Les régressions linéaires

Le problème linéaire

Soit *x* et *y* denote deux variables, comme le revenu et la consommation ou l'âge et la taille d'un enfant.

Avec un échantillon $\{(y_i, x_i) \in \mathrm{R}, i = 1, \ldots, N\}$ on souhaite résoudre le problème suivant:

Trouver $\hat{a}, \hat{b} \in \mathbb{R}$ tel q $\hat{y}_i = \hat{a} + \hat{b}x_i$ soit le plus proche possible de y pour tout i.

Ce modèle est dit :

"Simple" pour un seul régresseur *x* .
"Multiple« si il y en a plusieurs.

Impossibilité d'une équation exacte :

En général il n'existe pas de couple (a, b) solution de :

$$y_1 = a + bx_1$$
$$y_2 = a + bx_2$$
$$\ldots$$
$$y_N = a + bx_N$$

Si on a plus de deux équations : ($N \geq 2$).

La régression linéaire est donc une approximation.

La méthode OLS

L'estimateur OLS (ou moindres carrés) est un co $(\hat{a}, \hat{b})$ qui minimise le problème suivant :

$$SSR(a, b) = \sum_{i=1}^{N}(y_i - a - bx_i)^2:$$

$$SSR(\hat{a}, \hat{b}) = \min_{a,b} \sum_{i=1}^{N}(y_i - a - bx_i)^2$$

Les conditions de premier ordre(FOC) de ce problème (i.e. fixer la dérivée partielle à 0) sont :

$$\frac{\partial SSR(\hat{a}, \hat{b})}{\partial a} = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i) = 0$$

$$\frac{\partial SSR(\hat{a}, \hat{b})}{\partial b} = -2\sum_{i=1}^{N} x_i(y_i - \hat{a} - \hat{b}x_i) = 0$$

(*SSR*(*a, b*) est convexe ; donc les FOCs sont suffisantes.)

Solution

Soit :

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2, \quad \overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$$

On obtient :

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{a} - \widehat{b}x_i) = \overline{y} - \widehat{a} - \widehat{b}\overline{x}$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i(y_i - \widehat{a} - \widehat{b}x_i) = \overline{xy} - \widehat{a}\overline{x} - \widehat{b}\overline{x^2}$$

Ce qui impl . .

$(\widehat{a}, \widehat{b})$

$$\begin{cases} \overline{y} &= \widehat{a} + \widehat{b}\overline{x} \\ \overline{xy} &= \widehat{a}\overline{x} + \widehat{b}\overline{x^2} \end{cases}$$

Ce système est appe. . .quations
normales.

Estimateur OLS

On multiplie la première équation par X et on soustrait la seconde :

$$\overline{xy} - \overline{x}\,\overline{y} = \widehat{b}(\overline{x^2} - \overline{x}^2) \Leftrightarrow \widehat{b} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{Cov}_N(x_i, y_i)}{\mathrm{Var}_N(x_i)}$$

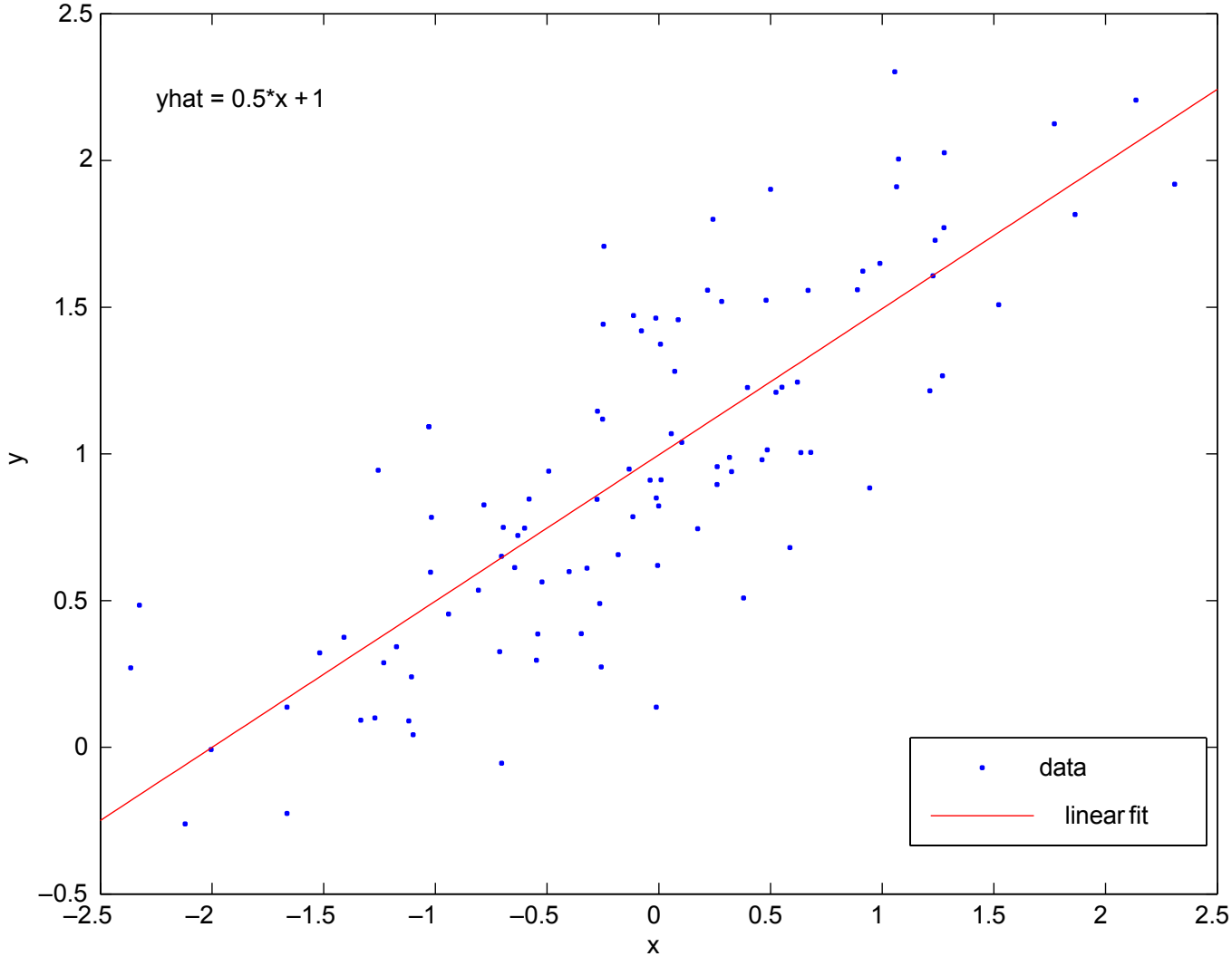
et

$$\widehat{a} = \overline{y} - \widehat{b}\overline{x}$$

Ceci est l'Estimateur « Ordinary Least Square » (OLS ou MCO en Français).

La ligne de régression : y= ax+b est une droite passant par le barycentre du graphique (x,y)

Ligne de Regression



Cas particuliers

Pas de régresseur :

$$\min_a \sum_{i=1}^{N} (y_i - a)^2 \rightarrow \widehat{a} = \frac{1}{N} \sum_{i=1}^{N} y_i = \overline{y}$$

Pas de constante (ordonnée à l'origines):

$$\min_b \sum_{i=1}^{N} (y_i - bx_i)^2 \rightarrow \widehat{b} = \frac{\frac{1}{N} \sum_{i=1}^{N} x_i y_i}{\frac{1}{N} \sum_{i=1}^{N} x_i^2} = \frac{\overline{xy}}{\overline{x^2}}$$

Centrer

On remplace chaque x et chaque y en leurs soustrayant leurs moyenne puis on régresse sans constante :

$$\widehat{b} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\mathrm{Cov}_N(x_i, y_i)}{\mathrm{Var}_N(x_i)}$$

Qui est l'estimateur de la pente de la droite de régression avec une constante.

Ajouter une constante revient au même que centrer toutes les variables de l'équation.

Réduire

On remplace chaque x et chaque y en le divisant par son écart type :

$$\widetilde{x}_i = \frac{x_i}{\sqrt{\mathrm{Var}_N\, x_i}} \text{ and } \widetilde{y}_i = \frac{y_i}{\sqrt{\mathrm{Var}_N\, y_i}}$$

La variance de x et y vaut désormais 1

Cette méthode rend notre régression invariante aux changements d'échelles (par exemple changer la mesure entre des jours et des semaines, ou d'euros en dollars).

Dans ce cas :

$$\widehat{b} = \frac{\mathrm{Cov}_N(\widetilde{x}_i, \widetilde{y}_i)}{\mathrm{Var}_N(\widetilde{x}_i)} = \frac{\mathrm{Cov}_N(\widetilde{x}_i, \widetilde{y}_i)}{1} = \frac{\mathrm{Cov}_N(x_i, y_i)}{\sqrt{\mathrm{Var}_N(x_i)}\sqrt{\mathrm{Var}_N(y_i)}} = \mathrm{Corr}_N(x_i, y_i)$$

L'OLS correspond à la corrélation.

Predictions et résidus

On définit la prédiction par $\widehat{y_i} = \widehat{a} + \widehat{b}x_i$ et le résidu $\widehat{u_i} = y_i - \widehat{y_i}$

Ainsi que : $SSR(\widehat{a}, \widehat{b}) = \sum_{i=1}^{N} \widehat{u_i}^2$.

Le résidu moyen est 0 selon la première équation normale. Cela implique que la prédiction moyenne est égale à la moyenne des yi:

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{y_i} = \widehat{a} + \widehat{b}\frac{1}{N}\sum_{i=1}^{N}x_i = \widehat{a} + \widehat{b}\overline{x} = \overline{y}$$

Correlation avec les résidus

Les équations normales impliquent l'absence de corrélation des résidus:

$$\mathrm{Cov}_N(x_i, \widehat{u}_i) = \frac{1}{N}\underbrace{\sum_{i=1}^{N} x_i \widehat{u}_i}_{=0} - \frac{1}{N}\sum_{i=1}^{N} x_i \frac{1}{N}\underbrace{\sum_{i=1}^{N} \widehat{u}_i}_{=0} = 0$$

Ce qui implique que prédiction et résidus ne sont pas corrélés:

$$\begin{aligned}\mathrm{Cov}_N(\widehat{y}_i, \widehat{u}_i) &= \frac{1}{N}\sum_{i=1}^{N} \widehat{y}_i \widehat{u}_i = \frac{1}{N}\sum_{i=1}^{N} (\widehat{a} + \widehat{b} x_i)\widehat{u}_i \\ &= \widehat{a}\frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i + \widehat{b}\frac{1}{N}\sum_{i=1}^{N} x_i \widehat{u}_i = 0\end{aligned}$$

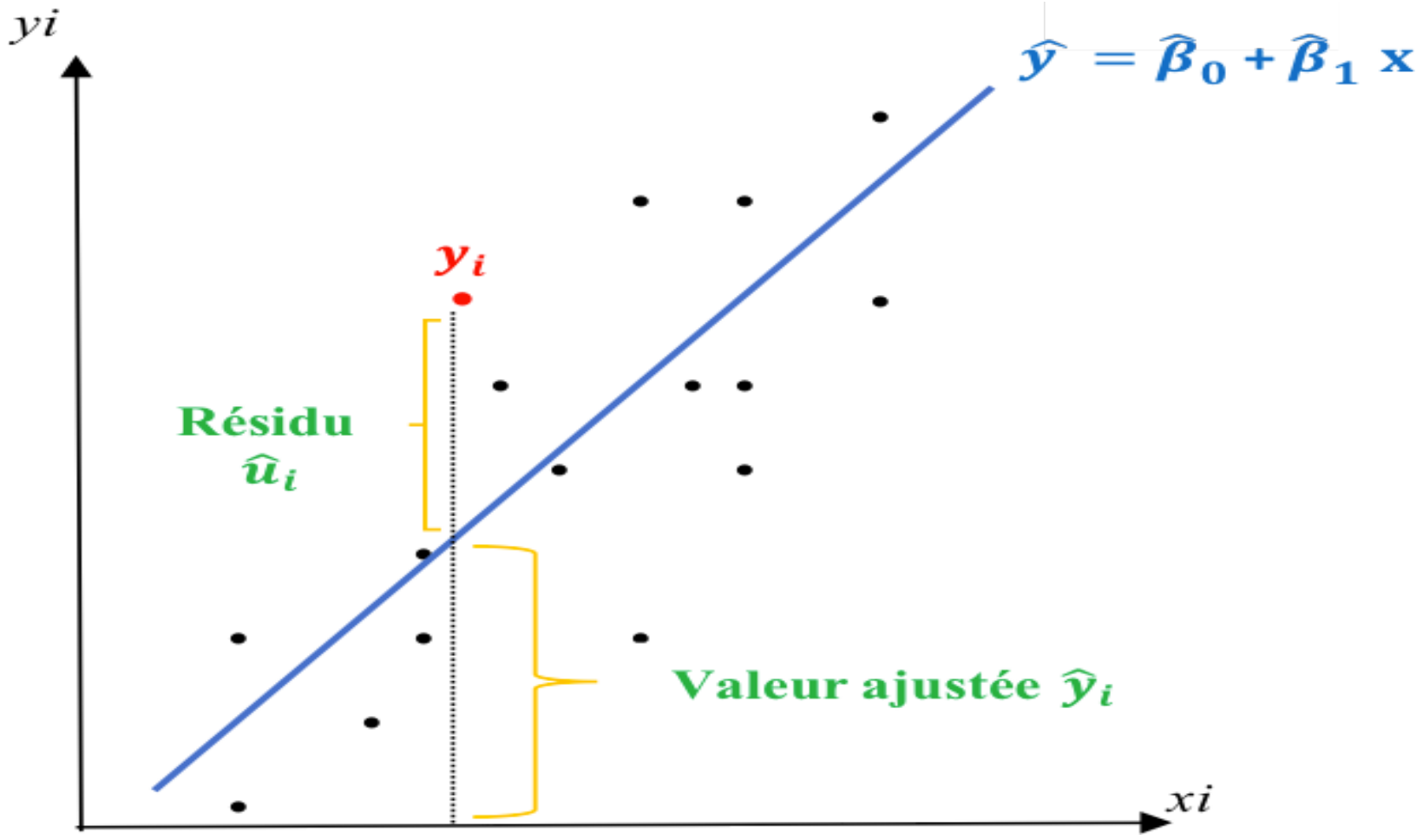
Orthogonalité et correlation

On dira que les résidus sont “orthogonaux” aux régresseurs et aux prédictions.

Formellement : le vecteur des résidus $(\hat{u}_1, \ldots, \hat{u}_N)$ est orthogonal au vecteur des 1 (1…. 1) et $(x_1, \ldots, x_N)$ des x il le sera donc aussi aux $(\hat{y}_1, \ldots, \hat{y}_N)$ des y qui est une combinaison linéaires de X et de 1. La vérification se fait par le produit scalaire :

$$(x_1, \ldots, x_N) \bullet (\hat{u}_1, \ldots, \hat{u}_N) = \sum_{i=1}^{N} x_i \hat{u}_i = 0.$$

L'ajustement linéaire en graphique



Interprétation

Que nous apprennent les coefficients a et b:

Conséquences directes :

- Si x et y sont positivement corrélés alors b est positif
- Si x et y sont négativement corrélés alors b est négatif

Interprétation, 2 manières :

- A une valeur de x, on associe une valeur de y (a+bx).
- tcepa, si x varie de 1 unité, y varie de b unités.

Nb : ceteris paribus (ou tcepa) = suppose que tous que les autres facteurs (inclus dans le terme d'erreur) soient fixes.

- estoler l'effet de la variable explicative : ce qui explique les variations de y sont les variations de x

Précision de l'ajustement

Quelle est la qualité de l'ajustement linéaire par rapport à y ?
Si $N = 2$ on peut résoudre parfaitement le système d'équations:

$$y_1 = a + bx_1$$

$$y_2 = a + bx_2$$

Mais avec $N \geq$ cela ne tient plus.

Comment évaluer la qualité de l'estimation ?

L'analyse de la variance (ANOVA)

Par définition : $y_i = \widehat{y}_i + \widehat{u}_i$ on va calculer la variance empirique de yi :

$$\begin{aligned} \mathrm{Var}_N\, y_i &= \mathrm{Var}_N(\widehat{y}_i + \widehat{u}_i) \\ &= \mathrm{Var}_N(\widehat{y}_i) + \mathrm{Var}_N(\widehat{u}_i) + 2\,\mathrm{Cov}_N(\widehat{y}_i, \widehat{u}_i) \\ &= \mathrm{Var}_N(\widehat{y}_i) + \mathrm{Var}_N(\widehat{u}_i) \end{aligned}$$

Rappel : $\mathrm{Cov}_N(\widehat{y}_i, \widehat{u}_i) = 0.$

On décompose ensuite cette variance en variance expliquée par le modèle et variance résiduelle :

- $\mathrm{Var}_N(y_i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$
- $\mathrm{Var}_N(\widehat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - \overline{y})^2$ because $\frac{1}{N}\sum_{i=1}^{N}\widehat{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}$
- $\mathrm{Var}_N(\widehat{u}_i) = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2$ because $\frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i = 0$

Decomposition en français :

Source de variation	Somme des carrés
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$

SCT = SCR + SCE

- SCT : somme des carrés totaux
- SCR : somme des carrés résiduels, non expliqués par le modèle
- SCE : somme des carrés expliqués par le modèle

Le coefficient de détermination

Le **coefficient de determination** est

$$R^2 = \frac{\text{Variance expliquée}}{\text{total variance}} = \frac{SCE}{SCT} = 1 - \frac{SC_R}{SC_T} \in [0,1]$$

Une bonne précision demande un R^2 élevé.

Le R^2 sera souvent plus élevé pour des séries temporelles que pour des données cross-sectionelles parce que l'échantillon sera plus petit.

R^2 tends vers zero quand la taille de l'échantillon augmente, donc l'interprétation doit être éclairée.

Un R^2 faible ne veut pas forcément dire que le modèle est mauvais, on préfèrera comparer des modèles similaires.

Les hypothèses classiques

- Deux types d'hypothèses
- Structurelles :
  - Sans lesquelles l'estimation n'est pas possible
  - Ou n'a pas de sens
- Comportementales:
  - Au sens du comportements des données
  - Les propriétés attendues des données
  - Qui permettent d'obtenir un estimateur désirable

- Ensemble = les hypothèses de Gauss-Markov

Les hypothèses classiques

- Les hypothèses portent :
  - Sur les données, échantillon et variables (Y, X)
  - Sur le terme aléatoire

- Ces hypothèses pèsent :
  - Sur les propriétés des estimateurs (biais, convergence)
  - Sur l'inférence statistique (loi de distribution de l'estimateur)

Les hypothèses classiques

- H1: Absence d'erreur de spécification
  - Pas d'erreur dans la liste des variables explicatives
  - La relation entre X et Y est linéaire
  - Sinon, pas d'estimation possible MCO
- H2: Echantillon aléatoire
  - Les probabilités d'inclusion/exclusion sont les mêmes pour tous les individus (sinon biais sélection …)
- H3: Y est aléatoire
  - Via le terme d'erreur
  - La seule erreur que l'on commet provient des insuffisances dans les X à expliquer les valeurs de Y

Les hypothèses classiques

- H4: le nombre d'observations dans l'échantillon doit être supérieur au nombre de paramètres à estimer
- H5 : $V(x) > 0$

  Les valeurs de X varient entre les individus,

  Sinon on ne peut pas estimer l'impact de X sur Y
- H6: $E(xi) = xi$

  Les variables explicatives sont certaines, exogènes (pas aléatoires)

Les hypothèses classiques

- H7: Cov (xi, uj)= 0
  - Les erreurs sont indépendantes des variables explicatives
  - Pas de corrélation entre x et résidus, ce qui permet d'isoler l'effet des xi
  - !!! Hypothèse d'exogénéité
- H8: E(ui)=0
  - La perturbation est d'espérance nulle
  - En moyenne les erreurs s'annulent (le modèle est bien spécifié)
  - Les aléas non pris en comptent n'affectent pas la valeur moyenne de y

Les hypothèses classiques

- H9a : V(ui)=$\sigma 2$
  - Homoscédasticité
  - La variance des perturbations est constante, i.e. la même quelles que soient les valeurs des variables explicatives
- H9b : Cov (ui, uj)= 0
  - Non autocorrélation linéaire des perturbations
  - Une erreur faite sur une observation ne dépend linéairement pas d'une erreur faite sur une autre observation
  - Surtout pour les séries chronologiques

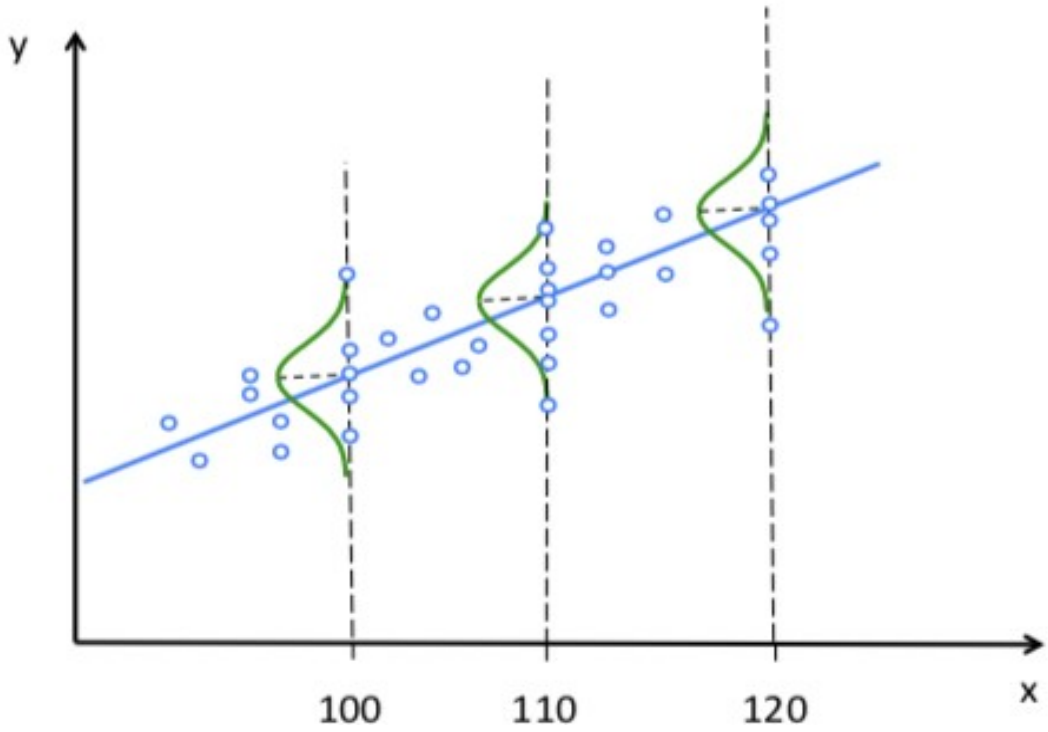
Les hypothèses classiques

- Homoscédasticité
  - La variance des résidus doit être la même tout le long de la droite de régression,
  - Elle ne doit pas varier en fonction des valeurs de X :
  - L'écart moyen des points à la droite de régression est constant

$$Var(u|x_1,x_2, \ \ldots \ ,x_k) \ = \ \sigma^2$$

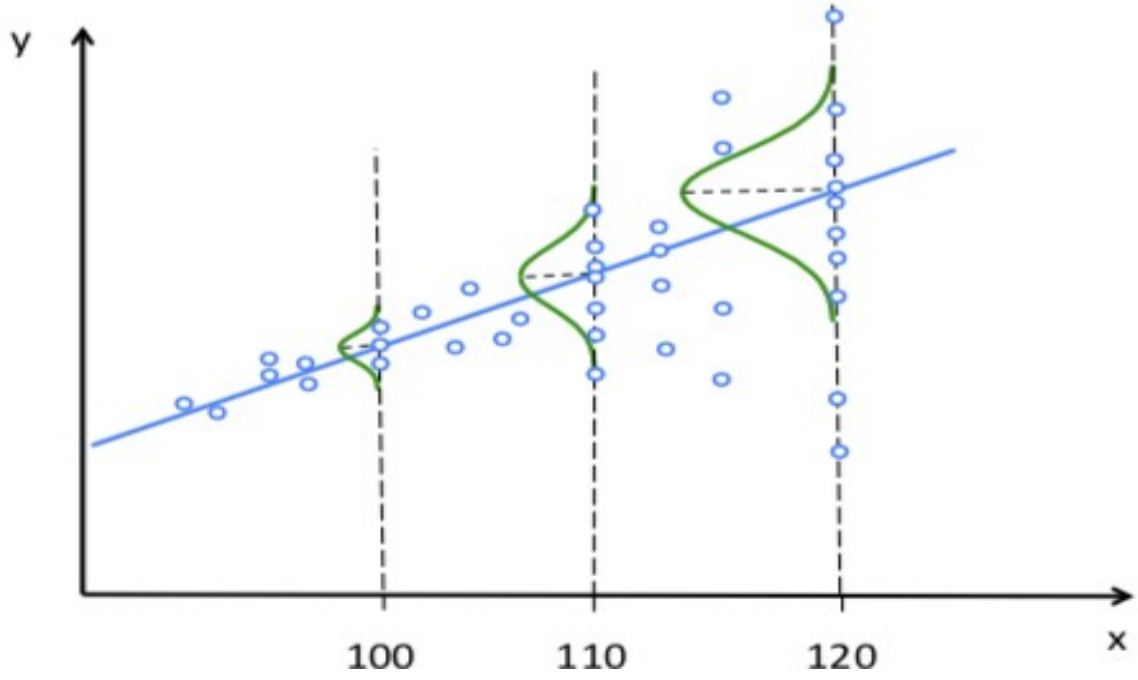
Les hypothèses classiques

V(u) est constante pour toute valeur de x : les résidus sont **homoscédastiques**



Les hypothèses classiques

V(u) n'est pas constante pour toute valeur de x : les résidus sont **hétéroscédastiques**

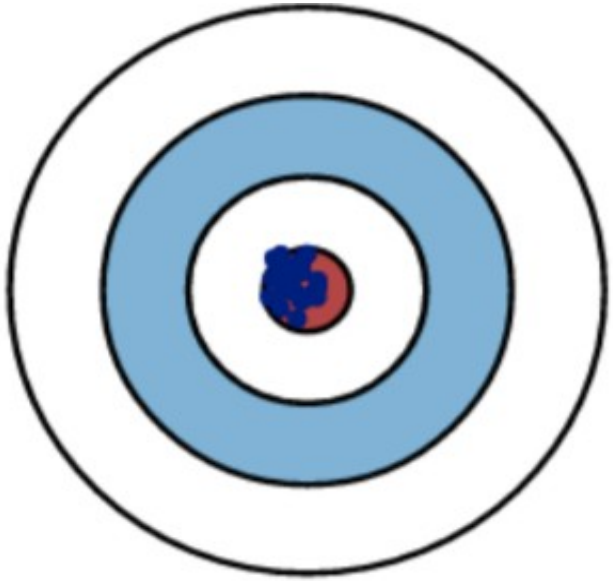


Un estimateur désirable

- Les hypothèses comportementales : des hypothèses à formuler en fonction des critères désirés pour l'estimateur

- Désirable =
  - Sans biais (espérance)
  - Précis (convergent, variance minimale)
  - De bonne famille (loi de distribution connue)

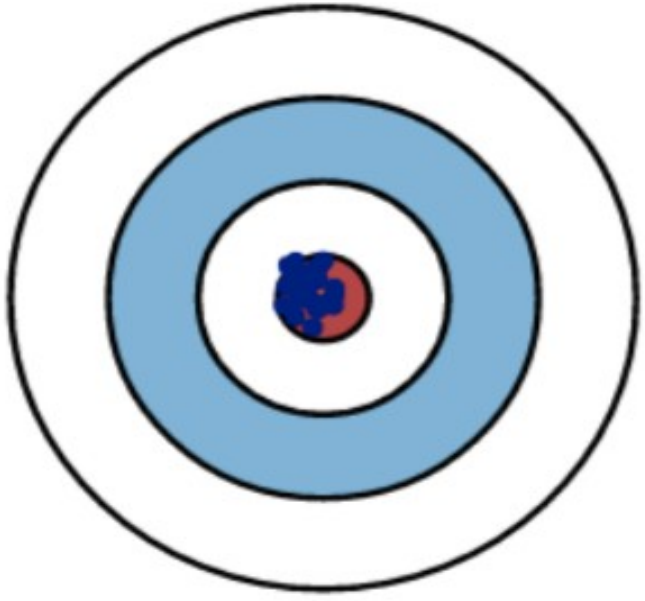
Biais et précision



Légende

- Le centre de la cible représente la « vraie » valeur β du modèle en population
- Chaque point correspond à la valeur de $\hat{\beta}$ propre à un échantillon (l'étude$_j$)

Biais et précision

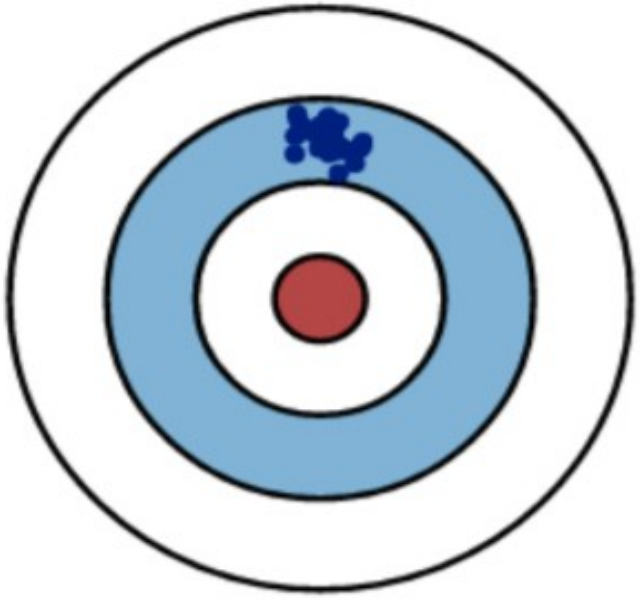


Faible biais, grande précision : l'estimateur idéal !

- Les estimations $\hat{\beta}$ sont proches de la « vraie » valeur β : $E(\hat{\beta}) \rightarrow \beta$

- Toutes ces estimations sont rapprochées : $V(\hat{\beta})$ est minimale

Biais et précision

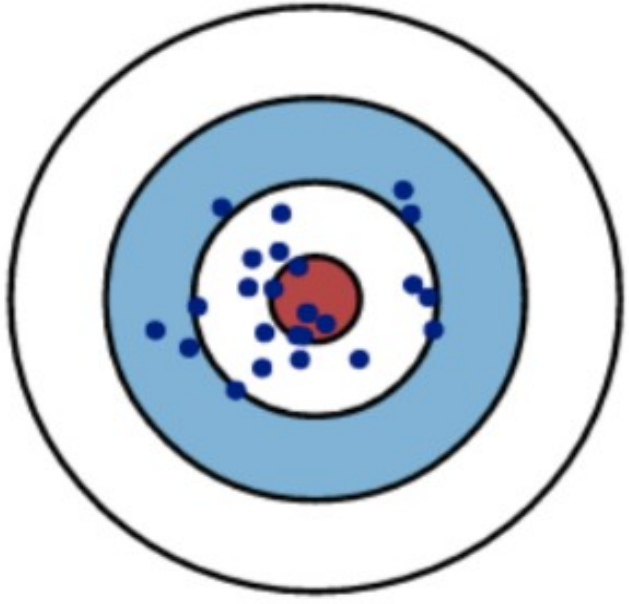


Fort biais, grande précision : l'estimateur faux que tout le monde trouve !

- Les estimations $\hat{\beta}$ sont loin de la « vraie » valeur β : $E(\hat{\beta}) \neq \beta$

- Mais toutes les estimations sont rapprochées : $V(\hat{\beta})$ est minimale

Biais et précision

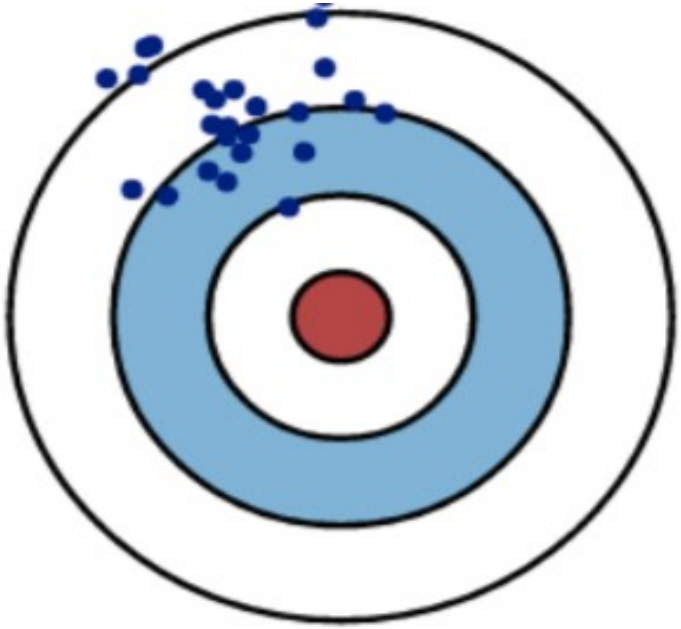


Faible biais, faible précision : l'estimateur est bon… dans une certaine mesure !

- Les estimations $\hat{\beta}$ sont proches de la « vraie » valeur β : $E(\hat{\beta}) \approx> \beta$

- Mais toutes les estimations sont éloignées : $V(\hat{\beta})$ est grande

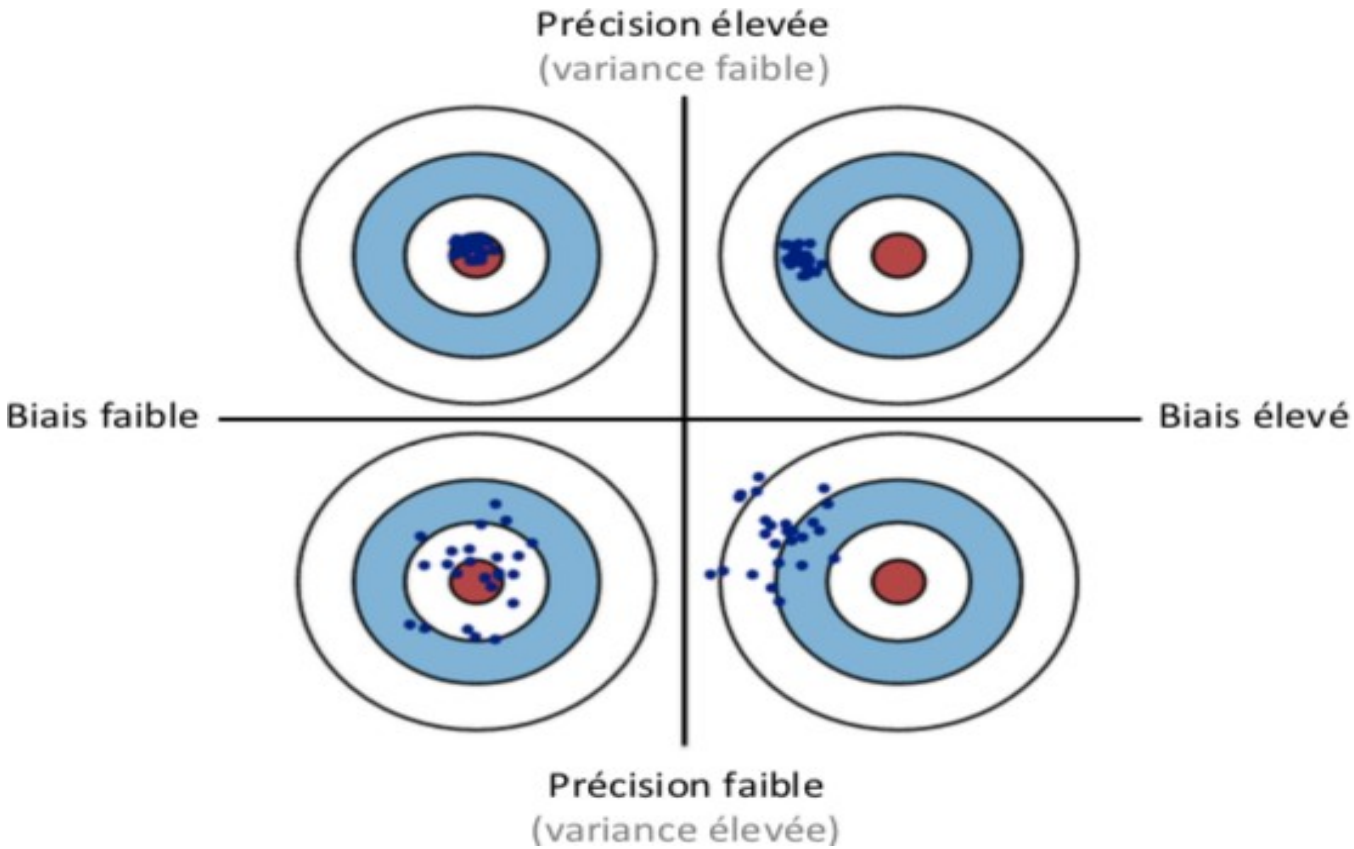
Biais et précision



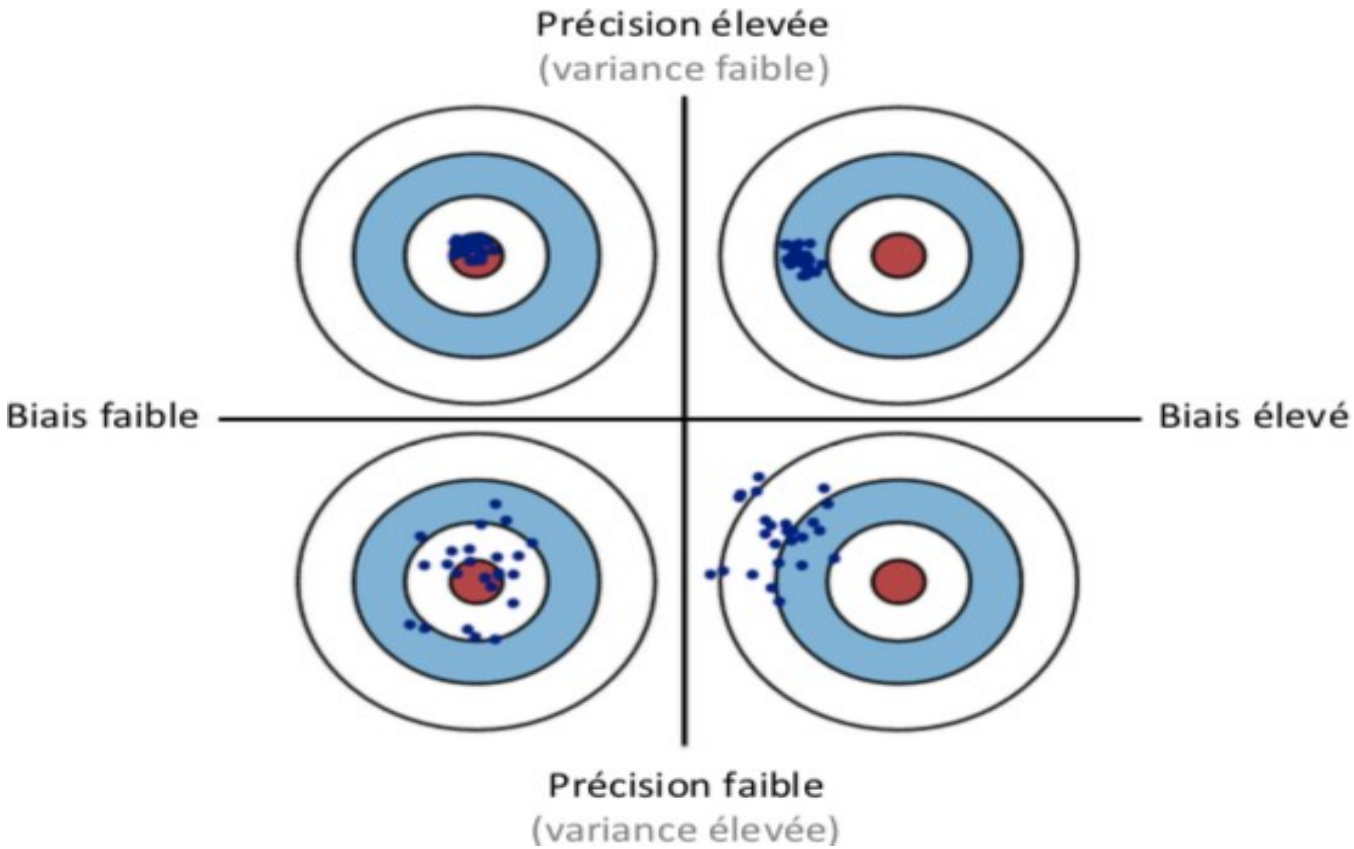
Fort biais, faible précision : le pire estimateur possible !

- Les estimations $\hat{\beta}$ sont éloignées de la « vraie » valeur β : $E(\hat{\beta}) \neq \beta$
- Et en plus, toutes les estimations sont éloignées : $V(\hat{\beta})$ est grande

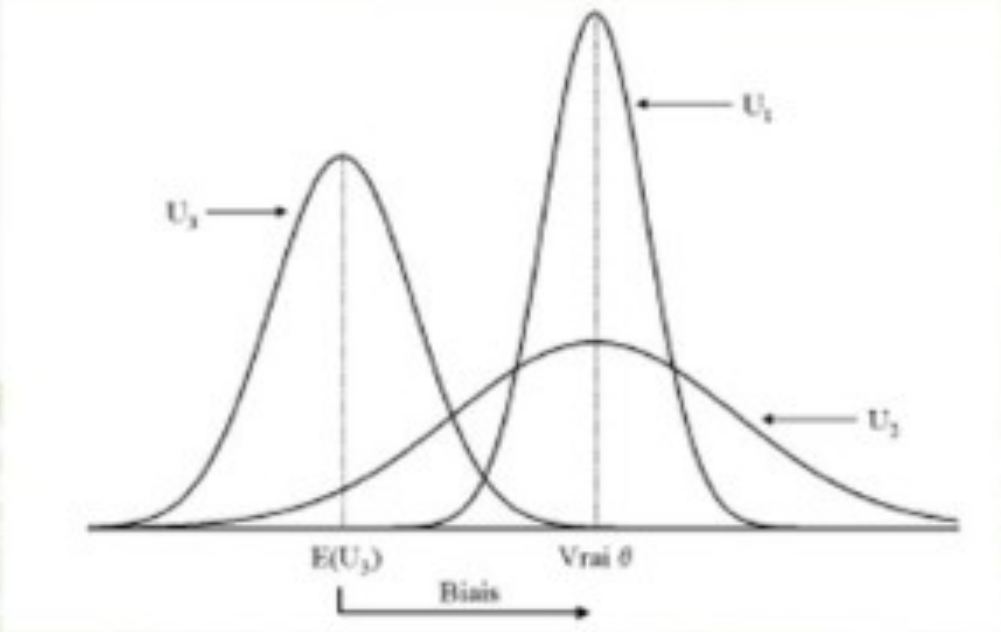
Biais et précision



Biais et précision



Biais et précision



Biais et variance pour 3 estimateurs d'un paramètre θ :

U1 et U2 sont 2 estimateurs sans biais avec Var(U1) < Var(U2)

U3 est un estimateur biaisé

Quantifier la précision

La précision $V(\hat{\beta}_1) = \frac{\sigma^2}{nV(x)}$ depend de :

La valeur de la variance des résidus = numérateur, Plus σ^2 est petite plus la précision est grande.

La taille de l'échantillon = dénominateur Plus n est grande plus $V\beta 1$ est petite, plus la précision est grande

i.e. C'est bien un estimateur convergent

La taille de la variance de x :

- Plus $V\,x$ est grande (x dispersées), plus V $\beta 1$ est petite et plus la précision est grande.

Quantifier la précision

$$V(\hat{\beta}_1) = \frac{\sigma^2}{nV(x)}$$

- Repose sur H9
- Cette formule est fausse si l'hypothèse d'homoscédasticité n'est pas vérifiée

- L'homoscédasticité permet de simplifier la formule de la précision
  et d'obtenir une variance minimale et donc l'estimateur le plus précis

Estimateur BLUE

- Best = variance minimale, le plus précis
- Linear = estimateur linéaire
- Unbiased = sans biais
- Estimator = des MCO

- Il réunit 3 conditions : c'est un estimateur linéaire, sans biais, qui a la plus petite variance
- Sous certaines hypothèses : relation linéaire, X exogènes et variance >0, espérance des perturbations nulle, homoscédasticité, etc.

Théorème Gauss-Markov

- Sous certaines hypothèses …
- Dans un modèle linéaire dans lequel les erreurs ont une espérance nulle, ne sont pas corrélées et de variance égales, le meilleur estimateur linéaire non biaisé des coefficients est l'estimateur MCO
  - Le meilleur estimateur non biaisé d'une combinaison linéaire des coefficients est son estimateur par les MCO
- Un des plus importants en économétrie
  - Avec les théorèmes asymptotiques (la LLN et le CLT)

Prédire

Avec x_{N+1} comment prédire

$$\widehat{y}_{N+1} = \widehat{a} + \widehat{b}x_{N+1}.$$

Prédiction dans et hors échantillon

In-sample prediction: on choisit x_{N+1} qui est observé dans l'échantillon.

La capacité du modèle à estimer tout les x présents dans l'échantillon est une mesure de sa qualité, cependant cela n'est pas une mesure absolue mais plutôt relative. Par exemple si un trader peut prédire les mouvements de bourse un petit peu mieux que ses compétiteurs il gagnera beaucoup d'argent

Out-of-sample prediction: lorsque x_{N+1} n'a pas encore été observée .Risqué!

Une manière de faire est de couper le jeu de données en deux et de prédire la 2eme moitié avec la première et le modèle, puis de comparer.

Liens

https://fr.wikipedia.org/wiki/Conditions_d%27optimalit%C3%A9#:~:text=On%20parle%20de%20conditions%20du,ordre%20si%20ces%20conditions%20font

https://economictheoryblog.com/2014/11/05/proof/