ÉCONOMETRIE THÉORIQUE

Chapitre 1

L'estimation par moindres carrés ordinaires (MCO)

Benoît Mulkay
Université de Montpellier
2023 - 2024

1

Chapitre I :

L'estimation par moindres carrés ordinaires (MCO)

2

I.1. Le modèle de régression linéaire classique

a) Le modèle de régression multiple

HAYASHI [2000], Chapitre I.

L'analyse de régression :

- Modélisation d'une relation pour « expliquer » une **variable dépendante** (régressand), notée y ,
- par des **variables explicatives** (régresseurs) , notées x

$$x \curvearrowright y$$

$$y = f(x)$$

Attention à la notion de causalité

supposée par hypothèse dans cette notation !!!

3

L'échantillon ou les données

- C'est un ensemble d'observations sur plusieurs variables ou caractéristiques.

- Echantillon de N observations indicées : $i = 1 , 2 , 3 , \ldots , N$;
  - Une **coupe instantanée** (données longitudinales) est observée sur différentes unités statistiques à une même période.
  - Elle est en général non ordonnée !

- Echantillon de T observations sur séries temporelles : $t = 1 , 2 , 3 , \ldots , T$.
  - Une **série temporelle** (ou chronologique) est observée sur une même unité statistique pour plusieurs périodes de temps (année, trimestres, jours,…)
  - Une série temporelle a un ordre naturel : passé → présent → futur

4

Les observations

- Ici, une seule variable dépendante (y_i)
- et plusieurs variables explicatives ($x_{k,i}$) : $k = 1 , 2 , 3 , \ldots , K$:

$$\boldsymbol{x}_i = \left(x_{1,i}, x_{2,i}, \ldots, x_{k,i}, \ldots, x_{K,i}\right)'$$

- En général, la première variable explicative est constante : $x_{1,i} = 1$
- Dans la régression simple : $K = 2 \rightarrow \boldsymbol{x}_i = \left(1, x_{2,i}\right)'$
- Dans la régression multiple : $K > 2 \rightarrow \boldsymbol{x}_i = \left(1, x_{2,i}, \ldots, x_{k,i}, \ldots, x_{K,i}\right)'$

- Régression* de y sur les variables explicatives (régresseurs) x_k

* : Francis GALTON (1886) : "Regression Towards Mediocrity in Hereditary Stature“, *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15: 246–263.

GALTON (1822 – 1911) était le cousin de Charles Darwin. Il a été un fondateur de la Société d’Eugénisme (Eugenics Society) qui voulait améliorer la qualité génétique des hommes.

Le modèle

- La variable dépendante (y_i) et les variables explicatives ($x_{k,i}$) sont considérées comme des **variables aléatoires**.

- Les observations sont considérées comme des **réalisations** d’un tirage aléatoire !

- Ici les régresseurs sont aussi des variables aléatoires parce que :
  - On ne contrôle pas les variables explicatives.
  - En sciences sociales et plus particulièrement en économie, on est rarement dans le cadre d’expériences contrôlées *(randomisée)* , mais plutôt avec des *données observationnelles*.
  - Le cas des régresseurs « fixes » est un cas particulier du cas des régresseurs aléatoires.

Un modèle

- Un **modèle économétrique** est une ensemble de restrictions sur la distribution conjointe des variables dépendante et explicatives.

- C'est un ensemble de distributions conjointes qui satisfont un ensemble d'hypothèses.

- Comme la relation n'est pas parfaite, on introduit une variable supplémentaire pour tenir compte d'autres éléments → **l'erreur** : ε !

$$\begin{tikzcd}
x \arrow[dr, bend left] & \\
& y \\
\varepsilon \arrow[ur, bend right] &
\end{tikzcd}$$

Commentaires sur l'erreur : ε_i (error , disturbance)

- L'erreur est aussi appelée aléas ou perturbation.

- Elle correspond à des éléments imprédictibles des aléas des comportements économiques, humains ou sociaux…
- Elle mesure les incertitudes du modèle.
- Elle incorpore des effets inobservables.
- Elle capture les effets d'un grand nombre de variables omises *(voir les hypothèses ci-après).*
- Elle prend en compte les erreurs de mesure sur la variable dépendante.

- L'erreur est une variable aléatoire avec des moments (espérance, variance,…) et une distribution à déterminer.

HYPOTHÈSE 1 : Linéarité du modèle

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i \qquad \text{pour } i = 1,2,\ldots,N.$$

- La relation entre les variables explicatives et la variable dépendante est **linéaire**.
- L'erreur est **additive**.
- Il n'y a pas d'erreurs de mesure sur les variables explicatives $\boldsymbol{x}$.
- → **Le modèle est correctement spécifié !**

HYPOTHÈSE 1 : Linéarité du modèle

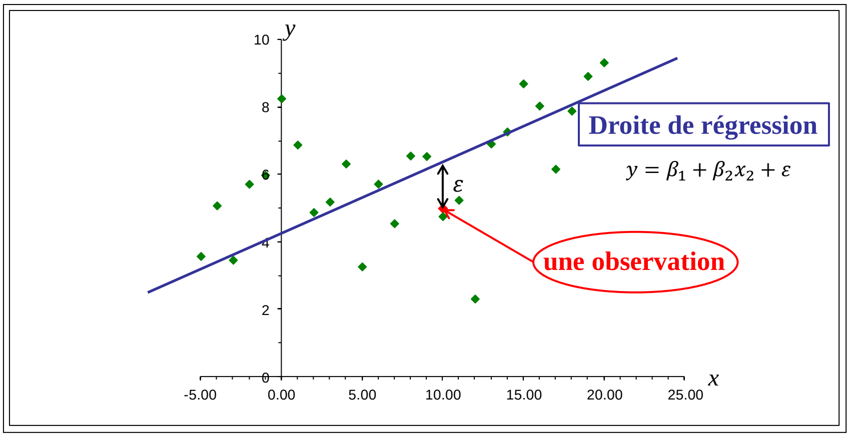
$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i = \sum_{k=1}^{K} \beta_k x_{K,i} + \varepsilon_i$$

- $\beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_K x_{K,i}$ est la **fonction de régression**.
- Les paramètres (ou coefficients) β_k sont les **coefficients de régression**.
- Ils représentent **l'effet marginal**, ***toutes choses égales par ailleurs***, d'une variation de la variable $x_{k,i}$ sur la variable dépendante y_i :

$$\beta_k = \frac{\partial y_i}{\partial x_{k,i}}$$

- La linéarité implique que les effets marginaux sont **constants** pour toutes les observations (individus).

Cas de la régression simple : la droite de régression



Cas de la régression multiple : un (hyper-)plan de régression

11

Commentaires sur la forme linéaire

$$\boxed{y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i}$$

Forme linéaire comme une **approximation locale** d'une forme non-linéaire inconnue…

ou comme une **projection linéaire** …

Modèle linéaire dans les paramètres, pas nécessairement dans les variables.

Celles-ci peuvent être transformées pour retrouver la linéarité.

Exemple : **le modèle double log**

$$y = \lambda x^{\beta} \eta \quad \rightarrow \quad y = \exp(\alpha) \times x^{\beta} \times \exp(\varepsilon) \qquad \textbf{avec} \begin{cases} \lambda = \exp(\alpha) \\ \eta = \exp(\varepsilon) \end{cases}$$

$$\rightarrow \quad \log(y) = \alpha + \beta \log(x) + \varepsilon$$

$$\beta = \frac{\partial log(y)}{\partial log(x)} = \frac{\partial y / y}{\partial x / x} = \frac{\partial y}{\partial x} \times \frac{x}{y}$$

Effet marginal

→ élasticité de y (par rapport) à x.

12

Autre exemple : **le modèle polynomial** :

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_p x^p + \varepsilon$$

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + \cdots + p\beta_p x^{p-1}$$

Ici l'effet marginal dépend de la valeur de la variable explicative x !

Autre exemple : **le modèle à paramètres variables ou avec interaction** :

$$y = (\alpha_1 + \alpha_2 z) + (\beta_1 + \beta_2 z)x + \varepsilon \quad \rightarrow \quad y = \alpha_1 + \alpha_2 z + \beta_1 x + \beta_2 zx + \varepsilon$$

$$\frac{\partial y}{\partial x} = \beta_1 + \beta_2 z$$

Ici l'effet marginal dépend de la valeur de l'autre variable explicative z !

Choix théorique et empirique pour la spécification du modèle

Il faut faire attention à l'interprétation des paramètres : effets marginaux, élasticités, effets constants ou variables,…

Extension aux modèles non linéaires :

$$y_i = f\left(x_{1,i}, x_{2,i}, x_{3,i}, \dots, x_{K,i}\ ;\ \theta\right) + \varepsilon_i$$

$$g\left(y_i, x_{1,i}, x_{2,i}, x_{3,i}, \dots, x_{K,i}\ ;\ \theta\right) = \varepsilon_i$$

Modèles paramétriques avec *f*(.) et / ou *g*(.) connues

Modèles non paramétriques : on doit aussi déterminer la **forme** des fonctions *f*(.) ou *g*(.)

→ *voir un cours d'économétrie non paramétrique…*

b) Notation matricielle du modèle de régression multiple

Utilisation des vecteurs et des matrices

Simplification de la notation → on évite les sommes…

Les données sont des tableaux de nombres

Propriétés du calcul matriciel (algèbre linéaire)

→ simplification des calculs et des démonstrations…

Pour une observation $i = 1,2,\dots,N$:

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i = \sum_{k=1}^{K} \beta_k x_{k,i} + \varepsilon_i$$

$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i$ avec $\boldsymbol{x_i}$ et $\boldsymbol{\beta}$ des vecteurs (colonnes) : $K \times 1$

$\boldsymbol{x}_i' \boldsymbol{\beta}$ → produit scalaire de 2 vecteurs de même dimension.

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

$\boldsymbol{x_i}$: un vecteur (K x 1) des <u>K variables explicatives</u> : $\boldsymbol{x_i} = \begin{pmatrix} 1 \\ x_{2,i} \\ x_{3,i} \\ \vdots \\ x_{K,i} \end{pmatrix}$

$\boldsymbol{\beta}$: un vecteur (K x 1) des <u>K paramètres inconnus</u> : $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix}$

En empilant les N observations de l'échantillon, on obtient le modèle de régression multiple en notation matricielle :

$$\begin{cases} y_1 = \boldsymbol{x}_1' \boldsymbol{\beta} + \varepsilon_1 \\ y_2 = \boldsymbol{x}_2' \boldsymbol{\beta} + \varepsilon_2 \\ \vdots \\ y_N = \boldsymbol{x}_N' \boldsymbol{\beta} + \varepsilon_N \end{cases} \rightarrow \quad \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\underbrace{\boldsymbol{y}}_{(N\times 1)} = \underbrace{\underset{(N\times K)}{\boldsymbol{X}}\underset{(K\times 1)}{\boldsymbol{\beta}}}_{(N\times 1)} + \underset{(N\times 1)}{\boldsymbol{\varepsilon}}$$

$\boldsymbol{y}$ et $\boldsymbol{\varepsilon}$: des vecteurs (N x 1) des variables dépendantes et des erreurs :

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

et $\boldsymbol{X}$: une matrice (N x K) des K variables explicatives :

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_N' \end{pmatrix} = \underbrace{\begin{bmatrix} 1 & x_{2,1} & x_{3,1} & \cdots & x_{K,1} \\ 1 & x_{2,2} & x_{3,2} & \cdots & x_{K,2} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2,N} & x_{3,N} & \cdots & x_{K,N} \end{bmatrix}}_{K \text{ variables}} \left.\vphantom{\begin{bmatrix} 1\\1\\1\\1\\1 \end{bmatrix}}\right\} N \text{ observations}$$

et toujours $\boldsymbol{\beta}$: un vecteur (K x 1) des K paramètres inconnus : $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix}$

HYPOTHÈSE 2 : Stricte exogénéité des régresseurs (H2)

$E(\varepsilon_i|\boldsymbol{X}) = 0$ pour tout $i = 1,2,\dots,N$.
$E(\varepsilon_i|\boldsymbol{x}_1,\boldsymbol{x}_2,\dots\boldsymbol{x}_N) = 0$ pour tout $i = 1,2,\dots,N$.
$E(\varepsilon_i|\boldsymbol{x}_j) = 0$ pour tout i et $j = 1,2,\dots,N$.

L'erreur doit être **exogène** par rapport à **toutes** les observations de **toutes** les variables explicatives.

Si on considère la distribution conjointe des $(NK + N)$ variables aléatoires : $f(\varepsilon_1,\varepsilon_2,\dots,\varepsilon_N,\boldsymbol{x}_1,\boldsymbol{x}_2,\dots\boldsymbol{x}_N)$ pour l'ensemble de l' échantillon,

la distribution conditionnelle pour chaque observation i : $f(\varepsilon_i|\boldsymbol{x}_1,\boldsymbol{x}_2,\dots\boldsymbol{x}_N)$ aura une espérance nulle : $\mathrm{E}(\varepsilon_i|\boldsymbol{x}_1,\boldsymbol{x}_2,\dots\boldsymbol{x}_N) = \mathrm{E}(\varepsilon_i|\boldsymbol{X}) = 0$.

Cette hypothèse est **fondamentale** pour qu'on puisse trouver un estimateur des paramètres $\boldsymbol{\beta}$ qui possède de bonnes propriétés statistiques.

On reviendra sur cette hypothèse dans le Chapitre VI.

Implications de l'hypothèse d'exogénéité stricte :

- **L'espérance inconditionnelle de l'erreur est nulle**

$$E(\varepsilon_i) = 0 \qquad \text{pour tout } i = 1,2,\dots,N\,.$$

par la loi de l'espérance itérée : $E(\varepsilon_i) = E_X[E(\varepsilon_i|\boldsymbol{X})] = 0$

- **Les régresseurs sont orthogonaux aux termes d'erreurs pour toutes les observations.**

Deux variables aléatoires sont dites orthogonales si et seulement si l'espérance de leur produit (scalaire) est égale à zéro !

$$E(x_{j,k}\varepsilon_i) = 0 \qquad \text{pour tout } i,j = 1,2,\dots,N \text{ et } k = 1,2,\dots,K.$$

$$E(\boldsymbol{x}_j\varepsilon_i) = \begin{pmatrix} E(x_{j,1}\varepsilon_i) \\ E(x_{j,2}\varepsilon_i) \\ \vdots \\ E(x_{j,K}\varepsilon_i) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \boldsymbol{0}_K \qquad \text{pour tout } i,j = 1,2,\dots,N\,.$$

DÉMONSTRATION

Comme $x_{j,k}$ est un élément de $\boldsymbol{X}$, la stricte exogénéité implique :

$$E(\varepsilon_i|x_{j,k}) = E\big[E(\varepsilon_i|\boldsymbol{X})|x_{j,k}\big] = 0$$

du fait de la loi des espérances itérées : $E[E(y|x,z)|x] = E(y|x)$

On aura alors avec la loi de l'espérance totale :

$$E(x_{j,k}\varepsilon_i) = E\big[E(x_{j,k}\varepsilon_i|x_{j,k})\big] = E\big[x_{j,k}E(\varepsilon_i|x_{j,k})\big] = 0$$

L'avant-dernière égalité s'obtient avec la propriété de linéarité des espérances conditionnelles : $E[f(x)y|x] = f(x)E(y|x)$

CQFD

On parlera alors de **conditions d'orthogonalité** pour cette hypothèse 2 :

$$E(\boldsymbol{X}'\boldsymbol{\varepsilon}) = \boldsymbol{0}_{\boldsymbol{K}}$$

- **Les conditions d'orthogonalité sont équivalentes à des conditions d'absence de corrélation.**

Avec la définition de la covariance et en utilisant l'espérance inconditionnelle de ε_i :

$$Cov(\varepsilon_i, x_{j,k}) = E(x_{j,k}\varepsilon_i) - E(x_{j,k})E(\varepsilon_i) = E(x_{j,k}\varepsilon_i) = 0$$

du fait des conditions d'orthogonalité précédentes.

- **L'hypothèse de stricte exogénéité n'est pas souvent satisfaite pour les modèles sur séries temporelles…**

Il se peut que la variable y au temps t influence en retour les variables explicatives X dans les périodes suivantes.

$$Corr(y_t, x_{t+s}) \neq 0 \quad \text{pour } s > 0$$

Exemple : Considérons le modèle (très courant) autorégressif simple en série temporelle, c'est-à-dire qu'il y a un seul régresseur, la variable dépendante retardée :

$$y_t = \beta y_{t-1} + \varepsilon_t$$

Supposons que le régresseur soit orthogonal à l'erreur : $E(y_{t-1}\varepsilon_t) = 0$, on aura alors :

$$E(y_t\varepsilon_t) = E\big((\beta y_{t-1} + \varepsilon_t)\varepsilon_t\big) = \beta E(y_{t-1}\varepsilon_t) + E(\varepsilon_t^2) = E(\varepsilon_t^2)$$

qui n'est pas nulle, sinon toutes les erreurs ε_t seraient nulles.

Mais y_t est le régresseur pour l'observation $t+1$. Celui-ci n'est donc pas orthogonal à l'erreur de la période précédente. Ce qui viole l'hypothèse de stricte exogénéité.

HYPOTHÈSE 3 : Absence de multicolinéarité parfaite (H3)

Le rang de la matrice $\boldsymbol{X}$ de dimension N x K est K avec une probabilité 1 :

$$\Rightarrow \quad rang(\boldsymbol{X}) = K$$

Cela veut dire qu'aucune colonne de la matrice $\boldsymbol{X}$ *(aucune variable)* n'est combinaison linéaire parfaite des autres colonnes de cette matrice *(des autres variables)*.

→ $\boldsymbol{X}$ est de rang-plein (colonne)

C'est une **condition suffisante**

Cette hypothèse permet l'**identification** des paramètres du modèle, c'est-à-dire cela permet d'obtenir une **estimation unique** des paramètres du modèle avec un critère donné.
Cette hypothèse (technique) rend possible l'estimation par moindres carrés ordinaires.

Une condition nécessaire mais pas suffisante est qu'il y ait au moins autant d'observations que de paramètres à estimer : $N \geq K$.

Les régresseurs sont dits « **parfaitement colinéaires** » si cette hypothèse n'est pas satisfaite. Cela se remarque souvent très facilement…

HYPOTHÈSE 4 : Les erreurs ont une matrice de variance-covariance sphérique (H4)

a) **Homoscédasticité conditionnelle** : Les termes d'erreur ε_i de toutes les observations ont la même variance conditionnelle :

$$V(\varepsilon_i|\boldsymbol{X}) = \sigma^2 \qquad \text{pour tout } i = 1,2,\dots,N\,.$$

Cette variance est strictement positive et finie : $0 < \sigma^2 < \infty$

On peut réécrire cette hypothèse comme : $V(\varepsilon_i|\boldsymbol{X}) = E(\varepsilon_i^2|\boldsymbol{X}) = \sigma^2$

$$\begin{aligned} V(\varepsilon_i|\boldsymbol{X}) &= E\left[\left(\varepsilon_i - E(\varepsilon_i|\boldsymbol{X})\right)^2\middle|\boldsymbol{X}\right] && \text{par définition de la variance} \\ &= E(\varepsilon_i^2|\boldsymbol{X}) - \underbrace{E(\varepsilon_i|\boldsymbol{X})^2}_{=\,0} && \text{parce que } E(\varepsilon_i|\boldsymbol{X}) = 0 \quad \text{(H2)} \\ &= E(\varepsilon_i^2|\boldsymbol{X}) = \sigma^2 \end{aligned}$$

---

b) **Absence de corrélation entre les observations**

Il y a **indépendance** « statistique » entre les erreurs.

En fait, on a besoin d'une hypothèse **plus faible** d'absence de corrélation entre les erreurs :

$$E(\varepsilon_i\varepsilon_j|\boldsymbol{X}) = 0 \qquad \text{pour tout } i,j = 1,2,\dots,N \text{ et } i \neq j\,.$$

En conséquence les erreurs ne sont pas corrélées entre elles.

$$Cov(\varepsilon_i,\varepsilon_j|\boldsymbol{X}) = E(\varepsilon_i\varepsilon_j|\boldsymbol{X}) - E(\varepsilon_i|\boldsymbol{X})E(\varepsilon_j|\boldsymbol{X}) = 0$$

$$Corr(\varepsilon_i,\varepsilon_j|\boldsymbol{X}) = \frac{Cov(\varepsilon_i,\varepsilon_j|\boldsymbol{X})}{\sqrt{V(\varepsilon_i|\boldsymbol{X}) \times V(\varepsilon_j|\boldsymbol{X})}} = \frac{E(\varepsilon_i\varepsilon_j|\boldsymbol{X})}{Var(\varepsilon_i|\boldsymbol{X})} = \frac{0}{\sigma^2} = 0$$

Conditionnellement à $\boldsymbol{X}$, il y a **non corrélation** entre les erreurs, et donc entre les observations !

→ ***absence d'autocorrélation***

Cette dernière hypothèse (H4) peut se réécrire en langage matriciel, sous la forme d'une **matrice de variance-covariance** (carrée de dimension N x N) :

$$V(\boldsymbol{\varepsilon}|\boldsymbol{X}) = E\left[\left(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}|\boldsymbol{X})\right)\left(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}|\boldsymbol{X})\right)'|\boldsymbol{X}\right]$$

$$V(\boldsymbol{\varepsilon}|\boldsymbol{X}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}] = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \boldsymbol{I}_N$$

La matrice identité de dimension (N x N).

Cette matrice de variance-covariance des erreurs est alors dite « **sphérique** ».

Cette matrice de variance-covariance est non-singulière parce que $\sigma^2 > 0$, donc elle est inversible.
Elle est aussi définie positive avec un déterminant positif.

DÉMONSTRATION : $V(\boldsymbol{\varepsilon}) = E\left[\left(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\right)\left(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\right)'\right] = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$

(en omettant le signe de conditionnalisation par rapport à X)

$$V(\boldsymbol{\varepsilon}) = E\left[\left(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\right)\left(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\right)'\right] = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$$

$$= E\left[\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix} (\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_N)'\right] = E\begin{bmatrix} \varepsilon_1{}^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_N \\ \varepsilon_2\varepsilon_1 & \varepsilon_2{}^2 & \cdots & \varepsilon_2\varepsilon_N \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_N\varepsilon_1 & \varepsilon_N\varepsilon_2 & \cdots & \varepsilon_N{}^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(\varepsilon_1{}^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_N) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2{}^2) & \cdots & E(\varepsilon_2\varepsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_N\varepsilon_1) & E(\varepsilon_2\varepsilon_N) & \cdots & E(\varepsilon_N{}^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \longleftarrow$$

Une matrice proportionnelle à la matrice identité de dimension (N x N).

Avec les hypothèses H4a : Homoscédasticité : $V(\varepsilon_i) = E(\varepsilon_i{}^2) = \sigma^2$

et H4b : Non –autocorrélation des erreurs : $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i\varepsilon_j) = 0$

HYPOTHÈSE 5 : Normalité des erreurs (H5)

Cette hypothèse n'est pas fondamentale pour l'obtention d'un bon estimateur, mais elle est **nécessaire** pour obtenir des propriétés en petits échantillons…

Si on suppose **les observations indépendantes**, la fonction de densité conjointe des erreurs est le produit des fonctions de densité marginale de chaque erreur :

$$f(\boldsymbol{\varepsilon}|\boldsymbol{X}) = f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N|\boldsymbol{X}) = f(\varepsilon_1|\boldsymbol{X}) \times f(\varepsilon_2|\boldsymbol{X}) \times \cdots \times f(\varepsilon_N|\boldsymbol{X}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} f(\varepsilon_i|\boldsymbol{X})$$

avec $f(.)$: la fonction de densité de la **loi normale univariée** :

$$\varepsilon_i|\boldsymbol{X} \approx \mathcal{N}(0\,,\sigma^2) \quad \rightarrow \quad f(\varepsilon_i|\boldsymbol{X}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{{\varepsilon_i}^2}{2\sigma^2}\right\}$$

Dans ce cas, les erreurs suivent une loi normale multivariée de dimension N:

$$\boldsymbol{\varepsilon}|\boldsymbol{X} \approx \mathcal{N}_N(\boldsymbol{0}\,,\sigma^2\boldsymbol{I}_N) \quad \rightarrow \quad f(\boldsymbol{\varepsilon}|\boldsymbol{X}) = (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left\{-\frac{{\varepsilon_i}^2}{2\sigma^2}\right\}$$

La fonction de densité des erreurs de la loi normale multivariée s'écrira alors :

$$\begin{aligned}
f(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} f(\varepsilon_i|\boldsymbol{X}) &= \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\varepsilon_i^2}{2\sigma^2}\right\}\right] \\
&= \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} \left[(2\pi)^{-1/2}(\sigma^2)^{-1/2} \exp\left\{-\frac{\varepsilon_i^2}{2\sigma^2}\right\}\right] \\
&= (2\pi)^{-N/2}(\sigma^2)^{-N/2} \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} \exp\left\{-\frac{\varepsilon_i^2}{2\sigma^2}\right\} \\
&= (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left\{\sum_{i=1}^{N}\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)\right\} \\
&= (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\varepsilon_i^2\right\}
\end{aligned}$$

(en langage matriciel) $$= (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left\{-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}\right\} \qquad \text{avec} \quad \sum_{i=1}^{N}\varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$$

I.2. L'estimation par Moindres Carrés Ordinaires (MCO).

a) *Le critère des moindres carrés*

Comment estimer les paramètres inconnus $\boldsymbol{\beta}$ et σ^2 ?

Utilisation de l'information de l'échantillon

Définir un critère ou un objectif à atteindre

Trouver une règle de décision ou de choix

Règle de décision → **Un estimateur**

Mise en œuvre – Faisabilité : **L'estimation** des paramètres du modèle

Critère des moindres carrés (Least-squares) :

→ ***Minimisation de la somme des carrés des erreurs (Q) :***

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^K} Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^K} Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

Avantages de ce critère :

- Donne les paramètres de la droite ou du plan de régression de y sur x
- Facilité de mise en œuvre (règle de décision → estimateur linéaire)
- Bonnes propriétés statistiques
- Mais forte pénalité aux erreurs « importantes »
- Pas de contraintes sur les paramètres estimés !

D'autres critères sont possibles. Par exemple :

- Minimiser la valeur absolue
- Minimiser le coefficient de Gini des erreurs

→ *voir des cours d'économétrie plus avancés…*

Propriétés du critère des moindres carrés Q :

$Q(\boldsymbol{\beta})$ est une fonction scalaire qui dépend d'un vecteur $(K \times 1)$ de paramètres $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) : \mathbb{R}^K \rightarrow \mathbb{R}^+$$

$Q(\boldsymbol{\beta})$ est une forme quadratique en $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \quad \text{avec} \quad \boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\beta}$$

qui peut se réécrire comme :

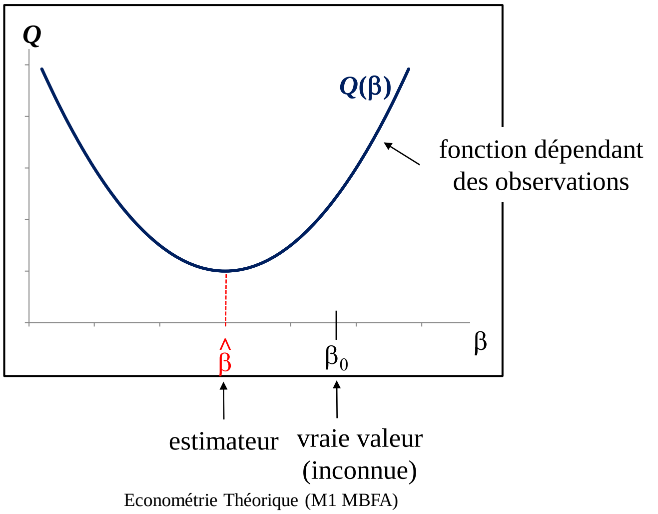
$$\begin{aligned} Q(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\ &= (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \end{aligned}$$

parce que $\boldsymbol{y}'\boldsymbol{X\beta}$ est un scalaire, et donc : $\boldsymbol{y}'\boldsymbol{X\beta} = (\boldsymbol{y}'\boldsymbol{X\beta})' = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$.

Propriétés du critère des moindres carrés Q :

$Q(\boldsymbol{\beta})$ a une forme en U avec un minimum unique *(voir plus loin)*.

Dans le cas où il y a un seul paramètre :



b) *L'estimateur des MCO*

Le critère : ***Minimisation de la somme des carrés des erreurs***

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^K} Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \boldsymbol{x_i}'\boldsymbol{\beta})^2$$

La solution de ce problème de minimisation est **l'estimateur des moindres carrés** :

$$\widehat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta}\in\mathbb{R}^K} Q(\boldsymbol{\beta})$$

On peut décomposer la forme quadratique $Q(\boldsymbol{\beta})$ avec la notation de sommation :

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} \beta_k x_{k,i}\right)^2$$

$$= \sum_{i=1}^{N} y_i^2 - 2\sum_{i=1}^{N}\sum_{k=1}^{K} \beta_k x_{k,i}\, y_i + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K} \beta_k \beta_l x_{k,i}\, x_{l,i}$$

(Voir démonstration ci-après…)

Il faut maintenant minimiser cette forme quadratique par rapport à tous les paramètres : $\beta_1, \beta_2, \dots, \beta_k, \dots, \beta_K$!!!

DÉMONSTRATION :

Décomposition de la forme quadratique $Q(\boldsymbol{\beta})$:

$$Q(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \underbrace{\boldsymbol{y}'\boldsymbol{y}}_{\text{Décomposition 1}} - \underbrace{2\boldsymbol{y}'\boldsymbol{X\beta}}_{\text{Décomposition 2}} + \underbrace{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}}_{\text{Décomposition 3}}$$

Décomposition 1 : $\boldsymbol{y}'\boldsymbol{y}$

$$\boldsymbol{y}'\boldsymbol{y} = (y_1 \quad y_2 \quad \cdots \quad y_N)\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \sum_{i=1}^{N} {y_i}^2$$

→ *Scalaire (1 x 1)*
→ *Ne dépend pas de* $\boldsymbol{\beta}$

Décomposition 2 : $-2\boldsymbol{y}'\boldsymbol{X\beta}$

$$\boldsymbol{y}'\boldsymbol{X} = (y_1 \quad y_2 \quad \cdots \quad y_N)\begin{bmatrix} 1 & x_{2,1} & \cdots & x_{K,1} \\ 1 & x_{2,2} & \cdots & x_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2,N} & \cdots & x_{K,N} \end{bmatrix} = \left(\sum_{\boldsymbol{i}} y_i \quad \sum_{\boldsymbol{i}} y_i\, x_{2,i} \quad \cdots \quad \sum_{\boldsymbol{i}} y_i\, x_{K,i}\right)$$

→ *Vecteur-ligne (1 x K)*

$$\boldsymbol{y}'\boldsymbol{X} = \begin{pmatrix} \sum_i y_i & \sum_i y_i\, x_{2,i} & \cdots & \sum_i y_i\, x_{K,i} \end{pmatrix} = (\boldsymbol{y}'\boldsymbol{x}_1 \quad \boldsymbol{y}'\boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{y}'\boldsymbol{x}_K)$$

avec : $\boldsymbol{y}'\boldsymbol{x}_k = \sum_{i=1}^{N} y_i x_{k,i}$

2ème étape : $-2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} = -2(\boldsymbol{y}'\boldsymbol{x}_1 \quad \boldsymbol{y}'\boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{y}'\boldsymbol{x}_K)\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} = -2\sum_{k=1}^{K} \beta_k \boldsymbol{y}'\boldsymbol{x}_k$

$$= -2\sum_{k=1}^{K} \beta_k \sum_{i=1}^{N} y_i x_{k,i}$$

$$= -2\sum_{i=1}^{N}\sum_{k=1}^{K} \beta_k y_i x_{k,i}$$

→ *Scalaire* (1 x 1)
→ *Combinaison linéaire en* $\boldsymbol{\beta}$

Décomposition 3 : $\boldsymbol{\beta}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{K,1} & x_{K,2} & \cdots & x_{K,N} \end{bmatrix}\begin{bmatrix} 1 & x_{2,1} & \cdots & x_{K,1} \\ 1 & x_{2,2} & \cdots & x_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2,N} & \cdots & x_{K,N} \end{bmatrix} = \begin{bmatrix} N & \sum_i x_{2,i} & \cdots & \sum_i x_{K,i} \\ \sum_i x_{2,i} & \sum_i x_{2,i}{}^2 & \cdots & \sum_i x_{2,i}x_{K,i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{K,i} & \sum_i x_{K,i}x_{2,i} & \cdots & \sum_i x_{K,i}{}^2 \end{bmatrix}$$

→ *Matrice symétrique (K x K)*

$$\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = \frac{\boldsymbol{X}'\boldsymbol{X}}{N} = \begin{bmatrix} 1 & \overline{x_2} & \cdots & \overline{x_K} \\ \overline{x_2} & \frac{1}{N}\sum_i x_{2,i}{}^2 & \cdots & \frac{1}{N}\sum_i x_{2,i}x_{K,i} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x_K} & \frac{1}{N}\sum_i x_{K,i}x_{2,i} & \cdots & \frac{1}{N}\sum_i x_{K,i}{}^2 \end{bmatrix}$$

avec : $\overline{x_k} = \frac{1}{N}\sum_{i=1}^{N} x_{k,i}$

→ *Matrice des moments des variables X*

Si on note $\boldsymbol{x}_i$ le vecteur (K x 1) des K variables explicatives pour un individu, la matrice des moments se réécrit comme :

$$\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i'$$

Maintenant on calcule la forme quadratique :

$$\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix}\beta_1 & \beta_2 & \cdots & \beta_K\end{pmatrix}\begin{bmatrix} N & \sum_i x_{2,i} & \cdots & \sum_i x_{K,i} \\ \sum_i x_{2,i} & \sum_i {x_{2,i}}^2 & \cdots & \sum_i x_{2,i}x_{K,i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{K,i} & \sum_i x_{K,i}x_{2,i} & \cdots & \sum_i {x_{K,i}}^2 \end{bmatrix}\begin{pmatrix}\beta_1 \\ \beta_2 \\ \vdots \\ \beta_K\end{pmatrix}$$

$$\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix}\beta_1 & \beta_2 & \cdots & \beta_K\end{pmatrix}\begin{pmatrix} \beta_1 N + \beta_2 \sum_i x_{2,i} + \cdots + \beta_K \sum_i x_{K,i} \\ \beta_1 \sum_i x_{2,i} + \beta_2 \sum_i {x_{2,i}}^2 + \cdots + \beta_K \sum_i x_{K,i}x_{2,i} \\ \vdots \\ \beta_1 \sum_i x_{K,i} + \beta_2 \sum_i x_{K,i}x_{2,i} + \cdots + \beta_K \sum_i {x_{K,i}}^2 \end{pmatrix}$$

$$= \begin{pmatrix}\beta_1 & \beta_2 & \cdots & \beta_K\end{pmatrix}\begin{pmatrix} \sum_{k=1}^K \beta_k \sum_i x_{k,i} \\ \sum_{k=1}^K \beta_k \sum_i x_{2,i}x_{k,i} \\ \vdots \\ \sum_{k=1}^K \beta_k \sum_i x_{K,i}x_{k,i} \end{pmatrix}$$

$$= \beta_1 \sum_{k=1}^K \beta_k \sum_i x_{k,i} + \beta_2 \sum_{k=1}^K \beta_k \sum_i x_{2,i}x_{k,i} + \cdots + \beta_K \sum_{k=1}^K \beta_k \sum_i x_{K,i}x_{k,i}$$

$$\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \sum_{l=1}^K \beta_l \sum_{k=1}^K \beta_k \sum_i x_{l,i}x_{k,i}$$

Ce qui donne finalement :

$$\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \sum_{i=1}^N \sum_{k=1}^K \sum_{l=1}^K \beta_k \beta_l x_{k,i} x_{l,i}$$

En reprenant les 3 décompositions, on obtient alors :

$$Q(\boldsymbol{\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\,\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

$$= \sum_{i=1}^N {y_i}^2 - 2 \sum_{i=1}^N \sum_{k=1}^K \beta_k y_i x_{k,i} + \sum_{i=1}^N \sum_{k=1}^K \sum_{l=1}^K \beta_k \beta_l x_{k,i} x_{l,i}$$

La minimisation du critère des moindres carrés en utilisant la notation matricielle :

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}$$

Formule de la dérivée d'une **combinaison linéaire** (2ème terme) :

$$\frac{\partial \boldsymbol{z}'\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}\left(\sum_{k=1}^{K} \theta_k z_k\right) = \boldsymbol{z} \quad \Rightarrow \quad \frac{\partial(-2\boldsymbol{y}'\boldsymbol{X\beta})}{\partial \boldsymbol{\beta}} = -2\frac{\partial(\boldsymbol{y}'\boldsymbol{X})\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y}$$

Formule de la dérivée d'une **forme quadratique symétrique** (3ème terme) :

$$\text{si } \boldsymbol{Z} \text{ est symétrique : } \frac{\partial \boldsymbol{\beta}'\boldsymbol{Z\beta}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{Z\beta} \quad \Rightarrow \quad \frac{\partial(\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}'\boldsymbol{X\beta}$$

Dérivée première du critère $Q(\boldsymbol{\beta})$ par rapport au vecteur des paramètres $\boldsymbol{\beta}$:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}) = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X\beta} = -2(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X\beta})$$

Un extrémum (ici le minimum) s'obtient en égalisant cette dérivée à 0 :

→ ***Condition du premier ordre :*** $\dfrac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{0}_K \quad \Rightarrow \quad -2(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X\beta}) = \boldsymbol{0}_K$

Il faut résoudre ce système d'équations pour obtenir **l'estimateur des moindres carrés** : $\widehat{\boldsymbol{\beta}}$

$$\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{0}_K \quad \Rightarrow \quad \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

Ce système est appelé **les équations normales** du problème des moindres carrés.

Ce système comprend K équations linéaires avec K inconnues : les composantes du vecteur $\boldsymbol{\beta}$ des paramètres.

Le système des K équations normales $\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ peuvent se réécrire comme :

$$\begin{bmatrix} N & \sum_i x_{2,i} & \cdots & \sum_i x_{K,i} \\ \sum_i x_{2,i} & \sum_i x_{2,i}^2 & \cdots & \sum_i x_{2,i}x_{K,i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{K,i} & \sum_i x_{K,i}x_{2,i} & \cdots & \sum_i x_{K,i}^2 \end{bmatrix} \begin{pmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \\ \vdots \\ \widehat{\beta_K} \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_{2,i} \\ \vdots \\ \sum_i y_i\, x_{K,i} \end{pmatrix}$$

Pour résoudre ce problème, il faut que la matrice (K x K) $\boldsymbol{X'X}$ soit de rang plein :

$$rang(\boldsymbol{X'X}) = K \quad \Leftrightarrow \quad \det(\boldsymbol{X'X}) = |\boldsymbol{X'X}| \neq 0 \quad (\text{ici } \det(\boldsymbol{X'X}) > 0)$$

Pour cela, il suffit que l'hypothèse d'absence de multicollinéarité parfaite (H3) soit satisfaite : Si $rang(\boldsymbol{X}) = K \leq N \quad \Rightarrow \quad rang(\boldsymbol{X'X}) = K$

Dans ce cas, l'inverse de $\boldsymbol{X'X}$ existe : $(\boldsymbol{X'X})^{-1}$

et on obtient une **solution unique** pour le système des équations normales : $\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'y}$:

l'estimateur des moindres carrés ordinaire (MCO) :

$$\boxed{\boldsymbol{\hat{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}}$$

En anglais : Ordinary Least-Squares (OLS) Estimator

Remarque : dans le livre de Hayashi, l'estimateur MCO est noté en lettre romaine minuscule : ***b*** !

Minimum de la somme des carrés des erreurs ?

L'estimateur des MCO est bien **le minimum** du critère parce que la dérivée seconde (*la matrice Hessienne*) :

$$\frac{\partial^2 Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \frac{\partial}{\partial\boldsymbol{\beta}'}(-2\boldsymbol{X'y} + 2\boldsymbol{X'X\beta}) = 2\boldsymbol{X'X}$$

est une **matrice symétrique définie-positive** si la condition de rang précédente est satisfaite (Hypothèse H3) :

$$\text{Si } rang(\boldsymbol{X}) = K \leq N \quad \Rightarrow \quad rang(\boldsymbol{X'X}) = K$$

La parabole du critère des moindres carrés est orientée vers le haut, elle a une forme en U.

Le minimum obtenu (dans le modèle linéaire) est **unique** *(identification)* !

On utilise souvent la notation : $\dfrac{\partial^2 Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = 2\boldsymbol{X'X} > \boldsymbol{0}$

Dérivation alternative de l'estimateur des MCO : la méthode des moments

On part de l'hypothèse 2 de stricte exogénéité

$$E(\boldsymbol{x}_i \varepsilon_i) = \boldsymbol{0}_K \quad \text{pour tout } i = 1,2,\dots,N.$$

On remplace ε_i par le vrai modèle: $y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$

$$E\left(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\right) = \boldsymbol{0}_K$$

Le principe d'analogie de la **méthode des moments** est de remplacer l'espérance sur la population (des individus) par la moyenne sur un échantillon observé :

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{b}) = \boldsymbol{0}_K$$

On obtient alors un système de K équations à K inconnues (les K éléments du vecteur $\boldsymbol{\beta}$) pour calculer **l'estimateur de la méthode des moments** $\boldsymbol{b}$.

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{b}) = \boldsymbol{0}_K$$

On peut réécrire ce système d'équation sous la forme :

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i'\,\boldsymbol{b} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i y_i$$

qui correspond aux équations normales de l'estimateur des MCO :

$$\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i'\,\boldsymbol{b} = \sum_{i=1}^{N} \boldsymbol{x}_i y_i \quad \Leftrightarrow \quad \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}$$

En effet : $\boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i'$ et $\boldsymbol{X}'\boldsymbol{y} = \sum_{i=1}^{N} \boldsymbol{x}_i y_i$.

L'estimateur de la méthode des moments :

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \widehat{\boldsymbol{\beta}}$$

est identique à l'estimateur des MCO.

c) *Remarques sur l'estimation par MCO*

DÉFINITION :

La **valeur calculée** (*fitted*) de $\boldsymbol{y}$: $\hat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

Le **résidus** (*residuals*) de l'estimation est l'écart entre la valeur observée et la valeur calculée de $\boldsymbol{y}$: $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

avec $\boldsymbol{e}$ le vecteur (N x 1) des ***résidus***, composé d'éléments $e_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$.

Il ne faut pas confondre les **résidus** $\boldsymbol{e}$ (calculables avec les observations) avec les **erreurs** $\boldsymbol{\varepsilon}$ (inconnues par définition).

PROPRIÉTÉ DES MCO :

La minimisation du critère des moindres carrés implique **toujours** que **le vecteur des résidus soit orthogonal à la matrice des variables explicatives**.

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}_K$$

DÉMONSTRATION : A partir des ***K* équations normales**, on a :

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0}_K$$
$$\boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{0}_K$$
$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}_K$$

Attention à ne pas confondre, cette propriété de l'estimation par MCO : $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}_K$ qui est toujours vérifiée avec l'hypothèse H2 : $E(\boldsymbol{X}'\boldsymbol{\varepsilon}) = \boldsymbol{0}_K$

Pourquoi ?

CONSÉQUENCE 1 :

Il y a une **corrélation nulle** entre les variables explicatives et les résidus :

$$\boldsymbol{x}_k'\boldsymbol{e} = \sum_i x_{k,i} e_i = 0 \qquad \text{pour tout } k = 1,2,\dots,K.$$

CONSÉQUENCE 2 :

La somme des résidus est toujours nulle s'il y a une constante dans la matrice de variables explicatives : (soit $\boldsymbol{x_1} = \boldsymbol{i}$ un vecteur de 1)

$$\boldsymbol{x}_1'\boldsymbol{e} = \boldsymbol{i}'\boldsymbol{e} = \sum_i e_i = 0$$

Donc il est **inutile** de vérifier cette propriété quand il y a une constante dans la régression !!!

CONSÉQUENCE 3 :

Le plan de régression passe par le point moyen du nuage de points s'il y a une constante dans la régression :

En prenant la première ligne des équations normales $\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$, on a :

$$\begin{pmatrix} N & \sum_i x_{2,i} & \cdots & \sum_i x_{K,i} \end{pmatrix} \widehat{\boldsymbol{\beta}} = \sum_i y_i$$

$$\begin{pmatrix} 1 & \frac{1}{N}\sum_i x_{2,i} & \cdots & \frac{1}{N}\sum_i x_{K,i} \end{pmatrix} \widehat{\boldsymbol{\beta}} = \frac{1}{N}\sum_i y_i$$

$$(1 \quad \overline{x_2} \quad \cdots \quad \overline{x_K})\widehat{\boldsymbol{\beta}} = \bar{y} \quad \Rightarrow \quad \bar{y} = \overline{\boldsymbol{x}}'\widehat{\boldsymbol{\beta}}$$

Dans la pratique, la majorité des logiciels économétriques n'utilisent pas ces formules qui nécessitent d'inverser la matrice $\boldsymbol{X}'\boldsymbol{X}$ …

Ils utilisent plutôt une méthode basée sur la décomposition en base orthogonale et triangulaire (dite décomposition **QR**) beaucoup moins coûteuse à calculer en temps et en nombre d'opérations élémentaires (*voir Davidson-McKinnon [1993], section I.5*)

Soit la matrice $N \times K$ des variables explicatives $\boldsymbol{X}$, on peut la réécrire comme :

$$\underset{N\times K}{\boldsymbol{X}} = \underset{N\times K}{\boldsymbol{Q}}\underset{K\times K}{\boldsymbol{R}} \quad \text{avec} \quad \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_K$$

Les colonnes de $\boldsymbol{Q}$ sont donc orthonormales.

$\boldsymbol{R}$ est une matrice carrée de dimension $K \times K$ triangulaire supérieure telle que :

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{R}$$

En effet : $\boldsymbol{X}'\boldsymbol{X} = (\boldsymbol{QR})'(\boldsymbol{QR}) = \boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{QR} = \boldsymbol{R}'\boldsymbol{R}$ parce que $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_K$

On peut montrer que l'estimateur des moindres carrés se calcule alors comme :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{y}$$

En effet :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{R}'\boldsymbol{R})^{-1}(\boldsymbol{Q}\boldsymbol{R})'\boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{R}'^{-1}\boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{y}$$

Cette formulation alternative nécessite seulement l'inversion de la matrice triangulaire supérieure $\boldsymbol{R}$, ce qui est facile récursivement …

mais au prix du calcul relativement simple (récursivement) de $\boldsymbol{Q}$ et de $\boldsymbol{R}$ à partir de la matrice $\boldsymbol{X}$!

Le nombre d'opérations élémentaires est réduit et les arrondis de calculs sont moindres avec cette méthode alternative !

En plus, cette méthode permet d'identifier la (les) variable(s) qui seraient parfaitement co-linéaires !

I.3. L'estimateur de la variance de l'erreur

On peut également estimer un second paramètre : **la variance de l'erreur** : σ^2

Comme dans le cas de la régression simple, on va utiliser la **somme des carrés des résidus.**

Le vecteur (N x 1) des **résidus** $\boldsymbol{e}$ est défini par : $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

En remplaçant l'estimateur MCO par sa formule :

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y} \quad \text{avec : } \boldsymbol{M} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\end{aligned}$$

La matrice $\boldsymbol{M}$ de dimension (N x N) est une matrice symétrique et idempotente :

$$\boldsymbol{M}' = \boldsymbol{M} \quad \text{et} \quad \boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$$ *(à démontrer ?)*

Cette matrice $\boldsymbol{M}$ est appelée « *residual – maker* » par W. Greene.

De plus, si on pré-multiplie $\boldsymbol{X}$ par cette matrice $\boldsymbol{M}$, on obtient une matrice nulle :

$$\boldsymbol{MX} = (\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}_{N\times K}$$

On aura avec le résultat précédent :

$$\boldsymbol{e} = \boldsymbol{My} = \boldsymbol{M}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{MX\beta} + \boldsymbol{M\varepsilon} = \boldsymbol{M\varepsilon}$$

Cependant cette expression est uniquement théorique parce qu'on ne connaît pas $\boldsymbol{\varepsilon}$. On ne peut pas trouver le vecteur des erreurs $\boldsymbol{\varepsilon}$ à partir du vecteur des résidus $\boldsymbol{e}$ en calculant : $\boldsymbol{\varepsilon} = \boldsymbol{M}^{-1}\boldsymbol{e}$, parce que la matrice $\boldsymbol{M}$ n'est pas inversible ! *(Pourquoi ?)*

Il n'est pas d'un intérêt pratique de calculer (ni d'imprimer) cette matrice $\boldsymbol{M}$, parce qu'elle peut être très grande : $N \times N$ (même si elle est symétrique).

Imaginez un échantillon de 1 000 observations : elle serait de dimension $1\,000 \times 1\,000$, *contenant 1 000 000 de nombres, soit 8 Mo !*
(en général : 1 nombre = 8 octets)

DÉFINITION : La **somme des carrés des résidus** (***SCR***) :

$$SCR = \boldsymbol{e}'\boldsymbol{e} = \sum_{i=1}^{N} e_i^2 \quad \text{avec} \quad e_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$$

Cette somme des carrés des résidus peut s'exprimer théoriquement en fonction du terme d'erreurs :

$$SCR = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{M\varepsilon})'(\boldsymbol{M\varepsilon}) = \boldsymbol{\varepsilon}'\boldsymbol{M}'\boldsymbol{M\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M\varepsilon}$$

On peut aussi réécrire la SCR de la manière suivante :

$$SCR = \boldsymbol{e}'\boldsymbol{e} = \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

Ce qui correspond à la **valeur minimisée** du critère des MCO !

On remplace dans le dernier terme, le premier estimateur MCO : $\widehat{\boldsymbol{\beta}}' = \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ pour obtenir :

$$SCR = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

Et finalement : $SCR = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y}'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{y}'\mathbf{e}$

On montrera plus loin (*Section II.1.d*) que :

$$E(SCR) = E(\boldsymbol{e}'\boldsymbol{e}) = E(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}) = (N-K)\sigma^2$$

On suppose maintenant que le nombre d'observations est strictement supérieur au nombre de variables : $N > K$.

DÉFINITION : L'estimateur des moindres carrés de **la variance de l'erreur** sera donné par :

$$\widehat{\sigma^2} = \frac{SCR}{N-K} = \frac{1}{N-K}\sum_{i=1}^{N} e_i^2$$

On a ainsi estimé la dispersion de la variable dépendante autour de la droite de régression.

Attention, ce n'est pas la moyenne des carrés des résidus : $\frac{1}{N}\sum_{i=1}^{N} e_i^2$!
ni la variance des résidus : $\frac{1}{N-1}\sum_{i=1}^{N} e_i^2$!

DÉFINITION : Le facteur $(N-K)$ est appelé le nombre de **degrés de liberté** (***ddl***) de la régression (ou tout simplement) les degrés de liberté…
(en anglais : degrees of freedom - df)

$$ddl = (N-K)$$

Remarquez que cet **estimateur de la variance** de l'erreur est mesuré dans le **carré** des unités de $\boldsymbol{e}$ (ou de $\boldsymbol{y}$) !

DÉFINITION : **L'écart-type de la régression** est la racine carrée de cet estimateur de la variance.
(En anglais : standard-error of the regression ou encore : Root mean-squared error – RMSE)

$$\hat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{SCR}{N-K}}$$

L'écart-type de la régression est mesuré avec les unités de $\boldsymbol{e}$ (ou de $\boldsymbol{y}$) !

I.4. La qualité de l'ajustement

a) *Le coefficient de détermination : R^2*

F. HAYASHI, *Econometrics* [2000], Section I.2. (p. 20-21)

W. GREENE, *Econométrie* [2018], Section III.4.

On a vu que l'estimation par MCO du modèle de régression impliquait par définition :

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e} = \widehat{\boldsymbol{y}} + \boldsymbol{e}$$

Pré-multiplions cette expression par sa transposée :

$$\boldsymbol{y}'\boldsymbol{y} = \left(\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}\right)'\left(\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}\right)$$

$$\boldsymbol{y}'\boldsymbol{y} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}\boldsymbol{X}'\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}'\boldsymbol{e}$$

Nous avons vu plus haut que les équations normales impliquaient : $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}_K$. En conséquence les deux termes centraux sont nuls …

$$\boldsymbol{y}'\boldsymbol{y} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}'\boldsymbol{e}$$

Soustrayons maintenant le carré de la moyenne de $\boldsymbol{y}$ multiplié par le nombre d'observations de chaque côté de l'égalité :

$$(\boldsymbol{y}'\boldsymbol{y} - N\bar{y}^2) = \left(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} - N\bar{y}^2\right) + \boldsymbol{e}'\boldsymbol{e}$$

Comme $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, il est évident que la moyenne de $\widehat{\boldsymbol{y}}$ est identique à la moyenne de $\boldsymbol{y}$.

$$\underbrace{(\boldsymbol{y}'\boldsymbol{y} - N\bar{y}^2)}_{SCT} \underset{=}{=} \underbrace{\left(\widehat{\boldsymbol{y}}'\widehat{\boldsymbol{y}} - N\bar{\widehat{y}}^2\right)}_{SCE} + \underbrace{\boldsymbol{e}'\boldsymbol{e}}_{+SCR}$$

avec :

- $SCT = \boldsymbol{y}'\boldsymbol{y} - N\bar{y}^2$: la somme des carrés totaux *(de la variable dépendante)*
- $SCE = \widehat{\boldsymbol{y}}'\widehat{\boldsymbol{y}} - N\bar{\widehat{y}}^2$: la somme des carrés expliqués *(par la régression)*
- $SCR = \boldsymbol{e}'\boldsymbol{e}$: la somme des carrés des résidus

La **somme des carrés totaux** (SCT) est définie par :

$$SCT = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \boldsymbol{y}'\boldsymbol{y} - N\bar{y}^2$$

La **somme des carrés expliqués** (*SCE*) par la régression est définie par :

$$SCE = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} - N\bar{y}^2$$

La **somme des carrés des résidus** (*SCR*) est définie par :

$$SCR = \sum_{i=1}^{N} e_i^2 = \boldsymbol{e}'\boldsymbol{e}$$

En notation matricielle

La moyenne se calcule comme : $\bar{y} = \dfrac{\boldsymbol{j}_N{}'\boldsymbol{y}}{N}$ avec $\boldsymbol{j}_N = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

Dès lors : $N\bar{\boldsymbol{y}}^2 = N\left(\dfrac{\boldsymbol{j}_N{}'\boldsymbol{y}}{N}\right)'\left(\dfrac{\boldsymbol{j}_N{}'\boldsymbol{y}}{N}\right) = \boldsymbol{y}'\dfrac{\boldsymbol{j}_N\boldsymbol{j}_N{}'}{N}\boldsymbol{y} = \boldsymbol{y}'\dfrac{\boldsymbol{J}_N}{N}\boldsymbol{y}$ avec $\boldsymbol{J}_N = \boldsymbol{j}_N\boldsymbol{j}_N'$

La matrice $\boldsymbol{J}_N$ est une matrice carrée de dimension $N \times N$ qui ne contient que des valeurs 1.

En conséquence : $SCT = \boldsymbol{y}'\boldsymbol{y} - N\bar{\boldsymbol{y}}^2 = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\dfrac{\boldsymbol{J}_N}{N}\boldsymbol{y} = \boldsymbol{y}'\left(\boldsymbol{I}_N - \dfrac{\boldsymbol{J}_N}{N}\right)\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{M}_0\boldsymbol{y}$

La matrice $\boldsymbol{M}_0$ effectue le centrage des observations $\boldsymbol{y}$ par rapport à leur moyenne :

$$\boldsymbol{M}_0\boldsymbol{y} = \left(\boldsymbol{I}_N - \frac{\boldsymbol{J}_N}{N}\right)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{j}_N\frac{\boldsymbol{j}_N{}'}{N}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{j}_N\bar{y}$$

La matrice $\boldsymbol{M}_0$ est une matrice symétrique et idempotente.

Remarque : elle donne le vecteur du résidu d'une régression où la seule variable explicative est la constante !

Si on pré-multiplie la régression par la matrice $\boldsymbol{M_0}$, on obtient :

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e} \quad \rightarrow \quad \boldsymbol{M_0 y} = \boldsymbol{M_0 X}\widehat{\boldsymbol{\beta}} + \boldsymbol{M_0 e} = \boldsymbol{M_0 X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}$$

parce que les résidus ont une moyenne nulle *s'il y a une constante parmi les variables explicatives* $\boldsymbol{X}$:

$$\boldsymbol{M_0 e} = \left(\boldsymbol{I}_N - \frac{\boldsymbol{J}_N}{N}\right)\boldsymbol{e} = \boldsymbol{e} - \boldsymbol{j}_N \frac{\boldsymbol{j}_N'\boldsymbol{e}}{N} = \boldsymbol{e} - \boldsymbol{j}_N \times 0 = \boldsymbol{e}$$

On effectue alors le produit scalaire de ce vecteur :

$$(\boldsymbol{M_0 y})'(\boldsymbol{M_0 y}) = \left(\boldsymbol{M_0 X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}\right)'\left(\boldsymbol{M_0 X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}\right)$$

$$\boldsymbol{y}'\boldsymbol{M_0 y} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{M_0 X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{M_0}\,\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{M_0 X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}'\boldsymbol{e}$$

$$\boldsymbol{y}'\boldsymbol{M_0 y} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{M_0 X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}'\boldsymbol{e} \qquad \textit{parce que}\ \ \boldsymbol{X}'\boldsymbol{M_0}\,\boldsymbol{e} = \boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$$

$$SCT = SCE + SCR$$

du fait des définitions de *SCT*, *SCE* et *SCR*.

Comme la somme des carrés totaux (la variabilité) de la variable $\boldsymbol{y}$ se décompose dans la somme de ce qui est expliqué par la régression (*SCE*) et de ce qui est inexpliqué (*SCR*) : $SCT = SCE + SCR$,
un bon ajustement voudrait que la part de la *SCE* soit importante, et la part de la *SCR* soit faible.

Si on divise partout par la *SCT*, on aura : $\dfrac{SCE}{SCT} + \dfrac{SCR}{SCT} = 1$

DEFINITION : On mesure la qualité de l'ajustement de la régression avec le **coefficient de détermination** : R^2 qui est définit comme :

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

$$R^2 = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{N} {e_i}^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

On peut aussi réécrire le R^2 sous la forme :

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{M_0}\boldsymbol{y}} \qquad \begin{cases} \boldsymbol{M} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\ \boldsymbol{M_0} = \boldsymbol{I}_N - \frac{\boldsymbol{J}_N}{N} = \boldsymbol{I}_N - \boldsymbol{j}_N(\boldsymbol{j}'_N\boldsymbol{j}_N)^{-1}\boldsymbol{j}'_N \end{cases}$$

PROPRIÉTÉ 1 : Pour autant qu'il y ait une constante dans le modèle, le coefficient de détermination R^2 est compris entre 0 et 1.

PROPRIÉTÉ 2 : Le coefficient de détermination R^2 est aussi le carré du coefficient de corrélation entre la valeur observée et la valeur calculée de la variable dépendante :

$$R^2 = Corr^2(y, \hat{y}) = \frac{Cov^2(y, \hat{y})}{V(y) \times V(\hat{y})}$$

(Voir démonstration ci-après ...)

Démonstration :

D'après la définition des variances :

$$V(y) = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{SCT}{N-1}$$

$$V(\hat{y}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = \frac{SCE}{N-1}$$

La covariance entre la variable dépendante $\boldsymbol{y}$ et la variable calculée $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ est par définition :

$$Cov(y, \hat{y}) = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{y}) = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i\hat{y}_i - N\bar{y}^2\right]$$

$$= \frac{1}{N-1}(\boldsymbol{y}'\hat{\boldsymbol{y}} - N\bar{y}^2) = \frac{1}{N-1}\left(\boldsymbol{y}'\hat{\boldsymbol{y}} - \boldsymbol{y}'\frac{\boldsymbol{J}_N}{N}\hat{\boldsymbol{y}}\right) = \frac{1}{N-1}\boldsymbol{y}'\boldsymbol{M}_0\hat{\boldsymbol{y}}$$

On peut aussi réécrire la covariance comme :

$$Cov(y,\hat{y}) = \frac{1}{N-1}\boldsymbol{y}'\boldsymbol{M}_0\hat{\boldsymbol{y}}$$

$$= \frac{1}{N-1}\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}\right)'\boldsymbol{M}_0\boldsymbol{X}\hat{\boldsymbol{\beta}} \qquad \text{parce que : } \boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}$$

$$= \frac{1}{N-1}\left(\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{M}_0\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}'\boldsymbol{M}_0\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$$

$$= \frac{1}{N-1}\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{M}_0\boldsymbol{X}\hat{\boldsymbol{\beta}} \qquad \text{parce que : } \boldsymbol{e}'\boldsymbol{M}_0\boldsymbol{X} = \boldsymbol{e}'\boldsymbol{X} = \boldsymbol{0}$$

$$= \frac{SCE}{N-1}$$

En conséquence, le carré de la corrélation entre variable dépendante et variable prédite devient le coefficient de détermination:

$$Corr^2(y,\hat{y}) = \frac{Cov^2(y,\hat{y})}{V(y)\times V(\hat{y})} = \frac{\left(\frac{SCE}{N-1}\right)^2}{\left(\frac{SCT}{N-1}\right)\left(\frac{SCE}{N-1}\right)} = \frac{SCE}{SCT} = R^2$$

CQFD.

b) Le R² non-centré

Problème du *R²* : Si il n'y a pas de constante dans le modèle,
le R^2 peut devenir négatif !!!
En effet on aura :

Si $SCR > SCT$ ou si $\sum_{i=1}^{N} {e_i}^2 > \sum_{i=1}^{N}(y_i - \bar{y})^2$ parce que $\sum_{i=1}^{N} e_i \neq 0$

alors $R^2 = 1 - \frac{SCR}{SCT} < 0$

Pour éviter cette valeur négative dans ce cas, on propose d'utiliser un **coefficient de détermination non centré** :

$$R_{nc}^2 = 1 - \frac{\sum_{i=1}^{N} {e_i}^2}{\sum_{i=1}^{N} {y_i}^2} = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y}}$$

En effet, on part de la décomposition de la variable dépendante : $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$.

Calculons son produit scalaire :

$$\begin{aligned} \boldsymbol{y}'\boldsymbol{y} = (\hat{\boldsymbol{y}} + \boldsymbol{e})'(\hat{\boldsymbol{y}} + \boldsymbol{e}) &= \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \boldsymbol{2}\hat{\boldsymbol{y}}'\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e} \\ &= \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \boldsymbol{2}\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e} \qquad \text{parce que : } \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} \qquad \text{parce que : } \boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0} \end{aligned}$$

Le coefficient de détermination non centré est alors :

$$R^2_{nc} = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y}} \qquad \text{avec } 0 \leq R^2_{nc} \leq 1$$

Il est toujours non-négatif et supérieur au coefficient de détermination centré : $R^2_{nc} \geq R^2$

Pourquoi ? Démontrez …

Si il n'y a pas de constante dans le modèle, **certains** logiciels (*Stata par exemple*) calculent automatiquement le coefficient de détermination non centré

$$R^2_{nc} = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y}}$$

plutôt que le coefficient de détermination centré :

$$R^2 = 1 - \frac{\sum_{i=1}^{N} {e_i}^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y} - N\bar{y}^2}$$

afin d'éviter d'obtenir une valeur négative pour ce dernier.

Mais attention au cas où il n'y a pas de constante, mais où une somme de variables explicatives est constante …!!!

c) Le R^2 ajusté pour les degrés de liberté

Problème du R^2 : Il augmente toujours lorsque l'on ajoute une variable explicative supplémentaire…
parce que la somme des carrés des résidus diminue…

Correction du R^2 : Ajustement pour le nombre de variables explicatives (K)

$$\bar{R}^2 = 1 - \frac{SCR/(N-K)}{SCT/(N-1)}$$

$$\bar{R}^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}/(N-K)}{\boldsymbol{y}'\boldsymbol{M}_0\boldsymbol{y}/(N-1)} = 1 - \frac{\hat{\sigma}^2}{V(y)}$$

On aura un **coefficient de détermination ajusté** (corrigé) pour les degrés de liberté.

Relation entre le R^2 et le R^2 ajusté :

A partir des définitions des coefficients de détermination, il est facile de montrer que :

$$\bar{R}^2 = 1 - \frac{N-1}{N-K}(1 - R^2)$$

De même le $\bar{R}^2$ ajusté sera toujours inférieur ou égal au R^2 classique :

$$\bar{R}^2 \leq R^2$$

L'égalité s'obtient lorsque l'ajustement est parfait : $\bar{R}^2 = R^2 = 1$!

Attention le $\bar{R}^2$ ajusté peut être négatif !

- supposons que le R^2 soit nul → $\bar{R}^2 = \dfrac{1-K}{N-K} \leq 0$
- Si le $R^2 \leq \frac{K-1}{N-1}$, alors $\bar{R}^2 \leq 0$

Le $\bar{R}^2$ ajusté peut diminuer lorsque l'on ajoute une variable explicative supplémentaire…

… si celle-ci n'apporte pas d'information suffisante pour un meilleur ajustement de la variable dépendante $\boldsymbol{y}$.

Cette variable explicative supplémentaire serait alors non pertinente !

(voir la démonstration dans la Section I.6)

d) Remarque sur l'utilisation du R^2

Le R^2 est souvent utiliser pour distinguer le pouvoir explicatif de deux régressions.

On peut choisir le modèle qui a le coefficient de détermination le plus élevé.

Cette stratégie de modélisation soulève cependant plusieurs remarques…

1) Comme le R^2 classique **augmente toujours** lorsque l'on ajoute une variable explicative à la régression, *même si elle est non pertinente (non significative)*, il est préférable d'utiliser le $\bar{R}^2$ ajusté pour les degrés de liberté.

2) Pour choisir entre des régresseurs, on doit comparer des régressions comparables : c'est-à-dire faites sur le **même échantillon**.
Donc il faut **vérifier que la taille et la composition de l'échantillon** ne changent pas entre les 2 régressions.

3) Le **choix d'un modèle <u>ne peut pas</u> se faire sur la seule base du coefficient de détermination *R^2*** . D'autres critères doivent être pris en compte :
- Significativité des paramètres estimés
- Interprétation économique des paramètres estimés
- Parcimonie
- Autres tests statistiques *(voir suite du cours)*

4) Il faut aussi que la **variable dépendante soit la même** dans les 2 modèles. En effet elle apparaît dans la ***SCT*** au dénominateur du R^2 .

I.5. Interprétation géométrique des MCO

On peut considérer chaque variable comme un point (ou un vecteur) dans un espace vectoriel à N dimensions : $\mathbb{R}^N$.

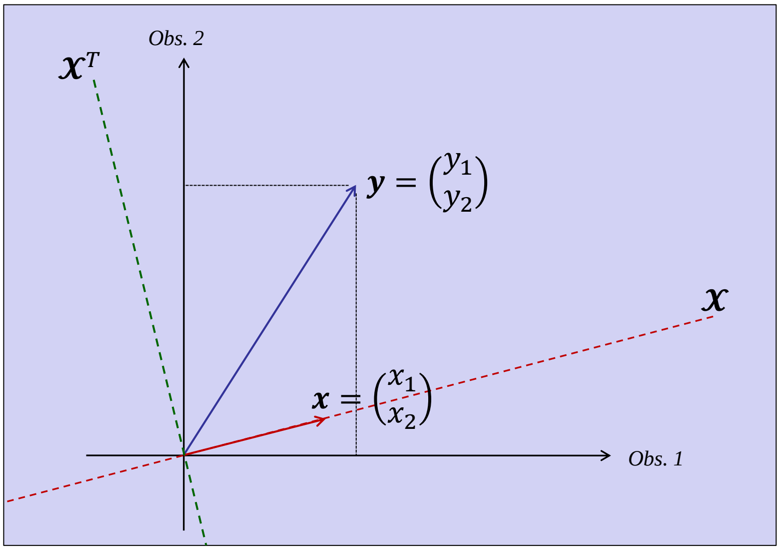
Donc la variable dépendante $\boldsymbol{y}$ représente un vecteur dans cet espace vectoriel, et les K variables explicatives $\boldsymbol{X}$ représentent K vecteurs dans ce même espace.

Ces K vecteurs définissent un sous-espace vectoriel (un hyperplan) $\boldsymbol{\mathcal{X}}$ de dimension K avec : $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^N$ avec $dim(\boldsymbol{\mathcal{X}}) = K$.

De même dans $\mathbb{R}^N$, on peut définir un espace vectoriel (un hyperplan) orthogonal à $\boldsymbol{\mathcal{X}}$ de dimension $N – K$, appelé $\boldsymbol{\mathcal{X}}^T$:

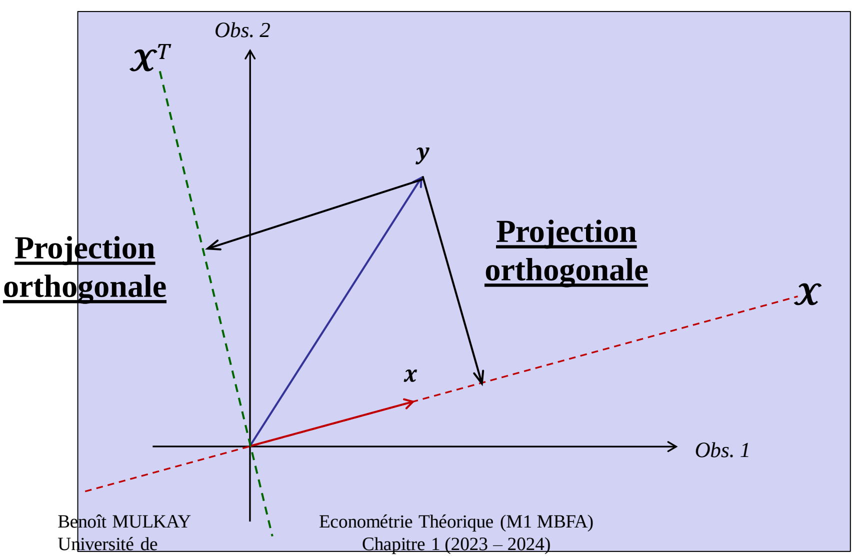
$$\boldsymbol{\mathcal{X}} \perp \boldsymbol{\mathcal{X}}^T \text{ et } \boldsymbol{\mathcal{X}}^T \subset \mathbb{R}^N \text{ avec } dim(\boldsymbol{\mathcal{X}}^T) = N - K$$
$$\text{tel que } \boldsymbol{\mathcal{X}}^T \cup \boldsymbol{\mathcal{X}} = \mathbb{R}^N$$

Exemple : ***N* = 2 observations** et K = 1, une seule variable explicative



La régression consiste à effectuer deux **projections orthogonales** du vecteur $\boldsymbol{y}$

- dans le sous-espace vectoriel $\boldsymbol{\mathcal{X}}$ engendré par les variables explicatives $\boldsymbol{X}$,
- dans le sous-espace vectoriel orthogonal $\boldsymbol{\mathcal{X}}^T \perp \boldsymbol{\mathcal{X}}$.



Mathématiquement, en géométrie analytique, une **projection orthogonale** est réalisée par des opérations matricielles sur les vecteurs.

Une matrice $\boldsymbol{P}$ de dimension $N \times N$ est une **matrice de projection** si elle est symétrique ($\boldsymbol{P}' = \boldsymbol{P}$) et idempotente ($\boldsymbol{PP} = \boldsymbol{P}$).

La dimension du sous-espace de projection est donnée par : $dim(\boldsymbol{\mathcal{X}}) = rang(\boldsymbol{P})$

Pour effectuer une projection dans le sous-espace $\boldsymbol{\mathcal{X}}$ engendré par les variables explicatives $\boldsymbol{X}$, on utilisera la **matrice de projection** :

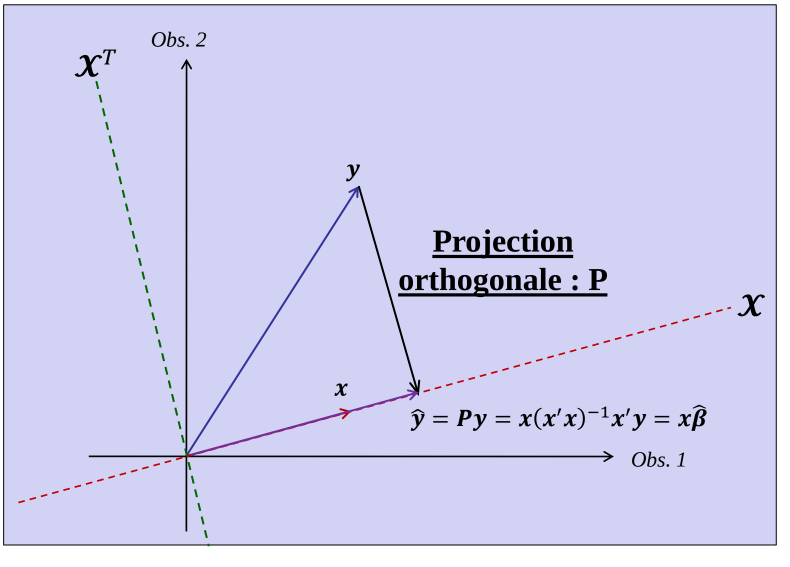
$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

Vérifiez que cette matrice est de dimension $N \times N$, symétrique et idempotente et que $dim(\boldsymbol{\mathcal{X}}) = rang(\boldsymbol{P}) = K$.

Le vecteur projeté dans le sous-espace $\boldsymbol{\mathcal{X}}$ de la variable dépendante $\boldsymbol{y}$ sera alors :

$$\boldsymbol{Py} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{y}}$$

Exemple : N **= 2 observations** et $K = 1$, une seule variable explicative



La **matrice de projection** dans le sous-espace orthogonal $\boldsymbol{\mathcal{X}}^T$ à celui engendré par les variables explicatives $\boldsymbol{\mathcal{X}}$, est la matrice $\boldsymbol{M}$ précédente :

$$\boldsymbol{M} = \boldsymbol{I}_N - \boldsymbol{P} = (\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$$

Cette matrice $N \times N$ est symétrique et idempotente et orthogonale à $\boldsymbol{P}$:

$$\boldsymbol{PM} = \boldsymbol{P}(\boldsymbol{I}_N - \boldsymbol{P})\boldsymbol{I}_N = \boldsymbol{P} - \boldsymbol{PP} = \boldsymbol{P} - \boldsymbol{P} = \boldsymbol{0}$$

La dimension du sous-espace $\boldsymbol{\mathcal{X}}^T$ est : $dim(\boldsymbol{\mathcal{X}}^T) = rang(\boldsymbol{M}) = N - K$

Si on applique cette projection $\boldsymbol{M}$ à la matrice des variables explicatives $\boldsymbol{X}$, on obtient le vecteur nul :

$$\boldsymbol{MX} = (\boldsymbol{I}_N - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{PX} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}$$

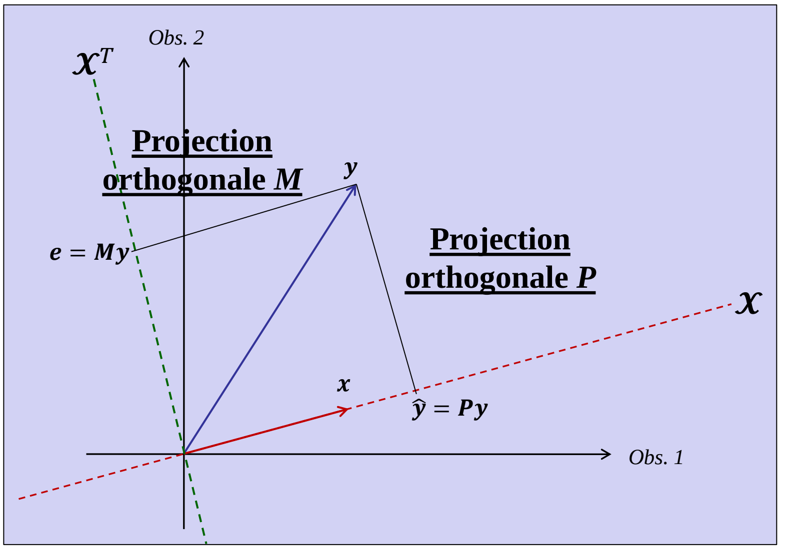
En effet le sous-espace $\boldsymbol{\mathcal{X}}^T$ est par construction orthogonal au sous-espace $\boldsymbol{\mathcal{X}}$ qui est supporté par les vecteurs de $\boldsymbol{X}$.

En revanche si on applique cette projection à la variable dépendante $\boldsymbol{y}$, on obtient le vecteur du résidu $\boldsymbol{e}$:

$$\boldsymbol{My} = (\boldsymbol{I}_N - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{Py} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{e}$$

Donc le résidu $\boldsymbol{e}$ est le vecteur $\boldsymbol{y}$ projeté orthogonalement dans le sous-espace vectoriel $\boldsymbol{\mathcal{X}}^T$ à celui engendré par les variables explicatives $\boldsymbol{X}$.

Exemple : $N = 2$ observations et $K = 1$, une seule variable explicative



On a ainsi décomposé la variable explicative $\boldsymbol{y}$ en deux composantes vectorielles orthogonales :

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} \quad \rightarrow \begin{cases} \hat{\boldsymbol{y}} \in \mathcal{X} \\ \boldsymbol{e} \in \mathcal{X}^T \end{cases} \quad \text{avec} \;\; \hat{\boldsymbol{y}} \perp \boldsymbol{e}$$

On a utilisé toute l'information contenue dans les variables explicatives $\boldsymbol{X}$ pour obtenir un $\boldsymbol{y}$ calculé.

En conséquence, il n'y a plus d'information disponible dans $\boldsymbol{X}$, contenue dans le résidu $\boldsymbol{e}$, afin d'améliorer la prévision de $\boldsymbol{y}$ **sans hypothèse supplémentaire**.

L'estimateur des MCO utilise ainsi l'ensemble de l'information disponible dans les vecteurs des variables explicatives…

b) Interprétation géométrique du R^2

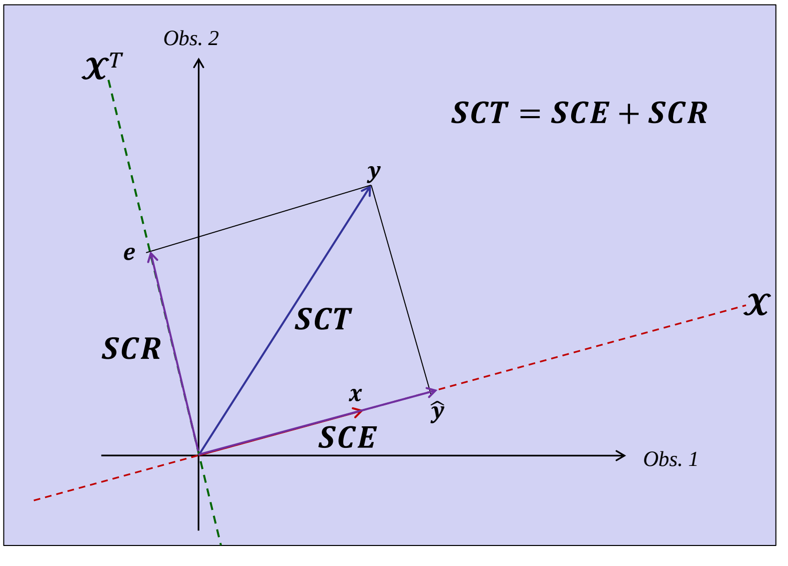
Retour sur l'interprétation géométrique de la régression par MCO
(*cfr. Section I.4*).

La décomposition $SCT = SCE + SCE$ provient aussi du théorème de Pythagore :

> *Dans un triangle rectangle, le carré de la longueur de l'hypoténuse est égal à la somme des carrés des longueurs des deux autres côtés.*

avec le fait qu'on a une décomposition orthogonale du vecteur $\boldsymbol{y}$:

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} \quad \text{avec} \quad \hat{\boldsymbol{y}} \perp \boldsymbol{e}$$



La régression consiste à faire une projection orthogonale du vecteur $\boldsymbol{y}$ dans le plan des régresseurs avec la matrice de projection $\boldsymbol{P}$.

On obtient le vecteur des variables calculées :

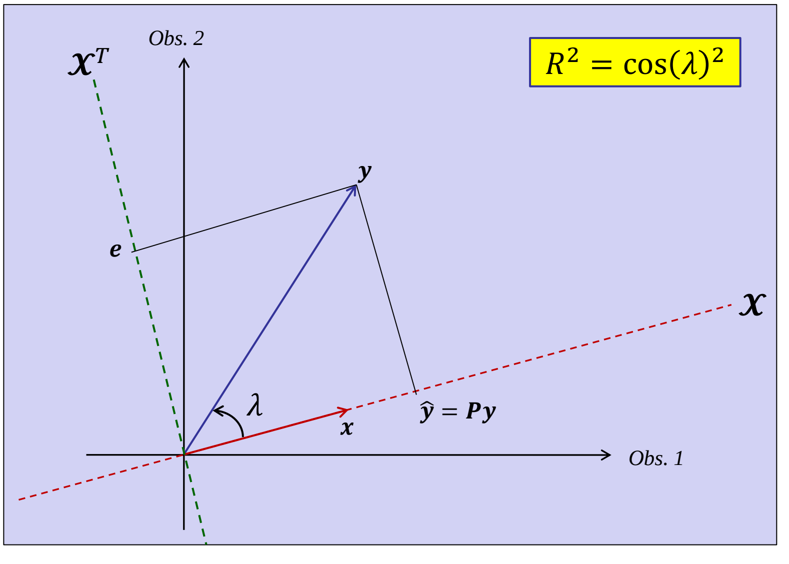
$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{y}}$$

Le coefficient de détermination R^2 est relié à l'angle λ entre le vecteur des variables dépendantes $\boldsymbol{y}$ et le vecteur des variables calculées $\widehat{\boldsymbol{y}}$:

$$R^2 = \cos(\lambda)^2$$

Plus le vecteur de la variable dépendante sera proche du plan de régression (engendré par les variables explicatives), plus le coefficient de détermination sera important.

Exemple : N = 2 observations et K = 1, une seule variable explicative



I.6. Le théorème de Frisch – Waugh – Lovell

Estimation séparée d'un sous-ensemble de paramètres :

→ les paramètres d'intérêt

→ les paramètres de « nuisance »

Comment enlever une tendance déterministe des données ?
Comment éviter de faire des inversions matricielles trop complexes ?

Voir : Ragnar FRISCH et Frederick WAUGH (1933) : « Partial Time Regressions as Compared to Individual Trends », *Econometrica*, I(4), pp. 387-401.

Michael C. LOVELL (1963) : « Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis », Journal of American Statistical Association, Vol. 58, N°304, pp. 996-1010.

L'intérêt pratique de ce théorème est assez limité de nos jours !
Mais il est très utile pour beaucoup de démonstration des propriétés de l'estimateur des MCO.

a) La régression partielle

Modèle de régression multiple avec K variables explicatives : $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

Les K variables explicatives sont **partitionnées** en 2 groupes de variables :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = (\boldsymbol{X}_1 \quad \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\varepsilon}$$

avec les matrices : $\boldsymbol{X}_1 : N \times K_1$ et $\boldsymbol{X}_2 : N \times K_2$ avec $K_1 + K_2 = K$
et les vecteurs : $\boldsymbol{\beta}_1 : K_1 \times 1$ et $\boldsymbol{\beta}_2 : K_2 \times 1$.

On s'intéresse plus particulièrement aux paramètres $\boldsymbol{\beta}_1$ (les paramètres d'intérêt).

Cela peut être le paramètre d'une seule variable explicative si : $K_1 = 1$ et $K_2 = K - 1$ ($\boldsymbol{X}_1$ devient alors un vecteur).

Voir : W. GREENE, Econométrie [2018], Théorème 3.3

Théorème de Frisch-Waugh-Lovell

Dans la régression linéaire par MCO de $\boldsymbol{y}$ sur deux groupes de variables explicatives $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$,

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

le sous-vecteur $\widehat{\boldsymbol{\beta}_1}$ des paramètres est obtenu quand les résidus ($\boldsymbol{u}$) d'une régression de $\boldsymbol{y}$ sur $\boldsymbol{X}_2$ sont régressés sur l'ensemble des résidus ($\boldsymbol{V}$) d'une régression de chaque colonne de $\boldsymbol{X}_1$ sur $\boldsymbol{X}_2$.

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{v} & \rightarrow \quad \boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}_2\widehat{\boldsymbol{\gamma}} \\ \boldsymbol{X}_1 = \boldsymbol{X}_2\boldsymbol{\Gamma} + \boldsymbol{\Upsilon} & \rightarrow \quad \boldsymbol{V} = \boldsymbol{X}_1 - \boldsymbol{X}_2\widehat{\boldsymbol{\Gamma}} \end{cases}$$

$$\Rightarrow \quad \boldsymbol{u} = \boldsymbol{V}\boldsymbol{\beta}_1 + \boldsymbol{\eta} \quad \rightarrow \quad \widehat{\boldsymbol{\beta}_1} = (\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'\boldsymbol{u}$$

qui peut être réécrit comme :

$$\widehat{\boldsymbol{\beta}_1} = (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y} \quad \text{avec} \quad \boldsymbol{M}_2 = I_N - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$$

DEMONSTRATION :

Considérons d'abord cette deuxième méthode.
On régresse la variable dépendante $\boldsymbol{y}$ sur les variables $\boldsymbol{X}_2$:

$$\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{v} \quad \rightarrow \quad \widehat{\boldsymbol{\gamma}} = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2{}'\boldsymbol{y}$$

Le résidu est alors égal à : $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}_2\widehat{\boldsymbol{\gamma}} = \boldsymbol{y} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{y}$

avec $\boldsymbol{M}_2 = \boldsymbol{I}_N - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$

On effectue aussi la régression (multivariée) de $\boldsymbol{X}_1$ sur $\boldsymbol{X}_2$:

$$\boldsymbol{X}_1 = \boldsymbol{X}_2\boldsymbol{\Gamma} + \boldsymbol{\Upsilon} \quad \rightarrow \quad \widehat{\boldsymbol{\Gamma}} = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2{}'\boldsymbol{X}_1$$

ce qui donne pour les résidus : $\boldsymbol{V} = \boldsymbol{X}_1 - \boldsymbol{X}_2\widehat{\boldsymbol{\Gamma}} = \boldsymbol{X}_1 - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1 = \boldsymbol{M}_2\boldsymbol{X}_1$

En fait, on projette $\boldsymbol{y}$ et $\boldsymbol{X}_1$ sur le plan orthogonal à celui engendré par $\boldsymbol{X}_2$.

On va maintenant régresser les résidus de la première régression ($\boldsymbol{u}$) sur les résidus de la seconde régression ($\boldsymbol{V}$) : $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{\beta}_1 + \boldsymbol{\eta}$

Ce qui donne : $\widehat{\boldsymbol{\beta}_1} = (\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'\boldsymbol{u} = (\boldsymbol{X}_1{}'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1{}'\boldsymbol{M}_2\boldsymbol{y}$

On a enlevé de la variable dépendante $\boldsymbol{y}$ et des régresseurs $\boldsymbol{X}_1$, l'effet des autres variables explicatives $\boldsymbol{X}_2$ (*partialling-out*) .

Calculons maintenant la régression globale :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = (\boldsymbol{X}_1 \quad \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\varepsilon}$$

Si on reprend les équations normales des moindres carrés, on aura :

$$\begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2\\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2\end{pmatrix}\begin{pmatrix}\widehat{\boldsymbol{\beta}_1}\\ \widehat{\boldsymbol{\beta}_2}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{y}\\ \boldsymbol{X}_2'\boldsymbol{y}\end{pmatrix}$$

Pour résoudre ce système, on peut utiliser les résultats de calcul matriciel sur les matrices partitionnées, ou procéder par substitution :

Prenons le second groupe d'équations : $(\boldsymbol{X}_2'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} + (\boldsymbol{X}_2'\boldsymbol{X}_2)\widehat{\boldsymbol{\beta}_2} = \boldsymbol{X}_2{}'\boldsymbol{y}$

dont la solution en fonction de $\widehat{\boldsymbol{\beta}}_1$ est : $\widehat{\boldsymbol{\beta}_2} = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} - (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1}$

Régression de $\boldsymbol{y}$ sur $\boldsymbol{X}_2$

Correction dépendante de la corrélation entre $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$

91

On substitue alors ce résultat dans le premier groupe d'équations normales :

$$(\boldsymbol{X}_1'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} + (\boldsymbol{X}_1'\boldsymbol{X}_2)\widehat{\boldsymbol{\beta}_2} = \boldsymbol{X}_1{}'\boldsymbol{y}$$

pour donner :

$$(\boldsymbol{X}_1'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} + (\boldsymbol{X}_1'\boldsymbol{X}_2)\left[(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} - (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2{}'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1}\right] = \boldsymbol{X}_1{}'\boldsymbol{y}$$

$$(\boldsymbol{X}_1'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} + (\boldsymbol{X}_1'\boldsymbol{X}_2)(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} + (\boldsymbol{X}_1'\boldsymbol{X}_2)(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} = \boldsymbol{X}_1{}'\boldsymbol{y}$$

$$(\boldsymbol{X}_1'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} + (\boldsymbol{X}_1'\boldsymbol{X}_2)(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} = \boldsymbol{X}_1'\boldsymbol{y} - (\boldsymbol{X}_1'\boldsymbol{X}_2)(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}$$

$$\boldsymbol{X}_1{}'(\boldsymbol{I}_N - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2')\boldsymbol{X}_1\widehat{\boldsymbol{\beta}_1} = \boldsymbol{X}_1'(\boldsymbol{I}_N - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2')\boldsymbol{y}$$

Du fait de la définition de la matrice de projection : $\boldsymbol{M}_2 = \boldsymbol{I}_N - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$, on obtient alors :

$$(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)\widehat{\boldsymbol{\beta}_1} = \boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}$$

Ce qui donne finalement exactement le même estimateur que précédemment, du fait que la matrice $(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)$ est inversible :

$$\widehat{\boldsymbol{\beta}_1} = (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}$$

→ **Les deux méthodes sont équivalentes.** **CQFD**

92

Le théorème de Frisch – Waugh permet de neutraliser l'influence d'un groupe de variables explicatives $\boldsymbol{X}_2$, pour se concentrer uniquement sur les paramètres d'intérêt $\boldsymbol{\beta}_1$.

On a enlevé de la variable dépendante $\boldsymbol{y}$ et des régresseurs $\boldsymbol{X}_1$, l'effet des autres variables explicatives $\boldsymbol{X}_2$

En anglais on appelle cette opération : *partialling – out* ou *netting – out* .

Ce théorème de Frisch – Waugh a une portée plus théorique que pratique… mais il permet d'effectuer de nombreuses démonstrations des propriétés des moindres carrés !

b) Exemple 1 : Le centrage des variables par rapport à leur moyenne

Soit $\boldsymbol{X}_2$ le vecteur de la constante : $\boldsymbol{X}_2 = \boldsymbol{j}_N = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

on aura pour la matrice de projection $\boldsymbol{M}_2$ sur l'espace orthogonal engendré par ce vecteur de constante :

$$\boldsymbol{M}_2 = \boldsymbol{I}_N - \boldsymbol{j}_N(\boldsymbol{j}_N{}'\boldsymbol{j}_N)^{-1}\boldsymbol{j}_N' = \boldsymbol{I}_N - \frac{\boldsymbol{j}_N\boldsymbol{j}_N'}{N} = \boldsymbol{I}_N - \frac{\boldsymbol{J}_N}{N}$$

avec $\boldsymbol{J}_N$: une matrice carrée $N \times N$ composée entièrement de 1.

Cette matrice $\boldsymbol{M}_2$ s'écrit alors :

$$\boldsymbol{M}_2 = \boldsymbol{I}_N - \frac{\boldsymbol{J}_N}{N} = \begin{bmatrix} 1-\frac{1}{N} & -\frac{1}{N} & -\frac{1}{N} & \cdots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & -\frac{1}{N} & \cdots & -\frac{1}{N} \\ -\frac{1}{N} & -\frac{1}{N} & 1-\frac{1}{N} & \cdots & -\frac{1}{N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{N} & -\frac{1}{N} & -\frac{1}{N} & \cdots & 1-\frac{1}{N} \end{bmatrix}$$

En prémultipliant un vecteur $\boldsymbol{y}$ par cette matrice $\boldsymbol{M}_2$, on **centre ce vecteur par rapport à sa moyenne** $\bar{y} = \boldsymbol{j}'_N\boldsymbol{y}/N$:

$$\boldsymbol{M}_2\boldsymbol{y} = \left(\boldsymbol{I}_N - \frac{\boldsymbol{J}_N}{N}\right)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{j}_N\frac{\boldsymbol{j}'_N\boldsymbol{y}}{N} = \boldsymbol{y} - \bar{y}\boldsymbol{j}_N = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{pmatrix} = \widetilde{\boldsymbol{y}}$$

L'estimateur des $K - 1$ paramètres d'intérêt d'un modèle où les variables sont centrées $\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}_1}\boldsymbol{\beta}_1 + \widetilde{\boldsymbol{v}} \;\rightarrow\; \boldsymbol{M}_2\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{M}_2\boldsymbol{v}$ est par le théorème de Frisch-Waugh :

$$\widehat{\boldsymbol{\beta}_1} = \left(\widetilde{\boldsymbol{X}'_1}\widetilde{\boldsymbol{X}_1}\right)^{-1}\widetilde{\boldsymbol{X}'_1}\widetilde{\boldsymbol{y}} = (\boldsymbol{X}'_1\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}'_1\boldsymbol{M}_2\boldsymbol{y}$$

Donc l'estimation des paramètres de pente n'est pas modifiée si on centre toutes les variables par rapport à leurs moyennes.

Remarquez qu'il ne serait pas nécessaire de « centrer » la variable dépendante y si on centre uniquement les variables explicatives. Pourquoi ?

Autres exemples : élimination d'une tendance,
ou d'effets saisonniers additifs,…

Il suffit d'éliminer ces effets de chacune des variables explicatives d'intérêt.

Cependant : Intérêt pratique du théorème quasiment nul actuellement !
Sauf modèle à effets fixes sur données de panel …

L'intérêt principal de ce théorème est d'ordre théorique…

c) Exemple 2 : L'introduction d'une variable supplémentaire

On reprend la régression partielle avec une seule variable supplémentaire pour la variable $\boldsymbol{X}_1 = \boldsymbol{z}$ et $\boldsymbol{X}_2 = \boldsymbol{X}$:

$$\boldsymbol{y} = (\boldsymbol{X}_1 \quad \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\varepsilon} \quad \rightarrow \quad \boldsymbol{y} = (\boldsymbol{z} \quad \boldsymbol{X})\begin{pmatrix}\gamma\\ \boldsymbol{\beta}\end{pmatrix} + \boldsymbol{\varepsilon}$$

L'utilisation du théorème de Frisch-Waugh permet d'estimer γ séparément de $\boldsymbol{\beta}$:

$$\hat{\gamma} = (\boldsymbol{z}'\boldsymbol{M}_X\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{M}_X\boldsymbol{y} \quad \text{avec : } \boldsymbol{M}_X = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$$

Ce qui donne en indiquant par une étoile (*) les variables pré-multipliées par $\boldsymbol{M}_X$:

$$\hat{\gamma} = (\boldsymbol{z}_*'\boldsymbol{z}_*)^{-1}\boldsymbol{z}_*'\boldsymbol{y}_* = \frac{\boldsymbol{z}_*'\boldsymbol{y}_*}{\boldsymbol{z}_*'\boldsymbol{z}_*} \qquad \text{avec } \boldsymbol{z}_* = \boldsymbol{M}_X\boldsymbol{z} \text{ et } \boldsymbol{y}_* = \boldsymbol{M}_X\boldsymbol{y}$$

C'est l'estimateur des MCO d'une régression simple de $\boldsymbol{y}_*$ sur la seule variable explicative $\boldsymbol{z}_*$:

$$\boldsymbol{y}_* = \gamma\boldsymbol{z}_* + \boldsymbol{v} \quad \rightarrow \quad \boldsymbol{y}_* = \hat{\gamma}\boldsymbol{z}_*$$

$\boldsymbol{y}_* = \boldsymbol{M}_X\boldsymbol{y}$ est le résidu de la régression de $\boldsymbol{y}$ sur les variables explicatives $\boldsymbol{X}$.
$\boldsymbol{z}_* = \boldsymbol{M}_X\boldsymbol{z}$ est le résidu de la régression de $\boldsymbol{z}$ sur les variables explicatives $\boldsymbol{X}$.

$$\boldsymbol{y}_* = \gamma\boldsymbol{z}_* + \boldsymbol{v}$$

C'est la régression partielle après avoir éliminé l'influence des variables explicative $\boldsymbol{X}$ de la variable dépendante $\boldsymbol{y}$ et de la variable explicative d'intérêt $\boldsymbol{z}$.

$$\hat{\gamma} = (\boldsymbol{z}_*'\boldsymbol{z}_*)^{-1}\boldsymbol{z}_*'\boldsymbol{y}_*$$

De même, selon le théorème de Frisch-Waugh, l'estimateur MCO de $\boldsymbol{\beta}$ dans le modèle complet $\widehat{\boldsymbol{\beta}}$ peut se réécrire en fonction de $\boldsymbol{\beta}^*$ (l'estimateur MCO de $\boldsymbol{\beta}$ sans la variable $\boldsymbol{z}$) :

$$\boldsymbol{\beta}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\hat{\gamma} = \boldsymbol{\beta}^* - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\hat{\gamma}$$

Remarquez que si $\hat{\gamma} \neq 0$, les estimateurs sont égaux ($\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$) seulement si $\boldsymbol{X}'\boldsymbol{z} = \boldsymbol{0}$, c'est-à-dire s'il n'y a pas de covariance (corrélation) entre les variables $\boldsymbol{X}$ et la variable $\boldsymbol{z}$!

Notons le résidu de cette **régression « complète »** avec les variables $\boldsymbol{X}$ et $\boldsymbol{z}$:

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{z}\hat{\gamma}$$

alors que le résidu de la **régression « simplifiée »** avec seulement les variables $\boldsymbol{X}$:

$$\boldsymbol{e}_* = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{M}_X\boldsymbol{y} = \boldsymbol{y}_*$$

En introduisant l'estimateur $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\hat{\gamma}$ dans l'expression du résidu $\boldsymbol{e}$:

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta}^* - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\hat{\gamma}) - \boldsymbol{z}\hat{\gamma} \\ &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\hat{\gamma} - \boldsymbol{z}\hat{\gamma} \\ &= \boldsymbol{e}_* - \boldsymbol{M}_X\boldsymbol{z}\hat{\gamma} = \boldsymbol{e}_* - \boldsymbol{z}_*\hat{\gamma}\end{aligned}$$

La somme des carrés des résidus dans la régression « complète » devient alors :

$$\begin{aligned}SCR = \boldsymbol{e}'\boldsymbol{e} &= (\boldsymbol{e}_* - \boldsymbol{z}_*\hat{\gamma})'(\boldsymbol{e}_* - \boldsymbol{z}_*\hat{\gamma}) \\ &= \boldsymbol{e}_*'\boldsymbol{e}_* + \hat{\gamma}^2\boldsymbol{z}_*'\boldsymbol{z}_* - 2\hat{\gamma}\boldsymbol{z}_*'\boldsymbol{e}_* \\ &= SCR^* - \hat{\gamma}^2\boldsymbol{z}_*'\boldsymbol{z}_*\end{aligned}$$

parce que : $\boldsymbol{z}_*'\boldsymbol{e}_* = \boldsymbol{z}_*'\boldsymbol{y}_* = \boldsymbol{z}_*'(\hat{\gamma}\boldsymbol{z}_*) = \hat{\gamma}\boldsymbol{z}_*'\boldsymbol{z}_*$

En conséquence, la SCR du modèle « complet » est inférieure ou égale à la SCR^* du modèle « simplifié » : $SCR = SCR^* - \hat{\gamma}^2\boldsymbol{z}_*'\boldsymbol{z}_* \leq SCR^*$

En insérant $\hat{\gamma} = \boldsymbol{z}_*'\boldsymbol{y}_*/\boldsymbol{z}_*'\boldsymbol{z}_*$, on obtient finalement :

$$SCR^* - SCR = \hat{\gamma}^2\boldsymbol{z}_*'\boldsymbol{z}_* = \frac{(\boldsymbol{z}_*'\boldsymbol{y}_*)^2}{(\boldsymbol{z}_*'\boldsymbol{z}_*)} = r_*^2(\boldsymbol{y}_*'\boldsymbol{y}_*)$$

où r_* est la **corrélation partielle** entre la variable dépendante $\boldsymbol{y}$ et la variable explicative $\boldsymbol{z}$, **en neutralisant les effets des autres variables explicatives $\boldsymbol{X}$** :

$$r_* = Corr(\boldsymbol{y}_*, \boldsymbol{z}_*) = \frac{(\boldsymbol{z}_*'\boldsymbol{y}_*)}{\sqrt{(\boldsymbol{z}_*'\boldsymbol{z}_*)(\boldsymbol{y}_*'\boldsymbol{y}_*)}}$$

Quand on ajoute une variable explicative supplémentaire $\boldsymbol{z}$ à un modèle,
Le gain en termes d'ajustement, ou la diminution de la SCR dépend de la corrélation partielle entre $\boldsymbol{z}$ et $\boldsymbol{y}$, compte tenu des autres variables explicatives $\boldsymbol{X}$.

Il faut que la variable supplémentaire $\boldsymbol{z}$ apporte une information nouvelle qui n'est pas encore contenue dans les variables explicatives $\boldsymbol{X}$.

Dans le modèle « simplifié », le coefficient de détermination est :

$$R_*^2 = 1 - \frac{SCR_*}{SCT}$$

et dans le modèle « complet » : $R^2 = 1 - \frac{SCR}{SCT}$

Comme $SCR \leq SCR^*$, on aura : $$1 - R^2 = \frac{SCR}{SCT} \leq \frac{SCR_*}{SCT} = 1 - R_*^2$$

$$-R^2 \leq -R_*^2$$

$$R^2 \geq R_*^2$$

Lorsqu'on ajoute une variable explicative supplémentaire, le R^2 du nouveau modèle « complet » augmente par rapport au R_*^2 du modèle « simplifié ».

$$R^2 = R_*^2 + r_*^2 \frac{\boldsymbol{y}_*'\boldsymbol{y}_*}{SCT} = R_*^2 + r_*^2 \frac{\boldsymbol{y}'\boldsymbol{M}_X\boldsymbol{y}}{SCT}$$

d) Corollaire : la décomposition du R²

(Voir Henri THEIL (1971) : *Principles of Econometrics*, John Wiley & Sons, Chapitre IV.)

On a vu que le coefficient de détermination pouvait s'écrire :

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{M}_0\boldsymbol{y}}$$

$$R^2 = \frac{SCE}{SCT} = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{\boldsymbol{y}'\boldsymbol{M}_0\boldsymbol{y}}$$

Supposons que <u>les variables soient centrées</u> (il suffit d'avoir une constante dans le modèle pour obtenir la même propriété), alors :

$$\boldsymbol{y}'\boldsymbol{M}_0\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{y} \quad \text{et} \quad \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

On aura alors pour le R^2 : $$R^2 = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}}$$

qui peut s'exprimer sous la forme : $$R^2 = \frac{\sum_{k=1}^{K} \widehat{\beta_k} \sum_{i=1}^{N} x_{k,i} y_i}{\sum_{i=1}^{N} y_i^2}$$

$$R^2 = \frac{\sum_{k=1}^{K} \widehat{\beta_k} \sum_{i=1}^{N} x_{k,i} y_i}{\sum_{i=1}^{N} y_i^2} = \sum_{k=1}^{K} \widehat{\beta_k} \frac{\sum_{i=1}^{N} x_{k,i} y_i}{\sum_{i=1}^{N} y_i^2}$$

Pour une variable explicative, le ratio $\widehat{\beta_k} \frac{\sum_{i=1}^{N} x_{k,i} y_i}{\sum_{i=1}^{N} y_i^2}$ pourrait être considéré comme la contribution de la $k^{ième}$ variable explicative au R^2.

MAIS ce ratio peut être négatif pour une variable explicative !

Ce qui empêche toute utilisation de cette mesure comme contribution individuelle de la la $k^{ième}$ variable explicative au R^2.

Explorons une autre voie…

On a vu plus haut que lorsqu'on ajoute une variable explicative supplémentaire, le R^2 du nouveau modèle « complet » augmente par rapport au R_*^2 du modèle « simplifié » :

$$R^2 = R_*^2 + r_*^2 \frac{\boldsymbol{y}_*' \boldsymbol{y}_*}{SCT} = R_*^2 + r_*^2 \frac{\boldsymbol{y}' \boldsymbol{M_X} \boldsymbol{y}}{SCT}$$

Avec : R_*^2 : le R^2 du modèle « simplifié » sans la $k^{ième}$ variable explicative,
r_*^2 : le coefficient de corrélation partielle entre cette $k^{ième}$ variable explicative et la variable dépendante y,
SCT : la somme des carrés totaux $SCT = \boldsymbol{y}'\boldsymbol{y}$,
$\boldsymbol{M_X} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (sans la $k^{ième}$ variable explicative)

On obtient alors :

$$R^2 = R_*^2 + r_*^2 \frac{\boldsymbol{y}' \boldsymbol{M_X} \boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}} = R_*^2 + r_*^2 \frac{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}} = R_*^2 + r_*^2 \left(1 - \frac{\boldsymbol{y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{\boldsymbol{y}'\boldsymbol{y}}\right)$$

$$R^2 = R_*^2 + r_*^2 (1 - R_*^2)$$

Finalement : $R^2 - R_*^2 = r_*^2(1 - R_*^2)$

Voir Theil, formule 3.12, p.175.

Ceci représente la contribution incrémentale (marginale) au R^2 de la $k^{ième}$ variable explicative à l'ajustement de la régression.

C'est le produit de deux facteurs positifs au plus égal à 1 :

- le carré de la corrélation partielle de la $k^{ième}$ variable explicative avec la variable dépendante y,
- la part inexpliquée (par les variables de la régression « simplifiée ») de la variation de la variable dépendante y.

On peut aussi ré-écrire cela sous la forme (*Theil, formule 3.1, p.172*) :

$$1 - R^2 = (1 - r_*^2)(1 - R_*^2)$$

I.7. Les observations influentes

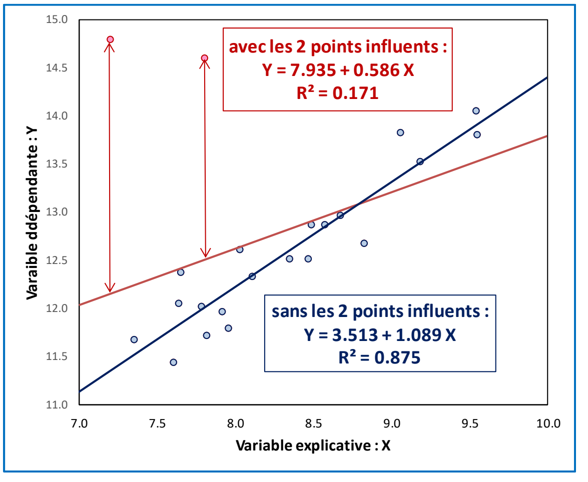
a) *L'importance d'une observation : le levier*

Dans la méthode d'estimation des MCO, plus l'erreur d'une observation est grande, plus elle influence les paramètres estimés.

$$\hat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \varepsilon_i^2$$

La méthode des MCO est sensible aux **observations extrêmes** (outliers) avec une forte erreur au détriment des observations avec une erreur faible …

Exemple



Peut-on détecter les observations les plus influentes dans les MCO ?

Dans quelle mesure l'estimation par MCO est-elle modifiée lorsqu'on retire ou on ajoute une observation ?

Voir le livre de David BELSLEY, Edwin KUH, et Roy WELSH (1980) : "Regression diagnostics: identifying influential data and sources of collinearity ". New York: John Wiley & Sons.

Remarque :

- Si on ajoute une observation qui correspond à la moyenne de l'échantillon $(\bar{y}, \bar{X})$, celle-ci est déjà sur le plan de régression (à condition d'avoir une constante).
- Cette observation n'influence pas la régression parce que son résidu sera nul !
- Donc il est important de savoir si une observation est éloignée de la moyenne de l'échantillon…

Estimateur *MCO* sur l'ensemble de l'échantillon : $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

Si on retire la $i^{ème}$ observation, on aura un estimateur des *MCO* :

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \left(\frac{1}{1-h_i}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i' e_i \quad \text{avec } e_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$$

(voir démonstration ci-dessous)

DÉFINITION : le scalaire h_i est appelé le « levier » (leverage) de l'observation i.

Il est calculé comme le $i^{ème}$ élément sur la diagonale principale de la matrice $\boldsymbol{H}$:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \quad \rightarrow \quad h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$$

DÉMONSTRATION

Si on enlève la $i^{ème}$ observation de l'échantillon, on définit $\boldsymbol{y}_{(i)}$ le vecteur de la variable dépendante, et $\boldsymbol{X}_{(i)}$ la matrice des variables explicatives sans l'observation $\boldsymbol{y}_i$ et $\boldsymbol{x}_i$.
En conséquence :

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)} + \boldsymbol{x}_i\boldsymbol{x}_i' \quad \text{et} \quad \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}_{(i)}'\boldsymbol{y}_{(i)} + \boldsymbol{x}_i y_i$$

On veut calculer l'inverse de $\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)} = \boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i'\boldsymbol{x}_i$ en utilisant la formule matricielle *(voir démonstration page suivante)* :

$$\text{si} \quad \boldsymbol{W} = \boldsymbol{A} - \boldsymbol{BDC} \quad \text{alors} \quad \boldsymbol{W}^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1}$$

Si on note ici : $\boldsymbol{W} = \boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)}$, $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{B} = \boldsymbol{x}_i$, $\boldsymbol{C} = \boldsymbol{x}_i'$, et $\boldsymbol{D} = \boldsymbol{I}_N$, on aura :

$$\left(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)}\right)^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i(1 - \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i)^{-1}\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

parce que $h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$

Démonstration : l'inverse de la matrice $\boldsymbol{W} = \boldsymbol{A} - \boldsymbol{BDC'}$ est :

$$\boldsymbol{W}^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1}$$

En effet :

$$\boldsymbol{WW}^{-1} = (\boldsymbol{A} - \boldsymbol{BDC})(\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1})$$

$$= \boldsymbol{I} + \boldsymbol{B}(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1} - \boldsymbol{BDCA}^{-1} - \boldsymbol{BDCA}^{-1}\boldsymbol{B}(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1}$$

$$= \boldsymbol{I} - \boldsymbol{B}[(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1} + \boldsymbol{D} + \boldsymbol{DCA}^{-1}\boldsymbol{B}(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}]\boldsymbol{CA}^{-1}$$

$$= \boldsymbol{I} - \boldsymbol{B}[\boldsymbol{D} - (\boldsymbol{I} - \boldsymbol{DCA}^{-1}\boldsymbol{B})(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}]\boldsymbol{CA}^{-1}$$

$$= \boldsymbol{I} - \boldsymbol{B}[\boldsymbol{D} - \boldsymbol{D}(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})(\boldsymbol{D}^{-1} - \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}]\boldsymbol{CA}^{-1}$$

$$= \boldsymbol{I} - \boldsymbol{B}[\boldsymbol{D} - \boldsymbol{D}]\boldsymbol{CA}^{-1}$$

$$= \boldsymbol{I}$$

115

Maintenant comme $\boldsymbol{X}'_{(i)}\boldsymbol{y}_{(i)} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i$, l'estimateur des MCO sur l'échantillon sans la $i^{ème}$ observation de l'échantillon sera :

$$\widehat{\boldsymbol{\beta}_{(i)}} = \left(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)}\right)^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{y}_{(i)} = \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\right](\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i)$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i + \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$- \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i$$

$$\widehat{\boldsymbol{\beta}_{(i)}} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i + \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i\widehat{\boldsymbol{\beta}} - \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i$$

$$\widehat{\boldsymbol{\beta}_{(i)}} = \widehat{\boldsymbol{\beta}} + \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i\widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i - \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i h_i y_i$$

$$\widehat{\boldsymbol{\beta}_{(i)}} = \widehat{\boldsymbol{\beta}} + \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i\widehat{\boldsymbol{\beta}} - \left(1 + \frac{h_i}{1-h_i}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i$$

$$\widehat{\boldsymbol{\beta}_{(i)}} = \widehat{\boldsymbol{\beta}} + \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}'_i\widehat{\boldsymbol{\beta}} - \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i$$

$$\widehat{\boldsymbol{\beta}_{(i)}} = \widehat{\boldsymbol{\beta}} - \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}'_i\widehat{\boldsymbol{\beta}}\right) = \widehat{\boldsymbol{\beta}} - \frac{1}{1-h_i}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i$$

CQFD.

116

La matrice de projection $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ est appelée la « hat matrix » en anglais...
En fait c'est notre matrice de projection $\boldsymbol{P}$ sur le plan engendré par les variables explicatives $\boldsymbol{\mathcal{X}}$.

Par analogie à la valeur calculée (*fitted*) de la variable dépendante : $\widehat{\boldsymbol{y}}$

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

Le levier d'une observation sera élevé si celle-ci diffère de la majorité des observations…

Prenons le cas d'une régression simple : $y_i = \alpha + \beta x_i + \varepsilon_i$, le levier sera :

$$h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i = (1 \quad x_i)\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$= (1 \quad x_i)\frac{1}{N\sum_i x_i^2 - \left(\sum_i x_i\right)^2}\begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{pmatrix}\begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$= \frac{\sum_i x_i^2 - 2x_i \sum_i x_i + N x_i^2}{N\sum_i (x_i - \bar{x})^2}$$

$$= \frac{\sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2 + \frac{1}{N}(\sum_i x_i)^2 - 2x_i \sum_i x_i + N x_i^2}{N\sum_i (x_i - \bar{x})^2}$$

$$= \frac{\sum_i (x_i - \bar{x})^2 + N\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)}{N\sum_i (x_i - \bar{x})^2} = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

Le levier d'une observation sera élevé si celle-ci diffère de la moyenne des observations.
Si $x_i = \bar{x}$, le deuxième terme est nul, et le levier sera : $h_i = 1/N$.

Propriétés des leviers $h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$

1) **Les leviers sont compris entre 0 et 1** : $0 < h_i \leq 1$

C'est une forme quadratique avec une matrice définie positive $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. En conséquence : $h_i > 0$.

De même, $h_i \leq 1$. Considérons le vecteur unitaire $\boldsymbol{\iota}_i$ composé de zéros, sauf une valeur 1 pour le i$^{\text{ème}}$ élément, et calculons :

$$h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i = \boldsymbol{\iota}_i'\boldsymbol{P}\boldsymbol{\iota}_i > 0$$

avec $\boldsymbol{H} = \boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Mais on a vu que $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$. Dès lors :

$$h_i = \boldsymbol{\iota}_i'\boldsymbol{P}\boldsymbol{\iota}_i = \boldsymbol{\iota}_i'(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{\iota}_i = \boldsymbol{\iota}_i'\boldsymbol{\iota}_i - \boldsymbol{\iota}_i'\boldsymbol{M}\boldsymbol{\iota}_i = 1 - \boldsymbol{\iota}_i'\boldsymbol{M}\boldsymbol{\iota}_i$$

parce que $\boldsymbol{\iota}_i'\boldsymbol{\iota}_i = 1$ par définition. Finalement comme $\boldsymbol{\iota}_i'\boldsymbol{M}\boldsymbol{\iota}_i \geq 0$ parce que $\boldsymbol{M}$ est une matrice semi-définie positive, on aura :

$$h_i = 1 - \boldsymbol{\iota}_i'\boldsymbol{M}\boldsymbol{\iota}_i \leq 1$$

2) **La somme des leviers est aussi égale au nombre de variables explicatives :**

$$\sum_{i=1}^{N} h_i = K.$$

Cette somme est la trace de la matrice $\boldsymbol{H}$:

$$\sum_{i=1}^{N} h_i = tr(\boldsymbol{H}) = tr[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] = tr[\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}]$$

par la propriété de circularité de la trace. En conséquence :

$$tr(\boldsymbol{H}) = tr[\boldsymbol{I}_K] = K.$$

Donc la moyenne des leviers est : $\bar{h} = \frac{1}{N}\sum_{i=1}^{N} h_i = \frac{K}{N}$

3) Les leviers sont supérieurs ou égaux à l'inverse du nombre d'observations.

$$\sum_{i=1}^{N} h_i = K.$$

On partitionne : $\boldsymbol{x}_i' = (1\ , \boldsymbol{z}_i')$. Sans perte de généralité, on peut remplacer $\boldsymbol{z}_i$ par les variables centrées autour de leur moyenne : $\widetilde{\boldsymbol{z}}_i = \boldsymbol{z}_i - \bar{\boldsymbol{z}}$.
Dès lors, comme $\widetilde{\boldsymbol{z}}_i$ est orthogonal à la constante :

$$\begin{aligned} h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i &= (1 \quad \widetilde{\boldsymbol{z}}_i') \begin{pmatrix} N & 0 \\ 0 & \widetilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{Z}} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \widetilde{\boldsymbol{z}}_i \end{pmatrix} \\ &= (1 \quad \widetilde{\boldsymbol{z}}_i') \begin{pmatrix} 1/N & 0 \\ 0 & (\widetilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{Z}})^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \widetilde{\boldsymbol{z}}_i \end{pmatrix} \\ &= \frac{1}{N} + \widetilde{\boldsymbol{z}}_i'(\widetilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{Z}})^{-1}\widetilde{\boldsymbol{z}}_i \geq \frac{1}{N} \end{aligned}$$

Dans la formule de l'estimateur *MCO* sans la *i^ème^* observation :

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \left(\frac{1}{1-h_i}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i' e_i$$

Plus le levier de la *i^ème^* observation sera proche de 1, plus l'observation aura du poids dans les MCO… et son absence modifiera fortement l'estimateur !

Elle sera donc très influente sur la régression.

Donc on peut calculer, pour chacune des *N* observations de l'échantillon, leurs leviers (leverage) respectifs h_i pour savoir quelles sont les observations les plus influentes…

On peut considérer que les observations avec un levier $h_i > 2\bar{h} = 2K/N$ sont « influentes » et méritent une investigation plus poussée…

Que faire si on détecte une ou plusieurs observations extrêmes (outliers) ?

1. Eliminer l'observation …

2. Corriger l'observation …

3. Mettre une indicatrice pour cette observation …
   … cela équivaut à l'éliminer !