ÉCONOMÉTRIE 2 : L3 MIASH, S2

RÉGRESSION LINÉAIRE
ESTIMATEUR DES MOINDRES CARRÉS

(RÉVISION DES PROPRIÉTÉS ET MÉTHODES D'INFÉRENCE À DISTANCE FINIE)

(CETTE VERSION : 28 FÉVRIER 2023)

MICHAL URDANIVIA [1]

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

TABLE DES MATIÈRES

1. INTRODUCTION

Dans ces premières notes de cours nous revenons sur l'estimateur des moindres carrés pour le modèle de régression linéaire étudié en économétrie 1 : ses propriétés et les méthodes courantes d'inférence qui lui sont associées dans un cadre non-asymptotique(i.e., à distance finie). Avant cela on s'intéresse aux interprétations du modèle.

2. ESTIMATEUR DES MOINDRES CARRÉS DANS LE MODÈLE DE RÉGRESSION LINÉAIRE

2.1. **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*),

$$U_i = Y_i - \mathrm{E}(Y_i|X_i) \tag{2.1}$$

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*

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.

régression linéaire on suppose que $\mathrm{E}(Y_i|X_i)$ est linéaire par rapport à un vecteur de paramètres inconnus,

$$\mathrm{E}(Y_i|X_i) = X_{i1}\beta_1 + X_{i2}\beta_2 + ... + X_{iK}\beta_K = X_i^\top \beta \tag{2.2}$$

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 lorsque $\mathrm{E}(Y_i|X_i)$ n'est pas linéaire il est possible de caractériser β de manière à ce que (2.2) 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 (2.1) et (2.2) permettent d'écrire,

$$Y_i = X_i^\top \beta + U_i \tag{2.3}$$

où par définition de (2.1)

$$\mathrm{E}(U_i|X_i) = 0 \tag{2.4}$$

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 $\sigma^2 > 0$.

2.2. **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]

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.

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 :

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

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 (2.3) 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.

2.3. **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,

$$\mathrm{E}(X_i U_i) = \mathrm{E}\left(X_i(Y_i - X_i^\top \beta)\right) = 0 \tag{2.5}$$

Un *estimateur des moments*(i.e., obtenu selon la MM) de β, $\hat{\beta}$, est obtenu en remplaçant l'espérance dans (2.5) par la moyenne empirique,

$$n^{-1}\sum_{i=1}^n X_i(Y_i - X_i^\top \hat{\beta}) = n^{-1}\sum_{i=1}^n X_i Y_i - n^{-1}\sum_{i=1}^n X_i X_i^\top \hat{\beta} = 0 \tag{2.6}$$

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

$$\hat{\beta} = \left(n^{-1}\sum_{i=1}^n X_i X_i^\top\right)^{-1} n^{-1}\sum_{i=1}^n X_i Y_i \tag{2.7}$$

qui peut s'écrire alternativement,

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i^\top \right)^{-1} \sum_{i=1}^{n} X_i Y_i \tag{2.8}$$

ou,

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{Y} \tag{2.9}$$

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 (2.6) le vecteur des résidus vérifie les *K équations normales*,

$$\sum_{i=1}^{n} \hat{U}_i X_i = \begin{pmatrix} \sum_{i=1}^{n} \hat{U}_i X_{i1} \\ \sum_{i=1}^{n} \hat{U}_i X_{i2} \\ \vdots \\ \sum_{i=1}^{n} \hat{U}_i X_{iK} \end{pmatrix} = 0 \tag{2.10}$$

ou en notation matricielle,

$$\mathbf{X}^\top \hat{\mathbf{U}} = 0 \tag{2.11}$$

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,

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta})^2 \tag{2.12}$$

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

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

2.4. **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 (2.2). 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.

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

Propriété 2.1. *$\hat{\beta}$ 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 2.2. *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,

$$\mathrm{E}\left(\hat{\beta}|\mathbf{X}\right) = \beta \tag{2.13}$$

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 (2.13) 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 2.3. *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 2.4. *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 2.5. ***(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,

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

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

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

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

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,

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta} - b|\mathbf{X}\right) &= \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) - \mathrm{Cov}\left(\hat{\beta}, b|\mathbf{X}\right) - \mathrm{Cov}\left(b, \hat{\beta}|\mathbf{X}\right) + \mathrm{Var}\left(b\mathbf{X}\right) \\
&= \mathrm{Var}\left(b|\mathbf{X}\right) - \mathrm{Var}\left(\hat{\beta}|\mathbf{X}\right) \qquad (2.14)
\end{aligned}$$

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

3. GÉOMÉTRIE DE MOINDRES CARRÉS

3.1. **Matrices de projection.** Nous pouvons penser à $\mathbf{Y}$ et aux colonnes $\mathbf{X}$ comme des éléments de l'espace euclidien à n dimensions, $\mathbb{R}^n$. On peut définir un sous-espace de $\mathbb{R}^n$ appelé l'*espace des colonnes* d'une matrice $n \times K$, $\mathbf{X}$. Il s'agit de la collection de tous les vecteurs dans $\mathbb{R}^n$ qui peuvent s'écrire comme des combinaisons linéaires des colonnes de $\mathbf{X}$,

$$\mathcal{S}(\mathbf{X}) = \left\{ z \in \mathbb{R}^n : z = \mathbf{X}b, b = (b_1, b_2, ..., b_K) \in \mathbb{R}^K \right\}$$

Étant donné deux vecteurs a, b, dans $\mathbb{R}^n$, la distance entre a et b est donné par la norme euclidienne [6] de leur différence $\|a - b\| = \sqrt{(a-b)^\top (a-b)}$. En conséquence, le problème des moindres carrés, à savoir la minimisation de la somme des carrés des erreurs, $(\mathbf{Y} - \mathbf{X}b)^\top (\mathbf{Y} - \mathbf{X}b)$, est celui de trouver, parmi tous les éléments de $\mathcal{S}(\mathbf{X})$, celui dont la distance par rapport à $\mathbf{Y}$ est la plus petite,

$$\min_{\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})} \|\mathbf{Y} - \tilde{\mathbf{Y}}\|^2$$

Le point le plus proche est obtenu en "traçant une perpendiculaire". Autrement dit, une solution au problème des moindres carrés, $\hat{Y} = \mathbf{X}\hat{\beta}$ doit être choisie de sorte que le vecteur des résidus, $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}}$ soit orthogonal(perpendiculaire) à chaque colonne de $\mathbf{X}$,

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

Un résultat de cela est que $\hat{\mathbf{U}}$ est orthogonal à chaque élément de $\mathcal{S}(\mathbf{X})$. En effet, si $z \in \mathcal{S}(\mathbf{X})$, alors il existe $b \in \mathbb{R}^K$ tel que $z = \mathbf{X}b$, et,

$$\begin{aligned} \hat{\mathbf{U}}^\top z &= \hat{\mathbf{U}}^\top \mathbf{X} b \\ &= 0 \end{aligned}$$

La collection des éléments de $\mathbb{R}^n$ orthogonaux à $\mathcal{S}(\mathbf{X})$ est appelée *complément orthogonal* de $\mathcal{S}(\mathbf{X})$,

$$\mathcal{S}^\perp(\mathbf{X}) = \left\{ z \in \mathbb{R}^n : z^\top \mathbf{X} = 0 \right\}$$

Soulignons que tout élément de $\mathcal{S}^\perp(\mathbf{X})$ est orthogonal à chaque élément de $\mathcal{S}(\mathbf{X})$.
Comme nous l'avions vu dans le cours précédent, la solution au problème des moindres carrés est donnée par,

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} \\ &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= \mathbf{P}_\mathbf{X} \mathbf{Y} \end{aligned}$$

où

$$\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

est appelée *matrice de projection orthogonale*. Pour tout vecteur $\mathbf{Y} \in \mathbb{R}^n$,

$$\mathbf{P}_\mathbf{X} \mathbf{Y} \in \mathcal{S}(\mathbf{X})$$

6. Pour un vecteur $x = (x_1, x_2, ..., x_n)$ sa norme euclidienne est définie comme $\|x\| = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2}$.

En outre, le vecteur des résidus est dans $\mathcal{S}^{\perp}(\mathbf{X})$,

$$\mathbf{Y} - \mathbf{P}_{\mathbf{X}}\mathbf{Y} \in \mathcal{S}^{\perp}(\mathbf{X}) \tag{3.1}$$

Pour montrer (3.1), notons d'abord, qu'étant donné que les colonnes de $\mathbf{X}$ sont dans $\mathcal{S}(\mathbf{X})$,

$$\begin{aligned}\mathbf{P}_{\mathbf{X}}\mathbf{X} &= \mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{X}\\ &= \mathbf{X}\end{aligned}$$

et comme $\mathbf{P}_{\mathbf{X}}$ est une matrice symétrique,

$$\mathbf{X}^{\top}\mathbf{P}_{\mathbf{X}} = \mathbf{X}^{\top}$$

Maintenant,

$$\begin{aligned}\mathbf{X}^{\top}(\mathbf{Y} - \mathbf{P}_{\mathbf{X}}\mathbf{Y}) &= \mathbf{X}^{\top}\mathbf{Y} - \mathbf{X}^{\top}\mathbf{P}_{\mathbf{X}}\mathbf{Y}\\ &= \mathbf{X}^{\top}\mathbf{Y} - \mathbf{X}^{\top}\mathbf{Y}\\ &= 0\end{aligned}$$

Ainsi, par définition, les résidus $\mathbf{Y} - \mathbf{P}_{\mathbf{X}}\mathbf{Y} \in \mathcal{S}^{\perp}(\mathbf{X})$. Les résidus peuvent s'écrire,

$$\begin{aligned}\hat{\mathbf{U}} &= \mathbf{Y} - \mathbf{P}_{\mathbf{X}}\mathbf{Y}\\ &= (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\\ &= \mathbf{M}_{\mathbf{X}}\mathbf{Y}\end{aligned}$$

où,

$$\begin{aligned}\mathbf{M}_{\mathbf{X}} &= \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}\\ &= \mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\end{aligned}$$

est une matrice de projection dans $\mathcal{S}^{\perp}(\mathbf{X})$.
Les matrices $\mathbf{P}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{X}}$ présentent les propriétés suivantes.

(1) $\mathbf{P}_{\mathbf{X}} + \mathbf{M}_{\mathbf{X}} = \mathbf{I}_n$. Ceci implique, que pour tout $\mathbf{Y} \in \mathbb{R}^n$,

$$\mathbf{Y} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} + \mathbf{M}_{\mathbf{X}}\mathbf{Y}$$

(2) $\mathbf{P}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{X}}$ sont symétriques,

$$\mathbf{P}_{\mathbf{X}}^{\top} = \mathbf{P}_{\mathbf{X}}, \quad \mathbf{M}_{\mathbf{X}}^{\top} = \mathbf{M}_{\mathbf{X}}$$

(3) $\mathbf{P}_{\mathbf{X}}$ et $\mathbf{M}_{\mathbf{X}}$ sont idempotentes,

$$\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}}, \quad \mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{X}} = \mathbf{M}_{\mathbf{X}}$$

En effet,

$$\begin{aligned}\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}} &= \left(\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\right)\left(\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\right)\\ &= \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\\ &= \mathbf{P}_{\mathbf{X}}\end{aligned}$$

de même,

$$\begin{aligned}\mathbf{M_X M_X} &= (\mathbf{I}_n - \mathbf{P_X})(\mathbf{I}_n - \mathbf{P_X}) \\ &= \mathbf{I}_n - 2\mathbf{P_X} + \mathbf{P_X P_X} \\ &= \mathbf{I}_n - \mathbf{P_X} \\ &= \mathbf{M_X}\end{aligned}$$

(4) $\mathbf{P_X}$ et $\mathbf{M_X}$ sont orthogonales,

$$\begin{aligned}\mathbf{P_X M_X} &= \mathbf{P_X}(\mathbf{I}_n - \mathbf{P_X}) \\ &= \mathbf{P_X} - \mathbf{P_X P_X} \\ &= \mathbf{P_X} - \mathbf{P_X} \\ &= 0\end{aligned}$$

Cette propriété implique que $\mathbf{M_X X} = 0$. En effet,

$$\begin{aligned}\mathbf{M_X X} &= (\mathbf{I}_n - \mathbf{P_X})\mathbf{X} \\ &= \mathbf{X} - \mathbf{P_X X} \\ &= \mathbf{X} - \mathbf{X} \\ &= 0\end{aligned}$$

Observons que, dans la discussion ci-dessus, aucune des hypothèses quant au modèle de régression n'ont été utilisées. Étant donné des données, $\mathbf{Y}$ et $\mathbf{X}$, nous pouvons toujours calculer l'estimateur des moindres carrés, indépendamment du processus générateur des données derrière les données. Néanmoins, nous avons besoin d'un modèle(i.e., d'hypothèses) pour pouvoir discuter des propriétés d'un estimateur(e.g., le fait qu'il soit ou non sans biais, etc).

3.2. **Propriétés de $\hat{\sigma}^2$.** Nous avions suggéré précédemment d'estimer σ^2 par,

$$\begin{aligned}\hat{\sigma}^2 &= n^{-1}\sum_{i=1}^{n}(Y_i - X_i^\top \hat{\beta})^2 \\ &= n^{-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= n^{-1}\sum_{i=1}^{n}\hat{U}_i^2 \\ &= n^{-1}\hat{\mathbf{U}}^\top\hat{\mathbf{U}}\end{aligned}$$

Il s'avère cependant que sous les hypothèses usuelles, (C1) - (C4), $\hat{\sigma}^2$ est un estimateur biaisé. Pour le voir, écrivons d'abord,

$$\begin{aligned}\hat{\mathbf{U}} &= \mathbf{M_X Y} \\ &= \mathbf{M_X}(\mathbf{X}\beta + \mathbf{U}) \\ &= \mathbf{M_X U}\end{aligned}$$

où la dernière égalité résulte de ce que $\mathbf{M_X X} = 0$. A présent,

$$\begin{aligned} n\hat{\sigma}^2 &= \hat{\mathbf{U}}^\top \hat{\mathbf{U}} \\ &= \mathbf{U}^\top \mathbf{M_X} \mathbf{M_X} \mathbf{U} \\ &= \mathbf{U}^\top \mathbf{M_X} \mathbf{U} \end{aligned}$$

Étant donné que $\mathbf{U}^\top \mathbf{M_X} \mathbf{U}$ est un scalaire,

$$\mathbf{U}^\top \mathbf{M_X} \mathbf{U} = \mathrm{Tr}\left(\mathbf{U}^\top \mathbf{M_X} \mathbf{U}\right)$$

où $\mathrm{Tr}(A)$ désigne la trace de la matrice A. Nous avons,

$$\begin{aligned} \mathrm{E}\left(\mathbf{U}^\top \mathbf{M_X} \mathbf{U} | \mathbf{X}\right) &= \mathrm{E}\left(\mathrm{Tr}(\mathbf{U}^\top \mathbf{M_X} \mathbf{U}) | \mathbf{X}\right) \\ &= \mathrm{E}\left(\mathrm{Tr}(\mathbf{M_X} \mathbf{U}\mathbf{U}^\top) | \mathbf{X}\right) (\text{car} \quad \mathrm{Tr}(ABC) = \mathrm{Tr}(BCA)) \\ &= \mathrm{Tr}\left(\mathbf{M_X}\, \mathrm{E}\left(\mathbf{U}\mathbf{U}^\top) | \mathbf{X}\right)\right) (\text{car l'opérateur trace et l'espérance sont linéaires }) \\ &= \sigma^2\, \mathrm{Tr}(\mathbf{M_X}) \end{aligned}$$

La dernière égalité résulte de ce que par l'hypothèse (C3), $\mathrm{E}(\mathbf{U}^\top \mathbf{U}) = \sigma^2 \mathbf{I}_n$. Maintenant,

$$\begin{aligned} \mathrm{Tr}(\mathbf{M_X}) &= \mathrm{Tr}\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right) \\ &= \mathrm{Tr}(\mathbf{I}_n) - \mathrm{Tr}\left(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right) \\ &= \mathrm{Tr}(\mathbf{I}_n) - \mathrm{Tr}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\right) \\ &= \mathrm{Tr}(\mathbf{I}_n) - \mathrm{Tr}(\mathbf{I}_K) \\ &= n - K \end{aligned}$$

Il s'en suit que,

$$\mathrm{E}(\hat{\sigma}^2) = \frac{n-K}{n}\sigma^2 \tag{3.2}$$

L'estimateur $\hat{\sigma}^2$ est biaisé, mais le résultat précédent suggère qu'il est aisé de le modifier afin d'obtenir un estimateur sans biais. Pour cela, définissons,

$$\begin{aligned} s^2 &= \hat{\sigma}^2 \frac{n}{n-K} \\ &= (n-K)^{-1} \sum_{i=1}^{n} \hat{U}_i^2 \end{aligned}$$

il résulte de (3.2) que,

$$\mathrm{E}(s^2) = \sigma^2$$

3.3. **Régression partitionnée.** Considérons la partition de la matrice des régresseurs, $\mathbf{X}$,

$$\mathbf{X} = (\mathbf{X}_1 \; \mathbf{X}_2)$$

et écrivons le modèle comme suit,

$$\mathbf{Y} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \mathbf{U}$$

où $\mathbf{X}_1$ est une matrice $(n \times K_1)$, $\mathbf{X}_2$ est une matrice $(n \times K_2)$, $K_1 + K_2 = K$, et,

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

β_1 et β_2 étant des vecteurs de paramètres, respectivement, $(K_1 \times 1)$ et $(K_2 \times 1)$. Partant de cette décomposition du modèle de régression concentrons nous sur un groupe de variables et leurs paramètres correspondants, par exemple $\mathbf{X}_1$ et β_1.

Soit l'estimateur des moindres carrés de β,

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

Nous pouvons écrire la version suivante des équations normales,

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

comme suit,

$$\begin{pmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^\top \mathbf{Y} \\ \mathbf{X}_2^\top \mathbf{Y} \end{pmatrix}$$

On peut obtenir des expressions pour $\hat{\beta}_1$ et $\hat{\beta}_2$ par inversion de la matrice partitionnée à gauche de l'équation ci-dessus. Alternativement, définissons $\mathbf{M}_2$ comme la matrice de projection sur l'espace orthogonal à l'espace $\mathcal{S}(\mathbf{X}_2)$,

$$\mathbf{M}_2 = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top$$

alors,

$$\hat{\beta}_1 = (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{Y} \tag{3.3}$$

Pour montrer cela, commençons par écrire,

$$\mathbf{Y} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \hat{\mathbf{U}} \tag{3.4}$$

Notons que par construction,

$$\begin{aligned} \mathbf{M}_2\hat{\mathbf{U}} &= \hat{\mathbf{U}} (\hat{\mathbf{U}} \text{ est orthogonal à } \mathbf{X}_2) \\ \mathbf{M}_2\mathbf{X}_2 &= 0 \\ \mathbf{X}_1^\top \hat{\mathbf{U}} &= 0 \\ \mathbf{X}_2^\top \hat{\mathbf{U}} &= 0 \end{aligned}$$

Substituons l'équation (3.4) dans la partie droite de l'équation (3.3),

$$\begin{aligned} \left(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{M}_2 \mathbf{Y} &= \left(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{M}_2 \left(\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \hat{\mathbf{U}}\right) \\ &= \left(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1\hat{\beta}_1 \\ &\quad + \left(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \hat{\mathbf{U}} \ (\text{car } \mathbf{M}_2\mathbf{X}_2 = 0 \text{ et } \mathbf{M}_2\hat{\mathbf{U}} = \hat{\mathbf{U}}) \\ &= \hat{\beta}_1 \end{aligned}$$

Étant donné que $\mathbf{M}_2$ est symétrique et idempotente, on peut écrire,

$$\begin{aligned}\hat{\beta}_1 &= \left((\mathbf{M}_2\mathbf{X}_1)^\top(\mathbf{M}_2\mathbf{X}_1)\right)^{-1}(\mathbf{M}_2\mathbf{X}_1)^\top(\mathbf{M}_2\mathbf{Y}) \\ &= \left(\tilde{\mathbf{X}}_1^\top\tilde{\mathbf{X}}_1\right)^{-1}\tilde{\mathbf{X}}_1\tilde{\mathbf{Y}}\end{aligned}$$

où,

$$\begin{aligned}\tilde{\mathbf{X}}_1 &= \mathbf{M}_2\mathbf{X}_1 \\ &= \mathbf{X}_1 - \mathbf{X}_2(\mathbf{X}_2^\top\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{X}_1\end{aligned}$$

à savoir les résidus de la régression de $\mathbf{X}_1$ sur $\mathbf{X}_2$. Et où,

$$\begin{aligned}\tilde{\mathbf{Y}} &= \mathbf{M}_2\mathbf{Y} \\ &= \mathbf{Y} - \mathbf{X}_2(\mathbf{X}_2^\top\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{Y}\end{aligned}$$

à savoir les résidus de la régression de $\mathbf{Y}$ sur $\mathbf{X}_2$.

Ainsi, pour obtenir les coefficients de K_1 premiers régresseurs, plutôt que de réaliser la régression avec les $K_1 + K_2 = K$ régresseurs, on peut régresser $\mathbf{Y}$ sur $\mathbf{X}_2$ pour obtenir les résidus $\tilde{\mathbf{Y}}$, régresser $\mathbf{X}_1$ sur $\mathbf{X}_2$ pour obtenir les résidus $\tilde{\mathbf{X}}_1$, et alors régresser $\tilde{\mathbf{Y}}$ sur $\tilde{\mathbf{X}}_1$ pour obtenir $\hat{\beta}_1$. Autrement dit, $\hat{\beta}_1$ décrit l'effet de $\mathbf{X}_1$ une fois que ceux de $\mathbf{X}_2$ ont été contrôlés.

De manière similaire que pour $\hat{\beta}_1$, nous avons pour $\hat{\beta}_2$,

$$\hat{\beta}_2 = (\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{Y}$$

où,

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top$$

Prenons comme exemple le modèle suivant,

$$Y_i = \beta_1 + \beta_2 X_i + U_i, \quad i = 1, 2, ..., n$$

Soit $\mathbf{1}_n$ le vecteur $(n \times 1)$ dont tous les éléments sont le nombre 1, i.e.,

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}$$

La matrice des régresseurs est alors,

$$(\mathbf{1}_n \;\; X) = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{pmatrix}$$

Considérons,

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n^\top \mathbf{1}_n)^{-1}\mathbf{1}_n^\top$$

et,

$$\hat{\beta}_2 = \frac{X^\top \mathbf{M}_1 \mathbf{Y}}{X^\top \mathbf{M}_1 X}$$

Nous avons, $\mathbf{1}_n^\top \mathbf{1}_n = n$, par conséquent,

$$\mathbf{M}_1 = \mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}$$

et,

$$\begin{aligned}
\mathbf{M}_1 X &= X - \mathbf{1}_n \frac{\mathbf{1}_n^\top X}{n} \\
&= X - \bar{X}\mathbf{1}_n \\
&= \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ . \\ . \\ . \\ X_n - \bar{X} \end{pmatrix}
\end{aligned}$$

où,

$$\begin{aligned}
\bar{X} &= \frac{\mathbf{1}_n^\top X}{n} \\
&= n^{-1} \sum_{i=1}^{n} X_i
\end{aligned}$$

Ainsi la matrice $\mathbf{M}_1$ transforme le vecteur X en un vecteur dont les éléments sont les écarts des observations X_i à leur moyenne. Et nous pouvons écrire,

$$\begin{aligned}
\hat{\beta}_2 &= \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}
\end{aligned}$$

3.4. **Qualité de l'ajustement et coefficient de détermination ou R^2.** Écrivons,

$$\begin{aligned}
\mathbf{Y} &= \mathbf{P}_\mathbf{X}\mathbf{Y} + \mathbf{M}_\mathbf{X}\mathbf{Y} \\
&= \hat{\mathbf{Y}} + \hat{\mathbf{U}}
\end{aligned}$$

où par construction,

$$\begin{aligned}
\hat{\mathbf{Y}}^\top \hat{\mathbf{U}} &= (\mathbf{P}_\mathbf{X}\mathbf{Y})^\top (\mathbf{M}_\mathbf{X}\mathbf{Y}) \\
&= \mathbf{Y}^\top \mathbf{P}_\mathbf{X} \mathbf{M}_\mathbf{X} \mathbf{Y} \\
&= 0
\end{aligned}$$

Supposons que le modèle contienne une constante, par exemple la première colonne de la matrice des régresseurs $\mathbf{X}$ est le vecteur unitaire $\mathbf{1}_n$. La *variation totale* dans $\mathbf{Y}$ est,

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} \\ &= (\hat{\mathbf{Y}} + \hat{\mathbf{U}})^\top \mathbf{M}_1 (\hat{\mathbf{Y}} + \hat{\mathbf{U}}) \\ &= \hat{\mathbf{Y}}^\top \mathbf{M}_1 \hat{\mathbf{Y}} + \hat{\mathbf{U}}^\top \mathbf{M}_1 \hat{\mathbf{U}} + 2\hat{\mathbf{Y}}^\top \mathbf{M}_1 \hat{\mathbf{U}}\end{aligned}$$

où $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. Comme le modèle contient une constante,

$$\mathbf{1}_n^\top \hat{\mathbf{U}} = 0$$

et,

$$\mathbf{M}_1 \hat{\mathbf{U}} = \hat{\mathbf{U}}$$

Cependant, $\hat{\mathbf{Y}}^\top \hat{\mathbf{U}} = 0$, et par conséquent,

$$\mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} = \hat{\mathbf{Y}}^\top \mathbf{M}_1 \hat{\mathbf{Y}} + \hat{\mathbf{U}}^\top \hat{\mathbf{U}}$$

ou,

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2 + \sum_{i=1}^{n} \hat{U}_i^2$$

où $\bar{\hat{Y}} = n^{-1} \sum_{i=1}^{n} \hat{Y}_i$. Notons que,

$$\begin{aligned}\bar{Y} &= \frac{\mathbf{1}_n^\top \mathbf{Y}}{n} \\ &= \frac{\mathbf{1}_n^\top \hat{\mathbf{Y}}}{n} + \frac{\mathbf{1}_n^\top \hat{\mathbf{U}}}{n} \\ &= \frac{\mathbf{1}_n^\top \hat{\mathbf{Y}}}{n} \\ &= \bar{\hat{Y}}\end{aligned}$$

Ainsi, la moyenne des Y_i et celle de leurs valeurs ajustées $\hat{Y}_i$ étant égales, nous pouvons écrire,

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{U}_i^2$$

ou,

$$SCT = SCE + SCR$$

où, $SCT := \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ est la *somme des carrés totale*, $SCE := \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ est la *somme des carrés expliqués*, et $SCR := \sum_{i=1}^{n} \hat{U}_i^2$ est la *somme des carrés des résidus*.

Le rapport de la SCE à la SCT est appelé coefficient de détermination [7] ou R^2,

$$\begin{aligned} R^2 &= \frac{SCE}{SCT} \\ &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{\sum_{i=1}^n \hat{U}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{\hat{\mathbf{U}}^\top \hat{\mathbf{U}}}{\mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y}} \end{aligned}$$

3.5. **Propriétés du R^2.**

(1) Le R^2 est borné entre 0 et 1 ainsi que cela est indiqué par sa décomposition. Remarquez néanmoins que ceci n'est plus vrai dans un modèle sans constante, et dans ce cas il est indiqué de ne pas utiliser la définition précédente du R^2. Remarquez aussi que si $R^2 = 1$ alors $\hat{\mathbf{U}}^\top \hat{\mathbf{U}} = 0$, ce qui sera vrai seulement si $\mathbf{Y} \in \mathcal{S}(X)$, i.e., $\mathbf{Y}$ est *exactement* une combinaison linéaire des colonnes de $\mathbf{X}$.

(2) Le R^2 augmente avec le nombre de régresseurs. Pour montrer cette propriété considérons une partition de la matrice des régresseurs $\mathbf{X} = (\mathbf{Z} \ \ \mathbf{W})$. Étudions l'effet d'ajouter $\mathbf{W}$ sur le R^2. Notons,

$$\begin{aligned} \mathbf{P_X} &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \\ \mathbf{P_Z} &= \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \end{aligned}$$

respectivement, la matrice de projection du modèle "complet"(i.e., avec $\mathbf{Z}$ et $\mathbf{W}$), et la matrice de projection du modèle avec uniquement $\mathbf{Z}$ comme matrice des régresseurs. Définissons aussi,

$$\begin{aligned} \mathbf{M_X} &= \mathbf{I}_n - \mathbf{P_X} \\ \mathbf{M_Z} &= \mathbf{I}_n - \mathbf{P_Z} \end{aligned}$$

Observons que comme $\mathbf{Z}$ est une partie de $\mathbf{X}$,

$$\mathbf{P_X Z} = \mathbf{Z}$$

et,

$$\begin{aligned} \mathbf{P_X P_Z} &= \mathbf{P_X Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \\ &= \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \\ &= \mathbf{P_Z} \end{aligned}$$

7. On l'appelle/prononce généralement "R deux".

Par conséquent,

$$\begin{aligned}\mathbf{M_X M_Z} &= (\mathbf{I}_n - \mathbf{P_X})(\mathbf{I}_n - \mathbf{P_Z}) \\ &= \mathbf{I}_n - \mathbf{P_X} - \mathbf{P_Z} + \mathbf{P_X P_Z} \\ &= \mathbf{I}_n - \mathbf{P_X} - \mathbf{P_Z} + \mathbf{P_Z} \\ &= \mathbf{M_X}\end{aligned}$$

Supposons que $\mathbf{Z}$ contienne un vecteur constant de sorte que les deux régressions(la complète et celle sans $\mathbf{W}$) contiennent chacune une constante. Définissons,

$$\hat{\mathbf{U}}_\mathbf{X} = \mathbf{M_X Y}\ ,\ \ \hat{\mathbf{U}}_\mathbf{Z} = \mathbf{M_Z Y}$$

Écrivons,

$$(\hat{\mathbf{U}}_\mathbf{X} - \hat{\mathbf{U}}_\mathbf{Z})^\top(\hat{\mathbf{U}}_\mathbf{X} - \hat{\mathbf{U}}_\mathbf{Z}) = \hat{\mathbf{U}}_\mathbf{X}^\top\hat{\mathbf{U}}_\mathbf{X} + \hat{\mathbf{U}}_\mathbf{Z}^\top\hat{\mathbf{U}}_\mathbf{Z} - 2\hat{\mathbf{U}}_\mathbf{X}^\top\hat{\mathbf{U}}_\mathbf{Z} \geq 0$$

Notons que,

$$\begin{aligned}\hat{\mathbf{U}}_\mathbf{X}^\top\hat{\mathbf{U}}_\mathbf{Z} &= \mathbf{Y}^\top\mathbf{M_X M_Z Y} \\ &= \mathbf{Y}^\top\mathbf{M_X Y} \\ &= \hat{\mathbf{U}}_\mathbf{X}^\top\hat{\mathbf{U}}_\mathbf{X}\end{aligned}$$

d'où,

$$\hat{\mathbf{U}}_\mathbf{Z}^\top\hat{\mathbf{U}}_\mathbf{Z} \geq \hat{\mathbf{U}}_\mathbf{X}^\top\hat{\mathbf{U}}_\mathbf{X}$$

(3) Le R^2 indique la part de la variation de $\mathbf{Y}$ dans l'échantillon qui est expliquée par $\mathbf{X}$. Cependant notre objectif n'est pas d'expliquer des variations dans l'échantillon mais celle de la population (dont est tiré l'échantillon). Il en résulte qu'un R^2 élevé n'est pas nécessairement un indicateur d'un bon modèle de régression et un R^2 faible n'est pas non plus un argument en défaveur du modèle considéré.

(4) Il est toujours possible de trouver une matrice de régresseurs $\mathbf{X}$ pour laquelle $R^2 = 1$, il suffit de prendre n vecteurs linéairement indépendants. En effet, un tel ensemble de vecteurs génère tout l'espace $\mathbb{R}^n$ de sorte que tout vecteur $\mathbf{Y} \in \mathbb{R}^n$ peut s'écrire comme une combinaison linéaire exacte des colonnes de $\mathbf{X}$.

3.6. R^2 **ajusté.** Étant donné que le R^2 augmente avec le nombre de régresseurs, une mesure alternative pour juger de la qualité de la régression est le R^2 *ajusté*,

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{n-1}{n-K}(1 - R^2) \\ &= 1 - \frac{\hat{\mathbf{U}}^\top\hat{\mathbf{U}}/(n-K)}{\mathbf{Y}^\top\mathbf{M}_1\mathbf{Y}/(n-1)}\end{aligned}$$

Le R^2 ajusté diminue la qualité de ajustement lorsque le nombre de régresseurs augmente relativement au nombre d'observations de sorte que $\bar{R}^2$ peut diminuer avec le nombre de régresseurs. Cependant il n'y a pas vraiment d'argument fort pour utiliser une telle mesure de l'ajustement.

4. INTERVALLES DE CONFIANCE DANS LE MODÈLE DE RÉGRESSION NORMAL

Dans cette section nous allons considérer le modèle de régression normal défini par les hypothèse C1-C5.

L'*estimateur ponctuel* de β, $\hat{\beta}$, n'est pas très informatif dans la mesure où $\mathrm{P}(\hat{\beta} = \beta) = 0$. Pour cette raisons nous allons nous intéresser ici à des intervalles(régions) aléatoires qui présentent la propriété d'inclure la vraie valeur du paramètre avec une certaine probabilité spécifiée $(1 - \alpha)$, où α est un nombre "petit" appelé *niveau de confiance*. Traditionnellement, les valeurs suivantes de α sont retenues, 0.01, 0.05, 0.10. Un intervalle de confiance avec une probabilité $(1 - \alpha)$ de couvrir β est noté $\mathrm{CI}_{1-\alpha}$.

4.1. **Cas scalaire.**

On cherche à construire un intervalle de confiance pour le paramètre β_1 dans la régression partitionnée,

$$\mathbf{Y} = \beta_1 \mathbf{X}_1 + \mathbf{X}_2 \beta_2 + \mathbf{U}$$

où à présent $\mathbf{X}_1$ est un vecteur $(n \times 1)$ contenant les valeurs observées du premier régresseur. L'estimateur des moindres carrés de β_1 est,

$$\hat{\beta}_1 = \frac{\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{Y}}{\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1}$$

où $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2$. Une méthode pour construire un intervalle de confiance consiste à considérer des intervalles symétriques autour de l'estimateur ponctuel,

$$\mathrm{CI}_{1-\alpha} = \left[\hat{\beta}_1 - c, \hat{\beta}_1 + c\right] \tag{4.1}$$

Comme $\hat{\beta}_1$ est une fonction de l'échantillon aléatoire, l'intervalle de confiance donné dans (4.1) l'est aussi. Le problème maintenant est de choisir c tel que,

$$\mathrm{P}(\beta_1 \in \mathrm{CI}_{1-\alpha} | \mathbf{X}) = 1 - \alpha$$

où $\mathbf{X} = (\mathbf{X}_1 \ \ \mathbf{X}_2)$. Pour choisir c, nous avons besoin de connaître la distribution de $\hat{\beta}_1|\mathbf{X}$. Sous les hypothèses C1-C5,

$$\hat{\beta}_1|\mathbf{X} \sim \mathcal{N}\left(\beta_1, \sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)\right)$$

et par conséquent,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}}|\mathbf{X} \sim \mathcal{N}(0,1) \tag{4.2}$$

Pour montrer ce résultat, notons que $\hat{\beta}_1$ est un estimateur linéaire, et écrivons $\hat{\beta}_1 = \beta_1 + (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U})/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)$.
Soit z_τ le quantile τ de la distribution normale standard ; autrement dit, si $Z \sim \mathcal{N}(0,1)$,

$$\mathrm{P}(Z \leq z_\tau) = \tau$$

Par exemple, pour $\tau = 0.5$ nous avons la médiane,

$$\mathrm{P}(Z \leq z_{0.5}) = 0.5$$

Notons qu'étant donné que la distribution normale standard est symétrique autour de zéro, nous avons,

$$z_\alpha = -z_{(1-\alpha)}$$

et par conséquent,

$$\mathrm{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

Par exemple, pour $\alpha = 0.05$, $z_{1-0.05/2} = z_{0.975} = 1.96$, et $z_{0.025} = -1.96$.

4.2. σ^2 **est connu.** Supposons pour le moment que σ^2 soit connu et que ce faisant nous puissions calculer la variance de $\hat{\beta}_1$(et non pas un estimateur). Posons,

$$c = z_{1-\alpha/2}\sqrt{\mathrm{Var}\left(\hat{\beta}_1|\mathbf{X}\right)} = z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}$$

Montrons maintenant que,

$$\mathrm{P}\left(\beta_1 \in \left[\hat{\beta}_1 - z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}, \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}\right]|\mathbf{X}\right) = 1 - \alpha$$

En effet,

$$\begin{aligned}
&\mathrm{P}\left(\hat{\beta}_1 - z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} \leq \beta_1 \leq \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}|\mathbf{X}\right) \\
&= \mathrm{P}\left(-z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} \leq \beta_1 - \hat{\beta}_1 \leq z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}|\mathbf{X}\right) \\
&= \mathrm{P}\left(-z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} \leq \hat{\beta}_1 - \beta_1 \leq z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}|\mathbf{X}\right) \\
&= \mathrm{P}\left(-z_{1-\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \leq z_{1-\alpha/2}|\mathbf{X}\right) \qquad (4.3)
\end{aligned}$$

Le résultat découle de (4.2), (4.3), et de la définition de $z_{1-\alpha/2}$.

4.3. σ^2 **est inconnu.** La construction précédente de $\mathrm{CI}_{1-\alpha}$ reposait sur l'hypothèse que σ^2 était connu. Lorsque ce n'est pas le cas, on peut néanmoins suivre une approche similaire à la précédente mais en remplaçant dans un premier temps σ^2 par son estimateur,

$$s^2 = \hat{\mathbf{U}}^\top \hat{\mathbf{U}}/(n-K)$$

Cependant $(\hat{\beta}_1 - \beta_1)/\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}$ n'est pas normalement distribué car c'est une fonction non-linéaire des termes aléatoires $\hat{\beta}_1$ et s^2. Il s'en suit que nous ne pouvons pas utiliser les quantiles de la distribution normale pour la construction des intervalles de confiance.
En fait, il s'avère que,

$$\frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}}|\mathbf{X} \sim t_{n-K} \qquad (4.4)$$

Rappelons que la distribution t_{n-K} est définie comme suit,

$$Z/\sqrt{V/(n-K)}$$

où $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_{n-K}$, et Z et V sont indépendantes.
Écrivons,

$$\begin{aligned}\frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} &= \left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right) / \frac{s^2}{\sigma^2} \\ &= \left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right) / \sqrt{\frac{\hat{\mathbf{U}}^\top \hat{\mathbf{U}}}{\sigma^2}/(n-K)} \end{aligned} \tag{4.5}$$

Nous savons déjà que, dans l'expression précédente, $\hat{\beta}_1 - \beta_1)/\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} | \mathbf{X} \sim \mathcal{N}(0,1)$.
Nous allons montrer maintenant que conditionnellement à $\mathbf{X}$,

$$\frac{\hat{\mathbf{U}}^\top \hat{\mathbf{U}}}{\sigma^2} | \mathbf{X} \sim \chi^2_{n-K} \tag{4.6}$$

Pour cela nous avons besoin du résultat suivant,

Lemme 4.1. *Supposons que le vecteur* $(n \times 1)$ $U \sim \mathcal{N}(0, \mathbf{I}_n)$. *Soit* A *une matrice* $(n \times n)$ *symétrique et idempotente avec* $\text{Rang}(A) = r \leq n$. *Alors* $U^\top A U \sim \chi^2_r$.

Démonstration. Il suffit de montrer que $U^\top A U = \sum_i^r Z_i^2$, où Z_i sont des variables aléatoires iid $\mathcal{N}(0,1)$.
Étant donné que A est symétrique, nous pouvons écrire,

$$A = C \Lambda C^\top$$

où Λ est une matrice diagonale $(n \times n)$ des valeurs propres de A, et $C^\top C = \mathbf{I}_n$. Étant donné que A est idempotente,

$$A = AA$$

et,

$$\begin{aligned} C\Lambda C^\top &= (C\Lambda C^\top)(C\Lambda C^\top) \\ &= C\Lambda^2 C^\top \end{aligned}$$

par conséquent,

$$\Lambda = \Lambda^2$$

ce qui implique que toutes les valeurs propres de Λ sont soit zéro, soit un. Comme le rang d'une matrice est égal au nombre de ses valeurs propres non-nulles, il doit y avoir r valeurs

propres non-nulles, λ_i, dans Λ,

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & . & . & . & 0 \\ 0 & \lambda_2 & 0 & . & . & 0 \\ . & 0 & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \lambda_n \end{pmatrix}$$

Définissons,

$$Z = C^\top U$$

Comme $U \sim \mathcal{N}(0, \mathbf{I}_n)$, nous avons,

$$\begin{aligned} Z &\sim \mathcal{N}(0, C^\top C) \\ &\sim \mathcal{N}(0, \mathbf{I}_n) \end{aligned}$$

Finalement, nous avons,

$$\begin{aligned} U^\top A U &= Z^\top \Lambda Z \\ &= \sum_{i=1}^{n} \lambda_i Z_i^2 \end{aligned}$$

Le résultat découle de ce que les Z_i sont iid distribuées $\mathcal{N}(0, 1)$, et qu'il y a r valeurs propres égales à un, et $n - r$ valeurs propres égales à zéro. □

A présent pour montrer (4.6), écrivons,

$$\frac{\hat{\mathbf{U}}^\top \hat{\mathbf{U}}}{\sigma^2} = \left(\frac{\mathbf{U}}{\sigma^2}\right)^\top \mathbf{M}_\mathbf{X} \left(\frac{\mathbf{U}}{\sigma^2}\right) \tag{4.7}$$

où,

$$\mathbf{M}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

Par l'hypothèse C5,

$$\frac{\mathbf{U}}{\sigma} | \mathbf{X} \sim \mathcal{N}(0, \mathbf{I}_n) \tag{4.8}$$

Étant donné que $\mathbf{M}_\mathbf{X}$ est symétrique et idempotente, ses valeurs propres sont soit zéro ou un. Par conséquent,

$$\begin{aligned} \text{Rang}(\mathbf{M}_\mathbf{X}) &= \text{Tr}(\mathbf{M}_\mathbf{X}) \\ &= n - K \end{aligned} \tag{4.9}$$

Le résultat dans (4.6) découle de (4.7), (4.8), (4.9), et du lemme 4.1.
Finalement, montrons que $\hat{\beta}_1 - \beta_1$ et $\hat{\mathbf{U}}^\top \hat{\mathbf{U}}$ dans (4.5) sont indépendants conditionnellement à $\mathbf{X}$. Écrivons,

$$\begin{aligned} \hat{\beta}_1 - \beta_1 &= (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}) / (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1) \\ \hat{\mathbf{U}}^\top \hat{\mathbf{U}} &= \mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U} \end{aligned}$$

Il suffit de montrer l'indépendance de $\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}$ et $\mathbf{M}_\mathbf{X}\mathbf{U}$. Comme $\hat{\beta}_1$ est une fonction de $\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}$, et $\hat{\mathbf{U}}^\top \hat{\mathbf{U}}$ est une fonction de $\mathbf{M}_\mathbf{X}\mathbf{U}$, l'indépendance de $\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}$ et $\mathbf{M}_\mathbf{X}\mathbf{U}$ implique l'indépendance de $\hat{\beta}_1$ et $\hat{\mathbf{U}}^\top \hat{\mathbf{U}}$. Premièrement, montrons que les termes ne sont pas corrélés,

$$\begin{aligned}
\mathrm{Cov}(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}, \mathbf{M}_\mathbf{X}\mathbf{U}|\mathbf{X}) &= \mathrm{E}(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}\mathbf{U}^\top \mathbf{M}_\mathbf{X}|\mathbf{X}) \\
&= \mathbf{X}_1^\top \mathbf{M}_2 \,\mathrm{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X})\mathbf{M}_\mathbf{X} \\
&= \mathbf{X}_1^\top \mathbf{M}_2 (\sigma^2 \mathbf{I}_n)\mathbf{M}_\mathbf{X} \\
&= \sigma^2 \mathbf{X}_1^\top \mathbf{M}_2 \mathbf{M}_\mathbf{X} \\
&= \sigma^2 \mathbf{X}_1^\top \mathbf{M}_\mathbf{X}(\text{ voir séction précédente}) \\
&= 0
\end{aligned}$$

Dans la mesure où $\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}$ et $\mathbf{M}_\mathbf{X}\mathbf{U}$ sont des fonctions linéaires de $\mathbf{U}$, elles sont normalement distribuées conditionnellement à $\mathbf{X}$. Étand donné qu'elles ne sont pas corrélées, la normalité implique qu'elles sont indépendantes. En conséquence, $\hat{\beta}_1 - \beta_1$, fonction de $\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}$, et $\hat{\mathbf{U}}^\top \hat{\mathbf{U}}$ sont aussi indépendants.
Nous avons montré (4.4). En conséquence, en construisant des intervalles de confiance, si l'on remplace l'inconnue σ^2 par s^2, on doit remplacer $z_{1-\alpha/2}$ par les quantiles de la t distribution, $t_{n-K,1-\alpha/2}$,

$$\mathrm{CI}_{1-\alpha} = \left[\hat{\beta}_1 - t_{n-K,1-\alpha/2}\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}, \hat{\beta}_1 + t_{n-K,1-\alpha/2}\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}\right]$$

L'expression $s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)$ qui apparaît dans l'équation ci-dessus est la variance estimé de $\hat{\beta}_1$,

$$\hat{\mathrm{Var}}(\hat{\beta}_1|\mathbf{X}) = s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)$$

Ainsi, on construit un intervalle de confiance de niveau α pour β_k, $k = 1, 2, ..., K$, comme suit,

$$\mathrm{CI}_{1-\alpha} = \left[\hat{\beta}_1 - t_{n-K,1-\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_1|\mathbf{X})}, \hat{\beta}_1 + t_{n-K,1-\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_1|\mathbf{X})}\right] \tag{4.10}$$

4.4. **Cas vectoriel.** Supposons que l'on s'intéresse aux vecteur des paramètres $\beta = (\beta_1, \beta_2, ..., \beta_K)^\top$. L'équation (4.10) décrit comment construire des intervalles de confiance "individuels" pour les éléments de β. Ces intervalles concernent les distributions marginales des éléments de β, et leur simple combinaison ne produit pas un ensemble qui inclue tout le vecteur β avec une probabilité souhaitée. Dans cette partie, nous considérons la construction de régions aléatoires qui incluent β avec une certaine probabilité pré-spécifiée $1 - \alpha$. Nous conservons la notation $\mathrm{CI}_{1-\alpha}$, malgré le fait que $\mathrm{CI}_{1-\alpha}$ est maintenant un sous-ensemble de $\mathbb{R}^K$.
Ce qui suit est une approche simple et conventionnelle pour construire des régions de confiance. Nous cherchons une région de confiance $\mathrm{CI}_{1-\alpha} = \{b \in \mathbb{R}^K\}$ tel que $\mathrm{P}(\beta \in$

$\text{CI}_{1-\alpha}|\mathbf{X}) = 1-\alpha$. Considérons une forme quadratique par rapport à $(\hat{\beta}-\beta)$,

$$\begin{aligned}(\hat{\beta}-\beta)^\top \left(\hat{\text{Var}}(\hat{\beta}|\mathbf{X})\right)^{-1}(\hat{\beta}-\beta)/K &= (\hat{\beta}-\beta)^\top \left(s^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1}(\hat{\beta}-\beta)/K \\ &= \frac{(\hat{\beta}-\beta)^\top \left(\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1}(\hat{\beta}-\beta)/K}{s^2/\sigma^2} \\ &= \frac{(\hat{\beta}-\beta)^\top \left(\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1}(\hat{\beta}-\beta)/K}{\left(\frac{\hat{\mathbf{U}}^\top\hat{\mathbf{U}}}{\sigma^2}\right)/(n-K)} \end{aligned} \tag{4.11}$$

Montrons maintenant que l'expression dans (4.11) possède une distribution $F_{K,n-K}$ conditionnellement à $\mathbf{X}$.
La distribution $F_{K,n-K}$ est définie comme la distribution de,

$$\frac{V/K}{W/(n-K)}$$

où $V \sim \chi^2_K$, et $W \sim \chi^2_{n-K}$ sont indépendantes. De la discussion dans la partie précédente nous savons que, $\hat{\mathbf{U}}^\top\hat{\mathbf{U}}/\sigma^2|\mathbf{X} \sim \chi^2_{n-K}$ qui est indépendant du numérateur dans (4.11). Il résulte de cela, que nous devons montrer que

$$(\hat{\beta}-\beta)^\top \left(\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1}(\hat{\beta}-\beta)|\mathbf{X} \sim \chi^2_K \tag{4.12}$$

Pour cela nous avons besoin du lemme suivant,

Lemme 4.2. *Supposons que le vecteur* $(K\times 1)$*,* $U \sim \mathcal{N}(0,\Sigma)$*, où* Σ *est une matrice définie positive de variances-covariances. Alors,* $U^\top\Sigma^{-1}U \sim \chi^2_K$*.*

Démonstration. Comme Σ est symétrique, $\Sigma = C\Lambda C^\top$, où Λ est une matrice diagonale des valeurs propres de Σ sur sa diagonale, et $C^\top C = CC^\top = \mathbf{I}_K$. Comme Σ est définie positive, ses valeurs propres sont positives, et par conséquent, $\Lambda^{1/2}$ peut être définie comme,

$$\Lambda^{1/2} = \begin{pmatrix} \lambda_1^{1/2} & 0 & . & . & . & 0 \\ 0 & \lambda_2^{1/2} & 0 & . & . & 0 \\ . & 0 & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \lambda_K^{1/2} \end{pmatrix}$$

et $\Lambda^{-1/2}$ peut être définie comme,

$$\Lambda^{-1/2} = \begin{pmatrix} \lambda_1^{-1/2} & 0 & . & . & . & 0 \\ 0 & \lambda_2^{-1/2} & 0 & . & . & 0 \\ . & 0 & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \lambda_K^{-1/2} \end{pmatrix}$$

A présent, étant donné que $C\Lambda^{-1}C^\top C\Lambda C^\top = \mathbf{I}_K$, nous avons,

$$\Sigma^{-1} = C\Lambda^{-1}C^\top$$

Définissons maintenant,

$$\Sigma^{1/2} = C\Lambda^{1/2}C^\top, \text{ et } \Sigma^{-1/2} = C\Lambda^{-1/2}C^\top$$

Nous avons que $(\Sigma^{1/2})^\top = \Sigma^{1/2}$ et $(\Sigma^{-1/2})^\top = \Sigma^{-1/2}$(symétrie). En outre,

$$\Sigma^{1/2}\Sigma^{1/2} = C\Lambda^{1/2}C^\top C\Lambda^{1/2}C^\top = C\Lambda^{1/2}\Lambda^{1/2}C^\top = C\Lambda C^\top = \Sigma$$

$$\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = C\Lambda^{-1/2}C^\top C\Lambda C^\top C\Lambda^{1/2}C^\top = C\Lambda^{-1/2}\Lambda\Lambda^{-1/2}C^\top = CC^\top = \mathbf{I}_K$$

La matrice $\Sigma^{1/2}$ est appelée racine carrée symétrique d'une matrice, et $\Sigma^{-1/2}$ est racine carrée symétrique négative. Définissons le vecteur$(K \times 1)$,

$$V = \Sigma^{-1/2}U$$

de sorte que

$$U^\top\Sigma^{-1}U = V^\top V \tag{4.13}$$

Étant donné que $U \sim \mathcal{N}(0, \Sigma)$, et que V est une transformation linéaire de U, nous avons,

$$\begin{aligned} V &\sim \mathcal{N}\left(0, \sigma^{-1/2}\operatorname{Var}(U)\Sigma^{-1/2}\right) \\ &= \mathcal{N}\left(0, \Sigma^{-1/2}\Sigma\Sigma^{-1/2}\right) \\ &= \mathcal{N}(0, \mathbf{I}_K) \end{aligned}$$

Ainsi, en raison de (4.13) et de la définition du χ^2_K,

$$U^\top\Sigma^{-1}U = V^\top V = \sum_{k=1}^{K} V_k^2 \sim \chi^2_K$$

□

Le résultat dans (4.12) découle du lemme 4.2. En conséquence,

$$\frac{(\hat\beta - \beta)^\top \left(s^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1}(\hat\beta - \beta)}{K}|\mathbf{X} \sim F_{K,n-K}$$

Soit, $F_{K,n-K,\tau}$ le quantile τ de la distribution F. La région de confiance de niveau α se construit comme suit,

$$\text{CI}_{1-\alpha} = \left\{b \in \mathbb{R}^K : (\hat\beta - b)^\top \left(s^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1}(\hat\beta - b)/K \leq F_{K,n-K,1-\alpha}\right\}$$

La discussion précédente implique que,

$$\begin{aligned} \text{P}(\beta \in \text{CI}_{1-\alpha}|\mathbf{X}) &= \text{P}\left((\hat\beta - \beta)^\top \left(s^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1}(\hat\beta - \beta)/K \leq F_{K,n-K,1-\alpha}|\mathbf{X}\right) \\ &= 1 - \alpha \end{aligned}$$

Remarque 1. La région/intervalle de confiance $\text{CI}_{1-\alpha}$ est une fonction de l'échantillon $\{(Y_i, X_i)\}_{i=1}^n$, et il est ce faisant aléatoire, ce qui nous permet de parler de la probabilité que $\text{CI}_{1-\alpha}$ contienne la vraie valeur de β. D'un autre côté, la réalisation de $\text{CI}_{1-\alpha}$ n'est pas aléatoire. Une fois que l'intervalle de confiance est calculé pour des observations données, il n'y a plus de sens à parler de la probabilité qu'il inclue β. C'est soit zéro, soit un.

5. TESTS D'HYPOTHÈSES DANS LE MODÈLE DE RÉGRESSION LINÉAIRE NORMAL

5.1. **Concepts de base.** Nous poursuivons notre discussion sur le modèle de régression linéaire normal, i.e., le modèle défini par les hypothèses C1-C5. Soit $\theta \in \Theta \subset \mathbb{R}^d$ un paramètre d'intérêt. Par exemple,

- Le paramètre β_k associé à un régresseur X_k, $k = 1, 2, ..., K$. Dans ce cas, $\theta = \beta_k$, $d = 1$, $\Theta = \mathbb{R}$.
- Un vecteur de L coefficients avec dans ce cas $\theta = (\beta_1, \beta_2, ..., \beta_L)^\top$, $d = L$, $\Theta = \mathbb{R}^L$.
- La variance des erreurs avec dans ce cas, $\theta = \sigma^2$, $d = 1$, $\Theta = \mathbb{R}_{++}$.

Une hypothèse statistique est une proposition concernant θ. Usuellement, on considère deux hypothèses concurrentes et nous voulons tirer une conclusion, sur la base des données observées, quant à savoir laquelle est vraie. Soit $\Theta_0 \subset \Theta$ et $\Theta_1 \subset \Theta$, tels que $\Theta_0 \cap \Theta_1 = \emptyset$, et $\Theta_0 \cup \Theta_1 = \Theta$. Les deux hypothèses concurrentes sont,

- L'*hypothèse nulle* $H_0 : \theta \in \Theta_0$. Il s'agit de l'hypothèse tenue comme vraie à moins que les données fournissent suffisamment d'information pour la rejeter.
- L'*hypothèse alternative* $H_1 : \theta \in \Theta_1$. Il s'agit de l'hypothèse contre laquelle l'hypothèse nulle est testée. Elle sera tenue comme vraie si l'on arrive à la conclusion que l'hypothèse nulle est fausse.

Notons que les sous-ensembles Θ_0 et Θ_1 sont choisis par l'analyste et sont donc connus. Notons aussi que les deux hypothèses H_0 et H_1 doivent être *disjointes*. Leur union correspond à une hypothèse implicite qui concerne l'espace des valeurs que θ peut prendre. Par exemple, quand $\Theta = \mathbb{R}$ on peut considérer $\Theta_0 = \{0\}$, et $\Theta_1 = \mathbb{R}\backslash\{0\}$. Un autre exemple est $\Theta_0 = (-\infty, 0]$, et $\Theta_1 = (0, \infty)$.
Lorsque Θ_0 contient un seul élément, on dit que $H_0 : \theta \in \Theta_0$ est une *hypothèse simple*. Autrement, on dit que H_0 est une *hypothèse composée*. De manière similaire H_1 peut être simple ou composée selon que Θ_1 contient ou non un seul élément.
L'analyste doit choisir entre H_0 et H_1. La *règle de décision* qui va conduire à accepter ou rejeter H_0 s'appuie sur une *statistique de test* qui est une fonction des données(de $\mathbf{X}$ et $\mathbf{Y}$ dans le cas d'un modèle de régression). Soit $S \in \mathcal{S}$ une statistique de test et l'ensemble des valeurs qu'elle peut prendre. Une règle de décision est définie comme une partition de $\mathcal{S}$ en une *région d'acceptation* $\mathcal{A}$ et une *région de rejet*(ou région critique) $\mathcal{R}$. Ces deux régions doivent être disjointes(i.e., $\mathcal{A} \cap \mathcal{R} = \emptyset$) et leur union doit être égale à l'ensemble des valeurs possible pour la statistique S(i.e., $\mathcal{A} \cup \mathcal{R} = \mathcal{S}$). On rejette H_0 lorsque la statistique de test prend des valeurs dans la région de rejet : $S \in \mathcal{R}$. Ainsi les tests peuvent être décrits par leurs règles de décision : rejeter H_0 quand $S \in \mathcal{R}$.
Deux types d'erreurs peuvent être faits,

- L'*erreur de type 1* est de rejeter H_0 alors que H_0 est vraie.
- L'*erreur de type 2* est d'accepter H_0 alors que H_1 est vraie.

Les probabilités d'erreur de type 1 et de type 2 peuvent être décrites en utilisant la *fonction puissance*. Soit un test qui s'appuie sur S et rejette H_0 quand $S \in \mathcal{R}$. La fonction de puissance de ce test est définie comme,

$$\pi(\theta) = \mathrm{P}_\theta(S \in \mathcal{R})$$

où $P_\theta(.)$ est la probabilité qui doit être calculée sous l'hypothèse que la vraie valeur du paramètre est θ. En conséquence, la fonction de puissance d'un test donne la probabilité de rejeter H_0 pour chaque valeur possible de θ. La plus grande probabilité d'erreur de type 1(i.e., rejeter H_0 quand elle est vraie) est,

$$\sup_{\theta\in\Theta_0}\pi(\theta) = \sup_{\theta\in\Theta_0} P_\theta(S \in \mathcal{R}) \tag{5.1}$$

L'expression ci-dessus est aussi appelée *taille du test*. Lorsque H_0 est simple, i.e. $\Theta = \{\theta_0\}$, la taille peut être simplement calculée comme $\pi(\theta_0) = P_{\theta_0}(S \in \mathcal{S})$.
La probabilité d'erreur de type 2(i.e., accepter H_0 quand elle est fausse) est,

$$1 - \pi(\theta) = 1 - P_\theta(S \in \mathcal{R}) \text{ for } \theta \in \Theta_1 \tag{5.2}$$

Typiquement, Θ_1 possède plusieurs éléments, et par conséquent la probabilité d'erreur de type 2 dépend de la vraie valeur de θ. Nous souhaiterions avoir des probabilités d'erreur de type 1 et de type 2 aussi petites que possible, mais malheureusement, comme cela apparaît dans (5.1) et (5.2), elles sont inversement reliées. Pour réduire la probabilité d'erreur de type 1 on doit faire en sorte de réduire $\mathcal{R}$. Cela va cependant augmenter la probabilité d'erreur de type 2.

Définition 1. Un test avec une fonction de puissance $\pi(\theta)$ est dit de niveau α si $\sup_{\theta\in\Theta_0}\pi(\theta) \leq \alpha$.
On dit que le test est de taille α si $\sup_{\theta\in\Theta_0}\pi(\theta) = \alpha$.

Remarquons que des tests de taille α sont des tests de niveau α. On considère qu'un test est valide si c'est un test de niveau α pour un $\alpha \in (0, 1)$ pré-sélectionné, où α est appelé *niveau de significativité* du test. Typiquement, le niveau de significativité est choisi comme un nombre petit et proche de zéro, par exemple, $\alpha = 0.01, 0.05, 0.10$ sont de niveaux fréquemment retenus.
Ce qui suit sont les étapes d'un test d'hypothèse,

(1) Spécifier H_0 et H_1.

(2) Choisir le niveau de significativité α.

(3) Définir une règle de décision(une statistique de test, et une région de rejet) de sorte que le test correspondant soit un test de niveau α.

(4) Exécuter le test.

La décision dépend des niveau de significativité. Il est facile de rejeter l'hypothèse nulle pour des valeurs grandes de α, étant donné qu'elles correspondent à de grandes régions de rejet. Étant donné les données, le plus petit niveau de significativité pour lequel l'hypothèse nulle peut être rejetée par un test est appelé p-value. Plutôt que de reporter les résultats des tests(acceptation ou rejet) pour un α spécifique, il est courant de reporter les p-values. Les étapes du test sont alors,

(1) Spécifier H_0 et H_1.

(2) Définir un test.

(3) Calculer la p-value.

(4) H_0 est rejetée pour toutes les valeurs de α plus grandes que la p-value.

La *puissance d'un test* de fonction de puissance $\pi(\theta)$ est définie comme,

$$\pi(\theta) \text{ pour } \theta \in \Theta_1$$

Étant donné deux tests de niveaux α, nous devrions préférer le test le plus puissant. On dit qu'un test de niveau α et de fonction de puissance $\pi_1(\theta)$ est *uniformément plus puissant* qu'un test de niveau α et de fonction de puissance $\pi_2(\theta)$ si $\pi(\theta_1) \geq \pi_2(\theta)$ pour tout $\theta \in \Theta_1$. Comme cela apparaîtra dans ce qui suit, des tests qui s'appuient sur des estimateurs avec des petites variances sont typiquement des tests uniformément plus puissants.

5.2. **Test d'une hypothèse par rapport à un seul coefficient.** Considérons le modèle partitionné vu dans les sections précédentes,

$$\mathbf{Y} = \beta_1 \mathbf{X}_1 + \mathbf{X}_2 \beta_2 + \mathbf{U}$$

où $\mathbf{X}_1$ est le vecteur $(n \times 1)$ d'observations du premier régresseur. Supposons que la variance des erreurs σ^2 soit connue. Soit $\hat{\beta}_1$ l'estimateur des moindres carrés de β_1. Cherchons à tester,

$$\begin{aligned} H_0 &: \beta_1 = \beta_{1,0} \\ H_1 &: \beta_1 \neq \beta_{1,0} \end{aligned} \tag{5.3}$$

Les intervalles de confiance et les tests d'hypothèses sont étroitement liés. En effet, une règle de décision pour un test de niveau α peut reposer sur l'intervalle de confiance $\text{CI}_{1-\alpha}$. L'intervalle de confiance de niveau $1-\alpha$ pour β_1 est,

$$\text{CI}_{1-\alpha} = \left[\hat{\beta}_1 - z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}, \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}\right]$$

Considérons le test suivant,

$$\text{Rejeter } H_0 \text{ si } \beta_{1,0} \notin \text{CI}_{1-\alpha}$$

Dans ce cas la région critique est donnée par le complément de $\text{CI}_{1-\alpha}$. On rejette ainsi H_0 si,

$$\beta_{1,0} \leq \hat{\beta}_1 - z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}$$

ou

$$\beta_{1,0} \geq \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}$$

De manière équivalente, on rejette si,

$$\left| \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right| > z_{1-\alpha/2} \tag{5.4}$$

Un tel test est appelé *bilatéral*, car sous l'hypothèse alternative, la vraie valeur β_1 peut être plus petite ou plus grande que $\beta_{1,0}$.

L'expression à gauche de l'inégalité est une statistique de test. Pour calculer la probabilité de rejet de l'hypothèse nulle supposons que la vraie valeur soit donnée par β_1. Écrivons,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} + \frac{\beta_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \tag{5.5}$$

Nous avons que,

$$\frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} | \mathbf{X} \sim \mathcal{N}\left(\frac{\beta_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}}, 1 \right)$$

Si l'hypothèse nulle est vraie alors $\beta_1 - \beta_{1,0} = 0$ et la statistique de test présente une distribution normale standard. Dans ce cas par définition de $z_{1-\alpha/2}$.

$$\begin{aligned} \mathrm{P}\left(\text{rejeter } H_0 | \mathbf{X}, H_0 \text{ est vraie}\right) &= \mathrm{P}\left(\left| \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right| > z_{1-\alpha/2} \right) \\ &= \alpha \end{aligned}$$

Ainsi, le test suggéré a la taille correcte α. Si l'hypothèse nulle est fausse, la distribution de la statistique de test n'est pas centrée autour de zéro, et l'on verra des taux de rejet supérieurs à α.

La probabilité de rejet est une fonction de la vraie valeur β_1 et dépend de la magnitude du deuxième terme dans (5.5), $\left|\beta_1 - \beta_{1,0}\right| / \sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}$. Supposons par exemple que,

$$\begin{aligned} \beta_{1,0} &= 0 \\ \sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} &= 1 \\ \alpha &= 0.05 (\text{et } z_{1-\alpha/2} = 1.96) \end{aligned}$$

Soit $Z \sim \mathcal{N}(0,1)$. Dans ce cas, la *fonction puissance* du test est,

$$\begin{aligned} \pi(\beta_1) &= \mathrm{P}\left(\left| \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right| > z_{1-\alpha/2} \right) \\ &= \mathrm{P}\left(\left| \frac{\hat{\beta}_1 - \beta_1 + \beta_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right| > 1.96 | \mathbf{X} \right) \\ &= \mathrm{P}\left(|Z + \beta_1| > 1.96 \right) \\ &= \mathrm{P}\left(Z < -1.96 - \beta_1 \right) + \mathrm{P}\left(Z > 1.96 - \beta_1 \right) \end{aligned}$$

Par exemple,

$$\pi(\beta_1) = \begin{cases} 0.52 \text{ pour } \beta_1 = -2 \\ 0.17 \text{ pour } \beta_1 = -1 \\ 0.05 \text{ pour } \beta_1 = 0 \\ 0.17 \text{ pour } \beta_1 = 1 \\ 0.52 \text{ pour } \beta_1 = 2 \end{cases}$$

Dans ce cas la fonction puissance est minimisée en $\beta_1 = \beta_{1,0}$ où $\pi(\beta_1) = \alpha$.
Pour le calcul des $p - values$ considérons l'exemple suivant. Supposons, qu'étant donné des données la statistique de test dans (5.4) soit égale à 1.88. Pour la distribution normale standard $\mathrm{P}(Z > 1.88) = 0.03$. Par conséquent la $p - value$ du test est 0.06. On rejeterait l'hypothèse nulle pour tous les tests avec un niveau de significativité supérieur à 0.06.
Dans le cas où σ_2 est inconnu, on peut tester (5.3) en considérant la t-statistique,

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \\ &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\hat{\beta}_1|\mathbf{X})}} \end{aligned} \tag{5.6}$$

Le test est donné par la règle de décision suivante,

$$\text{Rejeter } H_0 \text{ si } |T| > t_{n-K,1-\alpha/2}$$

Dans ce cas(voir section précédente) sous H_0, $\mathrm{P}(|T| > t_{n-K,1-\alpha/2}|\mathbf{X}, H_0 \text{ est vraie}) = \alpha$.
On peut aussi considérer des tests *unilatéraux*. Dans le cas de ces tests l'hypothèse nulle et l'hypothèse alternative peuvent être spécifiées comme suit,

$$\begin{aligned} H_0 &: \beta_1 \leq \beta_{1,0} \\ H_1 &: \beta_1 > \beta_{1,0} \end{aligned}$$

Notons que dans ce cas, et H_0 et H_1 sont composées, et la probabilité de rejet varie non seulement selon les valeurs de β_1 spécifiées sous H_1 mais aussi selon H_0. Dans ce cas cas un test valide devra satisfaire la condition,

$$\sup_{\beta_1 \leq \beta_{1,0}} \mathrm{P}(\text{rejeter } H_0|\mathbf{X}, \beta_1) \leq \alpha \tag{5.7}$$

i.e., la probabilité maximale de rejeter H_0 quand elle est vraie ne doit pas dépasser α. Soit T telle que définie dans (5.6) et considérons le test suivant(règle de décision) :

$$\text{Rejeter } H_0 \text{ quand } T > t_{n-K,1-\alpha}$$

Sous H_0, nous avons,

$$\begin{aligned}
\mathrm{P}\left(\text{rejeter } H_0 | \beta_1 \leq \beta_{1,0}\right) &= \mathrm{P}\left(T > t_{n-K,1-\alpha} | \mathbf{X}, \beta_1 \leq \beta_{1,0}\right) \\
&= \mathrm{P}\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} > t_{n-K,1-\alpha} | \mathbf{X}, \beta_1 \leq \beta_{1,0} \right) \\
&\leq \mathrm{P}\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} > t_{n-K,1-\alpha} | \mathbf{X}, \beta_1 \leq \beta_{1,0} \right) (\text{car } \beta_1 \leq \beta_{1,0}) \\
&= \alpha(\text{étant donné que } \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} | \mathbf{X} \sim t_{n-K})
\end{aligned}$$

Ainsi, la condition sur la taille (5.7) est satisfaite. Notons qu'étant donné qu'il s'agit d'un test unilatéral, la probabilité d'erreur de type 1 est portée uniquement par la queue droite de la distribution.

5.3. **Test d'une contrainte linéaire simple.** Considérons le modèle de régression linéaire normal défini par les hypothèse C1-C5,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$$

Supposons que l'on souhaite tester,

$$\begin{aligned}
H_0 &: c^\top \beta = r \\
H_1 &: c^\top \beta \neq r
\end{aligned}$$

Dans ce cas c est un vecteur $(K \times 1)$, r est un scalaire, et sous l'hypothèse nulle,

$$c_1\beta_1 + x_2\beta_2 + ... + c_K\beta_K - r = 0$$

Par exemple, en posant $c_1 = 1$, $c_2 = -1$, $c_3 = ... = c_K = 0$, et $r = 0$, nous pouvons tester l'hypothèse que $\beta_1 = \beta_2$.
Pour l'estimateur des moindres carrés nous avons,

$$\hat{\beta} | \mathbf{X} \sim \mathcal{N}\left(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right) \tag{5.8}$$

Alors,

$$\frac{c^\top \hat{\beta} - c^\top \beta}{\sqrt{\sigma^2 c^\top (\mathbf{X}^\top \mathbf{X})^{-1} c}} | \mathbf{X} \sim \mathcal{N}(0,1) \tag{5.9}$$

Par conséquent, sous H_0,

$$\frac{c^\top \hat{\beta} - r}{\sqrt{\sigma^2 c^\top (\mathbf{X}^\top \mathbf{X})^{-1} c}} | \mathbf{X} \sim \mathcal{N}(0,1) \tag{5.10}$$

Considérons la T statistique,

$$\begin{aligned} T &= \frac{c^\top \hat{\beta} - r}{\sqrt{s^2 c^\top (\mathbf{X}^\top \mathbf{X})^{-1} c}} \\ &= \left(\frac{c^\top \hat{\beta} - r}{\sqrt{\sigma^2 c^\top (\mathbf{X}^\top \mathbf{X})^{-1} c}} \right) / \sqrt{\frac{\mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U}}{\sigma^2} / (n-K)} \end{aligned}$$

Sous H_0, le résultat dans (5.10) est vérifié. En outre, conditionnellement à $\mathbf{X}$,

$$\mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U} / \sigma^2 | \mathbf{X} \sim \chi^2_{n-K} \text{ indépendant de } \hat{\beta} \tag{5.11}$$

Par conséquent sous H_0,

$$T|\mathbf{X} \sim t_{n-K}$$

Ainsi, le niveau de significativité α du test bilatéral de $H_0 : c^\top \beta = r$ est donné par,

$$\text{Rejeter } H_0 \text{ si } |T| > t_{n-K,1-\alpha/2}$$

En posant l'élément j de c, $c_j = 1$ et le restant des éléments de c égaux à zéro on obtient le test discuté dans la sous-section précédente,

$$\begin{aligned} H_0 &: \beta_j = r \\ H_1 &: \beta_j \neq r \end{aligned}$$

On rejette H_0 si,

$$\begin{aligned} |T| &= \left| \frac{\hat{\beta}_j - r}{\sqrt{s^2 \left[(\mathbf{X}^\top \mathbf{X})^{-1} \right]_{jj}}} \right| \\ &> t_{n-K,1-\alpha/2} \end{aligned}$$

où $\left[(\mathbf{X}^\top \mathbf{X})^{-1} \right]_{jj}$ est l'élément (j, j) de la matrice $(\mathbf{X}^\top \mathbf{X})^{-1}$.

5.4. **Tests de contraintes linéaires multiples.** Supposons que l'on souhaite tester,

$$\begin{aligned} H_0 &: \mathbf{R}\beta = r \\ H_1 &: \mathbf{R}\beta \neq r \end{aligned}$$

où $\mathbf{R}$ est une matrice $(q \times K)$ est r est un vecteur $(r \times 1)$. Par exemple,

- $\mathbf{R} = \mathbf{I}_K$, $r = 0$. Dans ce cas on teste que $\beta_1 = ... = \beta_K = 0$.
- $\mathbf{R} = \begin{pmatrix} 1 & 1 & 0 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & 0 & . & . & . & 0 \end{pmatrix}$, $r = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Dans ce cas, $H_0 : \beta_1 + \beta_2 = 1, \beta_3 = 0$.

Considérons la F statistique,

$$F = \left(\mathbf{R}\hat{\beta} - r \right)^\top \left(s^2 \mathbf{R} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right)^{-1} \left(\mathbf{R}\hat{\beta} - r \right) / q$$

Nous montrons dans ce qui suit que sous H_0,

$$F|\mathbf{X} \sim F_{q,n-K} \tag{5.12}$$

Premièrement, il résulte de (5.9) que,

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

Alors, sous H_0,

$$\mathbf{R}\hat{\beta} - r|\mathbf{X} \sim \mathcal{N}\left(0, \sigma^2\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)$$

En outre par le lemme (4.2) de la section précédente,

$$\left(\mathbf{R}\hat{\beta} - r\right)^\top \left(\sigma^2\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1} \left(\mathbf{R}\hat{\beta} - r\right) \sim \chi^2_q$$

Le résultat dans (5.12) est alors obtenu car en raison de (5.11) et de la définition de la F distribution. Par conséquent, le test est donné par,

$$\begin{aligned}\text{Rejeter } H_0 \text{ si } F = &\left(\mathbf{R}\hat{\beta} - r\right)^\top \left(s^2\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1} \left(\mathbf{R}\hat{\beta} - r\right)/q \\ &> F_{q,n-K,1-\alpha}\end{aligned}$$

5.5. **Moindres carrés contraints.** Une approche alternative pour le test d'hypothèses s'appuie sur l'estimation contrainte. On peut considérer la perte d'ajustement qui résulte du choix d'un autre valeurs que celles de $\hat{\beta}$ pour coefficients de la régression. Considérons, le problème des *moindres carrés contraints* suivants,

$$\min_b (\mathbf{Y} - \mathbf{X}b)^\top(\mathbf{Y} - \mathbf{X}b) \; s.c. \; \mathbf{R}b = r$$

Le lagrangien pour ce problème est,

$$L(b, \lambda) = (\mathbf{Y} - \mathbf{X}b)^\top(\mathbf{Y} - \mathbf{X}b) + 2\lambda^\top(\mathbf{R}b - r)$$

où λ est un vecteur $(q \times 1)$. Soit, $\tilde{\beta}$, $\tilde{\lambda}$, la solution, où $\tilde{\beta}$ est l'estimateur des moindres carrés contraint. Elle doit satisfaire les conditions du premier ordre,

$$\frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial b} = 2\mathbf{X}^\top\mathbf{X}\tilde{\beta} - 2\mathbf{X}^\top\mathbf{Y} + \mathbf{R}^\top\tilde{\lambda} = 0 \tag{5.13}$$

$$\frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial \lambda} = \mathbf{R}\tilde{\beta} - r = 0 \tag{5.14}$$

A partir de (5.13),

$$\begin{aligned}\tilde{\beta} &= (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y} - \mathbf{R}^\top\tilde{\lambda}) \\ &= \hat{\beta} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\tilde{\lambda}\end{aligned}$$

En combinant la dernière équation avec (5.14),

$$\begin{aligned}r &= \mathbf{R}\tilde{\beta} \\ &= \mathbf{R}\hat{\beta} - \mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\tilde{\lambda}\end{aligned}$$

et,

$$\tilde{\lambda} = \left(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1}(\mathbf{R}\hat{\beta} - r)$$

Par conséquent, l'estimateur des moindres carrés contraint est donné par,

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

Définissons les résidus contraints,

$$\begin{aligned}\tilde{\mathbf{U}} &= \mathbf{Y} - \mathbf{X}\tilde{\beta} \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta}) + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\left(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1}(\mathbf{R}\hat{\beta} - r) \\ &= \hat{\mathbf{U}} + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\left(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1}(\mathbf{R}\hat{\beta} - r)\end{aligned}$$

où $\hat{\mathbf{U}}$ est le vecteur des résidus non contraints. Considérons la somme *contrainte* des carrés des résidus,

$$\begin{aligned}SCR_r &= \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} \\ &= \hat{\mathbf{U}}^\top\hat{\mathbf{U}} + (\mathbf{R}\hat{\beta} - r)^\top\left(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1}(\mathbf{R}\hat{\beta} - r) \\ &+ 2\hat{\mathbf{U}}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\left(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1}(\mathbf{R}\hat{\beta} - r) \\ &= SCR + (\mathbf{R}\hat{\beta} - r)^\top\left(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1}(\mathbf{R}\hat{\beta} - r)\end{aligned}$$

où $SCR = \hat{\mathbf{U}}^\top\hat{\mathbf{U}}$ désigne la somme des carrés des résidus non contraints. Étant donné que $s^2 = \hat{\mathbf{U}}^\top\hat{\mathbf{U}}/(n-K)$, la F statistique discutée dans la sous-section précédente peut s'écrire,

$$F = \frac{(SCR_r - SCR)/q}{SCR/(n-K)} \tag{5.15}$$

Exemple 1. (**Significativité du modèle**) Considérons le modèle avec la constante,

$$Y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_K X_{iK} + U_i$$

Soit, l'hypothèse nuelle $H_0 : \beta_2 = ... = \beta_K = 0$. Le modèle contraint est donné par,

$$Y_i = \beta_1 + U_i$$

Dans ce cas l'estimateur des moindres carrés contraint est $\tilde{\beta}_1 = n^{-1}\sum_{i=1}^n Y_i = \bar{Y}$, et $SCR_r = SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Dans ce cas,

$$\begin{aligned}F &= \frac{(SCT - SCR)/(K-1)}{SCR/(n-K)} \\ &= \frac{SCE/(K-1)}{SCR/(n-K)} \\ &= \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \\ &\sim F_{K-1,n-K}\end{aligned}$$

Exemple 2. Considérons le modèle,

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i$$

et l'hypothèse nulle $H_0 : \beta_2 = \beta_3$. Le modèle contraint est donné par,

$$Y_i = \beta_1 + \beta_2(X_{i2} + X_{i3}) + U_i$$

Ainsi, pour tester si $\beta_2 = \beta_3$, on doit construire la nouvelle variable $W_i = (X_{i2} + X_{i3})$, calculer la SCR_r en prenant la SCR de la régression de Y_i sur une constante et W_i, calculer la SCR de la régression non contrainte, et construire la F statistique d'après (5.15).

6. PROPRIÉTÉS DU $\bar{R}^2$, MAUVAISES SPÉCIFICATIONS, TEST DE CHANGEMENT STRUCTUREL, VARIABLES INDICATRICES, PRÉVISIONS

6.1. **Propriétés du $\bar{R}^2$.** Nous allons montrer que quand on ajoute de nouveaux régresseurs, le $\bar{R}^2$ augmentera ou diminuera selon que la F-statistique associée aux nouveaux régresseurs est supérieure ou inférieure à 1, indépendamment du nombre de régresseurs ajoutés. Considérons le modèle non contraint avec $(K+q)$ régresseurs, et le modèle contraint avec K régresseurs,

$$\text{Modèle non contraint :} \beta_1 X_{i1} + ... + \beta_K X_{iK} + \beta_{K+1} X_{i,K+1} + ... + \beta_{K+q} X_{i,K+q} + U_i$$
$$\text{Modèle contraint :} \beta_1 X_{i1} + ... + \beta_K X_{iK} + U_I$$

où $q \geq 1$. Soit SCR et SCR_r, respectivement, la somme des carrés des résidus non contrainte, et la somme des carrés des résidus contrainte. Les coefficients de déterminaition ajustés qui leur sont associés sont donnés par,

$$\bar{R}^2 = 1 - \frac{n-1}{n-K-q}\frac{SCR}{SCT}$$
$$\bar{R}^2_r = 1 - \frac{n-1}{n-K}\frac{SCR_r}{SCT}$$

A présent,

$$\begin{aligned}
\bar{R}^2 - \bar{R}^2_r &= \frac{n-1}{n-K-q}\frac{SCR}{SCT} - \frac{n-1}{n-K}\frac{SCR_r}{SCT} \\
&= \frac{n-1}{SCT}\left(\frac{SCR_r}{n-K} - \frac{SCR}{n-K-q}\right) \\
&= \frac{n-1}{SCT}\frac{SCR}{n-K}\left(\frac{SCR_r}{SCR} - \frac{n-K}{n-K-q}\right) \\
&= \frac{n-1}{SCT}\frac{SCR}{n-K}\left(\frac{SCR_r}{SCR} - 1 + 1 - \frac{n-K}{n-K-q}\right) \\
&= \frac{n-1}{SCT}\frac{SCR}{n-K}\left(\frac{SCR_r - SCR}{SCR} - \frac{q}{n-K-q}\right) \\
&= \frac{n-1}{SCT}\frac{SCR}{n-K}\frac{q}{n-K-q}\left(\frac{(SCR_r - SCR)/q}{SCR/(n-K-q)} - 1\right)
\end{aligned}$$

Le résultat cherché est obtenu car

$$\frac{(SCR_r - SCR)/q}{SCR/(n-K-q)}$$

est la F-statistique associée à l'hypothèse nulle,

$$H_0 : \beta_{K+1} = ... = \beta_{K+q} = 0.$$

Pour comparaison, les valeurs critiques de la F-distribution dépassent 1. En conséquence, le choix d'un modèle en utilisant le coefficient de détermination ajusté peut conduire à ajouter des régresseur non pertinents.

6.2. **Mauvaise spécification du modèle.**

Exclusion de régresseurs pertinents. Supposons que le vrai modèle soit donné par,

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{U} \tag{6.1}$$

où $\mathbf{X}_1$ est une matrice $(n \times K_1)$, $\mathbf{X}_2$ est une matrice $(n \times K_2)$, $\beta_2 \neq 0$, et les hypothèses C1-C5 sont satisfaites avec $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2)$. Supposons que l'on régresse $\mathbf{Y}$ sur $\mathbf{X}_1$ uniquement, soit parce que $\mathbf{X}_2$ n'est pas disponible, ou parce que nous ne savons pas qu'il faut inclure ces régresseurs.
Dans un premier temps, étudions les propriétés de l'estimateurs des moindres carrés de β_1,

$$\begin{aligned}\tilde{\beta}_1 &= (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{Y} \\ &= (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{U}) \\ &= \beta_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\beta_2 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{U}\end{aligned}$$

Par l'hypothèse C2,

$$\mathrm{E}(\tilde{\beta}_1|\mathbf{X}_1) = \beta_1 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{X}_2\beta_2 \tag{6.2}$$

Dans ce cas, l'estimateur des moindres carrés de β_1 est biaisé, le biais étant donné par $(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{X}_2\beta_2$, et l'on remarque que ce biais disparaît dans le cas où $\mathbf{X}_1$ et $\mathbf{X}_2$ sont orthogonaux avec une probabilité égale à 1, i.e.,

$$\mathrm{P}(\mathbf{X}_1^\top\mathbf{X}_2 = 0) = 1$$

Considérons maintenant la variance conditionnelle de $\tilde{\beta}_1$. En raison de (6.2) et de l'hypothèse C3,

$$\begin{aligned}\mathrm{Var}(\tilde{\beta}_1|\mathbf{X}) &= (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\,\mathrm{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X})\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1} \\ &= \sigma^2(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\end{aligned}$$

En comparant avec la variance de $\hat{\beta}_1$, l'estimateur des moindres carrés de la régression qui inclut $\mathbf{X}_1$ et $\mathbf{X}_2$, soit,

$$\mathrm{Var}(\hat{\beta}_1|\mathbf{X}) = \sigma^2(\mathbf{X}_1^\top\mathbf{M}_2\mathbf{X}_1)^{-1}$$

où $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{P}_2$, et $\mathbf{P}_2 = \mathbf{X}_2(\mathbf{X}_2^\top\mathbf{X}_2)^{-1}\mathbf{X}_2^\top$. Considérons d'abord la différence,

$$\begin{aligned}\mathbf{X}_1^\top\mathbf{X}_1 - \mathbf{X}_1^\top\mathbf{M}_2\mathbf{X}_1 &= \mathbf{X}_1^\top\mathbf{P}_2\mathbf{X}_1 \\ &\geq 0\end{aligned}$$

L'inégalité résulte car $\mathbf{P}_2$ est symétrique et idempotente et par conséquent semi-définie positive. En conséquence,

$$(\mathbf{X}_1^\top\mathbf{X}_1)^{-1} - (\mathbf{X}_1^\top\mathbf{M}_2\mathbf{X}_1)^{-1} \leq 0$$

et

$$\mathrm{Var}(\tilde{\beta}_1|\mathbf{X}) - \mathrm{Var}(\hat{\beta}_1|\mathbf{X}) \leq 0$$

Ainsi, la variance augmente avec le nombre de régresseurs.
Sous l'hypothèse C5, nous obtenons que,

$$\tilde{\beta}_1|\mathbf{X} \sim \mathcal{N}\left(\beta_1 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\beta_2, \sigma^2(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\right)$$

Nous étudions maintenant l'effet d'une mauvaise spécification sur s^2, l'estimateur(sans biais) de σ^2. Nous avons,

$$s^2 = \frac{\mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y}}{n - K_1}$$

où K_1 est le nombre de colonnes dans $\mathbf{X}_1$, et $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top$. Comme le vrai modèle est (6.1),

$$s^2 = \frac{(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{U})^\top \mathbf{M}_1 (\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{U})}{n - K_1}$$

et,

$$\begin{aligned}
\mathrm{E}(s^2|\mathbf{X}) &= \mathrm{E}\left(\frac{\mathbf{U}^\top \mathbf{M}_1 \mathbf{U}}{n - K_1}|\mathbf{X}\right) + 2\,\mathrm{E}\left(\frac{\mathbf{U}^\top \mathbf{M}_1 \mathbf{X}_2\beta_2}{n - K_1}|\mathbf{X}\right) + \frac{\beta_2^\top \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 \beta_2}{n - K_1} \\
&= \sigma^2 + \frac{\beta_2^\top \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 \beta_2}{n - K_1} \\
&\geq \sigma^2
\end{aligned}$$

Le biais demeure, même si $\mathbf{X}_1$ et $\mathbf{X}_2$ sont orthogonaux car dans ce cas $\mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2$.
Une des conséquences de l'exclusion de variables pertinentes est que les test est les intervalles de confiance ne sont pas valides.

Inclusion de régresseurs non pertinents. Supposons que le vrai modèle soit,

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{U}$$

Cependant, nous incluons $\mathbf{X}_2$ et estimons β_1 par,

$$\tilde{\beta}_1 = (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{M}_2\mathbf{Y}$$

Dans ce cas, l'estimateur des moindres carrés est sans biais sous l'hypothèse C2,

$$\begin{aligned}
\mathrm{E}(\tilde{\beta}_1|\mathbf{X}) &= \mathrm{E}\left((\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{M}_2(\mathbf{X}_1\beta_1 + \mathbf{U})|\mathbf{X}\right) \\
&= \beta_1 + (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{M}_2\,\mathrm{E}(\mathbf{U}|\mathbf{X}) \\
&= \beta_1
\end{aligned}$$

Sous l'hypothèse C3, la variance de $\tilde{\beta}_1$ est donnée par la formule habituelle,

$$\mathrm{Var}(\tilde{\beta}_1|\mathbf{X}) = \sigma^2(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}$$

Cependant, $\tilde{\beta}_1$ est inefficace, en raison du théorème de Gauss-Markov. Comme, nous l'avons vu dans la section précédente, la variance augmente avec le nombre de régresseurs. Sous l'hypothèse C5, nous avons,

$$\tilde{\beta}_1|\mathbf{X} \sim \mathcal{N}\left(\beta_1, (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}\right)$$

Maintenant, considérons s^2 avec ici,

$$s^2 = \frac{\mathbf{Y}^\top \mathbf{M}_\mathbf{X} \mathbf{Y}}{n - K_1 - K_2}$$

où $\mathbf{M}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}$. Étant donné que $\mathbf{M}_\mathbf{X}\mathbf{X}_1 = 0$, il s'en suit que,

$$s^2 = \frac{\mathbf{U}^\top\mathbf{M}_\mathbf{X}\mathbf{U}}{n - K_1 - K_2}$$

et,

$$\mathrm{E}(s^2|\mathbf{X}) = \sigma^2$$

Naturellement, les tests et intervalles de confiance habituels demeurent valides en cas d'inclusion de variables non pertinentes. Toutefois, les régions de confiance pour β_1 seront plus grandes et les tests moins puissants par comparaison au modèle correctement spécifié. La discussion dans les sous-sections suivantes indique que pour choisir le modèle, on doit commencer par le modèle le plus général, et éliminer les régresseurs non pertinent en appliquant des F-tests.

6.3. **Test de changement structurel.** Supposons deux modèles de régression qui représentent, par exemple, des observations pour deux pays ou pour deux périodes différentes,

$$\mathbf{Y}_1 = \mathbf{X}_1\beta_1 + \mathbf{U}_1$$
$$\mathbf{Y}_2 = \mathbf{X}_2\beta_2 + \mathbf{U}_2$$

où $\mathbf{Y}_1$ est un vecteur $(n_1 \times 1)$ d'observations pour la variable dépendante dans la première population, $\mathbf{X}_1$ est un vecteur $(n_1 \times K)$ d'observations pour les régresseurs dans la première sous-population, $\mathbf{Y}_2$ est un vecteur $(n_2 \times 1)$ d'observations pour la variable dépendante dans la deuxième sous-population, et $\mathbf{X}_2$ est un vecteur $(n_2 \times K)$ d'observations pour les régresseurs dans la deuxième population.
On peut se demander si la réaction de la variable dépendante aux variations des régresseurs diffère entre les deux sous -populations en testant,

$$H_0 : \beta_1 = \beta_2 \tag{6.3}$$

Pour combiner deux équations dans une seule, il convient de définir,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix},\ \mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix},\ \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix},\ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

En utilisant ces définitions, le modèle non contraint peut s'écrire,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U} \tag{6.4}$$

qui le modèle de régression linéaire habituel. Nous supposons que ce modèle vérifie les hypothèse C1-C5. Dans ce cadre, les contraintes données par (6.3) peuvent s'écrire,

$$\mathbf{R}\beta = \begin{pmatrix} \mathbf{I}_K & -\mathbf{I}_K \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = 0$$

Notons que dans ce cas,

$$\begin{aligned}(\mathbf{X}^\top\mathbf{X})^{-1} &= \begin{pmatrix} \mathbf{X}_1^\top\mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2^\top\mathbf{X}_2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (\mathbf{X}_1^\top\mathbf{X}_1)^{-1} & 0 \\ 0 & (\mathbf{X}_2^\top\mathbf{X}_2)^{-1} \end{pmatrix}\end{aligned}$$

En conséquence,

$$\begin{aligned}
\mathbf{M}_{\mathbf{X}} &= \mathbf{I}_n - \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix}\begin{pmatrix} (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} & 0 \\ 0 & (\mathbf{X}_2^\top \mathbf{X}_2)^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{X}_1^\top & 0 \\ 0 & \mathbf{X}_2^\top \end{pmatrix} \\
&= \mathbf{I}_n - \begin{pmatrix} \mathbf{X}_1(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top & 0 \\ 0 & \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{n_1} - \mathbf{X}_1(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top & 0 \\ 0 & \mathbf{I}_{n_2} - \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{M}_1 & 0 \\ 0 & \mathbf{M}_2 \end{pmatrix}
\end{aligned}$$

Maintenant, la SCR non contrainte est donnée par les SCR dans les deux régression séparées,

$$\begin{aligned}
\hat{\mathbf{U}}^\top \hat{\mathbf{U}} &= \mathbf{Y}^\top \mathbf{M}_{\mathbf{X}} \mathbf{Y} \\
&= \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}^\top \begin{pmatrix} \mathbf{M}_1 & 0 \\ 0 & \mathbf{M}_2 \end{pmatrix}\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \\
&= \mathbf{Y}_1^\top \mathbf{M}_1 \mathbf{Y}_1 + \mathbf{Y}_2^\top \mathbf{M}_2 \mathbf{Y}_2 \\
&= SCR_1 + SCR_2
\end{aligned}$$

Notons qu'il y a, dans le modèle non contraint, $2K$ coefficients.
Maintenant, le modèle contraint peut s'écrire,

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}\beta_1 + \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix} \tag{6.5}$$

Par conséquent, la SCR contrainte doit être obtenue en empilant les deux sous populations. Définissons,

$$\mathbf{X}_r = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

et,

$$\mathbf{M}_r = \mathbf{I}_n - \mathbf{X}_r(\mathbf{X}_r^\top \mathbf{X}_r)^{-1}\mathbf{X}_r^\top$$

La SCR contrainte est donnée par,

$$SCR_r = \mathbf{Y}^\top \mathbf{M}_r \mathbf{Y}$$

Par conséquent, le test d'absence de changement structurel repose sur la statistique suivante,

$$F = \frac{(SCR_r - SCR_1 - SCR_2)/K}{(SCR_1 + SCR_2)/(n-2K)}$$

et on rejette, l'hypothèse nulle d'absence de changement structurel quand,

$$F > F_{K,n-2K,1-\alpha}$$

6.4. **Variables indicatrices.** Il est fréquent que l'on s'intéresse aux effets de variables qualitatives et qui ne sont pas quantifiées de la manière habituelle. Par exemple, on peut s'intéresser aux effets du sexe, du statut marital, de l'origine ethnique, de la religion sur d'autres variables telles que le revenu, ou le niveau d'études. Une approche courante pour quantifier ce type de variables consiste à introduire des variables artificielles qui indiquent si une modalité particulière est présente. Par exemple, supposons qu'une variable qualitative présente M modalités(e.g., pour le sexe il en a 2). Pour les observation $i = 1, ..., n$, définissons la variable indicatrice $D_{i,m}$, $m = 1, ..., M$ telle que,

$$D_{i,m} = \begin{cases} 1, & \text{si l'observation } i \text{ appartient à la modalité } m \\ & 0 \text{ autrement} \end{cases}$$

Ainsi, par exemple, si Y_i est le salaire de l'individu i, et que l'on veut étudier les effets du sexe sur Y_i, nous aurons deux indicatrices, à savoir,

$$D_{i,1} = \begin{cases} 1, & \text{si } i \text{ est un homme} \\ 0, & \text{si } i \text{ est une femme} \end{cases}$$

$$D_{i,2} = \begin{cases} 1, & \text{si } i \text{ est une femme} \\ 0, & \text{si } i \text{ est un homme} \end{cases}$$

Considérons le modèle de régression,

$$Y_i = \alpha_1 D_{i,1} + \alpha_2 D_{i,2} + X_i^\top \beta + U_i$$

où X_i est un vecteur d'autres régresseurs comme le nombre d'années d'étude, l'expérience, etc. Dans ce cas, α_1 et α_2 donnent le "salaire de départ"(i.e., en "fixant" $X_i = 0$, et $U_i = 0$) pour un homme et une femme respectivement. Alternativement, la spécification suivante peut être envisagée,

$$Y_i = \alpha_0 + \alpha_1 D_{i,1} + X_i^\top \beta + U_i$$

Dans ce cas, le salaire de départ pour une femme est α_0, et pour un homme c'est $\alpha_0 + \alpha_1$. Le coefficient α_1 donne la différence entre le salaire de départ d'homme et celui d'une femme. Nous pouvons alors tester l'existence d'une telle différence en testant l'hypothèse que $\alpha_1 = 0$.

Notons que dans cet exemple nous ne pouvons pas inclure à la fois la constante et les deux indicatrices car pour tout i,

$$D_{i,1} + D_{i,2} = 1$$

et l'hypothèse C4 ne sera pas vérifiée. La "règle générale" pour une variable qualitative à M modalités est : soit inclure les M indicatrices sans la constante, soit inclure la constante et $M - 1$ indicatrices.

On peut aussi considérer que les effets des régresseurs X_i différent selon les modalités. Ainsi, dans l'exemple précédent, ceci peut être modélisé en introduisant des effets d'interaction entre X_i et $D_{i,1}$,

$$Y_i = \alpha_0 + \alpha_1 D_{i,1} + X_i^\top \beta + (D_{i,1} X_i)^\top \delta + U_i$$

Et à présent, l'effet marginal de X_i est β pour les femmes, et $\beta + \delta$ pour les hommes. Nous pouvons alors tester si le modèle est différent pour les hommes et pour les femmes en testant

$H_0 : \alpha_1 = 0,\ \delta = 0.$
Considérons le test de changement structurel discuté dans la sous section précédente. Définissons,

$$D_i = \begin{cases} 0 & \text{pour } i = 1, ..., n_1 \\ 1 & \text{pour } i = n_1 + 1, ..., n \end{cases}$$

On peut écrire le modèle pour $i = 1, ..., n$ comme suit,

$$Y_i = X_i^\top \beta_1 + (D_i X_i)^\top \delta + U_i$$

où de manière équivalente,

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \beta_1 + \begin{pmatrix} 0 \\ \mathbf{X}_2 \end{pmatrix} \delta + \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix} \tag{6.6}$$

Dans ce cas, $\beta_2 = \beta_1 + \delta$, et le test d'absence de changement structurel revient à tester $H_0 : \delta = 0$. Afin de montrer que les deux approches, avec et sans variables indicatrices, sont équivalentes, il suffit de montrer que la matrice des régresseurs dans (6.4),

$$\begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix}$$

engendre le même sous espace linéaire que celle dans (6.6),

$$\begin{pmatrix} \mathbf{X}_1 & 0 \\ \mathbf{X}_2 & \mathbf{X}_2 \end{pmatrix}$$

6.5. **Prévision.** Considérons de nouveau le modèle de régression linéaire normal défini par les hypothèses C1-C5,

$$Y_i = X_i^\top \beta + U_i$$

Nous discutons à présent la question de la *prévision* de la variable dépendante Y_i, sachant des valeurs fixes x_f pour le vecteur $(K \times 1)$ des régresseurs, X_i. Soit $\hat{\beta}$ l'estimateur des moindres carrés de β à partir des données $\{(Y_i, X_i)\}_{i=1}^n$. Notons que x_f peut ou ne peut pas être une des valeurs réalisées des régresseurs dans l'échantillon observé. Étant donné que,

$$\mathrm{E}(Y_i | X_i = x_f) = x_f^\top \beta$$

il est naturel d'estimer l'espérance conditionnelle $\mathrm{E}(Y_i|X_i = x_f)$ par,

$$\hat{Y}_f = x_f^\top \hat{\beta} \tag{6.7}$$

Notons que $\hat{Y}_f$ est la valeur prédite d'un point sur la *droite de régression*. Comme x_f est fixe, $\hat{Y}_f$ est aléatoire par le biais de $\hat{\beta}$ seulement. En utilisant les résultats pour $\hat{\beta}$, nous obtenons,

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

L'intervalle de confiance de niveau α pour le point de la droite de régression qui correspond à x_f est donné par,

$$x_f^\top \hat{\beta} \pm t_{n-K,1-\alpha/2} \sqrt{s^2 x_f^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_f}$$

Maintenant, considérons la prédiction d'un point en dehors de la droite de régression. Définissons,

$$Y_f = x_f^\top \beta + U_f$$

où le vecteur $((n+1) \times 1)$, $(\mathbf{U}^\top, U_f)^\top$, vérifie,

$$\begin{pmatrix} \mathbf{U} \\ U_f \end{pmatrix} | \mathbf{X} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n+1}) \tag{6.8}$$

Étant donné que U_f ne peut pas être prédit à partir de $\mathbf{X}$, la valeur prédite de Y_f est donnée par (6.7). Soit maintenant, *l'erreur de prévision* donné par,

$$\begin{aligned} \hat{U}_f &= Y_f - x_f^\top \hat{\beta} \\ &= U_f - x_f^\top (\hat{\beta} - \beta) \end{aligned}$$

Le résultat dans (6.8) implique que,

$$\hat{U}_f | \mathbf{X} \sim \mathcal{N}(0, \sigma^2 + \sigma^2 x_f^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_f)$$

Par conséquent, l'intervalle de confiance de niveau α pour la valeur prédite de Y_f est donné par,

$$x_f^\top \hat{\beta} \pm t_{n-K,1-\alpha/2} \sqrt{s^2 (1 + x_f^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_f)}$$