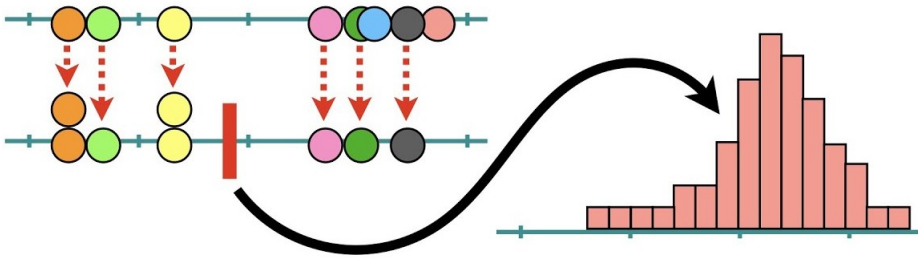
Faculté des Sciences
Aix*Marseille Université

Aix-Marseille Université

Mémoire de TER
M1 Mathématiques Appliquées, Statistiques

Bootstrap et Modèles de Régression



Moussa DIENG
Mohamed Bacar ABDOULANDHUM
Fatima SELMANI

Dirigé par : M. Pierre PUDLO

Avril 2024

Table des matières

1 Introduction

Les méthodes classiques d'inférence statistique ne permettent pas d'obtenir des réponses correctes à tous les problèmes concrets que l'on se pose. Elles ne sont en effet valables que sous des conditions d'application particulières. Ainsi, par exemple, le test t de Student d'égalité des moyennes suppose que les deux populations-parents sont normales, de même variance et que les deux échantillons sont aléatoires, simples et indépendants. Le calcul de l'intervalle de confiance d'une variance par l'intermédiaire des variables χ_2 suppose que la population-parent est normale et que l'échantillon est aléatoire et simple. L'inférence statistique classique en régression suppose, outre les conditions d'application relatives à la population et à l'échantillon, que le modèle est ajusté au sens des moindres carrés.
Que peut-on faire en pratique lorsque ces conditions d'application ne sont pas remplies ? Différentes attitudes sont possibles.
Dans certains cas, les méthodes classiques sont utilisées malgré le non-respect des conditions. Cette utilisation est alors justifiée par le caractère robuste des méthodes qui garantit que les résultats de l'inférence restent approximativement valables. Une autre attitude consiste à abandonner les méthodes paramétriques d'inférence statistique au profit de méthodes non paramétriques, pour lesquelles les conditions d'application sont bien moins restrictives. L'accès généralisé à des moyens de calcul puissants a permis le développement de méthodes d'inférence statistique basées sur l'utilisation intensive de l'ordinateur. Le **bootstrap** fait partie de ces méthodes. Le mot bootstrap provient de l'expression anglaise "to pull oneself up by one's bootstrap" [4], qui signifie littéralement "se soulever en tirant sur les languettes de ses bottes". Le mot bootstrap fait penser à des traductions telles que "à la force du poignet" ou "par soi-même" ou "passe partout" [1], mais en fait il n'est jamais traduit dans la littérature scientifique d'expression française.
Le principe général de la méthode est de rééchantillonner un grand nombre de fois l'échantillon initial qui a été réellement prélevé dans la population, l'inférence statistique étant basée sur les résultats des échantillons ainsi obtenus.
Dans ce projet, nous explorons l'association entre Bootstrap et la régression linéaire. Nous examinerons comment le bootstrap peut être utilisé pour améliorer la robustesse et la fiabilité des modèles de régression linéaire, notamment dans des contextes où les hypothèses classiques de la régression linéaire sont violées. Nous explorerons également les différentes méthodes de bootstrap, telles que le bootstrap paramétrique et non paramétrique, et nous illustrerons leur application à travers des exemples pratiques.

2 Méthodologie

Avant tout, il est important de présenter les méthodes que nous allons utiliser pour bien comprendre en quoi cela consiste. Dans cette partie, nous allons parler de la méthodologie du bootstrap, de la régression linéaire, ainsi que de l'utilisation du bootstrap en régression linéaire.

2.1 Méthodologie bootstrap

Dans cette section, nous discutons des techniques applicables à un seul échantillon homogène de données, noté par y_1,...,y_n. Les valeurs de l'échantillon sont considérées comme les résultats de variables aléatoires indépendantes et identiquement distribuées Y_1,...,Y_n dont nous noterons la fonction de densité de probabilité et la fonction de répartition par f et F, respectivement. L'échantillon est utilisé pour tirer des inférences sur une caractéristique de la population, généralement désignée par θ, en utilisant une statistique T dont la valeur dans l'échantillon est t. Nous supposons pour le moment que le choix de T a été fait et qu'il s'agit d'une estimation pour θ, que nous considérons comme un scalaire.
Nous concentrons notre attention sur les interrogations relatives à la distribution de probabilité de T. Par exemple, quel est son biais, son erreur-type, ou ses quantiles ? Quelles valeurs sont probables sous une hypothèse nulle spécifique d'intérêt ? Comment calculons-nous les intervalles de confiance pour θ en utilisant T ?
Il y a deux situations à distinguer, le paramétrique et le non paramétrique. Lorsqu'il existe un modèle mathématique particulier, avec des constantes ajustables ou des paramètres Φ qui déterminent entièrement f, un tel modèle est appelé paramétrique et les méthodes statistiques basées sur ce modèle sont des méthodes paramétriques. Dans ce cas, le paramètre d'intérêt θ est une composante ou une fonction de Φ. Lorsqu'aucun modèle mathématique de ce type n'est utilisé, l'analyse statistique est non paramétrique et utilise seulement le fait que les variables aléatoires Y_j sont indépendantes et identiquement distribuées.

2.1.1 Bootstrap non paramétrique

Le bootstrap non paramétrique est une technique utilisée lorsque nous ne disposons pas d'un modèle paramétrique, mais où il est raisonnable de supposer que $(Y_1, \ldots, Y_n)$ sont des observations indépendantes et identiquement distribuées selon une fonction de distribution inconnue F. Dans ce contexte, nous utilisons la fonction de répartition empirique $\hat{F}$ pour estimer la fonction de distribution inconnue F.
Prenons un exemple simple pour illustrer ce principe. Considérons un échantillon aléatoire de taille n observé à partir d'une distribution inconnue F :

$$Y_i = y_i, \quad Y_i \sim F, \quad i = 1, 2, \ldots, n$$

Définissons $Y = (Y_1, Y_2, \ldots, Y_n)$ et $y = (y_1, y_2, \ldots, y_n)$ comme l'échantillon aléatoire et sa réalisation observée, respectivement.
On souhaite estimer la distribution d'un estimateur $T(Y, F)$ d'un paramètre θ. T dépend éventuellement à la fois de Y et de la distribution inconnue F.
La méthode bootstrap pour le problème de l'échantillon unique est relativement simple et peut se résumer par les points suivants :

- Construire la distribution de probabilité de l'échantillon $\hat{F}$ en attribuant une masse de $1/n$ à chaque point y_i de l'échantillon $y_1, y_2, \ldots, y_n$.
- Une fois $\hat{F}$ fixée, tirer un échantillon aléatoire de taille n suivant $\hat{F}$. Cet échantillon est appelé l'échantillon bootstrap, noté $Y^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)$, avec $y^* = (y_1^*, y_2^*, \ldots, y_n^*)$. En d'autres termes, les valeurs de Y^* sont obtenues en effectuant des tirages avec remise à partir de l'ensemble $(y_1, y_2, \ldots, y_n)$.

— Approcher la distribution d'échantillonnage de $T(Y, F)$ par la distribution bootstrap de $T(Y^*, \hat{F})$.

Supposons que notre paramètre d'intérêt est la moyenne θ de l'échantillon $(Y_1, Y_2, \ldots, Y_n)$ et que nous souhaitons estimer l'écart-type d'un estimateur T pour la moyenne θ. Comme nous ne connaissons pas la distribution F de l'échantillon, nous allons donc utiliser $\hat{F}$.

Dans un premier temps, nous créons plusieurs échantillons bootstrap $y^* = (y_1^*, y_2^*, \ldots, y_n^*)$. Chaque échantillon bootstrap est obtenu en sélectionnant aléatoirement y_i^* avec remplacement à partir de l'ensemble $\{y_1, y_2, \ldots, y_n\}$. Pour chaque échantillon, on calcule une réplique bootstrap T^* de la statistique $T(y)$.

Ensuite, nous générons un grand nombre R d'échantillons bootstrap de manière indépendante. Pour chaque échantillon bootstrap, nous calculons la valeur de $T(y^*)$, que nous notons T_r^* pour $r = 1, 2, \ldots, R$.

Enfin, nous calculons l'estimation bootstrap de l'écart-type pour θ en prenant l'écart-type empirique de ces valeurs T_r^*. L'estimation bootstrap de l'erreur standard pour θ est calculée comme l'écart-type empirique des valeurs T_r^*, notée $\hat{\sigma}^*$, et calculée de la manière suivante :

$$\hat{\sigma}^* = \sqrt{\frac{\sum_{r=1}^{R}(T_r^* - T^*)^2}{R(R-1)}}$$

où T^* est la moyenne des valeurs T_r^* et est donnée par :

$$T^* = \frac{\sum_{r=1}^{R} T_r^*}{R}$$

L'exemple ci-dessus concerne le cas d'un échantillon simple, mais nous verrons par la suite de nombreuses extensions de cette méthode dans le cas d'échantillons multiples.

2.1.2 Bootstrap Paramétrique

Contrairement au bootstrap non paramétrique, qui repose sur des rééchantillonnages aléatoires des données sans faire d'hypothèses sur la distribution sous-jacente, le bootstrap paramétrique exploite des modèles statistiques pour générer des échantillons bootstrap.

L'une des principales caractéristiques du bootstrap paramétrique réside dans son utilisation de modèles statistiques pour estimer la distribution sous-jacente des données. Plutôt que de rééchantillonner directement à partir des données observées, cette méthode suppose l'existence d'un modèle théorique qui décrit la distribution des données. Cette approche permet d'obtenir des estimations plus précises des erreurs standards et des intervalles de confiance, notamment dans des cas où les données présentent une complexité ou des distributions non standardisées.

Pour appliquer le bootstrap paramétrique, plusieurs étapes sont nécessaires. Tout d'abord, il est crucial de choisir un modèle de distribution approprié qui corresponde aux données observées. Ensuite, les paramètres de ce modèle sont estimés à partir des données d'échantillon à l'aide de méthodes telles que la méthode des moindres carrés ou la maximisation de la vraisemblance. Une fois les paramètres estimés, des échantillons bootstrap sont générés en simulant des données à partir du modèle avec les nouveaux paramètres estimés. Ces échantillons bootstrap sont ensuite utilisés pour recalculer les estimations des paramètres ou des prévisions du modèle, ce qui permet d'obtenir une distribution empirique des estimations paramétriques.

Enfin, les erreurs standards des estimations sont évaluées en utilisant les échantillons bootstrap. Les valeurs obtenues à partir de ces échantillons sont utilisées pour calculer les erreurs standards, ce qui permet d'évaluer la précision des résultats obtenus à partir des données d'échantillon. En résumé, le bootstrap paramétrique constitue une approche puissante pour estimer la variabilité des estimations de paramètres ou des prévisions de modèles, en tenant compte de l'incertitude associée aux estimations des paramètres dans un contexte où un modèle statistique est préalablement spécifié. **Exemple d'application** :
Considérons un échantillon $X = (x_1, x_2, ..., x_n)$ de taille n provenant d'une distribution $\mathcal{N}(\mu, \sigma^2)$, où μ est la moyenne et σ^2 est la variance que l'on suppose connue.
L'estimateur du maximum de vraisemblance (EMV) pour la moyenne μ est donné par la moyenne empirique de l'échantillon :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Nous souhaitons estimer l'écart-type de $\hat{\mu}$ par la méthode du bootstrap paramétrique. On génère R échantillons bootstrap de la forme $X^* = (x_1^*, x_2^*, ..., x_n^*)$, en tirant aléatoirement n valeurs à partir de la distribution $\mathcal{N}(\hat{\mu}, \sigma^2)$, où $\hat{\mu}$ est l'estimateur de la moyenne à partir de l'échantillon original. Cela peut être formulé comme suit :

$$x_i^* \sim \mathcal{N}(\hat{\mu}, \sigma^2), \quad \text{pour } i = 1, 2, ..., n$$

Calcul de la statistique d'intérêt :
Pour chaque échantillon bootstrap X^*, calculons la moyenne bootstrap $\hat{\mu}_r^*$ (avec $r = 1, ..., R$), qui est également l'estimation de la moyenne à partir de cet échantillon bootstrap :

$$\hat{\mu}_r^* = \frac{1}{n}\sum_{i=1}^{n} x_i^*$$

Estimation de l'erreur standard :
Pour estimer l'erreur standard de notre estimation de la moyenne, calculons l'écart-type des moyennes bootstrap $\hat{\mu}_r^*$ obtenues à partir des échantillons bootstrap :

$$\hat{\sigma}^*_{\text{boots}} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_r^* - \hat{\mu})^2}$$

où R est le nombre d'échantillons bootstrap et $\hat{\mu}_r^*$ est l'estimateur bootstrap de la moyenne pour le r-ième échantillon bootstrap (après avoir généré R échantillons bootstrap, nous obtenons R estimations de la moyenne pour chaque échantillon bootstrap).

2.2 Rappel sur la regression linéaire simple et multiple

Considérons un ensemble de données $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, où x_i est une valeur de la variable indépendante X et y_i est une valeur de la variable dépendante Y . Le modèle de régression linéaire simple suppose que la relation entre X et Y peut être approximée par une droite :

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

où :

— β_0 et β_1 sont deux paramètres qui caractérisent le lien entre X et Y.
— ε représente le terme d'erreur aléatoire et est indépendant de X, tel que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
Sous les hypothèses du modèle linéaire simple gaussien, on a :

$$[Y \mid X = x] \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

et les estimations des paramètres par moindres carrés donnent :

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Il est d'usage d'estimer la variance σ^2 par :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

qui a des bonnes propriétés.
Ces estimations des paramètres sont normalement distribuées si les erreurs, c'est-à-dire les ε_i sont gaussiennes. Elles sont souvent approximativement normales pour d'autres distributions d'erreur, mais elles ne sont pas robustes face à une non-normalité prononcée des erreurs ou à des valeurs de réponse aberrantes.
Les résidus $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ peuvent être modifiés de différentes manières pour les rendre adaptés aux méthodes de diagnostic, mais la modification la plus utile pour nos besoins est de les changer pour avoir une variance constante, c'est-à-dire :

$$r_j = \frac{y_j - \hat{\mu}_j}{(1 - h_j)^{\frac{1}{2}}} \tag{2}$$

avec

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_j^n (x_j - \bar{x})^2}$$

Ainsi, nous utiliserons cette définition pour le bootstrap des résidus.
L'extension du modèle de régression linéaire simple au modèle à plusieurs variables explicatives est :

$$Y_j = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp} + \varepsilon_j, \quad \forall \quad j = 1, \ldots, n$$

On peut écrire matriciellement ces n équations :

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

où $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$, $X = \begin{pmatrix} 1 & X_{1,1} & \ldots & X_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & \ldots & X_{n,p} \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{pmatrix}$
Comme dans le cas de la régression linéaire simple, les Y_j sont supposés indépendants.
Si $\text{rang}(X) = p + 1$, l'estimation du vecteur des coefficients par moindres carrés est donnée par :

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

2.3 Bootstrap et régression linéaire

Le problème en régression linéaire est l'étude des effets des variables explicatives ou covariables sur une variable de réponse. Dans le cas où les erreurs aléatoires sont normales et ont une variance constante, la théorie des moindres carrés de l'estimation et de l'inférence fournit des méthodes précises et exactes pour l'analyse. Mais pour les généralisations aux erreurs non normales et à la variance non constante, des méthodes exactes existent rarement, et nous sommes confrontés à des méthodes approximatives basées sur des approximations linéaires aux estimateurs et sur les théorèmes de la limite centrale. Ainsi, les méthodes de rééchantillonnage ont le potentiel de fournir une analyse plus précise.
Dans cette partie, nous allons présenter trois méthodes de bootstrap en régression linéaire : le bootstrap paramétrique, le bootstrap des résidus et le bootstrap par paires.

2.3.1 Bootstrap paramétrique

La méthode de Bootstrap paramétrique en régression linéaire est utilisée lorsque les hypothèses gaussiennes du modèle et l'hypothèse d'homoscédasticité sont vérifiées. Ce qui correspond au modèle 1 introduit dans la section 2.2.
Dans cette méthode, nous générons des échantillons bootstrap en respectant cette distribution normale pour les résidus.
L'algorithme de la méthode de Bootstrap paramétrique en régression linéaire se présente comme suit :

1. Estimer les paramètres $\hat{\beta}_0$ et $\hat{\beta}_1$ du modèle de régression linéaire à partir de l'échantillon original.
2. Calculer les résidus $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.
3. Calculer la moyenne $\bar{e}$ et l'écart-type s des résidus.
4. Pour $r = 1, ..., R$
   — Générer des échantillons bootstrap $(e_1^*, ..., e_n^*)$ en tirant aléatoirement des résidus à partir de la distribution $\mathcal{N}(\bar{e}, s^2)$.
   — Calculer les valeurs prédites $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i^*$ pour $i = 1, \ldots, n$
   — Estimer les paramètres du modèle de régression $\hat{\beta}_0^*$ et $\hat{\beta}_1^*$ linéaire à partir des valeurs y_i^* en utilisant la méthode des moindres carrés ordinaires.

À la fin nous disposons des échantillons de taille R contenant les estimations des paramètres β_0 et β_1 , permettant ainsi de mener des analyses statistiques sur ces estimations.

2.3.2 Bootstrap des résidus

Le bootstrap des résidus généralement utilisés dans les problèmes de régression linéaire, est une méthode statistique qui repose sur le rééchantillonnage des résidus obtenus à partir du modèle de régression initial permettant ainsi d'estimer la distribution empirique des résidus et d'effectuer des inférences sur les paramètres du modèle. C'est une alternative pour avoir des estimations robustes des paramètres ainsi que des intervalles de confiances fiables sans avoir une idée a priori sur la distribution des données. Ce pendant comme le problème de la régression linéaire, le Bootstrap des résidus est aussi basé sur des hypothèses qui sont : La normalités des résidus et l'homoscédasticité. Pour l'illustrer, nous allons montrer les différents

méthodes de rééchantillonage du bootstrap des résidus et l'appliquer dans le problème de la régression linéaire.
On se place d'abord dans le cas de la régression linéaire simple. Considérant le modèle 1 de la partie 2.2
Pour étendre les algorithmes de bootstrap à la régression basée sur les résidus, notre première étape consiste à identifier le modèle sous-jacent F. Si l'équation 1 est exacte avec des erreurs homoscédastiques, alors ces erreurs sont effectivement tirées d'une seule distribution noté G. La méthode Bootstrap des résidus consiste à tirer de manière indépendante les résidus si les x_i sont fixes et puis créer des échantillons Bootstrap des résidus. Mais pour un usage pratique, il est préférable d'utiliser les résidus r_j car leur variance sont en accord avec celles des e_j. Notons que G est supposé avoir une moyenne nulle, nous estimons alors G par la distribution empirique de $r_j - \bar{r}$, oú $\bar{r}$ est la moyenne des r_j. Ces résidus centrés sont alors de moyenne nulle et dont leur distribution empirique est G, soit $r_j - \bar{r} \sim G$. Ainsi L'échantillon bootstrap des résidus noté $\epsilon^* = (\epsilon_1^*, \ldots, \epsilon_n^*)$ est obtenue par rééchantillonnage en appliquant la méthode de bootstrap non paramétrique vue à la partie 2.1.1 sur l'échantillon des $r_j - \bar{r}$, $j = \{1, 2, \ldots, n\}$.

Le modèle de rééchantillonnage complet est considéré comme ayant le même plan que les données, c'est-à-dire $x^* \equiv x_j$.
L'échantillon bootstrap pour les données devient :

$$Y_j^* = \hat{\mu}_j + \epsilon_j^*$$

avec $\hat{\mu}_j = \hat{\beta}_0 + \hat{\beta}_1 * x_j$, et $\epsilon_j^* \sim \hat{G}$. Donc l'algorithme pour générer des ensembles de données simulés et les estimations de paramètres correspondantes est le suivant.
For $r = 1, \ldots, R$,

1. For $j = 1, \ldots, n$,
   — $x_j^* = x_j$;
   — Échantillonner aléatoirement ϵ_j^* parmi les $r_j - \bar{r}$
   — $y_j^* = \hat{\beta}_0 + \hat{\beta}_1 * x_j + \epsilon_j^*$
2. Ajuster la régression des moindres carrés à $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$, fournissant des estimations $\hat{\beta}_{0,r}^*, \hat{\beta}_{1,r}^*, s_r^{*2}$.

Remarque : Cette méthode de rééchantillonage de bootstrap n'est valable que sous l'hypothèse d'homoscédasticité c'est à dire variance des résidus constante. Dans le cas contraire la variance des résidus est une fonction des observations x_i ce qui fait qu'on pourra plus tirer de manière indépendante les résidus pour en créer un nouveau échantillon. Ce-ci est une limite du boostrap des résidus. Néanmoins si l'hétéroscédasticité peut être modélisée, alors la simulation bootstrap en rééchantillonnant les erreurs reste possible.
Considérons notre modèle de régression linéaire précédent 1, l'erreur ε_j à $x = x_j$ a pour variance $\sigma_j^2 = f(x_j)$ ou $\sigma_j^2 = f(\mu_j)$ avec f une fonction connue. Nous n'avons besoin que les résidus modifiés

$$r_j = \frac{1}{f(x_j)(1 - h_j)^{\frac{1}{2}}} e_j \tag{3}$$

qui seront approximativement homoscédastiques. La fonction de répartition empirique (EDF) de ces résidus modifiés, après soustraction de leur moyenne, estimera la fonction de distribution

G des erreurs aléatoires homoscédastiques mises à l'échelle δ_j dans le modèle

$$Y_j^* = \hat{\beta}_0 + \hat{\beta}_1 * x_j + f_j^{\frac{1}{2}} \delta_j$$

oú $f_j = f(x_j)$. Et l'algorithme de rééchantillonnage est maintenant modifié comme suivant :
For $r = 1, \ldots, R$,

1. For $j = 1, \ldots, n$,
   — $x_j^* = x_j$;
   — Échantillonner aléatoirement ϵ_j^* parmi les $r_j - \bar{r}$
   — $y_j^* = \hat{\beta}_0 + \hat{\beta}_1 * x_j + \hat{f}_j^{\frac{1}{2}} \delta_j^*$
2. Ajuster la régression des moindres carrés à $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$, fournissant des estimations $\hat{\beta}_{0,r}^*, \hat{\beta}_{1,r}^*, s_r^{*2}$.

Cas de la régression linéaire généralisé : Les même conditions d'application du bootstrap des résidus dans le cas de la régression linéaire simple sont aussi valables dans le cas multiple c'est à dire : Les résidus doivent être indépendants et normalement distribués de variance constante (Homoscédasticité). Et pour l'algorithme d'échantillonage est une généralisation de l'algorithme défini précédemment dans le cas simple. Il est basé sur le modèle d'échantillonnage généralisé des données suivant :

$$Y_j^* = x_j^T \hat{\beta}_j + \epsilon_j^*$$

Les ϵ_j^* sont échantillonnés aléatoirement à partir des résidus $r_1, \ldots, r_n$ ou les résidus centrés $r_1 - \bar{r}, \ldots, r_n - \bar{r}$.

Conséquences : Lorsque l'hétéroscédasticité est présente et non modélisée correctement, le bootstrap des résidus peut produire des estimations biaisées de la distribution des erreurs. Les échantillons bootstrap ne refléteront pas correctement la variabilité accrue des erreurs observée dans les données réelles, ce qui peut conduire à des intervalles de confiance incorrects ou à des tests d'hypothèses inappropriés.

Dans de tels cas, il serait plus approprié d'utiliser des méthodes de bootstrap spécifiques pour traiter l'hétéroscédasticité, comme le bootstrap par blocs. De plus, des techniques de modélisation spécifiques, telles que les modèles linéaires avec termes d'erreur hétéroscédastiques, pourraient être nécessaires pour traiter correctement la structure de l'hétéroscédasticité dans les données.

2.3.3 Bootstrap par paires

Le bootstrap par paires est une approche qui ne nécessite pas l'hypothèse d'homoscédasticité, ce qui le distingue de la méthode précédente. Cette méthode peut être appliquée même si les termes d'erreur ont des variances différentes. En cas de l'hétéroscédasticité, on ne peut pas rééchantillonner les résidus de manière indépendante , le bootstrap par paires consiste à rééchantillonner directement à partir des données originales, comme s'est expliqué dans [3]. Cela signifie considérer les données comme un échantillon d'une certaine distribution bivariée F de (X, Y), où les coefficients de régression sont considérés comme des fonctions statistiques de F.

F étant la distribution bivariée de (X, Y), il est approprié de prendre $\hat{F}$ comme fonction densité empirique des paires de données, et le rééchantillonnage se fera à partir $\hat{F}$.
La simulation de rééchantillonnage implique donc d'échantillonner des paires avec remplacement à partir de $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Cela équivaut à prendre $(x_i^*, y_i^*) = (x_i, y_i)$, où i est uniformément distribué sur $\{1, 2, \ldots, n\}$. Les valeurs simulées $\hat{\beta}_0^*$, $\hat{\beta}_1^*$ des estimations des coefficients sont calculées à partir de $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$ en utilisant la méthode des moindres carrés. Ce qui correspond à l'algorithme de rééchantillonnage suivant.
Pour $r = 1$ à R,

- Échantillonner $i_1^*, \ldots, i_n^*$ de manière aléatoire avec remplacement dans $\{1, 2, \ldots, n\}$.
- Pour $j = 1, \ldots, n$, définir $x_j^* = x_{i_j^*}$, $y_j^* = y_{i_j^*}$.
- Ajuster la régression des moindres carrés avec $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$, fournissant des estimations $\hat{\beta}_{0,r}^*, \hat{\beta}_{1,r}^*, s_r^{*2}$.

Il est important de noter que Cette méthode ne suppose aucune hypothèse sur l'homogénéité de la variance et même la linéarité de la moyenne conditionnelle de Y pour un $X = x$ donné n'est pas présumée. Cela offre l'avantage d'une éventuelle robustesse à l'hétéroscédasticité, mais peut entraîner une inefficacité si le modèle de variance constante est correct. Dans d'autres situations, telles que celles que nous aborderons dans la section suivante sur le bootstrap, cette méthode peut ne pas être fiable.

2.4 Vérification d'un intervalle de confiance

Un intervalle de confiance d'un paramètre θ est un intervalle probabiliste dans lequel le paramètre θ a une probabilité $1 - \alpha$ ($\alpha \in [0; 1]$ est le seuil d'erreur choisi) d'être contenu. Mathématiquement, cela se traduit par :

$$\mathbb{P}(X : \{\theta \in [L(X), R(X)]\}) = 1 - \alpha$$

où $L(X)$ et $R(X)$ sont les bornes de l'intervalle de confiance et dépendent du jeu de données X.
Pour vérifier si un intervalle de confiance est correct, voici une méthodologie :
Soit $X = (X_1, \ldots, X_n)$ un échantillon suivant une distribution F_θ de paramètre θ et soit IC un intervalle de confiance du paramètre θ. Pour vérifier si cet intervalle est correct, on simule dans un premier temps N échantillons suivant la loi F_θ avec θ fixé. Ensuite, pour chaque échantillon, on construit un intervalle de confiance à $1 - \alpha$ et on vérifie si le paramètre θ initial est contenu dans chaque intervalle de confiance. Enfin, on calcule la fréquence α' du nombre d'échantillons pour lesquels le paramètre est contenu dans les intervalles de confiance. Si α' est proche de $1 - \alpha$, nous concluons que l'intervalle de confiance est valide. Si α' est significativement plus petit que $1 - \alpha$, nous disons que l'intervalle de confiance sous-estime l'erreur. En revanche, si α' est significativement plus grand que $1 - \alpha$, alors notre intervalle de confiance surestime l'erreur.
Un peu plus loin dans ce travail, nous aurons besoin de vérifier si les intervalles de confiance bootstrap, et éventuellement gaussiens, sont corrects. Nous utiliserons donc cette méthodologie pour tirer des conclusions.

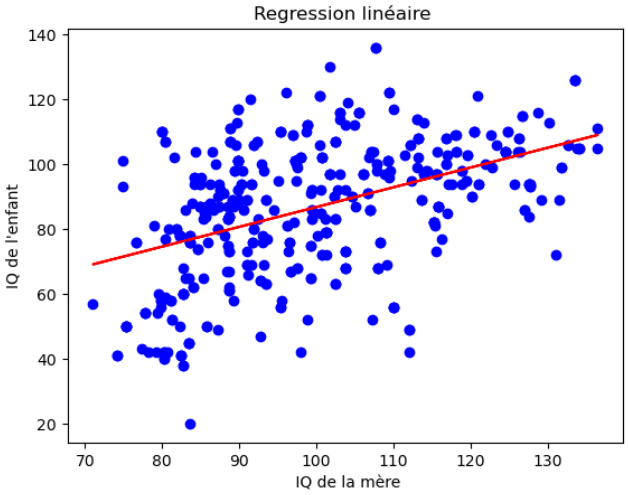


FIGURE 1 – Nuage des points des variables QI de l'enfant et QI de sa mère

3 Applications

Dans la partie précédente, nous avons parlé de la méthodologie du bootstrap ainsi que de ses principales variantes appliquées aux modèles linéaires. Après avoir abordé la robustesse et l'efficacité de cette méthode, nous allons maintenant, à travers des exemples concrets sous Python, explorer ce que fait le bootstrap, les cas où il est fiable et les cas où il est moins efficace.

3.1 Le boostrap un véritable couteau suisse

Le bootstrap est une méthode qui a fait ses preuves dans diverses applications. Dans cette partie, nous allons illustrer la performance de cette méthode sur des jeux de données réels. La plupart de nos illustrations se baseront sur le jeu de données "kidiq", qui contient des informations sur le quotient intellectuel (QI) des enfants ainsi que diverses caractéristiques démographiques et socio-économiques de leur famille.
Le jeu de données "kidiq" comprend les variables suivantes :

- **kid_score** : Le quotient intellectuel (QI) de l'enfant.
- **mom_hs** : Un indicateur binaire indiquant si la mère a obtenu son diplôme d'études secondaires.
- **mom_iq** : Le QI de la mère.
- **mom_work** : Un indicateur binaire indiquant si la mère travaille ou non.
- mom_age : L'âge de la mère.

La Figure 1 présente un nuage de points illustrant le QI des enfants par rapport au QI des mères pour un ensemble de données comprenant $n = 434$ observations. Cette visualisation suggère que les données peuvent être modélisées par une régression linéaire simple, décrite par l'équation :

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n,$$

où les ε_j sont non corrélés, avec des moyennes nulles et des variances constantes. Cette

constance de la variance, ou homoscédasticité, est approximativement correcte pour les données de l'exemple.

3.1.1 Application de bootstrap des résidus

Pour illustrer le boostrap des résidus nous allons utiliser le jeu de données Le jeu de données **mammals** inclus avec le package robustebase contient des mesures de la masse corporelle (body , en kg) et de la masse cérébrale (brain , en g) pour 65 espèces animales.
La relation entre ces deux grandeurs est visible sur le graphique suivant en log-log.

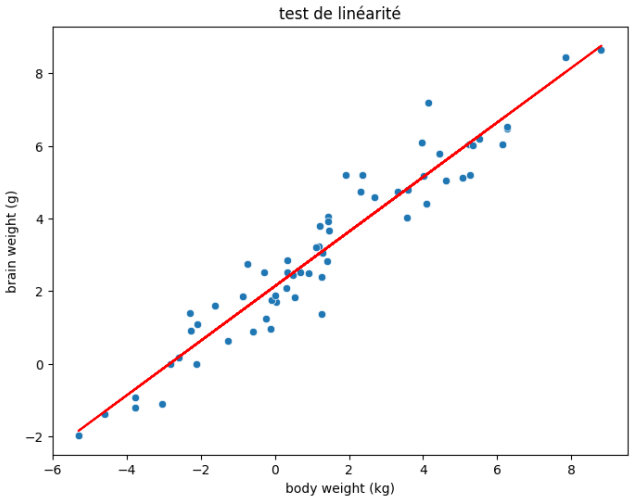


FIGURE 2 – Nuage des points des log des variables body et brain

La figure 2 présente un nuage de points illustrant la masse corporelle (log(body)) par rapport à la masse cérébral(log(brain)) pour un ensemble de données comprenant n=65 observations. Cette visualisation donne une idée à supposer l'existence d'une relation entre ces deux variables dont l'on peut modéliser par une régression linéaire simple décrite par l'équation :

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n,$$

Une régression linéaire basée sur l'ensemble des données donne une estimation $\beta_0 = 2,75$ et $\beta_1 = 0,75$.
Maintenant avant d'appliquer le bootstrap des résidus, nous allons vérifier les hypothèses de validité c'est à dire normalité des résidus et l'homoscédasticité. Comme illustrer à la partie **2.3.3**, nous utiliserons les résidus modifiés.
L'analyse standard suggère que les erreurs sont approximativement normales, bien qu'il y ait un léger soupçon d'hétéroscédasticité.
Donc on peut suppose les hypothèses valides pour l'application du bootstrap des résidus.

En appliquant l'algorithme de la section 2.3.2 avec R=1000 échantillons bootstrap, les estimations des écart types d'erreurs de β_0 et β_1 donne respectivement 0.095 et 0.027 comparé aux valeurs théoriques 0.096 et 0.028. Pour les intervalles de confiances bootstrap, l'échantillon bootstrap du paramètre β_1 permet d'obtenir par calcul $IC_{boost}(\beta_1) = [0.69, 0.80]$ comparé à

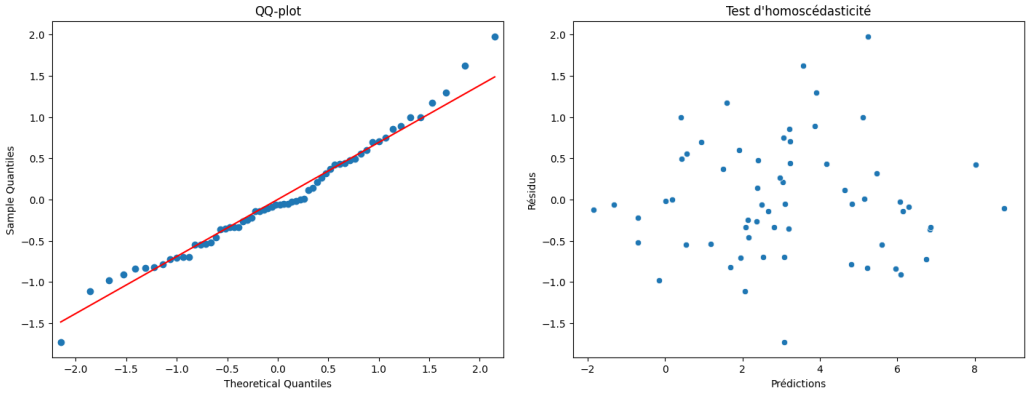


FIGURE 3 – Résidus modifiés en fonction des valeurs prédites et Q-Q plot des résidus modifiés

l'intervalle de confiance théorique obtenu par régression linéaire, $IC_{th}(\beta_1) = [0.69, 0.80]$.Les deux intervalles de confiances sont identiques. Mais de manière générale l'intervalle de confiance bootstrap est légèrement plus étroite que l'intervalle de confiance théorique si tous les hypothèses sont valides et cela dépend aussi du nombres d'échantillon bootstrap générés. En effet des expériences avec d'autres jeux de données simulés nous montre que, plus on augmente le nombre d'échantillons plus bootstrap est meilleur.

Un autre exemple qu'on peut utiliser c'est dans le cas ou l'hétéroscédascité peut être modifié. Par exemple si on considère deux échantillons x et y avec x un échantillon de taille 100 uniformément distribué sur l'intervalle $[0, 10]$.

La relation de variance des résidus en fonction de la covariable pour modéliser l'hétéroscédascité est définie par la fonction $f(x) = ax^2 + b$, avec a et b des réels.
Ensuite on génère des résidus gaussiens centrés en utilisant la relation de variance. Les erreurs ε_i suivent une distribution $\mathcal{N}(0, f(x_i))$. Et enfin notre variable dépendante $y = x + \varepsilon$

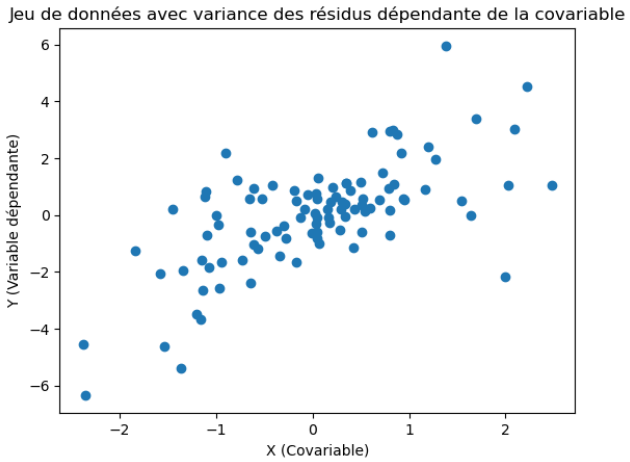


FIGURE 4 – Nuage des points de la variable y en fonction de x

Les nuages de point de la figure 4 donnent idée à modéliser la relation entre x et y par l'équation

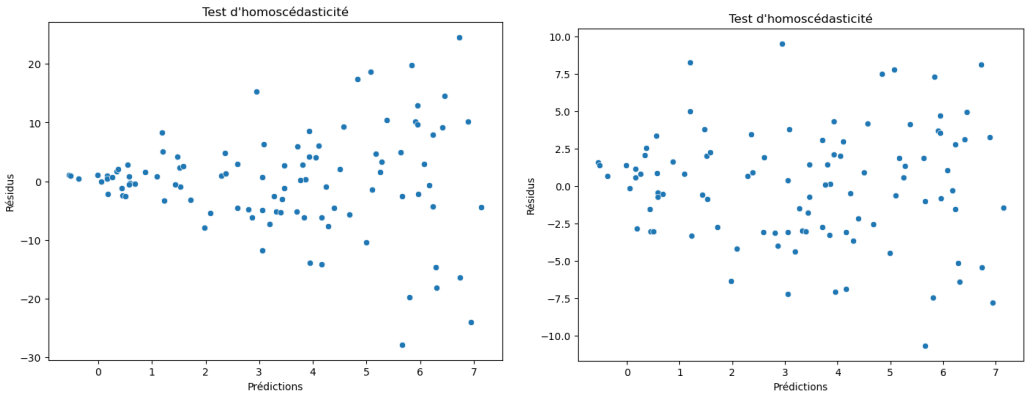


FIGURE 5 – Résidus et les résidus modifiés en fonction des valeurs prédites

suivante :

$$y = \beta_0 + \beta_1 x$$

En observant la figure 5 de dispersion des résidus en fonction des valeurs prédits, on voit nettement que les résidus avant modification ne vérifient pas l'hypothèse d'homoscédasticité. Cependant après modifications des résidus en utilisant les r_j de 3 on voit que l'homoscédasticité peut être validé même s'il y a une légère soupçon. Maintenant on peut utiliser les résultats obtenus par le bootstrap des résidus en utilisant l'algorithme dans la partie 2.3.2 si on fixe $a = 2$ et $b = 0.5$ qui donnent des estimations moyenne de $\hat{\beta}_0$ et $\hat{\beta}_1$ respectivement -0.05 et 1.35. On peut toute fois jouer sur les valeurs de a et b pour comparer les estimations bootstrap et celles de moindres carrés ordinaires.

3.1.2 Bootstrap paramétrique

Pour illustrer l'application du bootstrap paramétrique dans la régression linéaire simple , nous avons utilisé le jeu de données "Kidiq". Notre objectif était d'expliquer la variable "kid_score" (le quotient intellectuel) en fonction de la variable "mom_age" (l'âge de la mère).
Avant d'appliquer la méthode bootstrap paramétrique, il est important de vérifier l'hypothèse de normalité des erreurs. La figure 6, représentant l'histogramme des résidus, sous-entend que cette hypothèse est valide.

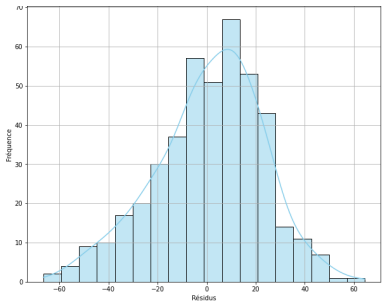


FIGURE 6 – Histogramme des résidus normalisées

En faisant une régression linéaire et en appliquant la méthode des moindres carrés sur notre jeu de données , on obtient $\hat{\beta}_0 = 70.95$ et $\hat{\beta}_1 = 0.69$.

Paramètre	Intervalle de confiance
β_0	[54.630660 : 87.283182]
β_1	[-0.016356 : 1.406729]

TABLE 1 – Intervalles de confiance théoriques de β_0 et β_1

En utilisant la méthode `conf_int()` de python, on obtient les intervalles de confiances théoriques affichés dans le tableau 1 .
Pour implémenter la méthode de bootstrap, nous nous sommes servis de l'algorithme décrit dans la section 2.3.1.
Nous avons généré des échantillons bootstrap des paramètres β_0 et β_1. Ensuite, nous avons calculé les bornes inférieures et supérieures des intervalles de confiance par percentile au niveau de 95% pour les deux paramètres, et nous avons obtenu les résultats affichés dans le tableau 2.

Paramètre	Intervalle de confiance
β_0	[52.47 : 86.25]
β_1	[0.02 : 1.52]

TABLE 2 – Intervalles de confiance bootstrap de β_0 et β_1

On observe que les intervalles de confiance obtenus après avoir effectué le bootstrap sont plus étroits que ceux de notre modèle original.

3.1.3 Application du bootstrap par paires

Nous avons utilisé le jeu de données "kidiq" introduit au début de la section pour illustrer la méthode bootstrap par paires. Sur cet exemple, nous avons estimé la distribution de l'estimateur $\hat{\beta}_1$ du paramètre β_1 de la régression entre le quotient intellectuel des enfants et celui de leur mère, afin de calculer des intervalles de confiance et estimer des écarts-types.
En utilisant la méthode des moindres carrés, nous avons estimé $\hat{\beta}_1 = 0.61$ avec un écart type de 0.059, et un intervalle de confiance théorique $IC_{th}(\beta_1) = [0.49, 0.72]$.
Ensuite, en appliquant la méthode bootstrap par paires en simulant $R = 100$ échantillons bootstrap, nous avons obtenu une estimation moyenne de $\hat{\beta}_1^*$ égale à 0.60 avec un écart type de 0.058.
En augmentant le nombre d'échantillons bootstrap à $R = 1000$, nous avons obtenu une moyenne de $\hat{\beta}_1^*$ égale à 0.61 un écart type de 0.055. En ce qui concerne les intervalles de confiance bootstrap percentile, l'intervalle obtenu pour $R = 100$ est $IC_{boots}(\beta_1) = [0.52, 0.70]$, légèrement plus étroit que l'intervalle théorique. Cela suggère une précision accrue avec la méthode bootstrap. Le tableau 3 résume les valeurs des estimateurs de β_1 ainsi que les écarts-types avec $R = 100$, $R = 300$ et $R = 1000$.
On constate que plus on augmente le nombre d'échantillons boostrap, plus la méthode et précise.
Ces résultats montrent à quel point la méthode Bootstrap peut être aussi efficace, voire plus efficace que la méthode théorique, pour estimer les intervalles de confiance des coefficients

Méthode	$\hat{\beta}_1$	**Écart-type**
Théorique	0.61	0.059
Bootstrap (R = 100)	0.60	0.058
Bootstrap (R = 300)	0.60	0.054
Bootstrap (R = 1000)	0.61	0.055

TABLE 3 – Valeurs des estimations de β_1 et des écarts-types pour différentes méthodes.

d'une régression linéaire. C'est donc une bonne alternative pour éviter les calculs théoriques et constitue une solution dans les cas où l'application de la théorie n'est pas possible.

3.2 Les limites du Bootsrap

Un cas où le Bootstrap n'est pas efficace est celui de la régression multiple, lorsque le nombre de covariables p est assez proche du nombre d'observations n. Le rééchantillonnage dans ce cas peut induire une quasi-collinéarité dans la matrice de conception X^*, ou de manière équivalente une quasi-singularité dans $X^{*T}X^*$, et donc produire des estimations bootstrap non correctes. Un exemple qui illustre un cas presque similaires à celui a été expliquer dans [2].

Pour illustrer ce phénomène, Nous avons simulé par machine une matrice aléatoire X de taille $n \times (p+1)$ et un vecteur Y de taille n tel que :

$$Y = X\beta + \epsilon$$

avec β un vecteur de taille p de coefficients fixés et ϵ un vecteur aléatoire de taille n.
Nous avons choisis $n = 150$ et avons fait varier p dans la grille $[60, 100, 140]$ et à chaque p fixé, nous avons calculé et estimé le coefficient β_1 par la méthode bootstrap par paires.
En suite, nous avons calculer les intervalles de confiance bootsrap pour chaque p et évalué la précision de l'intervalle bootstrap avec la méthodologie décrite dans la sous-section 2.4 .
Nous avons simulés $N = 1000$ échantillons pour le calcul de l'intervalle de confiance par percentile.
Les résultats montrent que plus p augmente, moins la méthode est précise. Pour $p = 60$, pour des intervalles de confiance à 55%, le paramètre tombe dans environ 68% des jeux de données. C'est à dire que l'intervalle couvre plus que ce qu'il ne faudrait.
Pour $p = 100$, pour des intervalles de confiance à 55%, le paramètre tombe dans environ 38% des jeux de données.
Pour $p = 140$, pour des intervalles de confiance à 55%, β_1 ne tombe presque jamais dans les intervalles de confiances bootstrap (environ 0% des jeux de données).
Le tableau 4 résume les résultats obtenus. On observe que les résultats que nous souhaitons démontrer deviennent plus claires quand on augmente le nombre de jeux de données. Cependant, lors de la vérification de l'intervalle de confiance, nous nous sommes limités à un nombre de jeux de données $N = 1000$ pour éviter des temps de calcul assez longs. En effet, pour $N = 10$, le processus a pris 4.6 secondes, tandis que pour $N = 1000$, il a nécessité 4600 secondes, soit environ une heure et quart.

Nombre de covariables (p)	**Précision de l'intervalle de confiance bootstrap (55%)**
60	68
100	38
140	0

TABLE 4 – Précision de l'intervalle de confiance Bootstrap à 55% pour différents nombres de covariables (p).

4 Conclusion

Dans notre travail, nous avons exploré en profondeur la méthodologie du bootstrap et ses différentes variantes appliquées à la régression linéaire. Nous avons mis en évidence les principes sous-jacents de chaque méthode, discuté de leurs avantages et évalué leurs limitations.
Nous avons constaté que le bootstrap paramétrique, dans le contexte de la régression linéaire, offre une approche robuste pour estimer la distribution des coefficients du modèle. Cette méthode permet d'évaluer l'incertitude associée à l'estimation des paramètres, enrichissant ainsi notre compréhension des modèles linéaires.
Ensuite, nous avons appliqué ces concepts à travers des exemples concrets. Nous avons illustré comment le bootstrap peut être utilisé pour estimer les intervalles de confiance des coefficients et évaluer la précision des paramètres dans des modèles linéaires réels. Ces exemples ont souligné l'efficacité du bootstrap comme un outil puissant et polyvalent pour l'analyse des données en régression linéaire.
Cependant, nos résultats ont également révélé que le bootstrap n'est pas toujours efficace, notamment dans des situations complexes. Par exemple, dans la régression multiple avec un nombre élevé de covariables par rapport au nombre d'observations, ou lorsque les erreurs du modèle sont hétéroscédastiques et que l'hétéroscédasticité ne peut pas être correctement modélisée.
En guise de perspectives futures, il convient de noter que des solutions existent pour surmonter ces défis. Par exemple, des méthodes de régularisation ou des techniques spécifiques de modélisation de l'hétéroscédasticité peuvent être explorées pour améliorer les performances du bootstrap dans des scénarios plus complexes. Bien que ces aspects n'aient pas été traités dans ce travail, ils représentent des pistes pour améliorer le bootstrap en régression linéaire.

Références

[1] Pierre Dagnelie. *Statistique théorique et appliquée : Statistique descriptive et bases de l'inférence statistique*, volume 1. De Boeck Supérieur, 1998.

[2] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

[3] Bradley Efron and Trevor Hastie. *Computer age statistical inference, student edition : algorithms, evidence, and data science*, volume 6. Cambridge University Press, 2021.

[4] Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1) :1–436, 1993.