ÉCONOMETRIE THÉORIQUE

Chapitre II

Propriétés en petits échantillons de l'estimateur des MCO

Benoît Mulkay
Université de Montpellier
2023 - 2024

1

Chapitre II :

Propriétés en petits échantillons de l'estimateur des MCO

2

II.1. Propriétés en échantillons finis

Voir : Fumio HAYASHI, *Econometrics* [2000], Section I.3.
Walter GREENE, *Econométrie* [2018], Chapitre IV.

a) *L'estimateur des MCO est linéaire*

Hypothèses H1 à H4 sur les erreurs : $\varepsilon_i|\boldsymbol{X} \sim i.i.d.(0\,,\sigma^2)$
Distribution conditionnelle de la variable dépendante :

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \qquad \Rightarrow \qquad \begin{cases} E(y_i|\boldsymbol{X}) = \boldsymbol{x}_i'\boldsymbol{\beta} \\ V(y_i|\boldsymbol{X}) = \sigma^2 \end{cases}$$

L'estimateur des MCO : $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = f(\boldsymbol{y}|\boldsymbol{X})$ est **une variable aléatoire**…
… parce que l'estimateur dépend de $\boldsymbol{y}$ qui est une variable aléatoire.

→ Quel est son espérance , sa variance, sa distribution ?

L'estimateur des MCO est une **combinaison linéaire** des variables aléatoires $\boldsymbol{y}$:

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

avec une matrice $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ de dimension $K \times N$ dépendant des variables explicatives $\boldsymbol{X}$.

Donc chaque composante de l'estimateur des MCO $\widehat{\beta_k}$ est **une combinaison linéaire des variables aléatoires *y*** avec des poids donnés par la $k^{ème}$ ligne de la matrice $\boldsymbol{A}$.
(Remarquez que ces poids sont aussi des variables aléatoires…).

$$\widehat{\beta_k} = \sum_{i=1}^{N} a_{k,i} y_i \qquad \text{pour } k = 1,2, \dots, K.$$

Donc on parle d'un **estimateur linéaire** pour l'estimateur des MCO !

b) *L'estimateur des MCO est sans biais.*

En utilisant le vrai modèle (**hypothèse H1**) : $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, on peut réécrire l'estimateur des MCO (*d'une manière théorique*) :

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

DÉFINITION : **l'erreur d'échantillonnage** (***sampling error***) dans l'estimation de $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} = \boldsymbol{A\varepsilon} \quad \text{avec} \quad \boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

Remarquez que cette erreur d'échantillonnage est inconnue parce que $\boldsymbol{\beta}$ et $\boldsymbol{\varepsilon}$ sont inconnus !

Une première propriété intéressante pour un estimateur est **l'absence de biais**, c'est-à-dire qu'en espérance (moyenne), l'estimateur donne la vraie valeur du paramètre inconnu, ou que l'espérance de l'erreur d'échantillonnage est nulle !

DÉFINITION : Un estimateur d'un paramètre $\boldsymbol{\theta}$ est SANS BIAIS (***unbiased***) si et seulement si $E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ ou $E(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \boldsymbol{0}$

PROPRIÉTÉ : L'estimateur des moindres carrés est **sans biais** (conditionnellement à $\boldsymbol{X}$) :

$$E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta} \quad \Leftrightarrow \quad \underbrace{E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X})}_{\textit{Biais}} = \boldsymbol{0}_K$$

En moyenne, l'erreur d'échantillonnage est nulle.
Pour cela, il suffit uniquement que les variables explicatives soient **strictement exogènes** (hypothèse **H2**) : $E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}$.

Avec la loi de l'espérance totale, l'espérance inconditionnelle de l'estimateur $\widehat{\boldsymbol{\beta}}$ est :

$$E(\widehat{\boldsymbol{\beta}}) = E_{\boldsymbol{X}}\left(E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})\right) = E_{\boldsymbol{X}}(\boldsymbol{\beta}) = \boldsymbol{\beta}$$

DÉMONSTRATION

L'erreur d'échantillonnage de l'estimateur des MCO s'écrit :

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

En prenant son espérance conditionnellement à X :

$$E\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \middle| \boldsymbol{X}\right) = E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right) = E(\underset{\uparrow}{\boldsymbol{A}}\boldsymbol{\varepsilon}|\boldsymbol{X})$$

Combinaison linéaire de V.A.

Du fait de l'hypothèse **H2** : $E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}_{\boldsymbol{N}}$, et de la linéarité des espérances conditionnelles, on obtient :

$$E\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \middle| \boldsymbol{X}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{0}_{\boldsymbol{K}}$$

En conséquence : $E\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \middle| \boldsymbol{X}\right) = \boldsymbol{0}_{\boldsymbol{K}} \quad \Rightarrow \quad E\left(\hat{\boldsymbol{\beta}} \middle| \boldsymbol{X}\right) = \boldsymbol{\beta}$

c) Une simulation

- Pour vérifier cette propriété d'absence de biais de l'estimateur des MCO, on va simuler un modèle dont on connaît les paramètres $\boldsymbol{\beta}$.

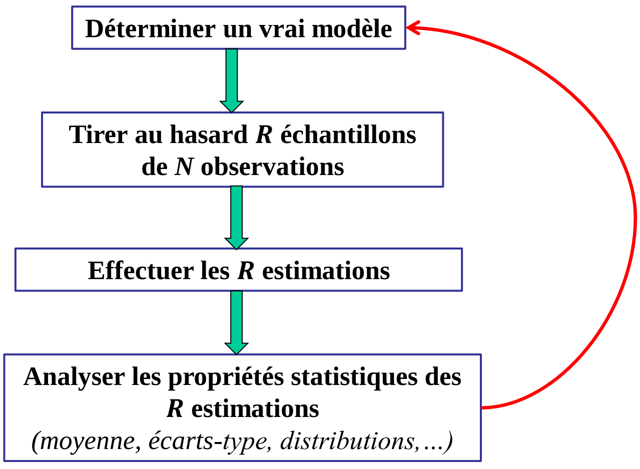
- On construit donc des variables explicatives $\boldsymbol{X}$,
- On tire au hasard un vecteur d'erreurs $\boldsymbol{\varepsilon}$.

- On calcule la variable dépendante en utilisant le vrai modèle : $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- On va ensuite estimer les paramètres par MCO sur cet échantillon construit.

- En recommençant un certain nombre de fois la simulation, on peut étudier les statistiques descriptives et la distribution des paramètres estimés…

→ SIMULATION DE MONTE-CARLO

Structure d’une simulation de Monte-Carlo



1) le « vrai » modèle

- Modèle de régression multiple avec deux variables explicatives et une constante :

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i \quad \text{avec} \quad \begin{cases} \beta_1 = 2.0 \\ \beta_2 = 1.0 \\ \beta_3 = -1.0 \end{cases}$$

- Le terme d’erreur est tiré aléatoirement dans une **loi uniforme** (non normale !).

$$\varepsilon_i \sim i.i.d.\, U(-3\,,3) \quad \rightarrow \quad \begin{cases} E(\varepsilon_i) = 0 \\ E(\varepsilon_i^2) = \sigma^2 = 3 \end{cases}$$

- Les variables explicatives x_1 et x_2 sont choisies **une fois pour toutes**. Elles ont une distribution normale bivariée :

$$\begin{pmatrix} x_{2,i} \\ x_{3,i} \end{pmatrix} \sim i.i.d.\, N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}\right] \quad \rightarrow \quad Corr(x_2, x_3) = \frac{3}{\sqrt{5 \times 2}} \cong 0.95$$

2) Les échantillons

- On choisit un nombre d'observations : $\underline{\boldsymbol{N} = \mathbf{20}}$ **observations**.
- On va tirer $\underline{\boldsymbol{R} = \mathbf{1\,000}}$ **échantillons** de 20 observations, avec des vecteurs d'erreurs $\boldsymbol{\varepsilon}$ uniformément distribués : $\varepsilon_i \sim i.i.d.\, U(-3\,,3)$

- On calcule les vecteurs de la variable dépendante $\boldsymbol{y}$ selon le vrai modèle :

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i \quad avec \quad \begin{cases} \beta_1 = 2.0 \\ \beta_2 = 1.0 \\ \beta_3 = -1.0 \end{cases}$$

3) Les estimations

- On estime alors le modèle par moindres carrés ordinaires sur chacun des 1 000 échantillons : $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- On obtient alors 1 000 vecteurs de paramètres estimés :

$$\widehat{\boldsymbol{\beta}_{(r)}} = \begin{pmatrix} \widehat{\beta_{1,(r)}} \\ \widehat{\beta_{2,(r)}} \\ \widehat{\beta_{3,(r)}} \end{pmatrix}, \qquad r = 1,2,\dots,1000.$$

dont on étudie les statistiques descriptives (*moyenne*, *écart-type*, *minimum, médiane, maximum, …*)

4) Les résultats

Simulations réalisées avec le logiciel Stata.
Voir le programme Simulation_1.do sur Moodle

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

	Vrai paramètre	Moyenne
$\widehat{\beta_1}$	2.0000	1.9990
$\widehat{\beta_2}$	1.0000	1.0066
$\widehat{\beta_3}$	– 1.0000	– 0.9972

1 000 échantillons de 20 observations

Très peu d'écart en moyenne entre le vrai paramètre et les paramètres estimés…

→ Absence de Biais

4) Les résultats (2)

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

	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$
Moyenne	1.9990	1.0066	– 0.9972
Ecart-type	0.4145	0.6115	1.0393
Minimum	0.7587	– 1.0903	– 4.3117
Médiane	2.0109	1.0036	– 0.9987
Maximum	3.1950	2.9568	2.2831
Ecart-type de la moyenne	*0.0131*	*0.0193*	*0.0329*

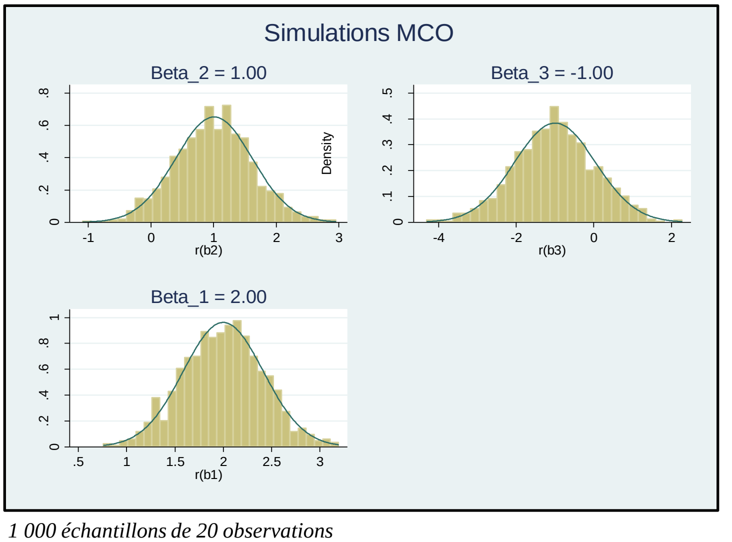
1 000 échantillons de 20 observations

Mais une très forte dispersion d'un échantillon à l'autre…

Remarquez que l'écart-type de la moyenne permet de construire des intervalles de confiance comprenant le vrai paramètre !

4) Les résultats (3)

Histogrammes et courbes de densité empirique



1 000 échantillons de 20 observations

d) *Le biais des variables omises*

Que se passe-t-il si on oublie une ou plusieurs variables explicatives pertinentes ? Soit le vrai modèle ou le modèle correct :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$$

mais on effectue la régression avec les seules variables explicatives $\boldsymbol{X}$, en **omettant** les variables explicatives $\boldsymbol{Z}$: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\omega}$

Le terme d'erreur $\boldsymbol{\omega} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$ contient alors l'effet des variables omises $\boldsymbol{Z}$:

$$\boldsymbol{y} = \boldsymbol{X\beta} + (\boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}) = \boldsymbol{X\beta} + \boldsymbol{\omega}$$

L'estimateur des moindres carrés de $\boldsymbol{\beta}$ est toujours : $\overline{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$

Mais cet estimateur *MCO* est désormais **<u>biaisé</u>** :

$$\begin{aligned} E(\overline{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{Z}) &= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'y}|\boldsymbol{X},\boldsymbol{Z}) \\ &= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon})|\boldsymbol{X},\boldsymbol{Z}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma} \\ &\neq \boldsymbol{\beta} \qquad \text{(en général)} \end{aligned}$$

Le biais de l'estimateur MCO de la régression **sans** les variables $\boldsymbol{Z}$ est alors :

$$Biais = E(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Z}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma}$$

Ce **biais de variables omises** dépend :

- de la valeur du paramètre $\boldsymbol{\gamma}$
- de la matrice $(\boldsymbol{X'X})^{-1}\boldsymbol{X'Z}$

Il n'y aura pas de biais (*le biais est nul*) si :

- les variables $\boldsymbol{Z}$ n'ont aucun effet sur la variable dépendante $\boldsymbol{y}$: $\boldsymbol{\gamma} = \boldsymbol{0}$,

ou

- les covariances (corrélations) entre $\boldsymbol{X}$ et $\boldsymbol{Z}$ sont nulles $\boldsymbol{X'Z} = \boldsymbol{0}$:

$$Biais = E(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Z}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma} = \boldsymbol{0}$$

<u>Remarque</u> : Cette seconde condition permet d'obtenir des estimateurs sans biais de $\boldsymbol{\beta}$ et de $\boldsymbol{\gamma}$ en faisant deux régression séparées :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\omega} \quad \text{et} \quad \boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\xi}$$

Si on écrit $\overline{\boldsymbol{\beta}}$ l'estimateur de $\boldsymbol{\beta}$ dans la régression « courte » **sans** les variables $\boldsymbol{Z}$:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\omega}$$

et $\widehat{\boldsymbol{\beta}}$ l'estimateur de $\boldsymbol{\beta}$ dans la régression « longue » **avec** les variables $\boldsymbol{Z}$:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$$

L'estimateur MCO de cette régression « longue » peut se réécrire comme :

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{pmatrix}$$

On peut montrer par le théorème de Frisch – Waugh (*Section I.6.*) que l'estimateur MCO de $\boldsymbol{\beta}$ peut s'écrire comme :

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

Le premier terme du membre de droite est l'estimateur de $\boldsymbol{\beta}$ dans la régression « courte » **sans** les variables $\boldsymbol{Z}$: $\overline{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$

alors que le second terme est l'estimateur d'une régression des variables omises $\boldsymbol{Z}$ sur les variables incluses dans la régression $\boldsymbol{X}$ que multiplie l'estimateur de $\boldsymbol{\gamma}$ dans la régression « longue » :

$$\boldsymbol{Z} = \boldsymbol{X\delta} + \boldsymbol{\nu} \quad \rightarrow \quad \widehat{\boldsymbol{\delta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z}$$

avec $\widehat{\boldsymbol{\delta}}$ la matrice de coefficients de la régression (multivariée) de $\boldsymbol{Z}$ sur $\boldsymbol{X}$.

17

On aura alors : $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} - (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z}\widehat{\boldsymbol{\gamma}} = \overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\gamma}}$

ou encore : $\overline{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\gamma}}$

Cela s'interprète comme :

> l'estimateur de $\boldsymbol{\beta}$ dans la régression « courte » $(\overline{\boldsymbol{\beta}})$ est égal à l'estimateur de $\boldsymbol{\beta}$ dans la régression « longue » $(\widehat{\boldsymbol{\beta}})$ plus l'effet des variables omises $(\widehat{\boldsymbol{\gamma}})$ multipliée par le vecteur des paramètres $(\widehat{\boldsymbol{\delta}})$ de la régression des variables omises $(\boldsymbol{Z})$ sur les variables incluses $(\boldsymbol{X})$.

Voir une dérivation intuitive du biais de variables omises dans le livre de Joshua ANGRIST et Jörn-Steffen PISCHKE : « Mastering 'Metrics : The Path From Cause to Effect » (2015), Princeton University Press, Chapter 2.

18

Dans le cas général, on ne peut pas prévoir le sens du biais : positif ou négatif ?
Mais dans le **cas particulier** où il y a une seule variable incluse dans $\boldsymbol{X} = \boldsymbol{x}$ et une seule variable omise dans $\boldsymbol{Z} = \boldsymbol{z}$ (car ce sont des vecteurs), on aura pour le biais :

$$E\left(\hat{\beta} - \beta \middle| \boldsymbol{X}, \boldsymbol{Z}\right) = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{z}\boldsymbol{\gamma} = \frac{\boldsymbol{x}'\boldsymbol{z}}{\boldsymbol{x}'\boldsymbol{x}}\boldsymbol{\gamma}$$

Au dénominateur : somme des carrés des $\boldsymbol{x}$ est toujours positive : $\boldsymbol{x}'\boldsymbol{x} > \boldsymbol{0}$
Au numérateur : somme des produits croisés des $\boldsymbol{x}$ et des $\boldsymbol{z}$ est positive ou négative selon le signe de la corrélation entre ces variables.

Le biais aura le signe de cette corrélation (si $\boldsymbol{\gamma}$ est positif) !

Si $\boldsymbol{x}$ et $\boldsymbol{z}$ sont corrélés positivement : $\boldsymbol{x}'\boldsymbol{z} > 0$
→ Omettre la variable $\boldsymbol{z}$ augmente l'estimation de $\boldsymbol{\beta}$ dans le cas où $\boldsymbol{\gamma} > \boldsymbol{0}$.

Le problème de l'omission des variables $\boldsymbol{Z}$ revient à imposer une restriction non pertinente :
→ prendre en compte la contrainte $\boldsymbol{\gamma} = \boldsymbol{0}$
→ l'estimateur de $\boldsymbol{\beta}$ est biaisé…
→ sa variance est incorrecte → pas d'inférence !

Simulations pour le biais de variable omise

On va procéder à des simulations pour démontrer le biais des estimateurs MCO si on exclut une variable explicative pertinente dans l'estimation.

Modèle « *correct* » avec 2 variables explicatives (sans constante) :

$$y_i = \beta x_i + \gamma z_i + \varepsilon_i \quad \text{avec} \quad \varepsilon_i \approx N(0; 1)$$

$$\beta = 1.00 \quad \text{et} \quad \gamma = 1.00$$

Les variables explicatives sont tirées au hasard dans une distribution uniforme entre 0 et 1.

Pour éviter trop d'erreurs d'échantillonnage, on prend un nombre élevé d'observations : ici $N = 10\ 000$

On va étudier les cas où la corrélation entre les deux variables est :
1) importante : $Corr(x, z) = 0.80$
2) nulle : $Corr(x, z) = 0.00$

cas 1 : $\rho = 0.80$

	y	x	z
Moyenne	1.2118	0.5023	0.7034
Ecart-type	0.7485	0.2878	0.2878
Minimum	– 1.3416	0.0001	0.0197
Maximum	3.5462	0.9996	1.3865

10 000 observations

Matrice de corrélation

	y	x	z
y	1.0000		
x	0.7025	1.0000	
z	0.7022	0.7994	1.0000

La corrélation empirique entre les deux variables explicatives (0.7994) est proche de la corrélation théorique (0.80).
On a aussi : $x'y = 7\,600.443$, $x'x = 3\,351.356$ et $x'z = 4\,192.936$

21

cas 1 : $\rho = 0.80$

On estime le modèle « correct » : $y_i = \beta x_i + \gamma z_i + \varepsilon_i$

$$\hat{y}_i = 1.0217\, x_i + 0.9960\, z_i\,, \qquad R^2 = 0.8753$$

Remarquez que les estimateurs sont proches de leur vraie valeur (1.00).
On peut considérer que le **biais est faible** !

On estime le modèle « *faux* » sans la variable z : $y_i = \beta x_i + \varepsilon_i$

$$\hat{y}_i = 2.2679\, x_i\,, \qquad R^2 = 0.8496$$

L'estimateur aurait pu être calculé comme : $\bar{\beta} = \frac{x'y}{x'x} = \frac{7\,600.443}{3\,351.356} = 2.2679$

qui est très éloigné de sa vraie valeur $\beta = 1.0000$.
On peut considérer que le **biais est important** !

22

cas 1 : $\rho = 0.80$

Dans ce cas, l'estimateur du modèle « *faux* » ($\bar{\beta} = 2.2679$) est relié à l'estimateur du modèle « *complet* » ($\hat{\beta} = 1.0217$) par la relation : $\bar{\beta} = \hat{\beta} + \hat{\delta}\hat{\gamma}$,

avec $\hat{\delta} = x'z/x'x = 4\,192.936/3\,351.351 = 1.2511$

$$\bar{\beta} = \hat{\beta} + \hat{\delta}\hat{\gamma} = 1.0217 + (1.2511 \times 0.9960) = 2.2679$$

Remarquez que l'estimateur MCO de z sur x : $\quad z_i = \delta x_i + v_i$

$$\hat{z}_i = 1.2511\, x_i\,, \qquad R^2 = 0.9092$$

cas 2 : $\rho = 0.00$

	y	***x***	***z***
Moyenne	1.0130	0.5023	0.5026
Ecart-type	0.6449	0.2878	0.2871
Minimum	– 1.4673	0.0001	0.0000
Maximum	3.3428	0.9996	0.9999

10 000 observations

Matrice de corrélation

	y	***x***	***z***
y	1.0000		
x	0.4434	1.0000	
z	0.4320	– 0.0061	1.0000

La corrélation empirique entre les deux variables explicatives (-0.0061) est proche de la corrélation théorique (0.00).
On a aussi (avec les variables centrées par rapport à leur moyenne) :

$$\tilde{x}'\tilde{y} = 822.141 \quad , \qquad \tilde{x}'\tilde{x} = 828.141 \quad \text{et} \quad \tilde{x}'\tilde{z} = -5.053$$

cas 2 : $\rho = 0.00$

On estime le modèle « correct » : $y_i = \beta x_i + \gamma z_i + \varepsilon_i$

$$\hat{y}_i = 1.0168\, x_i + 0.9937\, z_i\,, \qquad R^2 = 0.8228$$

Remarquez que les estimateurs sont proches de leur vraie valeur (1.00).
On peut considérer que le **biais est faible** !

On estime le modèle « *faux* » sans la variable *z* : $y_i = \beta x_i + \varepsilon_i$

$$\hat{y}_i = 0.5139 + 0.9937\, x_i\,, \qquad R^2 = 0.1966$$

L'estimateur aurait pu être calculé comme : $\bar{\beta} = \dfrac{\tilde{x}'\tilde{y}}{\tilde{x}'\tilde{x}} = \dfrac{822.931}{828.141} = 0.9937$

qui est très proche aussi de sa vraie valeur $\beta = 1.000$!
On peut considérer que le **biais est insignifiant** !

Ce qui montre que si les variables ne sont pas corrélées entre elles, il n'y a pas de biais en omettant la variable non corrélée dans la régression !

e) La matrice de variance-covariance de l'estimateur des MCO

Comme $\hat{\boldsymbol{\beta}}$ est un vecteur de variables aléatoires de dimension (*K* x *1*),
sa matrice de variance-covariance de dimension (*K* x *K*) est définie par :

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

parce que l'estimateur des MCO est sans biais : $E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta}$.

Cette variance dépend de l'erreur d'échantillonnage :

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} = \boldsymbol{A}\boldsymbol{\varepsilon} \quad \text{avec} \quad \boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

On obtient avec les hypothèses **H4a** (homoscédasticité) et **H4b** (absence d'autocorrélation) :

$$\boxed{V(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}}$$

DÉMONSTRATION

$V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = E[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'|\boldsymbol{X}] = E[(\boldsymbol{A\varepsilon})(\boldsymbol{A\varepsilon})'|\boldsymbol{X}]$

$= E[\boldsymbol{A\varepsilon\varepsilon' A'}|\boldsymbol{X}]$ $\longrightarrow$ *La transposée d'un produit est égal au produit des transposées dans l'ordre inverse !!!*

$= \boldsymbol{A}E[\boldsymbol{\varepsilon\varepsilon'}|\boldsymbol{X}]\boldsymbol{A'}$ $\longrightarrow$ *parce que A dépend uniquement de* $\boldsymbol{X}$

$= \boldsymbol{A}(\sigma^2 \boldsymbol{I_N})\boldsymbol{A'}$ $\longrightarrow$ *Hypothèses H4a et H4b sur les erreurs* $\boldsymbol{\varepsilon}$ *:*

$V[\boldsymbol{\varepsilon}|\boldsymbol{X}] = E[\boldsymbol{\varepsilon\varepsilon'}|\boldsymbol{X}] = \sigma^2 \boldsymbol{I_N}$

$= \sigma^2 \boldsymbol{AA'}$

$= \sigma^2 (\boldsymbol{X'X})^{-1}$ $\longrightarrow$ $\boldsymbol{AA'} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X})^{-1} = (\boldsymbol{X'X})^{-1}$

CQFD

Cette matrice de variance covariance est symétrique de dimension (K x K).
Elle contient les **variances des paramètres estimés** sur sa diagonale principale :

$$V(\widehat{\beta_k}|\boldsymbol{X}) = \sigma^2[(\boldsymbol{X'X})^{-1}]_{k,k}$$

et **les covariances entre deux paramètres estimés** dans son triangle inférieur (ou supérieur) :

$$Cov(\widehat{\beta_k},\widehat{\beta_l}|\boldsymbol{X}) = \sigma^2[(\boldsymbol{X'X})^{-1}]_{k,l}$$

Souvent on présente **l'écart-type des paramètres estimés** (*plutôt que leur variance*) :

$$s(\widehat{\beta_k}|\boldsymbol{X}) = \sqrt{V(\widehat{\beta_k}|\boldsymbol{X})} = \sqrt{\sigma^2[(\boldsymbol{X'X})^{-1}]_{k,k}} = \sigma\left([(\boldsymbol{X'X})^{-1}]_{k,k}\right)^{1/2}$$

f) L'estimateur de la variance des MCO

Pour pouvoir calculer la matrice de variance-covariance de l'estimateur, on doit estimer la variance de l'erreur : σ^2

On a proposé précédemment un estimateur des MCO de cette variance :

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

C'est un estimateur **sans biais** de σ^2 :

$$E\left(\widehat{\sigma^2}\middle|\boldsymbol{X}\right) = E\left(\frac{SCR}{N-K}\middle|\boldsymbol{X}\right) = \frac{E(SCR|\boldsymbol{X})}{N-K} = \frac{\sigma^2(N-K)}{N-K} = \sigma^2$$

Voir démonstration de l'avant-dernière égalité dans les pages suivantes

DÉMONSTRATION

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

L'espérance conditionnelle à X de la **somme des carrés des résidus** est proportionnelle à la trace de la matrice $\boldsymbol{M} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$:

$$E(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{X}) = E(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}) \quad \text{parce que} \quad \boldsymbol{e} = \boldsymbol{M}\boldsymbol{\varepsilon}$$

On a l'espérance d'un scalaire. Or un scalaire est égal à sa trace qui est un opérateur linéaire (c'est une somme). Donc :

$$E(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}) = tr\big(E(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X})\big) = E(tr(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon})|\boldsymbol{X})$$

Comme l'espérance et la trace sont des opérateurs linéaires, on peut intervertir leur ordre de calcul.

En utilisant la propriété de circularité de la trace d'une matrice : $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$, on obtient :

$$E(SCR|\boldsymbol{X}) = E(tr(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')|\boldsymbol{X}) = tr\big(E(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\big)$$

La matrice $\boldsymbol{M}$ ne dépendant que de $\boldsymbol{X}$, on peut la « sortir » du signe d'espérance :

$$E(SCR|\boldsymbol{X}) = tr\big(\boldsymbol{M}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\big)$$

Maintenant du fait des hypothèses H4a (homoscédasticité) et H4b (absence d'autocorrélation) sur les erreurs $\boldsymbol{\varepsilon}$: $V[\boldsymbol{\varepsilon}|\boldsymbol{X}] = E[\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X}] = \sigma^2 \boldsymbol{I}_N$

Ce qui donne : $E(SCR|\boldsymbol{X}) = tr(\boldsymbol{M}\sigma^2\boldsymbol{I}_N) = tr(\sigma^2\boldsymbol{M}) = \sigma^2 tr(\boldsymbol{M})$

Seconde étape : **quelle est la valeur de $tr(\boldsymbol{M})$?**

$$
\begin{aligned}
tr(\boldsymbol{M}) &= \boldsymbol{tr}\left(\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\\
&= \boldsymbol{tr}(\boldsymbol{I}_N) - \boldsymbol{tr}\left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\\
&= \boldsymbol{tr}(\boldsymbol{I}_N) - \boldsymbol{tr}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\right) \longrightarrow\\
&= \boldsymbol{tr}(\boldsymbol{I}_N) - \boldsymbol{tr}(\boldsymbol{I}_K)\\
&= N - K
\end{aligned}
$$

Propriété de circularité de la trace : $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$

Finalement : $E(SCR|\boldsymbol{X}) = \sigma^2 tr(\boldsymbol{M}) = \sigma^2(N-K)$

Donc : $E\left(\widehat{\sigma^2}|\boldsymbol{X}\right) = \dfrac{E(SCR|\boldsymbol{X})}{N-K} = \dfrac{\sigma^2(N-K)}{N-K} = \sigma^2$

$\widehat{\sigma^2}$ est un estimateur sans biais de σ^2. **CQFD**

Démonstration alternative de la seconde étape basée sur les propriétés des matrices idempotentes

Quelle est la valeur de tr(M) ?

La matrice $\boldsymbol{M}$ est une matrice $N \times N$ idempotente.
Une matrice idempotente a des valeurs propres λ_j égales à 0 ou 1.

Elle a des valeurs propres $\lambda_j = 1$ en nombre : $rang(\boldsymbol{M}) = N - K$,
et des valeurs propres $\lambda_j = 0$ en nombre : $ordre(\boldsymbol{M}) - rang(\boldsymbol{M}) = K$.

Comme la trace d'une matrice est la somme de ses valeurs propres :

$$tr(\boldsymbol{M}) = \sum_{j=1}^{N} \lambda_j = N - K$$

CQFD.

De cette façon, on prouve également que $\boldsymbol{M}$ n'est pas inversible (sauf la matrice identité) parce que le déterminant d'une matrice est le produit de ses valeurs propres :

$$\det(\boldsymbol{M}) = \prod_{j=1}^{N} \lambda_j = 0 \quad \Longrightarrow \quad \boldsymbol{M} \text{ est singulière.}$$

Résultats des simulations (1)

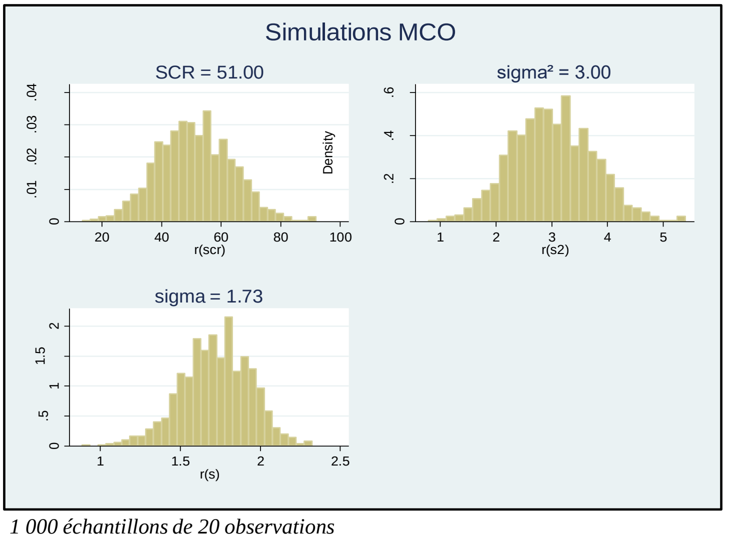
$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$$

	SCR	$\widehat{\sigma^2}$	$\hat{\sigma}$
Vraie valeur	***51.0000***	***3.0000***	***1.7321***
Moyenne	51.2395	3.0141	1.7225
Ecart-type	12.5544	0.7385	1.0393
Minimum	13.3680	0.7864	0.8868
Médiane	50.7432	2.9849	1.7277
Maximum	91.9049	5.4062	2.3251

1 000 échantillons de 20 observations

Résultats des simulations (2)

Histogrammes et courbes de densité empirique



1 000 échantillons de 20 observations

On peut alors avoir un **estimateur de la matrice de variance-covariance** de l'estimateur des MCO :

$$V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad \rightarrow \quad \widehat{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \widehat{\sigma^2}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

qui est un estimateur sans biais de la vraie variance de l'estimateur des MCO (sous les hypothèses H4a et H4b) :

$$E\left(\widehat{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})|\boldsymbol{X}\right) = E\left(\widehat{\sigma^2}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}\right) = E\left(\widehat{\sigma^2}|\boldsymbol{X}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})$$

La **<u>variance inconditionnelle de l'estimateur des MCO</u>** s'obtient en utilisant le théorème de la décomposition de la variance :

$$V(y) = V_X[E(y|x)] + E_X[V(y|x)]$$

que l'on applique ici : $V(\widehat{\boldsymbol{\beta}}) = \underbrace{V_X\left[E(\widehat{\boldsymbol{\beta}}|x)\right]}_{= V_X[\boldsymbol{\beta}] = \boldsymbol{0}} + \underbrace{E_X\left[V(\widehat{\boldsymbol{\beta}}|x)\right]}_{\substack{= E_X[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\ = \sigma^2 E_X(\boldsymbol{X}'\boldsymbol{X})^{-1}}} = \sigma^2 E_X(\boldsymbol{X}'\boldsymbol{X})^{-1}$

Le changement entre la variance conditionnelle et la variance inconditionnelle est subtil !

$$V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad \rightarrow \quad \widehat{V}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma^2}E_X(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

On utilise dans la variance inconditionnelle l'espérance de l'inverse de la matrice des moments des régresseurs, et non sa valeur observée…

→ *On se base sur le comportement « moyen » des régresseurs !*

Il faut donc faire des hypothèses supplémentaires sur le processus stochastique des régresseurs.

h) *Le théorème de Gauss – Markov*

Théorème important en économétrie qui justifie l'utilisation de l'estimateur des MCO sous les hypothèses **H1** à **H4** !

THÉORÈME DE GAUSS – MARKOV

L'estimateur des moindres carrés des paramètres $\boldsymbol{\beta}$ est le **meilleur parmi les estimateurs linéaires sans biais**.

En anglais : **BLUE** → Best Linear Unbiased Estimator

Comment dire, dans le cas d'un vecteur de paramètres, si un estimateur (sans biais) est **meilleur** qu'un autre estimateur (sans biais) ?

C'est celui qui a la **plus faible variance** ! On dit qu'il est efficace.

DÉFINITION GÉNÉRALE

Considérons 2 estimateurs $\tilde{\boldsymbol{\beta}}$ et $\bar{\boldsymbol{\beta}}$ qui sont **sans biais**,

$\tilde{\boldsymbol{\beta}}$ sera meilleur que $\bar{\boldsymbol{\beta}}$,

si pour tout vecteur non nul : $\boldsymbol{\alpha} \neq \mathbf{0}$, on a $V\left(\boldsymbol{\alpha}'\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) \leq V\left(\boldsymbol{\alpha}'\bar{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)$

ou encore : $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{\alpha} \leq \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\bar{\beta}}\boldsymbol{\alpha}$ avec $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}} = V\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)$ et $\boldsymbol{\Sigma}_{\bar{\beta}} = V\left(\bar{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)$.

Ce qui est vrai si $\left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}} - \boldsymbol{\Sigma}_{\bar{\beta}}\right)$ est une matrice semi-définie positive.

Cette dernière condition peut se réécrire :

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{\alpha} \leq \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\bar{\beta}}\boldsymbol{\alpha} \quad \Leftrightarrow \quad \boldsymbol{\alpha}'\left(\boldsymbol{\Sigma}_{\bar{\beta}} - \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\right)\boldsymbol{\alpha} \geq \mathbf{0} \qquad \text{pour tout } \boldsymbol{\alpha} \neq \mathbf{0}$$

DÉMONSTRATION DU THÉORÈME DE GAUSS – MARKOV

(par l'absurde) :

L'estimateur des moindres carrés est un estimateur linéaire : $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}$ dont la matrice de variance-covariance est donnée par

$$\Sigma_{\hat{\boldsymbol{\beta}}} = V\left(\hat{\boldsymbol{\beta}} \middle| \boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Considérons un autre estimateur linéaire : $\bar{\boldsymbol{\beta}} = \boldsymbol{D}\boldsymbol{y}$ avec $\boldsymbol{D}$ une matrice $K \times N$ de coefficients.

Sans perte de généralité, on peut définir une autre matrice ($K \times N$) de coefficients :

$$\boldsymbol{C} = \boldsymbol{D} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \quad \Leftrightarrow \quad \boldsymbol{D} = \boldsymbol{C} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

Ce qui permet d'écrire l'estimateur alternatif : $\bar{\boldsymbol{\beta}} = \boldsymbol{D}\boldsymbol{y} = (\boldsymbol{C} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y}$

On remplace $\boldsymbol{y}$ par le vrai modèle :

$$\bar{\boldsymbol{\beta}} = (\boldsymbol{C} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

Comme on considère uniquement des **estimateurs sans biais** et que l'hypothèse H3 (exogénéité des régresseurs) est maintenue, on aura :

$$E\left(\bar{\boldsymbol{\beta}} \middle| \boldsymbol{X}\right) = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta} + \boldsymbol{C}\underbrace{E(\boldsymbol{\varepsilon}|\boldsymbol{X})}_{=\boldsymbol{0}_N} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{E(\boldsymbol{\varepsilon}|\boldsymbol{X})}_{=\boldsymbol{0}_N} = \boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

L'estimateur alternatif sera **sans biais** si et seulement si la condition $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}_{K\times K}$ est satisfaite.

Calculons alors la matrice de variance-covariance de cet estimateur alternatif. Premièrement l'erreur d'échantillonnage de l'estimateur alternatif **sans biais** devient :

$$\bar{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\varepsilon} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} = \boldsymbol{C}\boldsymbol{\varepsilon} + \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

Cette matrice de variance-covariance est :

$$V\left(\bar{\boldsymbol{\beta}} \middle| \boldsymbol{X}\right) = E\left[\left(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \middle| \boldsymbol{X}\right] = E[(\boldsymbol{C}\boldsymbol{\varepsilon} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})(\boldsymbol{C}\boldsymbol{\varepsilon} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})'|\boldsymbol{X}]$$

$$= E[\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}'|\boldsymbol{X}] + E[\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}'|\boldsymbol{X}]$$
$$+ E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}]$$

$$= \boldsymbol{C}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}]\boldsymbol{C}' + \boldsymbol{C}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}]\boldsymbol{C}'$$
$$+ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Du fait des hypothèses H4 (homoscédasticité) et H5 (absence d'autocorrélation) sur les erreurs $\boldsymbol{\varepsilon}$: $V[\boldsymbol{\varepsilon}|\boldsymbol{X}] = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}] = \sigma^2 \boldsymbol{I}_N$

Cela donne :

$$V(\overline{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 \boldsymbol{CC}' + \sigma^2 \boldsymbol{CX}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}' + \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Mais l'estimateur $\overline{\boldsymbol{\beta}}$ doit respecter la condition $\boldsymbol{CX} = \boldsymbol{0}_{K\times K}$ ou $\boldsymbol{X}'\boldsymbol{C}' = \boldsymbol{0}_{K\times K}$ pour qu'il soit sans biais !

$$V(\overline{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 \boldsymbol{CC}' + \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Le second terme est la variance de l'estimateur MCO : $V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$

$$V(\overline{\boldsymbol{\beta}}|\boldsymbol{X}) - V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \Sigma_{\overline{\boldsymbol{\beta}}} - \Sigma_{\widehat{\boldsymbol{\beta}}} = \sigma^2 \boldsymbol{CC}'$$

Mais pour toute matrice $\boldsymbol{C}$, $\boldsymbol{CC}'$ est une matrice semi-définie positive.
A fortiori $\sigma^2 \boldsymbol{CC}' \geq \boldsymbol{0}$. Donc $\Sigma_{\widehat{\boldsymbol{\beta}}} - \Sigma_{\overline{\boldsymbol{\beta}}} = \sigma^2 \boldsymbol{CC}' \geq \boldsymbol{0}$

Donc l'estimateur des moindres carrés $\widehat{\boldsymbol{\beta}}$ est un meilleur estimateur que l'estimateur alternatif linéaire sans biais $\overline{\boldsymbol{\beta}}$.

Notons qu'on obtient des estimateurs identiques : $\widehat{\boldsymbol{\beta}} = \overline{\boldsymbol{\beta}}$ si et seulement si $\boldsymbol{C} = \boldsymbol{0}$.
CQFD

Implication du théorème de Gauss – Markov :

- **L'estimateur des MCO a une variance qui est la plus faible parmi les estimateurs linéaires sans biais.**
- Il est le plus précis dans cette classe d'estimateurs.
- L'estimateur des MCO est **efficace** (***efficient*** en anglais)
- Cela ne veut pas dire qu'il est le plus précis dans une classe d'estimateurs plus générale…
- … ou qu'il n'y ait pas un autre estimateur avec une plus faible variance.
- Mais soit cet autre estimateur est biaisé.
- Soit cet autre estimateur n'est pas linéaire.

i) *Absence de corrélation entre les résidus et l'estimateur des MCO*

PROPRIETE : La corrélation (la covariance) entre $\widehat{\boldsymbol{\beta}}$ l'estimateur MCO de $\boldsymbol{\beta}$ et le résidu $\boldsymbol{e}$ est **nulle** !

DEMONSTRATION :

$$Cov(\widehat{\boldsymbol{\beta}}, \boldsymbol{e}|\boldsymbol{X}) = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{e}'|\boldsymbol{X}\right) \longrightarrow E(\boldsymbol{e}|\boldsymbol{X}) = \boldsymbol{M}E(\boldsymbol{\varepsilon}|\boldsymbol{X}) \text{ et } E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})|\boldsymbol{X}\right)$$

$$= E\left(\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right)(\boldsymbol{M}\boldsymbol{\varepsilon})'|\boldsymbol{X}\right) \longrightarrow \text{parce que : } \boldsymbol{e} = \boldsymbol{M}\boldsymbol{\varepsilon}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{M} \longrightarrow \text{parce que } \boldsymbol{M} \text{ est uniquement fonction de } \boldsymbol{X}\text{ .}$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M} \longrightarrow \text{avec les hypothèses H4 et H5 : } E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}] = \sigma^2\boldsymbol{I}_N$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{0}_{K\times N} \longrightarrow \text{parce que : } \boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}_{N\times K}$$

$$= \boldsymbol{0}_{K\times N}$$

Attention : l'absence de corrélation entre 2 vecteurs aléatoires ne signifie pas nécessairement l'indépendance statistique entre ces 2 vecteurs !
(sauf s'ils sont normalement distribués)

j) *La distribution de l'estimateur des MCO*

On a vu que l'estimateur des MCO était une variable aléatoire,
qu'il était sans biais (espérance),
qu'on pouvait calculer sa variance,
qu'il était efficace (meilleur estimateur linéaire sans biais)
sous les Hypothèses H1 à H4 (Hypothèses de Gauss – Markov).

On va supposer maintenant que les erreurs $\boldsymbol{\varepsilon}$ sont normalement distribuées (**Hypothèse H5**) : $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim \mathcal{N}_N(\boldsymbol{0}_N, \sigma^2\boldsymbol{I}_N)$

On peut ainsi déterminer la **distribution de l'estimateur des MCO** en échantillons de taille donnée. *(taille de l'échantillon finie)*

Ce qui permet aussi d'effectuer des tests sur les paramètres estimés …

L'estimateur des moindres carrés est un estimateur linéaire :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y} \quad \text{avec} \quad \boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

Comme les erreurs $\boldsymbol{\varepsilon}$ sont normalement distribuées, la distribution de la variable dépendante $\boldsymbol{y}$ conditionnelle aux variables explicatives $\boldsymbol{X}$ s'écrit :

$$\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}_N(\boldsymbol{X\beta}\,, \sigma^2 \boldsymbol{I}_N) \quad \text{parce que} \quad \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

L'estimateur MCO est une combinaison linéaire de variables normalement distribuées $\boldsymbol{y}$.
En conséquence, l'estimateur des MCO $\widehat{\boldsymbol{\beta}}$ est distribué selon une **loi normale** (multivariée) :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}}\big|\boldsymbol{X} \sim \mathcal{N}_K(\boldsymbol{\beta}\,, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

k) La distribution de l'estimateur de la variance

L'estimateur des moindres carrés de la variance de l'erreur : $\widehat{\sigma^2} = \dfrac{\boldsymbol{e}'\boldsymbol{e}}{N-K}$
avec $\boldsymbol{e}$ le vecteur des résidus de la régression par MCO :

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{My} = \boldsymbol{M\varepsilon}$$

où $\boldsymbol{M} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ est une matrice symétrique et idempotente de rang $N-K$.

Le vecteur des résidus $\boldsymbol{e}$ est une combinaison linéaire de variable aléatoires normalement distribuées $\boldsymbol{\varepsilon}$:

$$\boldsymbol{e} = \boldsymbol{M\varepsilon}|\boldsymbol{X} \sim \mathcal{N}_N(\boldsymbol{0}_N, \sigma^2\boldsymbol{M})$$

Donc la somme des carrés des résidus est une forme quadratique de variables aléatoires indépendantes et normalement distribuées :

$$SCR = \boldsymbol{e}'\boldsymbol{e} = \boldsymbol{\varepsilon}'\boldsymbol{M\varepsilon} \quad \text{avec} \quad \boldsymbol{\varepsilon}|\boldsymbol{X} \sim \mathcal{N}_N(\boldsymbol{0}_N, \sigma^2\boldsymbol{I}_N)$$

DÉFINITION : Loi du Khi-deux

Soit le vecteur aléatoire $\boldsymbol{z} \sim \mathcal{N}_N(\boldsymbol{0}\,, \boldsymbol{I}_N)$ et une matrice symétrique et idempotente $\boldsymbol{A}$, la forme quadratique $\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z} \sim \chi^2\big(rang(\boldsymbol{A})\big)$ suit une loi du Khi-deux avec $rang(\boldsymbol{A})$ degrés de liberté.

Pour mémoire : l'espérance et la variance d'une variable aléatoire $\chi^2(R)$, avec R degrés de liberté, sont respectivement :

- $E\big(\chi^2(R)\big) = R$
- $V\big(\chi^2(R)\big) = 2R$

ce qui implique :

- $E\left(\frac{\chi^2(R)}{R}\right) = 1$
- $V\left(\frac{\chi^2(R)}{R}\right) = \frac{2}{R}$

Ici on aura : $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim \mathcal{N}_N(\boldsymbol{0}_N, \sigma^2 \boldsymbol{I}_N) \;\; \Rightarrow \;\; \boldsymbol{z} = \frac{1}{\sigma}\boldsymbol{\varepsilon}\Big|\boldsymbol{X} \sim \mathcal{N}_N(\boldsymbol{0}_N, \boldsymbol{I}_N)$

et $\boldsymbol{M}$ est symétrique, idempotente telle que $rang(\boldsymbol{M}) = tr(\boldsymbol{M}) = N - K$.

Donc :

$$\frac{\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \boldsymbol{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi^2(N-K)$$

L'estimateur des moindres carrés de la variance aura comme distribution :

$$\widehat{\sigma^2} = \frac{\boldsymbol{e}'\boldsymbol{e}}{N-K} = \frac{\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}}{N-K} = \left(\frac{\sigma^2}{N-K}\right) \underbrace{\left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}}{\sigma^2}\right)}_{\sim \chi^2(N-K)}$$

Donc cet estimateur est proportionnel à une loi du Khi-deux à $N-K$ degrés de liberté :

$$\widehat{\sigma^2}|\boldsymbol{X} \sim \left(\frac{\sigma^2}{N-K}\right) \chi^2(N-K)$$

Ce qui donne : $\frac{\widehat{\sigma^2}}{\sigma^2}\bigg|\boldsymbol{X} \sim \frac{\chi^2(N-K)}{N-K}$ ou encore $(N-K)\frac{\widehat{\sigma^2}}{\sigma^2}\bigg|\boldsymbol{X} \sim \chi^2(N-K)$

Donc on obtient comme précédemment (*sous une l'hypothèse plus forte de normalité*) :

$$E\left((N-K)\frac{\widehat{\sigma^2}}{\sigma^2}\middle|\boldsymbol{X}\right)=N-K \quad\to\quad (N-K)\frac{E\left(\widehat{\sigma^2}\middle|\boldsymbol{X}\right)}{\sigma^2}=N-K \quad\to\quad E\left(\widehat{\sigma^2}\middle|\boldsymbol{X}\right)=\sigma^2$$

L'estimateur de la variance $\widehat{\sigma^2}$ est sans biais.
et sa variance sera (*ce qu'on n'avait pas calculé auparavant*) :

$$V\left((N-K)\frac{\widehat{\sigma^2}}{\sigma^2}\middle|\boldsymbol{X}\right)=2(N-K) \quad\to\quad \frac{(N-K)^2}{\sigma^4}V\left(\widehat{\sigma^2}\middle|\boldsymbol{X}\right)=2(N-K) \quad\to\quad V\left(\widehat{\sigma^2}\middle|\boldsymbol{X}\right)=\frac{2\sigma^4}{N-K}$$

Comme les estimateurs $\widehat{\boldsymbol{\beta}}$ et $\boldsymbol{e}$ sont des combinaisons linéaires de $\boldsymbol{\varepsilon}$:

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}=\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \quad \text{et} \quad \boldsymbol{e}=\boldsymbol{M}\boldsymbol{\varepsilon}$$

deux vecteurs de variables aléatoires normalement distribuées,
ils sont conjointement normalement distribués, conditionnellement à $\boldsymbol{X}$.
Ils sont également non corrélés entre eux (voir II.1.h).
En conséquence, $\widehat{\boldsymbol{\beta}}$ et $\boldsymbol{e}$ sont statistiquement indépendants, conditionnellement à $\boldsymbol{X}$:

Ce qui s'écrit : $\widehat{\boldsymbol{\beta}} \perp \boldsymbol{e}|\boldsymbol{X} \quad\to\quad \begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \boldsymbol{e}\end{pmatrix}\middle|\boldsymbol{X}\sim\boldsymbol{N}\left(\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{0}_N\end{pmatrix},\begin{pmatrix}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{0}_{K\times N}\\ \boldsymbol{0}_{N\times K} & \sigma^2\boldsymbol{M}\end{pmatrix}\right)$

THÉORÈME DE STATISTIQUE MATHÉMATIQUE

Si $\boldsymbol{x}$ et $\boldsymbol{z}$ sont des variables aléatoires indépendantes,

alors les fonctions $f(\boldsymbol{x})$ et $f(\boldsymbol{x})$ sont également des variables aléatoires indépendantes entre – elles.

En conséquence, sous l'hypothèse de normalité des erreurs, et conditionnellement à $\boldsymbol{X}$: $\widehat{\boldsymbol{\beta}} \perp \widehat{\boldsymbol{\sigma^2}}|\boldsymbol{X}$ parce que $\widehat{\boldsymbol{\sigma^2}}=\boldsymbol{e}'\boldsymbol{e}/(N-K)$.

L'estimateur des MCO $\widehat{\boldsymbol{\beta}}$ et l'estimateur de la variance $\widehat{\sigma^2}$ sont statistiquement indépendants, conditionnellement à $\boldsymbol{X}$.

II.2. Le faux problème de la multicolinéarité

a) *La multicolinéarité parfaite*

On a vu que la matrice des variables explicatives $\boldsymbol{X}$ doit être de rang plein colonne (**H3**) :

→ **identification des paramètres**
→ **une seule solution aux équations normales**
→ **inversibilité de la matrice** $(\boldsymbol{X}'\boldsymbol{X})$

Il faut donc éviter qu'une variable explicative soit une combinaison linéaire parfaite des autres variables explicatives.

Comment ? → supprimer la variable parfaitement colinéaire aux autres.

La matrice des moments devient alors inversible, et on peut calculer l'estimateur des MCO.

Exemple 1 de colinéarité parfaite :

$$Dépenses = \alpha + \beta Salaires + \gamma Revenus\ non\ salariaux + \delta Revenus\ totaux + \varepsilon$$

mais $Revenus\ totaux = Salaires + Revenus\ non\ salariaux$

Remède : supprimer une des 3 variables ($Revenus\ totaux$, $Salaires$, ou $Revenus\ non\ salariaux$) parfaitement colinéaire aux autres.

Exemple 2 de colinéarité parfaite :

$$Salaire = \alpha + \beta Femme + \gamma Homme + \delta Education + \varepsilon$$

$$\text{avec } Femme = \begin{cases} 1 & si\ la\ personne\ est\ une\ femme \\ 0 & sinon \end{cases}$$

$$Homme = \begin{cases} 1 & si\ la\ personne\ est\ une\ homme \\ 0 & sinon \end{cases}$$

En supposant qu'une personne est soit un homme, soit une femme, on aura :

$$Femme = 1 - Homme$$

et il y aura une multicolinéarité parfaite.

Remède : supprimer une des 2 indicatrices ou la constante :

$$Salaire = \alpha + \beta Femme + \delta Education + \varepsilon$$
$$Salaire = \alpha + \gamma Homme + \delta Education + \varepsilon$$
$$Salaire = \beta Femme + \gamma Homme + \delta Education + \varepsilon$$

Ces 3 modèles donnent la même information…

b) Le problème de la multicolinéarité

Que se passe-t-il lorsque l'on a une très forte corrélation entre les variables explicatives ?

On peut montrer que :

→ la variance des estimateurs devient très grande !
→ les estimateurs deviennent imprécis et peu robustes !

Il est difficile d'estimer séparément les paramètres de variables très fortement corrélées…

On peut réécrire la variance de l'estimateur des MCO pour un paramètre dans une régression multiple :

$$V(\widehat{\beta_k}|\boldsymbol{X}) = \frac{\sigma^2}{SCT_k\left(1 - R^2_{(k)}\right)}$$

avec : $SCT_k = \sum_{i=1}^{N}\left(x_{k,i} - \overline{x_k}\right)^2$ la somme des carrés totaux de la $k^{ème}$ variable explicative,

$R^2_{(k)}$: le R^2 de la régression de x_k sur toutes les autres variables explicatives.

Voir démonstration dans le livre de Jeffrey WOOLDRIDGE (2013) : « Introduction à l'économétrie – Une approche moderne », De Boeck éditeur, Section III.4, p.148-150.

ou dans les pages suivantes …

$$V(\widehat{\beta_k}|\boldsymbol{X}) = \frac{\sigma^2}{SCT_k\left(1 - R^2_{(k)}\right)}$$

1) Plus la variance de l'erreur σ^2 est grande, plus la variance de l'estimateur des MCO est grande.

2) Plus la variabilité (somme des carrés totaux : SCT_k) de la variable explicative x_k est grande, plus faible sera la variance de l'estimateur des MCO.

3) Plus l'information contenue dans la variable explicative x_k est présente dans les autres variables explicatives, plus grande sera la variance de l'estimateur des MCO. Cela signifie que le $R^2_{(k)}$ est important.

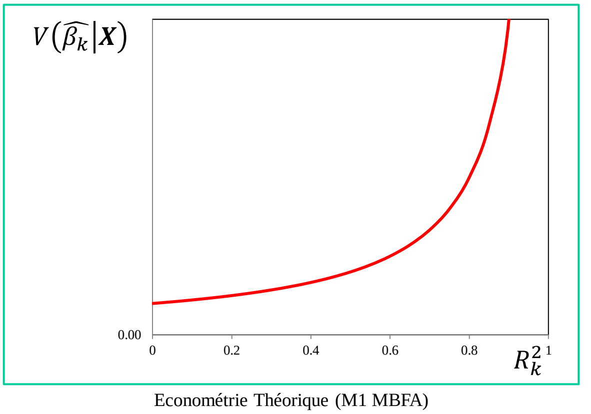
Donc pour améliorer la précision des estimateurs, il faut avoir :

- beaucoup d'observations (SCT_k augmente),
- des variables explicatives avec une grande variance (SCT_k important),
- mais qui ne sont pas fortement corrélées entre elles ($R^2_{(k)}$ faible) !

$$V\left(\widehat{\beta_k}|\boldsymbol{X}\right) = \frac{\sigma^2}{SCT_k\left(1 - R_{(k)}^2\right)}$$

Le cas où $R_{(k)}^2 = 1$ implique que la variance devient infinie : $V\left(\widehat{\beta_k}|\boldsymbol{X}\right) \to \infty$.
Mais ce cas est exclu, si l'hypothèse **H2** d'absence de multicolinéarité parfaite est vérifiée.

La variance augmente lorsque $R_{(k)}^2$ augmente :



Plus la corrélation entre x_k et les autres variables explicatives est élevée, plus la précision de l'estimateur des MCO se détériorera :

→ Problème de *multicolinéarité*

Mais le problème de multicolinéarité est un problème de manque d'observations (ou d'informations) : (faible SCT_k)

→ *micronumérosité* !

Voir la discussion sur la notion de « *micronumérosité** » dans le livre de l'économètre américain Arthur GOLDBERGER (1930 – 2009) : « *A Course in Econometrics* », 1991, Chapitre 23, pages 245-253.

→ * : micronumérosité : faible nombre d'observations

Un exemple :

- Régression multiple avec 2 variables explicatives **fortement positivement corrélée** : $\rho(x_2, x_3) = 0.99$.

- Donc avec une **très forte colinéarité** !!!

- La régression est la suivante :

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i \quad \text{avec} \begin{cases} \beta_1 = 2.00 \\ \beta_2 = 1.00 \\ \beta_3 = -1.00 \\ \varepsilon \sim N(0\,,1) \end{cases}$$

- On essaye différentes tailles d'échantillon : $N = 20, 50, 100, 500,$ et 1000.
- On fait 1 000 essais d'estimation.

- On présente dans le tableau suivant les moyennes des paramètres estimés et de leurs écarts-type, le minimum et le maximum des paramètres estimés et la proportion des cas où ils sont significatifs…

Moyennes sur 1000 essais

		N = 20	*N = 50*	*N = 100*	*N = 500*	*N = 1000*
$\beta_2 = 1.00$	Estimation	0.9727	1.0119	1.0026	0.9982	0.9975
	Écart-type	1.1802	0.7645	0.4937	0.2151	0.1529
	Minimum	-3.6237	-1.8460	-0.8678	0.1278	0.3918
	Maximum	5.7908	3.5889	2.5391	1.8006	1.5905
	% Signif.	4.0 %	27.2 %	51.6 %	99.0 %	100 %
$\beta_3 = -1.00$	Estimation	-0.9748	-1.0072	-1.0017	-0.9977	-0.9973
	Écart-type	0.2393	0.1449	0.1011	0.0448	0.0316
	Minimum	-5.0473	-3.3842	-2.4202	-1.7073	-1.5284
	Maximum	2.9406	1.5992	0.5935	-0.2640	-0.6547
	% Signif.	17.8 %	87.1 %	96.6 %	100 %	100 %

A partir d'une certaine taille d'échantillon, toutes les estimations donnent des paramètres statistiquement significatifs, parce que leurs écarts-type diminuent !

b) *Mesures de la multicolinéarité*

Comment détecter la multicolinéarité ?

- Le R^2 de la régression de x_k sur les autres variables explicatives $X_{(k)}$: $R^2_{(k)}$

- Le **facteur d'inflation de la variance** (***VIF : Variance Inflation Factor***) :

$$VIF_k = \frac{1}{1 - R^2_{(k)}} \qquad \text{pour tout } k = 2, \ldots, K.$$

Ce dernier indique la contribution à l'augmentation de la variance due à la non-orthogonalité entre cette variables et les autres variables explicatives. Plus le *VIF* est élevé, plus il y a multicolinéarité !

- Parfois on aura à la place du *VIF*, la *Tolérance* : $TOL_k = 1 - R^2_{(k)} = \frac{1}{VIF_k}$
Plus la tolérance est faible, plus il y a multicolinéarité !

- **La règle de Klein* :**

On compare le R^2 avec le carré des coefficients de corrélation simple entre les différentes variables explicatives : $r^2_{x_k,x_l}$ pour $k \neq l = 2, \ldots, K$.

Si $R^2 < r^2_{x_k,x_l}$, il y a présomption de multicolinéarité entre les deux variables !

* : Larry R. Klein (1962) : *An Introduction to Econometrics*. Prentice-Hall, p. 101.

- **L'indice de Theil* :**

On a vu dans la Section I.6.d que s'il y avait indépendance entre les variables explicatives :

$$\sum_{k=1}^{K}\left(R^2 - R_k^2\right) = R^2$$

Dans le cas où les variables explicatives sont corrélées entre elles : la différence est négative :

$$Theil = R^2 - \sum_{k=1}^{K}\left(R^2 - R_k^2\right)$$

Theil propose de mesurer l'importance de la multicolinéarité par cette différence.

* : Henri Theil (1971) : *Principles of Econometrics*, John Wiley & Sons.

- **Le nombre de conditionnement** (**Condition number***)

On considère uniquement les colonnes des variables explicatives $\boldsymbol{x}_k$ sans la constante.
On transforme les variables pour avoir une longueur unitaire telle que :

$$x_{k,i}^{+} = \frac{x_{k,i}}{\sqrt{\sum_{i=1}^{n} x_{k,i}^2}} \quad \rightarrow \quad \sum_{i=1}^{n}\left(x_{k,i}^{+}\right)^2 = 1$$

On considère les valeurs propres λ de la matrice des moments de ces variables transformées : $\boldsymbol{X}^{+\prime}\boldsymbol{X}^{+}$.

Le *nombre de conditionnement* de $\boldsymbol{X}'\boldsymbol{X}$ est : $CN = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} > 0$

* : David BELSLEY, Edwin KUH, et Roy WELSH (1980) : *Regression diagnostics: identifying influential data and sources of collinearity*, New York: John Wiley & Sons

Une collinéarité parfaite implique une valeur propre minimale nulle, ce qui implique un nombre de conditionnement tendant vers l'infini …

Belsley, Kuh et Welsch suggèrent qu'une valeur supérieure à 20 indique un problème de multicolinéarité !

- **<u>L'indicateur *Red** :</u>**

On part des valeurs propres λ_j de la matrice de corrélation ***R*** des variables explicatives (sauf la constante).

On calcule alors l'indicateur Red (*pour Redondance*) :

$$Red = \frac{\sum_{k=2}^{K}\left(\lambda_j - 1\right)^2}{k\sqrt{k-1}}$$

Si la valeur de cet indicateur est zéro, alors il y a une absence de redondance, ou de multicolinéarité entre les variables. Elles sont linéairement indépendantes.

Si la valeur de l'indicateur Red est proche de 1, cela indique une forte redondance et multicolinéarité !

* : P. Kovács, T. Petres, and Tóth (2005) : "A new measure of multicollinearity in linear regression models", *International Statistical Review / Revue Internationale de Statistique*, Vol. 73, N°3, pp. 405–412.

D’autres économètres ont proposés d’autres mesures pour détecter quelle variable explicative était plus ou moins colinéaires aux autres …

REMARQUE :

CE NE SONT PAS DES TESTS STATISTIQUES,

MAIS DES MESURES DE LA MULTICOLINEARITE !!!

c) « Test » de la multicolinéarité ?

- **Test de Farrar – Glauber* :**

On calcule la matrice de corrélation $\boldsymbol{R}$ des $K' = K - 1$ variables explicatives (évidemment sans la constante).

Si une variable est combinaison linéaire des autres variables explicatives, le déterminant de cette matrice de corrélation est nul : $|R| = 0$.

Si toutes les variables explicatives sont indépendantes les unes des autres, la matrice de corrélation est la matrice identité et : $|R| = 1$.

Donc on veut tester l’hypothèse nulle : $H_0 \colon |R| = 1$, les variables explicatives sont indépendantes, contre l’hypothèse alternative : $H_1 \colon |R| < 1$.

* : Donald E. Farrar et Robert R. Glauber (1967) "Multicollinearity in Regression Analysis: the Problem Revisited", *Review of Economics and Statistics*, Vol. 49, pp. 92-107.

Si les variables explicatives sont normalement distribuées, la matrice de corrélation suit une loi de Wishart*.
Wilks** a donné les moments du déterminant de cette matrice de corrélation, ce qui a permis à Bartlett*** d'approximer la distribution par une loi du Khi-deux.

Sur cette base, Farrar et Glauber on proposé un premier test :

$$FG = -\left[N - 1 - \frac{2K' + 5}{6}\right] \times \log|R|$$

Cette statistique est distribuée sous l'hypothèse nulle selon une loi du Khi-deux avec $K'(K'-1)/2$ degrés de liberté.

$$FG \approx \chi^2\left(\frac{K'(K'-1)}{2}\right)$$

* : John Wishart (1928) : « The Generalized Product Moment Distribution in Samples from a Multivariate Normal Population », *Biometrika*, Vol. 20A, N°1/2, pp. 32-52.
** : Samuel Wilks (1932) : « Certain Generalizations in the Analysis of Variance », *Biometrika*, Vol. 24. N°3/4, pp. 471-494.
*** : Maurice S. Bartlett (1950) : « Tests of Significance in Factor Analysis », *British Journal of Psychology, Statistical* Vol. 3., pp. 77-85.

Pour un niveau de test $\alpha\%$, on recherche dans les tables le seuil critique correspondant au quantile $\chi^2_{1-\alpha\%}(K'(K'-1)/2)$.

Si la statistique de Farrar et Glauber est supérieure à ce seuil critique, on rejette l'hypothèse nulle → les variables explicatives ne sont pas indépendantes.

Mais cela ne nous dit rien sur l'importance de la multicolinéarité !

Critique du test :

1. On teste l'indépendance des variables explicatives, et non la multicolinéarité (ce qui n'est pas la même hypothèse) !

2. Ce test est basé sur des variables explicatives normalement distribuées !

3. Ce test ne détecte pas vraiment les problèmes d'imprécision dûe à la multicolinéarité !
   (voir par exemple : Yoel Haitovky (1970) : « Multicollinearity in Regression Analysis: Comment », *Review of Economics and Statistics*, 51, pp. 486-489.)

f) Que faire en cas de multicolinéarité ?

- Eviter la colinéarité parfaite : une variable combinaison linéaire des autres !
- Méthodes d'estimations plus complexes :
  - Régression Ridge : mais estimateurs biaisés !!!
  - Transformer les variables explicatives en composantes principales (orthogonales par définition) : mais estimateurs non interprétable !!!
- Problème de données (**voir Greene, 2011**):
  - Augmenter le nombre d'observations…
  - Trouver d'autres variables explicatives « plus » orthogonales
  - Regrouper des variables explicatives
  - Revoir la spécification du modèle

Pour terminer, on peut citer l'économètre tchèque Jan KMENTA (1928-2016) dans son livre : ***Elements of Econometrics*** (2nd edition), McMillan, 1971 : page 431 :

> *Multicollinearity is a feature of the sample, not of the population.*
>
> *Therefore, we do not « test for multicollinearity » but can, if we wish, measure its degree in any particular sample.*

II.3. Test de significativité sous l'hypothèse de normalité

a) Le test t de significativité d'un paramètre

Si les erreurs $\boldsymbol{\varepsilon}$ sont normalement distribuées, on a :

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \mathcal{N}_K(\boldsymbol{\beta}\,,\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Donc pour un paramètre particulier : $\widehat{\beta_p}|\boldsymbol{X} \sim \mathcal{N}(\beta_p\,,\sigma^2 S_{pp})$ pour tout $p = 1,2,\ldots K$, où S_{pp} est le $p^{ième}$ élément de la diagonale principale de $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

On aura alors un ratio : $$z_p = \frac{\widehat{\beta_p} - \beta_p}{\sqrt{\sigma^2 S_{pp}}} \sim \mathcal{N}(0\,,1)$$

Mais comme dans la régression simple, on ne connaît pas σ^2, il faut estimer cette variance par : $\widehat{\sigma^2} = \boldsymbol{e}'\boldsymbol{e}/(N-K) = SCR/(N-K)$, que l'on substitue dans le ratio précédent pour obtenir le ratio t_p :

$$t_p = \frac{\widehat{\beta_p} - \beta_p}{\sqrt{\widehat{\sigma^2} S_{pp}}}$$

Le dénominateur est l'écart-type du $p^{ième}$ paramètre estimé par MCO :

$$\sqrt{\widehat{\sigma^2} S_{pp}} = \sqrt{\widehat{\sigma^2}[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{pp}} = \sqrt{\widehat{V}(\widehat{\beta_p}|\boldsymbol{X})} = s_{\widehat{\beta_p}}$$

Le ratio t_p devient alors $t_p = \frac{\widehat{\beta_p}-\beta_p}{s_{\widehat{\beta_p}}}$ qui suit alors une loi t de Student à $N-K$ degrés de liberté.

$$t_p = \frac{\widehat{\beta_p} - \beta_p}{s_{\widehat{\beta_p}}} \sim t(N-K)$$

Voir démonstration ci-après.

DÉMONSTRATION :

Soit le ratio t_p : $t_p = \frac{\widehat{\beta_p} - \beta_p}{s_{\widehat{\beta_p}}} = \frac{\widehat{\beta_p} - \beta_p}{\sqrt{\widehat{\sigma^2} S_{pp}}} = \frac{\widehat{\beta_p} - \beta_p}{\sqrt{\sigma^2 S_{pp}}} \times \frac{\sqrt{\sigma^2}}{\sqrt{\widehat{\sigma^2}}} = \frac{(\widehat{\beta_p} - \beta_p)/\sqrt{\sigma^2 S_{pp}}}{\sqrt{\widehat{\sigma^2}/\sigma^2}}$

Sous l'hypothèse de normalité des erreurs, le numérateur suit une loi normale standard *:* $z_p = (\widehat{\beta_p} - \beta_p)/\sqrt{\sigma^2 S_{pp}} \sim N(0\,,1)$.

De même, on a démontré ci-dessus que : $(N-K)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(N-K)$.

En conséquence au dénominateur (à l'intérieur de la racine carrée) :

$$\frac{\widehat{\sigma^2}}{\sigma^2} \sim \frac{\chi^2(N-K)}{(N-K)}$$

De plus on a montré ci-dessus que le numérateur, dépendant de $\widehat{\beta_p}$ était indépendant du dénominateur (fonction de $\widehat{\sigma^2}$).

DEFINITION : *Loi t de Student*

Le ratio d'une variable aléatoire normale standard sur la racine carrée d'une variable aléatoire du Khi-deux divisée par ses degrés de liberté est une **variable *t* de Student**, pour autant qu'elles soient **indépendantes l'une de l'autre**.

$$t(R) = \frac{N(0\,,1)}{\sqrt{\chi^2(R)/R}}$$

Donc le ratio : $t_p = \frac{z_p}{\sqrt{\widehat{\sigma^2}/\sigma^2}} = \frac{\sim \mathcal{N}(0\,,1)}{\sim \sqrt{\frac{\chi^2(N-K)}{(N-K)}}} \sim t(N-K)$

suit une loi t de Student à $N-K$ degrés de liberté : $t_p = \frac{\widehat{\beta_p}-\beta_p}{s_{\widehat{\beta_p}}} \sim t(N-K)$.

CQFD

Test d'une hypothèse simple : la significativité (statistique) d'un paramètre

On veut tester une hypothèse : la $p^{ième}$ variable explicative n'a pas d'effet sur la variable dépendante y.

Cela revient à vérifier si son vrai paramètre est nul : $\beta_p = 0$.

En revanche si cette la $p^{ième}$ variable explicative a un effet sur la variable dépendante, son vrai paramètre est différent de zéro : $\beta_p \neq 0$.

L'hypothèse principale qu'on veut tester est appelée l'**hypothèse nulle** (c'est celle qui nous intéresse) ! Elle est notée H_0 :

$$H_0 \,:\, \beta_p = 0$$

Elle est toujours sous une forme d'égalité !

Remarquez qu'on ne teste que l'effet d'une seule variable explicative, *toutes autres choses égales par ailleurs* !

Cette hypothèse nulle est testée contre une **hypothèse alternative**, notée H_1 ou H_A.

Cette hypothèse alternative peut être **unilatérale** si elle est sous forme d'une inégalité (deux possibilités) :

$$H_1 \,:\, \beta_p > 0 \qquad (\text{ ou bien } \quad H_1 \,:\, \beta_p < 0 \,)$$

Cette hypothèse alternative peut être **bilatérale** si elle rassemble les deux cas précédents :

$$H_1 \,:\, \beta_p > 0 \quad \text{et} \quad H_1 \,:\, \beta_p < 0$$

qui peut se réécrire sous la forme :

$$H_1 \,:\, \beta_p \neq 0$$

2ème étape : On définit alors une **statistique de test** qu'on peut calculer après l'estimation du modèle.

Ici on aura si $\beta_p = 0$: $$t_{\beta_p=0} = \frac{\widehat{\beta_p} - \beta_p}{\sqrt{\widehat{\sigma^2} S_{pp}}} = \frac{\widehat{\beta_p} - \beta_p}{s_{\widehat{\beta_p}}} = \frac{\widehat{\beta_p}}{s_{\widehat{\beta_p}}}$$

dont on connaît la **distribution sous l'hypothèse nulle** (si cette dernière est vraie):

$$\text{sous } H_0\,, \qquad t_{\beta_p=0} = \frac{\widehat{\beta_p}}{s_{\widehat{\beta_p}}} \sim t(N-K)$$

Si l'hypothèse nulle est vraie, la statistique de test $t_{\beta_p=0}$ est distribuée selon une loi t de Student avec $N - K$ degrés de liberté.

Si l'hypothèse nulle est fausse, la statistique de test $t_{\beta_p=0}$ suit une autre distribution, qui dépend de l'alternative choisie…

Un avantage de ce test t est que sa distribution ne dépend pas des variables explicatives $\boldsymbol{X}$, mais seulement du paramètre estimé et de son écart-type !

3ème étape : On choisit alors un **niveau de test** de α % définit comme :

$$\Pr(\text{rejeter } H_0 \mid H_0 \text{ est vraie}) = \alpha$$

C'est **l'erreur de 1ère espèce** (ou le risque de 1ère espèce).

Généralement, on la contrôle en prenant une faible probabilité d'occurrence.
Par exemple un niveau de test $\alpha = 5\ \%$ …
Mais on peut aussi choisir 10 %, ou 1 % ou 1/1000ème ou encore moins selon la décision à prendre…

Alternativement **l'erreur de 2ème espèce** sera :

$$\Pr(\text{accepter } H_0 \mid H_0 \text{ est fausse}) = \beta$$

Symétriquement, on appellera : $1 - \beta$ la **puissance du test** :

$$\Pr(\text{rejeter } H_0 \mid H_0 \text{ est fausse}) = 1 - \beta$$

Il est évident qu'on désire la puissance la plus élevée possible !
Ou un test qui a une puissance la plus élevée pour un niveau de test donné…

En résumé, on peut établir le tableau suivant :

	DECISION	
	H_0 acceptée	**H_0 rejetée**
Si H_0 est vraie	**Décision correcte** Probabilité = $1-\alpha$ % *(niveau de confiance)*	**Erreur de 1ère espèce** Probabilité = α % *(niveau du test)*
Si H_0 est fausse	**Erreur de 2ème espèce** Probabilité = β %	**Décision correcte** Probabilité = $1-\beta$ % *(puissance du test)*

On cherche à avoir la puissance maximale $1-\beta$ % pour un niveau de test α % donné, mais en général c'est contradictoire avec la recherche d'un faible niveau de test …

Voir un livre de Statistique sur la Théorie des Tests …
Par exemple : Saporta [2006], Wonnacott-Wonnacott [1995], Anderson-Sweeney-Williams [2009] ou Newbold-Carlson-Thorne [2010].

4ème étape : on finalement établit une **règle de décision** pour un niveau de test donné de α % .

En général, elle prend la forme de :

$$\begin{cases} \text{Si } |Statistique\ de\ test| < seuil\ critique & \Rightarrow \quad Accepter\ H_0 \\ \text{Si } |Statistique\ de\ test| \geq seuil\ critique & \Rightarrow \quad Rejeter\ \ H_0 \end{cases}$$

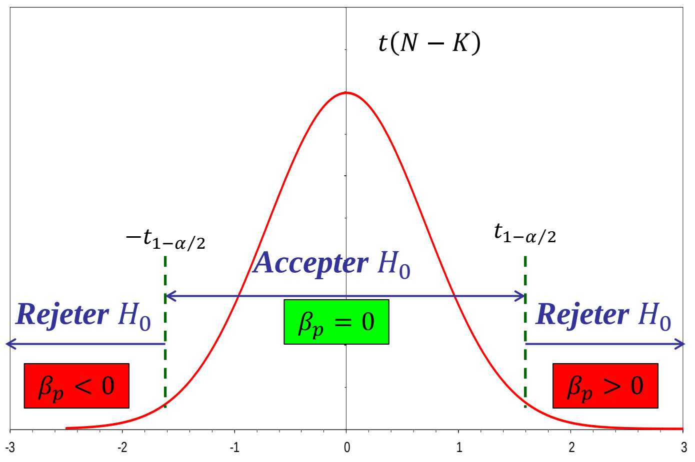
avec un ***seuil critique*** qui est le quantile de la loi de la statistique de test sous l'hypothèse nulle :

- $1-\alpha$ % pour un test unilatéral avec $H_1 :\ \beta_p > 0$ **ou** $H_1 :\ \beta_p < 0$
- $1-\alpha/2$ % pour un test bilatéral avec $H_1 :\ \beta_p \neq 0$

Règle de décision : dans le cas du test t de significativité : $t_{\beta_p=0} = \widehat{\beta_p}/s_{\widehat{\beta_p}}$

si $\left|t_{\beta_p=0}\right| < t_{1-\alpha/2}(N-K)$ $\longrightarrow$ On **accepte** H_0 : $\beta_p = 0$

si $\left|t_{\beta_p=0}\right| \geq t_{1-\alpha/2}(N-K)$ $\longrightarrow$ On **rejette** H_0 : $\beta_p = 0$



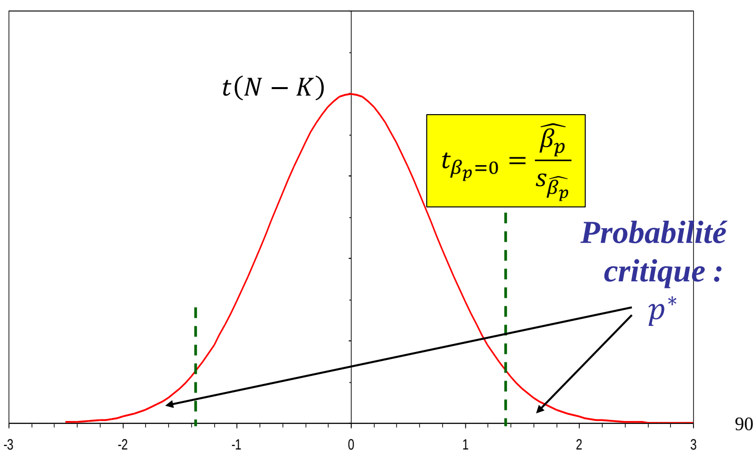
89

Alternativement, on peut utiliser la **probabilité critique** (**p-value**) :

$$\Pr(H_0 \text{ vraie}) = 2 \times \Pr\left(t(N-K) \geq \left|t_{\beta_p=0}\right|\right) = p^* = \text{probabilité critique}$$

test bilatéral

$si\ \ p^* > \alpha$ $\longrightarrow$ On **accepte** H_0 : $\beta_p = 0$



90

d) *Intervalle de confiance pour un paramètre*

Si les erreurs ε sont normalement distribuées, on sait que le ratio :

$$t_p = \frac{\widehat{\beta_p} - \beta_p}{s_{\widehat{\beta_p}}} \sim t(N-K)$$

suit une loi t de Student à $N-K$ degrés de liberté.

La théorie des probabilités nous donne par définition des quantiles d'une distribution : $\Pr\left[-t_{1-\alpha/2} \leq t(N-K) \leq t_{1-\alpha/2}\right] = 1-\alpha$

(pour la clarté, on n'indique plus, les degrés de liberté de la distribution de Student)

On peut construire un intervalle de confiance avec une probabilité $1-\alpha$:

$$\Pr\left[-t_{1-\alpha/2} \leq \frac{\widehat{\beta_p} - \beta_p}{s_{\widehat{\beta_p}}} \leq t_{1-\alpha/2}\right] = 1-\alpha$$

avec le quantile de la loi t de Student à $N-K$ degrés de liberté : $t_{1-\alpha/2}$, ou encore :

$$\Pr\left[\widehat{\beta_p} - \left(t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right) \leq \beta_p \leq \widehat{\beta_p} + \left(t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right)\right] = 1-\alpha$$

Cela signifie que le vrai paramètre β_p (inconnu) se trouve dans l'intervalle de confiance :

$$\left[\widehat{\beta_p} - \left(t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right) ; \widehat{\beta_p} + \left(t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right)\right]$$

avec une probabilité $1-\alpha$ %.

En d'autres termes : $\Pr\left\{\beta_p \in \left[\widehat{\beta_p} - \left(t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right) ; \widehat{\beta_p} + \left(t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right)\right]\right\} = 1-\alpha$

Remarquez que :

- le vrai paramètre appartient à un intervalle centré en $\widehat{\beta_p}$.
- l'intervalle est symétrique autour de $\widehat{\beta_p}$.
- l'étendue de l'intervalle dépend du produit : $t_{1-\alpha/2} \times s_{\widehat{\beta_p}}$
- l'intervalle est d'autant plus grand que l'écart-type estimé $s_{\widehat{\beta_p}}$ du paramètre est important.
- l'intervalle est d'autant plus grand que le quantile $t_{1-\alpha/2}$ est élevé, c'est-à-dire que le nombre de degré de liberté $N-K$ est faible ou que le niveau choisi $0 < \alpha < 1$ est faible.

DÉMONSTRATION :

Construction d'un **intervalle de confiance** à $1-\alpha$ % pour le vrai paramètre β_p .

$$\Pr\left[-t_{1-\alpha/2} \leq \frac{\widehat{\beta_p}-\beta_p}{s_{\widehat{\beta_p}}} \leq t_{1-\alpha/2}\right] = 1-\alpha$$

On multiplie partout par l'écart-type (positif) $s_{\widehat{\beta_p}}$:

$$\Pr\left[-t_{1-\alpha/2} \times s_{\widehat{\beta_p}} \leq \widehat{\beta_p}-\beta_p \leq t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right] = 1-\alpha$$

On soustrait partout le paramètre estimé $\widehat{\beta_p}$:

$$\Pr\left[-\widehat{\beta_p}-t_{1-\alpha/2} \times s_{\widehat{\beta_p}} \leq -\beta_p \leq -\widehat{\beta_p}+t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right] = 1-\alpha$$

On multiplie partout par -1, en inversant les inégalités :

$$\Pr\left[\widehat{\beta_p}+t_{1-\alpha/2} \times s_{\widehat{\beta_p}} \geq \beta_p \geq \widehat{\beta_p}-t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right] = 1-\alpha$$

En réarrangeant les termes, on obtient l'intervalle de confiance :

$$\Pr\left[\widehat{\beta_p}-t_{1-\alpha/2} \times s_{\widehat{\beta_p}} \leq \beta_p \leq \widehat{\beta_p}+t_{1-\alpha/2} \times s_{\widehat{\beta_p}}\right] = 1-\alpha$$ **CQFD**

e) *L'analyse de la variance et le test F de significativité conjointe*

On a vu précédemment (*Section I.5.a*) que la somme des carrés totaux (*SCT*) de la variable dépendante y pouvait se décomposer dans :

- la somme des carrés expliqués par la régression (*SCE*)
- la somme des carrés des résidus (*SCR*)

$$SCT = SCE + SCR \quad \text{avec :} \begin{cases} SCT = \sum_{i=1}^{N}(y_i-\bar{y})^2 \\ SCE = \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2 \\ SCR = \sum_{i=1}^{N} {e_i}^2 \end{cases}$$

On peut construire un tableau **d'analyse de la variance** (ANOVA) pour décomposer la variance de y .

Ensuite on va construire un test F de l'analyse de la variance pour les facteurs contenus dans les variables explicatives de la régression…

Source de Variation	Somme des Carrés	Degrés de Liberté	Carrés moyens
Régression	$SCE = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$	$K - 1$	$CME = \dfrac{SCE}{K-1}$
Résiduelle	$SCR = \sum_{i=1}^{N} {e_i}^2$	$N - K$	$CMR = \dfrac{SCR}{N-K} = \widehat{\sigma^2}$
Totale	$SCT = \sum_{i=1}^{N}(y_i - \bar{y})^2$	$N - 1$	$CMT = \dfrac{SCT}{N-1} = V(y)$

On construit alors un test F de l'analyse de la variance.

Ce test F donne la **significativité conjointe de tous les paramètres de pente** (ceux des variables explicatives utilisées dans la régression).

C'est un **test global de la significativité** de toutes les variables explicatives dans la régression.

Hypothèse nulle : $H_0 :\ \beta_2 = \beta_3 = \cdots = \beta_K = 0$

ou de manière équivalente : $H_0 :\ \beta_2 = 0$ et $\beta_3 = 0$ et ... et $\beta_K = 0$

Hypothèse alternative : H_1 : au moins un $\beta_p \neq 0$.

Statistique du Test *F* :

$$F = \frac{CME}{CMR} = \frac{SCE/(K-1)}{SCR/(N-K)}$$

Remarquez que le dénominateur est l'estimateur de la variance : $\widehat{\sigma^2} = SCR/(N-K)$

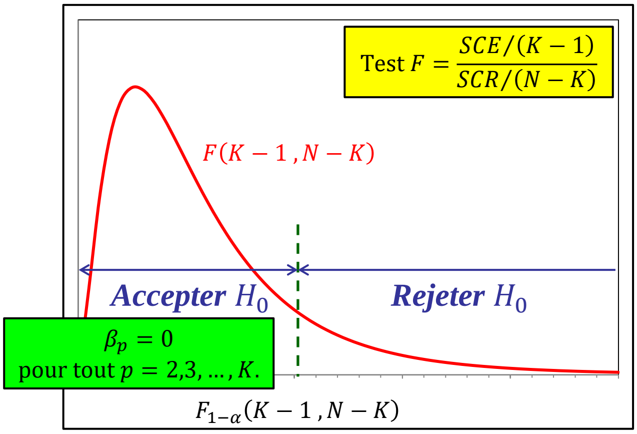
Cette statistique est distribuée sous l'hypothèse nulle selon une **loi *F* de Fisher**
avec $K-1$ degrés de liberté au numérateur
avec $\mathrm{N} - K$ degrés de liberté au dénominateur

sous H_0 : $F \sim F(K-1, N-K)$

La **règle de décision**, avec un niveau de test de α % :

si $F < F_{1-\alpha}(K-1, N-K) \longrightarrow$ **Accepter** $H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$
Tous les paramètres sont nuls !

si $F \geq F_{1-\alpha}(K-1, N-K) \longrightarrow$ **Rejeter** H_0
Au moins un des paramètres est différent de zéro !



Le test *F* peut se réécrire en terme de R^2 : $R^2 = 1 - \dfrac{SCR}{SCT} = \dfrac{SCE}{SCT}$

$$\rightarrow \quad SCE = R^2 \times SCT \quad \text{et} \quad SCR = (1-R^2) \times SCT$$

$$F = \frac{SCE/(K-1)}{SCR/(N-K)} \quad \Leftrightarrow \quad F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)} = \frac{R^2}{1-R^2} \times \frac{N-K}{K-1}$$

Ce test F de **significativité conjointe** est calculé automatiquement par de très nombreux logiciels d'économétrie.

La règle de décision du test F de **significativité conjointe** indique que si la statistique de test est inférieure au seuil critique, on accepte l'hypothèse nulle : tous les paramètres de pente sont nuls et aucune des variables explicatives joue dans la régression.

$$\text{si} \quad F < F_{1-\alpha}(K-1, N-K) \quad \rightarrow \quad \text{Accepter } H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

Réécrivons cette règle de décision d'acceptation de l'hypothèse nulle avec $F_{1-\alpha}$ le seuil critique :

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)} < F_{1-\alpha}$$

On peut chercher la valeur maximale du R^2 pour accepter cette hypothèse nulle :

$$\text{si } R^2 < \frac{F_{1-\alpha}}{F_{1-\alpha} + \frac{N-K}{K-1}} \quad \rightarrow \quad \text{Accepter } H_0$$

Dans notre exemple, pour un niveau de test de 5% : $F_{0.95}(2\,,17) = 3.592$. Donc :

$$\text{si } R^2 < \frac{3.592}{3.592 + \frac{17}{2}} = 0.297 \quad \rightarrow \quad \text{Accepter } H_0$$

Si $R^2 > 29.7\%$, la condition n'est pas vérifiée et on rejette l'hypothèse nulle.

II.4. Les moindres carrés contraints

a) Les contraintes linéaires

On veut estimer un modèle de régression linéaire classique : $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

avec J contraintes (ou combinaisons) linéaires entre les K paramètres, définies par :

$$\boldsymbol{R\beta} = \boldsymbol{r} \quad \text{ou} \quad \boldsymbol{R\beta} - \boldsymbol{r} = \boldsymbol{0}_J$$

où $\boldsymbol{R}$ est une matrice $(J \times K)$ de coefficients fixes connus et $\boldsymbol{r}$ un vecteur $(J \times 1)$ de valeurs données.

Il est évident qu'on ne peut pas avoir autant, voire plus, de contraintes que de paramètres à estimer : $J < K$.

En conséquence, la matrice $\boldsymbol{R}$ est de rang plein ligne : $rang(\boldsymbol{R}) = J < K$.

Les contraintes ne doivent pas être redondantes !

Exemples

1. La somme de deux paramètres est égale à 1 : $\beta_2 + \beta_3 = 1$
   On a une seule contrainte : $J = 1$

$$\boldsymbol{R\beta} = \boldsymbol{r} \quad \rightarrow \quad \boldsymbol{R} = \begin{pmatrix} 0 & 1 & 1 & 0 & \cdots & 0 \end{pmatrix} \quad \text{et} \quad \boldsymbol{r} = 1$$

2. Deux paramètres sont égaux : $\beta_2 = \beta_3$ ou $\beta_2 - \beta_3 = 0$
   On a une seule contrainte : $J = 1$

$$\boldsymbol{R\beta} = \boldsymbol{r} \quad \rightarrow \quad \boldsymbol{R} = \begin{pmatrix} 0 & 1 & -1 & 0 & \cdots & 0 \end{pmatrix} \quad \text{et} \quad \boldsymbol{r} = 0$$

3. Tous les paramètres de pentes sont égaux : $\beta_2 = \beta_3 = \cdots = \beta_K$
   On a $J = K - 2$ contraintes :

$$\boldsymbol{R\beta} = \boldsymbol{r} \quad \rightarrow \quad \boldsymbol{R} = \begin{bmatrix} 0 & 1 & -1 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & -1 & & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & & 1 & -1 \end{bmatrix} \quad \text{et} \quad \boldsymbol{r} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

b) Les Moindres Carrés Contraints

On va utiliser **l'estimateur des moindres carrés** pour estimer ce modèle contraint.

Maintenant, la minimisation du critère des moindres carrés se ferra sous les J contraintes linéaires entre les paramètres, en écrivant le Lagrangien :

$$\min_{\boldsymbol{\beta}} \Lambda = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) - 2\boldsymbol{\lambda}'(\boldsymbol{R\beta} - \boldsymbol{r})$$

avec un vecteur – colonne $\boldsymbol{\lambda}$ $(J \times 1)$ des J **multiplicateurs de Lagrange** associés aux J contraintes : $\boldsymbol{\lambda}' = (\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_J)$

Les conditions du premier ordre nécessaires pour la maximisation :

$$\begin{cases} \dfrac{\partial \Lambda}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta}^*) - 2\boldsymbol{R}'\boldsymbol{\lambda}^* = \boldsymbol{0}_K \\ \dfrac{\partial \Lambda}{\partial \boldsymbol{\lambda}} = -2(\boldsymbol{R\beta}^* - \boldsymbol{r}) = \boldsymbol{0}_J \end{cases}$$

Une étoile * indique la solution de ce problème :

→ **les estimateurs des Moindres Carrés Contraints** (**MCC**) : $\boldsymbol{\beta}^*$ et $\boldsymbol{\lambda}^*$.

Le système d'**équations normales** modifiées se réécrit comme :

$$\begin{cases} X'X\beta^* + R'\lambda^* = X'y \\ R\beta^* = r \end{cases} \Rightarrow \begin{pmatrix} X'X & R' \\ R & 0_J \end{pmatrix}\begin{pmatrix} \beta^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} X'y \\ r \end{pmatrix}$$

La solution de ces équations normales est obtenue en utilisant (par exemple) les résultats de calcul matriciel sur les matrices partitionnées :

$$\begin{pmatrix} \beta^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} X'X & R' \\ R & 0_J \end{pmatrix}^{-1}\begin{pmatrix} X'y \\ r \end{pmatrix}$$

$$= \begin{pmatrix} (X'X)^{-1} - (X'X)^{-1}R'\Gamma^{-1}R(X'X)^{-1} & (X'X)^{-1}R'\Gamma^{-1} \\ \Gamma^{-1}R(X'X)^{-1} & -\Gamma^{-1} \end{pmatrix}\begin{pmatrix} X'y \\ r \end{pmatrix}$$

avec $\Gamma = R(X'X)^{-1}R'$

$$\begin{pmatrix} \beta^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} (X'X)^{-1}X'y - (X'X)^{-1}R'\Gamma^{-1}R(X'X)^{-1}X'y + (X'X)^{-1}R'\Gamma^{-1}r \\ \Gamma^{-1}R(X'X)^{-1}X'y - \Gamma^{-1}r \end{pmatrix}$$

Ce qui donne l'estimateur des **Moindres Carrés Contraints** (***MCC***) des paramètres du modèle :

$$\Rightarrow \quad \beta^* = (X'X)^{-1}X'y - (X'X)^{-1}R'\Gamma^{-1}R(X'X)^{-1}X'y + (X'X)^{-1}R'\Gamma^{-1}r$$

On peut réécrire cet estimateur des ***MCC*** comme :

$$\beta^* = \underbrace{(X'X)^{-1}X'y}_{=\hat{\beta}} - (X'X)^{-1}R'\Gamma^{-1}R\underbrace{(X'X)^{-1}X'y}_{=\hat{\beta}} + (X'X)^{-1}R'\Gamma^{-1}r$$

$$= \hat{\beta} - (X'X)^{-1}R'\Gamma^{-1}R\hat{\beta} + (X'X)^{-1}R'\Gamma^{-1}r$$

$$= \hat{\beta} - (X'X)^{-1}R'\Gamma^{-1}\left(R\hat{\beta} - r\right)$$

ou en substituant : $\Gamma = R(X'X)^{-1}R'$

$$\beta^* = \hat{\beta} - (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right)$$

Le second groupe des équations normales modifiées donne l'estimateur des multiplicateurs de Lagrange :

$$\lambda^* = \Gamma^{-1}R\underbrace{(X'X)^{-1}X'y}_{=\hat{\beta}} - \Gamma^{-1}r = \Gamma^{-1}\left(R\hat{\beta} - r\right)$$

Ce qui donne : $\lambda^* = \left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right)$

Remarquez que si l'estimateur *MCO* satisfait la contrainte $(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r} = \boldsymbol{0})$, l'estimateur *MCC* du multiplicateur de Lagrange est nul…

On peut aussi réécrire l'estimateur des *MCC* de $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\lambda}^*$$

DÉMONSTRATION ALTERNATIVE PAR SUBSTITUTION

On part du système des équations normales modifiées : $\begin{cases} \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{R}'\boldsymbol{\lambda}^* = \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r} \end{cases}$

La solution pour la première partie de ce système (K équations) s'écrit :

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{R}'\boldsymbol{\lambda}^* = \boldsymbol{X}'\boldsymbol{y} \quad \Rightarrow \quad \boldsymbol{\beta}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\lambda}^*$$

On substitue cette expression dans la solution pour la deuxième partie de ce système (J équations) :

$$\boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r} \quad \Rightarrow \quad \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\lambda}^* = \boldsymbol{r}$$

Ce qui donne l'estimateur des multiplicateurs de Lagrange :

$$\boldsymbol{\lambda}^* = (\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - (\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\boldsymbol{r}$$

ou encore en remplaçant l'estimateur des MCO : $\boldsymbol{\lambda}^* = (\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$

On substitue alors cet estimateur de $\boldsymbol{\lambda}^*$ dans l'expression de l'estimateur de $\boldsymbol{\beta}^*$:

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

Les estimateurs de $\boldsymbol{\beta}^*$ et de $\boldsymbol{\lambda}^*$ sont identiques à ceux obtenus précédemment.

c) Propriétés des MCC.

L'estimateur des Moindres Carrés Contraints (MCC) :

$$\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right) = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\lambda}^*$$

est **sans biais** *si les contraintes sont correctes* : $E(\boldsymbol{\beta}^*|\boldsymbol{X}) = \boldsymbol{\beta}$

Ici on conserve encore les hypothèses **H2** (exogénéité forte) et **H3** (condition de rang).

Remarquez que l'estimateur MCO est aussi **sans biais** dans ce modèle :

$E\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) = \boldsymbol{\beta}$ → *Pourquoi ?*

DÉMONSTRATION : $E(\boldsymbol{\beta}^*|\boldsymbol{X}) = \boldsymbol{\beta}$

L'estimateur des MCC s'écrit :

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

En remplaçant la variable dépendante $\boldsymbol{y}$ par le vrai modèle : $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

$$\boldsymbol{\beta}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{r}$$

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R\beta} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$
$$+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{r}$$

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}(\boldsymbol{R\beta} - \boldsymbol{r}) + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

Si la **contrainte est correcte** ($\boldsymbol{R\beta} = \boldsymbol{r}$), on aura en prenant l'espérance conditionnelle à $\boldsymbol{X}$:

$$E(\boldsymbol{\beta}^*|\boldsymbol{X}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}|\boldsymbol{X}) - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}|\boldsymbol{X})$$

Avec l'hypothèse *H3* : $E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}_K$, l'estimateur des *MCC* est sans biais :

$E(\boldsymbol{\beta}^*|\boldsymbol{X}) = \boldsymbol{\beta}$ **CQFD**

Si la contrainte est correcte, la **matrice de variance – covariance** de l'estimateur des MCC est alors donnée par : *(voir ci-après)*

$$V(\boldsymbol{\beta}^*|\boldsymbol{X}) = \underbrace{\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}}_{V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})} - \underbrace{\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}}_{\textit{Matrice symétrique définie-positive}}$$

La variance de l'estimateur des MCC est **plus faible** que la variance de l'estimateur des MCO (au sens matriciel) : $V(\boldsymbol{\beta}^*|\boldsymbol{X}) \leq V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})$.

- L'estimateur des MCC est plus précis que l'estimateur des MCO (*pour autant que les contraintes soient correctes !*)
- parce qu'il inclut des informations (hors échantillon) supplémentaires → les contraintes imposées aux paramètres.

DÉMONSTRATION :

En supposant que les contraintes soient correctes : $\boldsymbol{R\beta} = \boldsymbol{r}$, l'erreur d'échantillonnage de l'estimateur *MCC* est :

$$\boldsymbol{\beta}^* - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

La variance conditionnelle de l'estimateur *MCC* (en tenant compte de l'absence de biais de cet estimateur sera :

$$V(\boldsymbol{\beta}^*|\boldsymbol{X}) = \boldsymbol{E}[(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})'|X]$$

$$= E\{\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right)$$
$$\times \left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right)'|\boldsymbol{X}\}$$

$$= E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}\right)$$
$$- E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X})$$
$$- E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X})$$
$$+ E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X})$$

Comme on prend l'espérance conditionnelle par rapport aux variables explicatives, on aura :

$$\begin{aligned} V(\boldsymbol{\beta}^*|\boldsymbol{X}) = & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & -(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & -(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & +(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

Avec l'hypothèse **H4** (homoscédasticité) et **H5** (absence d'autocorrélation) la matrice de variance – covariance des erreurs est sphérique : $E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I}_N$, on obtient :

$$\begin{aligned} V(\boldsymbol{\beta}^*|\boldsymbol{X}) = & \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & -\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & -\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & +\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

$$\begin{aligned} V(\boldsymbol{\beta}^*|\boldsymbol{X}) = & \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & -\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ & +\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\underbrace{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}_{=\boldsymbol{\Gamma}}\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

Mais : $\boldsymbol{\Gamma} = \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$

Cela permet de simplifier le 3ème avec le 4ème terme :

$$V(\boldsymbol{\beta}^*|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$V(\boldsymbol{\beta}^*|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Le premier terme du membre de droite est la variance conditionnelle de l'estimateur des *MCO* : $V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

On obtient alors la différence entre les variances des 2 estimateurs :

$$V(\boldsymbol{\beta}^*|\boldsymbol{X}) - V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = -\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{\Gamma}^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

qui est une matrice symétrique définie – négative (à cause du signe négatif).

En conséquence :

$$V(\boldsymbol{\beta}^*|\boldsymbol{X}) - V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) \leq \boldsymbol{0}_{K\times K}$$

ou :
$$V(\boldsymbol{\beta}^*|\boldsymbol{X}) \leq V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})$$

CQFD

De la même manière, la matrice de variance – covariance du vecteur des multiplicateurs de Lagrange s'obtient alors, en notant que :

$$\boldsymbol{\lambda}^* = (\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})$$

En utilisant les règles de calcul de la variance d'une combinaison linéaire de variables aléatoires, ici l'estimateur MCO $\widehat{\boldsymbol{\beta}}$:

$$V(\boldsymbol{\lambda}^*|\boldsymbol{X}) = (\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}$$

Comme $V(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, on aura : $V(\boldsymbol{\lambda}^*|\boldsymbol{X}) = \sigma^2(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}$

Si les erreurs sont normalement distribuées, l'estimateur des moindres carrés contraints $\boldsymbol{\beta}^*$ et $\boldsymbol{\lambda}^*$ est normalement distribué.

On peut alors faire les tests usuels de significativité…

On peut alors tester la significativité conjointe de $\boldsymbol{\lambda}^*$ pour savoir si l'estimateur *MCO* satisfait *de facto* la contrainte.

113

II.5. Le test *F* de contraintes linéaires

- Test de significativité sur un seul paramètre
- Maintenant : **Test conjoint sur plusieurs paramètres**
- En fait on veut tester plusieurs combinaisons linéaires des paramètres **simultanément**

114

a) *Test conjoint de plusieurs combinaisons linéaires*

On veut tester simultanément plusieurs hypothèses simples ou plusieurs combinaisons linéaires des paramètres :

$$\begin{cases} H_0 : \boldsymbol{R\beta} = \boldsymbol{r} \\ H_1 : \boldsymbol{R\beta} \neq \boldsymbol{r} \end{cases} \rightarrow \begin{cases} H_0 : \boldsymbol{R\beta} - \boldsymbol{r} = \boldsymbol{0} \\ H_1 : \boldsymbol{R\beta} - \boldsymbol{r} \neq \boldsymbol{0} \end{cases}$$

avec une matrice $\boldsymbol{R}$ $(J \times K)$ de coefficients fixes connus et un vecteur $\boldsymbol{r}$ $(J \times 1)$ de valeurs données.

La matrice $\boldsymbol{R}$ est de rang plein ligne : $rang(\boldsymbol{R}) = J < K$.

Donc on doit tester J combinaisons linéaires **mutuellement exclusives**.

Par exemple : si on veut tester l'égalité des 3 paramètres d'un modèle :

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

on aura que **2 contraintes mutuellement exclusives** $(J = 2)$: $\beta_1 = \beta_2$ et $\beta_2 = \beta_3$ parce que la troisième égalité $\beta_1 = \beta_3$ est implicite dans les deux premières.

Si on écrivait la matrice $\boldsymbol{R}$ pour les 3 contraintes :

$$\begin{cases} \beta_1 = \beta_2 \\ \beta_2 = \beta_3 \\ \beta_1 = \beta_3 \end{cases} \Rightarrow \begin{cases} \beta_1 - \beta_2 = 0 \\ \beta_2 - \beta_3 = 0 \\ \beta_1 - \beta_3 = 0 \end{cases} \Rightarrow \boldsymbol{R} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix} \quad \text{et} \quad \boldsymbol{r} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

On voit immédiatement que la 3$^{\text{ème}}$ ligne (la 3$^{\text{ème}}$ contrainte) est la somme des 2 premières lignes. Donc la matrice $\boldsymbol{R}$ est de rang 2 (et non pas 3).

La troisième contrainte n'est pas mutuellement exclusive, elle est redondante et on doit la supprimer :

$$\Rightarrow \quad \boldsymbol{R} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \quad \text{et} \quad \boldsymbol{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Rightarrow \quad J = 2.$$

A titre d'exercice : si on veut tester l'égalité à zéro des 3 paramètres d'un modèle :

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

- Combien a-t-on de **contraintes mutuellement exclusives** ?
- Donnez la forme de la matrice $\boldsymbol{R}$ et du vecteur $\boldsymbol{r}$?

On suppose que les erreurs sont normalement distribuées
→ résultats en échantillon finis.

La distribution de l'estimateur *MCO* est une loi normale multivariée :

$$\widehat{\boldsymbol{\beta}}\big|\boldsymbol{X} \sim \mathcal{N}_K(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Les J combinaisons linéaires des paramètres estimés seront donc normalement distribuées :

$$\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\big|\boldsymbol{X} \sim \mathcal{N}_J(\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r}, \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')$$

➢ 1er Test : Test de Wald

On va noter : $\boldsymbol{\Omega} = \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$

A un facteur de proportionnalité près (la variance de l'erreur σ^2), c'est la matrice de variance – covariance des J combinaisons linéaires :

$$V\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\middle|\boldsymbol{X}\right) = \boldsymbol{R}V\left(\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)\boldsymbol{R}' = \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = \sigma^2\boldsymbol{\Omega}$$

On peut calculer la racine carrée (matricielle) de cette matrice $\boldsymbol{\Omega}$ (c'est une matrice symétrique et définie-positive), notée $\boldsymbol{\Omega}^{1/2}$:

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}'^{1/2} \quad \text{et} \quad \boldsymbol{\Omega}^{-1} = \boldsymbol{\Omega}'^{-1/2}\boldsymbol{\Omega}^{-1/2}$$

Cette racine carrée d'une matrice symétrique définie positive est la décomposition de Cholesky en triangle inférieur. (*Attention à l'ordre de la transposition*)

Ce qui donne :

$$\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\big|\boldsymbol{X} \sim N_J(\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r}, \sigma^2\boldsymbol{\Omega})$$

$$\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\big|\boldsymbol{X} \sim N_J(\boldsymbol{0}, \sigma^2\boldsymbol{\Omega})$$

$$\frac{1}{\sigma}\boldsymbol{\Omega}^{-1/2}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right)\Big|\boldsymbol{X} \sim N_J\left(\boldsymbol{0}, \boldsymbol{I}_J\right)$$

parce que $\frac{1}{\sigma}\boldsymbol{\Omega}^{-1/2}\sigma^2\boldsymbol{\Omega}\frac{1}{\sigma}\boldsymbol{\Omega}'^{-1/2} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}'^{1/2}\boldsymbol{\Omega}'^{-1/2} = \boldsymbol{I}_J$

La somme des carrés de ces J variables normales indépendantes, centrées et réduites est, par définition, une variable aléatoires du Khi-deux à J degrés de liberté.

La somme des carrés de ces variables est le produit scalaire :

$$W = \left(\frac{1}{\sigma}\boldsymbol{\Omega}^{-1/2}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right)\right)'\left(\frac{1}{\sigma}\boldsymbol{\Omega}^{-1/2}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right)\right) \sim \chi^2(J)$$

Ce qui est la **statistique de test de Wald** *(Abraham WALD, statisticien hongrois, 1902 – 1950).*

On peut la réécrire sous la forme :

$$\begin{aligned} W &= \frac{1}{\sigma^2}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right)'\boldsymbol{\Omega}'^{-1/2}\boldsymbol{\Omega}^{-1/2}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right) \\ &= \frac{1}{\sigma^2}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right) \\ &= \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right)'(\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right) \\ &= \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right)'V\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta}\right) \end{aligned}$$

Si l'hypothèse nulle est vraie → $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$

$$W = \left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)'\left(\sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \sim \chi^2(J)$$

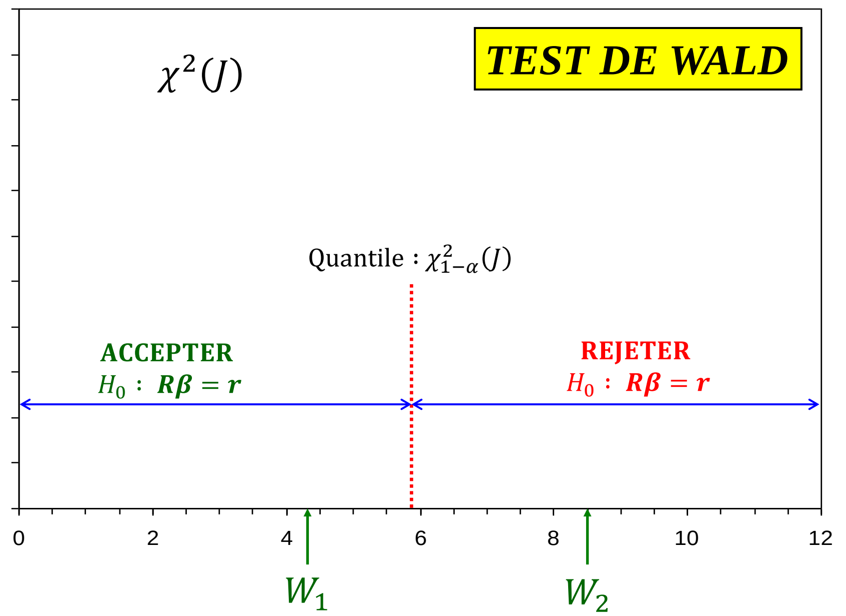
sera distribuée comme une variable Khi-deux avec J degrés de liberté.

En choisissant un niveau de test (la probabilité de rejeter H_0 si cette hypothèse est vraie) de α%, on cherche dans les tables le quantile à $1 - \alpha$% de la loi du Khi-deux à J degrés de liberté. C'est le *seuil critique*.

Si la statistique de test de Wald W est plus grande que ce seuil critique, on **rejette** l'hypothèse nulle au niveau α % :

$$\begin{array}{lcl} W \geq \chi^2_{1-\alpha}(J) & \rightarrow & \textbf{Rejeter } H_0 \\ W < \chi^2_{1-\alpha}(J) & \rightarrow & \textbf{Accepter } H_0 \end{array}$$

Mais on a supposé ici que σ^2 est connu !!!



➢ 2ème Test : Test F

Ce test F est la version **en échantillon fini** du test de Wald.

On va passer de la distribution du Khi – deux à la distribution F de Fisher, comme on est passé de la distribution normale à la distribution t dans le cas d'un test simple de significativité.

Si les erreurs sont normalement distribuées, la distribution de l'estimateur de la variance :

$$(N-K)\frac{\widehat{\sigma^2}}{\sigma^2}\bigg|\boldsymbol{X} \sim \chi^2(N-K)$$

Si on divise alors par $N-K$:

$$\frac{\widehat{\sigma^2}}{\sigma^2}\bigg|\boldsymbol{X} \sim \frac{\chi^2(N-K)}{(N-K)}$$

De même, si on divise la statistique de Wald par le nombre de combinaisons linéaires J à tester :

$$\frac{W}{J} = \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)' \frac{\left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}}{J \times \sigma^2} \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \sim \frac{\chi^2(J)}{J}$$

On divise alors W/J par le ratio précédent $\widehat{\sigma^2}/\sigma^2$:

$$\frac{W/J}{\widehat{\sigma^2}/\sigma^2} = \frac{\dfrac{\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)'\left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)}{J\cancel{\sigma^2}}}{\dfrac{\widehat{\sigma^2}}{\cancel{\sigma^2}}} \sim \frac{\dfrac{\chi^2(J)}{J}}{\dfrac{\chi^2(N-K)}{N-K}}$$

(on se rappelle que le numérateur est indépendant du dénominateur) ***Pourquoi ?***

Ce qui donne, sous l'hypothèse nulle H_0 : $\boldsymbol{R\beta} = \boldsymbol{r}$, **statistique de test F** :

$$F = \frac{1}{J}\frac{\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)'\left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)}{\widehat{\sigma^2}} \sim F(J\,,N-K)$$

qui est distribuée, sous l'hypothèse nulle, selon une loi F de Fisher à J degrés de liberté au numérateur et $N-K$ degrés de liberté au dénominateur.

La statistique F peut se réécrire en notant la variance de l'estimateur MCO :

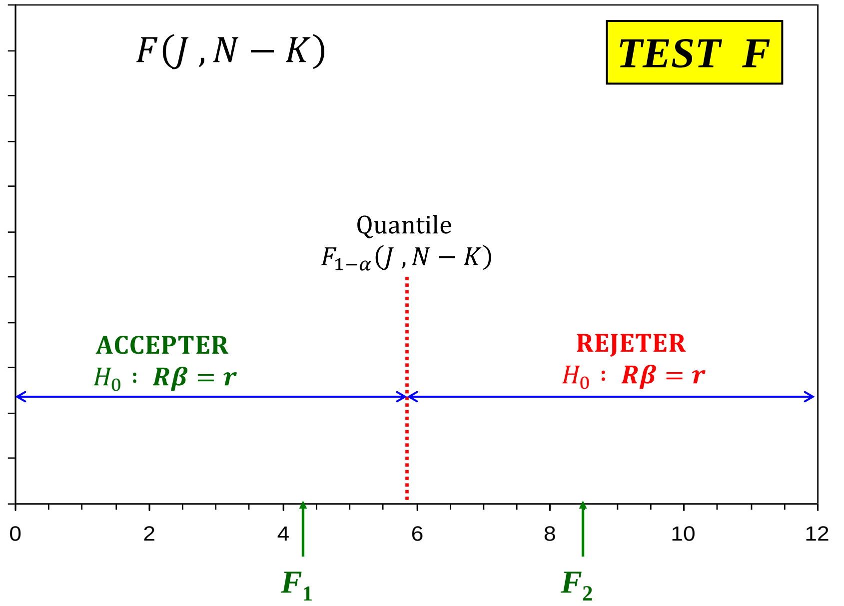
$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \hat{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \widehat{\sigma^2}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$F = \frac{1}{J}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})'\left(\widehat{\sigma^2}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}) = \frac{1}{J}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})'\left(\boldsymbol{R}\boldsymbol{\Sigma}_{\widehat{\beta}}\boldsymbol{R}'\right)^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})$$

Règle de décision

- Si la statistique F est **plus grande** que le seuil critique $F_{1-\alpha}(J\,, N-K)$, on **rejette** l'hypothèse nulle au niveau α % : $F \geq F_{1-\alpha}(J\,, N-K)$

- Si la statistique F est **plus petite** que le seuil critique $F_{1-\alpha}(J\,, N-K)$, on **accepte** l'hypothèse nulle au niveau α % : $F < F_{1-\alpha}(J\,, N-K)$



On peut relier le test F au test t de significativité vu dans le chapitre précédent :

	Une seule hypothèse ($J = 1$)	**Plusieurs hypothèses conjointes ($J > 1$)**
	Une seule combinaison linéaire $H_0 : \boldsymbol{R\beta} = r$	Plusieurs combinaisons linéaires $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$
Test asymptotique ou σ^2 connu	$z = \dfrac{\boldsymbol{R\widehat{\beta}} - r}{\sqrt{\sigma^2 \boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R'}}} \sim N(0,1)$	$W = (\boldsymbol{R\widehat{\beta}} - \boldsymbol{r})' V(\boldsymbol{R\widehat{\beta}})^{-1}(\boldsymbol{R\widehat{\beta}} - \boldsymbol{r}) \sim \chi^2(J)$
Test en petit échantillon ou σ^2 estimé	$t = \dfrac{\boldsymbol{R\widehat{\beta}} - r}{\sqrt{\widehat{\sigma^2} \boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R'}}} \sim t(N-K)$	$F = \frac{1}{J}(\boldsymbol{R\widehat{\beta}} - \boldsymbol{r})'(\boldsymbol{R\Sigma_{\widehat{\beta}}R'})^{-1}(\boldsymbol{R\widehat{\beta}} - \boldsymbol{r}) \sim F(J, N-K)$

Si $J = 1$, on aura : $W = z^2$ et $F = t^2$.

Notez en pratique : $W = J \times F$.

b) *Intervalle de confiance pour 2 paramètres : l'ellipsoïde de concentration*

Soit un modèle avec 2 paramètres : $\left.\begin{pmatrix}\widehat{\beta}_1\\ \widehat{\beta}_2\end{pmatrix}\right| \boldsymbol{X} \sim \mathcal{N}_2\left(\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}, \sigma^2(\boldsymbol{X'X})^{-1}\right)$

On veut tester l'hypothèse conjointe : $H_0 : \begin{cases}\beta_1 = b_1\\ \text{et}\\ \beta_2 = b_2\end{cases}$ contre $H_1 : \begin{cases}\beta_1 \neq b_1\\ \text{ou}\\ \beta_2 \neq b_2\end{cases}$

Dans ce cas : $\boldsymbol{R} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$ et $\boldsymbol{r} = \begin{pmatrix}b_1\\ b_2\end{pmatrix}$

$$\boldsymbol{R\beta} = \boldsymbol{r} \rightarrow \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}b_1\\ b_2\end{pmatrix} \rightarrow \begin{cases}\beta_1 = b_1\\ \beta_2 = b_2\end{cases}$$

On va noter la matrice : $\boldsymbol{\Omega} = (\boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R'})^{-1} = (\boldsymbol{X'X}) = \begin{pmatrix}\omega_{11} & \omega_{12}\\ \omega_{12} & \omega_{22}\end{pmatrix}$

avec $\omega_{kl} = \sum_{i=1}^{N} x_{k,i}x_{l,i}$

La statistique F est égale à :

$$F=\frac{\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}\right)'\left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}\right)}{J\widehat{\sigma^2}}=\frac{\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{b}\right)'(\boldsymbol{X}'\boldsymbol{X})\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{b}\right)}{2\widehat{\sigma^2}}$$

$$=\frac{1}{2\widehat{\sigma^2}}\left\{\left(\left(\widehat{\beta_1}-b_1\right)\quad\left(\widehat{\beta_2}-b_2\right)\right)\begin{pmatrix}\omega_{11} & \omega_{12}\\ \omega_{12} & \omega_{22}\end{pmatrix}\begin{pmatrix}\left(\widehat{\beta_1}-b_1\right)\\ \left(\widehat{\beta_2}-b_2\right)\end{pmatrix}\right\}$$

$$=\frac{1}{2\widehat{\sigma^2}}\left[\omega_{11}\left(\widehat{\beta_1}-b_1\right)^2+2\omega_{12}\left(\widehat{\beta_1}-b_1\right)\left(\widehat{\beta_2}-b_2\right)+\omega_{22}\left(\widehat{\beta_2}-b_2\right)^2\right]$$

Pour une probabilité donnée de $1-\alpha$ %, on peut alors définir une **ellipsoïde de concentration,** c'est-à-dire un ensemble de couple de point (b_1^*, b_2^*) tel que :

$$\underbrace{F_{1-\alpha}\left(2\,,(N-K)\right)}_{\text{Quantile à } 1-\alpha\ \%\ \text{donné dans les tables}}=\frac{1}{2\widehat{\sigma^2}}\left[\omega_{11}\left(\widehat{\beta_1}-b_1^*\right)^2+2\omega_{12}\left(\widehat{\beta_1}-b_1^*\right)\left(\widehat{\beta_2}-b_2^*\right)+\omega_{22}\left(\widehat{\beta_2}-b_2^*\right)^2\right]$$

Cette **ellipsoïde de concentration** correspond à un **intervalle de confiance bivarié** .

Pour chaque niveau de significativité α %, on aura une ellipse – un ensemble de points (b_1^*, b_2^*) qui satisfont l'équation ci-dessus.

Cette ellipse, centrée sur l'estimateur : $\left(\widehat{\beta_1}\,,\widehat{\beta_2}\right)$

- sera d'autant plus aplatie que les variances respectives des 2 estimateurs seront différentes,

- sera d'autant plus « inclinée » que la corrélation entre les 2 estimateurs sera importante.

DÉMONSTRATION

Notons $C = 2\widehat{\sigma^2}F_{1-\alpha}\big(2\,,(N-K)\big)$, on aura alors :

$$\omega_{11}\left(\widehat{\beta_1}-b_1^*\right)^2+2\omega_{12}\left(\widehat{\beta_1}-b_1^*\right)\left(\widehat{\beta_2}-b_2^*\right)+\omega_{22}\left(\widehat{\beta_2}-b_2^*\right)^2=\mathrm{C}$$

Pour une valeur donnée de $\left(\widehat{\beta_1}-b_1^*\right)$, on a une équation du second degré en $\widehat{\beta_2}-b_2^*$:

$$\omega_{22}\left(\widehat{\beta_2}-b_2^*\right)^2+\left(2\omega_{12}\left(\widehat{\beta_1}-b_1^*\right)\right)\left(\widehat{\beta_2}-b_2^*\right)+\left(\omega_{11}\left(\widehat{\beta_1}-b_1^*\right)^2-C\right)=0$$

La(les) solution(s) de cette équation donnera(ont) l'**ellipsoïde de concentration** :

$$\text{Si }\ 4\omega_{12}^2\left(\widehat{\beta_1}-b_1^*\right)^2-4\omega_{22}\left(\omega_{11}\left(\widehat{\beta_1}-b_1^*\right)^2-C\right)\begin{cases}>0 & \rightarrow\quad 2\,racines\\ =0 & \rightarrow\quad 1\,racine\\ <0 & \rightarrow\quad pas\ de\ \ racines\end{cases}$$

$$\left(\widehat{\beta_2}-b_2^*\right)=-\frac{\omega_{12}}{\omega_{22}}\left(\widehat{\beta_1}-b_1^*\right)\pm\frac{1}{\omega_{22}}\sqrt{\omega_{12}^2\left(\widehat{\beta_1}-b_1^*\right)^2-\omega_{22}\left(\omega_{11}\left(\widehat{\beta_1}-b_1^*\right)^2-C\right)}$$

En notant la corrélation : $\rho=\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}}$ et le rapport des écarts-type : $\theta=\sqrt{\omega_{11}/\omega_{22}}$, on aura finalement les points sur l'ellipse :

$$b_2^*=\widehat{\beta_2}+\rho\theta\left(\widehat{\beta_1}-b_1^*\right)\pm\theta\sqrt{(\rho^2-1)\left(\widehat{\beta_1}-b_1{}^*\right)^2+\frac{C}{\omega_{11}}}$$

En conséquence pour chaque valeur de b_1^*, on peut calculer 0, 1 ou 2 valeurs pour b_2^*, ce qui définit tous les points de l'ellipse.

➢ **Exemple avec 2 paramètres estimés sur 40 observations** :

$$\begin{pmatrix}\widehat{\beta_1}\\ \widehat{\beta_2}\end{pmatrix} = \begin{pmatrix}2\\ 1\end{pmatrix} \quad et \quad V\begin{pmatrix}\widehat{\beta_1}\\ \widehat{\beta_2}\end{pmatrix} = \begin{pmatrix}0.6 & -0.2\\ -0.2 & 0.1\end{pmatrix}$$

$$Corr\left(\widehat{\beta_1}, \widehat{\beta_2}\right) = \frac{-0.2}{\sqrt{0.6 \times 0.1}} = -0.816$$

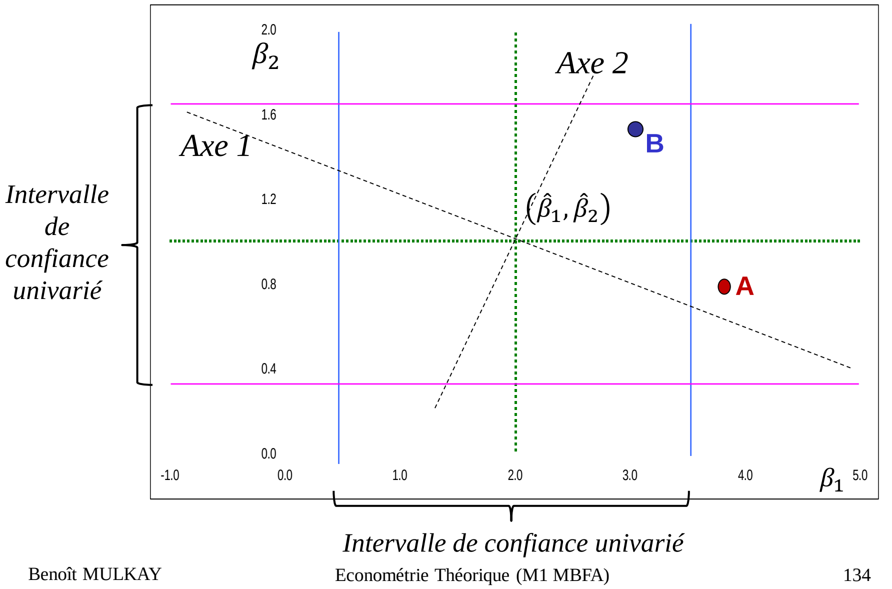
Statistique t pour chacun des paramètres avec le quantile $t_{0.975}(38) = 2.024$.

$$t_{\beta_1=0} = \frac{2}{\sqrt{0.6}} = \frac{2}{0.775} = 2.582$$

$$t_{\beta_2=0} = \frac{1}{\sqrt{0.1}} = \frac{2}{0.316} = 3.162$$

Ellipsoïde de concentration

Niveau : $\alpha = 5\ \%$



c) Autres formes du test F

Le test de Wald et le test *F* précédent sont calculés sur la régression non – contrainte (générale)…et on teste cette contrainte.

On aura une forme alternative si on peut estimer :

- le modèle non-contraint (*MCO* sans la restriction)
- et le modèle contraint (*MCC* avec la restriction)

On pourra alors utiliser une autre forme du test F qui est plus facilement calculable… parce qu'elle n'implique pas l'inversion de la matrice de variance – covariance des J combinaisons linéaires…

$$V\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r} \middle| \boldsymbol{X}\right) = \sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$$

Le résidu du modèle contraint $\boldsymbol{e}^*$ estimé par *MCC* s'écrit :

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

avec $\boldsymbol{e}$ le résidu du modèle estimé par *MCO* sans cette contrainte : $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

La somme des carrés des résidus du modèle contraint sera alors :

$$\begin{aligned} SCR^* = \boldsymbol{e}^{*\prime}\boldsymbol{e}^* &= \left(\boldsymbol{e} + \boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right)'\left(\boldsymbol{e} + \boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right) \\ &= \boldsymbol{e}'\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)'\boldsymbol{X}'\boldsymbol{e} + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)'\boldsymbol{X}'\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\ &= \boldsymbol{e}'\boldsymbol{e} + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)'\boldsymbol{X}'\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \qquad \text{parce que } \boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0} \end{aligned}$$

$$SCR^* - SCR = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)'\boldsymbol{X}'\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \geq 0$$

La somme des carrés des résidus du modèle contraint sera plus grande que la somme des carrés des résidus du modèle non contraint : $SCR^* \geq SCR$

parce qu'imposer une contrainte sur les paramètres à estimer, ne permet pas d'atteindre le minimum du critère sans contrainte !

Calculons la différence entre ces deux *SCR* pour des restrictions linéaires avec le résultat précédent : $\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})$

$$SCR^* - SCR = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$$

On obtient :

$$\begin{aligned} SCR^* - SCR &= \left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})\right)'\boldsymbol{X}'\boldsymbol{X}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})\right) \\ &= (\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}) \\ &= (\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}) \\ &= (\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}) \end{aligned}$$

qui est le numérateur de la statistique F :

$$F = \frac{(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r})}{J \times \widehat{\sigma^2}}$$

alors que le dénominateur de cette statistique est l'estimateur de la variance du **modèle non contraint** : $\widehat{\sigma^2} = \dfrac{SCR}{N-K} = \dfrac{\boldsymbol{e}'\boldsymbol{e}}{N-K}$

En conséquence la statistique F peut se réécrire en fonction de la différence des *SCR* des modèles contraint et non-contraint :

$$F = \frac{(SCR^* - SCR)/J}{SCR/(N-K)} \sim F(J, N-K)$$

Il suffit donc de calculer cette statistique à partir des *SCR* des 2 modèles : contraints et non contraints.
(si on peut les estimer !)

On peut aussi réécrire ce résultat en fonction des **coefficients de détermination** des 2 régressions :

$$R^{2*} = 1 - \frac{SCR^*}{SCT} \quad \text{et} \quad R^2 = 1 - \frac{SCR}{SCT}$$

qui peuvent être réécrits comme : $\frac{SCR^*}{SCT} = (1 - R^{2*})$ et $\frac{SCR}{SCT} = (1 - R^2)$

Le test F devient :

$$F = \frac{(SCR^* - SCR)/J}{SCR/(N-K)} = \frac{\frac{1}{J}\left(\frac{SCR^*}{SCT} - \frac{SCR}{SCT}\right)}{\frac{1}{N-K}\left(\frac{SCR}{SCT}\right)} = \frac{\frac{1}{J}\left(\left(\frac{SCR^*}{SCT} - 1\right) - \left(\frac{SCR}{SCT} - 1\right)\right)}{\frac{1}{N-K}\left(\frac{SCR}{SCT} - 1 + 1\right)}$$

$$= \frac{\frac{1}{J}(-R^{2*} + R^2)}{\frac{1}{N-K}(-R^2 + 1)} = \frac{\frac{1}{J}(R^2 - R^{2*})}{\frac{1}{N-K}(1 - R^2)}$$

$$\Rightarrow \quad F = \frac{R^2 - R^{2*}}{1 - R^2} \times \left(\frac{N-K}{J}\right)$$

d) Le Test du multiplicateur de Lagrange

Ici on veut tester simultanément plusieurs hypothèses simples ou plusieurs combinaisons linéaires des paramètres :

$$\begin{cases} H_0 : \boldsymbol{R\beta} - \boldsymbol{r} = \boldsymbol{0} \\ H_1 : \boldsymbol{R\beta} - \boldsymbol{r} \neq \boldsymbol{0} \end{cases}$$

On peut estimer le modèle linéaire $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ sous les contraintes $\boldsymbol{R\beta} - \boldsymbol{r} = \boldsymbol{0}$ par moindres carrés contraints (MCC) → *Voir Section III.1*

$$\begin{cases} \boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}) \\ \boldsymbol{\lambda}^* = (\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}) \end{cases}$$

Si les hypothèses du test sont vraies pour les MCO : $\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r} = \boldsymbol{0}$, alors l'estimateur des multiplicateurs de Lagrange est nul : $\boldsymbol{\lambda}^* = \boldsymbol{0}$.

Comme on connaît la variance de $\boldsymbol{\lambda}^*$, on peut construire un test de significativité (de nullité) de tous les éléments de $\boldsymbol{\lambda}^*$ sous la forme d'un test F ci-dessus :

$$\begin{cases} H_0 : \boldsymbol{\lambda} = \boldsymbol{0}_J \\ H_1 : \boldsymbol{\lambda} \neq \boldsymbol{0}_J \end{cases} \quad \rightarrow \quad J \text{ contraintes}$$

Ce test est appelé **<u>Test du Multiplicateur de Lagrange (LM)</u>**.

Il nécessite **uniquement** l'estimation du modèle **contraint** par *MCC*.

Cela revient à fixer : $\boldsymbol{R} = \boldsymbol{I}_J$ et $\boldsymbol{r} = \boldsymbol{0}_J$ dans un test F de $\boldsymbol{\lambda} = \boldsymbol{0}_J$:

$$LM = \frac{1}{J}(\boldsymbol{R}\boldsymbol{\lambda}^* - \boldsymbol{r})'\left(\boldsymbol{R}\hat{V}(\boldsymbol{\lambda}^*|\boldsymbol{X})\boldsymbol{R}'\right)^{-1}(\boldsymbol{R}\boldsymbol{\lambda}^* - \boldsymbol{r})$$

avec la variance : $\hat{V}(\boldsymbol{\lambda}^*|\boldsymbol{X}) = \widehat{\sigma^2}(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}$

Donc le test LM devient :

$$\begin{aligned}
LM &= \frac{1}{J}(\boldsymbol{\lambda}^*)'\left(\hat{V}(\boldsymbol{\lambda}^*|\boldsymbol{X})\right)^{-1}(\boldsymbol{\lambda}^*) \\
&= \frac{1}{J}\left((\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)\right)'\left(\widehat{\sigma^2}(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\right)^{-1}\left((\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)\right) \\
&= \frac{1}{J\widehat{\sigma^2}}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \\
&= \frac{1}{J\widehat{\sigma^2}}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \\
&= F \sim F(J, N-K)
\end{aligned}$$

Ce qui est équivalent au test F de la contrainte $\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r} = \boldsymbol{0}$.

141

141

<u>RÉSUMÉ</u>

- Les tests F et LM pour une contrainte d'exclusion sont strictement équivalents.
- Le test LM nécessite uniquement l'estimation du modèle **contraint** par *MCC*.
- La première forme du test F porte sur l'estimation unique du modèle **non contraint** par *MCO*.
- Alors que la seconde forme du test F basée sur les SCR demande les 2 estimations du modèle **contraint** (*MCC*) et du modèle **non contraint** (*MCO*) **.**

142

II.6. Les moindres carrés généralisés

Dans ce chapitre, on va remettre en question les hypothèses H4a et H4b sur les erreurs :

Hypothèse H4a : Homoscédasticité $E(\varepsilon_i^2|\boldsymbol{X}) = \sigma^2$ pour tout i

→ **Hypothèse H4a' : Hétéroscédasticité conditionnelle**

$$E(\varepsilon_i^2|\boldsymbol{X}) = {\sigma_i}^2$$

Hypothèse H4b : Absence d'autocorrélation $E(\varepsilon_i \varepsilon_j|\boldsymbol{X}) = 0$ pour tout $i \neq j$

→ **Hypothèse H4b' : Autocorrélation des erreurs**

$$E(\varepsilon_i \varepsilon_j|\boldsymbol{X}) \neq 0$$

a) *Transformation du modèle*

On considère un modèle linéaire général avec une matrice de variance – covariance des erreurs « non scalaire » ou « non sphérique » :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \text{avec } E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0} \text{ et } \boxed{E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X}) = \boldsymbol{\Phi}}$$

$\boldsymbol{\Phi}$ est la matrice de variance – covariance des erreurs.

Elle est carrée de dimension $N \times N$ symétrique et définie-positive et contient $N(N+1)/2$ éléments inconnus différents.

Sans perte de généralité, on peut toujours factoriser un élément de la matrice de variance-covariance. Par exemple : $E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X}) = \boldsymbol{\Phi} = \sigma^2 \boldsymbol{\Psi}$

En conséquence, on peut obtenir une matrice $N \times N$: $\boldsymbol{\Psi}^{1/2}$, appelée par convention la racine carrée de $\boldsymbol{\Psi}$, de telle sorte que :

$$\boldsymbol{\Psi} = \boldsymbol{\Psi}^{1/2} \boldsymbol{\Psi}^{1/2'}$$

Pour une matrice symétrique définie-positive, c'est la décomposition de Cholesky dans une matrice triangulaire inférieure.

Cette racine carrée est inversible $\left(\boldsymbol{\Psi}^{1/2}\right)^{-1} = \boldsymbol{\Psi}^{-1/2}$ telle que (*remarquez que la position de la matrice transposée !*) :

$$\boldsymbol{\Psi} = \boldsymbol{\Psi}^{1/2}\boldsymbol{\Psi}^{1/2'} \quad \rightarrow \quad \boldsymbol{\Psi}^{-1} = \boldsymbol{\Psi}^{-1/2'}\boldsymbol{\Psi}^{-1/2}$$

Plus loin, on va noter cette matrice inverse $\boldsymbol{P} = \boldsymbol{\Psi}^{-1/2}$ telle que $\boldsymbol{\Psi}^{-1} = \boldsymbol{P}'\boldsymbol{P}$

Attention : *Ici la matrice* $\boldsymbol{P}$ *n'est pas la matrice de projetcion vue dans la Section I.5 !!!*

Exemple 1 : Hétéroscédasticité (pour $N = 4$)

La matrice de variance-covariance des erreurs est proportionnelle à la matrice diagonale :

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

La racine carrée de cette matrice et son inverse sont évidemment :

$$\boldsymbol{\Psi}^{1/2} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \Rightarrow \quad \boldsymbol{\Psi}^{-1/2} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Exemple 2 : Autocorrélation (pour $N = 4$)

La matrice de variance-covariance des erreurs est proportionnelle à la matrice diagonale :

$$\boldsymbol{\Psi} = \begin{pmatrix} 1.000 & 0.800 & 0.640 & 0.512 \\ 0.800 & 1.000 & 0.800 & 0.640 \\ 0.640 & 0.800 & 1.000 & 0.800 \\ 0.512 & 0.640 & 0.800 & 1.000 \end{pmatrix}$$

La racine carrée de cette matrice (sa décomposition de Cholesky) est :

$$\boldsymbol{\Psi}^{1/2} = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.800 & 0.600 & 0.000 & 0.000 \\ 0.640 & 0.480 & 0.600 & 0.000 \\ 0.512 & 0.384 & 0.480 & 0.600 \end{pmatrix}$$

Et son inverse : $$\boldsymbol{\Psi}^{-1/2} = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ -1.333 & 1.667 & 0.000 & 0.000 \\ 0.000 & -1.333 & 1.667 & 0.000 \\ 0.000 & 0.000 & -1.333 & 1.667 \end{pmatrix}$$

On va transformer le modèle initial pour retrouver les hypothèses classiques d'application des *MCO* afin de satisfaire le théorème de Gauss – Markov.

On pré-multiplie le modèle par la matrice $\boldsymbol{P} = \boldsymbol{\Psi}^{-1/2}$:

$$\boldsymbol{Py} = \boldsymbol{PX\beta} + \boldsymbol{P\varepsilon} \quad \rightarrow \quad \tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \qquad \text{avec} \begin{cases} \tilde{\boldsymbol{y}} = \boldsymbol{Py} \\ \tilde{\boldsymbol{X}} = \boldsymbol{PX} \\ \tilde{\boldsymbol{\varepsilon}} = \boldsymbol{P\varepsilon} \end{cases}$$

Dans ce modèle de régression généralisée, l'espérance conditionnelle des erreurs transformées est nulle :

$$E(\tilde{\boldsymbol{\varepsilon}}|\tilde{\boldsymbol{X}}) = E(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = E(\boldsymbol{P\varepsilon}|\boldsymbol{X}) = \boldsymbol{P}E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}_N$$

du fait de l'hypothèse **H2** d'exogénéité forte et des propriétés de l'espérance conditionnelle.

Si l'hypothèse **H3** (non – multicolinéarité) est vérifiée, les variables explicatives transformées seront également linéairement indépendantes :

$$rang(\tilde{\boldsymbol{X}}) = rang(\boldsymbol{PX}) = rang(\boldsymbol{X}) = K$$ *Pourquoi ?*

Dans ce modèle de régression généralisée, la matrice de variance - covariance conditionnelle des erreurs transformées est « sphérique » :

$$\begin{aligned} V(\tilde{\boldsymbol{\varepsilon}}|\widetilde{\boldsymbol{X}}) = E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\widetilde{\boldsymbol{X}}) = E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\boldsymbol{X}) &= E(\boldsymbol{P}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{P}'|\boldsymbol{X}) \\ &= \boldsymbol{P}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{P}' \\ &= \boldsymbol{\Psi}^{-1/2}(\sigma^2\boldsymbol{\Psi})\boldsymbol{\Psi}^{-1/2\prime} \\ &= \sigma^2\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Psi}^{1/2}\boldsymbol{\Psi}^{1/2'}\boldsymbol{\Psi}^{-1/2\prime} \\ &= \sigma^2\boldsymbol{I}_N \end{aligned}$$

Ce qui correspond aux hypothèses **H4a** et **H4b** pour le modèle transformé.

On peut donc appliquer les *MCO* au modèle transformé…

b) *L'estimateur des Moindres Carrés Généralisés*

Les paramètres du modèle transformé vont être estimés par *MCO* :

$$\tilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \quad \text{avec} \begin{cases} E(\tilde{\boldsymbol{\varepsilon}}|\widetilde{\boldsymbol{X}}) = \boldsymbol{0}_N \\ V(\tilde{\boldsymbol{\varepsilon}}|\widetilde{\boldsymbol{X}}) = \sigma^2\boldsymbol{I}_N \end{cases}$$

L'estimateur MCO de ce modèle transformé :

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} &= (\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{y} \qquad \text{parce que } \boldsymbol{\Psi}^{-1} = \boldsymbol{P}'\boldsymbol{P} \end{aligned}$$

Cet estimateur est l'estimateur des ***Moindres Carrés Généralisés*** du modèle initial. (En anglais : Generalized Least Squares – GLS)

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

Il est **sans biais et efficace**.
Il est le meilleur estimateur linéaire sans biais de β pour le modèle transformé .
(*Théorème de Gauss – Markov, Section II.1(h)*).

Cet estimateur MCG est sans biais si les régresseurs sont strictement exogènes :

$$E\left(\widehat{\boldsymbol{\beta}_{MCG}}\middle|\boldsymbol{X}\right) = E((\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\beta}$$

L'erreur d'échantillonnage est : $\widehat{\boldsymbol{\beta}_{MCG}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon}$

Il est facile de montrer que la matrice de variance – covariance de l'estimateur *MCG* est donnée à partir du modèle transformé par :

$$V\left(\widehat{\boldsymbol{\beta}_{MCG}}\middle|\boldsymbol{X}\right) = E\left(\left(\widehat{\boldsymbol{\beta}_{MCG}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}_{MCG}} - \boldsymbol{\beta}\right)'\middle|\boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}$$

L'estimateur des MCG est le **meilleur estimateur linéaire sans biais** pour le modèle avec erreurs hétéroscédastiques et/ou autocorrélées :

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \boldsymbol{\Phi} = \sigma^2\boldsymbol{\Psi}$$

c) *Comparaison entre l'estimateur MCO et MCG*

On veut estimer le modèle : $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ avec $\begin{cases} E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}_N \\ E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \sigma^2\boldsymbol{\Psi} \end{cases}$

L'estimateur *MCG* : $\widehat{\boldsymbol{\beta}_{MCG}} = (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{y}$

est sans biais avec une variance minimale : $V\left(\widehat{\boldsymbol{\beta}_{MCG}}\middle|\boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}$

Alors que l'estimateur *MCO* : $\widehat{\boldsymbol{\beta}_{MCO}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

est aussi sans biais, parce que : $E\left(\widehat{\boldsymbol{\beta}_{MCO}}\middle|\boldsymbol{X}\right) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\beta}$

Mais sa variance n'est plus ici : $V\left(\widehat{\boldsymbol{\beta}_{MCO}}\middle|\boldsymbol{X}\right) \neq \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

En effet la démonstration de la *Section II.1.(e)* :

$$
\begin{aligned}
V(\widehat{\boldsymbol{\beta}_{MCO}}|\boldsymbol{X}) &= E\left((\widehat{\boldsymbol{\beta}_{MCO}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}_{MCO}}-\boldsymbol{\beta})'|\boldsymbol{X}\right) \\
&= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned}
$$

qui est la **vraie** variance de l'estimateur MCO si les erreurs sont hétéroscédastiques et/ou autocorrélées.

On parle généralement d'une formule « *sandwich* » pour la matrice de variance-covariance de l'estimateur MCO avec la matrice $\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}$ est encadrée à gauche et à droite par la matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Remarquez que si la matrice de covariance $\boldsymbol{\Psi} = \boldsymbol{I}_N$, on retrouve le résultat précédent :

$$V(\widehat{\boldsymbol{\beta}_{MCO}}|\boldsymbol{X}, \boldsymbol{\Psi} = \boldsymbol{I}_N) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

PROPOSITION :

la vraie matrice de variance-covariance de l'estimateur *MCO* est plus grande (au sens matriciel) que celle de l'estimateur *MCG* :

$$V(\widehat{\boldsymbol{\beta}_{MCO}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} > \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1} = V(\widehat{\boldsymbol{\beta}_{MCG}}|\boldsymbol{X})$$

En effet la différence des variances est une matrice semi-définie positive :

$$
\begin{aligned}
\underbrace{V(\widehat{\boldsymbol{\beta}_{MCO}}|\boldsymbol{X})}_{\boldsymbol{\Sigma}_{MCO}} - \underbrace{V(\widehat{\boldsymbol{\beta}_{MCG}}|\boldsymbol{X})}_{\boldsymbol{\Sigma}_{MCG}} &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1} \\
&= \sigma^2\boldsymbol{A}'\boldsymbol{\Psi}\boldsymbol{A} \qquad \text{avec } \boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}
\end{aligned}
$$

Cette matrice $\boldsymbol{A}$ est de rang K : $rang(\boldsymbol{A}) = K < N$

En conséquence, l'estimateur des MCO a une plus grande variance que celui des MCG, bien qu'ils soient tous les deux sans biais !

L'estimateur des MCO est **moins précis** que l'estimateur des MCG. Il faut cependant utiliser la vraie matrice de variance-covariance de l'estimateur. L'estimateur MCG est en conséquence préférable à l'estimateur MCO.

d) *Le critère d'estimation des MCG*

Le critère d'estimation des moindres carrés ordinaires s'applique ici au modèle transformé :

$$\min_{\beta} \sum_{i=1}^{N} \tilde{\varepsilon}_i^{\,2} = \min_{\beta} \tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}$$

Ce qui revient à : $\min_{\boldsymbol{\beta}} \boldsymbol{\varepsilon}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{\varepsilon} = \min_{\boldsymbol{\beta}} \boldsymbol{\varepsilon}'\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})$

Le critère d'estimation des moindres carrés généralisés est donc la minimisation d'une somme pondérée des carrés de l'erreur !

On effectue une projection oblique et non plus orthogonale sur le plan engendré par les régresseurs.

Conséquences :

1) Même avec une constante dans le modèle, le plan de régression ne passe plus par le point moyen de l'échantillon.

2) La somme des résidus n'est plus nulle.

3) Le R^2 peut devenir négatif !

4) Mais …

$$\begin{cases} \boldsymbol{e}_{MCG} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_{MCG}} \\ \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{e}_{MCG} = \boldsymbol{0}_K \quad \text{et} \quad \boldsymbol{j}_N{}'\boldsymbol{\Psi}^{-1}\boldsymbol{e}_{MCG} = 0 \\ \sum_{i=1}^{N} \psi_{x,i} x_i e_{MCG,i} = 0 \quad \text{et} \quad \sum_{i=1}^{N} \psi_{x,i} e_{MCG,i} = 0 \end{cases}$$

De même, il est sans intérêt de comparer le R^2 d'une régression *MCO* avec le R^2 d'une régression *MCG* !

e) Un estimateur de la variance de l'erreur

Pour calculer la variance de l'estimateur *MCG*, on doit estimer σ^2 :

$$V\left(\widehat{\boldsymbol{\beta}_{MCG}}\middle|\boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}$$

Si on utilise les résidus sur les données initiales : $\boldsymbol{e}_{MCG} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_{MCG}}$ pour calculer la somme des carrés des résidus et l'estimateur *MCO* de la variance :

$$\widehat{\sigma^2} = \frac{SCR}{N-K} = \frac{\boldsymbol{e}_{MCG}{}'\boldsymbol{e}_{MCG}}{N-K}$$

on peut démontrer que cet estimateur est biaisé !

$$E\left(\widehat{\sigma^2}\middle|X\right) = \frac{E(\boldsymbol{e}_{MCG}{}'\boldsymbol{e}_{MCG}|X)}{N-K} \neq \sigma^2$$

Cependant si on utilise les résidus du modèle transformé, on obtient un estimateur sans biais de σ^2 :

$$\widehat{\sigma^2_{MCG}} = \frac{\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}_{MCG}}\right)'\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}_{MCG}}\right)}{N-K}$$

que l'on peut réécrire sous la forme :

$$\begin{aligned}
\widehat{\sigma^2_{MCG}} &= \frac{\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}_{MCG}}\right)'\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}_{MCG}}\right)}{N-K} \\
&= \frac{1}{N-K}\left(\boldsymbol{Py} - \boldsymbol{PX}\widehat{\boldsymbol{\beta}_{MCG}}\right)'\left(\boldsymbol{Py} - \boldsymbol{PX}\widehat{\boldsymbol{\beta}_{MCG}}\right) \\
&= \frac{1}{N-K}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_{MCG}}\right)'\boldsymbol{P}'\boldsymbol{P}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_{MCG}}\right) \\
&= \frac{1}{N-K}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_{MCG}}\right)'\boldsymbol{\Psi}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_{MCG}}\right) \\
&= \frac{1}{N-K}\boldsymbol{e}_{MCG}{}'\boldsymbol{\Psi}^{-1}\boldsymbol{e}_{MCG} \qquad \text{avec } \boldsymbol{e}_{MCG} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_{MCG}}
\end{aligned}$$

On peut montrer, de la même manière que précédemment, que cet estimateur est sans biais :

$$E\left(\widehat{\sigma^2_{MCG}}|\boldsymbol{X}\right) = \sigma^2$$

DEMONSTRATION

$$E\left(\widehat{\sigma_{MCG}}^2|\boldsymbol{X}\right) = E\left(\frac{\boldsymbol{e}_{MCG}{}'\boldsymbol{\Psi}^{-1}\boldsymbol{e}_{MCG}}{N-K}\middle|\boldsymbol{X}\right)$$

Le résidus *MCG* s'écrit :

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

Remarquez que $\boldsymbol{M}_{X\Psi}$ n'est pas symétrique !
De plus : $\boldsymbol{M}_{X\Psi}\boldsymbol{y} = \boldsymbol{M}_{X\Psi}\boldsymbol{\varepsilon}$ pare que $\boldsymbol{M}_{X\Psi}\boldsymbol{X} = \boldsymbol{0}$

Donc l'espérance de l'estimateur *MCG* de la variance :

$$\begin{aligned}E\left(\widehat{\sigma_{MCG}}^2|\boldsymbol{X}\right) &= \frac{1}{N-K}E(\boldsymbol{\varepsilon}'\boldsymbol{M}'_{X\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{M}_{X\Psi}\boldsymbol{\varepsilon}|\boldsymbol{X})\\ &= \frac{1}{N-K}tr[E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{M}'_{X\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{M}_{X\Psi}]\\ &= \frac{\sigma^2}{N-K}tr[\boldsymbol{\Psi}\boldsymbol{M}'_{X\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{M}_{X\Psi}]\end{aligned}$$

$$tr[\boldsymbol{\Psi}\boldsymbol{M}'_{X\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{M}_{X\Psi}]$$

$$\begin{aligned}&= tr[\boldsymbol{\Psi}(\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1})'\boldsymbol{\Psi}^{-1}(\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1})]\\ &= tr[(\boldsymbol{\Psi} - \boldsymbol{\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1})]\\ &= tr\begin{bmatrix}\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\\ +\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\end{bmatrix}\\ &= tr[\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}]\\ &= tr(\boldsymbol{I}_N) - tr(\boldsymbol{I}_K)\\ &= N-K\end{aligned}$$

En conséquence : $E\left(\widehat{\sigma_{MCG}}^2|X\right) = \dfrac{\sigma^2}{N-K}tr[\Psi M'_{X\Psi}\Psi^{-1}M_{X\Psi}] = \dfrac{\sigma^2}{N-K}(N-K) = \sigma^2$

Cet estimateur de la variance de l'erreur σ^2 est sans biais !

CQFD