Institut Universitaire MathFinÉco

Machine Learning et Science des Données pour l'Économétrie

Master 1/2 — Économie, Finance, Statistique, Actuariat

Devoir 1

Étudiants : Wilfried AVADEME

Enseignants : Firmin Ayivodji

Chargé de TP : Gédéon Gbedonou

Année académique : 2025–2026

Partie A Théorique

A1 Estimateur MCO — sans biais, variance et théorème BLUE

Question 1 *(Sans biais)*

Preuve du caractère sans biais de l'estimateur MCO

On considère le modèle linéaire :

$$Y = X\beta + \varepsilon$$

où Y est le vecteur des observations de la variable expliquée, X la matrice des variables explicatives, β le vecteur des paramètres inconnus et ε le vecteur des erreurs.

L'estimateur des moindres carrés ordinaires est donné par :

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

On veut montrer que :

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

Preuve

On part de l'expression de l'estimateur MCO :

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

D'après l'hypothèse H1 (la linéarité du modèle), on a :

$$Y = X\beta + \varepsilon$$

On peut donc remplace Y par $X\beta + \varepsilon$ dans l'expression de $\hat{\beta}$:

$$\hat{\beta} = (X^\top X)^{-1} X^\top (X\beta + \varepsilon)$$

En développant, on obtient :

$$\hat{\beta} = (X^\top X)^{-1} X^\top X\beta + (X^\top X)^{-1} X^\top \varepsilon$$

Pour définir l'estimateur MCO $\hat{\beta} = (X^\top X)^{-1} X^\top Y$, il faut que $X^\top X$ soit inversible.

On a donc :

$$(X^\top X)^{-1} X^\top X = I$$

Par conséquent :

$$\hat{\beta} = I\beta + (X^\top X)^{-1} X^\top \varepsilon$$

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon$$

On prend maintenant l'espérance conditionnelle sachant X :

$$E[\hat{\beta} \mid X] = E\left[\beta + (X^\top X)^{-1} X^\top \varepsilon \mid X\right]$$

Comme β est un paramètre fixe et que X est connu lorsqu'on conditionne par rapport à X, on peut écrire :

$$E[\hat{\beta} \mid X] = \beta + (X^\top X)^{-1} X^\top E[\varepsilon \mid X]$$

D'après l'hypothèse H2 (l'exogénéité) :

$$E[\varepsilon \mid X] = 0$$

Donc :

$$E[\hat{\beta} \mid X] = \beta + (X^\top X)^{-1} X^\top 0$$
$$E[\hat{\beta} \mid X] = \beta$$

Passons à l'espérance non conditionnelle. En appliquant la loi de l'espérance totale, on obtient :

$$E[\hat{\beta}] = E\left[E[\hat{\beta} \mid X]\right]$$

Or :

$$E[\hat{\beta} \mid X] = \beta$$

Donc :

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

Comme β est un paramètre fixe :

$$E[\beta] = \beta$$

Finalement :

$$\boxed{E[\hat{\beta}] = \beta}$$

Donc l'estimateur des moindres carrés ordinaires est sans biais.

Question 2 *(Variance de $\hat{\beta}$)*

Variance conditionnelle de l'estimateur MCO

On considère le modèle linéaire :

$$Y = X\beta + \varepsilon$$

L'estimateur MCO est donné par :

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

On a déjà montré que :

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon$$

On veut montrer que :

$$\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}$$

Preuve

On part de :

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon$$

On prend la variance conditionnelle sachant X :

$$\mathrm{Var}(\hat{\beta} \mid X) = \mathrm{Var}\left(\beta + (X^\top X)^{-1} X^\top \varepsilon \mid X\right)$$

Comme β est un paramètre fixe, sa variance est nulle. Donc :

$$\mathrm{Var}(\hat{\beta} \mid X) = \mathrm{Var}\left((X^\top X)^{-1} X^\top \varepsilon \mid X\right)$$

Conditionnellement à X, la matrice $(X^\top X)^{-1} X^\top$ est constante.

Posons : $A = (X^\top X)^{-1} X^\top$

On a :

$$\begin{aligned}
\mathrm{Var}(\hat{\beta} \mid X) &= \mathrm{Var}(A\varepsilon \mid X) \\
&= E[A\varepsilon\varepsilon^\top A^\top \mid X] - E[A\varepsilon \mid X] E[A\varepsilon \mid X]^\top \\
&= A \cdot E[\varepsilon\varepsilon^\top \mid X] \cdot A^\top - A \cdot E[\varepsilon \mid X] \cdot E[\varepsilon \mid X]^\top \cdot A^\top \\
&= A(E[\varepsilon\varepsilon^\top \mid X] - E[\varepsilon \mid X] E[\varepsilon \mid X]^\top) A^\top \\
\mathrm{Var}(\hat{\beta} \mid X) &= A \cdot Var(\varepsilon \mid X) \cdot A^\top
\end{aligned}$$

Ainsi :

$$\mathrm{Var}(\hat{\beta} \mid X) = (X^\top X)^{-1} X^\top \, \mathrm{Var}(\varepsilon \mid X) \left[(X^\top X)^{-1} X^\top\right]^\top$$

Calculons $\left[(X^\top X)^{-1} X^\top\right]^\top$

$$\begin{aligned}
\left[(X^\top X)^{-1} X^\top\right]^\top &= (X^\top)^\top \cdot [(X^\top X)^{-1}]^\top \\
&= X \cdot [(X^\top X)^{-1}]^\top \\
&= X \cdot [(X^\top X)^\top]^{-1} \\
\left[(X^\top X)^{-1} X^\top\right]^\top &= X \cdot (X^\top X)^{-1} \qquad \text{car } X^\top X \text{ est symétrique}
\end{aligned}$$

Donc :

$$\mathrm{Var}(\hat{\beta} \mid X) = (X^\top X)^{-1} X^\top \mathrm{Var}(\varepsilon \mid X) X (X^\top X)^{-1}$$

Rappelons que $\mathrm{Var}(\varepsilon \mid X)$ est une matrice $n \times n$:

$$\mathrm{Var}(\varepsilon \mid X) = \begin{pmatrix} \mathrm{Var}(\varepsilon_1 \mid X) & \mathrm{Cov}(\varepsilon_1, \varepsilon_2 \mid X) & \cdots & \mathrm{Cov}(\varepsilon_1, \varepsilon_n \mid X) \\ \mathrm{Cov}(\varepsilon_2, \varepsilon_1 \mid X) & \mathrm{Var}(\varepsilon_2 \mid X) & \cdots & \mathrm{Cov}(\varepsilon_2, \varepsilon_n \mid X) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\varepsilon_n, \varepsilon_1 \mid X) & \mathrm{Cov}(\varepsilon_n, \varepsilon_2 \mid X) & \cdots & \mathrm{Var}(\varepsilon_n \mid X) \end{pmatrix}.$$

L'hypothèse H3 d'homoscédasticité fixe les éléments de la diagonale :

$$\mathrm{Var}(\varepsilon_i \mid X) = \sigma^2, \qquad \text{pour tout } i = 1, \ldots, n.$$

L'hypothèse H4 d'absence d'autocorrélation fixe les éléments hors diagonale :

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X) = 0, \qquad \text{pour tout } i \neq j.$$

Ainsi, sous H3 et H4, on obtient :

$$\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n.$$

Donc :

$$\begin{aligned} \mathrm{Var}(\hat{\beta} \mid X) &= (X^\top X)^{-1} X^\top \sigma^2 I_n X (X^\top X)^{-1} \\ &= \sigma^2 [(X^\top X)^{-1} X^\top X] (X^\top X)^{-1} \\ &= \sigma^2 [I_n] (X^\top X)^{-1} \\ \mathrm{Var}(\hat{\beta} \mid X) &= \sigma^2 (X^\top X)^{-1} \end{aligned}$$

Finalement :

$$\boxed{\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}}$$

Les hypothèses H3 et H4 sont nécessaires pour obtenir la formule classique de la variance :

$$\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}$$

Question 3 *(Théorème BLUE et conséquences)*

Théorème de Gauss-Markov et propriété BLUE des MCO

Énoncé du théorème de Gauss-Markov

On considère le modèle linéaire :

$$Y = X\beta + \varepsilon$$

On considère les hypothèses suivantes :

H1 : $Y = X\beta + \varepsilon$

H2 : $E[\varepsilon \mid X] = 0$

H3 : $\mathrm{Var}(\varepsilon_i \mid X) = \sigma^2, \quad \forall i$

H4 : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X) = 0, \quad \forall i \neq j$

Sous les hypothèses H1, H2, H3 et H4 (linéarité, exogénéité, homoscédasticité , absence d'autocorrélation), l'estimateur des MCO $\hat{\beta}$ est l'estimateur le plus précis (BLUE : Best Linear Unbiased Estimator) dans l'ensemble des estimateurs linéaires sans biais de β.

Autrement dit, $\hat{\beta}$ est BLUE :

$$\boxed{\text{BLUE} = \text{Best Linear Unbiased Estimator}}$$

c'est-à-dire :

- **Best** : il a la plus petite variance parmi les estimateurs linéaires sans biais.
- **Linear** : $\hat{\beta}$ est une combinaison linéaire de Y : $\hat{\beta} = (X^\top X)^{-1} X^\top Y$.
- **Unbiased** : l'estimateur est sans biais :$\mathbb{E}[\hat{\beta}] = \beta$.
- **Estimator** : C'est un estimateur du vecteur des paramètres β.

Les MCO sont BLUE parce que, sous les hypothèses de Gauss-Markov, ils sont linéaires en Y ; sans biais et de variance minimale parmi tous les estimateurs linéaires sans biais.

Cas où l'hypothèse H3 est violée

On suppose maintenant que l'hypothèse d'homoscédasticité H3 est violée.

Cela signifie que :

$$\mathrm{Var}(\varepsilon_i \mid X) \neq \sigma^2$$

ou encore que les erreurs n'ont pas toutes la même variance. On parle alors d'hétéroscédasticité.

(a) L'estimateur $\hat{\beta}$ reste-t-il sans biais ?

Oui, l'estimateur MCO reste sans biais si H3 est violée, à condition que l'hypothèse d'exogénéité H2 reste vraie :

$$E[\varepsilon \mid X] = 0$$

En effet, le caractère sans biais de $\hat{\beta}$ dépend principalement de H1 et H2, mais pas de H3.

On a :

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon$$

Ce qui nous a permis de prouver a la question 1 que :

$$\boxed{E[\hat{\beta}] = \beta}$$

Donc, même si H3 est violée :

$$\boxed{\hat{\beta} \text{ reste sans biais si } E[\varepsilon \mid X] = 0}$$

(b) L'estimateur $\hat{\beta}$ reste-t-il BLUE ?

Non, en général, $\hat{\beta}$ ne reste pas BLUE si H3 est violée.

La propriété BLUE signifie que l'estimateur est le meilleur estimateur linéaire sans biais, c'est-à-dire celui qui a la plus petite variance.

Or, l'homoscédasticité H3 est nécessaire pour garantir que les MCO ont la variance minimale parmi les estimateurs linéaires sans biais.

Si H3 est violée, les erreurs sont hétéroscédastiques. Dans ce cas, les MCO peuvent rester linéaires et sans biais, mais ils ne sont plus forcément les plus efficaces.

Donc :

$$\boxed{\hat{\beta} \text{ ne reste généralement pas BLUE si H3 est violée}}$$

(c) Conséquence pratique sur les intervalles de confiance et les tests t

Si H3 est violée, la formule classique de la variance des MCO devient incorrecte.

Normalement, sous H3 et H4, on utilise :

$$\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}$$

Mais en présence d'hétéroscédasticité, on n'a plus :

$$\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n$$

Donc la variance estimée de $\hat{\beta}$ est fausse.

Par conséquent :

- les erreurs standards peuvent être incorrectes ;
- les intervalles de confiance peuvent être trop larges ou trop étroits ;
- les statistiques de test t peuvent être incorrectes ;
- les conclusions des tests peuvent être fausses.

On risque donc de rejeter à tort ou de ne pas rejeter à tort une hypothèse.

Par exemple, on peut conclure qu'une variable est significative alors qu'elle ne l'est pas réellement, ou inversement.

La solution classique est d'utiliser des erreurs standards robustes à l'hétéroscédasticité.

En cas d'hétéroscédasticité, on utilise des erreurs standards robustes.

A2 Variable omise — preuve formelle et biais

Question 4 *(Dérivation du biais)*

Dérivation du biais

Le vrai modèle est :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Mais on estime le modèle réduit :

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + u_i$$

On veut montrer que :

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \delta_{12}$$

où δ_{12} est le coefficient de la régression auxiliaire de X_2 sur X_1.

1. Écriture matricielle du modèle réduit

Dans le modèle réduit, on régressse Y seulement sur une constante et X_1.

On note :

$$\tilde{X} = \begin{pmatrix} 1 & X_{11} \\ 1 & X_{12} \\ \vdots & \vdots \\ 1 & X_{1n} \end{pmatrix}$$

L'estimateur MCO du modèle réduit est :

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

avec :

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix}$$

2. Remplacer Y par le vrai modèle

Le vrai modèle s'écrit sous forme matricielle :

$$Y = \beta_0 \mathbf{1} + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Comme $\tilde{X}$ contient la constante et X_1, on peut écrire :

$$Y = \tilde{X} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \beta_2 X_2 + \varepsilon$$

Donc :

$$\tilde{\beta} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \left[\tilde{X} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \beta_2 X_2 + \varepsilon \right]$$

3. Développement

On développe :

$$\tilde{\beta} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{X} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \beta_2 (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top X_2 + (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon$$

Or :

$$(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{X} = I$$

Donc :

$$\tilde{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \beta_2 (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top X_2 + (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon$$

Interprétation du terme $(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top X_2$

Le terme :

$$(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top X_2$$

est exactement l'estimateur MCO de la régression auxiliaire de X_2 sur une constante et X_1 $(1, X_1)$).

Cette régression auxiliaire est :

$$X_{2i} = \delta_0 + \delta_{12} X_{1i} + v_i$$

Sous forme matricielle :

$$X_2 = \tilde{X}\delta + v$$

avec :

$$\delta = \begin{pmatrix} \delta_0 \\ \delta_{12} \end{pmatrix}$$

L'estimateur MCO de cette régression auxiliaire est :

$$\hat{\delta} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top X_2$$

Donc :

$$(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top X_2 = \begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_{12} \end{pmatrix}$$

Ainsi :

$$\tilde{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \beta_2 \begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_{12} \end{pmatrix} + (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon$$

4. Extraire la deuxième composante

La deuxième composante de $\tilde{\beta}$ correspond à $\tilde{\beta}_1$.

Donc :

$$\tilde{\beta}_1 = \beta_1 + \beta_2 \hat{\delta}_{12} + \left[(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon\right]_2$$

où :

$$\left[(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon\right]_2$$

désigne la deuxième composante du vecteur.

5. Prendre l'espérance

On prend maintenant l'espérance conditionnellement à X_1 et X_2.

On suppose que : $E[\varepsilon \mid X_1, X_2] = 0$ d'après l'hypothèse H2 (exogénéité).

Donc

$$\begin{aligned} E\left[\tilde{\beta}_1 \mid X_1, X_2\right] &= E\left[\beta_1 + \beta_2 \hat{\delta}_{12} + \left[(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon\right]_2 \mid X_1, X_2\right] \\ &= \beta_1 + \beta_2 \hat{\delta}_{12} + E\left[\left[(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon\right]_2 \mid X_1, X_2\right] \\ E\left[\tilde{\beta}_1 \mid X_1, X_2\right] &= \beta_1 + \beta_2 \hat{\delta}_{12} + (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top E\left[\varepsilon \mid X_1, X_2\right] \end{aligned}$$

Or :

$$E[\varepsilon \mid X_1, X_2] = 0$$

Donc :

$$E\left[(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon \mid X_1, X_2\right] = 0$$

Ainsi :

$$E[\tilde{\beta}_1 \mid X_1, X_2] = \beta_1 + \beta_2 \hat{\delta}_{12}$$

En appliquant la loi de l'espérance totale, on obtient

$$\begin{aligned} E[\tilde{\beta}_1] &= E[E[\tilde{\beta}_1 \mid X_1, X_2]] \\ E[\tilde{\beta}_1] &= E[\beta_1 + \beta_2 \hat{\delta}_{12}] \end{aligned}$$

On note δ_{12} le coefficient de la régression auxiliaire de X_2 sur X_1.

$\hat{\delta}_{12}$ est l'estimateur MCO de δ_{12} conditionnel à X_1, X_2, donc $E[\hat{\delta}_{12} \mid X_1, X_2] = \delta_{12}$ par sans-biais des MCO.

Ainsi $E[\beta_2\hat{\delta}_{12}] = \beta_2\delta_{12}$.

Donc :

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2\delta_{12}$$

Finalement :

$$\boxed{E[\tilde{\beta}_1] = \beta_1 + \beta_2\delta_{12}}$$

Question 5 *(Conditions d'absence de biais)*

Conditions d'annulation du biais de variable omise

On a montré que :

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2\delta_{12}$$

Donc le biais de l'estimateur $\tilde{\beta}_1$ est :

$$\text{Biais}(\tilde{\beta}_1) = E[\tilde{\beta}_1] - \beta_1$$

Ainsi :

$$\text{Biais}(\tilde{\beta}_1) = \beta_2\delta_{12}$$

Le biais est nul si et seulement si : $\beta_2\delta_{12} = 0$

Comme il s'agit d'un produit, on a : $\beta_2\delta_{12} = 0 \quad \Longleftrightarrow \quad \beta_2 = 0 \quad \text{ou} \quad \delta_{12} = 0$

Condition 1 : $\beta_2 = 0$

Cela signifie que, dans le vrai modèle : $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ la variable X_2 n'a aucun effet sur Y.

En effet, si : $\beta_2 = 0$ alors le vrai modèle devient : $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

Dans ce cas, le modèle réduit : $Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + u_i$ n'oublie pas une variable importante, car X_2 n'explique pas Y.

Ainsi pour $\beta_2 = 0$ on a :

$$\begin{aligned} E[\tilde{\beta}_1] &= \beta_1 + \beta_2\delta_{12} \\ &= \beta_1 + 0 \times \delta_{12} \\ E[\tilde{\beta}_1] &= \beta_1 \end{aligned}$$

Donc $\tilde{\beta}_1$ est sans biais.

Condition 2 :$\delta_{12} = 0$

Rappelons que δ_{12} est le coefficient de la régression auxiliaire de X_2 sur X_1 : $X_{2i} = \delta_0 + \delta_{12} X_{1i} + v_i$

C'est équivalent à dire que $\text{Cov}(X_1, X_2) = 0$ autrement dit X_1 et X_2 sont non corrélées

Si : $\delta_{12} = 0$ alors : $X_{2i} = \delta_0 + v_i$

Cela signifie que X_1 n'explique pas X_2. Autrement dit, la variable omise X_2 n'est pas corrélée avec la variable incluse X_1.

Ainsi pour $\delta_{12} = 0$ on a :

$$\begin{aligned} E[\tilde{\beta}_1] &= \beta_1 + \beta_2\delta_{12} \\ &= \beta_1 + \beta_2 \times 0 \\ E[\tilde{\beta}_1] &= \beta_1 \end{aligned}$$

Donc $\tilde{\beta}_1$ est sans biais.

Interprétation

Le biais de variable omise existe seulement si deux conditions sont réunies en même temps :

$$\boxed{\beta_2 \neq 0} \qquad \text{et} \qquad \boxed{\delta_{12} \neq 0}$$

Autrement dit :

— la variable omise X_2 doit avoir un effet réel sur Y ;
— la variable omise X_2 doit être liée à la variable incluse X_1.

Si l'une de ces deux conditions disparaît, le biais est nul.

Question 6 *(Application numérique)*

Application :

On a

$$\hat{\beta}_{\text{age}}^{\text{réduit}} = 273 \qquad \text{(modèle sans smoker)}$$

$$\hat{\beta}_{\text{age}}^{\text{complet}} = 267 \qquad \text{(modèle avec smoker)}$$

$$\hat{\beta}_{\text{smoker}} = 23855$$

$$\hat{\delta}_{\text{age}} = 0{,}00025$$

(a) Calcul du biais théorique

Le biais théorique de variable omise est donné par :

$$\begin{aligned}\text{Biais théorique} &= \hat{\beta}_{\text{smoker}} \times \hat{\delta}_{\text{age}} \\ &= 23855 \times 0{,}00025 \\ \text{Biais théorique} &= 5{,}96375\end{aligned}$$

Ainsi :

$$\boxed{\text{Biais théorique} \approx 5{,}96}$$

(b) Comparaison avec le biais empirique

Le biais empirique correspond à la différence entre le coefficient du modèle réduit et celui du modèle complet :

$$\begin{aligned}\text{Biais empirique} &= \hat{\beta}_{\text{age}}^{\text{réduit}} - \hat{\beta}_{\text{age}}^{\text{complet}} \\ &= 273 - 267 \\ \text{Biais empirique} &= 6\end{aligned}$$

On compare alors : Biais théorique $\approx 5{,}96 <$ Biais empirique $= 6$

Mais, ; les deux valeurs sont très proches : $5{,}96 \approx 6$

Donc :

$$\boxed{\text{Les deux valeurs concordent.}}$$

Le biais théorique (5,96) et le biais empirique (6) sont très proches. La légère différence s'explique par les arrondis appliqués aux coefficients $\hat{\beta}_{smoker}$ et $\hat{\delta}_{age}$. Les deux valeurs concordent, ce qui valide empiriquement la formule théorique $E[\tilde{\beta}_1] = \beta_1 + \beta_2\delta_{12}$ démontrée en Question 4.

(c) Sous-estimation ou surestimation de l'effet de l'âge

on a : $\hat{\beta}_{\text{age}}^{\text{réduit}} = 273 > \hat{\beta}_{\text{age}}^{\text{complet}} = 267$

Le modèle réduit attribue à l'âge un effet plus important que le modèle complet.

Ainsi, l'actuaire : **surestime l'effet de l'âge**

Le biais est positif : $273 - 267 = 6 > 0$

Cela signifie qu'en omettant la variable smoker, une partie de l'effet du tabagisme est attribuée à tort à la variable âge.

Conséquence pour la tarification

Si l'actuaire utilise le modèle réduit sans la variable smoker, la tarification devient moins précise.

L'omission de la variable smoker entraîne deux distorsions tarifaires :

Les assurés âgés sont surtarifés : on leur attribue un effet âge de 273$ par an au lieu de 267$

Les assurés jeunes fumeurs sont sous-tarifés : leur risque tabagique ($\approx 23855\$$ de charges supplémentaires) n'est pas capturé

Ces distorsions violent le principe actuariel d'équité tarifaire : chaque assuré doit payer une prime proportionnelle à son risque réel.

En pratique, cela peut entraîner une anti-sélection : les bons risques quittent le portefeuille car surtarifés, et les mauvais risques restent car sous-tarifés.

A3 Données manquantes — mécanismes et méthodes d'imputation

Question 7 *(Mécanismes MCAR, MAR et MNAR)*

Mécanismes de données manquantes : MCAR, MAR et MNAR

On considère une variable d'intérêt Y, une matrice de variables explicatives X, et un indicateur de réponse R.

Pour une observation i, on définit :

$$R_i = \begin{cases} 1 & \text{si la donnée est observée,} \\ 0 & \text{si la donnée est manquante.} \end{cases}$$

On distingue trois grands mécanismes de données manquantes :

$$\text{MCAR}, \quad \text{MAR}, \quad \text{MNAR}.$$

1. MCAR : Missing Completely At Random

(a) Définition formelle

Les données sont dites MCAR si la probabilité qu'une donnée soit manquante ne dépend ni des variables observées, ni des variables non observées.

Formellement :

$$P(R = 1 \mid Y_{\text{obs}}, Y_{\text{mis}}, X) = P(R = 1)$$

Cela signifie que le fait qu'une donnée soit manquante est totalement aléatoire.

Autrement dit, les individus avec données manquantes ne sont pas systématiquement différents des individus avec données complètes.

(b) Exemple concret

En assurance santé, supposons qu'un fichier contenant les charges médicales de certains assurés soit partiellement perdu à cause d'un problème informatique indépendant des caractéristiques des assurés.

Par exemple, une panne technique supprime aléatoirement 5% des valeurs de charges médicales, sans lien avec l'âge, le sexe, le statut fumeur ou le montant réel des charges.

Dans ce cas, les données manquantes sont MCAR.

(c) Effet de la suppression des lignes manquantes

Sous MCAR, la suppression des lignes manquantes, appelée *listwise deletion*, ne produit généralement pas un estimateur MCO biaisé.

En effet, l'échantillon restant est encore représentatif de la population initiale, puisque les observations supprimées l'ont été de manière totalement aléatoire.

Ainsi, si le modèle MCO était correctement spécifié avant suppression, on conserve :

$$E[\varepsilon \mid X, R = 1] = 0$$

Donc :

Sous MCAR, la suppression des lignes ne biaise pas l'estimateur MCO.

Cependant, elle réduit la taille de l'échantillon, donc elle augmente la variance des estimateurs.

Pas de biais, mais perte de précision.

2. MAR : Missing At Random

(a) Définition formelle

Les données sont dites MAR si la probabilité qu'une donnée soit manquante peut dépendre des variables observées, mais ne dépend pas de la valeur manquante elle-même une fois les variables observées contrôlées.

Formellement :

$$P(R = 1 \mid Y_{\text{obs}}, Y_{\text{mis}}, X) = P(R = 1 \mid Y_{\text{obs}}, X)$$

Cela signifie que le mécanisme de non-réponse est explicable par les variables observées.

(b) Exemple concret

En scoring bancaire, supposons que le revenu soit manquant plus souvent chez les jeunes clients ou chez les clients ayant un contrat de travail instable.

Si l'âge et le type de contrat sont observés dans la base, alors la probabilité que le revenu soit manquant dépend de variables observées.

Par exemple :

$$P(R_{\text{revenu}} = 1 \mid \text{âge}, \text{type de contrat})$$

Dans ce cas, le mécanisme peut être considéré comme MAR si, à âge et type de contrat donnés, l'absence du revenu ne dépend plus du revenu réel lui-même.

(c) Effet de la suppression des lignes manquantes

Sous MAR, la suppression des lignes manquantes peut produire un estimateur MCO biaisé si les variables qui expliquent la non-réponse sont liées à la variable dépendante ou aux variables explicatives du modèle.

En effet, l'échantillon complet n'est plus forcément représentatif de la population globale.

Par exemple, si les jeunes clients sont plus souvent supprimés parce que leur revenu est manquant, alors l'échantillon final contiendra proportionnellement moins de jeunes clients.

Le modèle estimé sur les seules lignes complètes peut donc être biaisé.

Formellement, après suppression des observations incomplètes, on estime le modèle sur les observations telles que :

$$R = 1$$

Pour que l'estimateur MCO reste sans biais, il faut :

$$E[\varepsilon \mid X, R = 1] = 0$$

Or, sous MAR, cette condition n'est pas automatiquement garantie.

Donc :

Sous MAR, la suppression des lignes peut biaiser l'estimateur MCO.

Le biais apparaît seulement si les variables qui expliquent le manque sont liées à Y et ne sont pas incluses dans le modèle. Si elles sont incluses dans X, le biais peut disparaître.

La bonne pratique consiste plutôt à utiliser des méthodes adaptées, comme : imputation multiple, pondération par probabilité de réponse, ou maximum de vraisemblance.

3. MNAR : Missing Not At Random

(a) Définition formelle

Les données sont dites MNAR si la probabilité qu'une donnée soit manquante dépend directement de la valeur manquante elle-même, même après avoir contrôlé les variables observées.

Formellement :

$$P(R = 1 \mid Y_{\text{obs}}, Y_{\text{mis}}, X) \neq P(R = 1 \mid Y_{\text{obs}}, X)$$

Cela signifie que le mécanisme de données manquantes dépend d'une information non observée.

(b) Exemple concret

En assurance santé, supposons que les personnes ayant des charges médicales très élevées refusent plus souvent de déclarer certaines informations de santé.

Par exemple, les assurés ayant une maladie grave non déclarée ont plus souvent des données médicales manquantes.

Dans ce cas, la probabilité que la donnée soit manquante dépend directement de la gravité réelle de l'état de santé, qui est précisément non observée.

On est donc dans un cas MNAR.

Autre exemple en scoring bancaire : les clients ayant un très fort endettement peuvent refuser plus souvent de déclarer leur dette réelle.

Ici, la probabilité que la dette soit manquante dépend du niveau réel de dette non observé.

(c) Effet de la suppression des lignes manquantes

Sous MNAR, la suppression des lignes manquantes produit généralement un estimateur MCO biaisé.

En effet, les observations supprimées sont systématiquement différentes des observations conservées selon une variable non observée.

L'échantillon final est donc sélectionné de manière non aléatoire.

Formellement, sous MNAR, on a généralement :

$$E[\varepsilon \mid X, R = 1] \neq 0$$

Donc l'hypothèse d'exogénéité nécessaire aux MCO est violée dans l'échantillon observé.

Ainsi :

Sous MNAR, la suppression des lignes produit généralement un biais.

Dans ce cas, les méthodes simples d'imputation peuvent également être insuffisantes, car le mécanisme de non-réponse dépend de valeurs non observées.

Il faut alors utiliser des modèles spécifiques de sélection ou des analyses de sensibilité.

Résumé comparatif

Mécanisme	**Dépend de quoi ?**	**Exemple**	**Listwise deletion**
MCAR	Ne dépend ni des données observées ni des données manquantes.	Panne informatique aléatoire.	Non biaisée, mais moins précise.
MAR	Dépend des données observées.	Revenu manquant selon l'âge ou le contrat.	Peut être biaisée.
MNAR	Dépend des données manquantes elles-mêmes.	Dette ou maladie grave non déclarée.	Généralement biaisée.

Question 8 *(Méthodes d'imputation)*

Méthodes d'imputation des données manquantes

L'imputation consiste à remplacer les valeurs manquantes par des valeurs estimées, afin de conserver les observations dans l'analyse statistique. Il existe plusieurs méthodes d'imputation, plus ou moins simples selon la nature des données et le mécanisme de manque.

1. Imputation par la moyenne

Principe

L'imputation par la moyenne consiste à remplacer chaque valeur manquante d'une variable quantitative par la moyenne des valeurs observées de cette variable.

Si X est une variable avec des valeurs manquantes, alors on remplace chaque valeur manquante par :

$$\bar{X}_{obs} = \frac{1}{n_{obs}} \sum_{i:R_i=1} X_i$$

où $R_i = 1$ signifie que la valeur X_i est observée.

Hypothèse requise

Cette méthode est surtout acceptable lorsque les données sont MCAR, c'est-à-dire manquantes complètement au hasard.

$$P(R = 1 \mid X_{obs}, X_{mis}) = P(R = 1)$$

Avantage principal

Son principal avantage est sa simplicité. Elle est facile à comprendre et à mettre en œuvre.

Limite principale

Elle réduit artificiellement la variance de la variable, car plusieurs valeurs manquantes sont remplacées par une même valeur moyenne.

Elle peut aussi affaiblir les relations entre variables, notamment les corrélations.

L'imputation par la moyenne est simple, mais elle déforme la variabilité des données.

2. Imputation par la médiane

Principe

L'imputation par la médiane consiste à remplacer chaque valeur manquante d'une variable quantitative par la médiane des valeurs observées.

La médiane est la valeur qui partage la distribution en deux parties égales.

$$X_i^{imp} = \text{Med}(X_{obs})$$

Hypothèse requise

Cette méthode est principalement adaptée lorsque les données sont MCAR, ou lorsque le taux de données manquantes est faible.

Avantage principal

Elle est plus robuste que la moyenne en présence de valeurs extrêmes.

Par exemple, en assurance santé, les charges médicales peuvent être très asymétriques avec quelques assurés ayant des coûts très élevés. Dans ce cas, la médiane est souvent plus stable que la moyenne.

Limite principale

Comme l'imputation par la moyenne, elle remplace toutes les valeurs manquantes par une seule valeur. Elle réduit donc la dispersion réelle des données et peut biaiser les relations entre variables.

L'imputation par la médiane est robuste, mais elle reste une méthode simpliste.

3. Imputation par la régression

Principe

L'imputation par la régression consiste à prédire la valeur manquante d'une variable à partir des autres variables observées.

Par exemple, si le revenu est manquant, on peut l'estimer à partir de l'âge, du statut professionnel, du niveau d'éducation et du score bancaire.

On estime un modèle du type :

$$X_j = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k + \eta$$

Puis on utilise ce modèle pour prédire les valeurs manquantes de X_j.

Hypothèse requise

Cette méthode suppose généralement que les données sont MAR, c'est-à-dire que le mécanisme de manque dépend des variables observées, mais pas directement de la valeur manquante elle-même.

$$P(R = 1 \mid X_{obs}, X_{mis}) = P(R = 1 \mid X_{obs})$$

Avantage principal

Elle utilise l'information contenue dans les autres variables. Elle est donc souvent plus pertinente que l'imputation par moyenne ou médiane.

Limite principale

Elle peut sous-estimer l'incertitude, car les valeurs imputées sont souvent trop proches de la droite de régression. Elle peut donc donner une impression excessive de précision.

L'imputation par régression est plus informative, mais elle peut sous-estimer la variabilité.

4. Imputation multiple par MICE

Principe

MICE signifie *Multiple Imputation by Chained Equations*. Cette méthode consiste à créer plusieurs bases de données complètes en imputant les valeurs manquantes plusieurs fois.

Chaque variable ayant des données manquantes est imputée à partir d'un modèle conditionnel basé sur les autres variables.

On obtient donc plusieurs jeux de données imputés :

$$D_1, D_2, \ldots, D_m$$

On estime ensuite le modèle statistique sur chaque base, puis on combine les résultats selon les règles de Rubin.

Hypothèse requise

MICE repose généralement sur l'hypothèse MAR :

$$R \perp X_{mis} \mid X_{obs}$$

Cela signifie que, conditionnellement aux variables observées, le fait qu'une donnée soit manquante ne dépend plus de sa vraie valeur non observée.

Avantage principal

Son principal avantage est qu'elle prend en compte l'incertitude liée à l'imputation. Contrairement à l'imputation simple, elle ne remplace pas une donnée manquante par une seule valeur fixe.

Limite principale

Elle est plus complexe à mettre en œuvre et nécessite de bien spécifier les modèles d'imputation. Si les modèles sont mal spécifiés, les résultats peuvent être biaisés.

MICE est une méthode puissante, mais elle demande une bonne modélisation.

5. Imputation par k plus proches voisins

Principe

L'imputation par k plus proches voisins, ou k-NN, consiste à remplacer une valeur manquante à partir des valeurs observées chez les individus les plus proches.

La proximité est généralement calculée à partir d'une distance entre observations, par exemple la distance euclidienne.

Pour une observation i ayant une valeur manquante, on identifie ses k voisins les plus proches, puis on impute par la moyenne ou la médiane de ces voisins.

Hypothèse requise

Cette méthode est surtout adaptée sous MCAR ou MAR, à condition que les variables observées permettent de trouver des individus réellement comparables.

Avantage principal

Elle est intuitive et flexible. Elle ne suppose pas forcément une relation linéaire entre les variables.

Limite principale

Elle peut être sensible au choix de k, à l'échelle des variables et à la qualité de la mesure de distance. Elle devient aussi moins efficace lorsque le nombre de variables est très grand.

k-NN est flexible, mais dépend fortement du choix des voisins.

Résumé comparatif

Méthode	Principe	Hypothèse	Avantage	Limite
Moyenne	Remplacer les valeurs manquantes par la moyenne observée.	MCAR	Très simple.	Réduit la variance et les corrélations.
Médiane	Remplacer les valeurs manquantes par la médiane observée.	MCAR	Robuste aux valeurs extrêmes.	Réduit aussi la dispersion des données.
Régression	Prédire la valeur manquante à partir des autres variables.	MAR	Utilise l'information disponible.	Sous-estime souvent l'incertitude.
MICE	Créer plusieurs bases imputées puis combiner les résultats.	MAR	Prend en compte l'incertitude d'imputation.	Plus complexe et dépend des modèles choisis.
k-NN	Imputer à partir des individus les plus proches.	MCAR ou MAR	Flexible et non nécessairement linéaire.	Sensible au choix de k et aux distances.

Conclusion

Les méthodes simples comme la moyenne ou la médiane sont faciles à appliquer, mais elles peuvent déformer la distribution des données.

Les méthodes plus avancées comme la régression, k-NN ou MICE utilisent davantage d'information et sont souvent préférables lorsque les données sont MAR.

En revanche, si les données sont MNAR, les méthodes classiques d'imputation peuvent rester biaisées, car le mécanisme de manque dépend directement des valeurs non observées.

Le choix de la méthode d'imputation dépend du mécanisme de manque et de l'objectif de l'analyse.

Question 9 *(Impact de l'imputation sur $\hat{\beta}$)*

Données manquantes MAR et biais d'imputation

On considère le modèle :

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

avec :

$$E(\varepsilon_i \mid X_i, Z_i) = 0$$

La variable X est manquante selon une variable auxiliaire observable Z. On est donc dans un cas MAR :

$$P(R_i = 1 \mid X_i, Z_i) = P(R_i = 1 \mid Z_i)$$

où :

$$R_i = \begin{cases} 1 & \text{si } X_i \text{ est observée,} \\ 0 & \text{si } X_i \text{ est manquante.} \end{cases}$$

On rappelle que, dans une régression simple, le coefficient de pente peut s'écrire :

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

Lorsqu'on impute X, on utilise plutôt :

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

(a) Imputation par la médiane inconditionnelle

Soit m la médiane inconditionnelle de X. L'imputation par la médiane consiste à poser :

$$X_i^{med} = R_i X_i + (1 - R_i)m$$

Ainsi, si X_i est observée, on garde sa vraie valeur. Si X_i est manquante, on la remplace par la même constante m.

Le coefficient estimé après imputation est :

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

Or le vrai modèle donne :

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

Donc :

$$\mathrm{Cov}(X^{med}, Y) = \mathrm{Cov}(X^{med}, \beta_0 + \beta_1 X + \varepsilon)$$

Comme β_0 est une constante :

$$\mathrm{Cov}(X^{med}, \beta_0) = 0$$

Donc :

$$\mathrm{Cov}(X^{med}, Y) = \beta_1 \,\mathrm{Cov}(X^{med}, X) + \mathrm{Cov}(X^{med}, \varepsilon)$$

Sous exogénéité :

$$\mathrm{Cov}(X^{med}, \varepsilon) = 0$$

Donc :

$$\mathrm{Cov}(X^{med}, Y) = \beta_1 \,\mathrm{Cov}(X^{med}, X)$$

Ainsi :

$$E(\hat{\beta}_1^{med}) = \beta_1 \frac{\mathrm{Cov}(X^{med}, X)}{\mathrm{Var}(X^{med})}$$

Le biais est donc :

$$\mathrm{Biais}(\hat{\beta}_1^{med}) = E(\hat{\beta}_1^{med}) - \beta_1$$

$$\mathrm{Biais}(\hat{\beta}_1^{med}) = \beta_1 \frac{\mathrm{Cov}(X^{med}, X)}{\mathrm{Var}(X^{med})} - \beta_1$$

Donc :

$$\boxed{\mathrm{Biais}(\hat{\beta}_1^{med}) = \beta_1 \left[\frac{\mathrm{Cov}(X^{med}, X)}{\mathrm{Var}(X^{med})} - 1 \right]}$$

Or, en général :

$$\mathrm{Cov}(X^{med}, X) \neq \mathrm{Var}(X^{med})$$

car l'imputation par une constante écrase la variabilité de X, surtout lorsque les données manquantes dépendent de Z.

Donc :

$$\boxed{E(\hat{\beta}_1^{med}) \neq \beta_1}$$

Ainsi, l'imputation par la médiane inconditionnelle produit généralement un estimateur biaisé de β_1.

Interprétation du biais

L'imputation par la médiane remplace toutes les valeurs manquantes par une même valeur. Elle ignore donc l'information contenue dans Z, alors que le mécanisme de manque dépend justement de Z.

Par conséquent, les individus avec données manquantes ne sont pas représentés correctement. La distribution de X imputée est artificiellement concentrée autour de la médiane.

Cela modifie :

$$\mathrm{Cov}(X, Y)$$

et :

$$\mathrm{Var}(X)$$

Donc la pente estimée est biaisée.

(b) Imputation de X par régression sur Z

Comme X est MAR selon Z, il est naturel d'utiliser l'information contenue dans Z.

On suppose que :

$$E(X \mid Z) = g(Z)$$

Par exemple, dans un modèle linéaire :

$$X_i = \alpha_0 + \alpha_1 Z_i + \eta_i$$

avec :

$$E(\eta_i \mid Z_i) = 0$$

On impute alors les valeurs manquantes par la valeur prédite :

$$\widehat{X}_i = E(X_i \mid Z_i)$$

ou, en pratique :

$$\widehat{X}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i$$

La variable imputée devient :

$$X_i^{reg} = R_i X_i + (1 - R_i)\widehat{X}_i$$

En population, on peut écrire :

$$X_i^{reg} = R_i X_i + (1 - R_i)E(X_i \mid Z_i)$$

Posons :

$$X_i = E(X_i \mid Z_i) + \eta_i$$

avec :

$$E(\eta_i \mid Z_i) = 0$$

Alors, pour les valeurs manquantes, on remplace X_i par $E(X_i \mid Z_i)$. L'erreur d'imputation est donc :

$$e_i = X_i - X_i^{reg}$$

Pour les observations complètes, $e_i = 0$. Pour les observations manquantes :

$$e_i = X_i - E(X_i \mid Z_i) = \eta_i$$

Donc :

$$X_i = X_i^{reg} + e_i$$

Le coefficient estimé avec X^{reg} est :

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

Or :

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

Donc :

$$\mathrm{Cov}(X^{reg}, Y) = \mathrm{Cov}(X^{reg}, \beta_0 + \beta_1 X + \varepsilon)$$

$$\mathrm{Cov}(X^{reg}, Y) = \beta_1 \,\mathrm{Cov}(X^{reg}, X) + \mathrm{Cov}(X^{reg}, \varepsilon)$$

Sous exogénéité :

$$\mathrm{Cov}(X^{reg}, \varepsilon) = 0$$

Donc :

$$\mathrm{Cov}(X^{reg}, Y) = \beta_1 \,\mathrm{Cov}(X^{reg}, X)$$

Or :

$$X = X^{reg} + e$$

Donc :

$$\mathrm{Cov}(X^{reg}, X) = \mathrm{Cov}(X^{reg}, X^{reg} + e)$$

$$\mathrm{Cov}(X^{reg}, X) = \mathrm{Var}(X^{reg}) + \mathrm{Cov}(X^{reg}, e)$$

Sous MAR et si le modèle d'imputation est correctement spécifié, l'erreur d'imputation e est orthogonale à X^{reg}. Donc :

$$\mathrm{Cov}(X^{reg}, e) = 0$$

Ainsi :

$$\mathrm{Cov}(X^{reg}, X) = \mathrm{Var}(X^{reg})$$

Donc :

$$\mathrm{Cov}(X^{reg}, Y) = \beta_1 \,\mathrm{Var}(X^{reg})$$

Par conséquent :

$$E(\hat{\beta}_1^{reg}) = \frac{\beta_1 \,\mathrm{Var}(X^{reg})}{\mathrm{Var}(X^{reg})}$$

Donc :

$$\boxed{E(\hat{\beta}_1^{reg}) = \beta_1}$$

Ainsi, une imputation de X par régression sur Z peut redonner un estimateur sans biais sous MAR, à condition que le modèle d'imputation soit correctement spécifié.

(c) Stratégie recommandée

Si les données sont MCAR, c'est-à-dire manquantes complètement au hasard, des méthodes simples comme la suppression des lignes ou une imputation simple peuvent parfois être acceptables, surtout si le taux de valeurs manquantes est faible.

Mais ici, les données sont MAR, car le manque de X dépend de la variable auxiliaire observable Z.

Dans ce cas, il ne faut pas imputer X par sa médiane inconditionnelle, car cette méthode ignore Z. Elle risque donc de produire un biais.

La stratégie recommandée est d'utiliser une méthode qui exploite Z, par exemple :

imputation par régression de X sur Z

ou mieux encore :

imputation multiple de type MICE

L'imputation par régression est adaptée si la relation entre X et Z est bien modélisée.

L'imputation multiple est préférable si l'on veut aussi tenir compte de l'incertitude liée à l'imputation.

Conclusion

Sous MAR, le mécanisme de données manquantes dépend d'une variable observable Z. Il faut donc utiliser cette variable dans l'imputation.

L'imputation par la médiane inconditionnelle est généralement biaisée car elle ignore Z.

En revanche, une imputation par régression sur Z est plus appropriée et peut fournir un estimateur sans biais si le modèle d'imputation est correctement spécifié.

Sous MAR, il faut imputer conditionnellement aux variables qui expliquent le manque.

Partie B : Pratique Python

B1 Exploration des données

Question 1

Interprétation

Le dataset contient **1338 observations** et **7 variables**.

Les variables disponibles sont :

- `age` : c'est une variable numérique entière ;
- `sex` : c'est une variable qualitative ;
- `bmi` : c'est une variable numérique continue ;
- `children` : c'est une variable numérique entière ;
- `smoker` : c'est une variable qualitative ;
- `region` : c'est une variable qualitative ;
- `charges` : c'est une variable numérique continue correspondant aux frais médicaux annuels.

Les statistiques descriptives montrent que la variable `charges` a une moyenne d'environ :

$$\overline{x}_{\texttt{charges}} = 13264{,}16$$

et une médiane d'environ :

$$\text{Med}(\texttt{charges}) = 9391{,}35$$

La moyenne est donc nettement supérieure à la médiane. Cela indique que la distribution des frais médicaux est asymétrique à droite. Autrement dit, une partie des assurés a des charges médicales très élevées, ce qui tire la moyenne vers le haut.

Cette situation est fréquente en assurance santé : la majorité des assurés ont des frais modérés, tandis qu'un petit nombre d'assurés présente des coûts médicaux très importants.

Suite voir fichier wilfried avademe devoir1.ipynb