Économétrie
Les MCO quand tout va bien

Stéphane Adjemian

stephane.adjemian@univ-lemans.fr

Septembre 2024

Le modèle de la nature

Un modèle linéaire

On suppose que les données (y) sont générées par le modèle suivant :

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_K x_{K,t} + \varepsilon_t$$

- y est la variable endogène (ou expliquée).
- x_k, $k = 1, \ldots, K$, sont les variables exogènes (ou explicatives).
- ε_t est une variable aléatoire, elle rend compte de ce qui ne peut être expliqué par les variables x_k.
- La nature nous donne un échantillon $\{y_t, x_{1,t}, \ldots, x_{K,t}\}_{t=1}^{T}$ (T est le nombre d'observations).
- Les K variables exogènes peuvent être déterministes (pour simplifier) ou aléatoires.

Le modèle de la nature

Variables exogènes déterministes ou aléatoires

- Les variables exogènes sont déterministes $\Leftrightarrow$ Lorsque l'économètre s'adresse à la nature afin d'obtenir un nouvel échantillon, elle lui renvoit toujours les mêmes valeurs pour les variables exogènes.

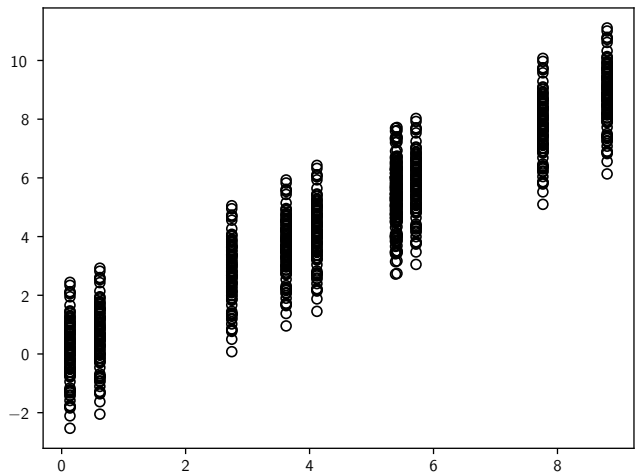
- Dans le cas de variables exogènes non stochastiques, ε est la seule source d'aléa.

$\Rightarrow$ La loi de y est directement déduite de celle de ε.

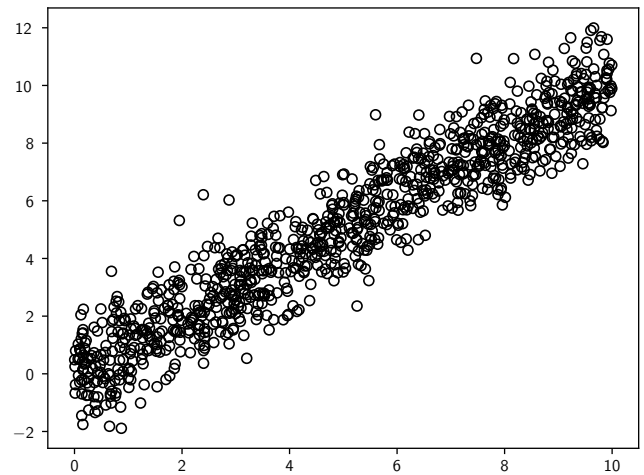
- Cette hypothèse simplifie grandement l'étude des propriétés de l'estimateur des Moindres Carrés Ordinaires, mais dans certaines circonstances elle est beaucoup trop forte (voire n'a aucun sens comme dans le cas des modèles dynamiques).

Le modèle de la nature

Variables exogènes déterministes ou aléatoires



(a) x déterministe.



(b) x stochastique.

Figure 1: Le modèle de la nature est $y_t = x_t + \varepsilon_t$ avec x_t une variable exogène prenant des valeurs dans l'intervalle $[0, 10]$ et ε_t une variable alléatoire gaussienne centrée réduite avec $\mathbb{E}[\varepsilon_t \varepsilon_s] = 0$ si $s \neq t$. Chaque figure représente 100 échantillons de 10 observations. Les codes pour reproduire ces graphiques sont disponibles ici.

Le modèle de la nature

Représentation matricielle

- Ce modèle peut être représenté matriciellement sous la forme :

$$\mathbf{y} = X\beta + \varepsilon$$

avec $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ et $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T)'$ des vecteurs $T \times 1$, $\beta = (\beta_1, \ldots, \beta_K)'$ un vecteur $K \times 1$ et

$$X = \begin{pmatrix} x_{1,1} & x_{2,1} & \ldots & x_{K,1} \\ x_{1,2} & x_{2,2} & \ldots & x_{K,2} \\ \vdots & \vdots & & \vdots \\ x_{1,T} & x_{2,2} & \ldots & x_{K,T} \end{pmatrix}$$

une matrice $T \times K$.

- On notera $\mathbf{x}_t = (x_{1,t}, x_{2,t}, \ldots, x_{K,t})$ la t-ième observation pour les exogènes (un vecteur $1 \times K$).

Le modèle de la nature

Hypothèses

$\mathscr{H}_1$ Les variables exogènes sont déterministes.

$\mathscr{H}_2$ X est une matrice de rang $K < T$ et vérifie :

$$\lim_{T \to \infty} \frac{X'X}{T} = Q$$

où Q est une matrice symétrique définie positive.

$\mathscr{H}_3$ ε suit loi normale multivariée d'espérance nulle et de variance $\sigma_\varepsilon^2 I_T$.

Le modèle de l'économètre

Pas de mauvaise spécification

- On suppose que l'économmètre connaît la forme du modèle de la nature.
- Mais il ne connaît pas les valeurs des paramètres β qu'il va chercher à estimer...
- ... En utilisant l'unique échantillon que lui donne la nature.
- Le modèle empirique est donc :

$$\mathbf{y} = X\mathbf{b} + \epsilon$$

où $\mathbf{b}$ est le vecteur des paramètres du modèle empirique.

Estimateur des Moindres Carrés Ordinaires

Définition.

L'estimateur des MCO de $\mathbf{b}$ minimise la somme des carrés des résidus:

$$\begin{aligned}\hat{\mathbf{b}} &= \arg\min_{\mathbf{b}} \sum_{t=1}^{T} \epsilon_t^2 \\ &= \arg\min_{\mathbf{b}} (\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b})\end{aligned}$$

- ϵ_t^2 est une mesure de la distance entre l'observation y_t et la prédiction $\mathbf{x}_t\mathbf{b}$.
- Le choix de la distance entre observation et prédiction est arbitraire.
- L'estimateur des MCO minimise la somme des carrés des erreurs de prédiction.

Estimateur des Moindres Carrés Ordinaires

Formule

Théorème 1

L'estimateur des MCO de $\mathbf{b}$ *dans le modèle* $\mathbf{y} = X\mathbf{b} + \epsilon$ *est :*

$$\hat{\mathbf{b}} = (X'X)^{-1} X'\mathbf{y}$$

Preuve. La forme quadratique que nous devons minimiser s'écrit en développant : $\mathcal{S}(\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'X\mathbf{b} - \mathbf{b}'X'\mathbf{y} + \mathbf{b}'X'X\mathbf{b}$, une fonction de $\mathbb{R}^K$ dans $\mathbb{R}^+$. En notant que $\mathbf{y}'X\mathbf{b} = \mathbf{b}'X'\mathbf{y}$ puisque la transposée d'un scalaire est égale au scalaire, on a : $\mathcal{S}(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'X'\mathbf{y} + \mathbf{b}'X'X\mathbf{b}$. En annulant la dérivée de $\mathcal{S}$ par rapport à $\mathbf{b}$ on obtient la condition du premier ordre :

$$-2X'\mathbf{y} + 2X'X\hat{\mathbf{b}} = 0$$

soit de façon équivalente :

$$\hat{\mathbf{b}} = (X'X)^{-1} X'\mathbf{y}$$

car $X'X$ est de plein rang par $\mathscr{H}_2$. C.Q.F.D.

Estimateur des Moindres Carrés Ordinaires

Remarques sur le théorème 1

- La condition du premier ordre qui permet d'identifier $\hat{\mathbf{b}}$ est linéaire car l'objectif est quadratique.

- La matrice hessienne de l'objectif est bien définie positive :

$$\frac{\partial^2 \mathcal{S}}{\partial \mathbf{b} \partial \mathbf{b}'} = 2X'X$$

- Notons aussi qu'il est possible de réécrire l'objectif de la façon suivante :

$$\mathcal{S}(\mathbf{b}) = \mathcal{S}(\hat{\mathbf{b}}) + \left(\mathbf{b} - \hat{\mathbf{b}}\right)' X'X \left(\mathbf{b} - \hat{\mathbf{b}}\right)$$

puisque le dernier terme est toujours positif (car la matrice hessienne $X'X$ est définie positive), on voit directement que la somme des carrés des résidus est minimale en $\hat{\mathbf{b}}$.

Estimateur des Moindres Carrés Ordinaires

Résidus estimés

Définition.

Les résidus estimés sont définis par :

$$\hat{\epsilon} = \mathbf{y} - X\hat{\mathbf{b}}$$

Proposition 1

Les résidus estimés sont orthogonaux aux variables explicatives :

$$X\hat{\epsilon} = 0$$

Corollaire 1

Si les variables explicatives contiennent une constante alors les résidus somment à zéro.

- **Preuve de la proposition 1.** En partant de la CNO pour $\hat{\mathbf{b}}$:

$$-2X'\mathbf{y} + 2X'X\hat{\mathbf{b}} = 0$$

et en factorisant on a directement :

$$X'(\mathbf{y} - X\hat{\mathbf{b}}) = 0$$

soit par définition des résidus estimés : $X'\hat{\epsilon} = 0$ C.Q.F.D.

- **Preuve du corollaire 1.** Supposons, sans perte de généralité, que les éléments de la première colonne de X soit tous égaux à 1. Le premier élément du vecteur $K \times 1$ $X'\hat{\epsilon}$ est alors :

$$\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_T \end{pmatrix} = \sum_{t=1}^{T} \hat{\epsilon}_t$$

qui doit être égal à zéro. C.Q.F.D.

- Le corollaire 1 implique en particulier que les résidus sont nécessairement de moyenne nulle dès lors que le modèle contient une constante.

Estimateur des Moindres Carrés Ordinaires

Sommes de carrés

Définition.

On note $\hat{y}_t = x_t \hat{\mathbf{b}}$ le prédicteur de y_t, $\bar{y} = T^{-1} \sum_{t=1}^{T} y_t$ la moyenne arithmétique, et on définit :

$$SSE = \hat{\epsilon}' \hat{\epsilon} = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2$$

$$SSR = \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2$$

$$SST = \sum_{t=1}^{T} (y_t - \bar{y})^2$$

Proposition 2

Si les variables explicatives contiennent une constante alors :

$$SST = SSR + SSE$$

- **Lemme.** La somme des carrés des résidus vérifie $\hat{\epsilon}'\hat{\epsilon} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}}$.

Preuve. Par définition des résidus estimés, on a :

$$\begin{aligned}
\hat{\epsilon}'\hat{\epsilon} &= (\mathbf{y} - X\hat{\mathbf{b}})'(\mathbf{y} - X\hat{\mathbf{b}}) \\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'X\hat{\mathbf{b}} - \hat{\mathbf{b}}'X'\mathbf{y} + \hat{\mathbf{b}}'X'X\hat{\mathbf{b}} \\
&= \mathbf{y}'\mathbf{y} - 2\hat{\mathbf{b}}'X'\mathbf{y} + \hat{\mathbf{b}}'X'X\hat{\mathbf{b}} \\
&= \mathbf{y}'\mathbf{y} - 2\hat{\mathbf{b}}'X'X\hat{\mathbf{b}} + \hat{\mathbf{b}}'X'X\hat{\mathbf{b}} \\
&= \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'X'X\hat{\mathbf{b}} \\
&= \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}}
\end{aligned}$$

où on passe à la troisième égalité en notant qu'un scalaire est égal à sa transposée, puis à la quatrième en utilisant la CNO pour $\hat{\mathbf{b}}$ qui nous dit que $X'\mathbf{y} = X'X\hat{\mathbf{b}}$. C.Q.F.D.

- **Preuve de la proposition 2.** On sait, voir le lemme donné au dessus, que :

$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\epsilon}'\hat{\epsilon}$$

En retranchant $T\bar{y}^2$ sur les deux membres, il vient :

$$\mathbf{y}'\mathbf{y} - T\bar{y}^2 = \hat{\mathbf{y}}'\hat{\mathbf{y}} - T\bar{y}^2 + \hat{\epsilon}'\hat{\epsilon}$$

Par ailleurs :

$$\begin{aligned} SST &= \sum_{t=1}^{T}(y_t - \bar{y})^2 \\ &= \sum_{t=1}^{T} y_t^2 - 2\bar{y}\sum_{t=1}^{T} y_t + T\bar{y}^2 \\ &= \mathbf{y}'\mathbf{y} - 2T\bar{y}^2 + T\bar{y}^2 \\ &= \mathbf{y}'\mathbf{y} - T\bar{y}^2 \end{aligned}$$

nous retrouvons donc SST sur le membre de gauche. On a aussi :

$$\begin{aligned} SSR &= \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 \\ &= \sum_{t=1}^{T} \hat{y}_t^2 - 2\bar{y}\sum_{t=1}^{T} \hat{y}_t + T\bar{y}^2 \end{aligned}$$

En rappelant que $y_t = \mathbf{x}_t\hat{\mathbf{b}} + \hat{\epsilon}_t = \hat{y}_t + \hat{\epsilon}_t$, on a encore :

$$SSR = \sum_{t=1}^{T} \hat{y}_t^2 - 2\bar{y}\sum_{t=1}^{T}(y_t - \hat{\epsilon}_t) + T\bar{y}^2$$

et puisque les résidus estimés somment à zéro quand X contient une constante :

$$SSR = \sum_{t=1}^{T} \hat{y}_t^2 - 2\bar{y} \sum_{t=1}^{T} y_t + T\bar{y}^2$$

et donc :

$$SSR = \hat{\mathbf{y}}'\hat{\mathbf{y}} - T\bar{y}^2$$

Ainsi, nous avons finalement :

$$SST = SSR + SSE$$

C.Q.F.D.

Estimateur des Moindres Carrés Ordinaires

Coefficient de détermination

Définition.

Le coefficient de détermination est :

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

Proposition 3

Si les variables explicatives contiennent une constante alors :

1. $R^2 = \frac{SSR}{SST}$
2. $0 \leq R^2 \leq 1$
3. $\sqrt{R^2} = \texttt{corr}(y, \hat{y})$

▶ **Preuve de la proposition 3. (1)** En utilisant la proposition 2 on a :

$$\frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

(2) Comme SSR et SST sont des sommes de carrés, le R^2 ne peut-être négatif. Comme, pour la même raison, $\frac{SSE}{SST} \geq 0$ le R^2 ne peut être supérieur à 1. **(3)** La corrélation entre y et $\hat{y}$ est définie par :

$$\texttt{corr}(y, \hat{y}) = \frac{\texttt{cov}(\texttt{y}, \hat{\texttt{y}})}{\sqrt{\mathbb{V}[y]\mathbb{V}[\hat{y}]}}$$

La variance de la variable expliquée est :

$$\mathbb{V}[y] = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 = \frac{SST}{T}$$

La variance de la prédiction, en notant que $\bar{y}$ est aussi la moyenne de $\hat{y}$ puisque les résidus estimés somment à zéro, est :

$$\mathbb{V}[\hat{y}] = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 = \frac{SSR}{T}$$

La covariance entre y et $\hat{y}$ est :

$$\texttt{cov}(y, \hat{y}) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})(\hat{y}_t - \bar{y})$$

En développant sous la somme :

$$\begin{aligned}\texttt{cov}(y,\hat{y}) &= \frac{1}{T}\sum_{t=1}^{T} y_t\hat{y}_t - \bar{y}y_t - \bar{y}\hat{y}_t + \bar{y}^2 \\ &= \frac{1}{T}\sum_{t=1}^{T} y_t\hat{y}_t - \bar{y}^2\end{aligned}$$

Puisque $y_t = \hat{y}_t + \hat{\epsilon}_t$ et $\sum_{t=1}^{T}\hat{\epsilon}_t\hat{y}_t = \hat{\mathbf{y}}'\hat{\epsilon} = \hat{\mathbf{b}}'X'\hat{\epsilon} = 0$ car les résidus estimés sont orthogonaux aux variables explicatives, nous avons :

$$\texttt{cov}(y,\hat{y}) = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^2 - \bar{y}^2 = \frac{SSR}{T}$$

Notons en passant que la covariance entre y et $\hat{y}$ est nécessairement positive (c'est heureux). La corrélation est donc :

$$\texttt{corr}(y,\hat{y}) = \sqrt{\frac{SSR}{SST}} = \sqrt{R^2}$$

C.Q.F.D.

Estimateur des Moindres Carrés Ordinaires

Remarques sur le R^2

- Par définition le R^2 est toujours plus petit que 1.
- Pour que le R^2 soit positif il faut que le modèle contienne une constante.
- Dans ce cas, le R^2 mesure la contribution de la variabilité de x à la variance de y.
- Les prédictions *in-sample* ($\hat{\mathbf{y}}$) sont d'autant meilleures (proches de $\mathbf{y}$) que le R^2 est proche de 1.

Estimateur des Moindres Carrés Ordinaires

Remarques sur le R^2 (suite)

- Un R^2 proche de 1 ne garantit pas que le modèle soit «bon».
- De bonnes prévisions *in-sample* ne garantissent pas de bonnes prévisions *out-of-sample*.
- Le R^2 ne préjuge pas de la relation entre y et x (ce n'est pas parce que le R^2 est faible que x n'a pas d'effet significatif sur y).
- Un grand R^2 proche de 1 ne veut pas dire que la variable x explique «bien»la variable y. Nous obtiendrions le même R^2 en inversant le modèle empirique $\Rightarrow$ Pas d'interprétation causale.

Estimateur des Moindres Carrés Ordinaires

- Échantillons (différents DGP) avec mêmes moments d'ordre 1 et 2.
- Même modèle empirique $\Rightarrow$ Estimations et $R^2 = 0,67$ identiques.

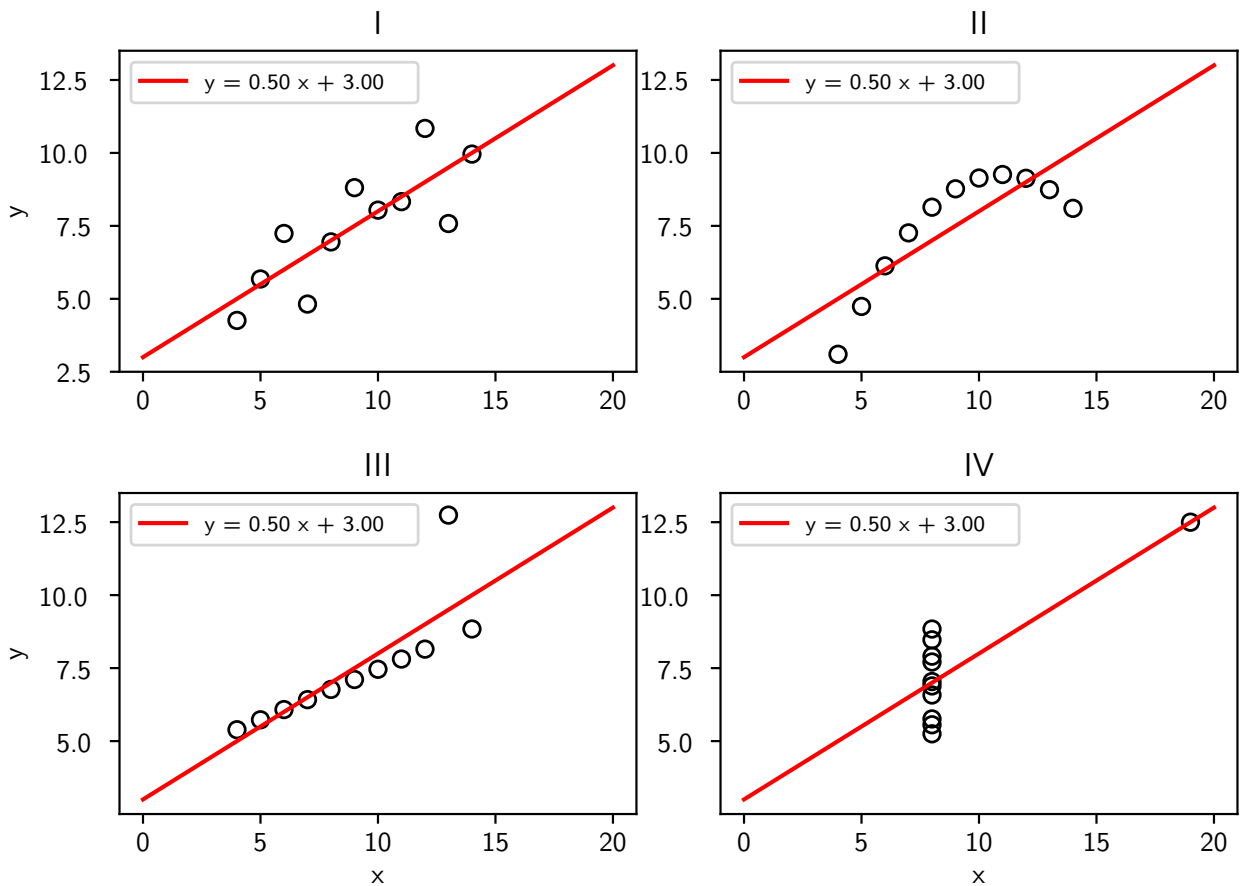


Figure 2: Les quatre échantillons d'Anscombe (1973)

Estimateur des Moindres Carrés Ordinaires

Propriétés statistiques de $\hat{\mathbf{b}}$

- L'estimateur $\hat{\mathbf{b}}$ est une fonction de X (supposé déterministe) et $\mathbf{y}$ (aléatoire).

$\Rightarrow$ L'estimateur $\hat{\mathbf{b}}$ est une variable aléaoire.

$\Rightarrow$ L'estimateur $\hat{\mathbf{b}}$ est toujours différent de β (sauf si $T \to \infty$).

Proposition 4

$\hat{\mathbf{b}}$ est un estimateur sans biais de β (en moyenne $\hat{\mathbf{b}}$ est égal à β), sa variance est :

$$\mathbb{V}\left[\hat{\mathbf{b}}\right] = \sigma_\varepsilon^2 (X'X)^{-1}$$

Preuve de la proposition 4. Pour étudier les propriétés de l'estimateur des MCO, il faut substituer le processus générateur des données dans l'expression de l'estimateur. On a :

$$\begin{aligned}\hat{\mathbf{b}} &= (X'X)^{-1}X'\mathbf{y} \\ &= (X'X)^{-1}X'(X\boldsymbol{\beta}+\varepsilon) \\ &= \boldsymbol{\beta}+(X'X)^{-1}X'\varepsilon\end{aligned}$$

Ainsi l'espérance de l'estimateur des MCO est :

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{b}}\right] &= \boldsymbol{\beta}+\mathbb{E}\left[(X'X)^{-1}X'\varepsilon\right] \\ &= \boldsymbol{\beta}+(X'X)^{-1}X'\mathbb{E}\left[\varepsilon\right] \quad \text{car } X \text{ est déterministe} \\ &= \boldsymbol{\beta} \quad \text{car } \varepsilon \text{ est d'espérance nulle.}\end{aligned}$$

L'estimateur est donc sans biais, en moyenne l'estimateur des MCO est égal à la vraie valeur des paramètres. La variance est l'espérance du carré de l'écart à l'espérance :

$$\mathbb{V}\left[\hat{\mathbf{b}}\right] = \mathbb{E}\left[\left(\hat{\mathbf{b}}-\boldsymbol{\beta}\right)\left(\hat{\mathbf{b}}-\boldsymbol{\beta}\right)'\right]$$

une matrice $K \times K$. En substituant l'espression de l'estimateur en fonction de ε, on obtient :

$$\begin{aligned}\mathbb{V}\left[\hat{\mathbf{b}}\right] &= \mathbb{E}\left[\left((X'X)^{-1}X'\varepsilon\right)\left((X'X)^{-1}X'\varepsilon\right)'\right] \\ &= \mathbb{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right]\end{aligned}$$

Puisque X est déterministe :

$$\begin{aligned}\mathbb{V}\left[\hat{\mathbf{b}}\right] &= (X'X)^{-1}X'\mathbb{E}\left[\varepsilon\varepsilon'\right]X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma_\varepsilon^2 I_T X(X'X)^{-1}\end{aligned}$$

par $\mathscr{H}_3$. Enfin :

$$\begin{aligned}\mathbb{V}\left[\hat{\mathbf{b}}\right] &= \sigma_\varepsilon^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma_\varepsilon^2(X'X)^{-1}\end{aligned}$$

C.Q.F.D.

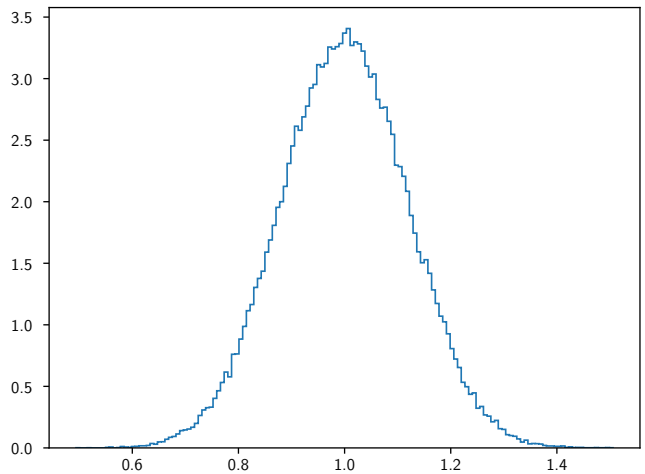
- L'estimateur des MCO est d'autant moins précis (sa variance est d'autant plus grande) que la variance des erreurs ε est importante.

- L'esimateur des MCO est d'autant moins précis que la variabilité des variables explicatives, la matrice $X'X$ «petite», est faible.

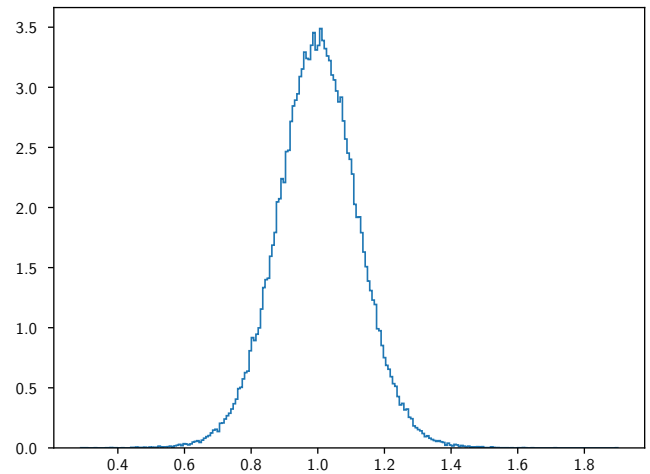
$\Rightarrow$ Il est plus facile d'identifier l'effet d'une variable explicative sur y si le signal (la variance de la variable explicative) est important relativement au bruit (σ_ε^2).

Estimateur des Moindres Carrés Ordinaires

Propriétés statistiques de $\hat{\mathbf{b}}$



(a) x déterministe.



(b) x stochastique.

Figure 3: Le modèle de la nature est $y_t = x_t + \varepsilon_t$ avec x_t une variable exogène (déterministe ou stochastique) prenant des valeurs dans l'intervalle $[0, 10]$ et ε_t un bruit blanc gaussien. On simule 100000 échantillons de 10 observations, pour chaque échantillon on estime le modèle empirique $y_t = b_0 + b_1 x_t + \epsilon_t$. Chaque figure représente la distribution (avec un histogramme) de l'esimateur $\hat{b}_1$. Comme attendu, $\hat{b}_1$ est centré sur 1, la vrai valeur de la pente. Les codes pour reproduire ces graphiques sont disponibles ici.

Estimateur des Moindres Carrés Ordinaires

Propriétés statistiques de $\hat{\mathbf{b}}$

Théorème 2 (Gauss Markov)

$\hat{\mathbf{b}}$ *est le meilleur estimateur linéaire sans biais (BLUE) de* β.

Remarque. L'estimateur $\lambda\hat{\mathbf{b}}$, où λ est un vecteur de paramètres $1 \times K$, est aussi le meilleur estimateur linéaire sans biais de $\lambda\beta$.

Preuve de la proposition 2. Nous avons déjà montré que $\hat{\mathbf{b}}$ est un estimateur sans biais de β. Soit un estimateur linéaire en $\mathbf{y}$: $\tilde{\mathbf{b}} = C\mathbf{y}$ avec, sans perte de généralité, $C = (X'X)^{-1}X'' + D$. L'absence de biais pose des contraintes sur la matrice D :

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathbf{b}}\right] &= \mathbb{E}\left[\left((X'X)^{-1}X' + D\right)(X\beta + \varepsilon)\right] \\ &= \beta + DX\beta\end{aligned}$$

pour que l'estimateur $\tilde{\mathbf{b}}$ soit non biaisé, il faut et il suffit que la matrice DX soit nulle. Calculons la variance de cet estimateur et montrons qu'elle est plus grande que la variance de l'estimateur des MCO, dans le sens où la matrice $\mathbb{V}\left[\tilde{\mathbf{b}}\right] - \mathbb{V}\left[\hat{\mathbf{b}}\right]$ est définie positive. Si $\tilde{\mathbf{b}}$ est sans biais, on a :

$$\begin{aligned}\mathbb{V}\left[\tilde{\mathbf{b}}\right] &= \mathbb{E}\left[\left(\tilde{\mathbf{b}} - \beta\right)\left(\tilde{\mathbf{b}} - \beta\right)'\right] \\ &= \mathbb{E}\left[\left(\left((X'X)^{-1}X' + D\right)(X\beta + \varepsilon) - \beta\right)\left(\left((X'X)^{-1}X' + D\right)(X\beta + \varepsilon) - \beta\right)'\right] \\ &= \mathbb{E}\left[\left((X'X)^{-1}X'\varepsilon + D\varepsilon + DX\beta\right)\left((X'X)^{-1}X'\varepsilon + D\varepsilon + DX\beta\right)'\right] \\ &= \mathbb{E}\left[\left((X'X)^{-1}X' + D\right)\varepsilon\varepsilon'\left((X'X)^{-1}X' + D\right)'\right] \\ &= \sigma_\varepsilon^2\left((X'X)^{-1}X' + D\right)\left((X'X)^{-1}X' + D\right)' \\ &= \sigma_\varepsilon^2\left((X'X)^{-1} + DD'\right) \\ &= \mathbb{V}\left[\hat{\mathbf{b}}\right] + \sigma_\varepsilon^2 DD'\end{aligned}$$

Comme DD' est une matrice définie positive, la variance de $\tilde{\mathbf{b}}$ est plus grande que celle

de $\hat{\mathbf{b}}$. L'estimateur des MCO est donc bien le meilleur, au sens de la réduction de la variance, estimateur linéaire sans biais de β. C.Q.F.D.

Estimateur des Moindres Carrés Ordinaires

Propriétés statistiques de $\hat{\mathbf{b}}$

Proposition 5

$\hat{\mathbf{b}}$ *est un estimateur convergent de* β *:*

$$\hat{\mathbf{b}} \underset{T \to \infty}{\overset{proba}{\longrightarrow}} \beta$$

- Ce résultat nous dit que la probabilité d'observer une différence arbitrairement petite entre l' estimateur des MCO et β tend vers zéro quand la taille de l'échantillon tend vers l'infini.
- La distribution de l'estimateur $\hat{\mathbf{b}}$ se concentre autour de β quand la taille de l'échantillon tend vers l'infini.

Estimateur des Moindres Carrés Ordinaires

Propriétés statistiques de $\hat{\mathbf{b}}$

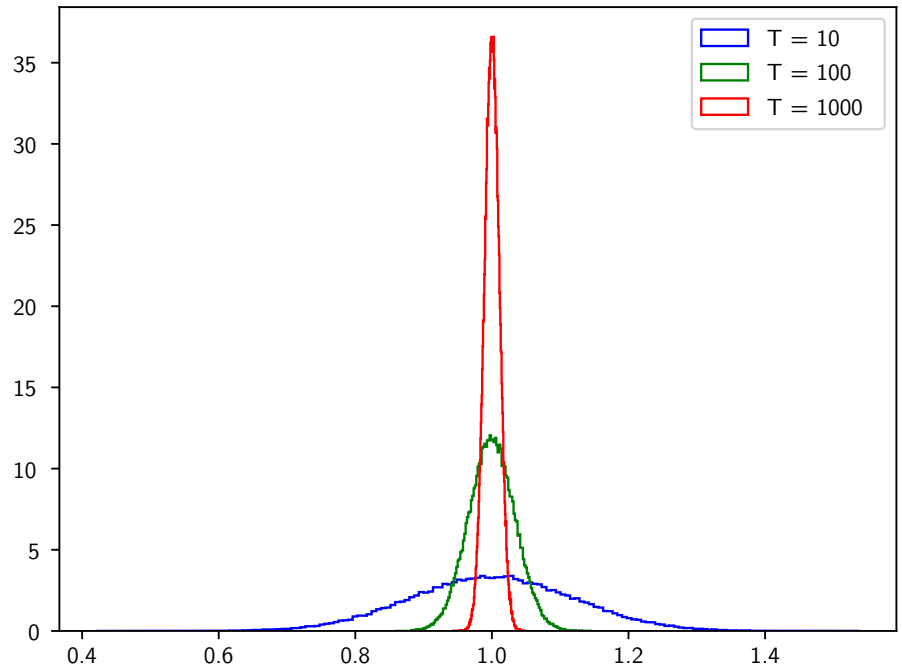


Figure 4: Le modèle de la nature est $y_t = x_t + \varepsilon_t$ avec x_t une variable exogène déterministe prenant des valeurs dans l'intervalle $[0, 10]$ et ε_t un bruit blanc gaussien. On simule 100000 échantillons de 10, 100 ou 1000 observations, pour chaque échantillon on estime le modèle empirique $y_t = b_0 + b_1 x_t + \epsilon_t$. Chaque figure représente la distribution de l'esimateur $\hat{b}_1$. Les codes pour reproduire ces graphiques sont disponibles ici.

Preuve de la proposition 5. Nous savons déjà que l'estimateur est sans biais, pour toute dimension de l'échantillon. Pour établir la convergence en probabilité de l'estimateur vers β, il suffit de montrer que sa variance tend vers 0 quand T tend vers l'infini. Par $\mathscr{H}_2$, nous savons que :

$$\lim_{T\to\infty} \frac{X'X}{T} = Q$$

où Q est une matrice définie positive. Ainsi, nous avons :

$$\begin{aligned}\lim_{T\to\infty} (X'X)^{-1} &= \lim_{T\to\infty} \frac{1}{T}\left(\frac{X'X}{T}\right)^{-1} \\ &= Q^{-1} \lim_{T\to\infty} \frac{1}{T} \\ &= 0\end{aligned}$$

et donc, par définition la variance de l'estimateur : $\lim_{T\to\infty} \mathbb{V}[\hat{\mathbf{b}}] = 0$. C.Q.F.D.

Estimateur des Moindres Carrés Ordinaires

Retour sur les résidus estimés

- On peut exprimer $\hat{\epsilon}$ comme une fonction linéaire de ε :

$$\begin{aligned}\hat{\epsilon} &= \mathbf{y} - X\hat{\mathbf{b}} \\ &= \mathbf{y} - X(X'X)^{-1}X'\mathbf{y} \\ &= X\beta + \varepsilon - X(X'X)^{-1}X'\left(X\beta + \varepsilon\right) \\ &= \left(I - X(X'X)^{-1}X'\right)\varepsilon\end{aligned}$$

- On notera : $P = X(X'X)^{-1}X'$, la matrice de projection orthogonale, et $M = I - P$. On a :

$$\hat{\epsilon} = M\varepsilon$$

- La matrice M est symétrique et idempotente :

$$M' = I - \left(X(X'X)^{-1}X'\right) = I - X(X'X)^{-1}X = M$$

$$M'M = I - 2X(X'X)^{-1}X' + \underbrace{X(X'X)^{-1}X'X(X'X)^{-1}X'}_{X(X'X)^{-1}X'} = M$$

- Les valeurs propres d'une matrice idempotente A sont égales à 1 ou 0. En effet, supposons que λ soit une valeur propre et x le vecteur propre associé. Alors, par définition d'une valeur propre, on doit avoir :

$$\lambda x = Ax$$

Comme A est idempotente, on a $\lambda x = A^2 x$, ou encore :

$$\begin{aligned} \lambda x &= A\lambda x \\ &= \lambda A x \end{aligned}$$

et donc :

$$\lambda x = \lambda^2 x$$

ce qui n'est possible que si λ est égal à un ou zéro.

- Ainsi le rang d'une matrice idempotente est égal à sa trace (la somme des éléments sur la diagonale).

- Nous avons donc :

$$\begin{aligned} \text{rang}(M) &= \text{trace}(I_T) - \text{trace}\left(X(X'X)^{-1}X'\right) \\ &= T - \text{trace}\left(X'X(X'X)^{-1}\right) \\ &= T - \text{trace}\left(I_K\right) \\ &= T - K \end{aligned}$$

Estimateur des Moindres Carrés Ordinaires

Retour sur les résidus estimés

- L'expression des résidus estimés en fonction de ε permet de déduire des propriétés probabilistes sur $\hat{\epsilon}$.
- On montre facilement que $\mathbb{E}[\hat{\epsilon}] = 0$ et $\mathbb{V}[\hat{\epsilon}] = \sigma_\varepsilon^2 M$.
- Même si la matrice de variance covariance de ε est diagonale, les résidus estimés sont corrélés (car la matrice M n'est pas diagonale).
- La normalité de ε implique la normalité de $\hat{\epsilon}$, **mais** la distribution de $\hat{\epsilon}$ est dégénére (sa variance n'est pas de plein rang)... Nous ne sommes pas surpris puisque nous avons déjà montré que la somme des résidus estimés doit être nulle (déterministe).
- On peut aussi exprimer la somme des carrés des résidus estimés comme une fonction de ε :

$$SSE = \hat{\epsilon}'\hat{\epsilon} = \varepsilon' M' M \varepsilon = \varepsilon' M \varepsilon$$

Estimateur des Moindres Carrés Ordinaires

Comment estimer σ_ε^2 ?

Proposition 6

L'espérance de somme des carrés des résidus estimés est donnée par :

$$\mathbb{E}[SSE] = \sigma_\varepsilon^2(T - K)$$

Corollaire 2

Un estimateur non biaisé de σ_ε^2 est :

$$s^2 = \frac{SSE}{T - K}$$

Proposition 7

s^2 est un estimateur convergent de σ_ε^2.

- **Preuve de la proposition 6.** On a :

$$\begin{aligned}\mathbb{E}[SSE] &= \mathbb{E}[\varepsilon' M \varepsilon] \\ &= \mathbb{E}\left[\text{trace}\left(\varepsilon' M \varepsilon\right)\right] \\ &= \mathbb{E}\left[\text{trace}\left(M \varepsilon \varepsilon'\right)\right] \\ &= \text{trace}\left(M \sigma_\varepsilon^2 I_T\right) \\ &= \sigma_\varepsilon^2 \text{trace}\left(M\right) \\ &= \sigma_\varepsilon^2 (T - K)\end{aligned}$$

C.Q.F.D.

En passant, notons que $\frac{SSE}{T}$, qui nous le verrons plus loin est l'estimateur du maximum de vraisemblance, est un estimateur biaisé à distance finie. Asymptotiquement, il n'y a cependant pas de différence entre l'estimateur du maximum de vraisemblance est s^2.

- **Lemme.** On a : $\underset{T\to\infty}{\text{plim}} \frac{\varepsilon' X}{T}$.

Preuve. Puisque X est déterministe, on a directement: $\mathbb{E}\left[\frac{\varepsilon' X}{T}\right] = 0$. Tout assi facilement on montre que la variance est :

$$\mathbb{V}\left[\frac{\varepsilon' X}{T}\right] = \frac{\sigma_\varepsilon^2}{T}\frac{X'X}{T} \underset{T\to\infty}{\longrightarrow} 0 \times Q = 0$$

$\frac{\varepsilon' X}{T}$ converge donc bien en probabilité vers 0. C.Q.F.D.

- **Preuve de la proposition 7.** Nous devons montrer que s^2 converge en probabilité vers σ_ε^2. On a :

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty} s^2 &= \operatorname*{plim}_{T\to\infty} \frac{\varepsilon' M \varepsilon}{T-K} = \operatorname*{plim}_{T\to\infty} \frac{\varepsilon' M \varepsilon}{T} \\
&= \operatorname*{plim}_{T\to\infty} \frac{\varepsilon'\varepsilon}{T} - \operatorname*{plim}_{T\to\infty} \frac{\varepsilon' X (X'X)^{-1} X'\varepsilon}{T} \\
&= \sigma_\varepsilon^2 - \operatorname*{plim}_{T\to\infty} \frac{\varepsilon' X}{T} \left(\frac{X'X}{T}\right)^{-1} \frac{X'\varepsilon}{T} \\
&\sigma_\varepsilon^2 - 0 \times Q^{-1} \times 0 = \sigma_\varepsilon^2
\end{aligned}$$

C.Q.F.D.

Estimateur des Moindres Carrés Ordinaires

Distribution de $\hat{\mathbf{b}}$ et s^2

Théorème 3

L'estimateur $\hat{\mathbf{b}}$ *est normalement distribué :*

$$\hat{\mathbf{b}} \sim \mathcal{N}\left(\beta, \sigma_\varepsilon^2 (X'X)^{-1}\right)$$

Théorème 4

L'estimateur $\frac{(T-K)s^2}{\sigma_\varepsilon^2}$ *est distribué comme un khi-2 :*

$$\frac{(T-K)s^2}{\sigma_\varepsilon^2} \sim \chi^2(T-K)$$

Corollaire 3

La variance de s^2 *est égale à* $\frac{2\sigma_\varepsilon^4}{T-K}$.

- **Preuve du théorème 3.** Nous avons déjà obtenu l'espérance et la variance de $\hat{\mathrm{b}}$. Comme l'estimateur des MCO est linéaire par rapport à $\mathbf{y}$ et donc ε (puisque le modèle de la nature est linéaire) qui suit une loi normale multivariée. L'estimateur des MCO est donc normalement distribué. C.Q.F.D.

- **Lemme.** Soit Q une matrice réelle $T \times T$ symétrique et idempotente, alors :

$$\frac{\varepsilon' Q \varepsilon}{\sigma_\varepsilon^2} \sim \chi^2 \left(\mathrm{trace}(Q)\right)$$

Preuve. Puisque la matrice réelle Q est symétrique, on sait qu'il existe une matrice orthogonale P telle que $P'QP = \Lambda$ est une matrice diagonale. On sait que les valeurs propres de Q sont égales à 1 ou 0. Sans perte de généralité, on suppose que les valeurs propres non nulles viennent en premier le long de la diagonale de Λ :

$$\Lambda = \begin{pmatrix} I_\star & 0 \\ 0 & 0 \end{pmatrix}$$

où la dimension de $I_\star$ est égale à la trace de la matrice Q (la trace d'une matrice est invariante au changement de base). Définissons $\nu = P'\varepsilon$, on a directement $\mathbb{E}[\nu] = 0$ et $\mathbb{V}[\nu] = \sigma_\varepsilon^2 I_T$ puisque P est une matrice orthogonale. Le vecteur ν est normalement distribué : $\nu \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 I_T\right)$. Puisque P est une matrice orthogonale, nous pouvons exprimer ε en fonction de ν :

$$P\nu = PP'\varepsilon \Leftrightarrow \varepsilon = P\nu$$

On a donc :

$$\begin{aligned}\frac{\varepsilon' Q \varepsilon}{\sigma_\varepsilon^2} &= \frac{\nu' P' Q P \nu}{\sigma_\varepsilon^2} \\ &= \frac{1}{\sigma_\varepsilon^2} \nu' \begin{pmatrix} I_\star & 0 \\ 0 & 0 \end{pmatrix} \nu \\ &= \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{\mathrm{tr}(Q)} \nu_i^2 \\ &= \sum_{i=1}^{\mathrm{tr}(Q)} \left(\frac{\nu_i}{\sigma_\varepsilon} \right)^2\end{aligned}$$

Puisque $\frac{\nu_i}{\sigma_\varepsilon} \perp \frac{\nu_j}{\sigma_\varepsilon} \sim \mathcal{N}(0,1)$ pour tout $i \neq j$, on a $\frac{\varepsilon' Q \varepsilon}{\sigma_\varepsilon^2} \sim \chi^2\left(\mathrm{tr}(Q)\right)$.

- **Preuve du théorème 4.** Nous avons :

$$\frac{(T-K)s^2}{\sigma_\varepsilon^2} = \frac{SSE}{\sigma_\varepsilon^2} = \frac{\varepsilon' M \varepsilon}{\sigma_\varepsilon^2}$$

où la matrice $M = I - X(X'X)^{-1}X'$ est une matrice de rang $T-K$. Le lemme précédent entraîne le résultat annoncé. C.Q.F.D.

- **Preuve du corollaire 3.** Direct en notant que la variance d'un χ^2 à $T-K$ degrés de liberté est $2(T-K)$. C.Q.F.D.

Estimateur des Moindres Carrés Ordinaires

Distribution de $\hat{\mathbf{b}}$ et s^2

Proposition 8

Les estimateurs $\hat{\mathrm{b}}$ et s^2 sont des variables aléatoires indépendantes.

- **Lemme** Soit Q une matrice réelle $T \times T$ symétrique et idempotente de rang q. Soit B une matrice réelle $m \times T$ telle que $BQ = 0$. Soit le vecteur aléatoire gaussien $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_T)$. Alors le vecteur aléatoire $B\varepsilon$ et la variable aléatoire $\varepsilon' Q \varepsilon$ sont indépendants.

Preuve. Comme dans la preuve du lemme précédent, on définit la matrice orthogonale P telle que :

$$P'QP = \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix}$$

et le vecteur gaussien $\nu = P'\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 I_T\right)$ (car P est une matrice orthogonale). On partitionne le vecteur ν sous la forme $(\nu_1', \nu_2')'$ où ν_1 et ν_2 sont des vecteurs aléatoires $q \times 1$ et $(T-q) \times 1$ indépendants. On a directement :

$$\varepsilon' Q \varepsilon = \nu_1' \nu_1$$

qui suit un $\chi^2(q)$. Posons $BP = C$ et partitionnons les colonnes de C conformément à la partition de ν:

$$C = \begin{pmatrix} C_1 & C_2 \end{pmatrix}$$

où C_1 est une matrice $m \times q$ et C_2 une matrice $m \times (T-q)$. Nous devons avoir $CP'QP = 0$ puisque $CP'QP = BPP'QP = BQP$ et $BQ = 0$ par hypothèse. Ainsi :

$$\begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} = 0$$

et donc $C_1 = 0$. Pour que la matrice BQ soit nulle, il faut éliminer toutes les

valeurs propres non nulles. En notant que :

$$B\varepsilon = BPP'\varepsilon = C\nu = C_2\nu_2$$

puisque C_1 doit être nul, on voit que $B\varepsilon$ ne dépend que de ν_2 qui est indépendant de ν_1 donc de $\varepsilon' Q\varepsilon$. C.Q.F.D.

- **Preuve de la proposition 8.** On a $s^2 = \frac{\varepsilon' M\varepsilon}{T-K}$ et $\hat{\mathbf{b}} - \beta = (X'X)^{-1}X'\varepsilon$. Puisque $(X'X)^{-1}X'M = (X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X' = 0$, le lemme implique que s^2 et $\hat{\mathbf{b}}$ sont indépendants. C.Q.F.D.

Vraisemblance

- La vraisemblance est la densité de l'échantillon.
- Comme le modèle est linéaire et que ε est normalement distribué, on a directement :

$$L(\beta, \sigma_\varepsilon^2; \mathbf{y}) = (2\pi\sigma_\varepsilon^2)^{-\frac{T}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)}$$

- En pratique, par la suite, on considérera souvent la log-vraisemblance notée $l(\beta, \sigma_\varepsilon^2; \mathbf{y})$:

$$l(\beta, \sigma_\varepsilon^2; \mathbf{y}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}(\mathbf{y} - X\beta)'(\mathbf{y} - X\beta)$$

- $\max\limits_{\{\beta,\sigma_\varepsilon^2\}} l(\beta, \sigma_\varepsilon^2; \mathbf{y}) \Leftrightarrow \max\limits_{\{\beta,\sigma_\varepsilon^2\}} L(\beta, \sigma_\varepsilon^2; \mathbf{y})$

- On peut écrire la vraisemnblance de façon équivalente sous la forme :

$$L(\beta, \sigma_\varepsilon^2; \mathbf{y}) = (2\pi\sigma_\varepsilon^2)^{-\frac{T}{2}} e^{-\frac{SSE + \left(\hat{\mathbf{b}} - \beta\right)' X' X \left(\hat{\mathbf{b}} - \beta\right)}{2\sigma_\varepsilon^2}}$$

Développons le numérateur sous l'exponentielle :

$$\begin{aligned}
(\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) &= \left(\mathbf{y} - X\hat{\mathbf{b}} + X(\hat{\mathbf{b}} - \beta)\right)' \left(\mathbf{y} - X\hat{\mathbf{b}} + X(\hat{\mathbf{b}} - \beta)\right) \\
&= (\mathbf{y} - X\hat{\mathbf{b}})'(\mathbf{y} - X\hat{\mathbf{b}}) + (\hat{\mathbf{b}} - \beta)' X' X (\hat{\mathbf{b}} - \beta) \\
&\qquad + 2(\hat{\mathbf{b}} - \beta) X' (\mathbf{y} - X\hat{\mathrm{b}}) \\
&= SSE + (\hat{\mathbf{b}} - \beta)' X' X (\hat{\mathbf{b}} - \beta)
\end{aligned}$$

où sur la deuxième ligne le dernier terme est nul car les résidus estimés sont orthogonaux aux variables explicatives.

- SSE et $\hat{\mathbf{b}}$ sont des statistiques suffisantes, du point de vue de la vraisemblance et donc de l'inférence, elles résument parfaitement les données.

- Comme l'estimateur des MCO pour β et σ_ε^2 ne dépendent que des statistiques suffisantes et que ces estimateurs sont sans biais, on peut montrer que ces estimateurs sont efficaces ($\nexists$ d'estimateur sans biais avec une variance plus faible).

Vraisemblance

- Les dérivées partielles de la log-vraisemblance :

$$\frac{\partial l(\beta, \sigma_\varepsilon^2; \mathbf{y})}{\partial \beta} = -\frac{1}{\sigma_\varepsilon^2} X'(\mathbf{y} - X\beta)$$

$$\frac{\partial l(\beta, \sigma_\varepsilon^2; \mathbf{y})}{\partial \sigma_\varepsilon^2} = -\frac{T}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4} (\mathbf{y} - X\beta)' (\mathbf{y} - X\beta)$$

- Les espérances des dérivées partielles sont nulles :

$$\mathbb{E}\left[\frac{\partial l(\beta, \sigma_\varepsilon^2; \mathbf{y})}{\partial \beta}\right] = -\frac{1}{\sigma_\varepsilon^2} X'\mathbb{E}[\varepsilon] = 0$$

$$\mathbb{E}\left[\frac{\partial l(\beta, \sigma_\varepsilon^2; \mathbf{y})}{\partial \sigma_\varepsilon^2}\right] = -\frac{T}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4}\mathbb{E}[\varepsilon'\varepsilon] = -\frac{T}{2\sigma_\varepsilon^2} + \frac{T\sigma_\varepsilon^2}{2\sigma_\varepsilon^4} = 0$$

Vraisemblance

Proposition 9

Les bornes inférieures de Cramer-Darmois-Fréchet-Rao pour les variances des estimateurs sans biais de β et σ_ε^2 sont $\sigma_\varepsilon^2(X'X)^{-1}$ et $\frac{2\sigma_\varepsilon^4}{T}$.

- La variance de $\hat{\mathbf{b}}$ est sur la borne CDFR, il s'agit donc bien d'un estimateur efficace (sans biais et de variance minimale).
- La variance de s^2 est strictement supérieure à la borne CDFR. Il s'agit tout de même d'un estimateur efficace car il ne dépend que d'une statistique suffisante (SSR, par le théorème de Lehmann-Scheffé).

- **Preuve de la proposition 9.** Calculons la matrice d'information de Fisher. Les dérivées secondes de la log-vraisemblance sont :

$$\frac{\partial^2 l(\beta, \sigma_\varepsilon^2; \mathbf{y})}{\partial \beta \partial \beta'} = -\frac{1}{\sigma_\varepsilon^2} X'X$$

$$\frac{\partial^2 l(\beta, \sigma_\varepsilon^2; \mathbf{y})}{\partial \sigma_\varepsilon^4} = \frac{T}{2\sigma_\varepsilon^4} - \frac{1}{\sigma_\varepsilon^6} (\mathbf{y} - X\beta)' (\mathbf{y} - X\beta)$$

$$\frac{\partial^2 l(\beta, \sigma_\varepsilon^2; \mathbf{y})}{\partial \beta \partial \sigma_\varepsilon^2} = -\frac{1}{\sigma_\varepsilon^4} X' (\mathbf{y} - X\beta)$$

La matrice d'information de Fisher est l'opposé de l'espérance de la matrice hessienne (des dérivées secondes) :

$$I = \begin{pmatrix} \frac{X'X}{\sigma_\varepsilon^2} & 0 \\ 0 & \frac{T}{2\sigma_\varepsilon^4} \end{pmatrix}$$

La borne CDFR est définie par la diagonale de l'inverse de la matrice d'information de Fisher. C.Q.F.D.

Vraisemblance

Proposition 10

Les estimateurs du maximum de vraisemblance de β et σ_ε^2 sont :

$$\hat{\beta} = (X'X)^{-1}X'\mathbf{y} \quad \textit{et} \quad \hat{\sigma}_\varepsilon^2 = \frac{SSE}{T}$$

- L'estimateur du MV de β est identique à l'estimateur des MCO $\hat{\mathbf{b}}$.
- L'estimateur du MV de σ_ε^2 est biaisé !
- La variance de l'estimateur du MV de σ_ε^2 est plus faible que la variance de s^2 :

$$\mathbb{V}\left[\hat{\sigma}_\varepsilon^2\right] = 2\sigma_\varepsilon^4 \frac{T-K}{T^2} < \underbrace{\frac{2\sigma_\varepsilon^4}{T}}_{\text{CDFR}} < \frac{2\sigma_\varepsilon^4}{T-K}$$

- **Mais** l'erreur quadratique moyenne de l'estimateur du MV est plus faible que celle de s^2.

- **Preuve de la proposition 10** Direct en annulant les dérivées partielles par rapport à β et σ_ε^2. Pour $\hat{\beta}$ on peut alternativement raisonner sur la représentation alternative (en fonction de SSE et $\hat{\mathbf{b}}$) de la fonction de vraisemblance. C.Q.F.D.

- **Calcul de la variance de $\hat{\sigma}_\varepsilon^2$.** Nous avons déjà montré, voir le corollaire 3, que :

$$\mathbb{V}\left[\frac{SSE}{T-K}\right] = \frac{2\sigma_\varepsilon^4}{T-K}$$

$$\Leftrightarrow \mathbb{V}\left[\frac{SSE}{T}\frac{T}{T-K}\right] = \frac{2\sigma_\varepsilon^4}{T-K}$$

$$\Leftrightarrow \mathbb{V}\left[\frac{SSE}{T}\right] = \frac{2\sigma_\varepsilon^4}{T-K}\frac{(T-K)^2}{T^2}$$

$$\Leftrightarrow \mathbb{V}\left[\hat{\sigma}_\varepsilon^2\right] = 2\sigma_\varepsilon^4\frac{T-K}{T^2}$$

C.Q.F.D.

- **Remarque 1.** La variance de l'estimateur du maximum de vraisemblance de σ_ε^2 est strictement inférieure à la borne de CDFR. Possible car cette borne inférieure ne concerne que les estimateurs sans biais (l'estimateur du MV est biaisé).

- L'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ est défini par :

$$MSE(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right]$$

- Si $\hat{\theta}$ est un estimateur sans biais de θ alors l'erreur quadratique moyenne est égale à la variance.

- Autrement, en notant $B(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta$ le biais de l'estimateur, on a :

$$\begin{aligned} MSE\left(\hat{\theta}\right) &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] + B\left(\hat{\theta}\right)\right)^2\right] \\ &= \mathbb{V}\left[\hat{\theta}\right] + B\left(\hat{\theta}\right)^2 + 2B\left(\hat{\theta}\right)\mathbb{E}\left[\hat{\theta} - E\left[\hat{\theta}\right]\right] \\ &= \mathbb{V}\left[\hat{\theta}\right] + B\left(\hat{\theta}\right)^2 \end{aligned}$$

- Arbitrage Biais-Variance. Au sens de la réduction de l'erreur quadratique moyenne, un estimateur peut compenser son biais par une plus faible variance (une plus grande précision)

- Dans le cas qui nous intéresse, on a :

$$MSE(s^2) = \mathbb{V}\left[s^2\right] = \sigma_\varepsilon^4 \frac{2}{T-K}$$

et

$$MSE(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^4 \frac{2(T-K)+K^2}{T^2}$$

Définissons :

$$r(s^2, \hat{\sigma}_\varepsilon^2) = \frac{MSE(s^2)}{MSE(\hat{\sigma}_\varepsilon^2)}$$

$$\Leftrightarrow r(s^2, \hat{\sigma}_\varepsilon^2) = \frac{2}{T-K}\frac{T^2}{2(T-K)+K^2}$$

On a :

$$\begin{aligned} r(s^2, \hat{\sigma}_\varepsilon^2) > 1 &\Leftrightarrow 2T^2 > (T-K)\left[2(T-K)+K^2\right] \\ &\Leftrightarrow 2T^2 > 2T^2 + 2K^2 - 4TK + K^2 \\ &\Leftrightarrow K(3K-4T) < 0 \\ &\Leftrightarrow T > \frac{3}{4}K \end{aligned}$$

La dernière inégalité est nécessairement vraie, autrement $X'X$ ne serait pas de plein rang (le nombre d'observations doit être supérieur aux nombre de variables explicatives). Le ratio $r(s^2, \hat{\sigma}_\varepsilon^2)$ doit donc être supérieur à 1, et l'erreur quadratique moyenne de l'estimateur du MV inférieure à l'erreur quadratique moyenne de s^2.

- **Remarque 2.** Même si en moyenne $\hat{\sigma}_\varepsilon^2$ est différent de σ_ε^2, cet estimateur est en moyenne moins éloigné de σ_ε^2 que s^2 (qui pourtant est égal à σ_ε^2 en moyenne).

Distribution asymptotique des estimateurs

Proposition 11

L'estimateur $\hat{\mathbf{b}}$ *est asymptotiquement normalement distribué :*

$$\sqrt{T}\left(\hat{\mathbf{b}} - \beta\right) \underset{T\to\infty}{\Longrightarrow} \mathcal{N}\left(0, \sigma_\varepsilon^2 Q^{-1}\right)$$

où $Q = \lim_{T\to\infty} \frac{X'X}{T}$.

Proposition 12

L'estimateur s^2 *est asymptotiquement normalement distribué :*

$$\sqrt{T}\left(s^2 - \sigma_\varepsilon^2\right) \underset{T\to\infty}{\Longrightarrow} \mathcal{N}\left(0, 2\sigma_\varepsilon^4\right)$$

- **Preuve de la proposition 11.** Nous avons déjà montré, voir le théorème 4, que $\hat{\mathbf{b}}$ est normalement distribué pour tout T. Nous avons :

$$\hat{\mathbf{b}} \sim \mathcal{N}\left(\beta, \sigma_\varepsilon^2 (X'X)^{-1}\right)$$

$$\Leftrightarrow \hat{\mathbf{b}} - \beta \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 (X'X)^{-1}\right)$$

$$\Leftrightarrow \sqrt{T}\left(\hat{\mathbf{b}} - \beta\right) \sim \mathcal{N}\left(0, T\sigma_\varepsilon^2 (X'X)^{-1}\right)$$

$$\Leftrightarrow \sqrt{T}\left(\hat{\mathbf{b}} - \beta\right) \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \left(\frac{X'X}{T}\right)^{-1}\right)$$

On obtient le résultat asymptotique en rapellant que, selon $\mathscr{H}_2$, $\frac{X'X}{T}$ converge vers Q lorsque T tend vers l'infini. C.Q.F.D.

- **Preuve de la proposition 12.** Une variable aléatoire du $\chi^2(n)$ est une somme de n variables aléatoires indépendantes (carrés de variables aléatoires normales centrées-réduites). Le théorème de la limite centrale nous dit que la variable du χ^2 doit être vers une variable aléatoire normale quand n tend vers l'infini.

  D'après le théorème 4, on a :

$$\frac{(T-K)s^2}{\sigma_\varepsilon^2} = \sum_{i=1}^{T-K} v_i^2$$

  où $(v_i, i = 1, \ldots, T-K)$ sont des variables aléatoires gaussiennes centrées réduites indépendantes, et donc $(v_i^2, i = 1, \ldots, T-K)$ sont des $\chi^2(1)$ indépendantes. Sachant que $\mathbb{E}\left[v_i^2\right] = 1$ et $\mathbb{V}\left[v_i^2\right] = 2$, le théorème de la limite

centrale nous dit que :

$$\frac{1}{\sqrt{T-K}} \sum_{i=1}^{T-K} \frac{v_i^2 - 1}{\sqrt{2}} \underset{T\to\infty}{\Longrightarrow} \mathcal{N}(0,1)$$

$$\Leftrightarrow \frac{1}{\sqrt{T-K}} \sum_{i=1}^{T-K} \left(v_i^2 - 1\right) \underset{T\to\infty}{\Longrightarrow} \mathcal{N}(0,2)$$

$$\Leftrightarrow \frac{1}{\sqrt{T-K}} \left(\sum_{i=1}^{T-K} v_i^2 - (T-K) \right) \underset{T\to\infty}{\Longrightarrow} \mathcal{N}(0,2)$$

$$\Leftrightarrow \frac{1}{\sqrt{T-K}} \left(\frac{(T-K)s^2}{\sigma_\varepsilon^2} - (T-K) \right) \underset{T\to\infty}{\Longrightarrow} \mathcal{N}(0,2)$$

$$\Leftrightarrow \sqrt{T-K} \left(s^2 - \sigma_\varepsilon^2\right) \underset{T\to\infty}{\Longrightarrow} \mathcal{N}\left(0, 2\sigma_\varepsilon^4\right)$$

$$\Rightarrow \sqrt{T} \left(s^2 - \sigma_\varepsilon^2\right) \underset{T\to\infty}{\Longrightarrow} \mathcal{N}\left(0, 2\sigma_\varepsilon^4\right)$$

C.Q.F.D.

Tests d'hypothèses

Restriction sur les paramètres

Théorème 5

Soient R un vecteur réel $1 \times K$ déterministe et r un scalaire réel déterministe. Sous l'hypothèse nulle $R\beta = r$, on a :

$$\frac{R\hat{\mathrm{b}} - r}{\sqrt{s^2 R(X'X)^{-1}R}} \sim t_{T-K}$$

où t_{T-K} est une loi de Student à $T - K$ degrés de liberté.

Exemple 6

Soit le modèle $y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_t$. Pour tester l'hypothèse $\beta_1 + \beta_2 = 1$, on pose $R = (1, 1)$ et $r = 1$. Notons que pour estimer le modèle contraint nous pouvons le réécrire en substituant la contrainte :

$$y_t - x_{2,t} = \beta_0 + \beta_1 (x_{1,t} - x_{2,t}) + \varepsilon_t$$

- **Preuve du théorème 5.** Le numérateur $R\hat{\mathbf{b}} - r$ est une variable aléatoire centrée sous l'hypothèse nulle :

$$\mathbb{E}\left[R\hat{\mathbf{b}} - r\right] = R\mathbb{E}\left[\hat{\mathbf{b}}\right] - r = R\beta - r = 0$$

Sa variance est :

$$\mathbb{V}\left[R\hat{\mathbf{b}} - r\right] = R\mathbb{V}\left[\hat{\mathbf{b}}\right]R' = \sigma_\varepsilon^2 R(X'X)^{-1}R'$$

Ainsi, sous l'hypothèse nulle, on a :

$$\frac{R\hat{\mathbf{b}} - r}{\sigma_\varepsilon^2 R(X'X)^{-1}R'} \sim \mathcal{N}(0,1)$$

Mais comme nous ne connaissons pas σ_ε^2, on remplace cette variance par s^2, un estimateur sans biais. Or nous savons, par le théorème 4, que :

$$\frac{(T-K)s^2}{\sigma_\varepsilon^2} \sim \chi^2(T-K)$$

Ainsi :

$$\frac{s^2 R(X'X)^{-1}R'}{\sigma_\varepsilon^2 R(X'X)^{-1}R'} = \frac{s^2}{\sigma_\varepsilon^2} \sim \frac{\chi^2(T-K)}{T-K}$$

et donc :

$$\frac{R\hat{\mathbf{b}} - r}{s^2 R(X'X)^{-1}R'} = \frac{\frac{R(\hat{\mathbf{b}}-\beta)}{\sqrt{\sigma_\varepsilon^2 R(X'X)^{-1}R'}}}{\sqrt{\frac{s^2 R(X'X)^{-1}R'}{\sigma_\varepsilon^2 R(X'X)^{-1}R'}}}$$

est le ratio d'une normale centrée réduite et d'un χ^2 à $T-K$ degrés de liberté rapporté à $T-K$. Puisque le numérateur et le dénominateur sont des variables aléaoires indépendantes, voir la proposition 8, le ratio suit une loi de student à $T-K$ degrés de liberté, voir la proposition 16 dans l'annexe A. C.Q.F.D.

Corollaire 4

Soit $s^2_{\hat{\mathrm{b}}_i} = s^2 \left[(X'X)^{-1}\right]_{ii}$ *la variance de l'estimateur des MCO du i-ème paramètre. La statistique :*

$$t = \frac{\hat{\mathbf{b}}_i}{s_{\hat{\mathrm{b}}_i}}$$

suit une loi de Student à $T - K$ *degrés de liberté sous l'hypothèse nulle* $\beta_i = 0$.

- Conséquence directe du théorème 5 avec $r = 0$ et R un vecteur de sélection du i-ème élément de β (un vecteur ligne nul à l'exception du i-ème élément égal à 1).

Théorème 7

Soit R une matrice réelle déterministe $m \times K$ de rang m. Soit r un vecteur réel $m \times 1$ déterministe. Alors sous l'hypothèse nulle $R\beta = r$, la statistique :

$$\frac{\left(r - R\hat{\mathbf{b}}\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(r - R\hat{\mathbf{b}}\right)/m}{SSE/(T-K)}$$

suit une loi de Fisher $F(m, T-K)$.

- Ici on teste simultanément m restrictions.
- Ces restrictions doivent être différentes, les m lignes de R sont linéairement indépendantes.
- Théorème 7 $\Longleftrightarrow$ $m \times$ Théorème 5

Tests d'hypothèses

Restrictions sur les paramètres

- $r - R\hat{\mathbf{b}}$ est généralement différent de zéro, même si $R\beta = r$, à cause de l'incertitude liée à l'estimateur des MCO.
- Le numérateur de la statistique de Fisher est une mesure de la distance aux m contraintes.
- Le dénominateur mesure la taille des perturbations ε.
- On rejette les m contraintes si la distance aux m contraintes est trop importante par rapport à la taille des perturbations.
- Pourquoi ne pas rajouter les contraintes $R\beta = r$ à la procédure d'estimation par les MCO ou par MV ? Cela permettrait de réduire la variance de l'estimateur (moins de paramètres à estimer).

Preuve du théorème 7. Sous l'hypothèse nulle, $r = R\beta$, on a :

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

Ainsi, la variance de $r - R\hat{\mathbf{b}}$ est :

$$\mathbb{V}\left[r - R\hat{\mathbf{b}}\right] = \sigma_\varepsilon^2 R(X'X)^{-1}X'X(X'X)^{-1}R' = \sigma_\varepsilon^2 R(X'X)^{-1}R'$$

dont nous retrouvons l'inverse au centre du numérateur, au facteur d'échelle σ_ε^2 près, en notant que cette matrice (comme son inverse) est symétrique et donc identique à sa transposée. Si $r - R\hat{\mathbf{b}}$ mesure l'écart aux m restrictions, la statistique donne relativement plus de poids aux restrictions qui sont mesurées de façon plus précises. Nous pouvons donc réécrire le numérateur sous la forme $\varepsilon' Q \varepsilon$ avec :

$$Q = X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X'$$

Clairement cette matrice est symétrique :

$$\begin{aligned}
Q' &= \left(X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X'\right)' \\
&= \left(R(X'X)^{-1}X'\right)'\left(R(X'X)^{-1}R'\right)^{-1}\left(X(X'X)^{-1}R'\right)' \\
&= X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X' \\
&= Q
\end{aligned}$$

elle est aussi idempotente :

$$\begin{aligned}
QQ &= X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X'X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X' \\
&= X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X' \\
&= X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X' = Q
\end{aligned}$$

Nous pouvons montrer que la trace de la matrice Q est égale à m (son rang) :

$$\begin{aligned}
\operatorname{tr}(Q) &= \operatorname{tr}\left(X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X'\right) \\
&= \operatorname{tr}\left(\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}X'X(X'X)^{-1}R'\right) \\
&= \operatorname{tr}\left(\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}R'\right) \\
&= \operatorname{tr}\left(I_m\right) = m
\end{aligned}$$

Nous savons donc que le numérateur est distribué comme $\sigma_\varepsilon^2\left(\chi^{2(m)}/m\right)$, voir la preuve du théorème 4 (le lemme utilisé dans la preuve). Par ailleurs, nous savons aussi que le dénominateur peut s'écrire comme $\varepsilon' M\varepsilon/(T-K)$ avec $M = I - X(X'X)^{-1}X'$ une matrice symétrique, idempotente et vérifiant $\operatorname{tr}(M) = T - K$. Le dénominateur est donc distribué comme $\sigma_\varepsilon^2\left(\chi^{2(T-K)}/(T-K)\right)$.

Comme le numérateur et le dénominateur sont deux variables aléatoires indépendantes, puisque le numérateur dépend de $\hat{\mathrm{b}}$ et le dénominateur de s^2 (voir la proposition 8, $\hat{\mathrm{b}}$ et s^2 sont indépendants), on sait par définition de la loi de Fisher (voir l'annexe A) que la statistique donnée dans le théorème, dite de Fisher, suit une loi de Fisher $F(m, T-K)$. C.Q.F.D.

MCO avec contraintes linéaires

- Le programme d'optimisation est maintenant :

$$\tilde{\mathbf{b}} = \arg\min_{\mathbf{b}} (\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b})$$
$$\underline{\text{s.c.}} \quad R\mathbf{b} = r$$

- Le lagrangien associé à ce programme est :

$$\mathscr{L} = (\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b}) + \boldsymbol{\lambda}' (R\mathbf{b} - r)$$

où $\boldsymbol{\lambda}$ est un vecteur $m \times 1$ de multiplicateurs de Lagrange.

- Les CNO sont :

$$\begin{cases} 0 &= -2X'(\mathbf{y} - X\tilde{\mathbf{b}}) + R'\boldsymbol{\lambda} \\ 0 &= R\tilde{\mathbf{b}} - r \end{cases}$$

MCO avec contraintes linéaires

- On peut réécrire les CNO matriciellement sous la forme :

$$\underbrace{\begin{pmatrix} 2X'X & R' \\ R & O \end{pmatrix}}_{\mathbb{X}} \underbrace{\begin{pmatrix} \tilde{\mathbf{b}} \\ \boldsymbol{\lambda} \end{pmatrix}}_{\mathbf{c}} = \underbrace{\begin{pmatrix} 2X'\mathbf{y} \\ r \end{pmatrix}}_{\mathbf{d}}$$

- Comme, par hypothèse, $X'X$ est une matrice $K \times K$ de plein rang et R une matrice $m \times K$ de rang m, la matrice $\mathbb{X}$ est nécessairement invesible. On a donc :

$$\begin{pmatrix} \tilde{\mathbf{b}} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} 2X'X & R' \\ R & O \end{pmatrix}^{-1} \begin{pmatrix} 2X'\mathbf{y} \\ r \end{pmatrix}$$

- Pour obtenir l'estimateur des MCO contraints $\tilde{\mathbf{b}}$ il faut inverser la matrice par blocs $\mathbb{X}$.

MCO avec contraintes linéaires

Proposition 13 (Inversion d'une matrice partitionnée)

Soient A_{11} et A_{22} des matrices réelles $n_1 \times n_1$ et $n_2 \times n_2$. On suppose que A_{11} est inversible. Soient A_{12} une matrice $n_1 \times n_2$ et A_{21} une matrice $n_2 \times n_1$. Si la matrice A de dimension $n \times n$, avec $n = n_1 + n_2$, définie par :

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

est inversible, alors on a :

$$A^{-1} = \begin{pmatrix} A_{11}^{-1}\left(I + A_{12}BA_{21}A_{11}^{-1}\right) & -A_{11}^{-1}A_{12}B \\ -BA_{21}A_{11}^{-1} & B \end{pmatrix}$$

avec $B = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}$.

- **Preuve de la proposition 13.** Il suffit de montrer que $A^{-1}A = AA^{-1}$ est une matrice identité. On montre que A^1A est une matrice identité (le reste est laissé au lecteur). Il suffit de montrer que les blocs $\left[A^{-1}A\right]_{11}$ et $\left[A^{-1}A\right]_{22}$ sont des matrices identité et que les blocs $\left[A^{-1}A\right]_{12}$ et $\left[A^{-1}A\right]_{21}$ sont nuls. On a :

$$
\begin{aligned}
\left[A^{-1}A\right]_{11} &= A_{11}^{-1}\left(I + A_{12}BA_{21}A_{11}^{-1}\right)A_{11} - A_{11}A_{12}BA_{21} \\
&= I + A_{11}^{-1}A_{12}BA_{21} - A_{11}^{-1}A_{12}BA_{21} \\
&= I \\
\left[A^{-1}A\right]_{22} &= BA_{21}A_{11}^{-1}A_{12} + BA_{22} \\
&= B\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \\
&= I \quad \text{par définition de } B \\
\left[A^{-1}A\right]_{12} &= A_{11}^{-1}\left(I + A_{12}BA_{21}A_{11}^{-1}\right)A_{12} - A_{11}^{-1}A_{12}BA_{22} \\
&= A_{11}^{-1}\left(A_{12} + A_{12}BA_{21}A_{11}^{-1}A_{12} - A_{12}BA_{22}\right) \\
&= A_{11}^{-1}\left(A_{12} - A_{12}B\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)\right) \\
&= A_{11}^{-1}\left(A_{12} - A_{12}\right) = 0 \\
\left[A^{-1}A\right]_{21} &= -BA_{21}A_{11}^{-1}A_{11} + BA_{21} \\
&= B\left(A_{21} - A_{21}\right) = 0
\end{aligned}
$$

C.Q.F.D.

MCO avec contraintes linéaires

Théorème 8

L'estimateur des MCO contraints est donné par :

$$\tilde{\mathbf{b}} = \hat{\mathbf{b}} - (X'X)^{-1}R' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\mathbf{b}} - r\right)$$

Sa variance est :

$$\mathbb{V}\left[\tilde{\mathbf{b}}\right] = \sigma_\varepsilon^2(X'X)^{-1} - \sigma_\varepsilon^2(X'X)^{-1}R' \left(R(X'X)^{-1}R'\right)^{-1} R(X'X)^{-1}$$

- La variance de $\tilde{\mathbf{b}}$ est plus petite que celle de $\hat{\mathbf{b}}$.
- L'estimateur est sans biais ssi la contrainte $R\beta = r$ est vraie.

- **Preuve du théorème 8.** En utilisant la formule d'inversion par bloc, voir la proposition 13, on obtient :

$$\mathbb{X}^{-1} = \begin{pmatrix} \frac{1}{2}(X'X)^{-1}\left[I - R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}\right] & (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1} \\ \left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1} & -2\left(R(X'X)^{-1}R'\right)^{-1} \end{pmatrix}$$

Le produit scalaire du premier bloc de ligne de $\mathbb{X}^{-1}$ avec le vecteur $\mathbf{d}$ donne l'estimateur $\tilde{\mathbf{b}}$:

$$\begin{aligned} \tilde{\mathbf{b}} &= \frac{1}{2}(X'X)^{-1}\left[I - R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}\right]2X'\mathbf{y} + (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}r \\ &= \hat{\mathbf{b}} - (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R\hat{\mathbf{b}} + (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}r \\ &= \hat{\mathbf{b}} - (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\mathbf{b}} - r\right) \end{aligned}$$

Pour calculer la variance de $\tilde{\mathbf{b}}$, il suffit de rappeler que si A et B sont deux variables aléatoires alors :

$$\mathbb{V}[A - B] = \mathbb{V}[A] + \mathbb{V}[B] - 2\text{cov}(A, B)$$

On connaît déjà la variance de $\hat{\mathbf{b}}$, la variance du second terme est :

$$\begin{aligned} \mathbb{V}\left[(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\mathbf{b}} - r\right)\right] &= \mathbb{V}\left[(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R\hat{\mathbf{b}}\right] \\ &= (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R\sigma_\varepsilon^2(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1} \\ &= \sigma_\varepsilon^2(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1} \end{aligned}$$

Enfin, on a :

$$\begin{aligned}\operatorname{cov}\left(\hat{\mathbf{b}}, (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R\hat{\mathbf{b}}\right) &= (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R\mathbb{V}\left[\hat{\mathbf{b}}\right] \\ &= \sigma_\varepsilon^2(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}R(X'X)^{-1}\end{aligned}$$

L'expression de la variance de $\tilde{\mathbf{b}}$ s'obtient directement en sommant les variances de $\hat{\mathbf{b}}$ et $(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\mathbf{b}}-r\right)$, puis en retranchant deux fois la covariance. C.Q.F.D.

MV avec contraintes linéaires

Théorème 9

L'estimateur du maximum de vraisemblance de β sous les contraintes $R\beta = r$ est donné par :

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

L'estimateur du maximum de vraisemblance de σ_ε^2 est :

$$\tilde{\sigma}_\varepsilon^2 = \frac{1}{T}\left(\mathbf{y} - X\hat{\beta}^c\right)' \left(\mathbf{y} - X\hat{\beta}^c\right)$$

- Comme dans le cas sans contrainte, l'estimateur du MV de β est identique à l'estimateur des MCO.

Preuve du théorème 9. La vraisemblance est donnée par :

$$l\left(\beta, \sigma_\varepsilon^2; \mathbf{y}\right) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\left(\mathbf{y} - X\beta\right)'\left(\mathbf{y} - X\beta\right)$$

puisque les contraintes ne concernent pas la variance de ε, on obtient l'estimateur du MV de σ_ε^2 en annulant la dérivée partielle de la log vraissemblance par rapport à σ_ε^2 :

$$\frac{\partial l\left(\beta, \sigma_\varepsilon^2; \mathbf{y}\right)}{\partial \sigma_\varepsilon^2} = -\frac{T}{2\tilde{\sigma}_\varepsilon^2} + \frac{1}{2\tilde{\sigma}_\varepsilon^4}\left(\mathbf{y} - X\beta\right)'\left(\mathbf{y} - X\beta\right) = 0$$

$$\Leftrightarrow \tilde{\sigma}_\varepsilon^2(\beta) = \frac{1}{T}\left(\mathbf{y} - X\beta\right)'\left(\mathbf{y} - X\beta\right)$$

En substituant l'estimateur de σ_ε^2 dans l'expression de la log vraisemblance, on obtient la log vraisemblance concentrée :

$$l\left(\beta; \mathbf{y}\right) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\left[\left(\mathbf{y} - X\beta\right)'\left(\mathbf{y} - X\beta\right)\right] - \frac{T}{2}$$

que nous devons maximiser, par rapport β, sous la contrainte $R\beta = r$ pour obtenir l'estimateur $\hat{\beta}^c$. Clairement cela revient à minimiser $\left(\mathbf{y} - X\beta\right)'\left(\mathbf{y} - X\beta\right)$ sous la contrainte $R\beta = r$. Posons le lagrangien associé à ce programme d'optimisation :

$$\mathscr{L} = \mathbf{y}'\mathbf{y} - 2\beta' X'\mathbf{y} + \beta' X' X\beta + \lambda'\left(R\beta - r\right)$$

où λ est un vecteur $m \times 1$ de multiplicateurs de Lagrange. Les conditions nécessaires d'optimalité sont :

$$\begin{cases} -2X'\mathbf{y} + 2X'X\hat{\beta}^c + R'\hat{\lambda} = 0 \\ R\hat{\beta}^c - r = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} -2R(X'X)^{-1}X'\mathbf{y} + 2R\hat{\beta}^c + R(X'X)^{-1}R'\hat{\lambda} = 0 \\ R\hat{\beta}^c - r = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} -2R\hat{\mathbf{b}} + 2R\hat{\beta}^c + R(X'X)^{-1}R'\hat{\lambda} = 0 \\ R\hat{\beta}^c - r = 0 \end{cases}$$

De la première équation nous déduisons $\hat{\lambda}$:

$$\hat{\lambda} = \left(R(X'X)^{-1}R'\right)^{-1}\left(2R\hat{\mathbf{b}} - 2R\hat{\beta}^c\right)$$

$$\Leftrightarrow \hat{\lambda} = -2\left(R(X'X)^{-1}R'\right)^{-1}\left(r - R\hat{\mathbf{b}}\right)$$

en substituant la seconde équation. Le multiplicateur de Lagrange est non nul si et seulement si l'estimateur des MCO ne satisfait les m contraintes linéaires. En substituant $\hat{\lambda}$ dans la première équation et en résolvant pour $\hat{\beta}^c$ on obtient :

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

C.Q.F.D.

Tests du ratio de vraisemblance

Théorème 10

La statistique de Fisher définie dans le théorème 7 s'interprête comme un test de ratio de vraisemblance de $R\beta = r$ contre $R\beta \neq r$.

Preuve du théorème 10. La fonction de vraisemblance est :

$$L\left(\beta, \sigma_\varepsilon^2; \mathbf{y}\right) = (2\pi)^{-\frac{T}{2}} \left(\sigma_\varepsilon^2\right)^{-\frac{T}{2}} \exp\left\{-\frac{1}{\sigma_\varepsilon^2}\left(\mathbf{y} - X\beta\right)'\left(\mathbf{y} - X\beta\right)\right\}$$

Sous l'hypothèse nulle, $R\beta = r$, la valeur du maximum de vraisemblance est :

$$(2\pi)^{-\frac{T}{2}} \left[\frac{1}{T}\left(\mathbf{y} - X\hat{\beta}^c\right)'\left(\mathbf{y} - X\hat{\beta}^c\right)\right]^{-\frac{T}{2}} \exp\left\{-\frac{T}{2}\right\}$$

Sous l'hypothèse alternative, $R\beta \neq r$, la valeur du maximum de vraisemblance est :

$$(2\pi)^{-\frac{T}{2}} \left[\frac{1}{T}\left(\mathbf{y} - X\hat{\beta}\right)'\left(\mathbf{y} - X\hat{\beta}\right)\right]^{-\frac{T}{2}} \exp\left\{-\frac{T}{2}\right\}$$

La statistique du ratio de vraisemblance est donc :

$$\frac{(2\pi)^{-\frac{T}{2}} \left[\frac{1}{T}\left(\mathbf{y} - X\hat{\beta}\right)'\left(\mathbf{y} - X\hat{\beta}\right)\right]^{-\frac{T}{2}} \exp\left\{-\frac{T}{2}\right\}}{(2\pi)^{-\frac{T}{2}} \left[\frac{1}{T}\left(\mathbf{y} - X\hat{\beta}^c\right)'\left(\mathbf{y} - X\hat{\beta}^c\right)\right]^{-\frac{T}{2}} \exp\left\{-\frac{T}{2}\right\}}$$

Le test du ratio de vraisemblance nous amènera à rejeter l'hypothèse nulle si la statistique est grande (les contraintes sont peu vraisemblables), c'est-à-dire si le ratio :

$$\frac{\left(\mathbf{y} - X\hat{\beta}^c\right)'\left(\mathbf{y} - X\hat{\beta}^c\right)}{\left(\mathbf{y} - X\hat{\beta}\right)'\left(\mathbf{y} - X\hat{\beta}\right)}$$

est grand (la somme des carrés des résidus du modèle contraint est grande par rapport à la somme des carrés des résidus non contraints).

Par ailleurs, on a :

$$\begin{aligned}
\left(\mathbf{y} - X\hat{\beta}^c\right)' \left(\mathbf{y} - X\hat{\beta}^c\right) &= \mathbf{y}'\mathbf{y} - 2\ \hat{\beta}^{c\,\prime}\, X'\mathbf{y} + \hat{\beta}^{c\,\prime}\, X'X\hat{\beta}^c \\
&= \mathbf{y}'\mathbf{y} - 2\hat{\beta}' X'\mathbf{y} + \hat{\beta}'\, X'X\hat{\beta} \\
&\quad + 2\left(\hat{\beta} - \hat{\beta}^c\right)' X'\mathbf{y} - \hat{\beta}'\, X'X\hat{\beta} + \hat{\beta}^{c\,\prime}\, X'X\hat{\beta}^c \\
&= \mathbf{y}'\mathbf{y} - 2\hat{\beta}' X'\mathbf{y} + \hat{\beta}'\, X'X\hat{\beta} \\
&\quad + 2\left(\hat{\beta} - \hat{\beta}^c\right)' X'X\hat{\beta} - \hat{\beta}'\, X'X\hat{\beta} + \hat{\beta}^{c\,\prime}\, X'X\hat{\beta}^c \\
&= \left(\mathbf{y} - X\hat{\beta}\right)' \left(\mathbf{y} - X\hat{\beta}\right) + \left(\hat{\beta} - \hat{\beta}^c\right)' X'X \left(\hat{\beta} - \hat{\beta}^c\right)
\end{aligned}$$

Le test du ratio de vraisemblance nous amènera donc à rejeter l'hypothèse nulle si le ratio :

$$\frac{\left(\hat{\beta} - \hat{\beta}^c\right)' X'X \left(\hat{\beta} - \hat{\beta}^c\right)}{\left(\mathbf{y} - X\hat{\beta}\right)' \left(\mathbf{y} - X\hat{\beta}\right)}$$

est grand. Finalement, en utilisant la définition de l'estimateur contraint, on a :

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

en substituant dans le numérateur on voit que le test du ratio de vraisemblance rejettera

la nulle lorsque le ratio :

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

au facteur $T-K/m$ près, on reconnaît la statistique donnée dans le théorème 7. C.Q.F.D.

Test joint sur les paramètres

Corollaire 5

Sous l'hypothèse nulle $\beta = \beta^\star$ la statistique :

$$\frac{\left(\hat{\mathbf{b}} - \beta^\star\right)' X'X \left(\hat{\mathbf{b}} - \beta^\star\right)}{SSE} \frac{T-K}{K}$$

suit une loi de Fisher de degrés de liberté K et $T-K$.

Preuve du corollaire 5 Direct par le théorème 7 avec R une matrice identité, $r = \beta^\star$ et $m = K$. C.Q.F.D.

Test joint sur un sous ensemble des paramètres

Proposition 14

Partitionnons les régresseurs et paramètres en considérant le modèle suivant :

$$\mathbf{y} = X_1\beta_1 + X_2\beta_2 + \epsilon$$

où X_1 et X_2 sont respectivement des matrices $T \times K_1$ et $T \times K_2$, les vecteurs β_1 et β_2 sont respectivement des vecteurs $K_1 \times 1$ et $K_2 \times 1$. Soient SSE_1 la somme des carrés des résidus quand $\mathbf{y}$ est régressé seulment sur X_1 (le modèle contraint) et SSE_{12} la somme des carrés des résidus quand $\mathbf{y}$ est régressé sur toutes les variables explicatives (le modèle non contraint). Alors, sous l'hypothèse nulle $\beta_2 = 0$ la statistique :

$$\frac{(SSE_1 - SSE_{12})/K_2}{SSE_{12}/T-K_1-K_2}$$

suit une loi de Fisher à K_2 et $T - K_1 - K_2$ degrés de liberté.

Preuve de la proposition 14. Nous avons :

$$SSE_1 - SSE_{12} = \mathbf{y}' M_1 \mathbf{y} - \mathbf{y}' M \mathbf{y}$$

avec $M_1 = I_T - X_1(X_1'X_1)^{-1}X_1'$ et $M = I_T - X(X'X)^{-1}X'$ où $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$. En développant, il vient :

$$\begin{aligned} SSE_1 - SEE_{12} &= (X_1\beta_1 + X_2\beta_2 + \varepsilon)' M_1 (X_1\beta_1 + X_2\beta_2 + \varepsilon) - \varepsilon' M \varepsilon \\ &= \beta_1' X_1' M_1 X_1 \beta_1 + \beta_1' X_1' M_1 X_2 \beta_2 + \beta_1' X_1' M_1 \varepsilon \\ &\quad + \beta_2' X_2' M_1 X_1 \beta_1 + \beta_2' X_2' M_1 X_2 \beta_2 + \beta_2' X_2' M_1 \varepsilon \\ &\quad + \varepsilon' M_1 X_1 \beta_1 + \varepsilon' M_1 X_2 \beta_2 + \varepsilon' M_1 \varepsilon - \varepsilon' M \varepsilon \\ &= \beta_2' X_2' M_1 X_2 \beta_2 + \beta_2' X_2' M_1 \varepsilon + \varepsilon' M_1 X_2 \beta_2 + \varepsilon' M_1 \varepsilon - \varepsilon' M \varepsilon \end{aligned}$$

en notant que $M_1 X_1 = 0$ et $X_1' M_1 = 0$. Sous l'hypothèse nulle $\beta_2 = 0$ nous avons donc :

$$SSE_1 - SSE_{12} = \varepsilon' (M_1 - M) \varepsilon$$

Puisque M_1 et M sont des matrices symétriques, $M_1 - M$ est nécessairement une matrice symétrique. On peut aussi montrer que cette matrice est idempotente. En effet :

$$(M_1 - M)^2 = M_1^2 - M_1 M - M M_1 + M^2 = M_1 - 2M_1 M + M$$

puisque les matrices M_1 et M sont symériques et idempotentes. Pour que la matrice $M_1 - M$ soit idempotente il faut et il suffit donc que $2M_1 M$ soit égal à $-M$. On a :

$$M_1 M = M - X_1(X_1'X_1)^{-1}X_1' M$$

On sait que $X'M = 0$, a fortiori $X_1' M = 0$ et donc $M_1 M = M$ et $(M_1 - M)^2 =$

$M_1 - M$. Puisque $M_1 - M$ a une trace éqale à $T - K_1 - (T - K_1 - K_2) = K_2$ il s'ensuit que $(SSE_1 - SSE_{12})/K_2$ comme $\frac{\sigma_\varepsilon^2}{K_2}\chi^2(K_2)$. Par ailleurs nous savons déjà que $SSE_{12}/T-K_1-K_2$ est distribué comme $\frac{\sigma_\varepsilon^2}{T-K_1-K_2}\chi^2(T - K_1 - K_2)$. Enfin, comme le numérateur et le dénominateur de la statistique de test sont indépendants, car :

$$(M_1 - M)M = M - M = 0$$

la statistique est un ratio de deux variables aléatoires du χ^2 indépendantes et suit donc une loi de Fisher comme annoncée dans la proposition. C.Q.F.D.

Test de significativité jointes des pentes

- Supposons que le modèle contienne une constante $\rightarrow X_1$
- On veut tester $\beta_2 = \beta_3 = \cdots = \beta_K = 0$ contre au moins un paramètre différent de zéro.
- On utilise la statistique définie dans la proposition 14.
- Ici on a $SSE_1 = SST$ et donc $SSE_1 - SSE_{12} = SST - SSE = SSR$.
- Ainsi la statistique s'écrit :

$$\frac{SSR/K-1}{SSE/T-K} = \frac{R^2}{1-R^2}\frac{T-K}{K-1}$$

puisque $SSE = (1-R^2)SST$ et $SSR = R^2 SST$. On accepte l'hypothèse nulle si le R^2 est assez proche de zéro.

Références

- Anscombe, F. J. (1973). "Graphs in Statistical Analysis". In: *The American Statistician* 27(1), pp. 17–21.
- Greene, William H. (2017). *Econometric analysis*. Pearson.
- Schmidt, Peter (1976). *Econometrics*. Taylor & Francis.

Annexe A
Probabilités

Convergence en probabilité

Définition.

Si $\{X_n\}$ une suite de variables aléatoires. On dit que X_n converge en probabilité vers X si :

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \varepsilon\right) = 1$$

pour tout $\varepsilon > 0$. On note :

$$X_n \xrightarrow[n\to\infty]{\text{proba}} X$$

ou

$$\operatorname*{plim}_{n\to\infty} X_n = X$$

Théorème 11

Soient $\{X_n\}$ une suite de variables aléatoires qui converge en probabilité vers X et g une fonction continue en X, alors :

$$g\left(X_n\right) \xrightarrow[n\to\infty]{\textit{proba}} g(X)$$

- Le théorème 11 est aussi valable pour des vecteurs de variables aléatoires (dans $\mathbb{R}^p$) et une fonction g de $\mathbb{R}^p$ dans $\mathbb{R}^q$ continue en X.

- **Exemple.** Soient deux suites de variables aléatoires $\{X_n\}$ et $\{Y_n\}$ qui convergent en probabilité vers X et Y. Le théorème 11 nous dit que :

$$X_n Y_n \xrightarrow[n\to\infty]{\text{proba}} XY$$

ou que

$$\frac{X_n}{Y_n} \xrightarrow[n\to\infty]{\text{proba}} \frac{X}{Y}$$

tant que $Y \neq 0$ (continuité).

- Ce résultat, fort utile pour établir les propriétés asymptotiques d'un estimateur, est remarquable. Les choses ne sont souvent pas aussi simple, penser par exemple à l'espérance et l'inégalité de Jensen.

La loi normale univariée

Définition.

Si X suit une normale d'espérance μ et de variance σ^2, on note $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, alors sa fonction de densité de probabilité est :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

La loi normale multivariée

Définition.

Le vecteur aléatoire $\boldsymbol{X} \in \mathbb{R}^n$ suit une normale d'espérance $\boldsymbol{\mu}$ et de variance Σ, on note $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, alors sa fonction de densité de probabilité est :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} \, |\Sigma|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

Loi du χ^2

Définition.

Soit $X_1, \ldots, X_k$ une suite de variable aléaoires normales centrées-réduites indépendantes. Alors $\sum_{i=1}^{k} X_i^2$ suit une loi du chi-deux à k degrés de liberté, notée : $\chi^2(k)$.

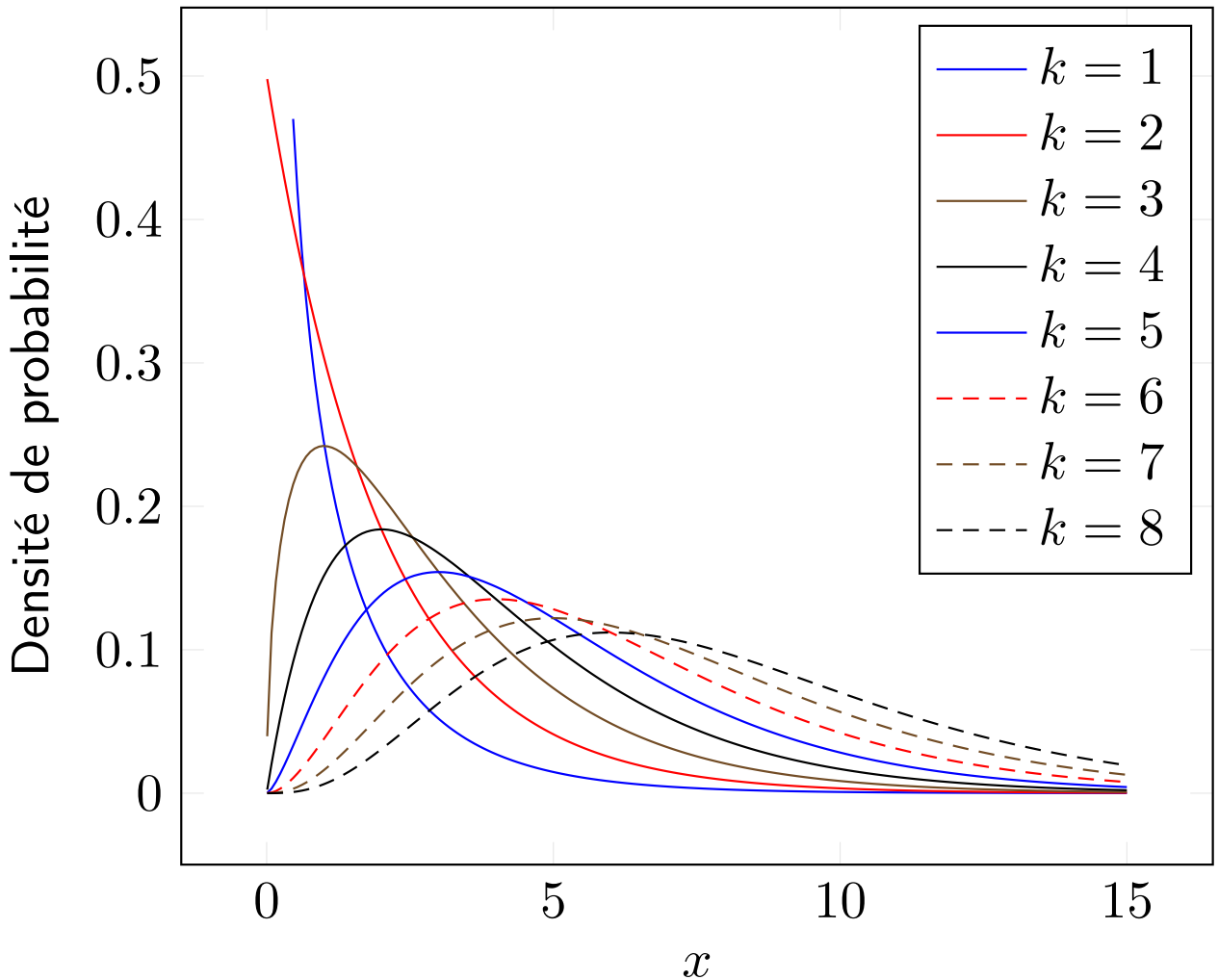
Proposition 15

La fonction de densité de probabilité de $X \sim \chi^2(k)$, avec $k \in \mathbb{N}$, est :

$$f_X(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$$

pour tout $x > 0$ et 0 sinon, où $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$ est la fonction Gamma. On a $\mathbb{E}[X] = k$ et $\mathbb{V}[X] = 2k$.

Loi du χ^2


Densité de probabilité
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
x
$k = 1$
$k = 2$
$k = 3$
$k = 4$
$k = 5$
$k = 6$
$k = 7$
$k = 8$

Loi de Student

Définition.

Soient Z une variable aléatoire normale centrée réduite et une variable aléatoire $U \perp Z$ distribuéesuivant la loi du χ^2 à k degrés de liberté. Alors

$$X = \frac{Z}{\sqrt{U/k}}$$

suit une loi de Student à k degrés de liberté, notée t_k.

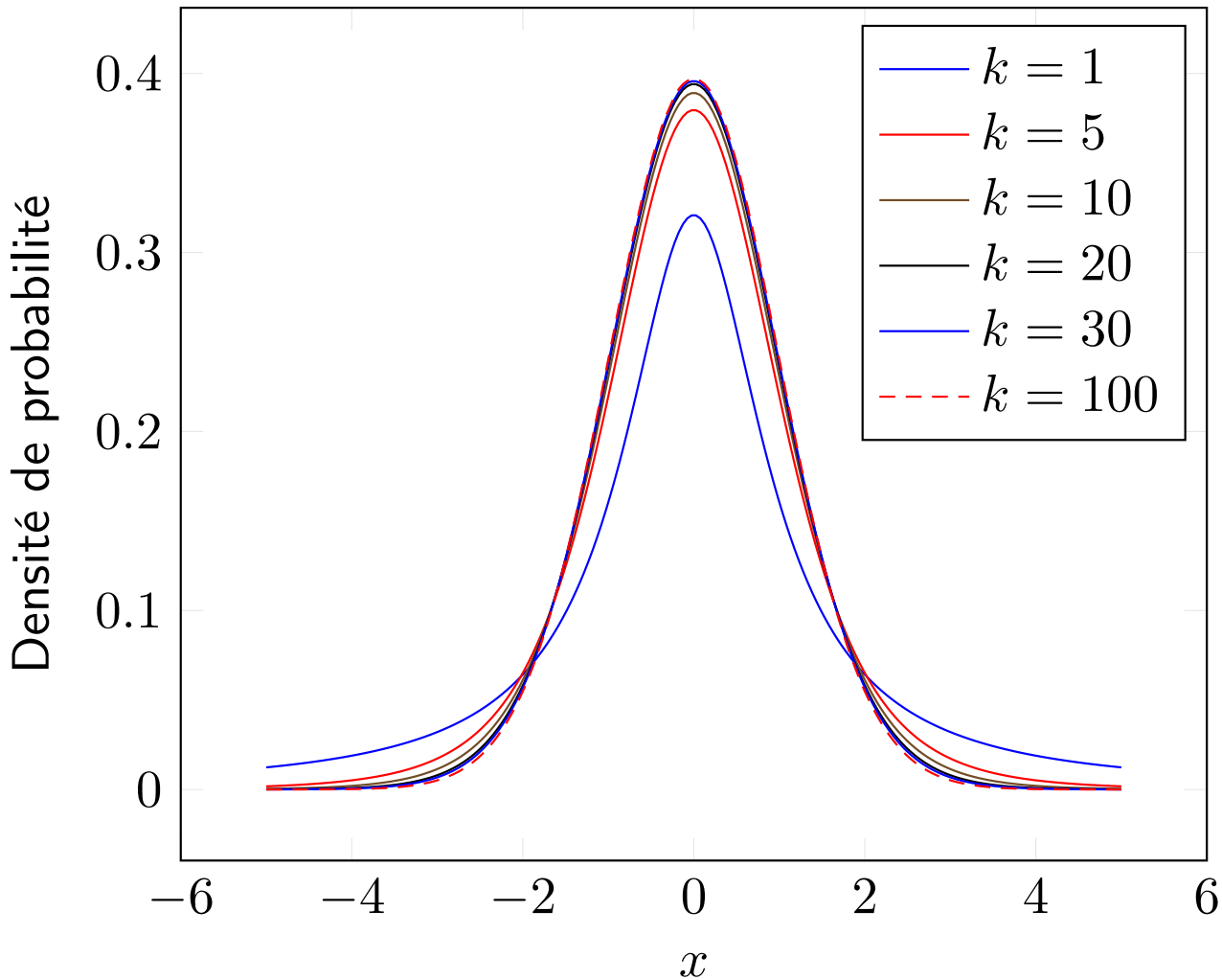
Proposition 16

La fonction de densité de probabilité de $X \sim t_k$, avec $k \in \mathbb{N}$, est :

$$f_X(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

pour tout $x \in \mathbb{R}$, où $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$ est la fonction Gamma. On a $\mathbb{E}[X] = 0$ si $k > 1$ et $\mathbb{V}[X] = \frac{k}{k-2}$ si $k > 2$. La loi de student converge vers la loi normale si k tend vers l'infini.

Loi de Student



Loi de Fisher

Définition.

Soient $U \sim \chi^2(p)$ et $V \sim \chi^2(q)$, avec $U \perp V$. $X = \frac{U/p}{V/q}$ une variable aléatoire réelle distribuée selon la loi de Fisher de degrés de liberté p et q. On note la distribution de Fisher : $F(p, q)$.

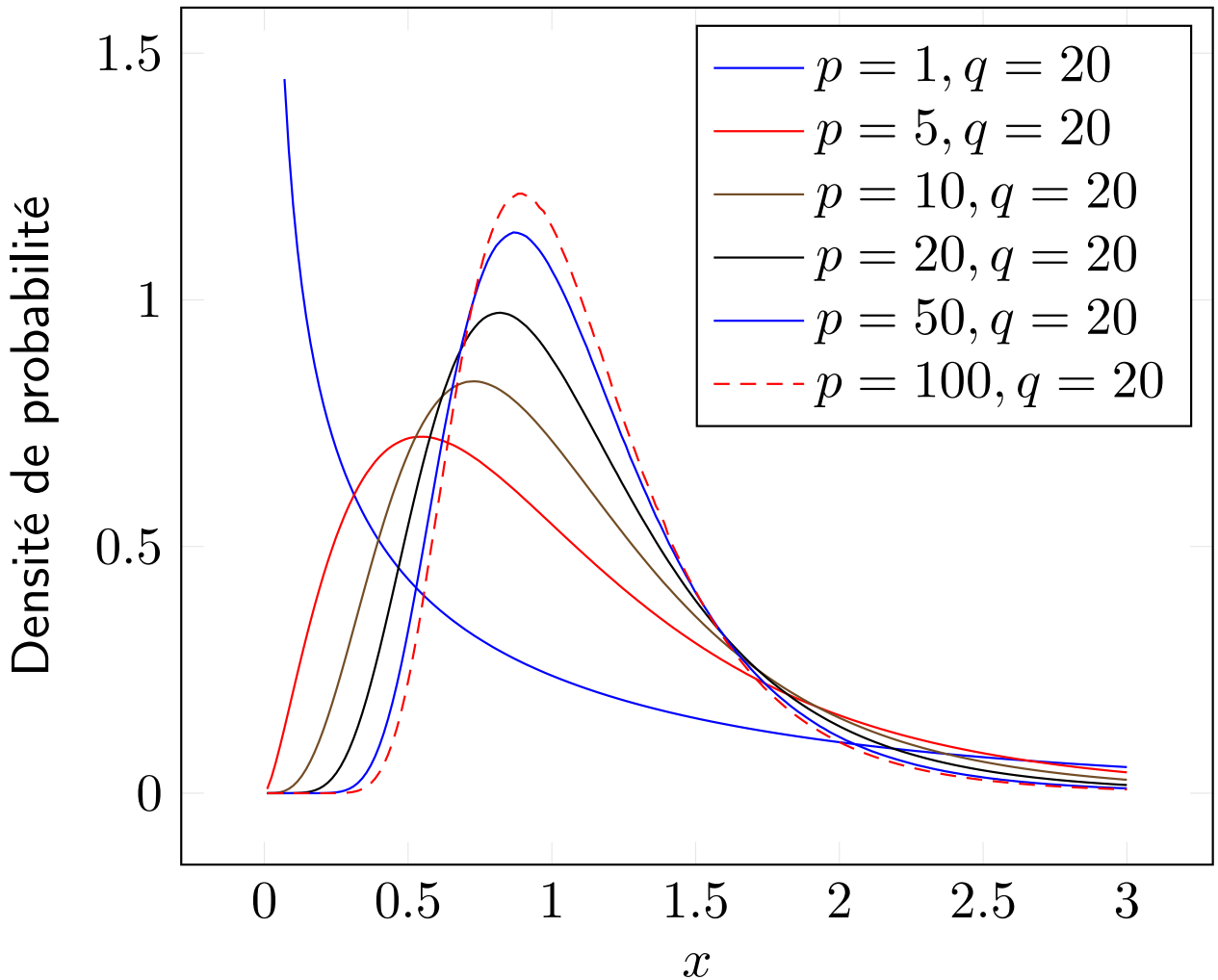
Proposition 17

La fonction de densité de probabilité de $X \sim F(p, q)$*, avec* $(p, q) \in \mathbb{N}^\star \times \mathbb{N}^\star$*, est :*

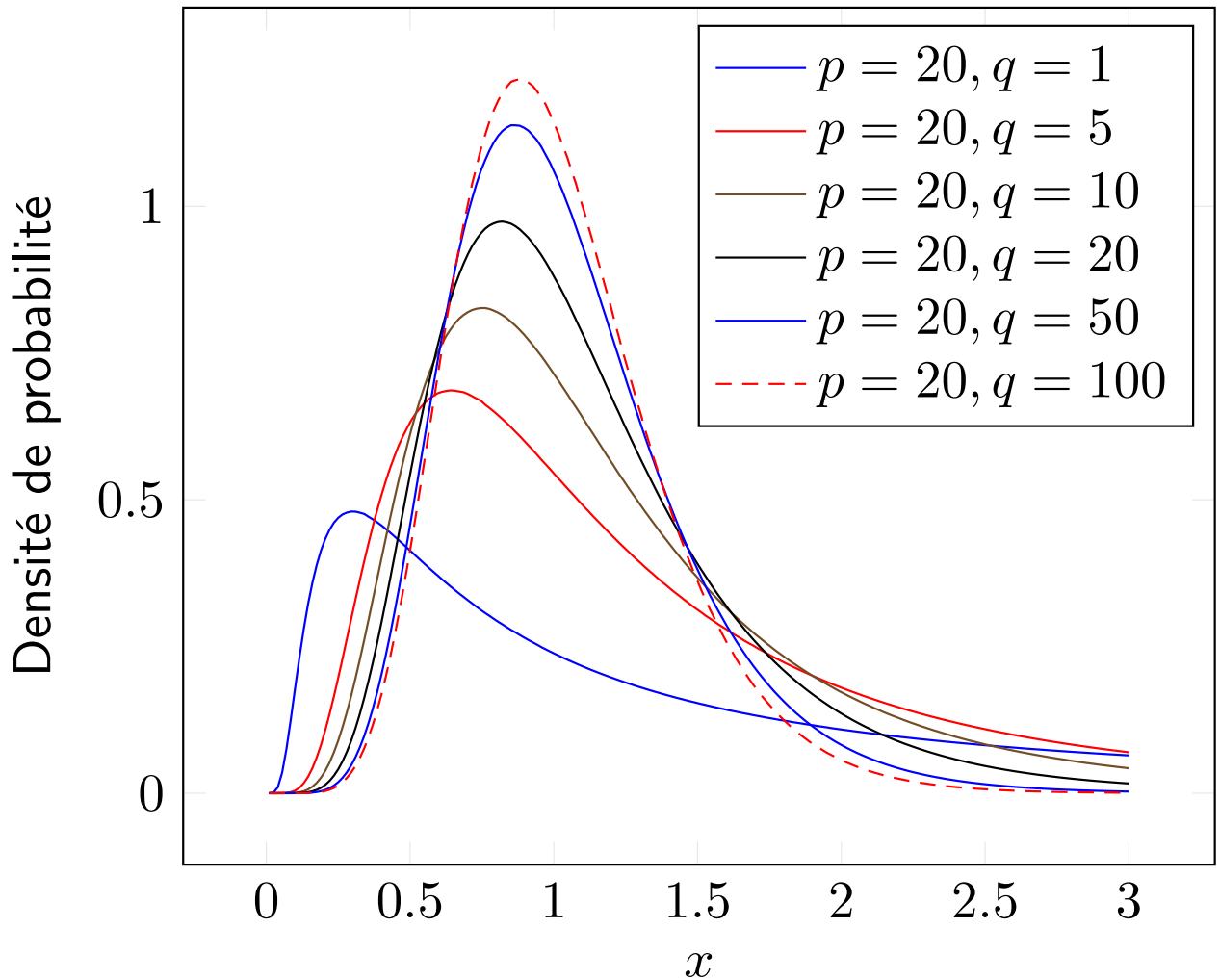
$$f_X(x) = \frac{\left(\frac{px}{px+q}\right)^{\frac{p}{2}} \left(1 - \frac{px}{px+q}\right)^{\frac{q}{2}}}{xB\left(\frac{p}{2}, \frac{q}{2}\right)}$$

pour tout $x \in \mathbb{R}_+$*, où* $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$ *est la fonction Beta. On a* $\mathbb{E}[X] = \frac{q}{q-2}$ *si* $q > 2$ *et* $\mathbb{V}[X] = \frac{2q^2(p+q-2)}{p(q-2)^2(q-4)}$ *si* $q > 4$.

Loi de Fisher



Loi de Fisher



Loi de Fisher