Université de Grenoble Alpes
(L3 MIASH, S2)

ÉCONOMÉTRIE

Régression Linéaire et Moindres Carrés à distance finie(1)

Estimateur des MCO

(Cette version : 11 février 2024)

Michal W. Urdanivia [1]

1. Contact : michal.wong-urdanivia@univ-grenoble-alpes.fr, Université de Grenoble Alpes, Faculté d'Économie, GAEL.

Table des matières

1. Définitions

Une question courante en économétrie concerne l'étude de l'effet d'un groupe de variables $X \in \mathcal{X} \subseteq \mathbb{R}^K$, traditionnellement appelées *régresseurs*, sur une autre variable $Y \in \mathcal{Y} \subseteq \mathbb{R}$ traditionnellement appelée *variable dépendante*. On dispose de données sur (Y, X), à savoir un *échantillon* de taille n, $\{(Y_i, X_i)\}_{i=1}^n$, où Y_i est une variable aléatoire et X_i est un vecteur $K \times 1$(de variables aléatoires), i.e.,

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iK} \end{pmatrix}$$

Une paire (Y_i, X_i) est appelée observation(sous entendu de (Y, X)). Le vecteur X_i contient les valeurs des K variables pour l'observation i. Pour des *données en coupe* [2] il est souvent supposé que toutes les observations sont tirées indépendamment les unes des autres à partir d'une même distribution. On dit dans ce cas que l'échantillon d'observations $\{(Y_i, X_i)\}_{i=1}^n$ est un échantillon aléatoire ou de manière équivalente que les observations sont identiquement et indépendamment distribuées(i.i.d. en abrégé). Remarquons que l'hypothèse d'observations i.i.d. ne signifie pas que Y_i et X_i soient indépendants, mais plutôt que l'observation (Y_i, X_i) est indépendante de toute autre observation (Y_j, X_j) pour $i \neq j$, n'excluant donc pas que Y_i et X_i puissent être liés

L'outil auquel nous allons nous intéresser dans ce cours pour étudier la relation entre la variable dépendante et les régresseurs est l'espérance conditionnelle de Y_i sachant X_i, $\mathrm{E}(Y_i|X_i)$, laquelle vue comme une fonction de X_i est appelée *fonction de régression*(ou plus succinctement régression) de Y_i sur X_i La différence entre Y_i et son espérance conditionnelle est appelée *terme d'erreur*(ou plus succinctement *erreur*),

$$U_i = Y_i - \mathrm{E}(Y_i|X_i) \tag{1}$$

et l'on note que contrairement à X_i et Y_i, l'erreur U_i n'est pas une variable observable par l'analyste étant donné que l'espérance conditionnelle lui est inconnue

Dans un cadre *paramétrique* ou *semi-paramétrique*, il est souvent supposé que l'espérance conditionnelle est connue à un ensemble de *paramètres* près. Ainsi dans le *modèle de régression linéaire* on suppose que $\mathrm{E}(Y_i|X_i)$ est linéaire par rapport à un vecteur de paramètres inconnus,

$$\mathrm{E}(Y_i|X_i) = X_{i1}\beta_1 + X_{i2}\beta_2 + ... + X_{iK}\beta_K = X_i^\top \beta \tag{2}$$

où,

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}$$

est un vecteur de K paramètres constants. La linéarité de $\mathrm{E}(Y_i|X_i)$ peut être justifiée, si par exemple, la distribution des observations $\{(Y_i, X_i)\}_{i=1}^n$ est une loi normale multivariée. Rappelons néanmoins que

2. Rappelons que des données en coupe sont des données où chaque observation ne concerne qu'une seule unité d'observation. Par exemple s'il s'agit d'observations sur des individus l'observation i concernera un individu différent de l'observation j.

lorsque $\mathrm{E}(Y_i|X_i)$ n'est pas linéaire il est possible de caractériser β de manière à ce que (2) constitue la *meilleure prédiction linéaire* de la variable dépendante par les régresseurs. Notons aussi que comme

$$\beta_k = \frac{\partial \mathrm{E}(Y_i|X_i)}{\partial X_{ik}}, \quad k = 1, 2, ..., K.$$

le vecteur β est le vecteur des *effets marginaux* des régresseurs, i.e., β_k donne la variation dans l'espérance conditionnelle de Y_i lorsque le régresseur X_{ik} varie, pour des valeurs fixes des autres régresseurs X_{il}, $l = 1, 2, ..., K$, $l \neq k$. Ceci est une des raisons pour lesquelles un des principaux objectifs est l'estimation du vecteur inconnu β à partir des données
Observons que les équations (1) et (2) permettent d'écrire,

$$Y_i = X_i^\top \beta + U_i \tag{3}$$

où par définition de (1)

$$\mathrm{E}(U_i|X_i) = 0 \tag{4}$$

Ceci implique que les régresseurs ne contiennent aucune information quant à l'écart entre Y_i et sont espérance conditionnelle. En outre, la *règle des espérances itérées* implique que les erreurs ont une espérance nulle : $\mathrm{E}(U_i) = 0$. Notons aussi qu'avec des observations i.i.d. les erreurs sont aussi i.i.d
Une hypothèse fréquente sur les erreurs consiste à supposer qu'ils sont *homoscédastiques*(on parle d'hypothèse d'homoscédasticité), par quoi on entend que leur variance est indépendante des régresseurs, et la même pour toutes les observations,

$$\mathrm{Var}(U_i|X_i) = \sigma^2$$

pour une constante $\sigma^2 > 0$.

2. Conditions

Avant de donner une définition formelle du modèle de régression linéaire, introduisons les notations vectorielles et matricielles suivantes,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1K} \\ X_{21} & X_{22} & \dots & X_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nK} \end{pmatrix}, \ \mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$$

Le modèle de régression linéaire consiste dans les hypothèses suivantes :

Condition C1. $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$

Condition C2. $\mathrm{E}(\mathbf{U}|\mathbf{X}) = 0$ p.s.

Condition C3. $\mathrm{Var}(\mathbf{U}|\mathbf{X}) = \sigma^2 \mathbf{I}_n$ p.s.

Condition C4. $\mathrm{Rang}(\mathbf{X}) = K$ p.s. [3]

3. Le rang colonne(ligne) d'une matrice est le nombre maximal de colonnes(lignes) lnéairement indépendantes). On peut montrer que pour toute matrice, le rang colonne et le rang ligne sont égaux. Si $\mathbf{A}$ est une matrice $n \times K$, alors $\mathrm{Rang}(\mathbf{A}) \leq \min(n, K)$. Si $\mathrm{Rang}(\mathbf{A}) = n$(ou $\mathrm{Rang}(\mathbf{A}) = K$), on dit que $\mathbf{A}$ est de rang ligne(colonne) plein. Quelques propriétés :

$$\mathrm{Rang}(\mathbf{A}) = \mathrm{Rang}(\mathbf{A}^\top) = \mathrm{Rang}(\mathbf{A}^\top\mathbf{A}) = \mathrm{Rang}(\mathbf{A}\mathbf{A}^\top),$$
$$\mathrm{Rang}(\mathbf{AB}) \leq \min\left(\mathrm{Rang}(\mathbf{A}), \mathrm{Rang}(\mathbf{B})\right),$$
$$\mathrm{Rang}(\mathbf{AB}) = \mathrm{Rang}(\mathbf{A}) \text{ si } \mathbf{B} \text{ est carrée ou de rang plein}$$

Plutôt que de conditionner par rapport aux valeurs observées des régresseurs, on peut supposer que $\mathbf{X}$ n'est pas aléatoire, i.e., supposer que la valeur de $\mathbf{X}$ est fixe dans des échantillons répétés. Dans ce cas là les hypothèses (C2) et (C3) peuvent être remplacés par, respectivement $\mathrm{E}(\mathbf{U}) = 0$ et $\mathrm{Var}(\mathbf{U}) = \sigma^2\mathbf{I}_n$. Dans la mesure où conditionner par rapport à $\mathbf{X}$ est équivalent à traiter les valeurs des régresseurs comme fixes, les deux ensembles d'hypothèses conduisent aux mêmes résultats Pour l'inférence on suppose parfois que,

Condition C5. $\mathbf{U}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$

Dans le cas de régresseur fixes, plutôt que (C5) il sera supposé que la distribution inconditionnelle des erreurs est normale. Les hypothèses (C1)-(C5) définissent alors le *modèle de régression linéaire normal* avec dans ce cas,

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$$

Remarquons qu'étant donné que les covariances dans (C5) sont toutes nulles, (C5) implique l'indépendance des erreurs. Les hypothèses (C1)-(C4) seules, n'impliquent pas l'indépendance entre les observations. En fait, plusieurs résultats importants n'exigent pas d'observations indépendantes. Néanmoins, nous supposerons parfois l'indépendance sans la normalité.

Condition C6. Les observations $\{(Y_i, X_i)\}_{i=1}^n$ sont i.i.d.

Dans le cas de régresseurs fixes cette hypothèse peut être remplacé par celle d'erreurs, $U_1, ..., U_n$, i.i.d L'hypothèse (C2) dit que $\mathbf{U}$ est indépendant de $\mathbf{X}$ en espérance, ce qui est une hypothèse forte. On appelle aussi qualifie aussi cette condition, *condition d'exogénéité forte*. Cependant, plusieurs résultats importants peuvent être obtenus avec une hypothèse plus faible d'absence de corrélation. Celle-ci est qualifiée de *condition d'exogénéité faible* :

Condition C7. Pour $i = 1, 2, ..., n$, $\mathrm{E}(X_i U_i) = 0$, et $\mathrm{E}(U_i) = 0$.

Toutefois sous cette condition $X_i^\top\beta$ ne peut pas s'interpréter comme une espérance conditionnelle, auquel cas (3) doit être vu comme un *processus générateur des données* L'hypothèse (C3) implique que les erreurs U_i ont la même variance pour tout i, et ne sont pas corrélés entre eux, i.e., $\mathrm{E}(U_i U_j|\mathbf{X}) = 0$ pour $i \neq j$. Notons que l'indépendance entre les erreurs peut aussi être obtenue avec la condition (C5) ou sous les conditions (C1) et (C6) L'hypothèse (C4) exige que le colonnes de $\mathbf{X}$ soient linéairement indépendantes. Que cette hypothèse ne soit pas vérifiée signifie qu'un ou plus de régresseurs duplique l'information contenue dans les autres, et ce faisant doit être écarté Souvent, une des colonnes de $\mathbf{X}$(souvent la première) est le vecteur unitaire et le paramètre qui lui est associé est appelé *constante*. La constante du modèle donne la valeur moyenne de la variable dépendante lorsque tous les régresseurs sont égaux à zéro.

3. Estimation par la méthode des moments

Nous allons à présent construire des estimateurs des paramètres β et σ^2. Rappelons qu'un estimateur est toute fonction des observations $\{(Y_i, X_i)\}_{i=1}^n$. Un estimateur peut dépendre des erreurs inconnues ou des paramètres inconnus β mais uniquement par le biais des variables observables $\mathbf{Y}$ et $\mathbf{X}$. Un estimateur n'est pas forcément unique en ce sens que pour un même paramètre plusieurs estimateur peuvent exister

Une des méthodes les plus anciennes pour construire des estimateurs est la *méthode des moments*(MM). La MM consiste à construire des estimateurs pour des paramètres définis par des moments théoriques en considérant les contreparties empiriques de ces moments appelées alors

moments empiriques. Par exemple si un paramètre est défini au travers d'une espérance(moment théorique), son estimateur sera construit à partir d'une moyenne(moment empirique) calculée sur les observations Dans le cas présent, les hypothèses (C1), et (C2) ou (C7) impliquent que la vraie valeur de β doit satisfaire,

$$\mathrm{E}(X_i U_i) = \mathrm{E}\left(X_i(Y_i - X_i^\top \beta)\right) = 0 \tag{5}$$

Un *estimateur des moments*(i.e., obtenu selon la MM) de β, $\hat{\beta}$, est obtenu en remplaçant l'espérance dans (5) par la moyenne empirique,

$$n^{-1}\sum_{i=1}^{n} X_i(Y_i - X_i^\top \hat{\beta}) = n^{-1}\sum_{i=1}^{n} X_i Y_i - n^{-1}\sum_{i=1}^{n} X_i X_i^\top \hat{\beta} = 0 \tag{6}$$

En résolvant par rapport à $\hat{\beta}$ on obtient,

$$\hat{\beta} = \left(n^{-1}\sum_{i=1}^{n} X_i X_i^\top\right)^{-1} n^{-1}\sum_{i=1}^{n} X_i Y_i \tag{7}$$

qui peut s'écrire alternativement,

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i^\top\right)^{-1} \sum_{i=1}^{n} X_i Y_i \tag{8}$$

ou,

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \tag{9}$$

où l'on note que la matrice $\sum_{i=1}^{n} X_i X_i^\top = \mathbf{X}^\top \mathbf{X}$ est inversible sous l'hypothèse (C4)[4]

Une fois $\hat{\beta}$ calculé, on définit les *valeurs ajustées* ou *prédictions* , ainsi qu'un vecteur $n \times 1$ des valeurs ajustées ou des prédictions, par respectivement,

$$\hat{Y}_i = X_i^\top \hat{\beta}, \quad \hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_n)^\top$$

De la même manière, on définit les *résidus* , et le vecteur $n \times 1$ des résidus, par respectivement,

$$\hat{U}_i = Y_i - X_i^\top \hat{\beta}, \quad \hat{\mathbf{U}} = (\hat{U}_1, \hat{U}_2, ..., \hat{U}_n)^\top$$

Notons que du fait de (6) le vecteur des résidus vérifie les *K équations normales*,

$$\sum_{i=1}^{n} \hat{U}_i X_i = \begin{pmatrix} \sum_{i=1}^{n} \hat{U}_i X_{i1} \\ \sum_{i=1}^{n} \hat{U}_i X_{i2} \\ \vdots \\ \sum_{i=1}^{n} \hat{U}_i X_{iK} \end{pmatrix} = 0 \tag{10}$$

ou en notation matricielle,

$$\mathbf{X}^\top \hat{\mathbf{U}} = 0 \tag{11}$$

4. Pour montrer que $\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^{n} X_i X_i^\top$ notons que,

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix}^\top \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & \dots & X_n \end{pmatrix} \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} = X_1 X_1^\top + X_2 X_2^\top + \ldots + X_n X_n^\top = \sum_{i=1}^{n} X_i X_i^\top$$

Remarquons aussi que si le modèle contient une constante alors il résulte des équations normales que $\sum_{i=1}^{n} \hat{U}_i = 0$(il suffit en effet de considérer que, par exemple, le premier régresseur est constant et égal à 1)

Afin d'estimer σ^2 considérons,

$$\sigma^2 = \mathrm{E}(U_i^2) = \mathrm{E}\left((Y_i - X_i^\top \beta)^2\right)$$

Dans la mesure où β, est inconnu un estimateur sera obtenu en remplaçant β par sont estimateur des moments,

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta})^2 \tag{12}$$

4. Moindres carrés

Pour motiver l'estimation par la méthode des moindres carrés prenons comme point de départ le problème consistant à minimiser l'erreur de prédiction quand on cherche à prédire Y_i par son espérance conditionnelle, $\mathrm{E}(Y_i|X_i)$, supposée être une fonction linéaire telle que (2). Plus précisément, $Y_i - \mathrm{E}(Y_i|X_i)$ étant l'erreur de prédiction on cherche β qui minimise un critère de perte quadratique,

$$\beta \in \arg\min_{b \in \mathbb{R}^K} S(b)$$

où $S(b) = \mathrm{E}\left((Y_i - X_i^\top b)^2\right)$. La contrepartie empirique de ce problème permet de définir un estimateur de β par,

$$\hat{\beta} \in \arg\min_{b \in \mathbb{R}^K} S_n(b)$$

où $S_n(b) = n^{-1} \sum_{i=1}^{n} \left((Y_i - X_i^\top \beta)^2\right)$, est la contrepartie empirique de la fonction objectif $S(b)$ Nous pouvons montrer que l'estimateur des moments de la section précédente est aussi l'estimateur des moindres carrés. Pour cela réécrivons la fonction objectif précédente,

$$\begin{aligned}
S_n(b) =& (\mathbf{Y} - \mathbf{X}b)^\top (\mathbf{Y} - \mathbf{X}b) \\
=& (\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}b)^\top (\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}b) \\
=& (\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\mathbf{X}\hat{\beta} - \mathbf{X}b)^\top (\mathbf{X}\hat{\beta} - \mathbf{X}b) + 2(\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{X}\hat{\beta} - \mathbf{X}b) \\
=& (\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - b)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - b) + 2\hat{\mathbf{U}}^\top \mathbf{X} (\hat{\beta} - b) \\
=& (\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - b)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - b)
\end{aligned}$$

On note que la minimisation de $S_n(b)$ équivaut à minimiser $(\hat{\beta} - b)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - b)$ car $(\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{Y} - \mathbf{X}\hat{\beta})$ ne fait pas intervenir b. Sous l'hypothèse (C4) la matrice $\mathbf{X}$ est de plein rang, et dans ce cas $\mathbf{X}^\top \mathbf{X}$ est définie positive,

$$(\hat{\beta} - b)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - b) \geq 0$$

et $(\hat{\beta} - b)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - b) = 0$ ssi $\hat{\beta} = b$ Alternativement, nous pouvons montrer que $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ est l'estimateur des moindres carrés de β(i.e., il minimise $S_n(b)$). Pour cela, écrivons,

$$S_n(b) = \mathbf{Y}^\top \mathbf{Y} - 2b^\top \mathbf{X}^\top \mathbf{Y} + b^\top \mathbf{X}^\top \mathbf{X} b$$

En utilisant le fait que pour une matrice symétrique $\mathbf{A}$,

$$\frac{\partial (x^\top \mathbf{A} x)}{\partial x} = 2\mathbf{A}x$$

la condition du premier ordre est,

$$\frac{\partial S_n(\hat{\beta})}{\partial b} = -2\mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\hat{\beta} = 0$$

ce qui permet d'obtenir,

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Remarquons aussi que les conditions du premier ordre peuvent s'écrire $\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0$, ce qui correspond aux équations normales vue précédemment.

5. Propriétés de l'estimateur des moindres carrés

Nous allons présenter un certain nombre de propriétés de l'estimateur des moindres carrés.

Propriété P1. *$\hat{\beta}$ est un estimateur linéaire.*

Démonstration. Un estimateur b est linéaire s'il peut s'écrire comme $b = \mathbf{AY}$, où $\mathbf{A}$ est une matrice quelconque qui dépend de $\mathbf{X}$ uniquement, et ne dépend pas de $\mathbf{Y}$. Pour l'estimateur des moindres carrés nous avons, $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. □

Propriété P2. *Sous les hypothèses* (C1)*,* (C2)*, et* (C4)*, $\hat{\beta}$ est sans biais, i.e.,*

$$\mathrm{E}(\hat{\beta}) = \beta$$

Démonstration. Pour montrer cette propriété écrivons, en utilisant l'hypothèse (C1),

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta + \mathbf{U}) = \beta + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{U}$$

Calculons l'espérance conditionnelle de $\hat{\beta}$,

$$\mathrm{E}\left(\hat{\beta}|\mathbf{X}\right) = \mathrm{E}\left(\beta + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{U}|\mathbf{X}\right) = \beta + \mathrm{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{U}|\mathbf{X}\right)$$

Notons que,

$$\mathrm{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{U}|\mathbf{X}\right) = (\mathbf{XX})^{-1}\,\mathbf{X}^\top\,\mathrm{E}(\mathbf{U}|\mathbf{X}) = 0$$

car sous l'hypothèse (C2), $\mathrm{E}(\mathbf{U}|\mathbf{X}) = 0$. Nous avons donc,

$$\mathrm{E}\left(\hat{\beta}|\mathbf{X}\right) = \beta \tag{13}$$

et par la règle des espérances itérées,

$$\mathrm{E}\left(\hat{\beta}\right) = \mathrm{E}\left(\mathrm{E}\left(\hat{\beta}|\mathbf{X}\right)\right) = \beta$$

□

L'équation (13) montre que $\hat{\beta}$ est conditionnellement sans biais sachant $\mathbf{X}$. On remarque aussi que pour que $\hat{\beta}$ soit sans biais l'hypothèse (C7) n'est pas suffisante.

Propriété P3. *Sous les hypothèses* (C1)*,* (C2)*, et* (C4)*,*

$$\mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{E}\left(\mathbf{U}\mathbf{U}^\top|\mathbf{X}\right)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$$

et avec des erreurs homoscédastiques(i.e., sous l'hypothèse (C3)*),*

$$\mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

Démonstration. Pour montrer ces résultats, partons de la définition de la variance conditionnelle de $\hat{\beta}$,

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) &= \mathrm{E}\left(\left(\hat{\beta}-\mathrm{E}(\hat{\beta}|\mathbf{X})\right)\left(\hat{\beta}-\mathrm{E}(\hat{\beta}|\mathbf{X})\right)^{\top}|\mathbf{X}\right)\\ &= \mathrm{E}\left(\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\top}|\mathbf{X}\right)\\ &= \mathrm{E}\left((\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{U}\mathbf{U}^{\top}\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}|\mathbf{X}\right)\\ &= (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\,\mathrm{E}\left(\mathbf{U}\mathbf{U}^{\top}|\mathbf{X}\right)\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\end{aligned}$$

Et avec des erreurs homoscédastiques, $\mathrm{E}(\mathbf{U}\mathbf{U}^{\top}|\mathbf{X}) = \sigma^2\mathbf{I}_n$, de sorte que,

$$\begin{aligned}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\,\mathrm{E}\left(\mathbf{U}\mathbf{U}^{\top}|\mathbf{X}\right)\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1} &= (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\\ &= \sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\\ &= \sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1}\end{aligned}$$

Notons qu'avec des régresseurs fixes $\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1}$. □

Propriété P4. *Sous les hypothèses* (C1) - (C5)*,*

$$\hat{\beta}|\mathbf{X} \sim \mathcal{N}\left(\beta, \sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1}\right)$$

Démonstration. Il est suffit ici de montrer ici que conditionnellement à $\mathbf{X}$ la distribution de $\hat{\beta}$ est normale. On aura alors que, $\hat{\beta}|\mathbf{X} \sim \mathcal{N}\left(\mathrm{E}(\hat{\beta}|\mathbf{X}), \mathrm{Var}(\hat{\beta}|\mathbf{X})\right)$. Néanmoins la normalité de $\hat{\beta}|\mathbf{X}$ résulte ici de ce que $\hat{\beta}$ est une fonction de linéaire de $\mathbf{Y}$, et que sous l'hypothèse (C5) $\mathbf{Y}|\mathbf{X}$ est normale. □

Notons que dans le cas de régresseur fixes, il suffit d'omettre le conditionnement par rapport à $\mathbf{X}$ et,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1}\right)$$

Propriété P5. *(**Éfficacité** [5]) Sous les hypothèses* (C1)-(C4)*, l'estimateur des moindres carrés est le meilleur estimateur linéaire sans biais de β, dans le sens où il s'agit de l'estimateur, dans la classe des estimateurs linéaires et sans biais, qui présente la plus petite variance. i.e., pour tout estimateur linéaire sans biais, b, la matrice* $\mathrm{Var}\,(b|\mathbf{X}) - \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right)$ *doit être semi-définie positive :*

$$\mathrm{Var}\,(b|\mathbf{X}) - \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) \geq 0$$

En outre, si $\tilde{\beta}$ *est un estimateur linéaire et sans biais et* $\mathrm{Var}\left(\tilde{\beta}|\mathbf{X}\right) = \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right)$*, alors* $\tilde{\beta} = \hat{\beta}$ *p.s.*

Avant de démontrer ce résultat notons qu'il discute la variance conditionnelle de l'estimateur des moindres carrés, et ce faisant il se réfère à des estimateurs conditionnellement sans biais.

Démonstration. Soit b un estimateur linéaire sans biais de β. Il doit ainsi vérifier,

$$b = \mathbf{A}\mathbf{Y}, \quad \mathrm{E}(b|\mathbf{X}) = \beta$$

5. Ce résultat est aussi connu sous le nom de *théorème de Gauss-Markov.*

Ces deux conditions impliquent que $\mathbf{AX} = \mathbf{I}_K$ p.s. En effet,

$$\begin{aligned}\mathrm{E}(b|\mathbf{X}) =& \mathrm{E}\left(\mathbf{A}\left(\mathbf{X}\beta + \mathbf{U}\right)\right)\\ =& \mathbf{AX}\beta + \mathbf{A}\,\mathrm{E}(\mathbf{U}|\mathbf{X})\end{aligned}$$

Par l'hypothèse (C2), $\mathrm{E}(\mathbf{U}|\mathbf{X}) = 0$, et par conséquent, pour que b soit sans biais nous avons besoin de $\mathbf{AX} = \mathbf{I}_K$ Montrons maintenant que $\mathrm{Cov}\left(\hat{\beta}, b|\mathbf{X}\right) = \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right)$,

$$\begin{aligned}\mathrm{Cov}\left(\hat{\beta}, b|\mathbf{X}\right) =& \mathrm{E}\left((\hat{\beta} - \beta)(b - \beta)^\top\right)\\ =& \mathrm{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{A}^\top|\mathbf{X}\right)\\ =& (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X})\mathbf{A}^\top)\\ =& \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{A}^\top(\text{car sous (C3)},\ \ \mathrm{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X}) = \sigma^2\mathbf{I}_n)\\ =& \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}(\text{car},\ \ \mathbf{X}^\top\mathbf{A}^\top = \mathbf{I}_K)\\ =& \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right)\end{aligned}$$

Finalement,

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta} - b|\mathbf{X}\right) =& \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) - \mathrm{Cov}\left(\hat{\beta}, b|\mathbf{X}\right) - \mathrm{Cov}\left(b, \hat{\beta}|\mathbf{X}\right) + \mathrm{Var}\left(b\mathbf{X}\right)\\ =& \mathrm{Var}\left(b|\mathbf{X}\right) - \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) \qquad (14)\end{aligned}$$

et notons que dans la mesure où toute matrice de variance-covariances est semi-définie positive, nous avons,

$$\mathrm{Var}\left(b|\mathbf{X}\right) - \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) \geq 0$$

Pour démontrer l'unicité, considérons un estimateur linéaire sans biais $\tilde{\beta}$ tel que $\mathrm{Var}\left(\tilde{\beta}|\mathbf{X}\right) = \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right)$. Alors, par (14), $\mathrm{Var}\left(\hat{\beta} - b|\mathbf{X}\right) = 0$, et par conséquent, $\tilde{\beta} = \hat{\beta} + c(\mathbf{X})$ pour une fonction $c(\mathbf{X})$ à valeurs dans $\mathbb{R}^K$ qui dépend uniquement de $\mathbf{X}$. Cependant, comme $\hat{\beta}$ et $\tilde{\beta}$ sont conditionnellement sans biais sachant $\mathbf{X}$, il s'en suit que $c(\mathbf{X}) = 0$ p.s. □

Notons que l'hypothèse (C3), $\mathrm{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X}) = \sigma^2\mathbf{I}_n$, joue un rôle crucial dans la démonstration du résultat précédent. Sans elle, il ne serait pas possible de tirer des conclusions quant à l'efficacité de l'estimateur des moindres carrés.