Inférence Causale : Impact de l'Activité Physique sur la Santé

Variables de contrôle, confondeurs et biais de surcontrôle

Kiyali Coulibaly
Économiste de la santé | Data Scientist
fas.coul@yahoo.fr
linkedin.com/in/kiyali-coulibaly
github.com/Faskos92

2025-07-31

Contents

List of Figures

List of Tables

1 Introduction

1.1 Contexte et enjeux

En santé publique, l'établissement de relations causales entre les comportements de santé (comme l'activité physique) et les résultats sanitaires représente un défi méthodologique majeur. Les données observationnelles, contrairement aux essais contrôlés randomisés, sont susceptibles de contenir plusieurs sources de biais qui peuvent fausser les conclusions :

- **Biais de confusion** : Variables qui influencent simultanément l'exposition et le résultat
- **Biais de médiation** : Contrôle inapproprié de variables intermédiaires
- **Biais de surcontrôle** : Inclusion de variables post-traitement ou de collideurs

Ce document présente une approche méthodologique rigoureuse pour identifier l'effet causal de l'activité physique sur la santé auto-déclarée à partir de données simulées.

1.2 Objectifs de l'étude

L'objectif principal est de démontrer l'importance cruciale d'un choix rigoureux des variables de contrôle dans les modèles statistiques pour obtenir des estimations causales fiables. Plus spécifiquement, nous cherchons à :

1. **Illustrer les différents types de biais** causés par un mauvais choix de variables de contrôle
2. **Comparer les estimations** obtenues avec différentes spécifications de modèles
3. **Proposer une méthode robuste** pour l'identification causale en données observationnelles

Question de recherche centrale : Quel est l'effet causal de l'activité physique sur le score de santé auto-déclarée ?

2 Cadre théorique : Hypothèse d'identification causale

Notre approche repose sur l'**hypothèse d'ignorabilité conditionnelle** (Rubin, 1974; Rosenbaum & Rubin, 1983) :

$$Y(x) \perp X \mid C$$

où :
- $Y(x)$ représente le résultat potentiel pour un niveau d'exposition x
- X est le niveau d'activité physique (variable continue)
- C est l'ensemble des variables de contrôle

Cette hypothèse stipule que, conditionnellement aux variables de contrôle C, l'exposition X est indépendante des résultats potentiels $Y(x)$. Bien que forte et non directement testable, elle permet d'interpréter les coefficients estimés comme des effets causaux.

2.1 Conditions d'identification

Pour une identification causale robuste, plusieurs conditions doivent être satisfaites :

1. **Absence de confondeurs non mesurés** : Toutes les causes communes de X et Y sont incluses dans C
2. **Spécification fonctionnelle correcte** : Le modèle capture les non-linéarités et interactions pertinentes

3. **Validité de la structure causale** : Le diagramme causal (DAG) représente fidèlement la réalité
4. **Absence de causalité inverse** : Le score de santé n'influence pas rétroactivement l'activité physique

Note méthodologique : Ces hypothèses doivent être évaluées dans chaque contexte spécifique pour garantir la validité des inférences causales.

3 Cadre conceptuel : Diagrammes causaux (DAGs)

Les **Diagrammes Acycliques Dirigés (DAGs)** constituent un outil fondamental pour visualiser les relations causales et guider le choix des variables de contrôle. Chaque DAG illustre une configuration causale spécifique pour l'estimation de l'effet de l'activité physique (X) sur le score de santé (Y).

3.1 Relations causales fondamentales

3.1.1 Configuration 1 : Confusion simple

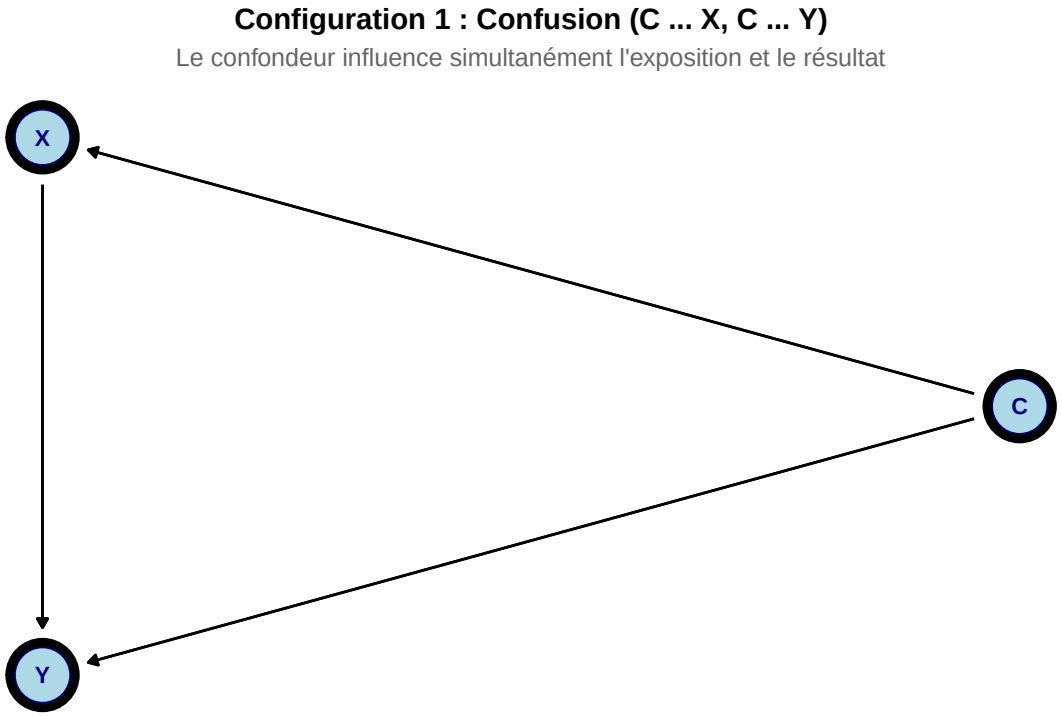

Figure 1: Structure causale avec confondeur simple

Interprétation : Dans cette configuration, le confondeur C (ex: âge, statut socio-économique) influence à la fois le niveau d'activité physique et le score de santé. Sans contrôle de C, l'association observée entre X et Y sera biaisée car elle inclut le chemin non-causal $X \leftarrow C \rightarrow Y$.

Stratégie d'identification : Contrôler C bloque le chemin de confusion et permet d'isoler l'effet causal direct $X \rightarrow Y$.

3.1.2 Configuration 2 : Médiation

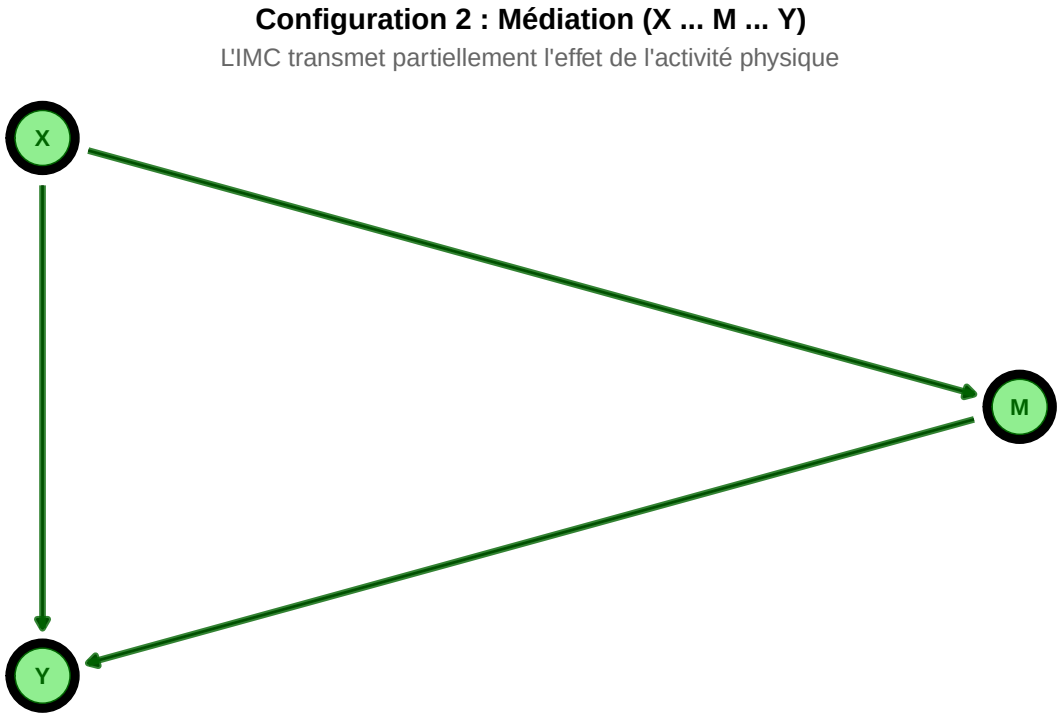

Figure 2: Structure causale avec médiateur

Interprétation : L'activité physique (X) influence l'IMC (M), qui à son tour affecte le score de santé (Y). L'IMC est donc un **médiateur** qui transmet une partie de l'effet causal.

Implications pour l'estimation :
- **Effet total** : $X \to Y$ (direct) + $X \to M \to Y$ (indirect)
- **Effet direct** : $X \to Y$ (en contrôlant M)
- **Erreur courante** : Contrôler M pour estimer l'effet total sous-estime l'effet réel

3.1.3 Configuration 3 : Surcontrôle

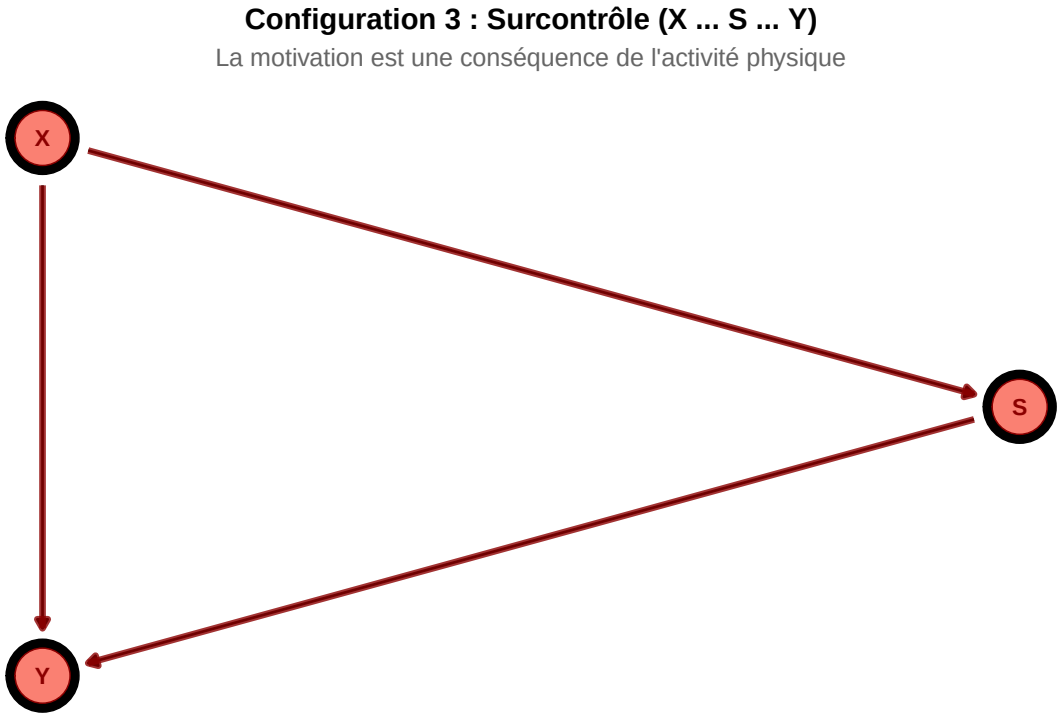

Figure 3: Structure causale avec variable de surcontrôle

Interprétation : La motivation santé (S) est influencée par l'activité physique (X) et affecte le score de santé (Y). S est une **variable post-traitement**.

Problème du surcontrôle : Contrôler S bloque le chemin causal $X \to S \to Y$, conduisant à une sous-estimation de l'effet total de X sur Y.

3.1.4 Modèle intégré : Configurations multiples

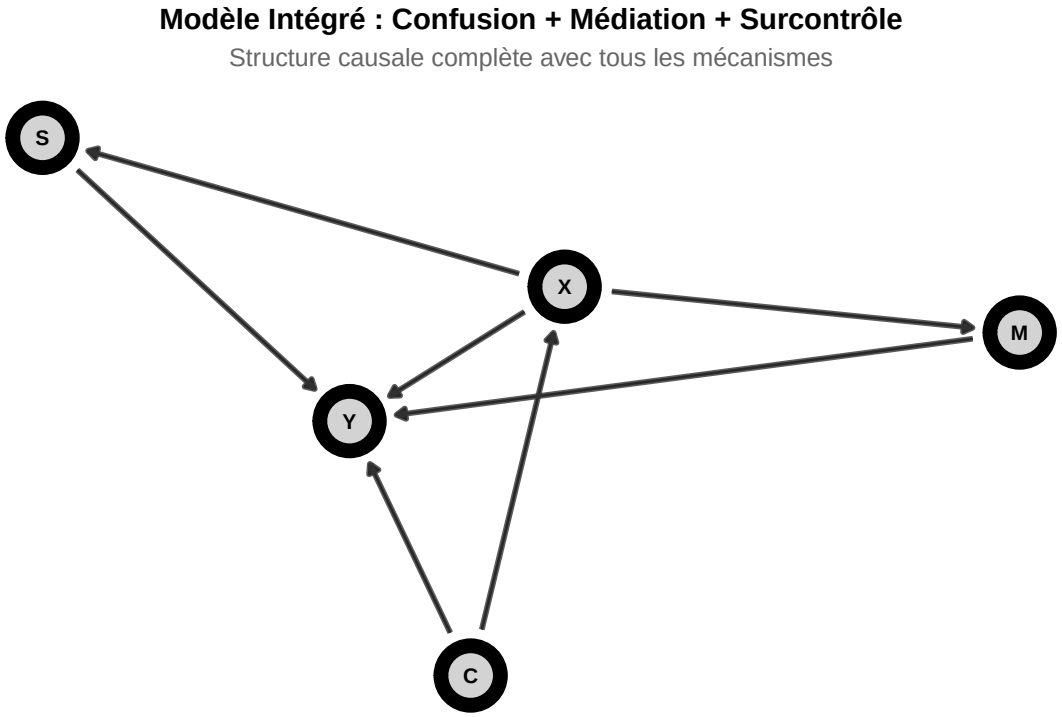

Figure 4: Structure causale complète

Synthèse des mécanismes : Ce modèle intégré combine :
- **Confusion** via C
- **Médiation** via M (IMC)
- **Surcontrôle** via S (motivation)

Stratégie d'identification optimale : Pour estimer l'effet causal total de X sur Y, seuls les confondeurs (C) doivent être contrôlés. Les médiateurs (M) et variables post-traitement (S) ne doivent pas être inclus.

4 Données et méthodologie

4.1 Présentation des données simulées

Les données utilisées proviennent d'une simulation réaliste conçue pour reproduire les caractéristiques d'une enquête de santé populationnelle. Cette approche permet de connaître la "vérité terrain" et d'évaluer la performance des différentes méthodes d'estimation.

```
##
## Variables de la base de données :

##  [1] "ScoreSante"          "ActivitePhysique"    "Sexe"
##  [4] "Tabagisme"           "Age"                 "Education"
##  [7] "RevenuFamilial"      "IMC"                 "MotivationSante"
## [10] "Region"              "Milieu"              "EtatSanteChronique"
## [13] "ConsommationAlcool"  "StressPsychologique" "SoutienSocial"
## [16] "QualiteSommeil"      "OccupationPhysique"  "AccesSoins"
## [19] "Alimentation"
```

```
##
## Structure des données :

## Rows: 10,000
## Columns: 19
## $ ScoreSante          <dbl> 79.4, 88.1, 100.0, 83.6, 89.0, 93.6, 91.1, 100.0, ~
## $ ActivitePhysique    <dbl> 8.1, 4.2, 9.7, 6.1, 6.3, 5.9, 9.1, 8.9, 3.2, 9.6, ~
## $ Sexe                <fct> Femme, Femme, Femme, Femme, Femme, Femme, Femme, F~
## $ Tabagisme           <fct> Non-fumeur, Non-fumeur, Fumeur, Non-fumeur, Non-fu~
## $ Age                 <dbl> 38, 50, 35, 36, 32, 49, 43, 51, 33, 34, 22, 28, 39~
## $ Education           <fct> Secondaire, Primaire, Superieur, Superieur, Second~
## $ RevenuFamilial      <dbl> 33.6, 29.9, 57.0, 43.0, 50.4, 33.5, 31.2, 67.2, 24~
## $ IMC                 <dbl> 18.7, 23.1, 23.5, 20.7, 20.4, 20.1, 21.4, 24.5, 25~
## $ MotivationSante     <dbl> 9.2, 7.4, 8.4, 6.2, 8.5, 6.8, 9.9, 9.3, 3.9, 8.5, ~
## $ Region              <fct> regionA, regionE, regionA, regionA, regionB, regio~
## $ Milieu              <fct> Urbain, Urbain, Urbain, Urbain, Urbain, Urbain, Ru~
## $ EtatSanteChronique  <int> 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1,~
## $ ConsommationAlcool  <fct> Buveur, Buveur, Buveur, Buveur, Buveur, Non-buveur~
## $ StressPsychologique <dbl> 6.1, 6.2, 2.7, 7.5, 6.2, 6.0, 5.1, 4.0, 6.5, 5.7, ~
## $ SoutienSocial       <dbl> 3.5, 3.2, 3.6, 5.7, 5.2, 5.1, 7.1, 3.0, 6.9, 6.8, ~
## $ QualiteSommeil      <dbl> 2.9, 3.7, 6.1, 2.6, 5.7, 3.0, 5.8, 5.7, 2.3, 4.5, ~
## $ OccupationPhysique  <dbl> 3.1, 1.2, 1.6, 2.9, 3.0, 4.2, 3.8, 3.0, 5.9, 5.0, ~
## $ AccesSoins          <dbl> 6.4, 6.2, 6.1, 4.8, 6.2, 5.5, 5.3, 6.2, 5.2, 7.1, ~
## $ Alimentation        <dbl> 4.0, 5.5, 9.2, 7.1, 5.0, 7.5, 5.5, 8.1, 4.1, 6.8, ~
```

4.2 Analyse descriptive

4.2.1 Statistiques univariées

Tableau 1: Statistiques descriptives univariées (N = 10000)

Characteristic	**N = 10,000**
ScoreSante	79.79 ± 17.03 [0.00 ; 100.00]
ActivitePhysique	6.50 ± 2.50 [0.00 ; 10.00]
Sexe	
Femme	5,391.0 (53.9%)
Homme	4,609.0 (46.1%)
Tabagisme	
Non-fumeur	8,372.0 (83.7%)
Fumeur	1,628.0 (16.3%)
Age	49.01 ± 14.18 [18.00 ; 80.00]
Education	
Secondaire	4,697.0 (47.0%)
Superieur	2,856.0 (28.6%)
Primaire	2,447.0 (24.5%)
RevenuFamilial	42.57 ± 12.60 [15.00 ; 88.40]
IMC	22.90 ± 3.70 [16.00 ; 35.90]
MotivationSante	6.86 ± 1.87 [0.00 ; 10.00]
Region	
regionB	2,054.0 (20.5%)
regionD	1,996.0 (20.0%)
regionE	1,996.0 (20.0%)
regionA	1,988.0 (19.9%)
regionC	1,966.0 (19.7%)

Tableau 1: Statistiques descriptives univariées (N = 10000) *(continued)*

Characteristic	N = 10,000
Milieu	
Urbain	7,011.0 (70.1%)
Rural	2,989.0 (29.9%)
EtatSanteChronique	7,672.0 (76.7%)
ConsommationAlcool	
Non-buveur	6,126.0 (61.3%)
Buveur	3,874.0 (38.7%)
StressPsychologique	5.27 ± 1.65 [0.00 ; 10.00]
SoutienSocial	4.83 ± 1.25 [0.40 ; 9.70]
QualiteSommeil	3.99 ± 2.04 [0.00 ; 10.00]
OccupationPhysique	3.47 ± 1.52 [0.00 ; 8.80]
AccesSoins	5.94 ± 1.03 [2.30 ; 9.90]
Alimentation	5.98 ± 1.37 [1.30 ; 10.00]

[1] Mean ± SD [Min ; Max]; n (%)

Observations :
- L'échantillon comprend 10000 individus avec un âge moyen de 49 ans
- Le score de santé moyen est de 79.8 points (écart-type : 17)
- Le niveau d'activité physique moyen est de 6.5 sur une échelle de 0 à 10

4.2.2 Analyse des relations bivariées

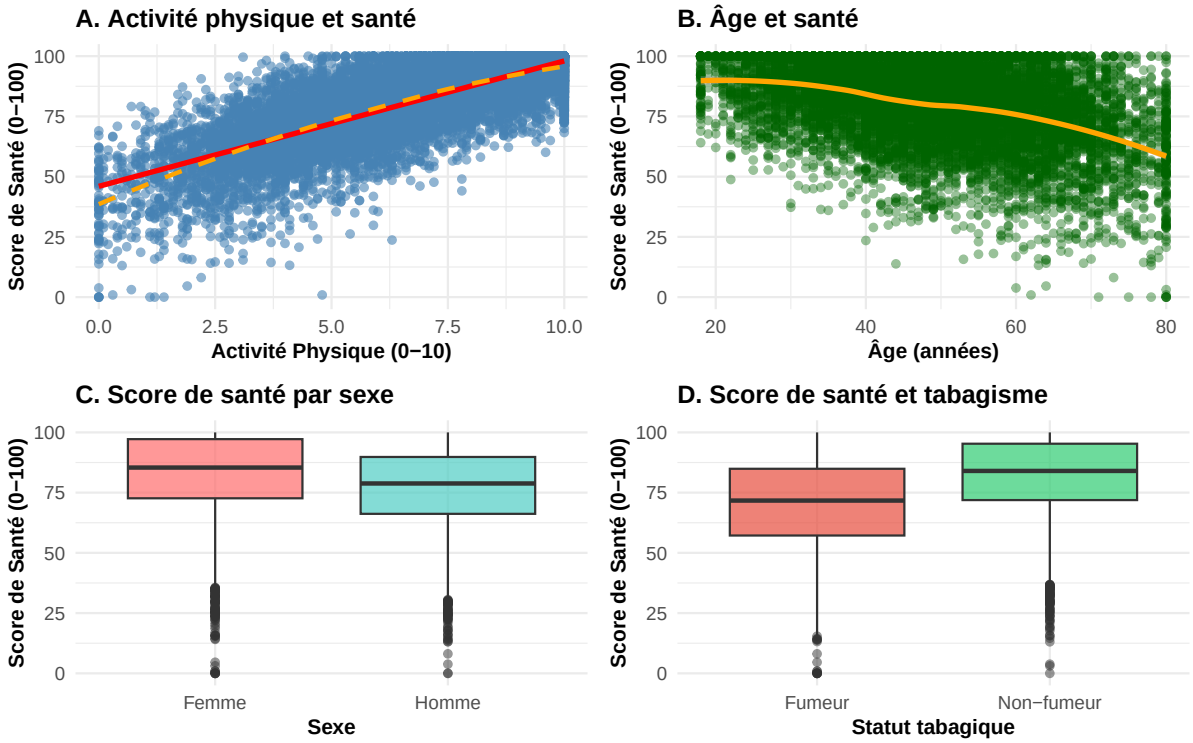

Figure 5: Relations principales entre variables explicatives et score de santé

Observations :
- **Relation positive** entre activité physique et score de santé (r = 0.763)
- **Effet non-linéaire de l'âge** : diminution du score de santé avec l'âge, plus marquée après 50 ans

- **Différences par sexe** : les femmes présentent un score médian légèrement supérieur
- **Impact du tabagisme** : écart moyen de 12.3 points entre non-fumeurs et fumeurs

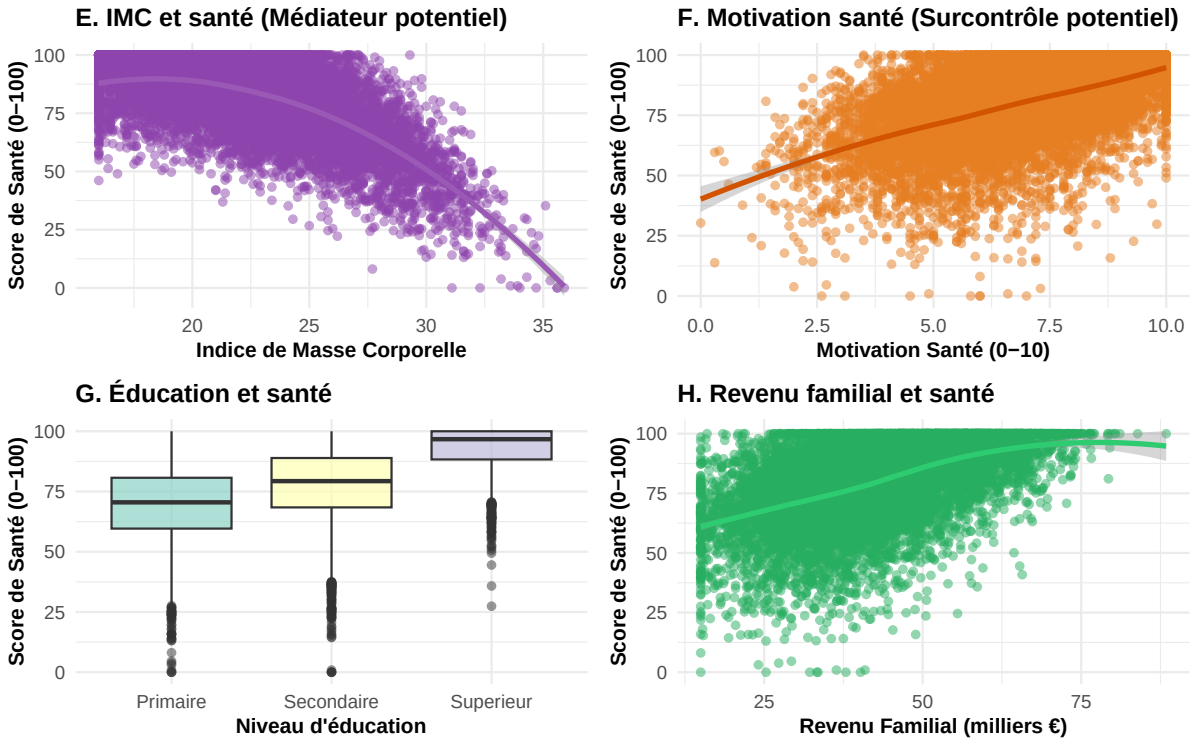

Figure 6: Relations secondaires et variables potentiellement problématiques

Observations :
- IMC : relation négative non linéaire (médiateur potentiel).
- Motivation santé : forte corrélation positive (risque de surcontrôle).
- Éducation : gradient positif.
- Revenu : association positive avec effet de plateau.

4.2.3 Analyse descriptive comparative et tests statistiques

Tableau 2: Score moyen de santé par modalité des variables catégorielles

Variable	Modalité	Effectif (n)	Score moyen	p-value
Sexe	Femme	5391	82.28	<0.001
	Homme	4609	76.87	
Tabagisme	Fumeur	1628	69.50	<0.001
	Non-fumeur	8372	81.79	
Education	Primaire	2447	69.17	<0.001
	Secondaire	4697	77.36	
	Superieur	2856	92.87	
Region	regionA	1988	76.54	<0.001
	regionB	2054	75.82	
	regionC	1966	76.56	
	regionD	1996	82.87	
	regionE	1996	87.19	
Milieu	Rural	2989	76.20	<0.001
	Urbain	7011	81.31	
ConsommationAlcool	Buveur	3874	77.89	<0.001

Non-buveur	6126	80.98

Les effectifs et les scores moyens de santé ont été présentés selon les modalités des variables catégorielles étudiées. Les tests statistiques appropriés ont été appliqués pour évaluer les différences entre groupes : test de Mann-Whitney pour les variables à deux modalités et test de Kruskal-Wallis pour celles à plus de deux modalités.
La dernière colonne du tableau présente les p-values issues des tests statistiques qui permettent d'évaluer la différence du score de santé entre les groupes.

4.2.4 Analyse des corrélations

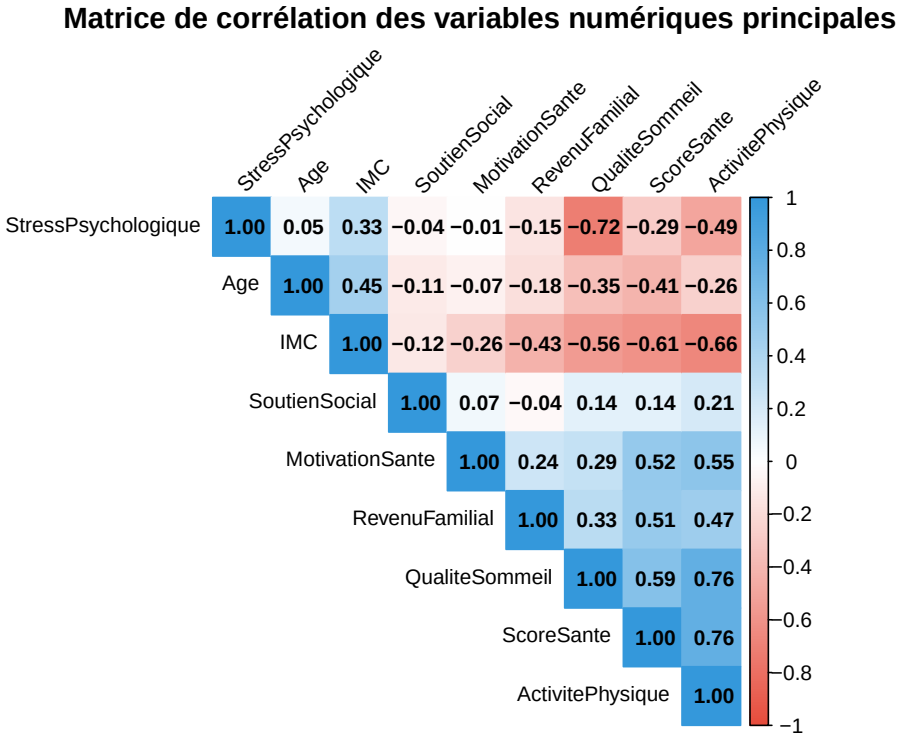

Figure 7: Matrice de corrélation des variables numériques principales

Tableau 3: Corrélations clés pour l'analyse causale

Paire.de.variables	Corrélation
Activité Physique Score Santé	0.763
Activité Physique IMC	-0.661
Activité Physique Motivation Santé	0.552
Motivation Santé Score Santé	0.520
IMC Score Santé	-0.608
Âge Score Santé	-0.413

Implications :
- Corrélation activité physique-motivation santé (r = 0.552).
- Corrélation négative activité physique-IMC (r = -0.661).

5 Analyse multivariée : Comparaison des stratégies d'estimation

5.1 Stratégie d'analyse

Nous comparons cinq modèles :
1. **Naïf** : Sans contrôles (biaisé).
2. **Contrôles légitimes** : Confondeurs seulement.
3. **Surcontrôle** : Inclusion de la motivation santé (biaisé).
4. **Médiation** : Contrôle de l'IMC (biaisé).
5. **Complet** : Contrôles légitimes + interactions.

5.2 Modèle 1 : Régression Naïve (BIAISÉE)

L'objectif est d'illustrer le biais de variables omises en estimant un modèle sans contrôles.

Tableau 4: Effet de l'activité physique (modèle naïf)

Termes	Estimate	IC 95% inf	IC 95% sup	p-value
Constante	45.983	45.382	46.584	<0.001
Activité physique	5.204	5.117	5.290	<0.001
Effectif (n)	10000.000	NA	NA	
R^2	0.583	NA	NA	

Commentaires : Une augmentation d'une unité d'activité physique augmente le score de santé de 5.204 points, mais l'effet est biaisé par l'absence de contrôle des confondeurs.
Conséquence attendue: L'effet de l'activité physique est probablement exagéré parce que ces variables oubliées sont liées à la fois à l'activité physique et au score de santé.

5.3 Modèle 2 : Contrôles Légitimes + effets fixes

L'objectif est d'estimer l'effet causal en contrôlant les confondeurs légitimes et les effets fixes.

Tableau 5: Résultats de la régression ajustée (modèle 2)

Variable	Estimate	IC 95% inf	IC 95% sup	p-value
Activité physique	3.974	3.854	4.093	<0.001
Sexe (Homme)	-4.755	-5.087	-4.423	<0.001
TabagismeNon-fumeur	3.785	3.310	4.260	<0.001
Âge (années)	0.750	0.680	0.820	<0.001
Âge^2	-0.010	-0.011	-0.009	<0.001
Éducation (Secondaire)	4.142	3.701	4.584	<0.001
Éducation (Supérieur)	4.072	3.356	4.789	<0.001
Revenu familial	0.130	0.110	0.150	<0.001
Accès aux soins	0.439	0.281	0.598	<0.001
Qualité alimentation	0.195	0.063	0.326	0.004
État de santé chronique	-3.120	-3.525	-2.716	<0.001
ConsommationAlcoolNon-buveur	2.423	2.097	2.749	<0.001
Stress psychologique	0.380	0.221	0.538	<0.001
Soutien social	-0.277	-0.415	-0.140	<0.001
Qualité du sommeil	0.283	0.119	0.446	<0.001
Occupation physique	0.084	-0.038	0.206	0.177
Région B	-0.949	-1.447	-0.452	<0.001
Région C	-0.552	-1.055	-0.048	0.032

Région D	5.561	5.059	6.063	<0.001
Région E	9.238	8.734	9.741	<0.001
Milieu urbain	1.514	1.045	1.983	<0.001

Commentaire Estimation non biaisé du coéf de l'activité sportive

5.4 Modèle 3 : Surcontrôle (biaisé)

Ce modèle inclut (MotivationSante) la motivation, une variable post-traitement, introduisant un biais de surcontrôle.

Tableau 6: Résultats de la régression avec surcontrôle (modèle 3)

Variable	Estimate	IC 95% inf	IC 95% sup	p-value
Activité physique	3.206	3.057	3.355	<0.001
Sexe (Homme)	-2.627	-3.040	-2.215	<0.001
TabagismeNon-fumeur	4.494	4.018	4.969	<0.001
Âge (années)	0.732	0.663	0.801	<0.001
Âge²	-0.010	-0.011	-0.009	<0.001
Éducation (Secondaire)	4.167	3.732	4.603	<0.001
Éducation (Supérieur)	3.577	2.868	4.286	<0.001
Revenu familial	0.144	0.124	0.164	<0.001
Accès aux soins	0.422	0.266	0.578	<0.001
Qualité alimentation	0.209	0.079	0.339	0.002
Motivation santé	1.234	1.088	1.379	<0.001
État de santé chronique	-3.138	-3.537	-2.739	<0.001
ConsommationAlcoolNon-buveur	2.429	2.107	2.751	<0.001
Stress psychologique	-0.166	-0.335	0.004	0.055
Soutien social	0.057	-0.084	0.199	0.427
Qualité du sommeil	0.273	0.112	0.434	<0.001
Occupation physique	0.091	-0.030	0.211	0.139
Région B	-0.931	-1.421	-0.440	<0.001
Région C	-0.481	-0.978	0.015	0.057
Région D	5.630	5.135	6.125	<0.001
Région E	9.301	8.804	9.798	<0.001
Milieu urbain	2.515	2.037	2.992	<0.001

Commentaire : Le coefficient de l'activité physique (3.206) est sous-estimé en raison du surcontrôle.

5.5 Modèle 4 : Contrôle du médiateur (biaisé)

Ce modèle inclut l'IMC, un médiateur, ce qui biaise l'estimation de l'effet total.

Tableau 7: Résultats de la régression avec médiateur (modèle 4)

Variable	Estimate	IC 95% inf	IC 95% sup	p-value
Activité physique	3.406	3.303	3.509	<0.001
Sexe (Homme)	-5.196	-5.467	-4.925	<0.001
TabagismeNon-fumeur	3.481	3.102	3.861	<0.001
Âge (années)	0.440	0.383	0.496	<0.001
Âge²	-0.007	-0.007	-0.006	<0.001

IMC	12.969	12.615	13.324	<0.001
IMC²	-0.286	-0.294	-0.279	<0.001
Éducation (Secondaire)	3.790	3.438	4.142	<0.001
Éducation (Supérieur)	5.576	5.003	6.148	<0.001
Revenu familial	0.114	0.098	0.131	<0.001
Accès aux soins	0.442	0.316	0.569	<0.001
Qualité alimentation	0.243	0.138	0.349	<0.001
État de santé chronique	-3.595	-3.919	-3.271	<0.001
ConsommationAlcoolNon-buveur	2.529	2.269	2.789	<0.001
Stress psychologique	0.299	0.172	0.425	<0.001
Soutien social	-0.141	-0.250	-0.031	0.012
Qualité du sommeil	0.338	0.208	0.468	<0.001
Occupation physique	0.132	0.035	0.229	0.008
Région B	-1.014	-1.411	-0.618	<0.001
Région C	-0.528	-0.929	-0.127	0.010
Région D	5.486	5.086	5.887	<0.001
Région E	9.136	8.735	9.538	<0.001
Milieu urbain	1.922	1.548	2.296	<0.001

Commentaire : Le coefficient de l'activité physique (3.406) sous-estime l'effet total en raison du contrôle de l'IMC.

5.6 Modèle 5 : Contrôles légitimes avec interactions + effets fixes

Tableau 8 :Résultats de la régression avec interactions (modèle 5)

Variable	Estimate	IC 95% inf	IC 95% sup	p-value
Activité physique	3.780	3.656	3.904	<0.001
Sexe (Homme)	-5.275	-6.087	-4.464	<0.001
TabagismeNon-fumeur	5.804	5.360	6.248	<0.001
Âge (années)	-0.133	-0.213	-0.053	0.001
Âge²	-0.009	-0.009	-0.008	<0.001
Éducation (Secondaire)	4.001	3.597	4.406	<0.001
Éducation (Supérieur)	4.813	4.156	5.470	<0.001
Revenu familial	0.128	0.109	0.146	<0.001
Accès aux soins	0.416	0.271	0.561	<0.001
Qualité alimentation	0.157	0.037	0.278	0.010
État de santé chronique	-3.622	-3.993	-3.251	<0.001
ConsommationAlcoolNon-buveur	2.494	2.196	2.793	<0.001
Stress psychologique	0.346	0.200	0.491	<0.001
Soutien social	-0.277	-0.403	-0.151	<0.001
Qualité du sommeil	0.346	0.196	0.495	<0.001
Occupation physique	0.070	-0.042	0.182	0.221
Région B	-0.846	-1.301	-0.391	<0.001
Région C	-0.590	-1.051	-0.130	0.012
Région D	5.514	5.055	5.973	<0.001
Région E	9.163	8.702	9.624	<0.001
Milieu urbain	1.617	1.188	2.046	<0.001
Interaction : Activité physique × Sexe (H)	0.055	-0.061	0.171	0.355
Interaction : Activité physique × Âge centré	0.047	0.043	0.052	<0.001
TabagismeNon-fumeur:I(Age - mean(Age))	0.507	0.474	0.539	<0.001

Commentaires : Le coefficient de l'activité physique (3.78) estimé est non biaisé et est le vrai effet.

5.7 Comparaison des modèles : résumé des 05 modèles

Tableau 9: Comparaison des coefficients de l'activité physique

	Estimate	SE	t value	P_value	R2_adj
Naïf (biaisé)	5.204	0.044	118.145	<0.001	0.583
Contrôles légitimes	3.974	0.061	65.092	<0.001	0.776
Surcontrôle (biaisé)	3.206	0.076	42.242	<0.001	0.782
Médiation (biaisé)	3.406	0.053	64.599	<0.001	0.858
Contrôles + interactions	3.78	0.063	59.886	<0.001	0.813

Tableau 10 : Effet de l'activité physique sur le score de santé

	Score de Santé				
	Naïf	Contrôles	Surcontrôle	Médiation	Contrôles + interaction
Activité Physique	5.204***	3.974***	3.206***	3.406***	3.780***
	(0.044)	(0.061)	(0.076)	(0.053)	(0.063)
Confondeurs	Non	Oui	Oui	Oui	Oui
Surcontrôle	Non	Non	Oui	Non	Non
Médiateur	Non	Non	Non	Oui	Non
Interactions	Non	Non	Non	Non	Oui

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Effet vrai simulé = 3.974 | 3.780

5.8 visualisation des coéf des 05 modèles et leurs IC (95%)

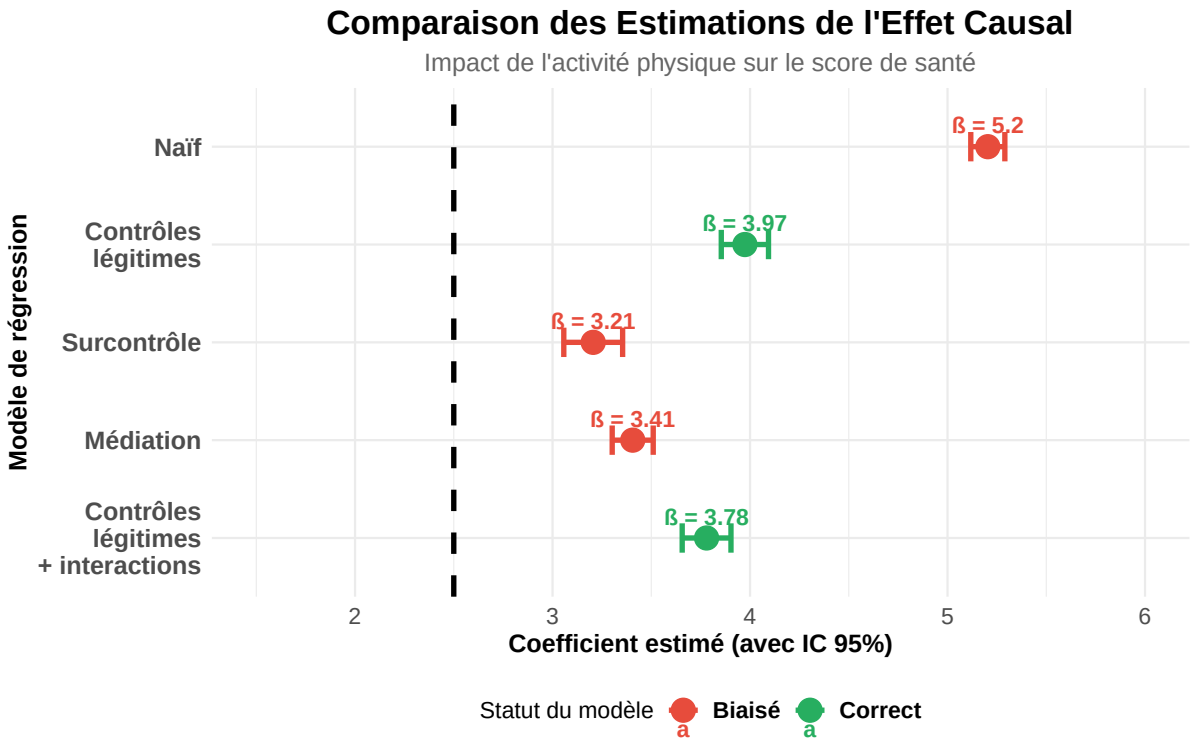

Figure 8: Comparaison des estimations de l'effet causal

Commentaire : Les modèles 2 et 5 sont non biaisés, contrairement aux modèles biaisés (1, 3, 4).

6 Sensibilité à la spécification fonctionnelle

Tableau 11: Tests de robustesse de l'effet de l'activité physique

	Estimate.Estimate	SE.Std. Error	t_value.t value	p_value	R2_adj
Modèle principal (linéaire)	3.974	0.061	65.092	<0.001	0.776
Spécification logarithmique	21.51	0.304	70.729	<0.001	0.788
Forme quadratique	7.718	0.147	52.356	<0.001	0.792

Commentaires :
- Tous les modèles confirment un effet positif significatif.
- Les modèles non linéaires (logarithmique, quadratique) suggèrent un effet décroissant à hauts niveaux d'activité physique.

6.1 Visualisation des effets marginaux des 03 spécifications

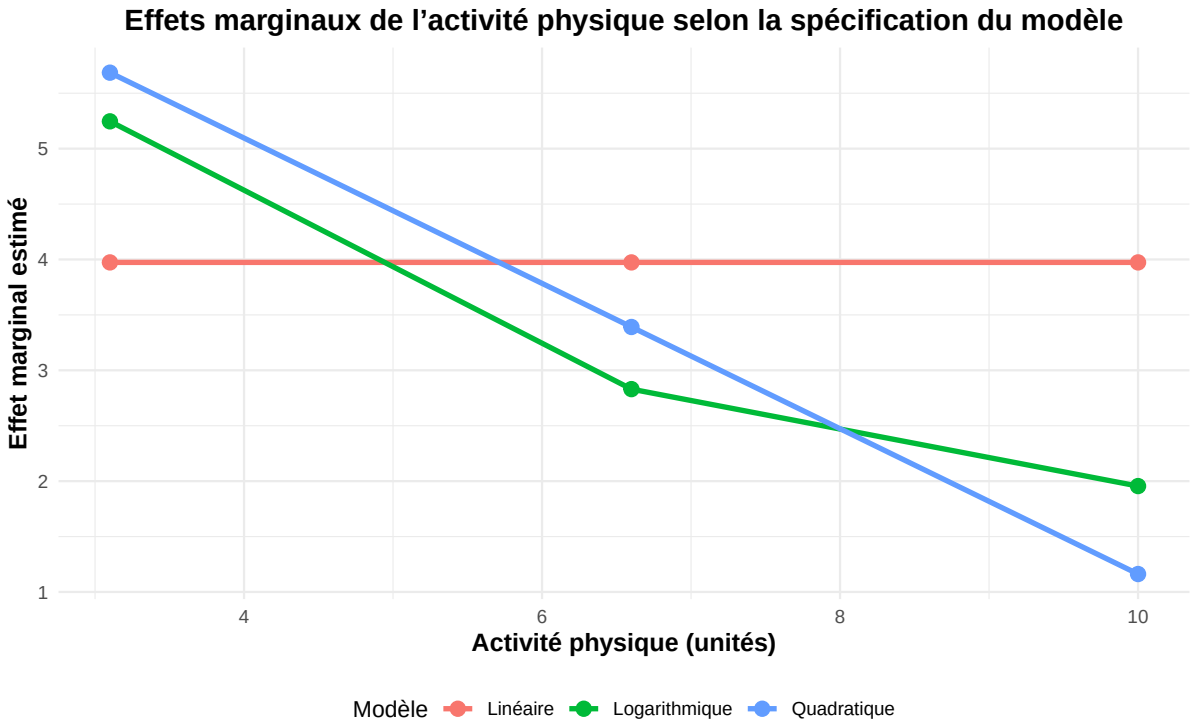

Figure 9: Effets marginaux de l'activité physique

Commentaires : Les modèles logarithmique et quadratique montrent une décroissance de l'effet marginal.

7 Analyse de médiation : Rôle de l'IMC

L'analyse de médiation permet de distinguer l'effet direct et indirect (via l'IMC) de l'activité physique sur le score de santé.

Tableau 12: Effets de médiation de l'IMC

Effet	Estimé	IC.95...Bas.	IC.95...Haut.	p.value
Effet total	3.974	3.788	4.169	NA
Effet direct (ADE)	3.708	3.586	3.843	<0.001
Effet indirect (ACME)	0.265	0.202	0.327	<0.001
Proportion médiée	0.067	0.051	0.082	NA

Commentaire :
- L'effet indirect via l'IMC représente une part significative de l'effet total, confirmant son rôle de médiateur.

8 Extension par apprentissage statistique

Les régressions linéaires supposent une forme fonctionnelle. Des méthodes non paramétriques comme les **forêts causales** ou le **double machine learning (DML)** permettent d'estimer les effets sans imposer de contraintes strictes.

9 Limites et implications

9.1 Limites du design

- Les données simulées sont transversales, limitant l'analyse des effets temporels.
- Les biais de simultanéité ou de sélection inversée peuvent persister dans des données réelles.

9.2 Implications pratiques

- Définir clairement les mécanismes causaux.
- Mesurer les confondeurs.
- Utiliser des méthodes robustes.
- Valider avec des approches alternatives (ex. variables instrumentales si biais sur les inobservables).

Message clé : Contrôler uniquement les confondeurs légitimes pour éviter les biais.

10 Références

- Angrist, J. D., & Pischke, J. S. (2008). *Mostly Harmless Econometrics.*
  Lien
- Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If.*
  Lien
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.*
  Lien
- Cunningham, S. (2021). *Causal Inference: The Mixtape.*
  Lien
- **Mon Github:**
  Lien