ÉCONOMÉTRIE
(L3 MIASH, S2)

COURS 1 : RÉGRESSION LINÉAIRE ET MCO (1)

Michal W. Urdanivia*

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

20 janvier 2025

1. Modèle linéaire

2. Estimation

3. Propriétés

4. Biais de $\hat{\sigma}^2$

5. Régression partitionnée

6. Qualité de l'ajustement

Outline

1. Modèle linéaire

2. Estimation

3. Propriétés

4. Biais de $\hat{\sigma}^2$

5. Régression partitionnée

6. Qualité de l'ajustement

1. Modèle linéaire

Introduction

- 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.,

$$X_i = \begin{pmatrix} X_{1,i} \\ X_{2,i} \\ \vdots \\ X_{K,i} \end{pmatrix}$$

- Une paire (Y_i, X_i) est appelée observation(sous entendu de (Y, X)).
- Il s'agit de *données en coupe* et les observations sont dans ce cas généralement supposé i.i.d(ce que nous supposerons ici).

Introduction

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

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

et l'on note que contrairement à X_i et Y_i, l'erreur U_i n'est pas une variable observable.

Introduction

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

$$\mathrm{E}(Y_i|X_i) = X_{1,i}\beta_1 + X_{2,i}\beta_2 + ... + X_{K,i}\beta_K = X_i^{\top}\beta \tag{2}$$

où,

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

est un vecteur de K paramètres constants.

- Notons aussi que comme

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

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 $X_{k,i}$ varie, pour des valeurs fixes des autres régresseurs $X_{l,i}$, $l = 1, 2, ..., K$, $l \neq k$.

- Observons que les équations (1) et (2) permettent d'écrire,

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

où par définition de (1)

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

- 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 : $\mathrm{E}(U_i) = 0$.

- Une hypothèse fréquente sur les erreurs consiste à supposer qu'ils sont *homoscédastiques*(on parle d'hypothèse d'homoscédasticité),

$$\mathrm{Var}(U_i|X_i) = \sigma^2 \tag{5}$$

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

Outline

1. Modèle linéaire

2. Estimation

3. Propriétés

4. Biais de $\hat{\sigma}^2$

5. Régression partitionnée

6. Qualité de l'ajustement

2. Estimation

Estimateur des moments

- Notre objectif est d'estimer les paramètres $\beta_1, \ldots, \beta_K$, ainsi que σ^2 pour (3).
- Pour cela nous allons commencer par employer la méthodes de moments(MM).
- La MM consiste à construire des estimateurs pour des paramètres définis par des moments théoriques en égalisant certains moments théoriques (qui dépendent de ces paramètres) avec leurs contreparties empiriques. Par exemple si une espérance(moment d'ordre 1) dépend d'un paramètre que l'on souhaite estimer, la MM consistera à égaliser cette espérance à la moyenne empirique et résoudre l'équation qui en résulte par rapport au paramètre à estimer. La solution sera l'estimateur des moments.
- Dans le cas présents nous allons nous appuyer sur la condition suivante que l'on supposera vérifiée par nos observations :

$$\mathrm{E}(X_i U_i) = 0, \text{ et } \mathrm{E}(U_i) = 0, \text{ pour } i = 1, \ldots, n. \tag{6}$$

Estimateur des moments

- Notons que cette condition est plus forte que supposer

$$\mathrm{E}(U_i|X_i) = 0, \text{ pour } i = 1, \ldots, n. \tag{7}$$

- Sous cette condition $X_i^\top \beta$ ne peut pas s'interpréter comme une espérance conditionnelle.

- La conditions (6) est une condition sur un moment théorique qui est ici une espérance conditionnelle. En particulier sous cette condition et avec (3) nous pouvons écrire :

$$\mathrm{E}(X_i U_i) = \mathrm{E}\left(X_i(Y_i - X_i^\top \beta)\right) = 0, \tag{8}$$

qui doit donc être vérifié par les paramètres $\beta_1, \ldots, \beta_K$.

- Un *estimateur des moments*(i.e., obtenu selon la MM) de β, $\hat{\beta}$, est obtenu en remplaçant l'espérance dans (8) par la moyenne empirique,

$$n^{-1}\sum_{i=1}^{n} X_i(Y_i - X_i^\top \hat{\beta}) = n^{-1}\sum_{i=1}^{n} X_i Y_i - n^{-1}\sum_{i=1}^{n} X_i X_i^\top \hat{\beta} = 0. \tag{9}$$

Estimateur des moments

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

$$\hat{\beta} = \left(n^{-1} \sum_{i=1}^{n} X_i X_i^{\top} \right)^{-1} n^{-1} \sum_{i=1}^{n} X_i Y_i = \left(\sum_{i=1}^{n} X_i X_i^{\top} \right)^{-1} \sum_{i=1}^{n} X_i Y_i = \left(\mathbf{X}^{\top} \mathbf{X} \right)^{-1} \mathbf{X}^{\top} \mathbf{Y}. \tag{10}$$

  où $\mathbf{X}$ est une matrice $n \times K$ ayant pour ligne $i = 1, \ldots, n$ le vecteur $X_i^{\top}$, et $\mathbf{Y}$ est un vecteur $n \times 1$ ayant pour élément $i = 1, \ldots, n$ Y_i.
- 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,

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - X_i^{\top} \hat{\beta})^2 \tag{11}$$

Estimateur des moments

- **Remarques** :

1. Pour pouvoir calculer $\hat{\beta}$, la matrice $\sum_{i=1}^{n} X_i X_i^\top = \mathbf{X}^\top \mathbf{X}$ doit admettre une inverse. Ce sera le cas sous la condition suivante :

$$\mathrm{Rang}(\mathbf{X}) = K, \text{ p.s.} \tag{12}$$

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

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

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

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

Moindres carrés ordinaires

- La construction de l'estimateur précédent repose notamment sur la condition $\mathrm{E}(X_i U_i) = 0$(avec $\mathrm{E}(U_i) = 0$) et l'emploi de la méthode des moments.
- Cependant, ce même estimateur peut être obtenu dans une *logique d'ajustement* où il s'agit de considérer le problème consistant à minimiser l'erreur de prédiction quand on cherche à prédire Y_i par son espérance conditionnelle, $\mathrm{E}(Y_i|X_i)$, supposée être une fonction linéaire telle que (2). Plus précisément, $Y_i - \mathrm{E}(Y_i|X_i)$ étant l'erreur de prédiction on cherche β qui minimise un critère de perte quadratique,

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

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

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

  où $S_n(b) = n^{-1} \sum_{i=1}^{n} \left((Y_i - X_i^\intercal \beta)^2\right)$, est la contrepartie empirique de la fonction objectif $S(b)$.

Moindres carrés ordinaires

- $\hat{\beta}$ est alors appelé estimateur des *moindres carrés ordinaires*.
- Nous pouvons montrer que l'estimateur des moments de la section précédente est aussi l'estimateur des moindres carrés(voir notes de cours).

Outline

1. Modèle linéaire

2. Estimation

3. Propriétés

4. Biais de $\hat{\sigma}^2$

5. Régression partitionnée

6. Qualité de l'ajustement

3. Propriétés

Linéarité

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

Estimateur sans biais

- Ici nous devons utiliser la condition plus forte que (6), soit (7).
- Avec celle-ci et (12) et (3) on peut montrer que $\hat{\beta}$ est un estimateur sans biais de β :

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

Variance

- En supposant (7), (12) et (3) :

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

et avec des erreurs homoscédastiques(i.e., sous l'hypothèse (5)),

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

Normalité

- En supposant (7), (12), (3), des erreurs homoscédastiques, et que,

$$U_i|X_i \sim \mathcal{N}(0, \sigma^2), \tag{13}$$

on montre que,

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

Efficacité

- En supposant (7), (12), (3), des erreurs homoscédastiques, 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.

Outline

1. Modèle linéaire

2. Estimation

3. Propriétés

4. Biais de $\hat{\sigma}^2$

5. Régression partitionnée

6. Qualité de l'ajustement

4. Biais de $\hat{\sigma}^2$

Biais de $\hat{\sigma}^2$

- Nous avons vu qu'un estimateur de σ^2 est :

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

  où $\hat{\beta}$ et l'estimateur des MCO dans un modèle de régression ayant pour équation $Y_i = X_i^{\mathsf{T}}\beta + U_i$.
- Cependant $\hat{\sigma}^2$ est un estimateur biaisé de σ^2. En effet on montre que(voir cours) :

$$\mathrm{E}(\hat{\sigma}^2) = \frac{n-K}{n}\sigma^2 \tag{14}$$

Biais de $\hat{\sigma}^2$

- Cela suggère l'estimateur sans biais :

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

il résulte de (14) que,

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

Outline

1. Modèle linéaire
2. Estimation
3. Propriétés
4. Biais de $\hat{\sigma}^2$
5. Régression partitionnée
6. Qualité de l'ajustement

5. Régression partitionnée

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 $(K_2 \times 1)$.
- Concentrons nous sur un groupe de variables et leurs paramètres correspondants, par exemple $\mathbf{X}_1$ et β_1.

Régression partitionnée

- Alors on peut montrer(voir cours) que :

$$\hat{\beta}_1 = (\mathbf{X}_1^\intercal \mathbf{M}_2 \mathbf{X}_1)^{-1} \mathbf{X}_1^\intercal \mathbf{M}_2 \mathbf{Y}, \tag{15}$$

  où $\mathbf{M}_2$ est la matrice de projection sur l'espace orthogonal à l'espace $\mathcal{S}(\mathbf{X}_2)$(espace des vecteurs de $\mathbb{R}^n$ s'écrivant comme des combinaison linéaires des colonnes de $\mathbf{X}_2$).

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

- 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 $\tilde{\mathbf{X}}_1$,
  - régresser $\tilde{\mathbf{Y}}$ sur $\tilde{\mathbf{X}}_1$ pour obtenir $\hat{\beta}_1$.

Outline

1. Modèle linéaire

2. Estimation

3. Propriétés

4. Biais de $\hat{\sigma}^2$

5. Régression partitionnée

6. Qualité de l'ajustement

6. Qualité de l'ajustement

- Nous pouvons obtenir la décomposition suivante :

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^{n}\hat{U}_i^2}_{\text{SCR}}$$

R^2

- Le rapport de la *SCE* à la *SCT* est appelé coefficient de détermination[1] ou R^2,

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

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

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.
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*,

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2)$$

- Le R^2 ajusté diminue la qualité de ajustement lorsque le nombre de régresseurs est grand 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.