \sigma^2 \sum_i (w_i + e_i)^2 \\
 &= \sigma^2 \sum_i (w_i^2 + 2w_i e_i + e_i^2) \\
 &= \underbrace{\sigma^2 \sum_i w_i^2}_{=V(\hat{\beta})} + 2\sigma^2 \underbrace{\sum_i w_i e_i}_{=0} + \sigma^2 \sum_i e_i^2 \\
 &= V(\hat{\beta}) + \sigma^2 \sum_i e_i^2 \geq V(\hat{\beta})
 \end{aligned}$$

On montre en effet que :

$$\sum_i w_i e_i = \sum_i \frac{(x_i - \bar{x})}{m_{xx}} e_i = \frac{1}{m_{xx}} \left(\underbrace{\sum_i x_i e_i}_{=0} - \bar{x} \underbrace{\sum_i e_i}_{=0} \right) = 0$$

par conséquent on a bien montré que les estimateurs des moindres carrés sont BLUE.

1.5 Estimateur de σ^2

Si les erreurs ε_i étaient observables, un estimateur évident de σ^2 serait :

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \implies E(\tilde{\sigma}^2) = \frac{1}{n} \sum_{i=1}^n E(\varepsilon_i^2) = \sigma^2$$

Puisque les erreurs ε_i ne sont pas observables, on utilisera les erreurs estimées $\hat{\varepsilon}_i$.

$$\begin{aligned}
 \hat{\varepsilon}_i &= y_i - \hat{y}_i \\
 &= y_i - (\hat{\alpha} + \hat{\beta}x_i), \quad \text{on remplace } \hat{\alpha} \text{ par son expression} \\
 &= y_i - (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i) \\
 &= (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})
 \end{aligned}$$

Considérons la somme des erreurs estimées :

$$\sum_{i=1}^n \hat{\varepsilon}_i = \sum_{i=1}^n (y_i - \bar{y}) - \hat{\beta} \sum_{i=1}^n (x_i - \bar{x}) = 0 \iff \sum_{i=1}^n \hat{\varepsilon}_i = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{y}_i \iff \sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$$

Considérons la somme pondérée des erreurs estimées :

$$\sum_{i=1}^n \hat{\varepsilon}_i x_i = \sum_{i=1}^n (y_i - \bar{y})x_i - \hat{\beta} \sum_{i=1}^n (x_i - \bar{x})x_i = m_{xy} - \hat{\beta}m_{xx} = m_{xy} - \frac{m_{xy}}{m_{xx}}m_{xx} = 0$$

Les erreurs estimées ε_i sont orthogonales aux valeurs prises par la variable explicative x_i .

Calculons la somme des carrés des erreurs estimées :

$$\begin{aligned} SSE &= \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2\hat{\beta} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \\ &= m_{yy} + \hat{\beta}^2 m_{xx} - 2\hat{\beta} m_{xy} \end{aligned}$$

Or $\hat{\beta} m_{xy} = \hat{\beta}^2 m_{xx}$ car $\hat{\beta} = m_{xy}/m_{xx}$, par conséquent $m_{xy} = \hat{\beta} m_{xx}$, donc :

$$\begin{aligned} SSE &= m_{yy} + \hat{\beta}^2 m_{xx} - 2\hat{\beta}^2 m_{xx} \\ &= m_{yy} - \hat{\beta}^2 m_{xx} \end{aligned}$$

Mais aussi :

$$\begin{aligned} SSE &= m_{yy} + \hat{\beta} m_{xy} - 2\hat{\beta} m_{xy} \\ &= m_{yy} - \hat{\beta} m_{xy} \end{aligned}$$

Calculons l'espérance mathématique de SSE :

$$\begin{aligned} SSE &= m_{yy} - \hat{\beta}^2 m_{xx} \\ E(SSE) &= E(m_{yy}) - m_{xx} E(\hat{\beta}^2) \\ &= E(m_{yy}) - m_{xx} [V(\hat{\beta}) + E(\hat{\beta})^2] \\ &= E(m_{yy}) - m_{xx} \left[\frac{\sigma^2}{m_{xx}} + \beta^2 \right] \\ &= E(m_{yy}) - \sigma^2 - \beta^2 m_{xx} \end{aligned}$$

Récrivons m_{yy} :

$$\begin{aligned} m_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\alpha + \beta x_i + \varepsilon_i - \alpha - \beta \bar{x} - \bar{\varepsilon})^2 = \sum_{i=1}^n [\beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]^2 \\ &= \beta^2 \sum_{i=1}^n (x_i - \bar{x})^2 + 2\beta \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 \\ &= \beta^2 m_{xx} + 2\beta \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^n \varepsilon_i^2 - n\bar{\varepsilon}^2 \end{aligned}$$

Calculons maintenant l'espérance mathématique de m_{yy} :

$$\begin{aligned} E(m_{yy}) &= \beta^2 m_{xx} + 2\beta \sum_{i=1}^n (x_i - \bar{x}) E(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^n E(\varepsilon_i^2) - nE(\bar{\varepsilon}^2) \\ &= \beta^2 m_{xx} + n\sigma^2 - n \frac{\sigma^2}{n} \\ &= \beta^2 m_{xx} + (n-1)\sigma^2 \end{aligned}$$

Par conséquent, on obtient pour l'espérance mathématique de SSE :

$$\begin{aligned} E(SSE) &= \beta^2 m_{xx} + (n-1)\sigma^2 - \sigma^2 - \beta^2 m_{xx} \\ &= (n-2)\sigma^2 \end{aligned}$$

L'estimateur centré de σ^2 sera par conséquent :

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

On note que SSE est divisé par le nombre d'observations dans l'échantillon n moins le nombre de paramètres à estimer (α et β) : c'est le degré de liberté de SSE .

Propriétés de $\hat{\sigma}^2$

- Par construction, $\hat{\sigma}^2$ est un estimateur centré de σ^2
- Sous l'hypothèse de normalité, on montre que $\hat{\sigma}^2$ est BQUE (Best Quadratic Unbiased Estimator) : c'est l'estimateur à variance minimale dans la classe des estimateurs quadratiques centrés.

Variances, écarts-types et covariance estimés des estimateurs des MCO

$$\begin{aligned} \hat{V}(\hat{\alpha}) &= \frac{\hat{\sigma}^2}{m_{xx}} & \hat{\sigma}_{\hat{\alpha}} &= \sqrt{\hat{V}(\hat{\alpha})} \\ \hat{V}(\hat{\beta}) &= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{m_{xx}} \right) & \hat{\sigma}_{\hat{\beta}} &= \sqrt{\hat{V}(\hat{\beta})} \\ \widehat{cov}(\hat{\alpha}, \hat{\beta}) &= -\hat{\sigma}^2 \frac{\bar{x}}{m_{xx}} \end{aligned}$$

1.6 Qualité de l'ajustement et tableau d'analyse de la variance

Le coefficient de détermination R^2 :

On se pose la question du pouvoir explicatif du modèle de régression : dans quelle mesure les variations de la variable dépendante expliquent-elles la variation de la variable dépendante ? Elle est évaluée à l'aide du coefficient de détermination R^2 fondé sur la décomposition de la variance de la variable expliquée : il mesure la proportion de la variance de y expliquée par la régression. Il s'écrit :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

où SST est la variabilité totale, SSR est la variabilité expliquée par la régression et SSE est la variabilité résiduelle :

$$\begin{aligned} \text{Variabilité totale } SST &= \sum_{i=1}^n (y_i - \bar{y})^2 = m_{yy} \\ \text{Variabilité résiduelle } SSE &= \sum_{i=1}^n \hat{\varepsilon}_i^2 = m_{yy} - \hat{\beta}^2 m_{xx} \\ \text{Variabilité expliquée par la régression } SSR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 m_{xx} \end{aligned}$$

En effet :

$$\left. \begin{array}{l} \hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \\ \bar{\hat{y}}_i = \hat{\alpha} + \hat{\beta}\bar{x} \end{array} \right\} \implies \hat{y}_i - \bar{\hat{y}}_i = \hat{\beta}(x_i - \bar{x})$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}}_i)^2 = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\beta}^2 m_{xx}$$

On vérifie évidemment la formule de décomposition de la variabilité de y :

$$SST = SSR + SSE$$

Le tableau d'analyse de la variance :

Variabilité	source	degrès de liberté	variance
Régression	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = \hat{\beta}^2 m_{xx}$	$k - 1$	$\frac{SSR}{k-1}$
Résiduelle	$SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2 = m_{yy} - \hat{\beta}^2 m_{xx}$	$n - k$	$\frac{SSE}{n-k}$
Totale	$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = m_{yy}$	$n - 1$	$\frac{SST}{n-1}$

Le coefficient de détermination R^2 s'écrit alors pour la régression simple :

$$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}^2 m_{xx}}{m_{yy}} = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{m_{yy}}$$

Le coefficient de corrélation linéaire entre la variable explicative et la variable expliquée s'écrit :

$$\rho_{xy} = \frac{cov(x, y)}{\sqrt{V(x)}\sqrt{V(y)}} = \hat{\beta} \frac{\sqrt{V(x)}}{\sqrt{V(y)}} = \sqrt{R^2}$$

Propriétés et interprétation du coefficient de détermination R^2 :

On montre que $0 \leq R^2 \leq 1$.

En effet, $SSR \geq 0$ et $SST > 0$, ce qui implique que $\frac{SSR}{SST} \geq 0$. Par ailleurs $SSE = SST - SSR \leq SST$, donc $\frac{SSE}{SST} \leq 1$. Par conséquent $0 \leq 1 - \frac{SSE}{SST} \leq 1$, ce qui implique que $0 \leq R^2 \leq 1$. Si $SSE = 0$ on a $SSR = SST$ et $R^2 = 1$, il s'agit d'un ajustement parfait. La variabilité de la variable indépendante explique parfaitement la variabilité de la variable dépendante. Le pouvoir explicatif du modèle de régression est maximal. Si $SSR = 0$ on a $SSE = SST$ et $R^2 = 0$. Notons que $SSR = 0 \implies \hat{\beta}^2 m_{xx} = 0$, or $m_{xx} > 0$ ce qui implique que $\hat{\beta} = 0$. La variabilité de la variable indépendante n'explique aucunement la variabilité de la variable dépendante. Le pouvoir explicatif du modèle est nul.

En pratique, R^2 sera strictement compris entre 0 et 1. Ainsi un $R^2 = 0,80$ signifiera que 80% de la variabilité de la variable dépendante est expliquée par la variabilité de la variable indépendante. Le pouvoir explicatif du modèle de régression pourra être jugé satisfaisant.

Remarque 12 Une condition nécessaire pour que $0 \leq R^2 \leq 1$ est que la matrice X contienne un terme constant. Les logiciels calculent généralement R^2 pour le modèle contenant un terme constant, l'exclure du modèle a pour conséquence l'obtention d'un R^2 incorrect. En pratique, il faut toujours inclure un terme constant dans la régression, dans le cas contraire on peut obtenir soit un R^2 supérieur à 1, soit un R^2 négatif suivant la formule utilisée pour le calculer. Eviews utilise la formule : $R^2 = 1 - \frac{SSE}{SST}$.

Remarque 13 On peut noter que dans des régressions en séries temporelles les R^2 peuvent souvent atteindre des valeurs proches de 1 du fait de la présence de tendances temporelles communes à la variable expliquée et aux variables explicatives. Dans une régression en coupe transversale, c'est rarement le cas et les R^2 peuvent paraître faibles sans que ceci ne signifie nécessairement que le pouvoir explicatif de la spécification adoptée soit insatisfaisant.

2 Moindres Carrés Ordinaires : régression multiple

Considérons le modèle de régression suivant avec k variables explicatives, si on dispose d'un échantillon de taille n , on écrit pour la i ème observation :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, \dots, n$$

Si dans le modèle il y a un terme constant, cela signifie $x_{i1} = 1, \forall i$. k est alors le nombre de paramètres à estimer et il y a $k - 1$ variables explicatives dans le modèle.

Sous forme matricielle, en empilant les n observations, on écrit :

$$y = X\beta + \varepsilon$$

où y est la variable expliquée d'ordre $n \times 1$, X est la matrice des variables explicatives d'ordre $n \times k$, β est le vecteur des coefficients inconnus d'ordre $k \times 1$ et ε est le vecteur d'erreurs d'ordre $n \times 1$.

2.1 Hypothèses

Sur la nature de la relation entre les variables :

- H_1 : forme fonctionnelle linéaire stable

Sur les variables explicatives :

- H_2 : les variables explicatives sont supposées non stochastiques (fixes) ou :
- H_4 : indépendantes du terme d'erreur : $cov(X, \varepsilon) = 0$ (H_4).
- La matrice X est de rang complet : $rg(X) = k < n$ (les colonnes de X sont linéairement indépendantes).
- De plus on suppose $\lim_{n \rightarrow \infty} \frac{1}{n} X'X = Q$, matrice définie positive, pour obtenir les propriétés asymptotiques.

Sur les erreurs :

- $H_{3a} : \varepsilon \sim N(0, \sigma^2 I) \Rightarrow H_{3b} : \varepsilon_i \sim i.i.d. (0, \sigma^2) \Rightarrow H_{3c} : \begin{cases} E(\varepsilon) = 0 \\ E(\varepsilon\varepsilon') = \sigma^2 I \end{cases}$

2.2 Estimateurs des MCO

L'estimateurs de β est obtenus en appliquant la méthode des Moindres Carrés qui consiste à minimiser la somme des carrés des erreurs.

Soit le vecteur d'erreur : $\hat{\varepsilon} = y - X\hat{\beta}$. La somme des carrés des erreurs s'écrit :

$$SS = \hat{\varepsilon}'\hat{\varepsilon} = (y - X\hat{\beta})'(y - X\hat{\beta})$$

Le problème est donc de minimiser cette somme des carrés des erreurs par rapport au vecteur des coefficients inconnus β :

$$\min_{\beta} SS = \min_{\beta} (y - X\hat{\beta})'(y - X\hat{\beta})$$

Développons SS :

$$\begin{aligned} SS &= y'y - \hat{\beta}'X'y - yX\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

En effet $\hat{\beta}'X'y = yX\hat{\beta}$, ces expressions représentant le même scalaire. Notons au passage que $2\hat{\beta}'X'y$ est une forme linéaire en $\hat{\beta}$ et que $\hat{\beta}'X'X\hat{\beta}$ est une forme quadratique en $\hat{\beta}$.

Écrivons les conditions du premier ordre :

$$\frac{\partial SS}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0$$

Si $\text{rg}(X) = k$, $X'X$ est définie positive, donc $X'X$ est non singulière, son inverse existe. On peut donc écrire :

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'y$$

2.2.1 Propriétés en échantillon de taille finie des estimateurs des MCO

- Considérons l'hypothèse H_{3c} , la moins restrictive sur la spécification des erreurs. On postule l'homoscédasticité et l'absence de corrélation des erreurs. On a vu que l'estimateur des MCO, noté $\hat{\beta}_{MCO}$ s'écrit $\hat{\beta}_{MCO} = (X'X)^{-1}X'y$, par conséquent : $\hat{\beta}_{MCO} = Ay$ avec $A = (X'X)^{-1}X'$ et $\hat{\beta}_{MCO}$ est un estimateur linéaire en y .
 $E(\hat{\beta}_{MCO}) = AE(y) = AX\beta = (X'X)^{-1}X'X\beta = \beta$ donc $\hat{\beta}_{MCO}$ est un estimateur centré de β .

$V(\hat{\beta}_{MCO}) = AV(y)A' = \sigma^2AA' = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$ est la matrice des variances-covariances d'ordre $k \times k$ de l'estimateur $\hat{\beta}_{MCO}$.

Si on appelle x^{ij} l'élément d'ordre (i, j) de la matrice $(X'X)^{-1}$:

$$\begin{aligned} V(\hat{\beta}_i) &= \sigma^2 x^{ii} \\ \text{cov}(\hat{\beta}_i, \hat{\beta}_j) &= \sigma^2 x^{ij} \end{aligned}$$

- $\hat{\beta}_{MCO}$, est BLUE (Best Linear Unbiased Estimator) : c'est-à-dire que c'est l'estimateur à variance minimale dans la classe des estimateurs linéaires centrés (Théorème de Gauss-Markov).
- $\hat{\sigma}_{MCO}^2 = \frac{SS}{n-k}$ estimateur centré de σ^2 car $E(SS) = \sigma^2(n-k) \Rightarrow E(\hat{\sigma}_{MCO}^2) = \sigma^2$
En effet la somme des carrés des erreurs estimées SS s'écrit :

$$\begin{aligned} SS &= \hat{\varepsilon}'\hat{\varepsilon} = (y - X\hat{\beta}_{MCO})' (y - X\hat{\beta}_{MCO}) \\ &= y'y - \hat{\beta}'_{MCO}X'y - y'X\hat{\beta}_{MCO} + \hat{\beta}'_{MCO}X'X\hat{\beta}_{MCO} \\ &= y'y - 2\hat{\beta}'_{MCO}X'y + \hat{\beta}'_{MCO}X'X(X'X)^{-1}X'y \\ &= y'y - 2\hat{\beta}'_{MCO}X'y + \hat{\beta}'_{MCO}X'y \\ &= y'y - \hat{\beta}'_{MCO}X'y \end{aligned}$$

Mais aussi :

$$SS = y'M_X y = \varepsilon'M_X \varepsilon$$

où $M_X = I - X(X'X)^{-1}X'$ est une matrice idempotente et symétrique de rang $n - k$. SS est donc une forme quadratique idempotente en un vecteur normal y ou ε .

En effet :

$$\begin{aligned}\hat{\varepsilon} &= y - X\hat{\beta}_{MCO} = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = M_X y \\ &= M_X(X\beta + \varepsilon) = M_X \varepsilon \\ &\quad \text{car } M_X X = 0\end{aligned}$$

Le théorème suivant nous sera utile pour certaines démonstrations :

Théorème 1

Soit Y un vecteur aléatoire tel que $Y \sim N(0, \Omega)$ avec Ω une matrice définie positive. Soit la forme quadratique $q = Y'AY$ avec A une matrice symétrique. Alors :

$$E(q) = \text{tr}(A\Omega) \quad V(q) = 2 \text{tr} A\Omega A\Omega$$

Montrons que $\hat{\sigma}_{MCO}^2$ est un estimateur centré de σ^2 , calculons l'espérance mathématique de SS :

$$E(SS) = E(\text{tr} \varepsilon' M_X \varepsilon) = E(\text{tr} M_X \varepsilon \varepsilon') = \text{tr} M_X E(\varepsilon \varepsilon') = \sigma^2 \text{tr} M_X = \sigma^2 (n - k)$$

Ce qui implique que $\hat{\sigma}_{MCO}^2 = \frac{SS}{n-k}$ est un estimateur centré de σ^2 car $E(\hat{\sigma}_{MCO}^2) = \sigma^2$.

La matrice des variances-covariances estimée de $\hat{\beta}_{MCO}$ s'écrit alors :

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

Calculons la variance de $\hat{\sigma}_{MV}^2$ pour référence ultérieure :

$$V(\hat{\sigma}_{MCO}^2) = 2 \text{tr} \left(\frac{1}{(n-k)} M_X \sigma^2 I_n \frac{1}{(n-k)} M_X \sigma^2 I_n \right) \quad (4)$$

$$V(\hat{\sigma}_{MCO}^2) = \frac{2\sigma^4}{(n-k)^2} \text{tr} M_X = \frac{2\sigma^4}{(n-k)} \quad (5)$$

- Sous l'hypothèse supplémentaire de normalité des erreurs, on montre que :

$$\hat{\beta}_{MCO} \sim N(\beta, \sigma^2 (X'X)^{-1}).$$

$$SS/\sigma^2 \sim \chi_{n-k}^2, \text{ indépendant de } \hat{\beta}_{MCO}.$$

2.3 Qualité de l'ajustement

2.3.1 Le coefficient de détermination R^2 et le tableau d'analyse de la variance

On se pose la question du pouvoir explicatif du modèle de régression : dans quelle mesure les variations des variables indépendantes expliquent-elles la variation de la variable dépendante ? Elle est évaluée à l'aide du coefficient de détermination R^2 fondé sur la décomposition de la variance de la variable expliquée : il mesure la proportion de la variance de y expliquée par la régression. Il s'écrit :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SS}{SST}$$

où SST est la variabilité totale, SSR est la variabilité due à la régression et SSE est la variabilité résiduelle :

$$\begin{aligned} SST &= m_{yy} = \sum_i (y_i - \bar{y})^2 = y'y - n\bar{y}^2 \\ SSR &= \sum_i (\hat{y}_i - \bar{y})^2 = \hat{y}'\hat{y} - n\bar{y}^2 = \hat{\beta}'X'X\hat{\beta} - n\bar{y}^2 = \hat{\beta}'X'X(X'X)^{-1}X'y - n\bar{y}^2 \\ &= \hat{\beta}'X'y - n\bar{y}^2 \end{aligned}$$

car, dans le modèle avec constante $\bar{\hat{y}} = \bar{y}$.

$$SSE = SS = \hat{\varepsilon}'\hat{\varepsilon} = y'y - \hat{\beta}'X'y$$

On vérifie évidemment la formule de décomposition de la variabilité de y :

$$SST = SSR + SSE$$

Tableau d'analyse de la variance :

Variabilité	source	degrès de liberté	variance
Régression	$SSR = \hat{\beta}'X'y - n\bar{y}^2$	$k - 1$	$\frac{SSR}{k-1}$
Résiduelle	$SSE = \hat{\varepsilon}'\hat{\varepsilon}$	$n - k$	$\frac{SSE}{n-k}$
Totale	$SST = y'y - n\bar{y}^2$	$n - 1$	$\frac{SST}{n-1}$

Remarque 14 (recherche de spécifications) On montre que le R^2 croît mécaniquement lorsqu'on rajoute des variables explicatives dans le modèle. On peut croire ainsi qu'on améliore le modèle, alors qu'il n'en est rien. C'est pourquoi on calcule souvent un R^2 ajusté du nombre de degrés de liberté :

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

Ce \bar{R}^2 ajusté peut décroître lorsqu'une variable explicative est rajouté au modèle, il peut même être négatif. En effet considérons le cas extrême d'un modèle de régression simple où y et x ne serait pas corrélés, le \bar{R}^2 ajusté serait alors égal à $\frac{-1}{(n-2)}$. Pour que \bar{R}^2 augmente, il faut que la contribution de la variable explicative supplémentaire à l'ajustement de la variable expliquée fasse plus que compenser la perte d'un degré de liberté supplémentaire. Le résultat suivant peut être démontré : dans une régression multiple, \bar{R}^2 décroît lorsqu'une variable explicative est éliminée de la régression avec une statistique t supérieure à 1 et croît lorsqu'une variable explicative est éliminée de la régression avec une statistique t inférieure à 1. Le critère de la maximisation du \bar{R}^2 conduit donc à une démarche totalement différente de la démarche qui consiste à ne retenir une variable explicative dans une régression que si, par exemple, le coefficient associé est significatif au seuil de 5%.

Remarque 15 D'une manière générale la recherche du meilleur R^2 , ou même \bar{R}^2 ne constitue pas l'objectif primordial à atteindre du fait des remarques précédentes. On préfère souvent se concentrer sur les résultats de l'inférence statistique, c'est-à-dire la significativité des coefficients associés aux diverses variables explicatives pour évaluer qualité d'une régression. Le R^2 (ou \bar{R}^2) doit être présenté à titre indicatif plutôt que comme une mesure indiscutable de la qualité de l'ajustement.

Remarque 16 Une autre manière de concevoir la qualité de l'ajustement serait d'utiliser le carré du coefficient de corrélation linéaire entre y et \hat{y} :

$$R^2 = \frac{[\text{cov}(y, \hat{y})]^2}{V(y)V(\hat{y})} = \frac{[\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2}{[\sum_i (y_i - \bar{y})^2] [\sum_i (\hat{y}_i - \bar{\hat{y}})^2]}$$

2.3.2 Critères d'information : Akaike et Schwarz

On peut utiliser le critère de maximisation du \bar{R}^2 dans la recherche d'une spécification satisfaisante, on peut montrer d'ailleurs que ce critère n'est autre que celui de la minimisation de la variance des erreurs estimées $\hat{\sigma}^2$. Cependant, il semblerait que \bar{R}^2 n'apporte pas une correction suffisante en termes de degrés de liberté. On peut à la place utiliser le critère de minimisation d'un critère d'information. Deux critères sont utilisés dans les logiciels :

- Le critère d'information d'Akaike (1973) :

$$AIC = \ln\left(\frac{SS}{n}\right) + \frac{2k}{n}$$

- Le critère d'information de Schwarz (1978) :

$$SC = BIC = \ln\left(\frac{SS}{n}\right) + \frac{k}{n} \ln(n)$$

- Le critère d'information de Hannan-Quinn (1979) :

$$HQ = \ln\left(\frac{SS}{n}\right) + \frac{2k}{n} \ln[\ln(n)]$$

On retiendra alors la spécification qui minimise le critère d'Akaike et/ou de Schwarz. Cependant ces critères n'échappent pas non plus à certaines critiques et doivent être considérés conjointement avec l'inférence statistique dans le cas de modèles emboîtés. Dans le cas de modèles non emboîtés, d'autres procédures de test de spécification sont envisageables.

2.4 Estimateurs du maximum de vraisemblance

Sous l'hypothèse $H_{3d} : \varepsilon \sim N(0, \sigma^2 I)$, la fonction de densité de probabilité de ε qui est la fonction de densité de probabilité jointe normale multivariée de l'échantillon s'écrit :

$$f(\varepsilon|\sigma^2) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \varepsilon' \varepsilon\right] = L(\sigma^2|\varepsilon)$$

Cette fonction considérée comme fonction du paramètre σ^2 est appelée fonction de vraisemblance pour ε .

La transformation inverse de ε en y est donnée par $\varepsilon = y - X\beta = H^{-1}(y)$, la matrice jacobienne de la transformation inverse étant ici la matrice identité, $J = \frac{\partial \varepsilon}{\partial y'} = I_n$ et son

le jacobien étant par conséquent égale à 1 ($\det J = 1$), la fonction de vraisemblance pour y s'écrit alors :

$$L(\beta, \sigma^2 | y, X) = f(H^{-1}(y)) \cdot |\det J|$$

$$L(\beta, \sigma^2 | y, X) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \right]$$

Il est souvent plus facile de travailler sur le logarithme de la fonction de vraisemblance :

$$\ln L(\beta, \sigma^2 | y, X) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \quad (6)$$

Les estimateurs du maximum de vraisemblance sont obtenus en maximisant cette fonction de log-vraisemblance par rapport aux paramètres β et σ^2 . La fonction logarithme étant une fonction monotone croissante, il est équivalent de maximiser la fonction de vraisemblance ou la fonction de log-vraisemblance.

Nous sommes donc confrontés à un problème d'optimisation libre :

$$\underset{\beta, \sigma^2}{Max} \ln L(\beta, \sigma^2 | y, X)$$

Les conditions nécessaires s'écrivent :

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\beta) = 0 \quad (7)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)'(y - X\beta) = 0 \quad (8)$$

Les estimateurs du maximum de vraisemblance sont alors solutions de ce système d'équations qu'on appelle les équations de vraisemblance :

$$\hat{\beta}_{MV} = (X'X)^{-1} X'y \quad (9)$$

$$\hat{\sigma}_{MV}^2 = \frac{SS}{n} \quad (10)$$

Définition 1 (Score) Soit Y un vecteur aléatoire de dimension $n \times 1$ et θ un vecteur de paramètres inconnus de dimension $k \times 1$, on appelle score le vecteur de dimension $k \times 1$ des dérivées partielles premières de la log-vraisemblance par rapport à θ , autrement dit le gradient de la log-vraisemblance, il s'agit d'un vecteur aléatoire :

$$S(\theta) = \text{grad} \ln L(Y|\theta) = \frac{\partial \ln L(Y|\theta)}{\partial \theta} \quad (11)$$

On montre que, sous certaines conditions de régularité, satisfaites pour la distribution normale (Amemiya, 1985 ; Ruud, 2000) :

- $E[S(\theta)] = E \left[\frac{\partial \ln L(Y|\theta)}{\partial \theta} \right] = 0$, l'espérance mathématique du score est nulle.
- $V[S(\theta)] = E[S(\theta) \cdot S(\theta)'] = E \left[\frac{\partial \ln L(Y|\theta)}{\partial \theta} \cdot \frac{\partial \ln L(Y|\theta)}{\partial \theta'} \right] = -E \left[\frac{\partial^2 \ln L(Y|\theta)}{\partial \theta \partial \theta'} \right]$

Définition 2 (Matrice d'Information) On appelle matrice d'information, de dimension $k \times k$, l'opposé de l'espérance mathématique de la matrice Hessienne de la log-vraisemblance :

$$I(\theta) = -E \left[\frac{\partial^2 \ln L(Y|\theta)}{\partial \theta \partial \theta'} \right] \quad (12)$$

Théorème 2 (Inégalité de Cramer-Rao)

Soit $\tilde{\theta}$ un estimateur centré du vecteur de paramètres inconnus θ de dimension $k \times 1$. Alors, sous certaines conditions de régularité, on montre que la matrice de variance-covariance de tout estimateur centré de θ est borné inférieurement par l'inverse de la matrice d'information de θ :

$$V(\tilde{\theta}) \geq I(\theta)^{-1} \quad \text{autrement dit} \quad V(\tilde{\theta}) - I(\theta)^{-1} \quad \text{est définie non négative.}$$

($k \times k$)

Remarque 17 L'inégalité de Cramer-Rao montre l'existence de la borne inférieure de la variance de n'importe quel estimateur centré, linéaire ou non, indépendamment de la méthode d'estimation utilisée pour l'obtenir.

Si la variance d'un estimateur centré atteint la borne inférieure de l'inégalité de Cramer-Rao, on dit que l'estimateur est efficient : il est centré à variance minimale.

Seulement si un estimateur centré n'atteint pas cette borne inférieure, on ne sait pas (sauf dans certains cas particuliers), si cet estimateur est efficient ou pas.

Propriétés 3

Sous l'hypothèse supplémentaire de normalité des erreurs :

- (a) On constate que $\hat{\beta}_{MCO}$ est l'estimateur du maximum de vraisemblance. On montre qu'il atteint la borne inférieure de l'inégalité de Cramer-Rao. Il est donc efficient parmi tous les estimateurs centrés (linéaires ou non).
- (b) On constate que $\hat{\sigma}_{MCO}^2$ est proportionnel à l'estimateur du maximum de vraisemblance : $\hat{\sigma}_{MV}^2 = \frac{SS}{n}$ qui est biaisé, et $\hat{\sigma}_{MCO}^2 = \frac{n}{n-k} \hat{\sigma}_{MV}^2$.
On montre que $\hat{\sigma}_{MCO}^2$ n'atteint pas la limite inférieure de l'inégalité de Cramer-Rao. On admettra cependant que $\hat{\sigma}_{MCO}^2$ est efficient parmi tous les estimateurs centrés quadratiques : il est BQUE (Best Quadratic Unbiased Estimator) sous l'hypothèse de normalité des erreurs.

Démontrons maintenant (a). Soit $\theta = (\beta, \sigma^2)$ le vecteur, de dimension $k \times 1$, des paramètres inconnus et soit $\hat{\theta}_{MV} = (\hat{\beta}_{MV}, \hat{\sigma}_{MV}^2)$ l'estimateur du maximum de vraisemblance.

Calculons la matrice Hessienne de la log-vraisemblance (6) :

$$\frac{\partial^2 \ln L(Y|\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{\partial^2 \ln L(Y|\theta)}{\partial \beta \partial \beta'} & \frac{\partial^2 \ln L(Y|\theta)}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 \ln L(Y|\theta)}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \ln L(Y|\theta)}{\partial^2 \sigma^2} \end{bmatrix} \quad (13)$$

($k+1$) \times ($k+1$) ($k \times k$) ($k \times 1$)
($1 \times k$) (1×1)

$$\frac{\partial^2 \ln L(Y|\theta)}{\partial \beta \partial \beta'} = -\frac{1}{\sigma^2} X' X \quad (14)$$

$$\frac{\partial^2 \ln L(Y|\theta)}{\partial^2 \sigma^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} (y - X\beta)' (y - X\beta) \quad (15)$$

$$\frac{\partial^2 \ln L(Y|\theta)}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} X' (y - X\beta) \quad (16)$$

$$\frac{\partial^2 \ln L(Y|\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} -\frac{1}{\sigma^2} X'X & -\frac{1}{\sigma^4} X'\varepsilon \\ \frac{(k \times k)}{(k \times 1)} & \frac{(k \times 1)}{(1 \times 1)} \\ -\frac{1}{\sigma^4} \varepsilon'X & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \varepsilon'\varepsilon \\ \frac{(k+1) \times (k+1)}{(1 \times k)} & \frac{(1 \times 1)}{(1 \times 1)} \end{bmatrix} \quad (17)$$

La matrice d'information s'écrit alors :

$$I(\theta) = -E \begin{bmatrix} -\frac{1}{\sigma^2} X'X & -\frac{1}{\sigma^4} X'\varepsilon \\ \frac{(k \times k)}{(k \times 1)} & \frac{(k \times 1)}{(1 \times 1)} \\ -\frac{1}{\sigma^4} \varepsilon'X & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \varepsilon'\varepsilon \\ \frac{(k+1) \times (k+1)}{(1 \times k)} & \frac{(1 \times 1)}{(1 \times 1)} \end{bmatrix} \quad (18)$$

Sachant que nous supposons que X est fixe, et que par hypothèse $E(X'\varepsilon) = 0$ et $E(\varepsilon'\varepsilon) = n\sigma^2$, nous obtenons :

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} X'X & \mathbf{0} \\ \frac{(k \times k)}{(k \times 1)} & \frac{(k \times 1)}{(1 \times 1)} \\ \mathbf{0} & \frac{n}{2\sigma^4} \\ \frac{(k+1) \times (k+1)}{(1 \times k)} & \frac{(1 \times 1)}{(1 \times 1)} \end{bmatrix} \quad (19)$$

La borne inférieure de l'inégalité de Cramer-Rao peut alors être obtenue en inversant la matrice d'information :

$$I(\theta)^{-1} = \begin{bmatrix} \sigma^2 (X'X)^{-1} & \mathbf{0} \\ \frac{(k \times k)}{(k \times 1)} & \frac{(k \times 1)}{(1 \times 1)} \\ \mathbf{0} & \frac{2\sigma^4}{n} \\ \frac{(k+1) \times (k+1)}{(1 \times k)} & \frac{(1 \times 1)}{(1 \times 1)} \end{bmatrix} \quad (20)$$

Ainsi l'estimateur centré $\hat{\beta}_{MCO} \equiv \hat{\beta}_{MV}$ dont la matrice de variance-covariance est égale à $\sigma^2 (X'X)^{-1}$ atteint la borne inférieure de l'inégalité de Cramer-Rao. Nous avons donc montré que $\hat{\beta}_{MCO} \equiv \hat{\beta}_{MV}$ est un estimateur efficace dans la classe des estimateurs centrés (linéaires ou non).

Démontrons (b). Calculons l'espérance mathématique de $\hat{\sigma}_{MV}^2$:

$$\hat{\sigma}_{MV}^2 = \frac{SS}{n} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} = \frac{\varepsilon' M_X \varepsilon}{n} \text{ forme quadratique en un vecteur normal standard} \quad (21)$$

$$E(\hat{\sigma}_{MV}^2) = \frac{1}{n} \text{tr}(M_X E(\varepsilon'\varepsilon)) = \frac{1}{n} \text{tr}(M_X \sigma^2 I_n) = \frac{\sigma^2}{n} \text{tr}(M_X) = \frac{(n-k)}{n} \sigma^2 \quad (22)$$

Calculons maintenant le biais de σ_{MV}^2 :

$$B(\hat{\sigma}_{MV}^2) = E(\hat{\sigma}_{MV}^2) - \sigma^2 = \frac{(n-k)}{n} \sigma^2 - \sigma^2 = -\frac{k}{n} \sigma^2 < 0 \quad (23)$$

Le biais de $\hat{\sigma}_{MV}^2$ est négatif : on sous-estime systématiquement la vraie variance σ^2 . L'estimateur du maximum de vraisemblance de σ^2 étant biaisé, l'inégalité de Cramer-Rao ne s'applique pas. On note cependant que ce biais tend vers 0 lorsque n tend vers l'infini : l'estimateur du maximum de vraisemblance de σ^2 est asymptotiquement sans biais.

On remarque que la variance de $\hat{\sigma}_{MCO}^2$ n'atteint pas la borne inférieure de l'inégalité de Cramer-Rao puisque $V(\hat{\sigma}_{MCO}^2) = \frac{2\sigma^4}{(n-k)} > \frac{2\sigma^4}{n}$.

3 Inférence statistique dans le modèle de régression

3.1 Motivation

On a établi les propriétés de l'estimateur des MCO : $\widehat{\beta}_{MCO}$ dans le cadre de la régression multiple. Il nous faut maintenant préciser comment utiliser cet estimateur pour réaliser des tests d'inférence statistique sur β , le vecteur de paramètres inconnus. Ces procédures d'inférence statistiques jouent un rôle primordial dans l'évaluation des implications testables des théories économiques.

Considérons une théorie économique qui implique que le facteur primordial affectant le comportement d'investissement soit le taux d'intérêt réel, comment cette implication peut-elle faire l'objet d'un test ? On pourrait considérer d'abord le modèle simple où l'investissement dépend du taux d'intérêt nominal et du taux d'inflation :

$$\ln I_t = \beta_1 + \beta_2 \ln r_t + \beta_3 \ln p_t + \varepsilon_t \quad (1)$$

L'implication précédente apparaît alors comme une restriction imposée au modèle (1) : $\beta_2 + \beta_3 = 0$. Cette restriction constitue une hypothèse pouvant faire l'objet d'un test d'inférence statistique. Par ailleurs, elle peut être intégrée directement dans le modèle (1) qu'on pourrait alors reformuler de la manière suivante :

$$\ln I_t = \beta_1 + \beta_2(\ln r_t - \ln p_t) + \varepsilon_t \quad (2)$$

le modèle (1) est alors appelé le modèle libre ou non contraint et le modèle (2) le modèle contraint.

On peut donc dégager à partir de cet exemple deux approches possibles pour s'attaquer au problème du test d'hypothèses :

- la première approche consiste à se demander si les estimations obtenues dans le modèle (1) satisfont raisonnablement la restriction imposée par l'hypothèse nulle ;
- la deuxième approche consiste à se demander si la somme des carrés des erreurs estimées dans le modèle contraint (2) est raisonnablement plus grande que la somme des carrés des erreurs estimées dans le modèle non contraint (1), sachant qu'avec une contrainte on fait toujours moins bien que sans la contrainte sauf dans le cas où la contrainte est satisfaite automatiquement et donc inopérante (cf. optimisation).

Considérons quelques exemples d'hypothèses qu'il peut être intéressant de tester :

- $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$, on fait donc l'hypothèse que la variable explicative X_i n'a pas d'influence sur la variable dépendante Y : c'est le test de significativité bien connu (test bilatéral).
- $H_0 : \beta_i = \beta_{i0}$ contre $H_1 : \beta_i \neq \beta_{i0}$, ici β_{i0} est une valeur donnée a priori du paramètre. Si par exemple β_i représente une élasticité prix, on peut vouloir tester si cette élasticité est égale à -1 ou pas (test bilatéral).
- $H_0 : \beta_2 + \beta_3 = 1$ contre $H_1 : \beta_2 + \beta_3 \neq 1$, si β_2 est l'élasticité du travail et β_3 l'élasticité du capital dans une fonction de production, on peut vouloir tester la nature des rendements d'échelle : constants ou pas.
- $H_0 : \beta_3 = \beta_4$ contre $H_1 : \beta_3 \neq \beta_4$, ici on veut tester l'hypothèse que les variables explicatives X_3 et X_4 exercent la même influence sur Y .

- $H_0 : \beta_2 = \dots = \beta_k = 0$ contre $H_1 : \text{au moins un des } \beta_i \neq 0 \text{ pour } i = 2, \dots, k$, ici on teste conjointement la significativité des coefficients associés aux variables explicatives sauf la constante, c'est le test de la significativité de la régression dans sa globalité.
- $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$, ici on partage β en deux sous-vecteurs β_1 et β_2 contenant respectivement k_1 et $k_2 = k - k_1$ éléments, l'hypothèse à laquelle on s'intéresse est alors qu'un sous-ensemble des variables explicatives introduites dans la régression n'influence pas la variable dépendante Y .
- $H_0 : \beta_i \geq \beta_{i0}$ contre $H_1 : \beta_i < \beta_{i0}$, (β_{i0} peut être éventuellement égal à zéro), c'est un test qu'on peut adopter si, par exemple, on n'est pas prêt d'accepter que la demande est élastique ($\beta_i < \beta_{i0}$) sauf si les données nous présente une preuve raisonnable du contraire (test unilatéral à gauche).
- $H_0 : \beta_i \leq \beta_{i0}$ contre $H_1 : \beta_i > \beta_{i0}$, (β_{i0} peut être éventuellement égal à zéro), c'est un test qu'on peut adopter dans le cas contraire (test unilatéral à droite).

Remarque 18 Du fait que l'échantillon est aléatoire, la statistique du test est elle-même aléatoire. La même procédure de test peut conduire à des conclusions différentes dans différents échantillons. La mise en oeuvre d'un test peut conduire à retenir une hypothèse conforme à la réalité, mais on peut aussi commettre deux types d'erreur : l'erreur de première espèce : on rejette l'hypothèse nulle à tort et l'erreur de seconde espèce : on accepte l'hypothèse nulle à tort.

- on appelle seuil du test ou niveau de significativité, noté α , la probabilité de rejeter l'hypothèse nulle à tort (risque de 1ère espèce).
- on appelle puissance d'un test la probabilité de rejeter l'hypothèse nulle avec raison : puissance = $1 - \beta$, où β est la probabilité d'accepter l'hypothèse nulle à tort (risque de seconde espèce). Le seuil d'un test est la valeur maximale de la puissance du test lorsque l'hypothèse nulle est vraie.

Il est clair que le test idéal minimiserait à la fois α et β . Cependant minimiser α tend à augmenter β et réciproquement : il faut faire un arbitrage entre les deux. La procédure habituelle consiste alors à fixer α et à minimiser β pour cette valeur du seuil, ce qui revient à maximiser la puissance du test pour un seuil donné. Pour un seuil α , un test est dit sans biais si la probabilité de rejeter l'hypothèse nulle est plus grande lorsque cette est hypothèse est fausse que lorsqu'elle est vraie : si sa puissance est supérieure ou égale à son seuil pour toute valeur admissible du paramètre. On dit qu'un test est convergent si sa puissance tend vers 1 lorsque la taille de l'échantillon tend vers l'infini. Pour un seuil donné, on dit qu'un test est uniformément plus puissant s'il a une plus grande puissance que tout autre test au même seuil pour toute valeur admissible du paramètre.

Rappelons que l'hypothèse nulle joue par construction un rôle privilégié dans la procédure de test, puisqu'on ne peut la rejeter que si l'échantillon apporte une preuve raisonnable et suffisante du contraire. La procédure de test semble ainsi conservatrice, on conserve l'hypothèse nulle sauf si les données conduisent à la rejeter : analogie avec la présomption d'innocence : tout suspect est présumé innocent et on doit apporter la preuve de sa culpabilité avant que la justice ne décide de le condamner. Quand on accepte l'hypothèse nulle, on ne prouve pas qu'elle est vraie, on accepte de la conserver car on n'a pas pu accumuler suffisamment de preuve contre elle. Accepter l'hypothèse nulle c'est en fait "acquitter faute de preuve". C'est pour cette raison qu'on préfère dire "ne pas rejeter" au lieu de dire "accepter" l'hypothèse nulle. C'est aussi pour cette raison qu'il faut faire

d'autant plus attention à la spécification des hypothèses nulles et alternatives dans les procédures de tests.

La démarche scientifique procède du même mode de fonctionnement : l'hypothèse nulle représente la théorie largement admise par la communauté scientifique, on la conserve sauf si une expérience répétable prouve qu'elle est fausse (au moins dans certaines conditions). On avance donc par un mécanisme de réfutation des théories par l'expérimentation.

3.2 Réalisation des tests d'hypothèses

C'est l'hypothèse de normalité des erreurs qui nous permet d'effectuer des tests d'inférence statistique dans le modèle de régression. L'hypothèse de normalité de ε implique en effet la normalité de y qui à son tour implique la normalité de $\hat{\beta}_{MCO}$ étant donné sa linéarité en y .

$$H_{3a} : \varepsilon \sim N(0, \sigma^2 I) \implies y \sim N(X\beta, \sigma^2 I) \implies \hat{\beta}_{MCO} \sim N(\beta, \sigma^2(X'X)^{-1})$$

3.2.1 Tests individuels

Sous l'hypothèse de normalité des erreurs, c'est-à-dire l'hypothèse $H_{3a} : \varepsilon \sim N(0, \sigma^2 I)$, la statistique de test pour les tests individuels (bilatéraux et unilatéraux) s'écrit :

$$T = \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\frac{SS}{\sigma^2} x^{ii}}} = \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\hat{\sigma}^2 x^{ii}}} = \frac{\hat{\beta}_i - \beta_{i,0}}{\hat{\sigma}_{\hat{\beta}_i}} \sim t_{n-k}$$

où x^{ii} est l'élément en position (i, i) de $(X'X)^{-1}$. Elle est distribuée sous l'hypothèse nulle suivant la loi de Student à $n - k$ degrés de liberté.

En effet on a d'une part :¹

$$\hat{\beta}_{MCO} \sim N(\beta, \sigma^2(X'X)^{-1}) \implies \hat{\beta}_i \sim N(\beta_{i,0}, \sigma^2 x^{ii}) \implies \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\sigma^2 x^{ii}}} \sim N(0, 1)$$

Si σ^2 était connue, cette statistique pourrait servir de base à l'inférence, mais cela n'est pas le cas. Il faut estimer σ^2 . On peut utiliser l'estimateur centré $\hat{\sigma}^2 = \frac{SS}{n-k}$ à cette fin et construire une statistique faisant intervenir $\hat{\sigma}^2$ au lieu de σ^2 pour servir de base à l'inférence.

On note en effet que $\frac{SS}{\sigma^2} = \tilde{\varepsilon}' M_X \tilde{\varepsilon}$ avec $\tilde{\varepsilon} = \frac{1}{\sigma} \varepsilon \implies \tilde{\varepsilon} \sim N(0, I)$ est un vecteur Normal standard. $\frac{SS}{\sigma^2}$ est donc une forme quadratique idempotente en un vecteur normal standard $\tilde{\varepsilon}$

1. Cette démonstration nécessite l'utilisation des 2 théorèmes de statistique mathématiques suivants :

Théorème 1 (Rao, 1973, p.186) :

Une CNS pour qu'une forme quadratique $Y'AY$ suive la loi du χ^2 est que la matrice A soit idempotente ($A^2 = A$) auquel cas le nombre de degrés de liberté du χ^2 est égal au $rg(A) = tr(A)$

Théorème 2 (Rao, 1973, p.187) :

Soit $Y'A_1Y \sim \chi_a^2$ et $Y'A_2Y \sim \chi_b^2$. Une CNS pour que ces 2 formes quadratiques soit indépendamment distribuées est que $A_1A_2 = 0$.

et suit par conséquent une loi du χ^2 à un nombre de degré de liberté égal à $\text{tr}(M_X) = n - k$.

$$\frac{SS}{\sigma^2} \sim \chi_{n-k}^2$$

Par ailleurs $\widehat{\beta}_{MCO}$ est une forme linéaire en $\tilde{\varepsilon}$:

$$\begin{aligned}\widehat{\beta}_{MCO} &= (X'X)^{-1} X'y = (X'X)^{-1} X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1} X'\varepsilon = \beta + \sigma (X'X)^{-1} X'\tilde{\varepsilon} \\ \frac{\widehat{\beta}_{MCO} - \beta}{\sigma} &= (X'X)^{-1} X'\tilde{\varepsilon}\end{aligned}$$

où $(X'X)^{-1} X'$ est la matrice de la forme linéaire.

On montre que la matrice de la forme linéaire $(X'X)^{-1} X'$ s'annule en produit avec la matrice de la forme quadratique M_X :

$$(X'X)^{-1} X'M_X = 0 \quad \text{car } X'M_X = 0$$

La forme linéaire et la forme quadratique sont donc indépendante, c'est-à-dire que $\frac{SS}{\sigma^2}$ est indépendant de $\widehat{\beta}_{MCO}$. Par conséquent la statistique T est bien distribuée suivant une loi de Student à $n - k$ degrés de liberté (v.a. normale centrée réduite au numérateur et racine carrée d'une v.a. du χ_{n-k}^2 divisée par son degré de liberté au dénominateur, les deux v.a. étant indépendantes).

- Test bilatéral au seuil de $\alpha\%$

On teste l'hypothèse nulle $H_0 : \beta_i = \beta_{i,0}$ contre l'hypothèse alternative $H_A : \beta_i \neq \beta_{i,0}$.
On calcule la statistique de test :

$$|t_{obs}| = \left| \frac{\widehat{\beta}_i - \beta_{i,0}}{\widehat{\sigma}_{\widehat{\beta}_i}} \right| \sim t_{n-k}$$

Règles de décision : Soit $t_{(n-k),\alpha}^*$, la valeur critique de la loi de Student à $n - k$ degrés de liberté au seuil de $\alpha\%$.

si $|t_{obs}| > t_{(n-k),1-\frac{\alpha}{2}}^* \implies$ on rejette l'hypothèse nulle

si $|t_{obs}| < t_{(n-k),1-\frac{\alpha}{2}}^* \implies$ on ne peut rejeter l'hypothèse nulle

Remarque 19 Le test de significativité d'un coefficient de régression est le cas particulier d'un test bilatéral au seuil de $\alpha\%$ avec $\beta_{i,0} = 0$.

- Test unilatéral à droite au seuil de $\alpha\%$

On teste l'hypothèse nulle $H_0 : \beta_i \leq \beta_{i,0}$ contre l'hypothèse alternative $H_A : \beta_i > \beta_{i,0}$.
On calcule la statistique de test :

$$t_{obs} = \frac{\widehat{\beta}_i - \beta_{i,0}}{\widehat{\sigma}_{\widehat{\beta}_i}} \sim t_{n-k}$$

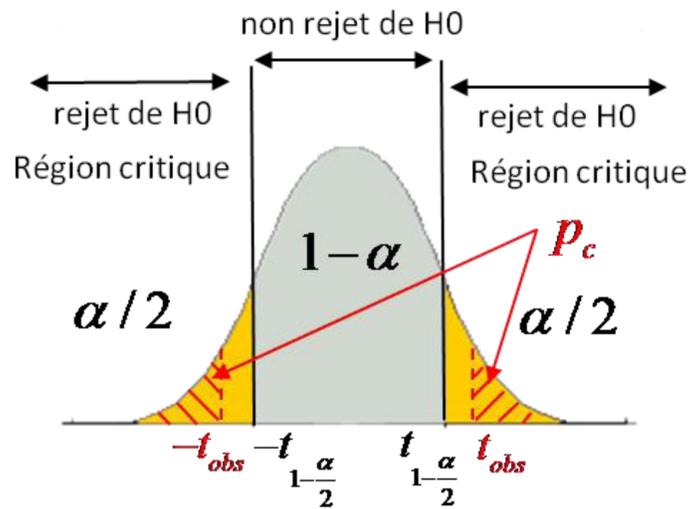
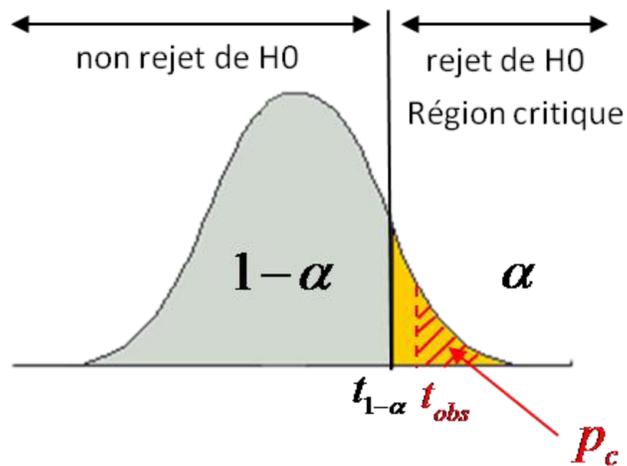
FIGURE 3 – Zone de rejet du test bilatéral au seuil de $\alpha\%$ 

FIGURE 4 – Zone de rejet du test unilatéral à droite

Règles de décision : Soit $t_{(n-k),\alpha}^*$, la valeur critique de la loi de Student à $n-k$ degrés de liberté au seuil de $\alpha\%$.

si $t > t_{(n-k),1-\alpha}^* \implies$ on rejette l'hypothèse nulle

si $t < t_{(n-k),1-\alpha}^* \implies$ on ne peut rejeter l'hypothèse nulle

- Test unilatéral à gauche au seuil de $\alpha\%$

On teste l'hypothèse nulle $H_0 : \beta_i \geq \beta_{i,0}$ contre l'hypothèse alternative $H_A : \beta_i < \beta_{i,0}$.

On calcule la statistique de test :

$$t_{obs} = \frac{\hat{\beta}_i - \beta_{i,0}}{\hat{\sigma}_{\hat{\beta}_i}} \sim t_{n-k}$$

Règles de décision : Soit $t_{(n-k),\alpha}^*$, la valeur critique de la loi de Student à $n-k$ degrés de liberté au seuil de $\alpha\%$.

si $t < -t_{(n-k),\alpha}^* \implies$ on rejette l'hypothèse nulle

si $t > -t_{(n-k),\alpha}^* \implies$ on ne peut rejeter l'hypothèse nulle

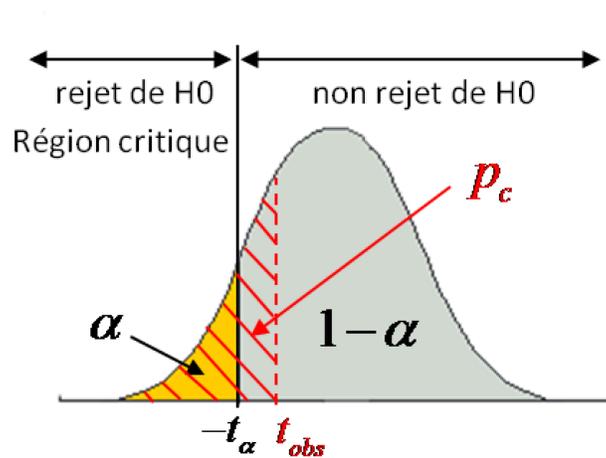


FIGURE 5 – Zone de rejet du test unilatéral à gauche

3.2.2 Probabilité critique ou p-value

On appelle probabilité critique ou p-value d'une statistique de test suivant par exemple la loi Normale centrée réduite, la probabilité que sous l'hypothèse nulle, la variable aléatoire normale centrée réduite prenne en valeur absolue une valeur supérieure ou égale à la valeur calculée de la statistique :

$$p_c(z) = P[|Z| \geq z | H_0] = 2[1 - F(|z|)]$$

où F est la fonction de répartition de la Loi normale centrée réduite.

Dans le cas d'un test bilatéral au seuil de $\alpha\%$, la probabilité critique s'écrit :

$$p_c(t) = P[|T| \geq t | H_0]$$

Tandis que dans le cas d'un test unilatéral à droite au seuil de $\alpha\%$, la probabilité critique s'écrit :

$$p_c(t) = P[T \geq t | H_0]$$

Finalement pour un test unilatéral à gauche au seuil de $\alpha\%$, la probabilité critique s'écrit :

$$p_c(t) = P[T \leq t | H_0]$$

La règle de décision s'écrit alors :

si $p_c(t) < \alpha \implies$ on rejette l'hypothèse nulle

si $p_c(t) > \alpha \implies$ on ne peut rejeter l'hypothèse nulle

Ainsi dans la Figure 2, est illustrée le cas d'un test bilatéral où la probabilité critique p_c , correspondant à la zone hachurée en rouge, est inférieure au seuil du test α . On rejette dans ce cas l'hypothèse nulle $H_0 : \beta_i = \beta_{i,0}$. De même dans la Figure 3 est illustrée le cas d'un test unilatéral à droite où la probabilité critique p_c est également inférieure au seuil du test α . On rejette dans ce cas l'hypothèse nulle $H_0 : \beta_i \leq \beta_{i,0}$. Finalement, dans la Figure 4, est illustrée le cas d'un test unilatéral à gauche où la probabilité critique p_c est supérieure au seuil du test α . On ne peut dans ce cas rejeter l'hypothèse nulle correspondante, $H_0 : \beta_i \geq \beta_{i,0}$.

Remarque 20 Les logiciels d'économétrie présentent systématiquement les statistiques t de significativité individuelle pour chaque coefficient de la régression. Ils donnent aussi les probabilités critiques ou p-values associées aux tests individuels de significativité permettant de faire les tests sans avoir recours à des tables statistiques.

Dans Eviews, la probabilité critique est par exemple présentée dans la colonne prob. des résultats de la régression par les MCO.

3.2.3 Test de l'hypothèse globale de linéarité

Sous l'hypothèse de normalité des erreurs, c'est-à-dire l'hypothèse $H_{3a} : \varepsilon \sim N(0, \sigma^2 I)$, le tableau d'analyse de la variance conduit à un premier test : le test de l'hypothèse globale

de linéarité. On teste en fait l'hypothèse nulle jointe suivant laquelle tous les coefficients de la régression sont nuls excepté le terme constant :

$$\begin{aligned} H_0 & : \beta_2 = \dots = \beta_k = 0 \\ H_1 & : \text{au moins un des } \beta_i \neq 0 \text{ pour } i = 2, \dots, k \end{aligned}$$

Ce test s'effectue à l'aide de la statistique F distribuée, sous l'hypothèse nulle, suivant la loi de Fisher à $k - 1$ et $n - k$ degrés de liberté :

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Si $F > F_{0.05}^*(k-1, n-k)$, on rejette l'hypothèse nulle. Sinon, on ne peut pas la rejeter. Ce test permet donc d'évaluer le pouvoir explicatif global de la régression.

3.2.4 Tests d'un ensemble de restrictions linéaires (test de Wald) : première approche

Tous les tests bilatéraux envisagés peuvent être construits dans le cadre du test d'un ensemble de restrictions linéaires en utilisant le principe de Wald. Intuitivement, on cherche à évaluer si la distance à l'hypothèse nulle est significativement non nulle dans l'échantillon, si c'est le cas on rejettera l'hypothèse nulle, sinon on ne pourra pas la rejeter.

Considérons le test de l'hypothèse nulle $H_0 : R\beta - r = 0$ contre $H_1 : R\beta - r \neq 0$ où R est une matrice de constantes connues de dimension $m \times k$ avec $m < k$ et r est un vecteur de constantes connues de dimension $k \times 1$.

La statistique du test s'écrit alors :

$$F = \frac{(R\hat{\beta} - r)' [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - r) / m}{SS / (n - k)} \sim F_{m, n-k}$$

elle est distribuée sous l'hypothèse nulle suivant la Loi de Fisher à m et $n - k$ degrés de liberté.

Règles de décision : avec $F_{m, n-k}^*$ valeur critique de la loi de Fisher à m et $n - k$ degrés de liberté au seuil de $\alpha\%$:

si $F > F_{m, n-k, 1-\alpha}^*$ ou si $p_c(F) < \alpha \implies$ on rejette l'hypothèse nulle

si $F < F_{m, n-k, 1-\alpha}^*$ ou si $p_c(F) > \alpha \implies$ on ne peut pas rejeter l'hypothèse nulle

Démonstration :

Considérons l'espérance mathématique et la variance de $R\hat{\beta}$:

$$\begin{aligned} E(R\hat{\beta}) & = R\beta \\ V(R\hat{\beta}) & = \sigma^2 R(X'X)^{-1}R' \end{aligned}$$

sous l'hypothèse de normalité des erreurs, c'est-à-dire $H_{3a} : \varepsilon \sim N(0, \sigma^2 I)$, la distribution de $R\hat{\beta}$ est normale comme combinaison linéaire d'une variable aléatoire normale.

$$R\hat{\beta} \sim N(\beta, \sigma^2 R(X'X)^{-1}R')$$

$$R(\hat{\beta} - \beta) \sim N(0, \sigma^2 R(X'X)^{-1} R')$$

Sous l'hypothèse nulle $H_0 : R\beta = r$, on a alors :

$$R\hat{\beta} - r \sim N(0, \sigma^2 R(X'X)^{-1} R')$$

Nous pouvons alors construire une forme quadratique en une variable aléatoire normale centrée réduite qui suit une loi du χ^2 dont le nombre de degrés de liberté est égale à $tr(R(X'X)^{-1} R') = m$ en appliquant le principe de Wald :

$$(R\hat{\beta} - r)' \left[\sigma^2 R(X'X)^{-1} R' \right]^{-1} (R\hat{\beta} - r) \sim \chi_m^2$$

Le seul problème est encore une fois que σ^2 est inconnu, mais on peut l'estimer avec $\hat{\sigma}^2$ et on sait que :

$$\frac{SS}{\sigma^2} \sim \chi_{n-k}^2$$

Or ces deux formes quadratiques sont indépendantes puisqu'on a montré que $\frac{SS}{\sigma^2}$ est indépendant de $\hat{\beta}$, par conséquent la statistique F est distribuée suivant la loi de Fisher à m et $n - k$ degrés de liberté (v.a. du χ_m^2 divisée par son degré de liberté au numérateur et v.a. du χ_{n-k}^2 divisée par son degré de liberté au dénominateur, les deux v.a. étant indépendantes).

On peut toutefois vérifier l'indépendance directement en notant que :

$$\frac{R(\hat{\beta} - \beta)}{\sigma} = R(X'X)^{-1} X' \tilde{\varepsilon}$$

par conséquent la statistique F est bien le rapport de deux formes quadratiques en $\tilde{\varepsilon}$, de plus ces deux formes quadratiques sont bien indépendantes puisque :

$$R(X'X)^{-1} X' M_X = 0 \quad \text{car } X' M_X = 0$$

Dans certain cas, il peut être utile de réécrire la statistique F de la manière suivante :

$$F = \frac{(R\hat{\beta} - r)' \left[R \left[\hat{\sigma}^2 (X'X)^{-1} \right] R' \right]^{-1} (R\hat{\beta} - r)}{m} \sim F_{m, n-k}$$

où $\hat{\sigma}^2 (X'X)^{-1}$ est la matrice de variances-covariances estimée de $\hat{\beta}$.

Remarque 21 Les tests individuels, le test d'une restriction linéaire et le test global de linéarité sont évidemment des cas particuliers du test F précédent.

En ce qui concerne le test de significativité d'un coefficient β_i par exemple, on n'a qu'une seule restriction $H_0 : \beta_i = 0$ et la statistique F s'écrit :

$$F = \frac{\hat{\beta}_i^2}{\hat{\sigma}_{\hat{\beta}}^2} \sim F_{1, n-k}$$

sa racine carrée redonnant le résultat précédent :

$$t = \sqrt{F} = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}_{\hat{\beta}}^2 x^{ii}}} = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}}} \sim t_{n-k}$$

3.2.5 Tests fondés sur les modèle contraint et non contraint : deuxième approche

Par ailleurs tous les tests envisagés peuvent également être construits en spécifiant le modèle contraint intégrant les restrictions, qui correspond au modèle sous l'hypothèse nulle et le modèle non contraint qui correspond au modèle sous l'hypothèse alternative². Intuitivement, on cherche à évaluer si la distance entre la somme des carrés des erreurs estimées dans le modèle contraint (obtenu en imposant les restrictions) et la somme des carrés des erreurs estimées dans le modèle non contraint est significativement non nulle dans l'échantillon, si c'est le cas on rejettera l'hypothèse nulle, sinon on ne pourra pas la rejeter.

On sait qu'avec une contrainte on fait obligatoirement moins bien que sans contrainte (sauf si la contrainte est inopérante). La solution du problème des MCO sous la contrainte $R\beta_c - r = 0$ s'écrit :

$$\begin{aligned} \min_{\beta_c} SS &= \min_{\beta_c} (y - X\beta_c)'(y - X\beta_c) \\ \text{s.c. } R\beta_c - r &= 0 \end{aligned}$$

On écrit le Lagrangien du problème et on résoud les conditions du premier ordre, on obtient (*démonstration à faire*) (G p. 205, JD p. 96-97) :

$$\hat{\beta}_c = \hat{\beta} - (X'X)^{-1} R' [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - r)$$

Soit $\hat{\varepsilon}_c$ l'erreur estimée dans le modèle contraint, soit $\hat{\varepsilon}$ l'erreur estimée dans le modèle non contraint :

$$\hat{\varepsilon}_c = y - X\hat{\beta}_c$$

on peut écrire $\hat{\varepsilon}_c$ de la manière suivante :

$$\hat{\varepsilon}_c = y - X\hat{\beta} - X(\hat{\beta}_c - \hat{\beta}) = \hat{\varepsilon} - X(\hat{\beta}_c - \hat{\beta})$$

La somme des carrés des erreurs estimées dans le modèle contraint peut alors s'écrire :

$$\begin{aligned} \hat{\varepsilon}'_c \hat{\varepsilon}_c &= [\hat{\varepsilon} - X(\hat{\beta}_c - \hat{\beta})]' [\hat{\varepsilon} - X(\hat{\beta}_c - \hat{\beta})] \\ &= \hat{\varepsilon}'\hat{\varepsilon} - \underbrace{\hat{\varepsilon}'X}_{=0} (\hat{\beta}_c - \hat{\beta}) - (\hat{\beta}_c - \hat{\beta})' \underbrace{X'\hat{\varepsilon}}_{=0} + (\hat{\beta}_c - \hat{\beta})' X'X (\hat{\beta}_c - \hat{\beta}) \\ &= \hat{\varepsilon}'\hat{\varepsilon} + \underbrace{(\hat{\beta}_c - \hat{\beta})' X'X (\hat{\beta}_c - \hat{\beta})}_{\geq 0} \geq \hat{\varepsilon}'\hat{\varepsilon} \end{aligned}$$

en remplaçant $\hat{\beta}_c - \hat{\beta}$ par son expression dans la solution des MCO contraints, on obtient :

$$SS_C - SS = \hat{\varepsilon}'_c \hat{\varepsilon}_c - \hat{\varepsilon}'\hat{\varepsilon} = (R\hat{\beta} - r)' [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - r)$$

2. Ceci n'est vrai que pour les restrictions linéaires

qu'on trouve également au numérateur de la statistique F obtenue précédemment qu'on peut donc réécrire de la manière suivante :

$$F = \frac{(SS_C - SS)/m}{SS/(n-k)} \sim F_{m,n-k} \quad \text{Loi de Fisher à } m \text{ et } n-k \text{ degrés de liberté sous } H_0$$

qu'on peut aussi écrire en divisant le numérateur et le dénominateur par $SST = m_{yy}$:

$$F = \frac{(SS_C - SS)/m}{SS/(n-k)} = \frac{(1 - R_c^2 - 1 + R^2)/m}{(1 - R^2)/(n-k)} = \frac{(R^2 - R_c^2)/m}{(1 - R^2)/(n-k)}$$

On retrouve alors la statistique du test de l'hypothèse globale de linéarité, car $R_c^2 = 0$ sous l'hypothèse nulle jointe correspondante. Les règles de décision sont les mêmes que précédemment.