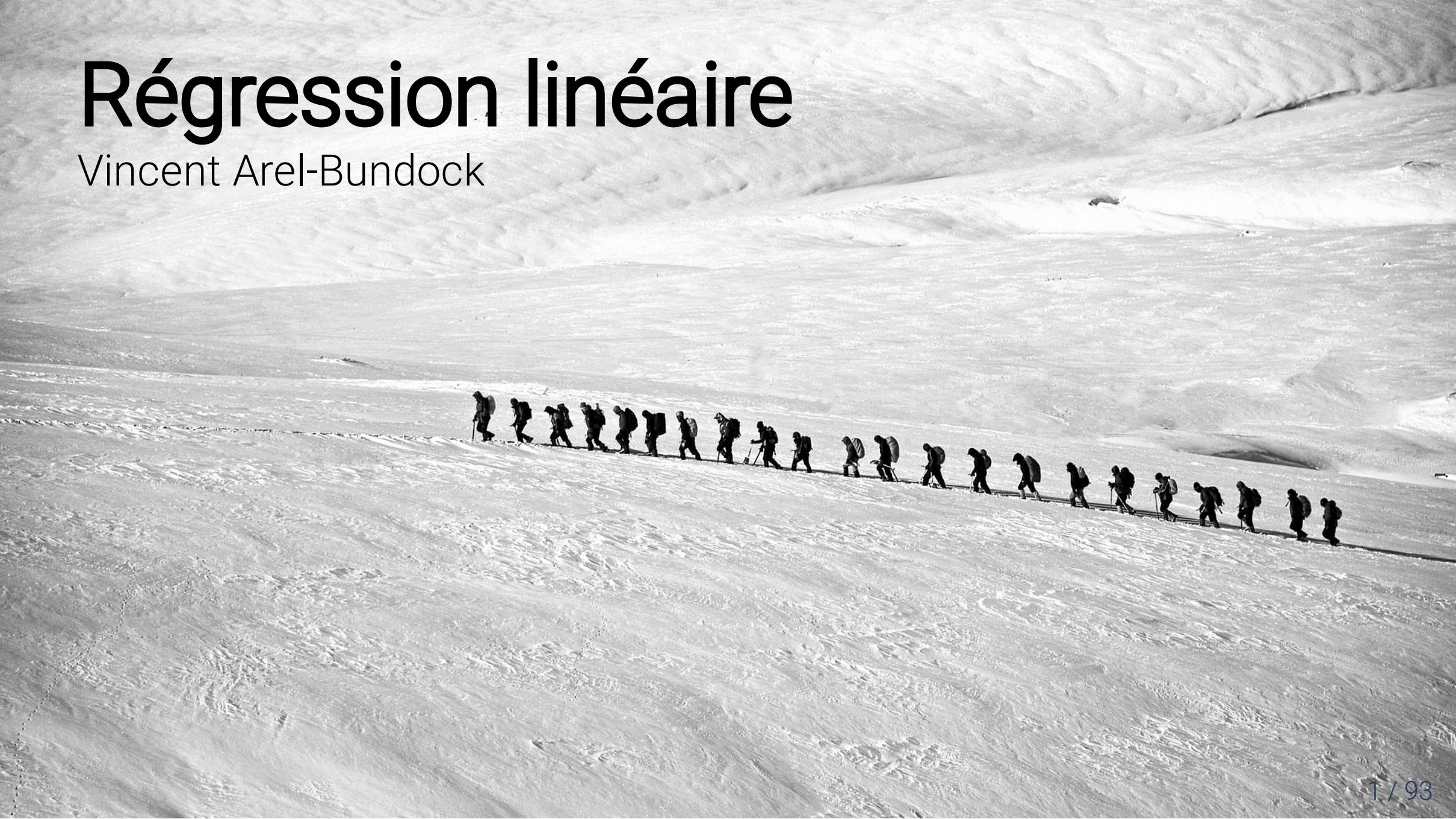


Régression linéaire

Vincent Arel-Bundock



Plan

1. Équation linéaire
2. Relation entre 2 variables
3. Modèle linéaire
4. Moindres carrés ordinaires
5. Biais
6. Incertitude
7. Test d'hypothèse
8. Régression multiple
9. Boîte à outils



Modèle linéaire

Objectif:

- Résumer la relation entre deux variables

Stratégie:

- Équation d'une droite

Rappel:

- $Y = a + b \cdot X$
 - a: ordonnée à l'origine, ou constante
 - b: pente, ou coefficient

Modèle linéaire: Composantes

Relation entre la taille d'une personne et la taille de sa mère:

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Modèle linéaire: Composantes

Relation entre la taille d'une personne et la taille de sa mère:

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

- 118: Constante
- 0,3: Coefficient

Modèle linéaire: Prédictions

Taille d'une personne = $118 + 0,3 \cdot$ Taille de sa mère

Modèle linéaire: Prédiction

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Si une mère mesure 160 cm, la grandeur prédictive de son enfant à l'âge adulte est 166 cm:

$$166 = 118 + 0,3 \cdot 160$$

Modèle linéaire: Prédictions

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Si une mère mesure 160 cm, la grandeur prédictive de son enfant à l'âge adulte est 166 cm:

$$166 = 118 + 0,3 \cdot 160$$

Si une mère mesure 161 cm, la grandeur prédictive de son enfant à l'âge adulte est 166,3 cm:

$$166,3 = 118 + 0,3 \cdot 161$$

Modèle linéaire: Prédictions

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Si une mère mesure 160 cm, la grandeur prédictive de son enfant à l'âge adulte est 166 cm:

$$166 = 118 + 0,3 \cdot 160$$

Si une mère mesure 161 cm, la grandeur prédictive de son enfant à l'âge adulte est 166,3 cm:

$$166,3 = 118 + 0,3 \cdot 161$$

Une augmentation de 1 cm de la taille d'une mère est associée à une augmentation de 0,3 cm de la taille de son enfant ($166,3 - 166 = 0,3$).

Modèle linéaire: Erreurs

Taille d'une personne = $118 + 0,3 \cdot$ Taille de sa mère

Modèle linéaire: Erreurs

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Si une mère mesure 160 cm, la grandeur prédictive de son enfant à l'âge adulte est 166 cm:

$$166 = 118 + 0,3 \cdot 160$$

Modèle linéaire: Erreurs

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Si une mère mesure 160 cm, la grandeur prédite de son enfant à l'âge adulte est 166 cm:

$$166 = 118 + 0,3 \cdot 160$$

Si la *vraie* grandeur de l'individu est de 170cm, le modèle fait une erreur de 4cm:

$$170 - 166 = 4$$

Modèle linéaire: Prédictions et Erreurs

$$\text{Enfant} = 118 + 0,3 \cdot \text{Mère}$$

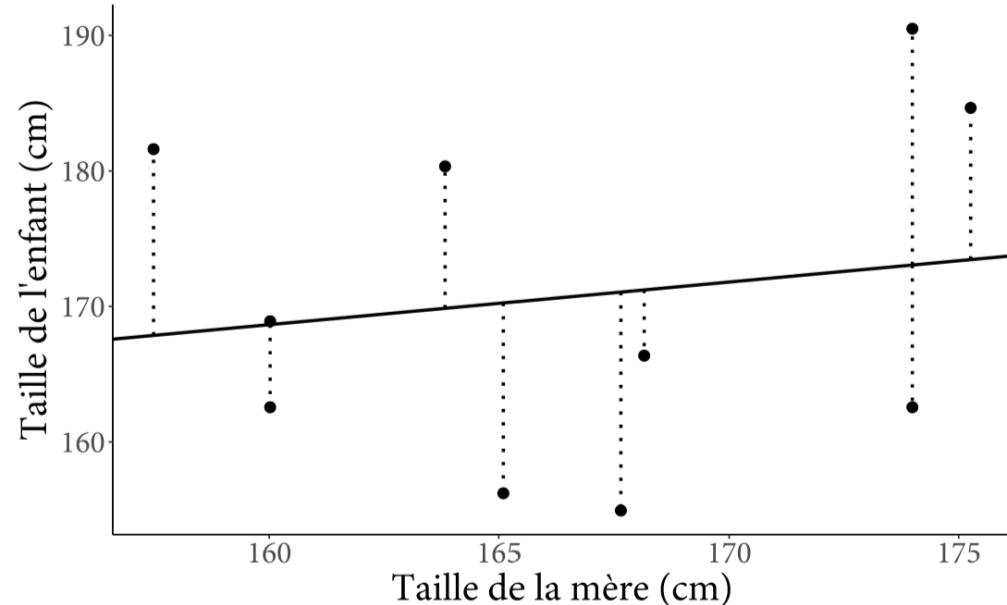
Banque données de 10 observations

Individu	Enfant	Mère	Prédiction	Erreur	Erreur carrée
1	163	174	170,2	-7,2	51,8
2	163	160	166,0	-3,0	9,0
3	169	160	166,0	3,0	9,0
4	156	165	167,5	-11,5	132,2
5	182	157	165,1	16,9	285,6
6	185	175	170,5	14,5	210,2
7	190	174	170,2	19,8	392,0
8	155	168	168,4	-13,4	179,6
9	180	164	167,2	12,8	163,8
10	166	168	168,4	-2,4	5,8

Modèle linéaire: Graphiquement

Mêmes données, représentées graphiquement

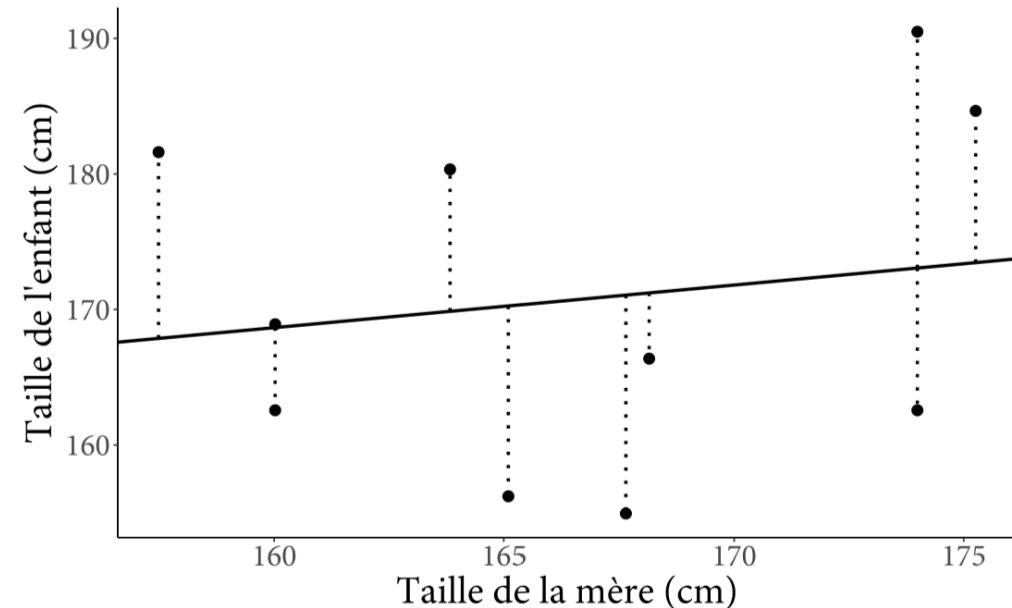
Individu	Enfant	Mère	Prédiction	Erreur	Erreur carrée
1	163	174	170,2	-7,2	51,8
2	163	160	166,0	-3,0	9,0
3	169	160	166,0	3,0	9,0
4	156	165	167,5	-11,5	132,2
5	182	157	165,1	16,9	285,6
6	185	175	170,5	14,5	210,2
7	190	174	170,2	19,8	392,0
8	155	168	168,4	-13,4	179,6
9	180	164	167,2	12,8	163,8
10	166	168	168,4	-2,4	5,8



Modèle linéaire: Graphiquement

- Points: Tailles observées
- Droite: Tailles prédictes
- Pente: Coefficient (0,3)
- Ordonnée à l'origine: Constante (118)
- Pointillés: Erreurs de mesure

$$\text{Enfant} = 118 + 0,3 \cdot \text{Mère}$$



Modèle linéaire: Équation générale

Le modèle de régression linéaire bivariée prend cette forme générale:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon,$$

où

- Y est la variable dépendante ou la variable à expliquer.

Modèle linéaire: Équation générale

Le modèle de régression linéaire bivariée prend cette forme générale:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon,$$

où

- Y est la variable dépendante ou la variable à expliquer.
- X est la variable indépendante ou explicative.

Modèle linéaire: Équation générale

Le modèle de régression linéaire bivariée prend cette forme générale:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon,$$

où

- Y est la variable dépendante ou la variable à expliquer.
- X est la variable indépendante ou explicative.
- β_1 est la pente de la droite ou le coefficient de régression.
 - $\beta_1 > 0$, la droite de prédiction a une pente positive.
 - $\beta_1 < 0$, la droite de prédiction a une pente négative.

Modèle linéaire: Équation générale

Le modèle de régression linéaire bivariée prend cette forme générale:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon,$$

où

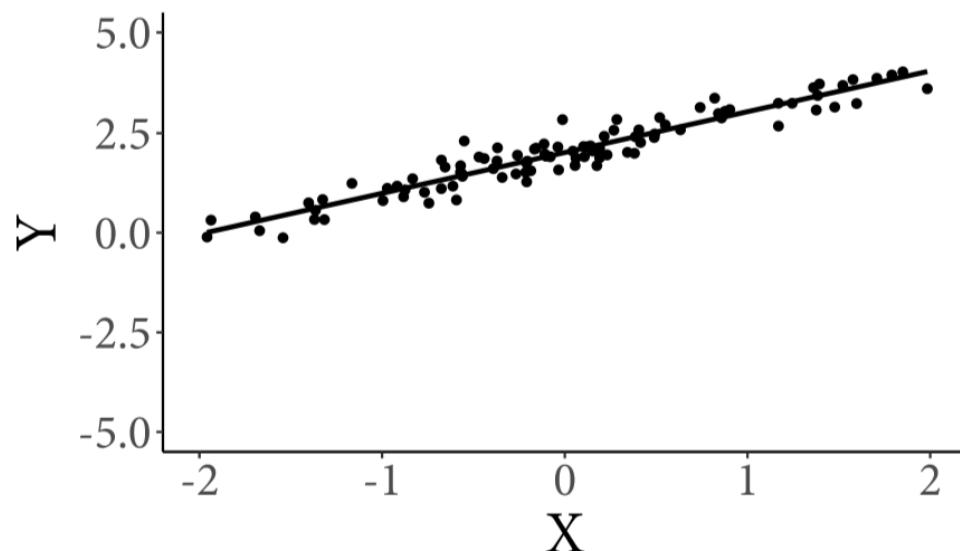
- Y est la variable dépendante ou la variable à expliquer.
- X est la variable indépendante ou explicative.
- β_1 est la pente de la droite ou le coefficient de régression.
 - $\beta_1 > 0$, la droite de prédiction a une pente positive.
 - $\beta_1 < 0$, la droite de prédiction a une pente négative.
- ε
 1. « Erreur de prédiction »: différence entre les prédictions du modèle et les valeurs observées de la variable Y .
 2. « Bruit »: la valeur de Y est déterminée par des facteurs aléatoires qui ne peuvent pas être capturés.
 3. « Résidu »: toutes les variables qui déterminent la valeur de Y mais qui sont ignorées par le modèle sont reléguées au terme résiduel.

Modèle linéaire: Graphiquement

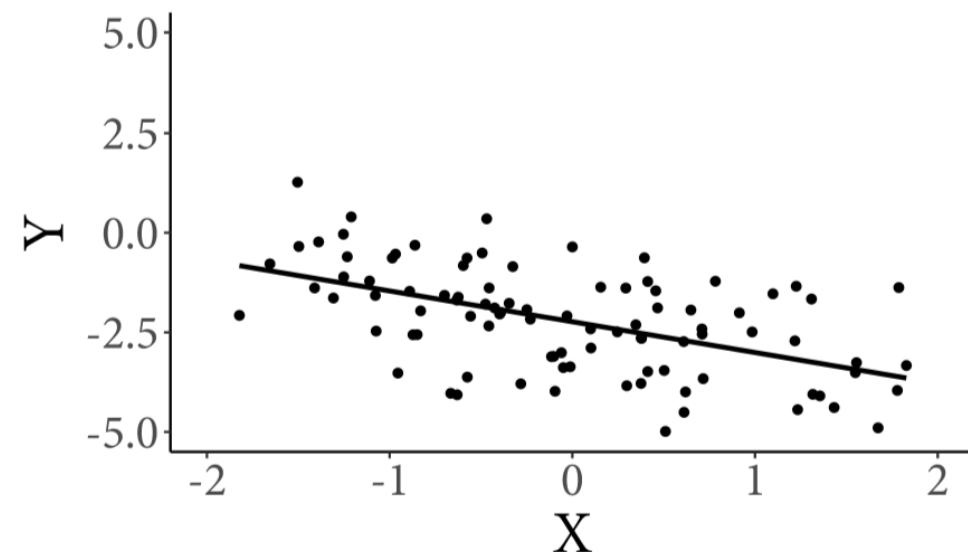
GRAPHIQUE 6.2. –

Deux modèles de régression linéaire. Les points représentent les observations faites sur les variables X et Y. Les lignes correspondent aux droites de prédictions obtenues par la méthode des moindres carrés ordinaires.

$\beta_0 > 0, \beta_1 > 0$, petit ε .



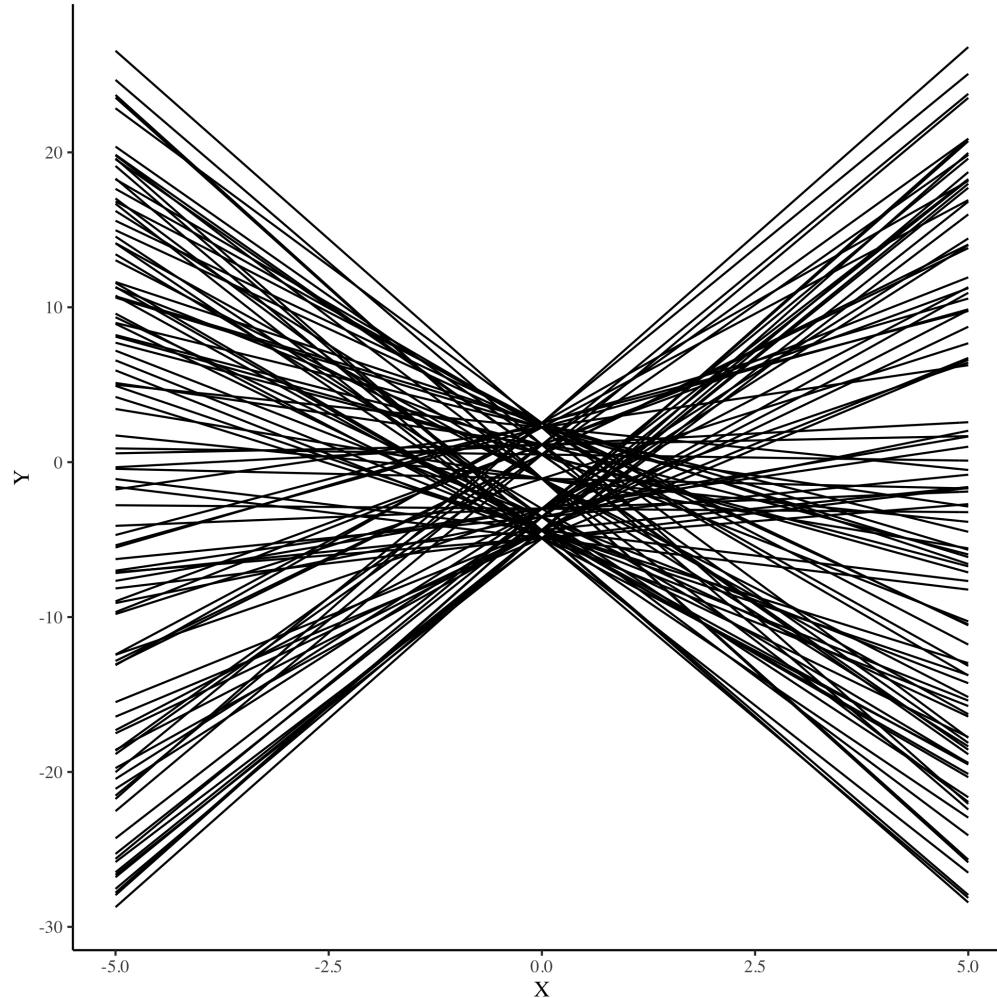
$\beta_0 < 0, \beta_1 < 0$, grand ε .



The background of the image is a wall constructed from rectangular stones of various sizes and shades of brown, tan, and reddish-brown. The stones are laid in a staggered pattern. At the bottom of the image, there is a small cluster of green plants growing in a dark, textured soil or rock crevice.

Moindres carrés ordinaires

Modèle linéaire optimal



Problème:

- Il existe une infinité de modèles linéaires.
Comment choisir le "meilleur"?

Solution:

- La méthode des moindres carrés ordinaires.

Modèle linéaire optimal

Erreurs de prédition ε :

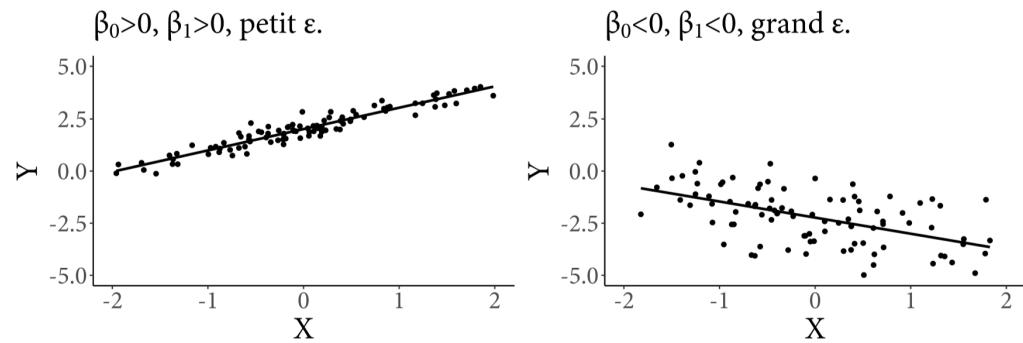
- Différence entre la taille observée d'un individu (un point) et la taille prédite pour cet individu (la droite).

Modèle linéaire optimal:

- Droite qui passe au coeur du nuage de point
- Droite qui minimise les erreurs de prédictions (élevées au carré)

GRAPHIQUE 6.2. –

Deux modèles de régression linéaire. Les points représentent les observations faites sur les variables X et Y. Les lignes correspondent aux droites de prédictions obtenues par la méthode des moindres carrés ordinaires.



Modèle linéaire optimal

Erreurs de prédiction ϵ :

- Différence entre la taille observée d'un individu (un point) et la taille prédictive pour cet individu (la droite).

Modèle linéaire optimal:

- Droite qui passe au coeur du nuage de point
- Droite qui minimise les erreurs de prédictions (élevées au carré)

Individu	Enfant	Mère	Prédiction	Erreur	Erreur carrée
1	163	174	170,2	-7,2	51,8
2	163	160	166,0	-3,0	9,0
3	169	160	166,0	3,0	9,0
4	156	165	167,5	-11,5	132,2
5	182	157	165,1	16,9	285,6
6	185	175	170,5	14,5	210,2
7	190	174	170,2	19,8	392,0
8	155	168	168,4	-13,4	179,6
9	180	164	167,2	12,8	163,8
10	166	168	168,4	-2,4	5,8

Méthode des moindres carrés ordinaires

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

Méthode des moindres carrés ordinaires:

- Identifie la droite de prédiction qui minimise: $\sum \varepsilon^2$
- β_0 : Constante optimale
- β_1 : Coefficient optimal

Méthode des moindres carrés ordinaires

Si \bar{Y} et \bar{X} représentent les moyennes des variables Y et X , alors ces formules permettent de trouver la constante et le coefficient qui minimisent la somme des erreurs de prédiction élevées au carré ($\sum \varepsilon^2$):

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

Méthode des moindres carrés ordinaires

Si \bar{Y} et \bar{X} représentent les moyennes des variables Y et X , alors ces formules permettent de trouver la constante et le coefficient qui minimisent la somme des erreurs de prédiction élevées au carré ($\sum \varepsilon^2$):

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

Lorsque $\text{Cov}(Y, X)$ est positive, $\hat{\beta}_1$ est positif.

Lorsque $\text{Cov}(Y, X)$ est négative, $\hat{\beta}_1$ est négatif.

Interprétation

- Quand $X = 0$, la valeur prédite (c.-à-d. moyenne) de Y est β_0 .
- Une augmentation d'une unité de X est associée à un changement de $\hat{\beta}_1$ unités de Y .

Important!

Biais



Biais

Rappel:

- Si nous estimions le même modèle dans un très grand nombre d'échantillons différents, est-ce que le modèle produirait la bonne réponse *en moyenne*?
- Un estimateur est biaisé s'il produit la mauvaise réponse *en moyenne*.

Distinction:

- β_1 : Le *vrai* coefficient
- $\hat{\beta}_1$: Le coefficient *estimé*

Un estimateur est non biaisé si:

$$E[\hat{\beta}_1] = \beta_1$$

Biais: Variables omises

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

ε représente la somme de toutes les variables qui déterminent la valeur de Y , mais qui sont ignorés par le modèle. Ce sont les "variables omises".

Ici, X est la seule variable explicative incluse dans le modèle. Tous les autres déterminants de Y le reste sont des "variables omises".

Biais: Variables omises

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

ε représente la somme de toutes les variables qui déterminent la valeur de Y , mais qui sont ignorés par le modèle. Ce sont les "variables omises".

Ici, X est la seule variable explicative incluse dans le modèle. Tous les autres déterminants de Y le reste sont des "variables omises".

- Y : Taille d'un individu
- X : Taille de sa mère
- ε :
 - Taille du père
 - Sexe
 - Nutrition
 - etc.

Biais: Condition à remplir

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

Pour que β_1 soit non biaisé, il faut que la variable explicative X soit indépendante des variables qui sont ignorées par le modèle :

$$X \perp \varepsilon$$

Biais: Exemple 1

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

Pour que β_1 soit non biaisé, il faut que la variable explicative X soit indépendante des variables qui sont ignorées par le modèle :

$$X \perp \varepsilon$$

Exemple:

- Y : Taille d'un individu
- X : Taille de sa mère
- ε :
 - Sexe
 - Nutrition
 - Taille du père
 - etc.

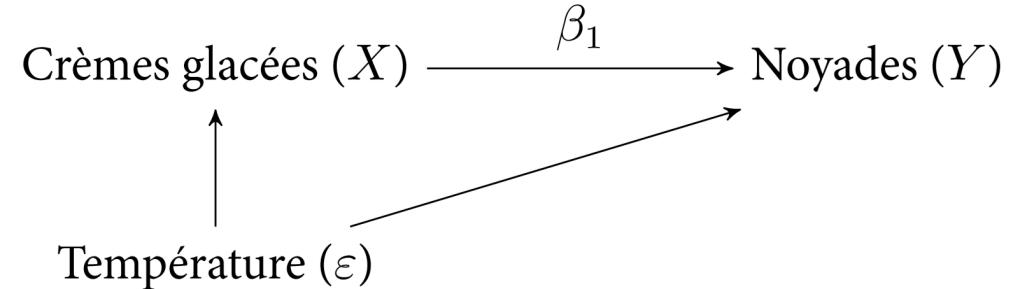
Biais: Exemple 2

Estimer la relation entre le nombre de crèmes glacées vendues et le nombre de noyades mensuelles au Canada :

$$\text{Noyades} = \beta_0 + \beta_1 \text{Crèmes glacées} + \varepsilon$$

Ce modèle ignore l'effet de la température sur le nombre de noyades : plus il fait chaud, plus les gens nagent, et plus ils se noient.

Plus il fait chaud, plus les gens consomment des produits rafraîchissants comme la crème glacée.



Biais: Condition à remplir

Lorsque $X \not\perp \varepsilon$,

- β_1 , l'estimé de la relation entre X et Y , est biaisé.
- On a un "biais par variable omise."
- Il est impossible de tirer des conclusions causales du modèle de régression.

La condition d'indépendance, $X \perp \varepsilon$, est essentielle pour éviter les biais par variable omise, les biais de sélection, les biais de mesure, et les biais de simultanéité.

4 chapitres du livre portent sur cette conditions.

Incertitude



Incertitude

Est-ce que notre estimé du coefficient de régression serait très différent si nous estimions le même modèle dans un nouvel échantillon? Pour répondre à cette question,

Variance échantillonnale du coefficient de régression:

$$Var(\hat{\beta}_1)$$

Erreur type:

$$\sigma_{\hat{\beta}_1} = \sqrt{Var(\hat{\beta}_1)}$$

Ceci nous permettra de calculer:

- Statistique t
- Valeur p
- Intervalle de confiance
- Hypothèse nulle.

Erreur type

Rappel:

- Racine carrée de la variance échantillonnale.
- Incertitude scientifique qui découle du hasard échantillonnal.
- Dispersion des coefficients qui seraient obtenus si l'analyste estimait un même modèle dans un très grand nombre d'échantillons aléatoires distincts.

Sous certaines conditions, l'erreur type de $\hat{\beta}_1$ est égale à :

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_X \cdot \sqrt{n}}$$

où

- n : taille de l'échantillon
- $\hat{\sigma}_X$: écart type de X
- $\hat{\sigma}_\varepsilon$: écart type des erreurs de prédiction

Erreur type: 3 intuitions

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_X \cdot \sqrt{n}}$$

L'incertitude est plus grande lorsque

1. L'échantillon est petit (n).
2. Le modèle fait de grandes erreurs de prédiction (σ_ε).
3. La variable explicative contient peu de variation, et donc peu d'information (σ_X).

Quand n est petit, nos résultats sont sensibles au hasard échantillonnaux, et nos conclusions pourraient facilement être dues à des facteurs aléatoires, plutôt que systématiques.

Erreur type

Ce tableau rassemble de l'information sur la taille de 10 personnes et sur la taille de leurs mères :

Individu	Enfant	Mère	Prédiction	Erreur	Erreur carrée
1	163	174	170,2	-7,2	51,8
2	163	160	166,0	-3,0	9,0
3	169	160	166,0	3,0	9,0
4	156	165	167,5	-11,5	132,2
5	182	157	165,1	16,9	285,6
6	185	175	170,5	14,5	210,2
7	190	174	170,2	19,8	392,0
8	155	168	168,4	-13,4	179,6
9	180	164	167,2	12,8	163,8
10	166	168	168,4	-2,4	5,8

- $\hat{\sigma}_\epsilon$ est l'écart type des erreurs dans la 5^e colonne du tableau
- $\hat{\sigma}_X$ est l'écart type de la variable explicative dans la 3^e colonne
- n est égal à 10

Statistique t

Rappel:

- La statistique t permet de mesurer à quel point les données observées sont "compatibles" avec une hypothèse nulle.

$$t = \frac{\hat{\beta}_1 - H_0}{\hat{\sigma}_{\hat{\beta}_1}}$$

où

- $\hat{\beta}_1$ est l'estimé du coefficient de régression
- H_0 est l'hypothèse nulle que l'analyste tente de rejeter
- $\hat{\sigma}_{\hat{\beta}_1}$ est l'erreur type du coefficient

Statistique t

$$t = \frac{\hat{\beta}_1 - H_0}{\hat{\sigma}_{\hat{\beta}_1}}$$

Souvent, l'hypothèse nulle qui nous intéresse est qu'il n'existe pas de relation linéaire entre X et Y ($\beta_1 = 0$):

- Un médicament n'a aucun effet sur les symptômes d'une maladie
- Jouer à des jeux vidéos n'a aucun effet sur le comportement violent.

Lorsque l'hypothèse nulle est égale à zéro, la statistique t correspond au ratio du coefficient de régression et de son erreur type :

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

Valeur p

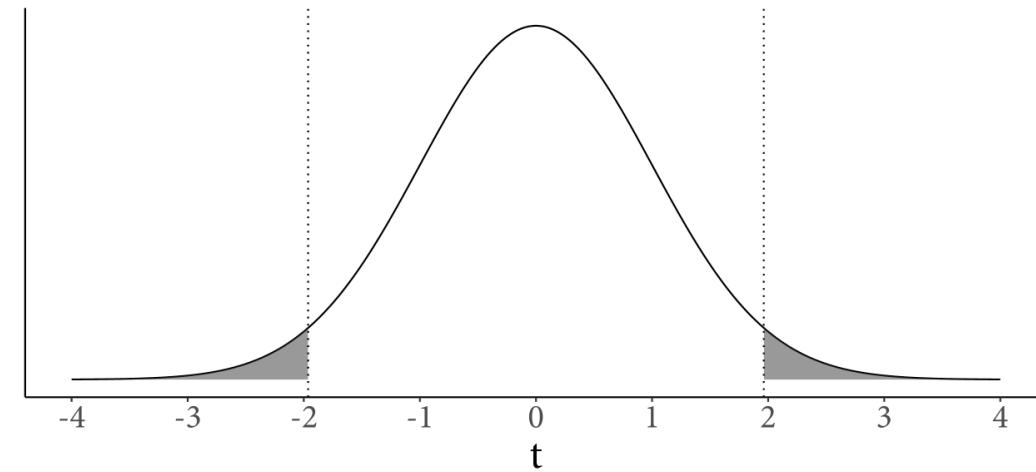
Procédure hypothétique:

1. Tirer un très grand nombre d'échantillons aléatoires de 1000 observations
2. Estimer un même modèle de régression bivarié dans chacun
3. Calculer l'erreur type de chaque coefficient
4. Calculer la statistique t associée à chaque coefficient

Si l'hypothèse nulle était vraie ($\beta_1 = 0$), les statistiques t obtenues suivraient la loi de Student tracée dans le graphique 6.3.

GRAPHIQUE 6.3. –

Loi de Student avec 998 degrés de liberté. La région blanche correspond à l'intervalle $[-1,96, 1,96]$. L'aire sous la courbe dans la région blanche est égale à 0.95. L'aire sous la courbe dans les régions grises est égale à 0.05.



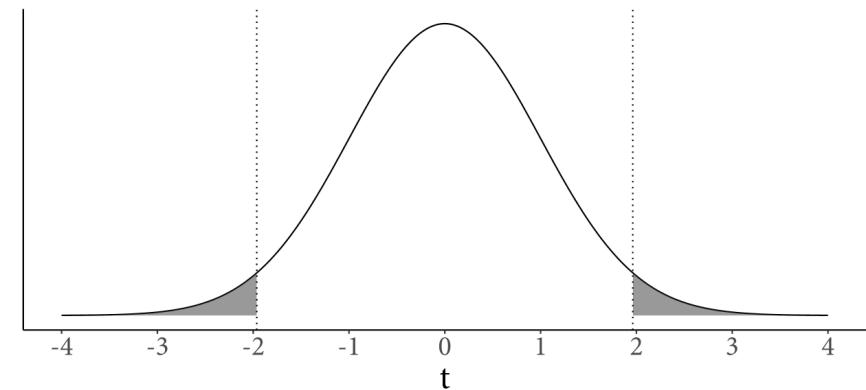
Valeur p

Valeur p :

- Probabilité d'estimer une statistique t plus extrême que la nôtre, c.-à-d. plus loin de l'hypothèse nulle, si elle était vraie.
- Probabilité d'observer un échantillon plus "étrange" que le nôtre, si l'hypothèse nulle était vraie.
- Plus p est petite, moins les données observées sont "compatibles" avec l'hypothèse nulle et plus il est justifiable de rejeter l'hypothèse nulle.

GRAPHIQUE 6.3. –

Loi de Student avec 998 degrés de liberté. La région blanche correspond à l'intervalle $[-1,96, 1,96]$. L'aire sous la courbe dans la région blanche est égale à 0.95. L'aire sous la courbe dans les régions grises est égale à 0.05.



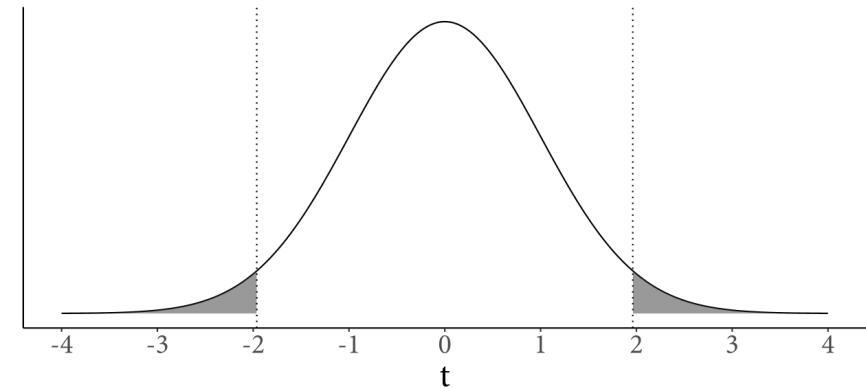
Valeur p

La valeur p mesure l'aire sous la courbe du graphique, dans les ailes à gauche de $-|t|$ et à droite de $|t|$.

Si $t = \hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1} = 1,96$, alors la valeur p est égale à la probabilité d'observer $t < -1,96$ ou $t > 1,96$, soit l'aire des régions grises.

GRAPHIQUE 6.3. –

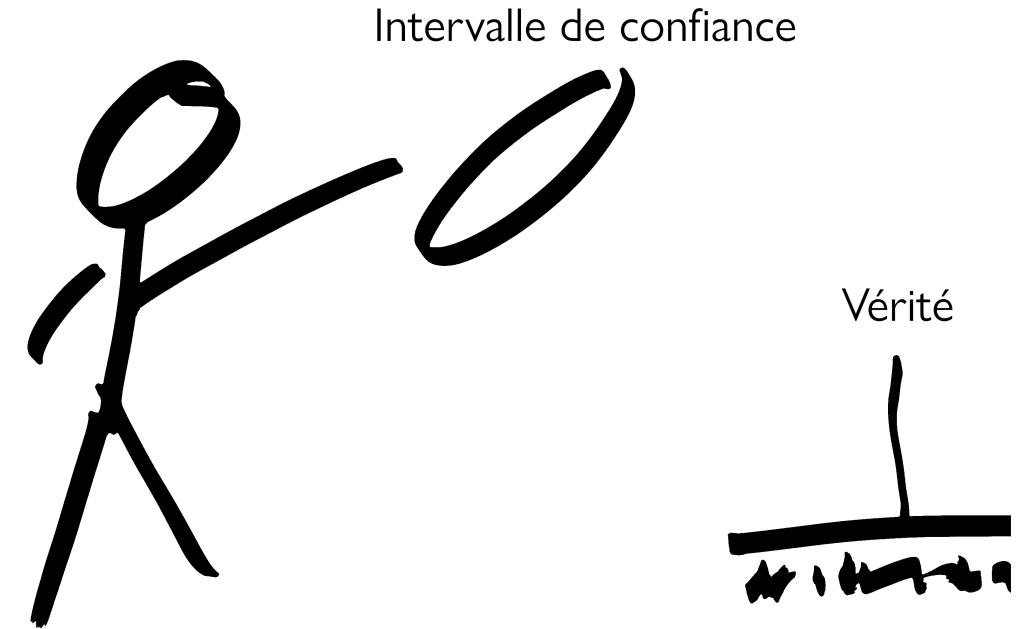
Loi de Student avec 998 degrés de liberté. La région blanche correspond à l'intervalle $[-1,96, 1,96]$. L'aire sous la courbe dans la région blanche est égale à 0.95. L'aire sous la courbe dans les régions grises est égale à 0.05.



Intervalle de confiance

Rappels:

1. Façon alternative de représenter l'incertitude qui entoure notre coefficient estimé.
2. Chaque intervalle de confiance est lié à un niveau de signification statistique spécifié *a priori*.
3. Un intervalle de confiance de 99% signale que nous sommes prêts à rejeter l'hypothèse nulle si la valeur p est inférieure à 0.01.



Intervalle de confiance

Construction:

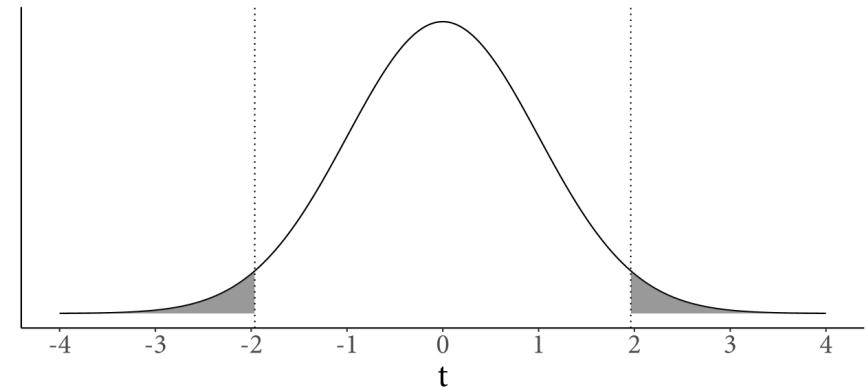
1. Identifier les seuils critiques
 - 95% central de la distribution t si H_0 était vraie
2. Multiplier par l'erreur type
3. Additionner/Soustraire à l'estimé

Intervalle de 95%:

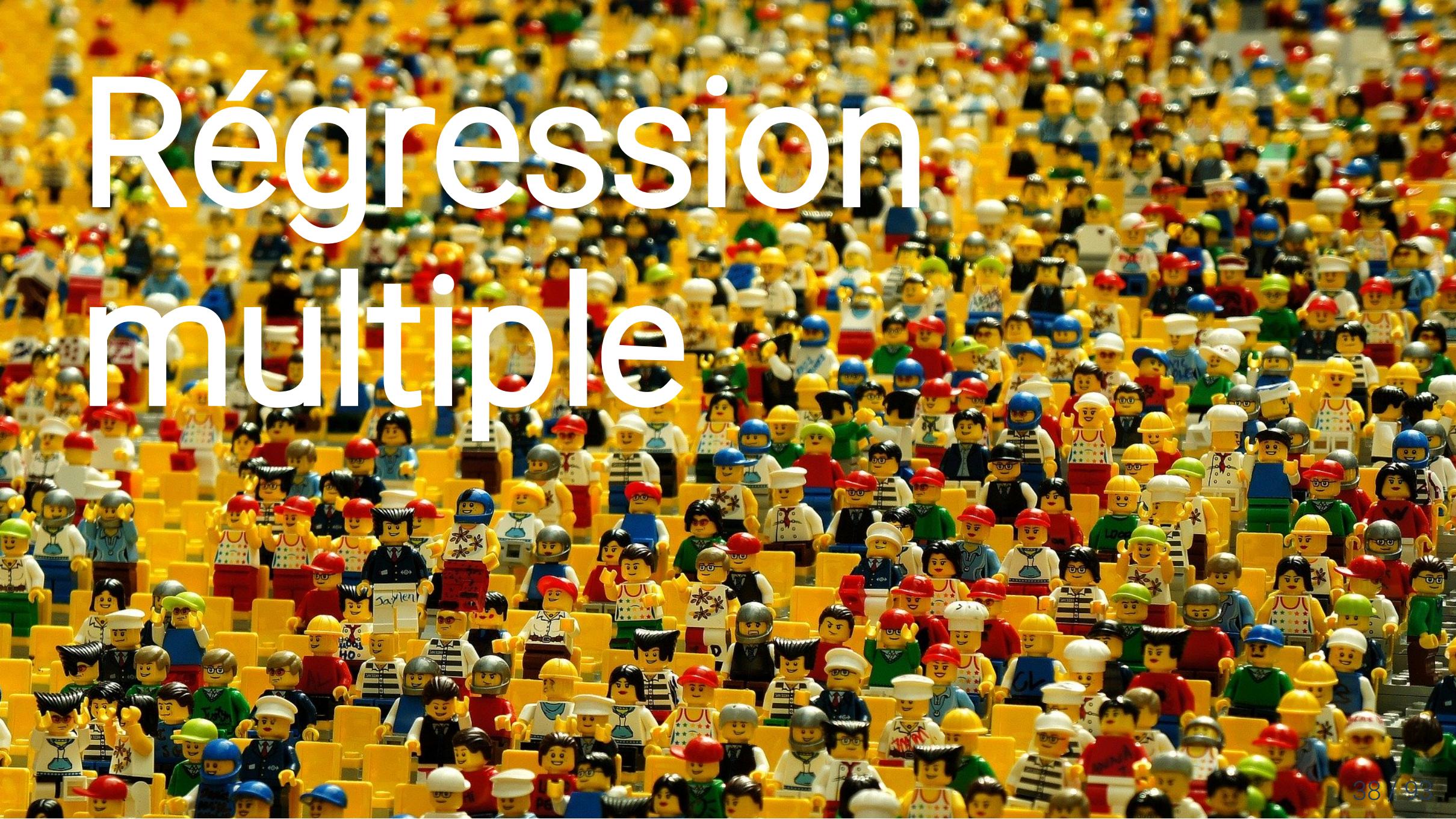
$$[\hat{\beta}_1 - 1,96 \cdot \hat{\sigma}_{\hat{\beta}_1}; \hat{\beta}_1 + 1,96 \cdot \hat{\sigma}_{\hat{\beta}_1}]$$

GRAPHIQUE 6.3. –

Loi de Student avec 998 degrés de liberté. La région blanche correspond à l'intervalle $[-1,96, 1,96]$. L'aire sous la courbe dans la région blanche est égale à 0.95. L'aire sous la courbe dans les régions grises est égale à 0.05.



Régression multiple



Régression multiple: Motivation

Modèle bivarié:

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Autres déterminants relégués au terme résiduel ε :

- Taille du père
- Nutrition
- Sexe
- etc.

Régression multiple:

- Association entre *une* variable dépendante et *plusieurs* variable indépendantes

Régression: Bivariée vs. Multiple

Régression bivariée:

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

Régression multiple:

$$\text{Taille d'une personne} = \beta_0 + \beta_1 \cdot \text{Taille de sa mère} + \beta_2 \cdot \text{Femme} + \varepsilon$$

où "Femme" est une variable binaire égale à 1 pour une femme, 0 autrement.

Estimer ce modèle à l'aide des données de Galton produit ces résultats :

$$\text{Taille d'une personne} = 120 + 0,34 \cdot \text{Taille de sa mère} - 13 \cdot \text{Femme}$$

Régression multiple: Prédictions

$$\text{Taille d'une personne} = \beta_0 + \beta_1 \cdot \text{Taille de sa mère} + \beta_2 \cdot \text{Femme} + \varepsilon$$

Ce modèle permet de faire des prédictions différentes en fonction du genre :

La taille prédite d'un homme dont la mère mesure 155cm est :

$$172,7 = 120 + 0,34 \cdot 155 - 13 \cdot 0$$

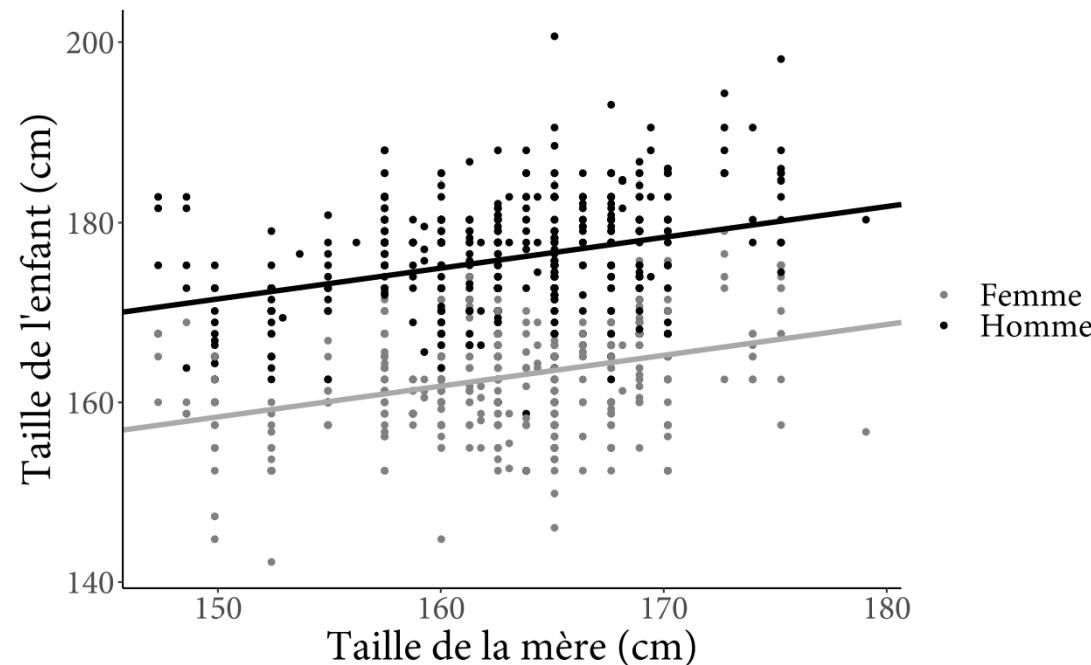
La taille prédite d'une femme dont la mère mesure 155cm est :

$$159,7 = 120 + 0,34 \cdot 155 - 13 \cdot 1$$

Régression multiple

GRAPHIQUE 6.4. –

Relation entre la taille d'une mère, la taille de son enfant à l'âge adulte, et le genre de l'enfant. La droite grise représente les prédictions du modèle pour les enfants féminins. La droite noire représente les prédictions du modèle pour les enfants masculins.



Régression multiple

Forme générale:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Ce modèle a k variables explicatives, dénotées X_1, X_2, \dots, X_k .

Chaque variable explicative est associée à un coefficient de régression distinct $\beta_1, \beta_2, \dots, \beta_k$.

Même si ce modèle inclut plusieurs variables indépendantes et plusieurs coefficients, il contient seulement une variable dépendante Y et une constante β_0 .

Régression multiple: Avantages

Régression multiple: Avantages

Précision et Biais

Régression multiple: Précision

Ajouter des variables au modèle hausse (souvent) son pouvoir explicatif.

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

$$\text{Taille d'une personne} = \beta_0 + \beta_1 \cdot \text{Taille de sa mère} + \beta_2 \cdot \text{Femme} + \varepsilon$$

Le 1er modèle risque de faire de moins bonnes prédictions que le 2e.

Régression multiple: Précision

Ajouter des variables au modèle hausse (souvent) son pouvoir explicatif.

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

$$\text{Taille d'une personne} = \beta_0 + \beta_1 \cdot \text{Taille de sa mère} + \beta_2 \cdot \text{Femme} + \varepsilon$$

Le 1er modèle risque de faire de moins bonnes prédictions que le 2e.

Petites erreurs de prédiction $\downarrow \hat{\sigma}_\varepsilon$

Régression multiple: Précision

Ajouter des variables au modèle hausse (souvent) son pouvoir explicatif.

$$\text{Taille d'une personne} = 118 + 0,3 \cdot \text{Taille de sa mère}$$

$$\text{Taille d'une personne} = \beta_0 + \beta_1 \cdot \text{Taille de sa mère} + \beta_2 \cdot \text{Femme} + \varepsilon$$

Le 1er modèle risque de faire de moins bonnes prédictions que le 2e.

Petites erreurs de prédiction $\downarrow \hat{\sigma}_\varepsilon$

Erreur type:

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_X \cdot \sqrt{n}}$$

Plus petites erreurs types!

Régression multiple: Biais

Le coefficient de régression est non biaisé lorsque la variable explicative est indépendante des variables qui causent Y , mais qui sont ignorées par le modèle :

$$X_1 \perp \varepsilon$$

Régression multiple: Biais

Le coefficient de régression est non biaisé lorsque la variable explicative est indépendante des variables qui causent Y , mais qui sont ignorées par le modèle :

$$X_1 \perp \varepsilon$$

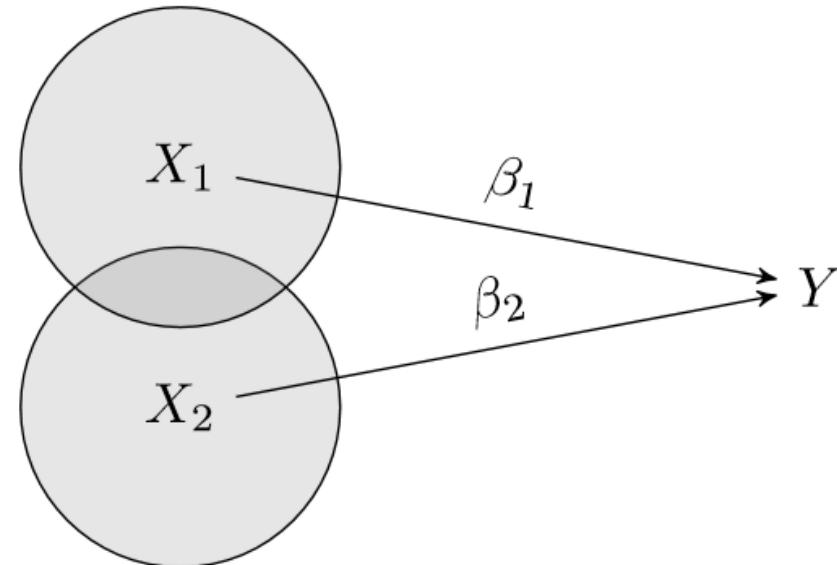
Avec la régression multiple, nous cessons d'ignorer les variables qui se cachent dans ε , et nous les intégrons directement au modèle.

Ceci nous permet de relaxer la condition d'indépendance. Pour que notre estimé de β_1 soit non biaisé, il suffit que X_1 soit indépendant d' ε , après avoir contrôlé pour les autres variables du modèle:

$$X_1 \perp \varepsilon | X_2, X_3, \dots, X_n$$

Contrôle statistique

"Contrôler": Estimer l'association entre X_1 et Y en *ignorant* la variable "commune" entre X_1 et X_2 .



Contrôle statistique

Comment un modèle de régression multiple peut-il « contrôler » pour une variable?

Il est utile de considérer un modèle simple avec une seule variable explicative:

$$Y = \pi_0 + \pi_1 X_1 + \nu$$

π_0 est la constante du modèle

π_1 est le coefficient de régression

ν est le résidu

Contrôle statistique

$$Y = \pi_0 + \pi_1 X_1 + \nu$$

Contrôle statistique

$$Y = \pi_0 + \pi_1 X_1 + \nu$$

Problème:

- Si une seconde variable X_2 cause Y , ce modèle relègue cette variable au terme résiduel ν .
- Si X_2 est liées à X_1 , nous savons que $X_1 \not\perp \nu$, et que notre estimé du coefficient π_1 est biaisé.

Contrôle statistique

$$Y = \pi_0 + \pi_1 X_1 + \nu$$

Problème:

- Si une seconde variable X_2 cause Y , ce modèle relègue cette variable au terme résiduel ν .
- Si X_2 est liées à X_1 , nous savons que $X_1 \not\perp \nu$, et que notre estimé du coefficient π_1 est biaisé.

Solution:

- Intégrer la variable X_2 directement au modèle, plutôt que de la traiter comme du bruit:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Ce modèle estime le coefficient β_1 en considérant seulement la variation *indépendante* de X_1 , c'est-à-dire la variation dans X_1 qui n'est pas « explicable » par X_2 .

Contrôle statistique: Mécanique

Pour purger X_1 de son association avec X_2 , nous procérons en deux étapes :

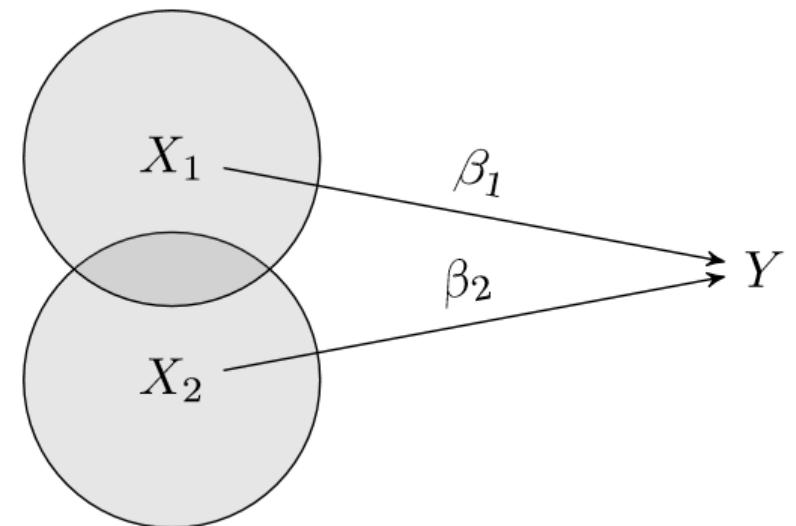
Étape 1

Estimer un modèle avec X_1 comme variable dépendante et X_2 comme variable indépendante:

$$X_1 = \lambda_0 + \lambda_1 X_2 + \tilde{X}_1$$

\tilde{X}_1 : "bruit" ou "variation dans X_1 qui ne peut pas être expliquée par X_2 ".

\tilde{X}_1 = Demi-lune!



Contrôle statistique: Mécanique

.column60-left[Étape 2

Utiliser le bruit \tilde{X}_1 dans le modèle final:

$$Y = \beta'_0 + \beta_1 \tilde{X}_1 + \varepsilon'$$

β_1 estimée par ce modèle est exactement égal au β_1 du modèle de base :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Régression multiple: Interprétation

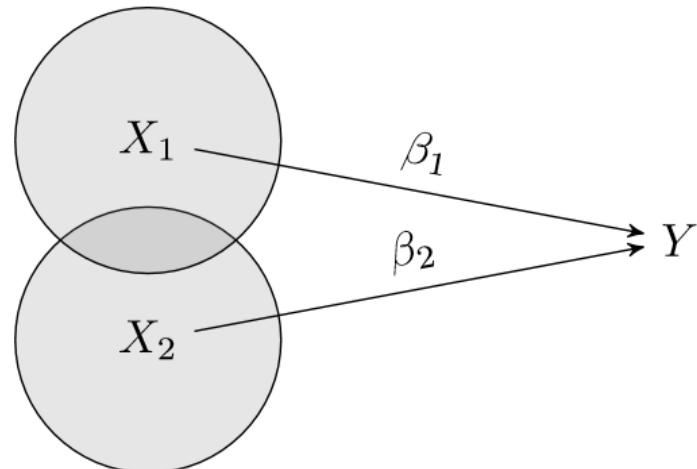
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

β_1 :

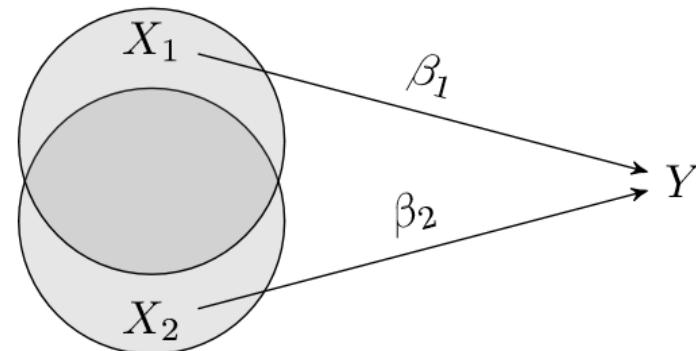
- Une mesure de l'association entre Y et X_1 après qu'on ait retranché la variation commune entre X_1 et X_2 .
- Une mesure de l'association entre Y et X_1 , après avoir contrôlé pour X_2 .
- β_1 mesure l'association entre Y et X_1 quand on tient X_2 constant.

Multicollinéarité

(A)
Faible collinéarité



(B)
Forte collinéarité

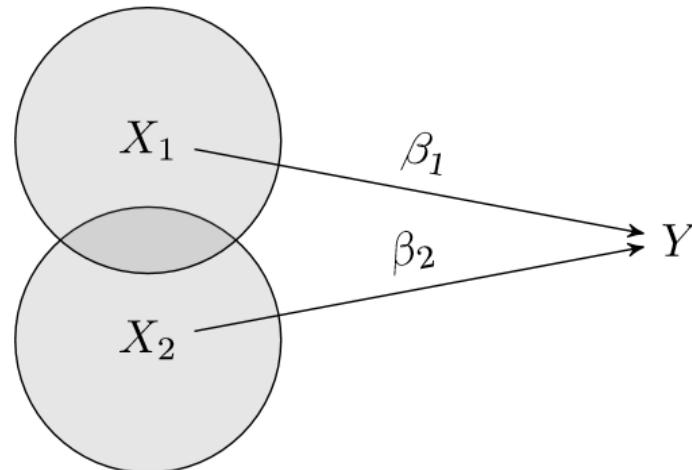


Lorsque deux (ou plusieurs) variables explicatives d'un modèle sont corrélées, on dit que le modèle souffre de « multicollinéarité ».

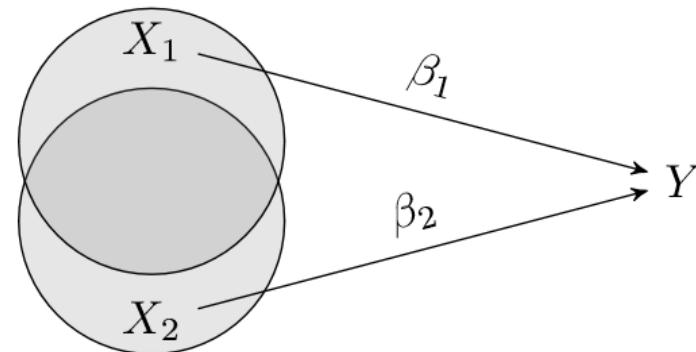
- À gauche, les variables X_1 et X_2 sont faiblement corrélées.
- À droite, les variables X_1 et X_2 sont fortement corrélées.

Multicollinéarité

(A)
Faible collinéarité



(B)
Forte collinéarité



Lorsque deux variables sont fortement associées, l'intersection du diagramme de Venne est grande, et l'information indépendante qui est disponible pour estimer le coefficient de régression est pauvre.

La pauvreté de l'information disponible se traduit par une hausse du niveau d'incertitude, et cette hausse est (correctement) capturée par une augmentation de l'erreur type.



Boîte à outils

Plan

- Variable dépendante binaire
- Variable indépendante binaire
- Variable indépendante ordinaire ou nominale
- Normalisation
- Qualité de l'ajustement statistique
- Hétéroscédasticité, autocorrélation
- erreurs types robustes
- Données influentes
- Effet marginal
- Transformations



Variables catégoriques



Variable dépendante binaire

Régression linéaire + Variable dépendante binaire = "modèle de probabilité linéaire"

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \text{ où } Y \in \{0, 1\}$$

Interprétation:

- Une augmentation d'une unité de X est associée à un changement de $100 \cdot \beta_1$ points de pourcentage dans la probabilité que Y soit égale à 1.

Alternatives:

- Logit
- Probit

Variable indépendante binaire

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \text{ où } X \in \{0, 1\}$$

Interprétation:

- β_0 : moyenne de Y quand $X = 0$;
- β_1 : différence entre la moyenne de Y quand $X = 0$ et la moyenne de Y quand $X = 1$.

Alternative:

- Test t de différence de moyennes entre deux échantillons

Variable indépendante binaire: Exemple

Dans une étude clinique sur l'aspirine, on divise les participants en deux groupes aléatoires :

- Le groupe-traitement consomme de l'acide acétylsalicylique ($X = 1$),
- Le groupe-contrôle consomme un placebo ($X = 0$).

Après une heure, on mesure l'intensité des maux de tête de chaque participant (Y).

$$\text{Mal de tête} = \beta_0 + \beta_1 \cdot \text{Aspirine} + \varepsilon, \text{ où } X \in \{0, 1\}$$

Interprétation:

- β_1 : Différence entre l'intensité moyenne des maux de tête dans les deux groupes.
- Valeur p associée à β_1 permet de vérifier si nous pouvons rejeter l'hypothèse nulle selon laquelle il n'y a pas de différence entre les deux groupes.

Variable indépendante ordinaire ou nominale

Parfois, nous pouvons traiter une variable explicative ordinale "comme si" elle était continue.

- Beaucoup de catégories
- Conceptuellement "équidistante"

Exemples:

- Niveau d'approbation du Premier Ministre: Échelle de 0 à 10
 - Passer de 0 à 1 a le même effet que passer de 8 à 9.
 - OK
- Pauvre, Classe moyenne, Riche
 - Passer de pauvre à moyen n'a pas toujours le même effet que passer de moyen à riche
 - ~~OK~~

Toujours ~~OK~~ pour les variables nominales.

Variable indépendante ordinaire ou nominale

Stratégie alternative en 2 étapes:

1. Créer une nouvelle variable dichotomique pour chaque catégorie de la variable ordinaire ou nominale:
2. Estimer un modèle de régression en incluant toutes ces nouvelles variables dichotomiques, sauf une.

Définition:

- La variable dichotomique omise s'appelle la "catégorie de référence" du modèle.

Interprétation:

- Les coefficients associés aux autres variables dichotomiques doivent être interprétés en référence à cette catégorie omise.

Variable indépendante ordinaire ou nominale

TABLEAU 6.2. —

Six passagers du Titanic, avec deux nouvelles variables dichotomiques qui serviront à estimer l'association entre la classe de la cabine d'un passager et son taux de survie.

Nom	Survie	Femme	Classe	C1	C2	C3
Wright, Mr George	0	0	1	1	0	0
Bird, Ms Ellen	1	1	1	1	0	0
Beane, Mr Edward	1	0	2	0	1	0
Mack, Mrs Mary	0	1	2	0	1	0
Asim, Mr Adola	0	0	3	0	0	1
Lang, Mr Fang	1	0	3	0	0	1

La variable ordinale "Classe" indique la classe de la cabine dans laquelle chaque passager voyage. Pour utiliser "Classe" comme variable indépendante, il faut créer trois nouvelles variables dichotomiques: "C1," "C2" et "C3."

Variable indépendante ordinaire ou nominale

Relation entre la survie et la classe de cabine :

$$\text{Survie} = \beta_0 + \beta_1 \cdot \text{C2} + \beta_2 \cdot \text{C3} + \varepsilon$$

Catégorie de référence: C1

Interprétation:

- Variable dépendante est binaire: les coefficients sont interprétés en termes de probabilité de survie.
- β_1 : association entre un changement de "C1" à "C2", et la probabilité de survie.
- β_2 , mesure l'association entre un changement de "C1" à "C2", et la probabilité de survie.

On aurait aussi pu omettre une des autres variables. Si on omettait C3, les coefficients C1 et C2 seraient interprétés en relation avec la catégorie de référence C3.

Variable indépendante ordinaire ou nominale

Le modèle suivant estime la relation entre le taux de survie et la classe de cabine :

$$\text{Survie} = \beta_0 + \beta_1 \cdot \text{C2} + \beta_2 \cdot \text{C3} + \varepsilon$$

Pour l'estimer, nous importons les données du Titanic dans R :

```
dat <- read.csv('data/titanic.csv')
```

La fonction **factor** permet de créer automatiquement une variable binaire par catégorie. Cette fonction se charge aussi d'omettre une catégorie de référence. Nous estimons donc le modèle ainsi :

```
mod <- lm(survie ~ factor(classe), data = dat)
summary(mod)
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.59938    0.02468  24.284  < 2e-16 ***
## factor(classe)2 -0.17286    0.03623  -4.772 2.03e-06 ***
## factor(classe)3 -0.40529    0.02975 -13.623  < 2e-16 ***
```

Normalisation



Normalisation

Définition:

- Transformer une variable à l'aide d'une fonction mathématique

Objectifs:

- Faciliter l'interprétation des résultats de régression.
- Certains modèles complexes sont plus faciles à estimer avec des variables transformées ou imposent certaines distributions aux données.

Types de normalisation

- À l'unité
- Student
- etc.

Normalisation à l'unité

Ramener les variables d'un modèle sur une échelle de 0 à 1:

$$\tilde{X} = \frac{X - \min_X}{\max_X - \min_X},$$

où \min_X est le minimum de la variable X , et \max_X est son maximum.

Normalisation à l'unité

Ramener les variables d'un modèle sur une échelle de 0 à 1:

$$\tilde{X} = \frac{X - \min_X}{\max_X - \min_X},$$

où \min_X est le minimum de la variable X , et \max_X est son maximum.

$$Y = \beta_0 + \beta_1 \tilde{X} + \varepsilon$$

Interprétation:

- Une augmentation du minimum jusqu'au maximum de la variable X est associée à un changement de β_1 unités de Y .

Normalisation de Student

Moyenne de 0 et écart type de 1:

$$\check{X} = \frac{X - \bar{X}}{\sigma_X},$$

où \bar{X} est la moyenne de X , σ_X son écart type, et \check{X} la version normalisée de la variable.

Normalisation de Student

Moyenne de 0 et écart type de 1:

$$\check{X} = \frac{X - \bar{X}}{\sigma_X},$$

où \bar{X} est la moyenne de X , σ_X son écart type, et \check{X} la version normalisée de la variable.

Cette transformation est utile pour interpréter nos coefficients de régression sur une échelle "neutre" qui fait abstraction de l'unité originale de mesure.

$$\check{Y} = \beta_0 + \beta_1 \check{X} + \varepsilon$$

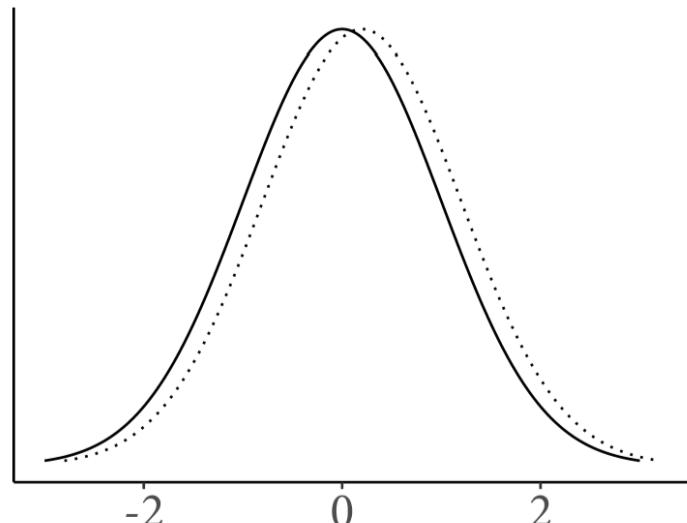
Le coefficient s'interprète ainsi : une augmentation d'un écart type dans la variable X est associée à un changement de β_1 écarts types dans la variable Y .

Normalisation de Student

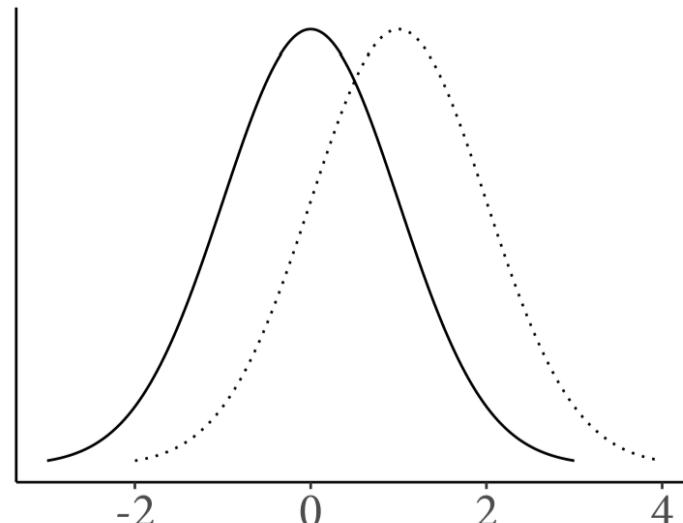
GRAPHIQUE 6.6. –

Chaque courbe représente une variable distribuée normalement. Dans le panneau de gauche, les deux variables sont décalées de 0.2 écart type. Dans le panneau de droite, les deux variables sont décalées de 1 écart type.

(a) 0.2 écart type de différen



(b) 1 écart type de différence



A close-up photograph of a man's torso and hands. He is wearing a dark grey suit jacket over a white shirt and a patterned orange tie. His hands are adjusting the knot of his tie. The lighting is dramatic, coming from the side to create strong shadows and highlights on his skin and the fabric of his suit.

Ajustement statistique

Qualité de l'ajustement statistique

Définition:

- Capacité à faire des bonnes prédictions *dans* l'échantillon

Distinction:

- Le coefficient est une mesure d'association, pas de la qualité des "prédictions"
- Prédictions *dans* l'échantillon vs. Prédictions *hors* échantillon

Alternatives:

- Erreur quadratique moyenne
- Coefficient de détermination R^2
- etc.

Erreur quadratique moyenne

Modèle linéaire bivarié:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Erreur de prédiction estimée:

$$\hat{\varepsilon} = Y - \hat{\beta}_0 + \hat{\beta}_1 X$$

Erreur quadratique moyenne

Modèle linéaire bivarié:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Erreur de prédiction estimée:

$$\hat{\varepsilon} = Y - \hat{\beta}_0 + \hat{\beta}_1 X$$

Pour un échantillon de taille n , l'erreur quadratique moyenne est:

$$\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

Plus l'erreur quadratique moyenne est faible, plus notre modèle fait de bonnes prédictions dans l'échantillon.

Coefficient de détermination: R^2

Pour le même modèle:

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2},$$

où \bar{Y} est égal à la moyenne de Y .

R^2 mesure la proportion de la variance de Y qui est "capturée" ou "expliquée" par le modèle.

Une haute valeur de R^2 signifie que le modèle est "bien ajusté" aux données de l'échantillon.

Coefficient de détermination: R^2

Attention!

- "Expliqué" \neq Causal
- $\uparrow R^2 \neq$ Non biaisé
- $\uparrow R^2 \neq$ Prédictions hors échantillon
- Ajouter des variables augmente le R^2 mécaniquement (voir le *Adj. R^2*)

Utile mais très limité.

Hétéroscédasticité, autocorrélation, et erreurs types robustes

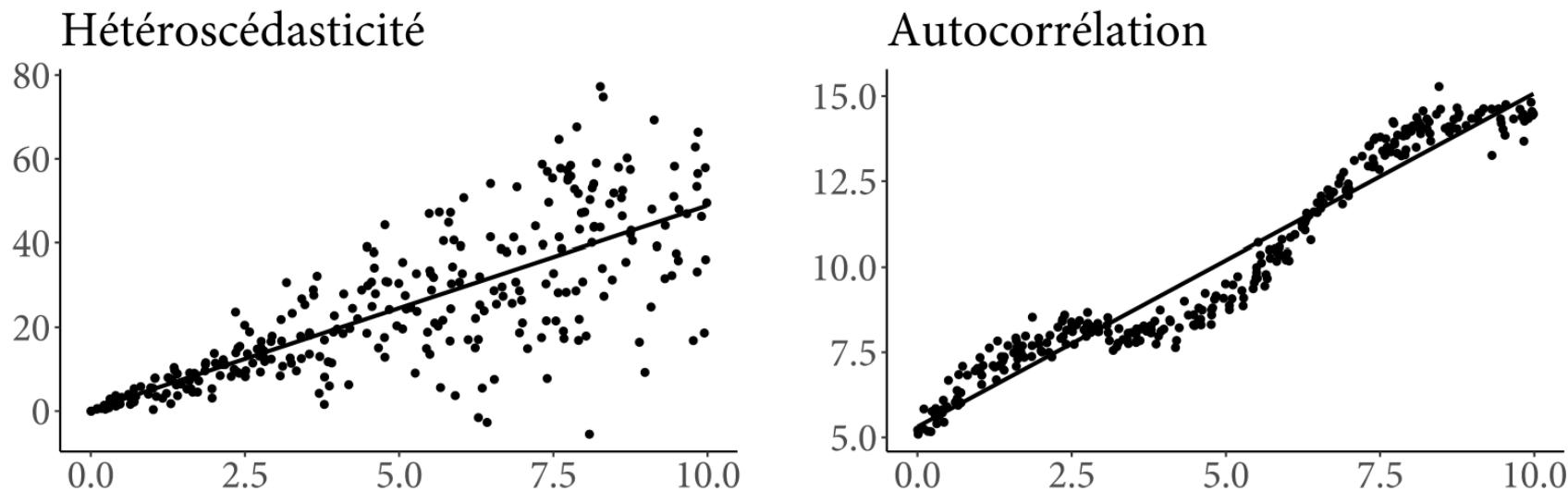
Considérons la formule qui permet d'estimer l'erreur type "classique" :

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_X \cdot \sqrt{n}}$$

Cette équation produit un estimé adéquat de l'incertitude sous deux conditions très restrictives :

1. Les erreurs de prédiction doivent être "homoscédastiques," c'est-à-dire qu'elles doivent toutes être tirées de distributions ayant la même variance.
2. Les erreurs de prédiction "classiques" ne doivent pas être corrélées entre elles.

Hétéroscédasticité, autocorrélation, et erreurs types robustes



1. Hétéroscédasticité. Les erreurs n'ont pas toutes la même variance : les observations ne sont pas dispersées de façon uniforme autour de la droite de régression.
2. L'autocorrélation survient dans plusieurs contextes, dont l'analyse de séries temporelles.

Hétéroscédasticité, autocorrélation, et erreurs types robustes

L'hétéroscédasticité peut survenir si notre modèle n'a pas le même pouvoir de prédiction pour tous les types d'observation dans notre échantillon (p. ex., autocratie vs. démocratie, Québec vs. reste du Canada).

Ces problèmes n'affectent pas les coefficients de régression, mais seulement les erreurs types.

Tous les logiciels statistiques modernes peuvent facilement calculer des erreurs types "robustes", qui tiennent compte de l'hétéroscédasticité et de l'autocorrélation.

Données influentes

La moyenne est sensible aux valeurs extrêmes ou aberrantes, alors que la médiane y est robuste.

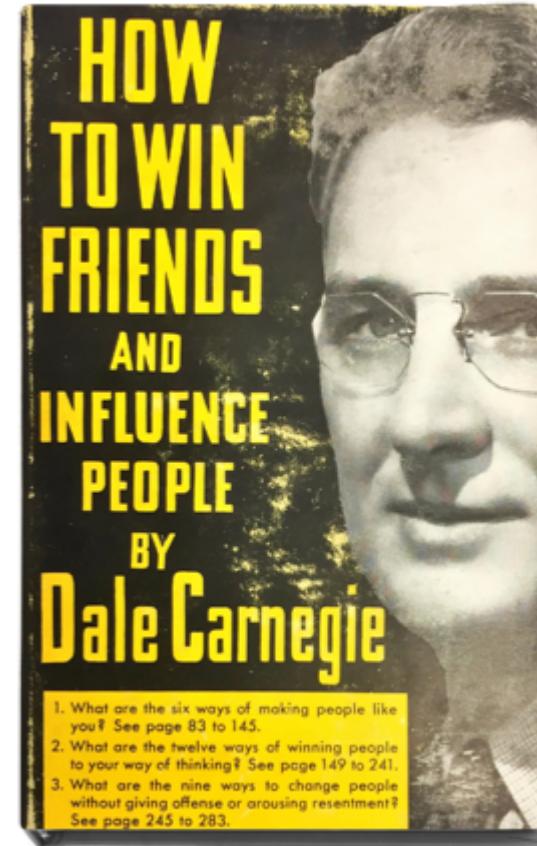
Le coefficient de régression linéaire peut aussi être affecté par les observations extrêmes.

Observation aberrante ou extrême:

- Le modèle prédit mal la variable dépendante pour une unité et son résidu est extrême.

Observation influente:

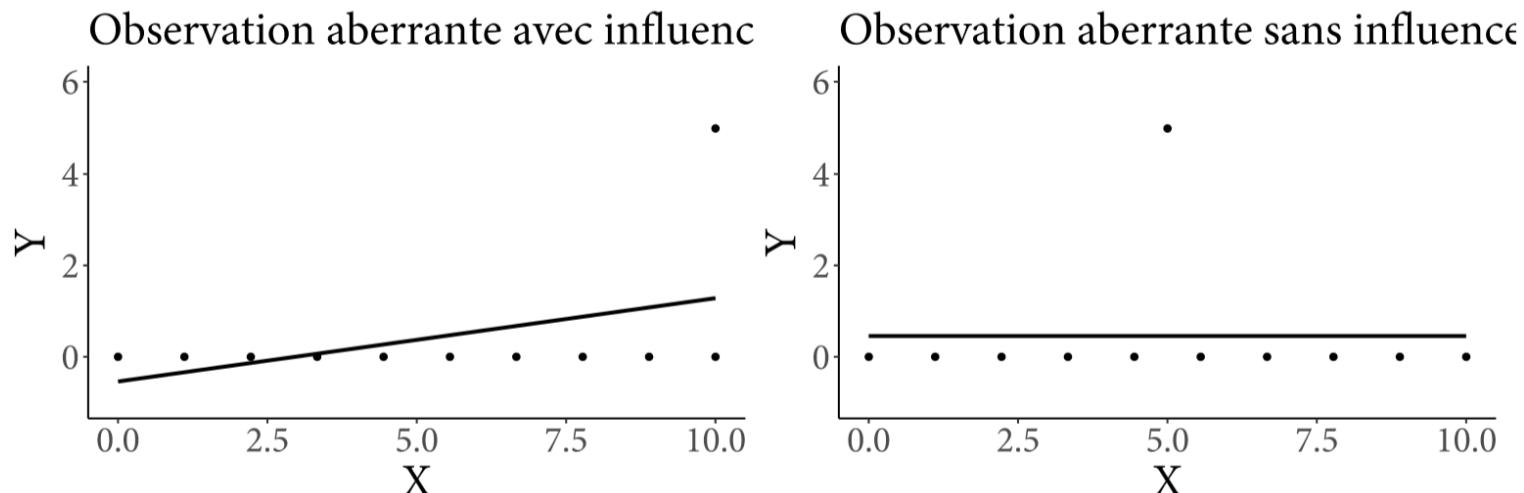
- Exclure cette observation de la banque de données a un effet important sur le coefficient de régression.



Données influentes

GRAPHIQUE 6.8. –

Certaines valeurs extrêmes peuvent influencer la pente de la droite de régression (c.-à-d. le coefficient de régression). La ligne noire représente la droite de prédiction du modèle de régression linéaire.

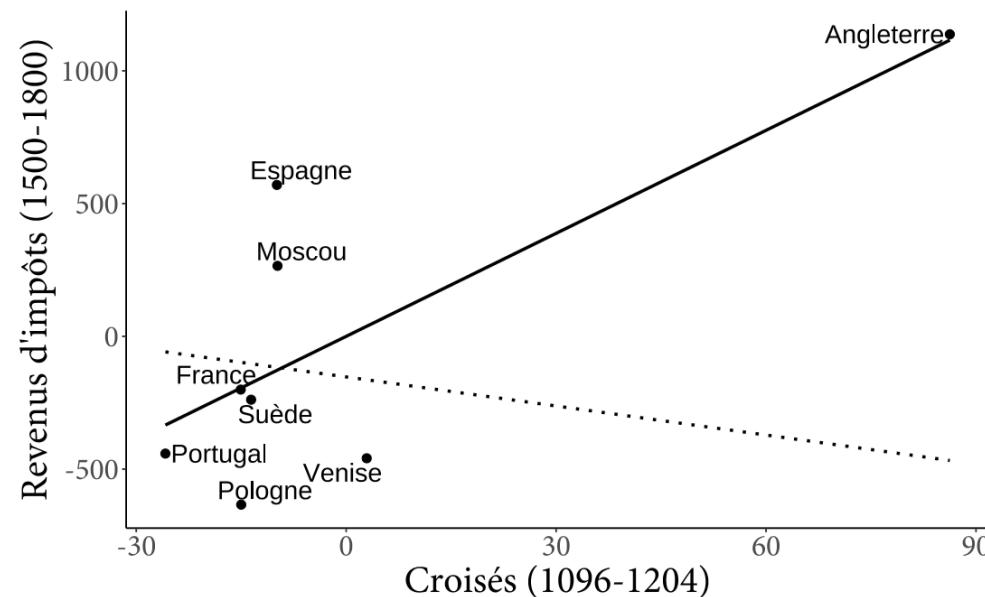


Données influentes

Diagnostique: Visualisation, DFBETA, DFFITS, D de Cook.

GRAPHIQUE 6.9. –

Relation entre le nombre de croisés issus d'une région et les revenus d'impôts du gouvernement de cette région. La ligne pleine reproduit la droite de régression de Blaydes et Paik (2016). La ligne pointillée représente un modèle de régression alternatif estimé en excluant l'Angleterre et le Portugal.



Effet marginal



Effet marginal

Mesure:

- Force de l'association entre une variable indépendante et une variable dépendante.
- Si X augmente, est-ce que Y augmente, diminue, ou reste constante?

Définition:

- Dérivée partielle de la variable dépendante par rapport à la variable explicative.
- $\frac{\partial Y}{\partial X}$

Interprétation:

- $\frac{\partial Y}{\partial X}$ mesure l'effet d'un petit changement dans X sur Y , lorsque toutes les autres variables demeurent constantes.

Effet marginal

Modèle linéaire avec deux variables explicatives :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

L'effet marginal de X_1 sur Y est égal à la dérivée partielle de Y par rapport à X_1 , soit:

$$\frac{\partial Y}{\partial X_1} = \beta_1$$

Dans un modèle linéaire simple, l'effet marginal de X_1 est exactement égal au coefficient de régression β_1 .

Effet marginal

Utile pour:

- Modèles avec variables transformées
- Modèles de modération, avec interactions multiplicatives
- Modèles linéaires généralisés
 - Logit
 - Probit



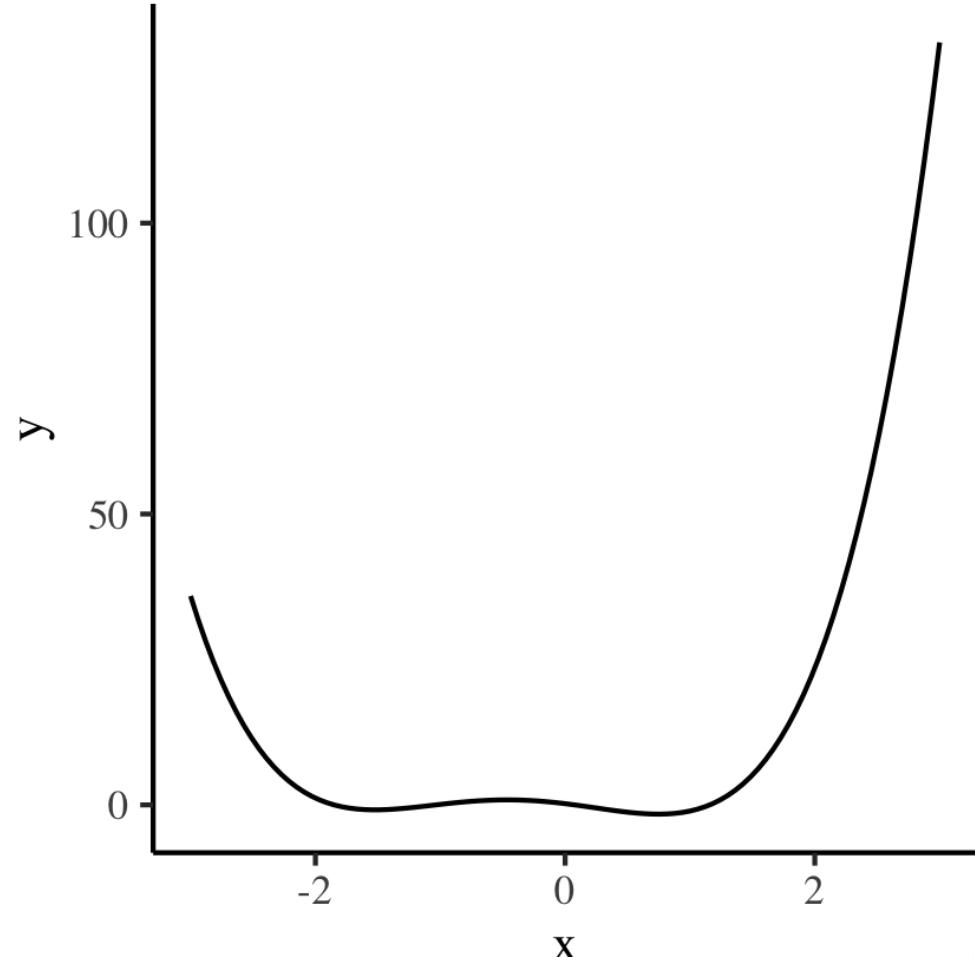
Transformations

Le modèle de régression de base représente la relation entre X et Y par une équation linéaire.

Même si l'équation est linéaire en ses paramètres, les variables du modèle peuvent être transformées par des fonctions arbitraires. Ceci rend le modèle linéaire flexible.

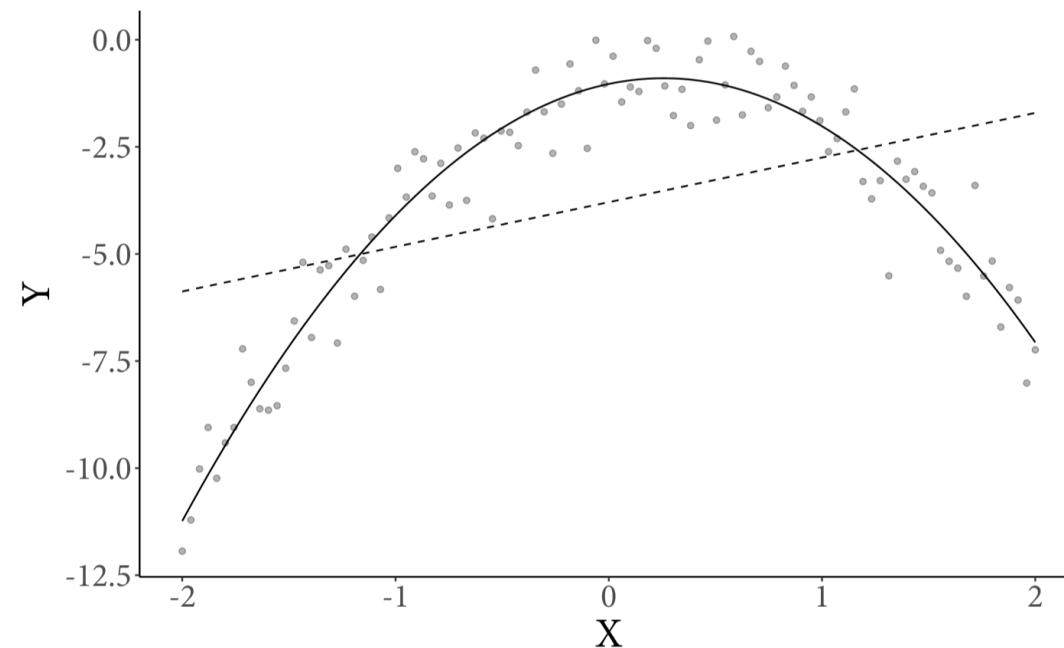
Nous allons considérer 2 transformations:

1. Quadratique
2. Logarithmique



Transformation quadratique

Le graphique montre la relation entre deux variables X et Y :



Cette relation est quadratique: lorsque X est basse, une augmentation de X est associée à une augmentation de Y . Lorsque X est élevée, une augmentation de X est associée à une diminution de Y .

Transformation quadratique

Un simple modèle linéaire bivarié ne tient pas compte de la courbure dans la relation.

Nous pouvons faire mieux en ajoutant une nouvelle variable à notre banque de données. Cette nouvelle variable, X^2 , est égale au carré de la variable explicative X :

Y	X	X^2
-3.078	1.434	2.057
-1.754	0.626	0.392
-1.147	1.152	1.326
-3.401	1.717	2.949
-5.371	-1.354	1.832

Avec ces variables, nous pouvons estimer une relation quadratique:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \varepsilon$$

Transformation quadratique

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \varepsilon$$

Sur la base des données dessinées dans ce graphique, le modèle estime les coefficients suivants:

$$\hat{\beta}_0 = -1, \hat{\beta}_1 = 1, \hat{\beta}_2 = -2$$

Insérer ces valeurs dans l'équation donne:

$$Y = -1 + 1 \cdot X - 2 \cdot X^2 + \varepsilon$$

Comment interpréter ces coefficients? Dans le modèle quadratique, l'effet marginal de X sur Y est égal à:

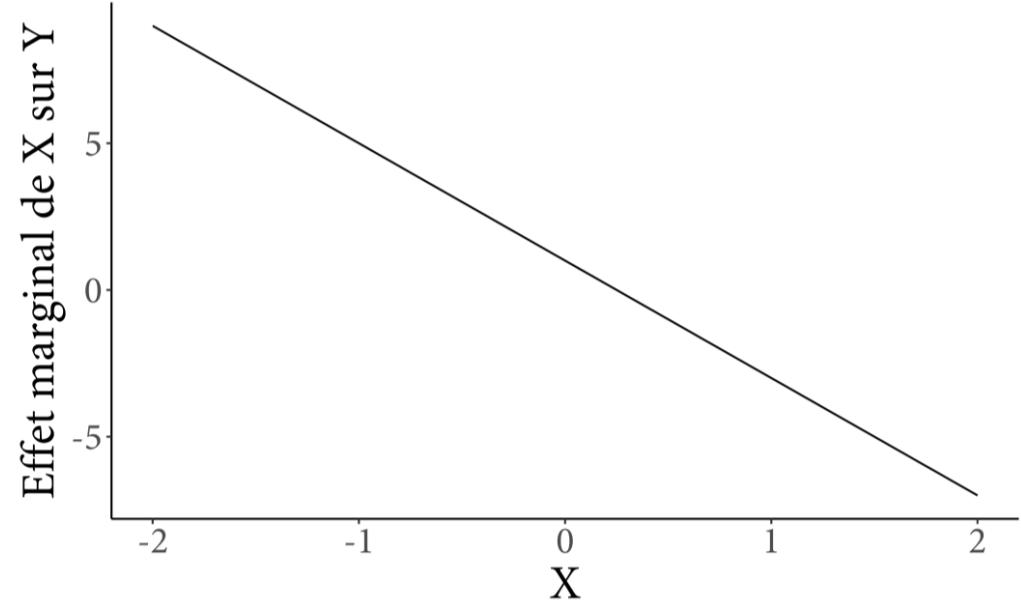
$$\begin{aligned}\frac{\partial Y}{\partial X} &= \beta_1 + \beta_2 \cdot 2 \cdot X \\ &= 1 - 2 \cdot 2 \cdot X \\ &= 1 - 4 \cdot X\end{aligned}$$

Transformation quadratique

$$\frac{\partial Y}{\partial X} = 1 - 4 \cdot X$$

Contrairement aux effets marginaux dans un modèle linéaire, l'effet marginal de X sur Y dans le modèle quadratique est conditionnel : l'effet d'un *changement* dans X sur Y dépend du *niveau* de X .

Plus le niveau de X est élevé, moins l'effet d'un changement dans X sur Y est élevé.

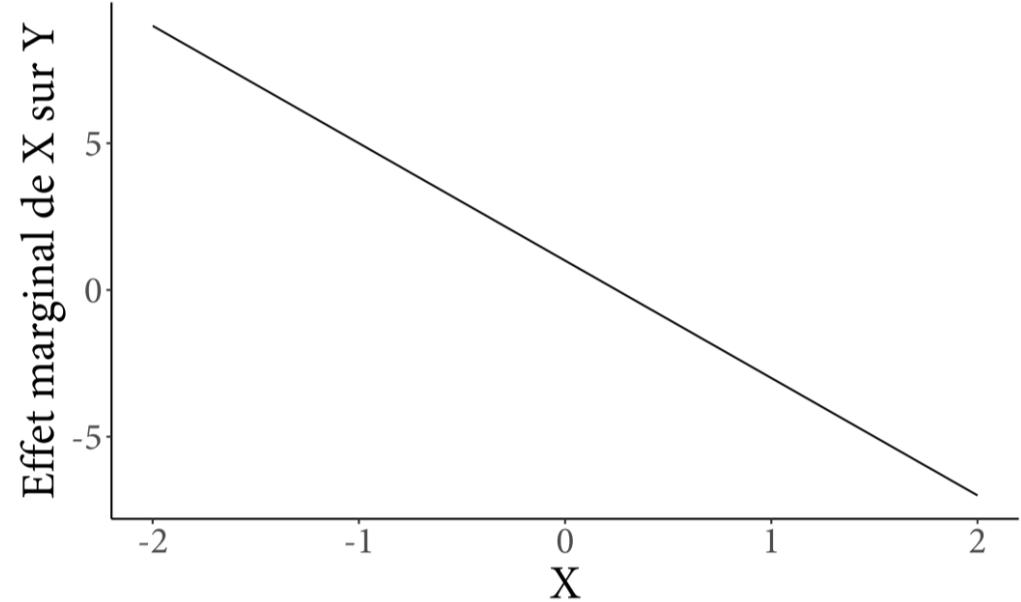


Transformation quadratique

Lorsque $X = -1$, augmenter X d'une unité est associé à une augmentation de 5 unités dans la variable Y .

Lorsque $X = 0.25$, augmenter X d'une unité n'est associé à aucun changement dans la variable Y .

Lorsque $X = 1$, augmenter X d'une unité est associé à une diminution de 3 unités dans la variable Y .



Transformation logarithmique

Une fonction pratique pour transformer nos variables est le logarithme naturel.

Grâce à cette fonction, nous pouvons estimer trois types de modèles:

$$Y = \beta_0 + \beta_1 \cdot \ln(X) + \varepsilon \quad \text{Lin-Log}$$

$$\ln(Y) = \lambda_0 + \lambda_1 \cdot X + \varepsilon \quad \text{Log-Lin}$$

$$\ln(Y) = \pi_0 + \pi_1 \cdot \ln(X) + \varepsilon \quad \text{Log-Log}$$

Dans le modèle Lin-Log, une augmentation de 1% de X est associée à un changement de $\beta_1/100$ unités de Y .

Dans le modèle Log-Lin, une augmentation de 1 unité de X est associée à un changement de $\lambda_1 \cdot 100\%$ de Y .

Dans le modèle Log-Log, une augmentation de 1% de X est associée à un changement de $\pi_1\%$ de Y .

Merci!

