UNIVERSITÉ Grenoble Alpes

UGA L3 Miash:
Économétrie 1
Modèle de régression linéaire et moindres carrés

Michal W. Urdanivia*

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

18 septembre 2017

UNIVERSITÉ Grenoble Alpes

Contenu

UNIVERSITÉ Grenoble Alpes

Contenu

UNIVERSITÉ Grenoble Alpes

Contenu

UNIVERSITÉ Grenoble Alpes

Contenu

Section 1

Modèle de régression linéaire et moindres carrés

Définitions

- On s'intéresse à l'effet d'un groupe de variables $X \in \mathbb{R}^K$, traditionnellement appelées *régresseurs*, sur une autre variable $Y \in \mathbb{R}$ traditionnellement appelée *variable dépendante*.
- On dispose de données $\{(Y_i, X_i)\}_{i=1}^n$ où Y_i est une variable aléatoire et X_i est un vecteur $K \times 1$,

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

- (Y_i, X_i) est appelée observation(sous entendu de (Y, X)), et l'ensemble des n observations est un *échantillon*.
- Le vecteur X_i contient les valeurs des K variables pour l'observation i.

Définitions

- Pour des *données en coupe* il est souvent supposé que toutes les observations sont tirées indépendamment les unes des autres à partir d'une même distribution(i.e., loi de probabilité).
- 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.).
- 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.
- Un outil pour étudier la relation entre la variable dépendante et les régresseurs est l'espérance conditionnelle de Y_i sachant X_i, $\mathbb{E}(Y_i|X_i)$, laquelle vue comme une fonction de X_i est appelée *fonction de régression*.
- La différence entre Y_i et son espérance conditionnelle(i.e., sa fonction de régression) est appelée *terme d'erreur*(ou plus succinctement *erreur*),

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

Définitions

- U_i n'est pas une variable observable par l'analyste étant donné que l'espérance conditionnelle lui est inconnue.
- Dans un cadre *paramétrique* ou *semi-paramétrique*, il est souvent supposé que l'espérance conditionnelle est connue à un ensemble de *paramètres* près.
- Ainsi dans le modèle de régression linéaire on suppose que $\mathbb{E}(Y_i|X_i)$ est linéaire par rapport à un vecteur de paramètres inconnus,

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

où,

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

Définitions

est un vecteur de K paramètres constants.

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

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

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.

Définitions

- Les équations (1) et (2) permettent d'écrire,

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

où par définition de (1)

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

- Ceci implique que les régresseurs ne contiennent aucune information quant à l'écart entre Y_i et sont espérance conditionnelle.
- En outre, *la loi des conditionnements successifs* implique que les erreurs ont une espérance nulle : $\mathbb{E}(U_i) = 0$. 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 $\sigma^2 > 0$.

Hypothèses

- Introduisons les notations vectorielles et matricielles suivantes,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} X_1^\top \\ X_2^\top \\ . \\ . \\ . \\ X_n^\top \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & . & . & . & X_{1K} \\ X_{21} & X_{22} & . & . & . & X_{2K} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ X_{n1} & X_{n2} & . & . & . & X_{nK} \end{pmatrix}$$

Hypothèses

$$\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ . \\ . \\ . \\ U_n \end{pmatrix}$$

- Le modèle de régression linéaire consiste dans les hypothèses suivantes :

Hypothèse H1

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

Hypothèse H2

$\mathbb{E}(\mathbf{U}|\mathbf{X}) = 0$ *p.s.*

Hypothèse H3

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

Hypothèse H4

$\mathrm{Rang}(\mathbf{X}) = K$ *p.s.*

Hypothèses

- Pour l'inférence nous supposerons parfois que,

Hypothèse H5

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

- Les hypothèses H1-H5 définissent alors le *modèle de régression linéaire normal* avec,

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

- Remarquons qu'étant donné que les covariances dans H5 sont toutes nulles, H5 implique l'indépendance des erreurs.
- Les hypothèses H1-H4 seules, n'impliquent pas l'indépendance entre les observations.
- En fait, plusieurs résultats importants n'exigent pas d'observations indépendantes.

Hypothèses

- Néanmoins, nous supposerons parfois l'indépendance sans la normalité.

Hypothèse H6

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

Dans le cas de régresseurs fixes cette hypothèse peut être remplacé par celle d'erreurs i.i.d.

- L'hypothèse H2 dit que **U** est indépendant de **X** en espérance, ce qui est une hypothèse forte.
- Cependant, plusieurs résultats importants peuvent être obtenus avec une hypothèse plus faible de non corrélation.

Hypothèse H7

Pour $i = 1, 2, ..., n$, $\mathbb{E}(U_i X_i) = 0$, *et* $\mathbb{E}(U_i) = 0$.

- Toutefois sous cette condition $X_i^\top \beta$ ne peut pas s'interpréter comme une espérance conditionnelle, auquel cas (3) doit être vu comme un *processus générateur des données*.

Hypothèses

- L'hypothèse H3 implique que les erreurs U_i ont la même variance pour tout i, et ne sont pas corrélés entre eux, i.e., $\mathbb{E}(U_i U_j|\mathbf{X}) = 0$ pour $i \neq j$.
- Notons que l'indépendance entre les erreurs peut aussi être obtenue avec la condition H5 ou sous les conditions H1 et H6.
- L'hypothèse H4 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 $\mathbf{X}$ 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.

Estimation par la méthode des moments

- Nous allons construire des estimateurs des paramètres β et σ^2.
- Une des méthodes les plus ancienne 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.
- Les hypothèses H1, et H2 ou H7 impliquent que la vraie valeur de β doit satisfaire,

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

Estimation par la méthode des moments

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

$$n^{-1}\sum_{i=1}^{N}(Y_i - X_i^\top \widehat{\beta})X_i = n^{-1}\sum_{i=1}^{N}X_i Y_i - n^{-1}\sum_{i=1}^{N}X_i X_i^\top \widehat{\beta} = 0 \qquad (6)$$

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

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

qui peut s'écrire alternativement,

$$\widehat{\beta} = \left(\sum_{i=1}^{N}X_i X_i^\top\right)^{-1} \sum_{i=1}^{N}X_i Y_i \qquad (8)$$

Estimation par la méthode des moments

ou,

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

où l'on note que la matrice $\sum_{i=1}^{N} X_i X_i^\top = \mathbf{X}^\top \mathbf{X}$ est inversible sous l'hypothèse H4.

- On définit les *valeurs ajustées* ou *prédictions* , ainsi qu'un vecteur $n \times 1$ des valeurs ajustées ou des prédictions, par respectivement,

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

- On définit les *résidus* , et le vecteur $n \times 1$ des résidus, par respectivement,

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

Estimation par la méthode des moments

- Du fait de (6) le vecteur des résidus vérifie les *K équations normales*,

$$\sum_{i=1}^{N} \widehat{U}_i X_i = \begin{pmatrix} \sum_{i=1}^{N} \widehat{U}_i X_{i1} \\ \sum_{i=1}^{N} \widehat{U}_i X_{i2} \\ . \\ . \\ . \\ \sum_{i=1}^{N} \widehat{U}_i X_{iK} \end{pmatrix} = 0 \tag{10}$$

ou en notation matricielle,

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

Estimation par la méthode des moments

- Si le modèle contient une constante alors il résulte des équations normales que $\sum_{i=1}^{N} \widehat{U}_i = 0$(il suffit en effet de considérer que, par exemple, le premier régresseur est constant et égal à 1).
- Afin d'estimer σ^2 considérons,

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

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

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

Moindres carrés

- Soit le problème consistant à minimiser l'erreur de prédiction quand on cherche à prédire Y_i par son espérance conditionnelle, $\mathbb{E}(Y_i|X_i)$, supposée être une fonction linéaire telle que (2).
- Plus précisément, $Y_i - \mathbb{E}(Y_i|X_i)$ étant l'erreur de prédiction on cherche β qui minimise un critère de perte quadratique,

$$\beta \in \underset{b \in \mathbb{R}^K}{\arg\min}\, S(b)$$

où $S(b) = \mathbb{E}\left((Y_i - X_i^\top b)^2\right)$.

- La contrepartie empirique de ce problème permet de définir un estimateur de β par,

$$\widehat{\beta} \in \underset{b \in \mathbb{R}^K}{\arg\min}\, S_n(b)$$

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 $S(b)$.

Moindres carrés

- Nous pouvons montrer que l'estimateur des moments de la section précédente est aussi l'estimateur des moindres carrés.
- La fonction objectif précédente peut s'écrire(voir notes),

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

- La minimisation de $S_n(b)$ équivaut à minimiser $(\widehat{\beta} - b)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\beta} - b)$ car $(\mathbf{Y} - \mathbf{X}\widehat{\beta})^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta})$ ne fait pas intervenir b.
- Sous l'hypothèse H4 la matrice $\mathbf{X}$ est de plein rang, et dans ce cas $\mathbf{X}^\top \mathbf{X}$ est définie positive,

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

et $(\widehat{\beta} - b)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\beta} - b) = 0$ ssi $\widehat{\beta} = b$.

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

Moindres carrés

- Pour cela, écrivons,

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

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

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

la condition du premier ordre est,

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

ce qui permet d'obtenir,

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

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

Propriétés de l'estimateur des moindres carrés

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

Proposition 1

$\widehat{\beta}$ *est un estimateur linéaire.*

- 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, $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$.

Proposition 2

Sous les hypothèses H1, H2, et H4, $\widehat{\beta}$ *est sans biais, i.e.,*

$$\mathbb{E}(\widehat{\beta}) = \beta$$

Propriétés de l'estimateur des moindres carrés

- Pour montrer cette propriété écrivons, en utilisant l'hypothèse H1,

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

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

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

Notons que,

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

car sous l'hypothèse H2, $\mathbb{E}(\mathbf{U}|\mathbf{X}) = 0$. Nous avons donc,

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

et par la loi des espérances itérées,

$$\mathbb{E}\left(\widehat{\beta}\right) = \mathbb{E}\left(\mathbb{E}\left(\widehat{\beta}|\mathbf{X}\right)\right) = \beta$$

Propriétés de l'estimateur des moindres carrés

- L'équation (13) montre que $\widehat{\beta}$ est conditionnellement sans biais sachant $\mathbf{X}$.
- On remarque aussi que pour que $\widehat{\beta}$ soit sans biais l'hypothèse H7 n'est pas suffisante.

Proposition 3

Sous les hypothèses H1, H2, et H4,

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

et avec des erreurs homoscédastiques(i.e., sous l'hypothèse H3),

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

Propriétés de l'estimateur des moindres carrés

- Pour montrer ces résultats, partons de la définition de la variance conditionnelle de $\widehat{\beta}$,

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

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

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

Propriétés de l'estimateur des moindres carrés

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

Proposition 4

Sous les hypothèses H1 - H5,

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

- Il est suffit ici de montrer ici que conditionnellement à $\mathbf{X}$ la distribution de $\widehat{\beta}$ est normale.
- On aura alors que, $\widehat{\beta}|\mathbf{X} \sim \mathcal{N}\left(\mathbb{E}(\widehat{\beta}|\mathbf{X}), \mathrm{Var}(\widehat{\beta}|\mathbf{X})\right)$.
- Néanmoins la normalité de $\widehat{\beta}|\mathbf{X}$ résulte ici de ce que $\widehat{\beta}$ est une fonction de linéaire de $\mathbf{Y}$, et que sous l'hypothèse H5 $\mathbf{Y}|\mathbf{X}$ est normale.

Propriétés de l'estimateur des moindres carrés

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

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

Proposition 5

*(**Éfficacité ou théorème de Gauss-Markov.**) Sous les hypothèses H1-H4, 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 $\mathrm{Var}\left(b|\mathbf{X}\right) - \mathrm{Var}\left(\widehat{\beta}|\mathbf{X}\right)$ doit être semi-définie positive :*

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

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

Propriétés de l'estimateur des moindres carrés

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

$$b = \mathbf{AY}, \quad \mathbb{E}(b|\mathbf{X}) = \beta$$

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

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

- Par l'hypothèse H2, $\mathbb{E}(\mathbf{U}|\mathbf{X}) = 0$, et par conséquent, pour que b soit sans biais nous avons besoin de $\mathbf{AX} = \mathbf{I}_K$.

Propriétés de l'estimateur des moindres carrés

- Montrons maintenant que $\text{Cov}\left(\widehat{\beta}, b|\mathbf{X}\right) = \text{Var}\left(\widehat{\beta}|\mathbf{X}\right)$,

$$\begin{aligned}
\text{Cov}\left(\widehat{\beta}, b|\mathbf{X}\right) =& \mathbb{E}\left((\widehat{\beta}-\beta)(b-\beta)^\top\right)\\
=& \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{A}^\top|\mathbf{X}\right)\\
=& (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X})\mathbf{A}^\top)\\
=& \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{A}^\top(\text{car sous H3},\ \ \mathbb{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X}) = \sigma^2\mathbf{I}_n)\\
=& \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}(\text{car},\ \ \mathbf{X}^\top\mathbf{A}^\top = \mathbf{I}_K)\\
=& \text{Var}\left(\widehat{\beta}|\mathbf{X}\right)
\end{aligned}$$

Finalement,

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

Propriétés de l'estimateur des moindres carrés

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

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

- Pour démontrer l'unicité, considérons un estimateur linéaire sans biais $\tilde{\beta}$ tel que $\text{Var}\left(\tilde{\beta}|\mathbf{X}\right) = \text{Var}\left(\widehat{\beta}|\mathbf{X}\right)$.
- Alors, par (14), $\text{Var}\left(\widehat{\beta} - b|\mathbf{X}\right) = 0$, et par conséquent, $\tilde{\beta} = \widehat{\beta} + c(\mathbf{X})$ pour une fonction $c(\mathbf{X})$ à valeurs dans $\mathbb{R}^K$ qui dépend uniquement de $\mathbf{X}$.
- Cependant, comme $\widehat{\beta}$ 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 H3, $\mathbb{E}(\mathbf{U}\mathbf{U}^\top|\mathbf{X}) = \sigma^2\mathbf{I}_n$, 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.

Section 2

Géométrie des moindres carrés

Matrices de projection

- Nous pouvons penser à $\mathbf{Y}$ et aux colonnes $\mathbf{X}$ comme des éléments de l'espace euclidien à n dimensions, $\mathbb{R}^n$.
- Considérons le sous-espace de $\mathbb{R}^n$ appelé l'*espace des colonnes* de la matrice $n \times K$, $\mathbf{X}$.
- C'est la collection de tous les vecteurs dans $\mathbb{R}^n$ qui peuvent s'écrire comme des combinaisons linéaires des colonnes de $\mathbf{X}$,

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

- Rappelons maintenant qu'étant donné deux vecteurs a, b, dans $\mathbb{R}^n$, la distance entre a et b est donné par la norme euclidienne de leur différence $||a - b|| = \sqrt{(a-b)^\top (a-b)}$.
- Par conséquent, le problème de la minimisation de la somme des carrés des erreurs, $(\mathbf{Y} - \mathbf{X}b)^\top (\mathbf{Y} - \mathbf{X}b)$, consiste à trouver, parmi tous les éléments de $\mathcal{S}(\mathbf{X})$, celui dont la distance par rapport à $\mathbf{Y}$ est la plus petite,

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

Matrices de projection

- Une solution au problème des moindres carrés, $\widehat{Y} = \mathbf{X}\widehat{\beta}$ doit être choisie de sorte que le vecteur des résidus, $\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{Y}}$ soit orthogonal(perpendiculaire) à chaque colonne de $\mathbf{X}$,

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

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

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

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

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

- Tout élément de $\mathcal{S}^\perp(\mathbf{X})$ est orthogonal à chaque élément de $\mathcal{S}(\mathbf{X})$.

Matrices de projection

- La solution au problème des moindres carrés est donnée par,

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

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 $\mathbf{Y} \in \mathbb{R}^n$,

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

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

$$\mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y} \in \mathcal{S}^\perp(\mathbf{X}) \tag{15}$$

Matrices de projection

- Pour montrer (15), 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,

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

Matrices de projection

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

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

où,

$$\begin{aligned}\mathbf{M_X} &= \mathbf{I}_n - \mathbf{P_X} \\ &= \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\end{aligned}$$

est une matrice de projection dans $\mathcal{S}^{\perp}(\mathbf{X})$.

- Les matrices $\mathbf{P_X}$ et $\mathbf{M_X}$ présentent les propriétés suivantes.
  1. $\mathbf{P_X} + \mathbf{M_X} = \mathbf{I}_n$. Ceci implique, que pour tout $\mathbf{Y} \in \mathbb{R}^n$,

$$\mathbf{Y} = \mathbf{P_X Y} + \mathbf{M_X Y}$$

Matrices de projection

2. $\mathbf{P_X}$ et $\mathbf{M_X}$ sont symétriques,

$$\mathbf{P_X}^\top = \mathbf{P_X}, \quad \mathbf{M_X}^\top = \mathbf{M_X}$$

3. $\mathbf{P_X}$ et $\mathbf{M_X}$ sont idempotentes,

$$\mathbf{P_X}\mathbf{P_X} = \mathbf{P_X}, \quad \mathbf{M_X}\mathbf{M_X} = \mathbf{M_X}$$

En effet,

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

de même,

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

Matrices de projection

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,

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

- 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.

Matrices de projection

- 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).

Propriétés de $\widehat{\sigma}^2$

- σ^2 est estimé par,

$$\begin{aligned}\widehat{\sigma}^2 &= n^{-1}\sum_{i=1}^{N}\widehat{U}_i^2 \\ &= n^{-1}\widehat{\mathbf{U}}^\top\widehat{\mathbf{U}}\end{aligned}$$

- Mais sous H1 - H4 $\widehat{\sigma}^2$ est biaisé.
- En effet,

$$\begin{aligned}\widehat{\mathbf{U}} &= \mathbf{M}_\mathbf{X}\mathbf{Y} \\ &= \mathbf{M}_\mathbf{X}(\mathbf{X}\beta + \mathbf{U}) \\ &= \mathbf{M}_\mathbf{X}\mathbf{U}\end{aligned}$$

où la dernière égalité résulte de ce que $\mathbf{M}_\mathbf{X}\mathbf{X} = 0$.

Propriétés de $\widehat{\sigma}^2$

- En outre,

$$\begin{aligned} n\widehat{\sigma}^2 &= \widehat{\mathbf{U}}^\top \widehat{\mathbf{U}} \\ &= \mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{M}_\mathbf{X} \mathbf{U} \\ &= \mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U} \end{aligned}$$

- Étant donné que $\mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U}$ est un scalaire,

$$\mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U} = \mathrm{Tr}\left(\mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U}\right)$$

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

Propriétés de $\widehat{\sigma}^2$

- Nous avons,

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

- La dernière égalité résulte de ce que par l'hypothèse H3, $\mathbb{E}(\mathbf{U}^\top\mathbf{U}) = \sigma^2\mathbf{I}_n$.

Propriétés de $\widehat{\sigma}^2$

- Maintenant,

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

- Il s'en suit que,

$$\mathbb{E}(\widehat{\sigma}^2) = \frac{n-K}{n}\sigma^2 \tag{16}$$

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

Propriétés de $\widehat{\sigma}^2$

- Pour cela, définissons,

$$s^2 = \widehat{\sigma}^2 \frac{n}{n-K}$$
$$= (n-K)^{-1} \sum_{i=1}^{N} \widehat{U}_i^2$$

- Et il résulte de (16) que,

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

Régression partitionnée

- Considérons,

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

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, $(K_1 \times 1)$ et $(K_2 \times 1)$.

- Concentrons nous sur $\mathbf{X}_1$ et β_1.
- Soit l'estimateur des moindres carrés de β,

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

Régression partitionnée

- Nous pouvons écrire la version suivante des équations normales,

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

comme suit,

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

- On peut obtenir des expressions pour $\widehat{\beta}_1$ et $\widehat{\beta}_2$ 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,

$$\widehat{\beta}_1 = (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{Y} \tag{17}$$

Régression partitionnée

- Pour montrer cela, commençons par écrire,

$$\mathbf{Y} = \mathbf{X}_1\widehat{\beta}_1 + \mathbf{X}_2\widehat{\beta}_2 + \widehat{\mathbf{U}} \tag{18}$$

- Notons que par construction,

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

Régression partitionnée

- Substituons l'équation (18) dans la partie droite de l'équation (17),

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

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

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

Régression partitionnée

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 $\widehat{\beta}_1$.
- Autrement dit, $\widehat{\beta}_1$ décrit l'effet de $\mathbf{X}_1$ une fois que ceux de $\mathbf{X}_2$ ont été contrôlés.

Régression partitionnée

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

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

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 unitaire $(n \times 1)$, i.e.,

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

Régression partitionnée

- La matrice des régresseurs est alors,

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

- Considérons,

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

et,

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

Régression partitionnée

- 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,

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

Régression partitionnée

où,

$$\overline{X} = \frac{\mathbf{1}_n^\top X}{n}$$
$$= n^{-1} \sum_{i=1}^{N} X_i$$

- 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.

Régression partitionnée

- Et nous pouvons écrire,

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})Y_i}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

$$= \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

Qualité de l'ajustement et coefficient de détermination ou R^2

- Écrivons,

$$\begin{aligned}\mathbf{Y} &= \mathbf{P_X Y} + \mathbf{M_X Y} \\ &= \widehat{\mathbf{Y}} + \widehat{\mathbf{U}}\end{aligned}$$

où par construction,

$$\begin{aligned}\widehat{\mathbf{Y}}^\top \widehat{\mathbf{U}} &= (\mathbf{P_X Y})^\top (\mathbf{M_X Y}) \\ &= \mathbf{Y}^\top \mathbf{P_X M_X Y} \\ &= 0\end{aligned}$$

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

Qualité de l'ajustement et coefficient de détermination ou R^2

- La *variation totale* dans $\mathbf{Y}$ est,

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y}$$
$$= (\widehat{\mathbf{Y}} + \widehat{\mathbf{U}})^\top \mathbf{M}_1 (\widehat{\mathbf{Y}} + \widehat{\mathbf{U}})$$
$$= \widehat{\mathbf{Y}}^\top \mathbf{M}_1 \widehat{\mathbf{Y}} + \widehat{\mathbf{U}}^\top \mathbf{M}_1 \widehat{\mathbf{U}} + 2\widehat{\mathbf{Y}}^\top \mathbf{M}_1 \widehat{\mathbf{U}}$$

où $\overline{Y} = n^{-1}\sum_{i=1}^{N} Y_i$.

- Comme le modèle contient une constante,

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

et,

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

Qualité de l'ajustement et coefficient de détermination ou R^2

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

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

ou,

$$\sum_{i=1}^{N} (Y_i - \overline{Y})^2 = \sum_{i=1}^{N} (\widehat{Y}_i - \overline{\widehat{Y}})^2 + \sum_{i=1}^{N} \widehat{U}_i^2$$

où $\overline{\widehat{Y}} = n^{-1} \sum_{i=1}^{N} \widehat{Y}_i$.

Qualité de l'ajustement et coefficient de détermination ou R^2

- Notons que,

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

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

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{N}\widehat{U}_i^2$$

Qualité de l'ajustement et coefficient de détermination ou R^2

ou,

$$SCT = SCE + SCR$$

où, $SCT := \sum_{i=1}^{N}(Y_i - \overline{Y})^2$ est la *somme des carrés totale*,

$SCE := \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2$ est la *somme des carrés expliqués*, et $SCR := \sum_{i=1}^{N}\widehat{U}_i^2$ est la *somme des carrés des résidus*.

Qualité de l'ajustement et coefficient de détermination ou R^2

- Le rapport de la SCE à la SCT est appelé coefficient de détermination(On l'appelle/prononce généralement "R deux") ou R^2,

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

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 $\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}} = 0$, 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.
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 $\mathbf{X}$.

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*,

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

- Le R^2 ajusté diminue la qualité de ajustement lorsque le nombre de régresseurs augmente relativement au nombre d'observations de sorte que $\overline{R}^2$ 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.

Section 3

Intervalles de confiance

Introduction

- On considère le modèle de régression normal défini par les hypothèse H1-H5.
- L'*estimateur ponctuel* $\widehat{\beta}$, n'est pas très informatif dans la mesure où $\mathbb{P}(\widehat{\beta} = \beta) = 0$.
- C'est pour cela qu'on s'intéresse 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*(e.g., 0.01, 0.05, 0.10).
- Un intervalle de confiance avec une probabilité $(1 - \alpha)$ de couvrir β est noté $\mathrm{CI}_{1-\alpha}$.

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 $\mathbf{X}_1$ est un vecteur $(n \times 1)$ contenant les valeurs observées du premier régresseur.

L'estimateur des moindres carrés de β_1 est,

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

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,

$$\mathrm{CI}_{1-\alpha} = \left[\widehat{\beta}_1 - c, \widehat{\beta}_1 + c \right] \tag{19}$$

Cas scalaire

Comme $\widehat{\beta}_1$ est une fonction de l'échantillon aléatoire, l'intervalle de confiance donné dans (19) l'est aussi.

Le problème maintenant est de choisir c tel que,

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

où $\mathbf{X} = (\mathbf{X}_1 \quad \mathbf{X}_2)$.

Pour choisir c, nous avons besoin de connaître la distribution de $\widehat{\beta}_1 | \mathbf{X}$.

Cas scalaire

Sous les hypothèses H1-H5,

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

et par conséquent,

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} | \mathbf{X} \sim \mathcal{N}(0,1) \tag{20}$$

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

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

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

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

Cas scalaire

et par conséquent,

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

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

σ^2 est connu

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

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

- Montrons maintenant que,

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

σ^2 est connu

- En effet,

$$\mathbb{P}\left(\widehat{\beta}_1 - z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} \leq \beta_1 \leq \widehat{\beta}_1 + z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}|\mathbf{X}\right)$$
$$= \mathbb{P}\left(-z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} \leq \beta_1 - \widehat{\beta}_1 \leq z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}|\mathbf{X}\right)$$
$$= \mathbb{P}\left(-z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)} \leq \widehat{\beta}_1 - \beta_1 \leq z_{1-\alpha/2}\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}|\mathbf{X}\right)$$
$$= \mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \leq z_{1-\alpha/2}|\mathbf{X}\right) \tag{21}$$

- Le résultat découle de (20), (21), et de la définition de $z_{1-\alpha/2}$.

σ^2 est inconnu

- On peut ici suivre une approche similaire à la précédente mais en remplaçant dans un premier temps σ^2 par son estimateur,

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

- Cependant $(\widehat{\beta}_1 - \beta_1)/\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}$ n'est pas normalement distribué car c'est une fonction non-linéaire des termes aléatoires $\widehat{\beta}_1$ 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,

$$\frac{(\widehat{\beta}_1 - \beta_1)}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} | \mathbf{X} \sim t_{n-K} \tag{22}$$

- 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.

σ^2 est inconnu

- Écrivons,

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} = \left(\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right) / \frac{s^2}{\sigma^2}$$

$$= \left(\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right) / \sqrt{\frac{\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}}}{\sigma^2}/(n-K)} \qquad (23)$$

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

$$\frac{\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}}}{\sigma^2} | \mathbf{X} \sim \chi^2_{n-K} \qquad (24)$$

σ^2 est inconnu

- Pour cela nous avons besoin du résultat suivant,

Lemme 3.1

- *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* $\mathrm{Rang}(A) = r \leq n$.
- *Alors* $U^\top AU \sim \chi^2_r$.

Démonstration.

(voir notes de cours) □

- A présent pour montrer (24), écrivons,

$$\frac{\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}}}{\sigma^2} = \left(\frac{\mathbf{U}}{\sigma^2}\right)^\top \mathbf{M}_\mathbf{X} \left(\frac{\mathbf{U}}{\sigma^2}\right) \tag{25}$$

où,

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

σ^2 est inconnu

- Par l'hypothèse H5,

$$\frac{\mathbf{U}}{\sigma}|\mathbf{X} \sim \mathcal{N}(0, \mathbf{I}_n) \tag{26}$$

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

$$\begin{aligned} \text{Rang}(\mathbf{M}_\mathbf{X}) &= \text{Tr}(\mathbf{M}_\mathbf{X}) \\ &= n - K \end{aligned} \tag{27}$$

- Le résultat dans (24) découle de (25), (26), (27), et du lemme 3.1.
- Finalement, montrons que $\widehat{\beta}_1 - \beta_1$ et $\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}}$ dans (23) sont indépendants conditionnellement à $\mathbf{X}$.
- Écrivons,

$$\widehat{\beta}_1 - \beta_1 = (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U})/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)$$
$$\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}} = \mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U}$$

σ^2 est inconnu

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

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

- 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 $\mathbf{X}$.

σ^2 est inconnu

- Étand donné qu'elles ne sont pas corrélées, la normalité implique qu'elles sont indépendantes.
- En conséquence, $\widehat{\beta}_1 - \beta_1$, fonction de $\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}$, et $\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}}$ sont aussi indépendants.
- Nous avons montré (22).
- 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, $t_{n-K,1-\alpha/2}$,

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

- 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 $\widehat{\beta}_1$,

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

σ^2 est inconnu

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

$$\mathrm{CI}_{1-\alpha} = \left[\widehat{\beta}_1 - t_{n-K,1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_1|\mathbf{X})}, \widehat{\beta}_1 + t_{n-K,1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_1|\mathbf{X})}\right] \tag{28}$$

UNIVERSITÉ Grenoble Alpes

Cas vectoriel

- Supposons que l'on s'intéresse aux vecteur des paramètres $\beta = (\beta_1, \beta_2, ..., \beta_K)^\top$.
- L'équation (28) 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 $\mathbb{P}(\beta \in \mathrm{CI}_{1-\alpha} | \mathbf{X}) = 1 - \alpha$.

Cas vectoriel

- Considérons une forme quadratique par rapport à $(\widehat{\beta} - \beta)$,

$$
\begin{aligned}
(\widehat{\beta} - \beta)^\top \left(\widehat{\mathrm{Var}}(\widehat{\beta}|\mathbf{X})\right)^{-1} (\widehat{\beta} - \beta)/K &= (\widehat{\beta} - \beta)^\top \left(s^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1} (\widehat{\beta} - \beta)/K \\
&= \frac{(\widehat{\beta} - \beta)^\top \left(\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1} (\widehat{\beta} - \beta)/K}{s^2/\sigma^2} \\
&= \frac{(\widehat{\beta} - \beta)^\top \left(\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)^{-1} (\widehat{\beta} - \beta)/K}{\left(\frac{\widehat{\mathbf{U}}^\top\widehat{\mathbf{U}}}{\sigma^2}\right)/(n-K)}
\end{aligned}
\tag{29}
$$

- Montrons maintenant que l'expression dans (29) 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.

Cas vectoriel

- De la discussion dans la partie précédente nous savons que, $\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}}/\sigma^2 | \mathbf{X} \sim \chi^2_{n-K}$ qui est indépendant du numérateur dans (29).
- Il résulte de cela, que nous devons montrer que

$$(\widehat{\beta} - \beta)^\top \left(\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right)^{-1} (\widehat{\beta} - \beta) | \mathbf{X} \sim \chi^2_K \tag{30}$$

- Pour cela nous avons besoin du lemme suivant,

Lemme 3.2

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.

(voir notes de cours) □

- Le résultat dans (30) découle du lemme 3.2.
- En conséquence,

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

Cas vectoriel

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

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

- La discussion précédente implique que,

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

Remarque 1

- *La région/intervalle de confiance $\mathrm{CI}_{1-\alpha}$ 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 $\mathrm{CI}_{1-\alpha}$ contienne la vraie valeur de β.*
- *D'un autre côté, la réalisation de $\mathrm{CI}_{1-\alpha}$ 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.*

Section 4

Tests d'hypothèses

Test d'une hypothèse par rapport à un seul coefficient

Nous poursuivons notre discussion sur le modèle de régression linéaire normal, i.e., le modèle défini par les hypothèses H1-H5. On considère,

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

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 $\widehat{\beta}_1$ l'estimateur des moindres carrés de β_1.

Cherchons à tester,

$$\begin{aligned} H_0 &: \beta_1 = \beta_{1,0} \\ H_1 &: \beta_1 \neq \beta_{1,0} \end{aligned} \tag{31}$$

Une règle de décision pour un test de niveau α peut reposer sur l'intervalle de confiance $\mathrm{CI}_{1-\alpha}$.

Test d'une hypothèse par rapport à un seul coefficient

Nous avons,

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

Considérons le test suivant,

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

Dans ce cas la région critique est donnée par le complément de $\mathrm{CI}_{1-\alpha}$.

On rejette ainsi H_0 si,

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

ou

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

Test d'une hypothèse par rapport à un seul coefficient

De manière équivalente, on rejette si,

$$\left| \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \right| > z_{1-\alpha/2} \tag{32}$$

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,

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} + \frac{\beta_1 - \beta_{1,0}}{\sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \tag{33}$$

Nous avons que,

Test d'une hypothèse par rapport à un seul coefficient

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

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 $z_{1-\alpha/2}$.

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

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.

Test d'une hypothèse par rapport à un seul coefficient

La probabilité de rejet est une fonction de la vraie valeur β_1 et dépend de la magnitude du deuxième terme dans (33), $|\beta_1 - \beta_{1,0}| / \sqrt{\sigma^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}$.

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,

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

Test d'une hypothèse par rapport à un seul coefficient

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 (32) soit égale à 1.88.

Pour la distribution normale standard $\mathbb{P}(Z > 1.88) = 0.03$.

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.

Test d'une hypothèse par rapport à un seul coefficient

Dans le cas où σ_2 est inconnu, on peut tester (31) en considérant la t-statistique,

$$\begin{aligned} T &= \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{s^2/(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)}} \\ &= \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_1|\mathbf{X})}} \end{aligned} \tag{34}$$

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,
$\mathbb{P}(|T| > t_{n-K,1-\alpha/2}|\mathbf{X}, H_0 \text{ est vraie}) = \alpha$.

On peut aussi considérer des tests *unilatéraux*.

Test d'une hypothèse par rapport à un seul coefficient

Dans le cas de ces tests l'hypothèse nulle et l'hypothèse alternative peuvent être spécifiées comme suit,

$$H_0 : \beta_1 \leq \beta_{1,0}$$
$$H_1 : \beta_1 > \beta_{1,0}$$

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,

$$\sup_{\beta_1 \leq \beta_{1,0}} \mathbb{P}(\text{rejeter } H_0 | \mathbf{X}, \beta_1) \leq \alpha \tag{35}$$

i.e., la probabilité maximale de rejeter H_0 quand elle est vraie ne doit pas dépasser α. Soit T telle que définie dans (34) et considérons le test suivant(règle de décision) :

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

Test d'une hypothèse par rapport à un seul coefficient

Sous H_0, nous avons,

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

Ainsi, la condition sur la taille (35) 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.

Test d'une contrainte linéaire simple

Considérons le modèle de régression linéaire normal défini par les hypothèse H1-H5,

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

Supposons que l'on souhaite tester,

$$H_0 : c^\top \beta = r$$
$$H_1 : c^\top \beta \neq r$$

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 $\beta_1 = \beta_2$.

Test d'une contrainte linéaire simple

Pour l'estimateur des moindres carrés nous avons,

$$\widehat{\beta}|\mathbf{X} \sim \mathcal{N}\left(\beta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right) \tag{36}$$

Alors,

$$\frac{c^\top\widehat{\beta} - c^\top\beta}{\sqrt{\sigma^2 c^\top(\mathbf{X}^\top\mathbf{X})^{-1}c}}|\mathbf{X} \sim \mathcal{N}(0,1) \tag{37}$$

Par conséquent, sous H_0,

$$\frac{c^\top\widehat{\beta} - r}{\sqrt{\sigma^2 c^\top(\mathbf{X}^\top\mathbf{X})^{-1}c}}|\mathbf{X} \sim \mathcal{N}(0,1) \tag{38}$$

Test d'une contrainte linéaire simple

Considérons la t statistique,

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

Sous H_0, le résultat dans (38) est vérifié.

En outre, conditionnellement à $\mathbf{X}$,

$$\mathbf{U}^\top \mathbf{M}_\mathbf{X} \mathbf{U} / \sigma^2 | \mathbf{X} \sim \chi^2_{n-K} \text{ indépendant de } \widehat{\beta} \tag{39}$$

Par conséquent sous H_0,

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

Test d'une contrainte linéaire simple

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,

$$H_0 : \beta_j = r$$
$$H_1 : \beta_j \neq r$$

On rejette H_0 si,

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

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

Tests de contraintes linéaires multiples

Supposons que l'on souhaite tester,

$$H_0 : \mathbf{R}\beta = r$$
$$H_1 : \mathbf{R}\beta \neq r$$

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, $H_0 : \beta_1 + \beta_2 = 1$, $\beta_3 = 0$.

Tests de contraintes linéaires multiples

Considérons la F statistique,

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

On peut alors montrer(voir notes de cours) que sous H_0,

$$F|\mathbf{X} \sim F_{q,n-K} \tag{40}$$

Par conséquent, le test est donné par,

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