ÉCOLE NATIONALE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION


PROJET METHODOLOGIQUE 3A

---

# Correction du biais de sélection dans les enquêtes multimodes

---

*Autrices :*
Cléo BOREL HERBERT
Maha DAO
Fanny DURAND

*Encadrant :*
Stéphane LEGLEYE

Année 2023-2024

Mixed-mode surveys use different collection channels, which increases statistical bias. In particular, selection bias refers to disparities in the composition of respondents due to the collection mode chosen. The aim of our project is to propose adjustment methods to correct this bias.

We focus on a sequential multi-mode internet then telephone, and assume that there is no measurement bias. We simulate populations and samples by simple random sampling for three different scenarios : 1) No selection bias, 2) Ignorable selection bias, for which response probabilities depend on observable characteristics and 3) Non-ignorable selection bias, for which response probabilities depend on non-observable characteristics. In our model, the variable of interest is available to respondents only and depends on observable and non-observable factors. These factors can be modelled using one or more variables depending on the survey framework. We assume that observable characteristics are available for all individuals in our sampling frame. The probabilities of response by mode of collection are not observed.

Various probability distributions can be used to simulate the different variables ; we have mainly used normal distributions for quantitative variables and Bernouilli distributions for qualitative variables. To compare the different methods, we rely in particular on calculations of bias and root mean square error (RMSE). We simulated survey frameworks echoing two French multi-mode surveys : the Information and Communication Technologies (TIC) survey and the Epidemiology and Living Conditions under Covid-19 (EpiCov) survey.

Theoretically, in the absence of selection bias the Horvitz-Thompson estimator is suitable, but imputation methods help to reduce the variance. Where selection bias is not negligible, imputation model and reweighting methods are supposed to be the most effective. Finally, Heckman's method is the best for non-ignorable selection bias. These results are confirmed in the implementation on data simulating the TIC and EpiCov surveys. For TIC, the reweighting method that takes both modes into account is the most efficient, and the literature reviews suggest that indeed there was little or no selection bias. For EpiCov, the Heckman methods in one or two stages are significantly more effective, dividing the RMSE by three compared with a homogeneous correction by strata.

KEYWORDS : mixed-mode, sequential, selection bias, Heckman

# Table des matières

# Introduction

En 2010, l'Insee a réalisé une enquête-test Logement qui a été comparée à l'édition 2006 de l'Enquête National Logement (ENL). Le protocole incluait trois modes de collecte différents : une enquête par internet pour 10 000 ménages issus du fichier taxe d'habitation de 2009, suivie d'une enquête téléphonique pour 1 000 non-répondants et d'une enquête en face à face auprès de 500 ménages. Le taux de réponse par internet était de 19,5%, contre 70% pour l'enquête face à face, et contre 76% pour ENL 2006 dont la collecte se faisait par assistance ordinateur (CAPI). Les résultats de l'enquête internet présentaient également des caractéristiques particulières, notamment une déclaration de revenus élevés et des résultats différents de ENL 2006, ce qui a mis en exergue des effets liés au mode.

Ces effets sont aujourd'hui particulièrement étudiés dans le cadre des enquêtes multimodes. En effet, si ce type d'enquêtes n'est pas une nouveauté, elles se singularisent depuis quelques années par le recours à internet. Ce développement répond à plusieurs impératifs adressés à la statistique publique : faire face à la baisse du taux de réponse, réduire les coûts de collecte et diminuer l'erreur totale de l'enquête (erreur de couverture, erreur d'échantillonnage, etc). Il s'agit aussi de répondre à la stratégie Ambition 2015 avec la recherche de méthodes innovantes.

Le recours à ces enquêtes apporte néanmoins son lot de questions. En effet, un même individu peut ne pas répondre de la même façon lorsqu'il s'agit d'un questionnaire auto-administré que lorsqu'il s'agit d'un questionnaire hétéro-administré, nécessitant la présence d'une personne tierce. De même, une personne qui va préférer répondre par téléphone ne présente certainement pas les mêmes caractéristiques socio-démographiques qu'un individu préférant répondre en ligne. Ces deux biais, appelés biais de sélection et biais de mesure, sont constitutifs des enquêtes multimodes.

Le multimode complexifie le processus d'enquête et engendre de nouvelles contraintes pour le traitement des données. Dans le cadre de notre projet méthodologique, nous avons alors étudié l'impact du biais de sélection sur la précision et l'adéquation des résultats de l'enquête. Pour cela, nous avons cherché à quantifier l'impact des corrélations entre les variables sur différentes méthodes de redressement.

Le présent rapport décrit les différentes étapes de notre cheminement. Après avoir défini les concepts théoriques que nous mobilisons, nous présentons le cadre expérimental et la méthodologie mise en œuvre. Nous détaillons ensuite les différents redressements implémentés avant de proposer deux exemples d'application, basés sur les enquêtes TIC et EpiCov.

# 1 Le cadre théorique

Dans cette partie nous allons définir les concepts statistiques que nous avons mobilisé tout au long de notre projet.

## 1.1 Les enquêtes multimodes

Les enquêtes de la statistique publique interrogent des populations différentes selon la thématique étudiée, et répondent à des contraintes financières et géographiques particulières selon le contexte. Le choix du ou des mode(s) de collecte se fait au regard de l'objectif de l'étude, tout comme le protocole de contact qui est une étape clé dans la réussite à mobiliser l'échantillon retenu. De façon générale, le multimode désigne le fait de retenir plusieurs façons de contacter et/ou d'interroger les individus. Cependant, la définition se restreint souvent au fait de recourir à différents modes de collecte.

**Définition 1: Le multimode**

**Enquête multimode** : Une enquête est dite multimode lorsque le recueil d'information, soit la collecte des données, se fait par au moins deux moyens différents.
Parmi les moyens les plus utilisés, nous pouvons citer les entretiens en face à face, les questionnaires en ligne et les appels téléphoniques.

Le recours au multimode peut se faire selon plusieurs méthodologies, choisies elles aussi en fonction du contexte et des objectifs de l'enquête :

- **Multimode séquentiel** : Un seul mode est proposé initialement, puis après un temps déterminé les non-répondants se voient proposer au moins une autre alternative. En général, le premier mode proposé est celui qui minimise les coûts de la collecte.

- **Multimode concurentiel** : Dès le début de l'enquête, les répondants ont le choix du mode de réponse.

- **Multimode intégré** : Au cours de la collecte, il y a recours à un autre mode pour tous les répondants. Par exemple, un enquêteur pose les premières questions, puis un questionnaire auto-administré est donné aux répondants. Ce choix peut notamment être fait en raison de la sensibilité des thématiques évoquées.

- **Multimode disjoint** : Les répondants sont divisés en (au moins) deux groupes et chaque groupe se voit proposer un mode de collecte différent.

Notre projet méthodologique est restreint aux enquêtes multimodes séquentielles dont la collecte se déroule par internet puis téléphone.

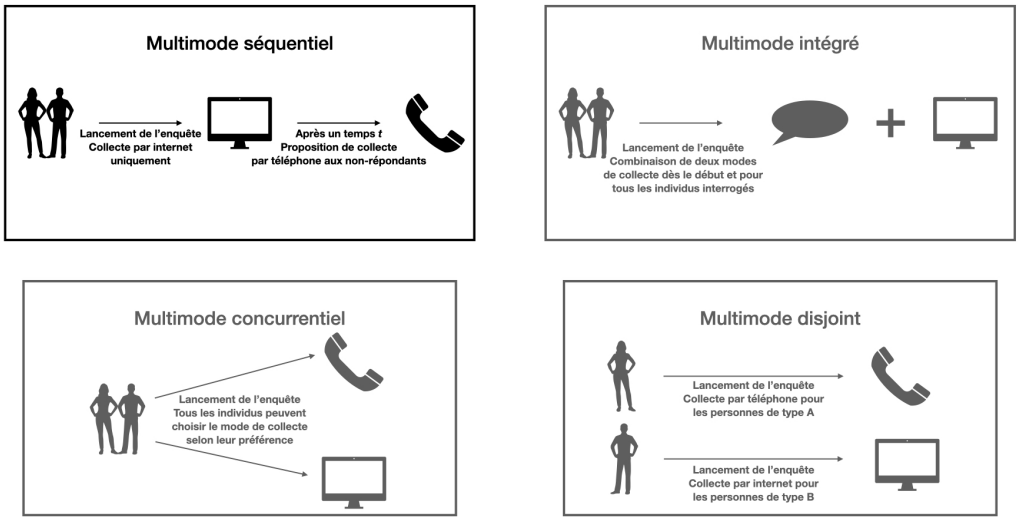


FIGURE 1 – Illustration des différents types de multimode

## 1.2 Définition des biais

Dans le cas d'une enquête multimode, les données récoltées par un mode ne sont pas forcément comparables à celles collectées par une autre méthode. Par exemple, un individu qui répond en ligne à un questionnaire fera sans doute moins preuve de désirabilité sociale qu'un individu qui échange au téléphone avec un enquêteur : il s'agit de l'effet mode.

**Définition 2: Effet mode et les différents biais**

**Effet de mode** : Désigne le fait qu'il n'est pas possible de comparer directement les résultats de deux modes de collecte différents.

Nous pouvons en distinguer deux :

— **Biais de mesure** : Le biais de mesure se réfère au fait qu'un même individu ne répond pas de la même façon selon le mode de collecte qui lui est proposé. C'est une conséquence directe du choix du mode de collecte.
Ce biais illustre l'importance du contexte de l'enquête et fait écho à différents mécanismes humains et sociaux tels que la désirabilité sociale ou le satisficing.

— **Biais de sélection** : Le biais de sélection renvoie aux disparités dans la composition des répondants en raison du mode de collecte choisi.
Par exemple, la propension à répondre par internet des personnes plus âgées est plus faible que pour les jeunes : les personnes âgées pourront alors être sous-représentées parmi les répondants de ce mode.

Notre projet intègre uniquement un biais de sélection. Il se réfère aux caractéristiques structurelles des populations, qui peuvent être observées ou non : c'est pourquoi nous distinguons deux types de biais de sélection.

**Définition 3: Les types de biais de sélection**

— **Biais de sélection ignorable** : Un biais de sélection est dit ignorable lorsqu'il est conditionnel à des variables observables, et qu'il peut donc être corrigé. Par exemple, l'âge est un facteur observable et connu dans les enquêtes statistiques.

— **Biais de sélection non-ignorable** : A l'inverse, un biais de sélection non-ignorable dépend de l'inobservé ou de variables non mesurables. Il ne peut donc pas être directement corrigé. Par exemple le degré d'adhésion à des théories complotistes sur le Covid a probablement désincité certaines personnes à répondre à l'enquête EpiCov, mais c'est un effet qui est difficilement quantifiable.

Nous souhaitons donc corriger les biais de sélection au moyen de plusieurs méthodes de redressement.

# 2 Le cadre expérimental

Le but de cette partie est de formaliser le problème traité dans ce rapport, tout en définissant les cas envisagés et le protocole mis en place. En effet, dans la mesure où nous ne travaillons pas sur des données réelles mais sur des simulations nous avons dû faire un certain nombre de choix, tant du point de vue de la modélisation que du point de vue de l'implémentation. Il nous semble donc nécessaire de les présenter au préalable.

## 2.1 Formalisation mathématique du problème

Tout au long de notre rapport, nous utiliserons les notations suivantes :

— $Y$ est la variable d'intérêt,

— $X$ regroupe les caractéristiques socio-démographiques des individus,

— $U$ désigne les caractéristiques inobservables,

— $P_i$ est la propension à répondre à l'enquête via internet,

— et $P_t$ est la propension à répondre à l'enquête par téléphone.

Nous supposons que $X$ est une variable auxiliaire, c'est-à-dire que les caractéristiques socio-démographiques sont présentes dans la base de sondage et disponibles pour tous les individus (répondants et non-répondants). La variable Y n'est quant à elle observée que pour les répondants. Enfin, la variable U et les probabilités $P_i$ et $P_t$ ne sont pas observées.

### 2.1.1 Décomposition de Y

Nous supposons que la variable d'intérêt $Y$ peut dépendre non seulement de caractéristiques socio-démographiques observées, mais aussi de caractéristiques inobservables. Ainsi, nous faisons l'hypothèse que $Y$ peut s'écrire de la façon suivante :

$$Y = f(X) + g(U) \tag{1}$$

Les fonctions $f$ et $g$ sont quelconques, et peuvent être en particulier la fonction nulle.

### 2.1.2 Probabilités de réponse

Dans la mesure où nous travaillons sur des enquêtes multimodes, nous disposons pour chaque individu de sa probabilité de réponse par internet $P_i$ et de sa probabilité de réponse par téléphone $P_t$. Dans la phase de simulation des données elles sont utilisées pour déterminer le comportement de réponse de chaque individu enquêté. Le détail de la procédure de simulation de la non-réponse est renseigné dans la partie 2.3.
Enfin, nous faisons l'hypothèse que ces probabilités de réponse sont toutes les deux liées à un comportement global propre à chaque individu (et que nous ne simulons pas) : c'est pourquoi nous avons choisi de considérer une corrélation non-nulle entre $P_i$ et $P_t$. La valeur de cette dernière a été fixée arbitrairement à 0,5.

### 2.1.3 Matrice de corrélation et scénarios envisagés

Cette matrice permet de spécifier les corrélations entre toutes les variables du modèle. Elle est de la forme suivante [1] :

$$Corr_{Y,X,U,P_i,P_t} = \begin{pmatrix} C_Y & C_{YX} & C_{YU} & C_{YP_i} & C_{YP_t} \\ C_{XY} & C_X & C_{XU} & C_{XP_i} & C_{XP_t} \\ C_{UY} & C_{UX} & C_U & C_{UP_i} & C_{UP_t} \\ C_{P_iY} & C_{P_iX} & C_{P_iU} & C_{P_i} & C_{P_iP_T} \\ C_{P_tY} & C_{P_tX} & C_{P_tU} & C_{P_tP_i} & C_{P_t} \end{pmatrix}$$

1. De manière évidente, tous les éléments diagonaux sont égaux à 1.

Les différents scénarios concernant la présence ou l'absence de biais de sélection se manifesteront ensuite au travers de cette matrice de corrélations.

**Scénario 1 : absence de biais de sélection**

Dans le cas d'une absence de biais de sélection, on suppose qu'aucune variable (observée ou inobservée) n'influe sur le comportement de réponse à l'enquête. Comme on peut le voir sur la matrice ci-dessous, on supposera donc que les corrélations entre les probabilités de réponse et les variables $X$ et $U$ sont égales à 0.

$$Corr_{Y,X,U,P_i,P_t} = \begin{pmatrix} 1 & C_{YX} & C_{YU} & 0 & 0 \\ C_{XY} & 1 & C_{XU} & 0 & 0 \\ C_{UY} & C_{UX} & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & C_{P_iP_T} \\ 0 & 0 & 0 & C_{P_tP_i} & 1 \end{pmatrix}$$

Notons que, compte-tenu de l'équation 1, les corrélations entre $Y$ et les probabilités de réponse sont nulles également.

**Scénario 2 : biais de sélection ignorable**

Dans le cas de l'existence d'un biais de sélection ignorable, on suppose que les probabilités de réponse dépendent des caractéristiques socio-démographiques observées. On obtient la matrice suivante, avec des corrélations non-nulles entre $X$ et $P_i$, et entre $X$ et $P_t$.

$$Corr_{Y,X,U,P_i,P_t} = \begin{pmatrix} 1 & C_{YX} & C_{YU} & C_{YP_i} & C_{YP_t} \\ C_{XY} & 1 & C_{XU} & C_{XP_i} & C_{XP_t} \\ C_{UY} & C_{UX} & 1 & 0 & 0 \\ C_{P_iY} & C_{P_iX} & 0 & 1 & C_{P_iP_T} \\ C_{P_tY} & C_{P_tX} & 0 & C_{P_tP_i} & 1 \end{pmatrix}$$

**Scénario 3 : biais de sélection non-ignorable**

Enfin, dans le cas d'un biais de sélection non-ignorable, on suppose que les probabilités de réponse dépendent également de caractéristiques inobservables. Nous obtenons donc la matrice suivante, avec des corrélations non nulles entre les probabilités de réponse et la variable inobservée $U$ :

$$Corr_{Y,X,U,P_i,P_t} = \begin{pmatrix} 1 & C_{YX} & C_{YU} & C_{YP_i} & C_{YP_t} \\ C_{XY} & 1 & C_{XU} & C_{XP_i} & C_{XP_t} \\ C_{UY} & C_{UX} & 1 & C_{UP_i} & C_{UP_t} \\ C_{P_iY} & C_{P_iX} & C_{P_iU} & 1 & C_{P_iP_T} \\ C_{P_tY} & C_{P_tX} & C_{P_tU} & C_{P_tP_i} & 1 \end{pmatrix}$$

### 2.1.4 Modélisation des variables

Nous souhaitons avoir une certaine flexibilité dans la modélisation de nos variables $Y$, $X$, $U$, $P_i$ et $P_t$, aussi, nous avons fait en sorte que ces dernières puissent être simulées suivant différentes lois de probabilité. Nous avons très régulièrement modélisé les variables quantitatives par des lois normales, mais il est possible d'envisager d'autres lois continues comme la loi uniforme ou la loi gamma. En ce qui concerne les variables discrètes, nous avons essentiellement travaillé avec des variables binaires, que nous avons donc modélisées par des lois de Bernoulli.

De plus, afin de pouvoir se rapprocher au plus d'une situation réelle, nous avons permis une importante liberté dans la modélisation des caractéristiques socio-démographiques observées $X$. Ainsi, il est possible de se placer dans l'un des trois cas suivants :

1) $X$ représente une seule variable. Il s'agit du scénario le plus rudimentaire, qui est assez peu cohérent avec notre cadre d'application. En effet, dans une enquête, les caractéristiques socio-démographiques observées se limitent rarement à une seule variable.

2) $X$ synthétise l'information contenue dans plusieurs variables. Par exemple, $X$ pourrait correspondre aux coordonnées d'une Analyse en Composantes Principales (ACP). Cette possibilité permet d'enrichir notre modélisation en ajoutant des informations, tout en conservant une certaine simplicité due au fait qu'on ne manipule *in fine* qu'une seule variable.

3) $x$ regroupe plusieurs variables distinctes. Ainsi, nous pourrions par exemple vouloir inclure dans notre modèle à la fois l'âge, le sexe, la situation familiale et la catégorie socio-professionnelle. Cette solution offre une grande flexibilité, mais elle complexifie les calculs en alourdissant fortement la matrice de corrélations.

Enfin, on retrouve ces trois cas de figure pour la modélisation des caractéristiques inobservables.

## 2.2 Méthodologie

Notre approche se décompose en cinq étapes.

**Etape 1 :** Nous commençons par simuler une population de taille $N$ à partir des lois de nos variables et de leur matrice de corrélations.

**Etape 2 :** Nous sélectionnons ensuite un échantillon de $n$ individus issus de cette population, en suivant un certain plan de sondage. Cette étape correspond au tirage de l'échantillon sur lequel l'enquête sera réalisée.

**Etape 3 :** Nous introduisons de la non-réponse selon une procédure présentée dans la partie 2.3 en tenant compte des scores $P_i$ et $P_t$. Ceci permet de modéliser la phase de passation du questionnaire auprès des individus enquêtés.

A l'issue de ces trois étapes, nous disposons donc d'une table de données semblable à celle dont disposerait un chargé d'enquête après la collecte des données.

**Etape 4 :** Nous testons plusieurs méthodes de redressement afin d'approcher au mieux le total de la variable d'intérêt Y, dont nous connaissons la valeur sur l'ensemble de la population.

**Etape 5 :** Enfin, pour départager ces méthodes et déterminer laquelle est la meilleure, nous utilisons différents critères à savoir le biais, la variance, l'erreur quadratique et l'erreur quadratique moyenne de chaque estimateur.

## 2.3 Choix d'implémentation

Tout d'abord, il convient de préciser la taille de la population avec laquelle nous travaillons. Nous avons choisi de simuler systématiquement une population de taille $N = 2000000$ à l'étape 1, puis de sélectionner au sein de cette population un échantillon de taille $n = 10000$ à l'étape 2.

Concernant la méthode d'échantillonnage mobilisée à l'étape 2, nous avons utilisé un sondage aléatoire simple pour toutes les modélisations présentées dans la partie 3. Ainsi, tous les individus de la population ont comme probabilité d'être tiré dans l'échantillon $\pi_k = \frac{n}{N} = \frac{10000}{2000000} = 0.005$.

Penchons-nous maitenant sur la simulation des comportements de réponse. Dans la majorité des cas, nous avons supposé une non-réponse proportionnelle à la propension de réponse. Plus précisément, nous avons procédé de la manière suivante :

1. Simulation de deux vecteurs aléatoires suivant une loi uniforme sur l'intervalle [0, 1], notés $Seuil_{internet}$ et $Seuil_{telephone}$. Ces deux vecteurs prennent une valeur différente pour chacun des individus.
2. Grâce aux seuils précédents, on détermine les indicatrices de réponse par internet et/ou par téléphone. Nous comparons alors respectivement les valeurs de $P_i$ et $P_t$ aux seuils $Seuil_{internet}$ et $Seuil_{telephone}$ : si la valeur du score est supérieure à la valeur du seuil alors nous supposons que l'individu répond, sinon nous supposons qu'il ne répond pas.
3. Nous terminons en traitant les cas où un individu répond par internet et par téléphone. Ceci n'est pas envisageable car un individu ne peut pas répondre deux fois à l'enquête. De plus, comme nous

sommes dans le cas d'un protocole séquentiel internet-téléphone, nous considérons que si la probabilité de réponse par internet d'un individu est supérieure à la valeur de $Seuil_{internet}$, alors il répond par internet, et ce même si sa probabilité de réponse par téléphone était également supérieure à la valeur de $Seuil_{telephone}$. A l'issue de ces deux dernières étapes, nous avons ainsi déterminé pour chaque individu s'il était répondant internet, répondant téléphone ou non-répondant (dans le cas où les deux probabilités sont inférieures à leurs seuils respcetifs).

Les détails concernant les différentes méthodes de redressement étant détaillées dans la partie 3, il nous reste à préciser notre démarche d'évaluation de ces dernières. Concrètement, nous avons implémenté une fonction qui prend en argument notre table de données, la taille de l'échantillon, le plan de sondage utilisé pour l'obtenir, la méthode d'ajout de la non-réponse, la méthode de redressement que l'on veut évaluer et le nombre d'échantillons sur lesquels nous allons évaluer cette dernière. Elle renvoie ensuite la valeur du biais empirique, de la variance empirique et de l'erreur quadratique moyenne associés à la méthode en question. Enfin, nous rassemblons les résultats dans un tableau récapitulatif afin de pouvoir les comparer.

# 3 Méthodes de redressement

## 3.1 Estimation d'un total avec un échantillon

Toutes les méthodes que nous avons implémentées ont le même objectif : utiliser l'information récoltée sur l'échantillon enquêté pour estimer au mieux le total de Y sur la population, sachant qu'il n'y a pas de biais de mesure. Notons $t_y$ ce total. On remarque également que la plupart des autres estimations que l'on peut souhaiter faire (ratio, corrélation, etc) peuvent être écrites comme des fonctions de totaux. En passant par un estimateur par substitution on pourrait alors généraliser les méthodes présentées pour des cas plus complexes.

Si on revient au cadre de notre projet, nous avons une population initiale $P$ de taille $N = 2000000$ dans laquelle nous tirons un échantillon $S$ de taille $n = 10000$ selon le plan de sondage $p$. Chaque individu $k$ a une probabilité $\pi_k$ d'être tiré dans l'échantillon. En l'absence de non-réponse et en supposant que tous les $\pi_k$ sont connues et strictement positives, nous pouvons estimer sans biais le total de Y grâce à l'estimateur de Horvitz-Thompson :

$$\hat{t_y} = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{2}$$

Dans le cas particulier d'un plan de sondage aléatoire simple sans remise, les probabilités d'inclusion sont toutes égales à $n/N$, ce qui nous donne l'estimateur de Horvitz-Thompson suivant :

$$\hat{t_y} = \frac{N}{n} \sum_{k \in S} y_k \tag{3}$$

Cependant, lorsqu'il y a des valeurs manquantes, la valeur de $y$ n'est pas connue pour tout l'échantillon. On pose ainsi le sous-échantillon $S_r$ des individus qui répondent à l'enquête. Il est de taille $n_r < n$ et sa composition dépend implicitement des probabilités de réponse $p_k$.
Lorsque les $p_k$ sont connues et strictement positives on peut alors estimer $t_y$ sans biais avec l'estimateur par expansion :

$$\hat{t_y} = \sum_{k \in S_r} \frac{y_k}{\pi_k p_k} \tag{4}$$

C'est notamment le cas lorsqu'on suppose que la non-réponse est aléatoire (mécanisme de réponse MCAR ou "Missing Completely At Random"), puisque pour tout $k \in S$ on a $p_k = n_r/n$. En l'injectant dans la formule 4 nous obtenons ainsi **HT répondants** la première méthode codée. Nous avons choisi ce nom car en modifiant légèrement l'écriture on voit qu'il s'agit en fait d'adapter un estimateur de Horvitz-Thompson sur $S_r$ au nombre de répondants $n_r$ :

$$\hat{t_y} = \frac{n}{n_r} \sum_{k \in S_r} \frac{y_k}{\pi_k} \tag{5}$$

L'avantage de cette méthode est qu'elle est simple à mettre en œuvre, en particulier car nous n'avons pas besoin de connaître $X$ pour tout le monde, tout en restant sans biais sous l'hypothèse de non-réponse aléatoire. Il en résulte seulement une augmentation de la variance car le nombre de données exploitables a diminué. Il est tout de même appréciable d'avoir des informations auxilaires puisqu'elles permettent d'améliorer nos estimateurs, via du calage sur marges par exemple.

Néanmoins l'hypothèse de non-réponse aléatoire manque de réalisme : en présence d'un biais de sélection la non-réponse est ignorable (mécanisme MAR ou "Missing At Random") voire même non-ignorable (mécanisme NMAR ou "Non Missing At Random"), ce qui nécessite des méthodes plus sophistiquées. Pour finir, nous considérons que la non-réponse partielle est une manière comme une autre de répondre à l'enquête et donc que son traitement n'est pas lié au biais de sélection mais au biais de mesure. Etant donné que les effets de mesure ne sont pas considérés dans ce rapport nous avons choisi de ne pas aborder ce type de données manquantes, mais comme nos méthodes interviennent en amont dans le traitement des données cela n'aura pas d'impact pour la suite.

## 3.2 Correction du biais de sélection par imputation

Supposons désormais que le mécanisme de réponse est lié à $X$. A l'exception du cas où $Y$ et $X$ sont indépendantes, la non-réponse est alors reliée à $Y$ et nous sommes en présence d'un biais de sélection. Notons tout de même que ce biais est ignorable car il dépend uniquement des variables observées. Il existe alors deux approches pour y faire face :

— repondérer l'échantillon pour qu'il respecte à nouveau les hypothèses de la partie 3.1,

— et se débarasser des valeurs manquantes en imputant $Y$ pour les non répondants.

Nous commencerons ici par l'imputation des valeurs manquantes, puis nous discuterons des méthodes de repondération dans la partie 3.3.

Le principe des méthodes d'imputation est qu'on utilise l'ensemble des répondants $S_r$ pour estimer le lien général entre la variable d'intérêt $Y$ et les variables auxiliaires $X$. La connaissance d'une telle relation permet d'estimer $y_k$ pour tous les individus $k \in S$, indépendamment de leur réponse à l'enquête, et par conséquent indépendamment du biais de sélection. De plus, comme $X$ est connu dans toute la base de sondage, on peut aussi ignorer le biais d'échantillonnage en prédisant $\hat{y_k}$ pour tout $P$ :

$$\hat{t_y} = \sum_{k \in P} \hat{y_k} \tag{6}$$

D'un point de vue théorique il est néanmoins difficile de garantir que ces méthodes seront sans biais. En effet, au delà de la validité interne (est-ce qu'il est possible de représenter $Y$ uniquement à l'aide de $X$ et est-ce que le modèle choisi est adapté) il faut également prendre en compte la validité externe et éviter le surapprentissage. Un second point à mettre en avant est que si $X$ est mal représentée dans l'échantillon des répondants alors il ne sera pas possible de construire un modèle qui soit pertinent pour toute la population. C'est par exemple le cas en présence de non-répondants irréductibles, c'est-à-dire d'individus ayant une probabilité nulle de répondre à l'enquête ($p_k$=0).

Nous avons choisi d'implémenter trois méthodes, soit trois modélisations de $Y$ :

1. **Modèle linéaire** : On suppose que $\forall k \in P, y_k = \alpha + \beta x_k + \epsilon_k$ avec $Cov(\epsilon_k) = \sigma^2 I$ et on estime :

$$\hat{t_y} = \sum_{k \in P} (\hat{\alpha} + \hat{\beta} x_k) \tag{7}$$

   La fonction permet de forcer $\alpha = 0$, d'intégrer seulement une partie des variables contenues dans $X$, de transformer certaines variables et d'ajouter des termes d'interactions.

2. **Modèle ratio** : On suppose que $\forall k \in P, y_k = \beta x_k + \epsilon_k$ avec $Cov(\epsilon_k) = \sigma^2 x_k$ et on estime :

$$\hat{t_y} = \sum_{k \in P} \hat{\beta} x_k \tag{8}$$

   La fonction nécessite d'avoir une unique variable dans X mais il est possible de la définir comme une combinaison de plusieurs autres variables.

3. **Modèle homogène par strates** : On définit une partition $(P_1, ..., P_H)$ de $P$ à l'aide des $H$ quantiles de $X_{strate}$ et on suppose que $\forall h \in [\![1; H]\!], \forall k \in P_h, y_k = c_h$. En notant $(n_1, ..., n_H)$ les tailles respectives de $(P_1, ..., P_H)$ on estime :

$$\hat{t_y} = \sum_{h=1}^{H} n_h \hat{c_h} \tag{9}$$

   On remarque qu'il y a plusieurs façons de définir la partition $(P_1, ..., P_H)$ et les constantes $(c_1, ..., c_H)$. Ainsi, même si notre fonction ne le permet pas, il serait intéressant de créer des versions alternatives pour prendre en compte plusieurs variables, fixer des seuils arbitraires, utiliser la médiane de $Y$ au lieu de la moyenne, etc.

Aucune des méthodes présentées ci-dessus ne prend en compte l'aspect multimode de l'enquête, car même en construisant un modèle pour les répondants internet et un second pour les répondants téléphones il serait impossible de savoir lequel utiliser pour les non-répondants. Nous avons envisagé plusieurs solutions comme par exemple prédire un "mode de réponse privilégié" pour chaque individu ou utiliser la moyenne des deux prédictions, mais accumuler les modèles pourrait nuire à la robustesse des résultats.

## 3.3 Correction du biais de sélection par re pondération

### 3.3.1 Cas d'un biais de sélection ignorable

Dans le cas de la non-réponse aléatoire, on pouvait utiliser l'estimateur par expansion 4 car les $p_k$ étaient connues conditionnellement à $S$. Lorsque la non-réponse est ignorable, nous n'avons plus de formule exacte pour $p_k$, mais puisqu'elles dépendent uniquement de $X$ qui est observé, nous pouvons les estimer. Si toutes les probabilités $\pi_k$ sont connues et strictement positives, et si toutes les estimations $\hat{p_k}$ sont strictement positives, nous pouvons utiliser l'estimateur corrigé de la non-réponse pour le total de $Y$ :

$$\hat{t_y} = \sum_{k \in S_r} \frac{y_k}{\pi_k \hat{p_k}} \tag{10}$$

L'enjeu des méthodes de repondération consiste alors à estimer correctement $p_k$. A l'instar des méthodes par imputation de la partie 3.2, il est donc nécessaire de faire attention à la qualité du modèle, à la robustesse des résultats et à la présence potentielle de non-répondants irréductibles. Nous avons ainsi utilisé des groupes homogènes de réponse (GHR) afin de diminuer la variance des estimations. En d'autres termes, au lieu d'utiliser directement les résultats renvoyés par les régressions logistiques, nous avons regroupé les observations ayant des valeurs proches de $\hat{p_k}$. Dans chaque groupe nous avons ensuite remplacé $\hat{p_k}$ par le taux de réponse observé car il est moins volatil.
La construction des GHR est codée de façon à pouvoir utiliser les quantiles des $\hat{p_k}$, une classification ascendante hiérarchique (CAH) ou un algorithme des moyennes mobiles (algorithme k-means). Ces trois possibilités ont pour point commun d'utiliser les prédictions individuelles pour rapprocher les observations, ce qui est caractéristique des méthodes dites "des scores". Elles s'opposent aux méthodes "des croisements" qui forment les groupes par stratification de $X$.

Tout d'abord, nous pouvons estimer $p_k$ à l'aide de l'indicatrice de réponse $R$ pour toutes les enquêtes, qu'elles soient multimodes ou monomodes. La fonction correspondante est **Sans mode**, et comme évoqué précédemment elle ajuste une régression logistique sur $S$ avant de prédire $\hat{p_k}$ via des groupes homogènes de réponse. Cependant, dans le cas particulier des enquêtes multimodes nous pouvons exploiter, non pas une, mais deux variables de réponse : $R_i$ pour indiquer si l'individu a répondu par internet et $R_{tnoni}$ pour indiquer si l'individu a répondu par téléphone. Comme le protocole de collecte est séquentiel internet/téléphone on rappelle que :

— si un individu répond par internet cela ne veut pas dire qu'il n'aurait pas répondu par téléphone,
— et qu'au contraire un individu qui répond par téléphone a refusé de répondre par internet.

Nous repondérons alors nos données en utilisant $R_i$ et $Rtnoni$. La première manière de faire consiste à repondérer différemment les individus qui ont répondu par internet (ensemble $S_{ri}$) et ceux qui ont répondu par téléphone (ensemble $S_{rt}$). Pour cela nous avons appliqué deux méthodologies différentes :

1. **Repondération de chaque mode** : Nous construisons pour tous les individus de l'échantillon $S$ deux probabilités de réponse estimées : $\hat{p_{ki}}$ la probabilité de répondre par internet et $\hat{p_{kt\hat{noni}}}$ la probabilité de répondre par téléphone. Il serait alors possible d'utiliser uniquement $S_{ri}$ ou uniquement $S_{rt}$ pour prédire $t_y$ :

$$\hat{t_y} = \sum_{k \in S_{ri}} \frac{y_k}{\pi_k \hat{p_{ki}}} = \sum_{k \in S_{rt}} \frac{y_k}{\pi_k \hat{p_{kt\hat{noni}}}} \tag{11}$$

   Supposons qu'il y a $n_i$ individus dans $S_{ri}$ et $n_t$ dans $S_{rt}$. Nous avons choisi d'intégrer les deux modes de collecte proportionnellement à leur part dans l'échantillon des répondants $S_r$, ce qui nous mène à utiliser l'estimateur suivant :

$$\hat{t_y} = \frac{n_i}{n_r} \sum_{k \in S_{ri}} \frac{y_k}{\pi_k \hat{p_{ki}}} + \frac{n_t}{n_r} \sum_{k \in S_{rt}} \frac{y_k}{\pi_k \hat{p_{kt\hat{noni}}}} \tag{12}$$

2. **Repondération des répondants téléphone** : Dans une enquête multimode séquentielle, comme les répondants téléphones sont aussi des non-répondants internet, nous pouvons dire que $S_{rt}$ forme l'ensemble des non-répondants internet ayant répondu à l'enquête.
En parallèle, on peut partitionner $P$ avec d'un côté les personnes qui répondraient par internet (ensemble $P_i$) et de l'autre celles qui ne répondraient pas par internet (ensemble $\bar{P}_i$). Ainsi, le total $t_y$ à estimer est égal à $t_{yi} + t_{ynoni}$, où $t_{yi}$ et $t_{ynoni}$ sont les totaux respectifs de $Y$ dans $P_i$ et $\bar{P}_i$.

   Nous en déduisons une partition de $S$ sous la forme $(S_i, \bar{S}_i)$, avec $S_i \subset P_i$ et $\bar{S}_i \subset \bar{P}_i$. Par construction, puisque $S_i$ est l'ensemble des répondants internet, toutes les probabilités de réponse sont connues et égales à 1. Nous pouvons alors les injecter dans l'estimateur par expansion 4 et tout se passe comme si nous n'avions pas repondéré les répondants internet :

$$\hat{t_{yi}} = \sum_{k \in S_i, S_i = S_{ri}} \frac{y_k}{\pi_k} \tag{13}$$

   Inversement, $\bar{S}_i$ contient de la non-réponse puisqu'il regroupe à la fois les répondants téléphones $S_{rt}$ et ceux qui ne répondent pas du tout à l'enquête. Nous procèdons donc à l'estimation des probabilités de réponse $\hat{p_{knoni}}$ avec une régression logistique et des groupes homogènes de réponse sur $\bar{S}_i$, avant de calculer :

$$\hat{t_{ynoni}} = \sum_{k \in S_{rt}} \frac{y_k}{\pi_k \hat{p_{knoni}}} \tag{14}$$

   Finalement, comme $t_y = t_{yi} + t_{ynoni}$ nous pouvons sommer les estimateurs 13 et 14 pour obtenir notre estimateur d'intérêt :

$$\hat{t_y} = \sum_{k \in S_{ri}} \frac{y_k}{\pi_k} + \sum_{k \in S_{rt}} \frac{y_k}{\pi_k \hat{p_{knoni}}} \tag{15}$$

Nous venons de voir deux manières de repondérer une enquête multimode qui se basent sur un traitement différencié de chaque mode. La deuxième approche pour estimer la probabilité de répondre à l'enquête à l'aide de $R_i$ et $R_{tnoni}$ utilise une seule formule pour tout l'échantillon, et repose sur l'hypothèse que la réponse par téléphone n'est pas impactée par le fait d'avoir préalablement proposé de répondre par internet. On peut alors utiliser la formule des probabilités totales pour écrire : $p_k = p_{ki} + (1 - p_{ki})p_{kt}$, avec $p_{ki}$ la probabilité que $k$ réponde par internet et $p_{kt}$ la probabilité qu'il réponde par téléphone. L'objectif étant toujours d'estimer $p_k$, la fonction **Avec modes** estime séparément $p_{ki}$ et $p_{kt}$ avant de déduire :

$$\hat{t_y} = \sum_{k \in S_r} \frac{y_k}{\pi_k \hat{p_k}} = \sum_{k \in S_r} \frac{y_k}{\pi_k (\hat{p_{ki}} + (1 - \hat{p_{ki}})\hat{p_{kt}})} \tag{16}$$

D'un côté, on estime $p_{ki}$ pour tout l'échantillon $S$ avec l'indicatrice de réponse par internet $R_i$. Encore une fois, on commence par une régression logistique et on détermine $\hat{p_{ki}}$ avec des GHR.
De l'autre côté la situation est plus compliquée car on ne connaît pas l'indicatrice de réponse $R_{tnoni}$ pour les répondants internet. On va alors modéliser $R_{tnoni}$ uniquement pour les individus n'ayant pas répondu par internet. Comme on a supposé que le comportement de réponse par téléphone ne dépendait pas des propositions de réponse préalables, on peut ensuite appliquer la régression logistique aux répondants internet avant de créer les $\hat{p_{kt}}$ via des groupes homogènes de réponse.

Notons pour finir que nous nous sommes inspirées ici des travaux conduits par Louise Kozlowski à la division logement et à la division sondage de l'Insee (cf B.1). Les méthodes présentées sont nommées de la manière suivante dans son rapport de stage : "méthode sans distinguer les modes", "méthode multiplicative sans stratification", "méthode multiplicative avec stratification" et "méthode additive".

### 3.3.2 Cas d'un biais de sélection non ignorable

Supposons maintenant que la non-réponse ne dépend plus seulement de $X$ mais également des facteurs inobservés $U$. Il s'agit de l'hypothèse qui a le plus de chance d'être vérifiée en pratique, en particulier pour les enquêtes ménages où le nombre de variables auxiliaires est restreint. Dans cette situation il y a de grandes chances pour que les méthodes précédentes soient mauvaises car il est de plus en plus difficile d'estimer sans

biais la variable d'intérêt et les probabilités de réponse.

Sous réserve d'avoir trouvé un bon instrument et dans certaines conditions, le modèle de Heckman permet de contrôler la dépendance entre $Y$ et les probabilités de réponse. Il devient alors envisageable d'obtenir un estimateur convergent de $t_y$. Nous en présentons ici les idées principales, et on pourra trouver plus de détails dans le document de travail "Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman" (cf B.1) de Laura Castell et Patrick Sillard.

Si $Y$ est une variable continue, le modèle de Heckman est de la forme suivante : $\begin{cases} y_k &= c^1 + x_k\chi + \epsilon_k^1 \\ r_k^* &= c^0 + x_k\beta + w_k\psi + \epsilon_k^0 \\ r_k &= \mathbb{1}_{r_k^*>0} \end{cases}$

avec :

- $R^*$ la propension à répondre, que l'on observe uniquement à travers l'indicatrice de réponse $R$,
- $c^0$, $c^1$,$\beta$, $\chi$ et $\psi$ des inconnues,
- $\epsilon^0$ et $\epsilon^1$ des aléas,
- et W un instrument observé sur toute la population.

On ajoute deux conditions d'identification : $E(\begin{pmatrix}\epsilon^0\\ \epsilon^1\end{pmatrix}|x_i, w_i) = 0$ et $\begin{pmatrix}\epsilon^0\\ \epsilon^1\end{pmatrix} \sim \mathcal{N}(\begin{pmatrix}0\\ 0\end{pmatrix}, \Sigma)$ avec $\Sigma = \begin{pmatrix}1 & \rho\sigma\\ \rho\sigma & \sigma^2\end{pmatrix}$.

Le choix de l'instrument est essentiel pour garantir la convergence de l'estimateur : $W$ doit être lié à $R^*$, indépendant de $Y$ et monotone, c'est-à-dire qu'une augmentation de $W$ doit toujours avoir le même effet sur la propension à répondre $R^*$. Ici nous avons choisi de séparer $S$ en deux sous-échantillons aléatoires, un qui se voit proposer une collecte multimode séquentielle et un deuxième qui ne peut répondre que par internet. On définit alors $w_k$ comme l'indicatrice de l'échantillon auquel $k$ appartient. $W$ est relié à $R^*$ par construction et il est bien monotone puisque dans tous les cas on propose aux individus de répondre par internet. S'il n'y a pas de biais de mesure la condition d'indépendance par rapport à $Y$ est respectée également, étant donné que l'attribution des modes de collecte est aléatoire.

L'estimation des paramètres du modèle de Heckman peut se faire en maximisant la vraisemblance (on dit qu'elle est faite en une étape) ou avec deux modèles successifs. On peut alors s'intéresser aux probabilités de réponse estimées :

$$\hat{p_k} = \Phi\left(\frac{\hat{c^0} + x_k\hat{\beta} + w_k\hat{\psi} + \frac{\hat{\rho}}{\hat{\sigma}}(y_k - \hat{c^1} - z_k\hat{\chi})}{\sqrt{1-\hat{p^2}}}\right) \tag{17}$$

On retrouve ensuite $t_y$ à l'aide de l'estimateur corrigé de la non-réponse 10, comme dans la partie 3.3.1. On aurait pu également procéder comme dans la partie 3.2 et imputer une valeur $\hat{y_k}$ à partir du modèle de Heckman, cette possibilité n'a pas été implémentée. La fonction **Repondération Heckman** du code permet néanmoins une certaine souplesse puisqu'on peut choisir la modélisation de $Y$, celle de $R^*$ et la méthode d'estimation.

## 3.4 Exemples

Nous présentons ici trois cas abstraits dans lesquels nous avons évalué nos méthodes. Pour chacun d'eux nous avons utilisé $Y \sim \mathcal{N}(10, 2)$, $X \sim \mathcal{N}(0, 1)$, $U \sim \mathcal{N}(0, 1)$ et $P_i, P_t \sim \mathcal{N}(0.5, 0.19^2)$. La loi suivie par $P_i$ et $P_t$ est paramétrée de telle sorte qu'au moins 99% des valeurs sont comprises strictement entre 0 et 1.

Les trois matrices de corrélation sont les suivantes :

1. Un exemple sans biais de sélection :

$$Corr_{Y,X,U,Pi,Pt} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}$$

2. Un exemple de biais de sélection ignorable :

$$Corr_{Y,X,U,Pi,Pt} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.25 & 0.25 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0 & 0 \\ 0.25 & 0.5 & 0 & 1 & 0.5 \\ 0.25 & 0.5 & 0 & 0.5 & 1 \end{pmatrix}$$

3. Un exemple de biais de sélection non-ignorable :

$$Corr_{Y,X,U,Pi,Pt} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$$

Précisons également que ce sont des exemples et non des cas généraux. Ainsi, le fait qu'une méthode soit adaptée ou non pour une de ces matrices ne signifie pas qu'elle le sera toujours "à biais de sélection constant". On verra notamment dans la suite de ce rapport deux exemples de biais de sélection non-ignorables pour lesquels les meilleures méthodes retenues sont différentes (cf parties 4 et 5).

Nous avons ensuite appliqué et comparé les méthodes codées, à l'exception de la méthode ratio qui n'est pas pertinente pour une telle distribution des données. Les groupes homogènes de réponse ont été construit à partir des déciles et nous avons utilisé 2 strates pour X dans la méthode homogène par strates.

— Dans le cas 1 c'est-à-dire sans biais de sélection, la méthode HT répondants est adaptée. Cependant il est préférable d'utiliser les méthodes d'imputation car elles font nettement diminuer la variance.

— Dans le second cas il n'est plus pertinent de considérer que la non-réponse est aléatoire. La comparaison des erreurs moyennes nous amène plutôt à envisager un modèle linéaire ou des méthodes de repondération (hors modèle de Heckman).

— Dans le cas 3 toutes les méthodes sont biaisées et sont associées à de fortes variances. La meilleure méthode est alors le modèle de Heckman estimé en une étape, suivi par la repondération séparée des répondants téléphone.

Pour finir, on remarque que pour ces trois exemples il y a une bonne adéquation entre les erreurs moyennes et les conditions théoriques de validité des méthodes.

# 4 Un premier exemple d'application : l'enquête TIC ménages

## 4.1 Présentation de l'enquête

L'enquête TIC est une enquête annuelle produite par l'Insee, dont le but est d'étudier les pratiques et l'équipement des ménages dans le domaine des technologies de l'information et de la communication. Depuis sa mise en place en 2007, la collecte de cette enquête est réalisée par internet, téléphone et questionnaire papier. Il s'agit donc d'une enquête multimode.

De plus, celle-ci se base sur un échantillon de 39000 ménages tirés dans le fichier Fidéli de l'année n-2, et que l'on peut séparer en deux groupes : les ménages dont le numéro de téléphone peut être retrouvé dans l'annuaire d'une part (qui sont au nombre de 31000 dans l'enquête 2024), et les ménages dont le numéro de téléphone ne peut pas être retrouvé dans l'annuaire d'autre part (8000 ménages en 2024). Les ménages du second groupe sont tous interrogés par internet ou papier, tandis que parmi les ménages du premier groupe on tire deux sous-échantillons : un premier de 3 890 ménages interrogés par téléphone et un second de 21 000 ménages interrogés par internet-papier. Notons que si les ménages sollicités par téléphone n'ont pas répondu au questionnaire au bout de 7 semaines, alors ils sont relancés par internet-papier. On peut donc considérer qu'il s'agit pour l'échantillon téléphone d'une procédure séquentielle téléphone/internet-papier.

Parmi les sujets particulièrement intéressants que cette enquête permet d'étudier, nous pouvons par exemple citer l'illectronisme. Une personne est considérée comme étant en situation d'illectronisme dès lors qu'elle n'a pas utilisé internet au cours des 12 derniers mois, ou dès qu'elle ne possède aucune compétence dans les domaines du numérique (recherche d'information, communication, maîtrise de logiciels, résolution de problèmes nécessitant par exemple des démarches administratives en ligne). Aussi, dans un contexte où l'évolution vers le "tout numérique" pourrait contribuer à creuser les inégalités, la mesure de l'illectronisme est un enjeu important.

**Remarque :**

Comme nous l'avons précisé, l'enquête TIC suit un protocole multimode séquentiel téléphone/internet-papier pour l'échantillon téléphone et un protocole multimode internet-papier pour l'échantillon internet. Il ne s'agit donc pas d'une enquête multimode séquentielle internet/téléphone comme nous l'avons modélisé. Cependant, afin de conserver une certaine cohérence dans ce rapport, nous choisissons de traiter cet exemple en faisant comme si l'enquête TIC ménages suivait un protocole séquentiel internet/téléphone. Nous gardons alors en tête que les résultats que nous obtenons ne permettent pas de tirer des conclusions quant au choix de la méthode la plus adaptée pour TIC. En effet, en choisissant d'imposer internet comme premier mode de collecte, on s'attend à amplifier l'effet de sélection, puisque le mode de collecte est précisément lié à ce que l'on cherche à mesurer.

## 4.2 Modélisation

Dans cet exemple, notre variable d'intérêt est l'indicatrice d'illectronisme. Celle-ci est déterminée par la réponse à plusieurs questions de l'enquête, concernant à la fois l'utilisation d'internet dans les 12 derniers mois et le niveau de compétences dans les domaines du numérique (que nous ne modéliserons pas ici).
Nous disposerons également de quatre variables socio-démographiques : l'âge, le sexe, le revenu disponible (ou niveau de vie), et le niveau de diplôme.
Enfin, nous considérerons deux variables inobservables : le niveau de défiance vis-à-vis d'internet, et le niveau global d'équipement du ménage (englobant l'équipement au sein du foyer et à proximité).

### 4.2.1 Choix des lois

Nous faisons les hypothèses suivantes concernant la modélisation des quatre premières variables :

- $Y \sim \mathcal{B}(0.15)$ : Le choix d'une loi de Bernoulli est assez naturel dans la mesure où il s'agit d'une indicatrice. Le choix du paramètre $p = 0,15$ provient du fait que le taux d'illectronisme au niveau

national est de 15%, selon une publication de l'Insee datant de 2019 [2].

- $X_{Age} \sim \mathcal{U}([15; 85])$ : Nous choisissons de nous limiter à une loi uniforme dans un souci de simplification, bien qu'une loi gamma aurait pu mieux refléter la structure de la population par âge.
- $X_{Sexe} \sim \mathcal{B}(0.5)$ : Le choix d'une loi de Bernoulli est naturel ici encore puisqu'on considère deux modalités.
- $X_{Revenu} \sim \mathcal{N}(23160, 8500)$ : Nous choisissons d'utiliser une loi normale centrée en 23160, correspondant au niveau de vie médian en 2021, bien qu'une loi gamma aurait aussi pu être envisagée.

Puis nous faisons le choix de considérer les trois variables suivantes, qui sont des variables de niveau, comme des scores. Aussi, la modélisation par des lois normales centrées réduites nous semble être plutôt adaptée dans la mesure où l'on peut utiliser le signe de ces variables pour distinguer les valeurs plutôt élevées des valeurs plutôt faibles. On aura donc :

- $X_{Diplome} \sim \mathcal{N}(0, 1)$ : Plus la valeur de $X_{Diplome}$ est élevée, plus l'individu en question a un niveau de diplôme élevé.
- $U_{DefianceInternet} \sim \mathcal{N}(0, 1)$ : Plus la valeur de $U_{DefianceInternet}$ est élevée pour un individu donné, plus celui-ci est défiant vis-à-vis d'internet.
- $U_{Equipement} \sim \mathcal{N}(0, 1)$ : De la même manière que précédemment, plus la valeur de $U_{Equipement}$ est élevée, plus l'individu dispose d'équipements numériques chez lui ou à proximité.

Enfin, nous avons choisi de modéliser les probabilités de réponse par des lois normales d'écart-type 0,19. Notons cependant que nous avons imposé une moyenne plus élevée pour $P_t$ afin d'être cohérent avec ce que nous avons pu lire à ce sujet.

- $P_i \sim \mathcal{N}(0.4, 0.19^2)$
- $P_t \sim \mathcal{N}(0.7, 0.19^2)$

### 4.2.2 Matrice de corrélations

On suppose qu'on obtient la matrice de corrélations suivante entre nos neuf variables :

$$Corr_{Y,X_{Age},X_{Sexe},X_{Revenu},X_{Diplome},U_{DefianceInternet},U_{Equipement},Pi,Pt} = \begin{pmatrix} 1 & 0.8 & 0.1 & 0.7 & 0.7 & 0.5 & -0.9 & -0,9 & 0 \\ . & 1 & 0.1 & 0.6 & -0.3 & 0.6 & -0.7 & -0.8 & -0.7 \\ . & . & 1 & -0.2 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & 1 & 0.6 & 0 & 0.5 & 0.5 & 0 \\ . & . & . & . & 1 & 0 & 0.6 & 0.5 & 0 \\ . & . & . & . & . & 1 & -0.7 & -0.9 & 0 \\ . & . & . & . & . & . & 1 & 0.2 & 0 \\ . & . & . & . & . & . & . & 1 & 0.5 \\ . & . & . & . & . & . & . & . & 1 \end{pmatrix}$$

Des éléments de justification concernant cette matrice sont donnés en annexe A.2. Nous pouvons cependant d'ores et déjà noter que nous supposons l'existence d'un biais de sélection non-ignorable, car la probabilité de réponse par internet est corrélée avec le niveau d'équipement et avec le niveau de défiance vis-à-vis d'internet, qui sont des inobservables.

### 4.2.3 Ajustement de la matrice de corrélations

La simulation des données avec le package *simstudy* à partir des lois de probabilités et de la matrice de corrélations entre les variables nécessite que celle-ci soit définie positive. Or la matrice de corrélations que nous avons présentée dans la section précédente ne vérifie pas cette condition. Nous avons donc utilisé la fonction *nearPD* du package R *Matrix*, qui permet de calculer la matrice définie positive la plus proche de la matrice saisie en argument.

2. BENDEKKICHE Hayet et VIARD-GUILLOT Louise, "15 % de la population est en situation d'illectronisme en 2021", *Insee Première*, No 1953, 22/06/2023.

Après avoir effectué cette correction, nous obtenons la matrice suivante pour la base de données simulée de taille N = 2000000 :

$$Corr_{Y,X_{Age},X_{Sexe},X_{Revenu},X_{Diplome},U_{DfianceInternet},U_{Equipement},Pi,Pt} = \begin{pmatrix} 1 & 0.5 & 0 & 0.3 & 0.2 & 0.4 & -0.4 & -0.3 & -0.1 \\ . & 1 & 0 & 0.3 & -0.1 & 0.6 & -0.6 & -0.7 & -0.5 \\ . & . & 1 & -0.1 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & 1 & 0.6 & -0.1 & 0.2 & 0.2 & 0 \\ . & . & . & . & 1 & -0.1 & 0.4 & 0.3 & 0.1 \\ . & . & . & . & . & 1 & -0.6 & -0.8 & 0 \\ . & . & . & . & . & . & 1 & 0.5 & 0 \\ . & . & . & . & . & . & . & 1 & 0.5 \\ . & . & . & . & . & . & . & . & 1 \end{pmatrix}$$

Globalement, on observe que cette correction a tendance à diminuer l'importance des corrélations que nous avions fixées, à quelques exceptions près. Ceci est particulièrement marqué pour les corrélations entre la variable d'intérêt $Y$ et les autres variables, que nous avions quasiment toutes supposées comme étant plutôt fortes. A l'inverse on peut relever que la corrélation entre la probabilité de réponse par internet et le niveau d'équipement a été augmentée par cette correction, passant ainsi de 0,2 à 0,5. Cependant, le signe de toutes les corrélations est conservé, ce qui est rassurant.
En revanche on peut noter que certaines corrélations que nous avions supposées nulles entre la probabilité de réponse par téléphone et d'autres variables sont désormais significativement différentes de 0. C'est le cas des corrélations entre $P_t$ et $Y$, et entre $P_t$ et $X_{Diplome}$. Une faible corrélation négative (de l'ordre de -0,1) que nous avions supposée nulle, est également ajoutée entre les variables $U_{Defiance}$ et $X_{Revenu}$ ainsi qu'entre les variables $U_{Defiance}$ et $X_{Diplome}$. Enfin, on relève que les corrélations que nous avions supposées égales à 0,1 sont désormais nulles. Cette nouvelle matrice conduit donc à supposer qu'il n'y a pas de corrélations entre le sexe et le fait d'être en situation d'illectronisme, ni entre le sexe et l'âge.

Ainsi, on conclut que cette correction ne diffère que légèrement des suppositions que nous avions faites. Nous simulons donc notre population de 2 000 000 d'individus à partir des lois présentées dans la partie 4.2.1 et de cette matrice corrigée. A ce stade, nous pouvons calculer le total exact de la variable $Y$ sur l'ensemble de la population : on obtient un total de 300 971, ce qui donne bien un taux d'illectronisme de 15%.

### 4.2.4 Tirage de l'échantillon et ajout de non-réponse

Pour le tirage de l'échantillon des 10 000 individus enquêtés, nous procédons selon un sondage aléatoire simple comme présenté dans la partie 2.3. Cependant, la manière dont nous choisissons d'ajouter de la non-réponse est légèrement différente de celle que nous avons présentée dans cette partie. Ainsi, nous faisons les deux hypothèses suivantes :

1. Si un individu fait partie de 10% les plus défiants vis-à-vis d'internet, alors il est exclu qu'il réponde via internet.
2. Si un individu fait partie des 10% les moins équipés, alors il est également exclu qu'il réponde via internet.

A l'issue de cette procédure, nous obtenons les effectifs suivants de répondants par mode et de non-répondants :

| Internet | Téléphone | Non-répondants |
|---|---|---|
| 3687 | 4266 | 2047 |

FIGURE 2 – Effectifs de répondants par mode et de non-répondants

De plus, le taux de réponse global est de 80%.

## 4.3 Résultats

Les méthodes de redressement dont nous avons cherché à évaluer la performance sont les suivantes :

- Estimateur de Horvitz-Thompson sur les répondants.
- Modèle linéaire en régressant Y sur l'âge et le revenu.
- Modèle homogène par strate (avec deux strates).
- Modèle de repondération sans les modes en estimant le comportement de réponse à partir de l'âge, et du revenu et en constituant les groupes homogènes de réponse par CAH.
- Modèle de repondération séparée des deux modes en estimant le comportement de réponse par internet et par téléphone à partir de l'âge et du revenu, et en constituant les groupes homogènes de réponse par CAH.
- Modèle de repondération pour les répondants téléphone uniquement (non-réponse portée par les répondants téléphone) en estimant le comportement de réponse par téléphone à partir de l'âge et du revenu, et en constituant les groupes homogènes de réponse par CAH.
- Méthode de repondération prenant en compte les deux modes, en estimant le comportement de réponse par internet à partir de l'âge et du revenu, puis le comportement de réponse téléphone des non-répondants internet grâce à l'âge et au revenu également. Les groupes homogènes de réponses sont constitués par CAH.
- Modèle de Heckman en une étape en estimant la fonction de réponse implicite et Y à partir de l'âge et du revenu.
- Modèle de Heckman en deux étapes en estimant la fonction de réponse implicite et Y à partir de l'âge et du revenu.

On obtient les résultats visibles dans le tableau ci-dessous :

| Méthode | Biais | Variance | MSE | RMSE |
|---|---|---|---|---|
| Homogène par strates (2 strates) | -2349 | 60622991 | 66140558 | 8133 |
| Repondération deux modes | 393 | 72714461 | 72868822 | 8536 |
| Repondération sans modes | 2508 | 72310178 | 78599804 | 8866 |
| Linéaire | 5578 | 55605377 | 86718906 | 9312 |
| Heckman 1 étape | 12365 | 188159726 | 341058053 | 18468 |
| Repondération que téléphone | 21524 | 75280011 | 538583901 | 23207 |
| Repondération séparée deux modes | -26255 | 70210299 | 759542261 | 27560 |
| Horvitz-Thompson sur répondants | -30496 | 58773661 | 988791097 | 31445 |
| Heckman 2 étapes | 115098 | 1070947072 | 14318395286 | 119659 |

FIGURE 3 – Résultats des méthodes de redressements pour l'exemple TIC par ordre croissant de RMSE

On constate que la meilleure méthode de redressement du point de vue du biais est la méthode de repondération prenant en compte les deux modes (estimation des probabilités de réponse par internet sur tout l'échantillon, puis estimation des probabilités de réponse par téléphone à partir des non-répondants internet). Cette méthode est également celle qui donne la deuxième erreur quadratique moyenne la plus faible, après la méthode homogène par strates dont le biais est bien plus élevé. Il est intéressant de noter que les méthodes de Heckman ne font pas partie des méthodes les plus performantes, alors même que nous avions supposé un biais de sélection non-ignorable et que ces méthodes sont justement utilisées pour corriger ce type de biais. Ceci semble donc indiquer qu'il n'y aurait pas de réel biais de sélection non-ignorable dans notre exemple.

Concernant l'enquête TIC ménages, une expérimentation en face-à-face a eu lieu au quatrième trimestre de 2021 dans le but de mettre en évidence l'existence d'un potentiel biais (de sélection en particulier) [3]. L'utilisation d'un protocole en face-à-face permettant d'obtenir de meilleurs taux de réponse, et présentant l'avantage de proposer un mode de collecte qui n'est pas en lien avec le sujet de l'enquête, les auteurs ont

3. LEGLEYE Stéphane, VIARD-GUILLOT Louise, NOUGARET Amandine, "Correction des effets de mode et biais de sélection : les apports d'une expérimentation de l'enquête TIC (Technologies de l'Information et de la Communication) en face-à-face", *Journées de méthodologie statistique de l'Insee*, 2022

pu mettre en évidence un biais de sélection relativement faible (et légèrement plus élevé chez les personnes âgées). En particulier, les auteurs relèvent qu'il n'y a pas de fort biais de sélection au sein du sous-échantillon internet-papier pour lequel on ne dispose pas des coordonnées téléphoniques ou électroniques des individus interrogés, contrairement à ce à quoi on pourrait d'attendre.

Ainsi, bien que nos résultats ne soient pas directement comparables car nous considérons un protocole de collecte différent de celui de l'enquête TIC classique, ces éléments vont dans le même sens que nos observations, ce qui nous conforte dans l'idée qu'il pourrait *in fine* ne pas y avoir d'effet de sélection non-ignorable dans notre exemple.

# 5 Un second exemple d'application : l'enquête EpiCov

## 5.1 Présentation de l'enquête

EpiCov, pour Epidémiologie et Conditions de Vie, est une enquête Inserm - Drees (respectivement l'Institut national de la santé et de la recherche médicale, et la Direction de la recherche, des études, de l'évaluation et des statistiques du ministère des Solidarités et de la Santé). L'Insee et Santé publique France ont apporté leur expertise méthodologique.

EpiCov a été mise en place dès le premier confinement lié au Covid-19, en mars 2020. Son caractère innovant tient aux informations collectées : il y a eu la combinaison d'un questionnaire et d'un autotest sérologique. L'objectif était d'étudier les conséquences de la pandémie et du confinement sur la santé générale (physique, mentale) et sur les conditions de vie de la population française. Ces préoccupations sanitaires et sociales ont d'ailleurs conduit à sur-représenter trois caractéristiques dans l'échantillon : (1) les zones géographiques très touchées par le Covid-19, (2) les départements peu peuplés et (3) les ménages à bas revenu. Dans un souci de simplicité et pour ne pas choisir des valeurs arbitraires, nous avons choisi de procéder à un sondage aléatoire simple dans notre modélisation, mais c'est là une piste d'approfondissement de notre travail.

Plus précisément, l'échantillon choisi pour EpiCov couvre l'ensemble du territoire français (métropole et Drom). Là encore, nous avons décidé de nous restreindre au territoire métropolitain dont l'échantillon était constitué de 350 000 individus. EpiCov a connu plusieurs vagues : les questionnaires étaient sensiblement différents selon l'évolution du contexte sanitaire et devaient permettre de suivre l'évolution des conditions de vie. Nous nous sommes concentrées sur la première vague, qui correspond à mai 2020. Le taux de réponse a été de 37,6%, soit à peu près 130 000 personnes.

Ce taux varie considérablement selon le mode de collecte utilisé : 35,3% pour les lots internet et de 46,7% pour les lots multimodes. En effet, les répondants à l'enquête ont été répartis en vingt lots. Les quatre premiers lots se sont vus proposer du multimode, soit séquentiel avec une relance téléphonique, soit concurrentiel. Les autres lots ne pouvaient répondre que par internet. Ces différences de taux de réponse nous ont convaincues qu'EpiCov était un exemple intéressant pour notre projet.

Nous pouvons également mentionner que deux questionnaires, un court et un long ont été testé. Parmi les quatre lots multimodes le premier lot a reçu le questionnaire long. Cependant, le temps de réponse moyen diffère seulement d'une dizaine de minutes donc nous avons décidé de ne pas prendre en compte ce facteur.

## 5.2 Modélisation

Pour cette application, nous avons choisi d'étudier un score de santé générale ressentie (mêlant donc santé physique, mentale, voire financière). Nous avons sélectionné quatre variables socio-démographiques disponibles à partir de Fidéli, la base de sondage utilisée pour EpiCov, et qui nous semblaient pertinentes : le sexe, l'âge, le revenu et le lieu de vie. S'ajoutent ensuite deux inobservables : le degré d'adhésion à des théories complotistes et les prédispositions liées à l'état de santé (étant donné que des facteurs à risque tels l'obésité morbide avaient été identifiés).

### 5.2.1 Choix des lois

Concernant la modélisation statistique, nous avons repris les mêmes lois pour les variables `age`, `sexe`, et `revenu` que pour l'exemple TIC. Voici les hypothèses de modélisation que nous avons faites pour les variables choisies :

Pour la variable d'intérêt :

— $Y \sim \mathcal{N}(0,1)$ : Il s'agit là d'une loi qui facilite les modélisations et les analyses statitiques, mais qui peut aussi refléter les disparités de vécu de la pandémie.

Pour les variables observées :

— $X_{Sexe} \sim \mathcal{B}(0.5)$

— $X_{Age} \sim \mathcal{U}([15; 85])$

— $X_{Revenu} \sim \mathcal{N}(23160, 8500)$

— $X_{LieuVie} \sim \mathcal{B}(0, 08)$ : Pour modéliser le lieu de vie, nous avons décidé de créer une variable qui prend la valeur 1 si l'individu habite dans un Quartier prioritaire de la Ville (QPV) et 0 sinon. Cette information est disponible dans Fidéli et est un facteur distinctif dans les résultats de l'enquête EpiCov. 8% de la population métropolitaine habite un QPV.

Pour les inobservées :

— $U_{Complotisme} \sim \mathcal{B}(0, 3)$. Il est estimé qu'un tiers des Français adhère à au moins une théorie du complot (source Ifop, 2023). Bien sûr, la nature de l'objet étudié et le mode de collecte peuvent faire l'objet de questionnements, mais il sagit là d'une approximation.

— $U_{PredispositionsRisquesSante} \sim \mathcal{B}(0, 15)$. Nous avons choisi ce paramètre après s'être renseignées sur les taux d'obésité et de cancer en France, deux facteurs identifiants des personnes atteintes de comorbidité lors de la pandémie. Là encore, il s'agit d'approximer une valeur.

Pour les probabilités de réponse :

— $P_i \sim \mathcal{N}(0.35, 0.19^2)$

— $P_t \sim \mathcal{N}(0.73, 0.19^2)$

### 5.2.2 Matrice de corrélations

Voici la matrice de corrélation que nous proposons. Les choix de ces corrélations se sont faits suite à la lecture d'articles de recherche.

$$Corr_{Y,X_{Age},X_{Sexe},X_{Revenu},X_{LieuVie},U_{Complotisme},U_{PredispositionsRisquesSante},Pi,Pt} = \begin{pmatrix} 1 & 0.6 & 0.3 & 0.7 & 0.7 & 0.5 & 0.9 & 0.6 & 0.5 \\ . & 1 & 0.1 & 0.7 & 0.3 & -0.3 & 0.7 & -0.8 & -0.7 \\ . & . & 1 & 0.7 & 0 & 0 & -0.1 & 0 & 0 \\ . & . & . & 1 & -0.8 & -0.4 & 0.4 & 0.2 & 0 \\ . & . & . & . & 1 & 0.2 & 0.2 & -0.1 & -0.1 \\ . & . & . & . & . & 1 & 0.1 & -0.1 & 0 \\ . & . & . & . & . & . & 1 & -0.2 & -0.2 \\ . & . & . & . & . & . & . & 1 & 0.5 \\ . & . & . & . & . & . & . & . & 1 \end{pmatrix}$$

**Corrélations avec la variable d'intéret Y, représentant un score de santé**

Les personnes de 80 ans et plus avaient cinq fois plus de chances d'avoir un Covid-19 grave. Les personnes âgées étaient donc considérées comme particulièrement à risques, d'autant plus qu'il s'agit souvent de personnes isolées. Cependant, les jeunes ont aussi beaucoup souffert mentalement et financièrement de cette crise [4], ce qui explique le choix d'une corrélation de 0.6, et pas d'un niveau plus élevé.

La corrélation entre Y et la variable de sexe traduit la gestion du quotidien qui a incombé aux femmes. Par exemple, 80% des femmes passaient plus de quatre heures par jour avec les enfants contre 52% des hommes [5]. De même, 85% des familles monoparentales sont celles d'une mère, et ces familles ont connu des difficultés financières considérables.

Concernant les corrélations avec le revenu, nous pouvons évoquer le fait que 37% des individus interrogés appartenant au premier quintile de revenu ont jugé le confinement pénible, contre 17% pour le dernier quintile[4]. De même, la séroprévalence était presque deux fois plus élevée en QPV que sur l'ensemble du territoire

4. HAZO J.-B., COSTEMALLE V. : « Confinement du printemps 2020 : une hausse des syndromes dépressifs, surtout chez les 15-24 ans. Résultats issus de la 1re vague de l'enquête EpiCov et comparaison avec les enquêtes de santé européennes (EHIS) de 2014 et 2019 » *Etudes et Résultats*, Drees, 2021

5. ALBOUY V., LEGLEYE S. 2020 « Conditions de vie pendant le confinement : des écarts selon le niveau de vie et la catégorie socioprofessionnelle », *Insee Focus* 197

(8,2% contre 4,2% respectivement) [6]

Pour les variables inobservables, il y avait peu de chiffres disponibles concernant le complotisme. Nous avons estimé que le fait d'adhérer à des théories du complot pouvait autant réfréner les individus de répondre que les enjoindre à partager leur ressenti sur une situation modulée par leurs spéculations. Pour les prédispositions augmentant les risques de comorbidité, nous avons logiquement choisi une corrélation élevée.

Enfin, concernant les probabilités de réponse selon le mode de collecte nous avons repris les mêmes que pour TIC.

**Corrélations avec l'âge**

Nous n'avons pas mis à 0 la corrélation entre âge et sexe après étude de la pyramide des âges par sexe, mais celle-ci reste peu élevée.

Nous avons choisi une corrélation évidemment élevée entre l'âge et le revenu, mais faible pour l'âge et le lieu de vie : nous savons seulement que les jeunes sont un peu sur-représentés en QPV par rapport au territoire national.

La corrélation négative l'âge et le degré d'adhésion aux croyances complotistes renvoie au fait que 28% des 18-24 ans adhèrent à cinq théories ou plus, contre seulement 9% des 65 ans et plus [7]. La corrélation entre l'âge et les facteurs à risques liés au Covid-19 est évidemment positive, bien que l'obésité touche de plus en plus les jeunes.

Enfin, nous avons repris les mêmes corrélations entre l'âge et les probabilités de réponse que pour l'enquête TIC.

**Corrélations avec le sexe**

Nous avons choisi une forte corrélation avec le revenu pour refléter les inégalités financières qui se sont accrues lors de la pandémie. La corrélation concernant les prédispositions de santé renvoie à des indicaterurs de santé liés au sexe [8].

En revanche, nous avons trouvé peu d'information sur les liens entre le sexe et les probabilités de réponse, ou les QPV, ou l'adhésion aux théories du complot ; c'est pourquoi nous avons choisi de mettre les corrélations concernées à 0.

**Corrélations avec le revenu**

De façon logique, nous avons choisi une corrélation négative élevée entre le revenu et le fait d'habiter en QPV. En effet, par construction les habitants des QPV sont plus fragiles financièrement. Par exemple, ils sont moins en emploi et lorsqu'ils le sont il s'agit plus souvent d'emplois à temps partiel ou limité [9].

La corrélation avec le degré d'adhésion à des théories du complot est négative et élevée : parmi les Français croyant à cinq théories ou plus, 38% sont de catégorie sociale "pauvre" et ce chiffre est décroissant pour atteindre 7% de personnes issues de catégories aisées[6].

La corrélation entre le revenu et les facteurs à risques reflète les inégalités financières qui peuvent conduire à des choix d'alimentation déifférenciés selon la catégorie sociale, et les inégalités d'accès aux soins [7].

Encore une fois nous avons utilisé les mêmes probabilités de réponse que pour la modélisation TIC.

**Corrélations avec le lieu de vie**

6. WARSZAWSKI J. et al, 2020 «En mai 2020, 4,5% de la population vivant en France métropolitaine a développé des anticorps contre le SARS-CoV-2»

7. Enquête Complotisme 2019, Fondation Jean Jaurès `https://www.jean-jaures.org/publication/enquete-complotisme-2019-les-grands-enseignements/`

8. Haute Autorité de Santé «Sexe, Genre et Santé» *Rapport d'analyse prospective 2020*

9. S DURIEUX, P ROUAUD (Insee), R BELLE (Drees), 2020 «Dans les quartiers les plus en difficulté, seulement un habitant sur trois en emploi»

La corrélation positive avec la variable reflétant l'adhésion à des théories complotistes renvoie aux caractéristiques des populations QPV qui sont plus fragiles financièrement, et sur-représentées parmi les adhérents au complotisme.

La corrélation avec la variable des facteurs à risque renvoie ici encore aux inégalité de santé et de soins.

Cocnernant les probabilités de réponse, nous avons suivi l'hypothèse que les populations plus défavorisées sont moins sensisbles aux enquêtes et répondent moins.

**Corrélations avec le degré d'adhésion à des théories du complot**
Nous avons corrélé positivement l'adhésion à des théories du complot et le fait d'avoir un état de santé présentant plus de facteurs risques. Cette hypothèse nous est parue d'autant plus valide compte tenu de la thématique sanitaire de l'étude. Par ailleurs, les populations défavorisées et qui subissent des inégalités d'accès au soins sont surreprésentées parmi les croyants à des théories du complot.

Nous avons fait l'hypothèse d'une corrélation faiblement négative entre adhésion à des théories du complot et réponse sur internet, en raison de possibles croyances ou défiances justement. Nous n'avions cependant pas réellement d'informations concernant les croyants à des théories du complot et leurs taux de réponse, c'est pourquoi ces corrélations sont relativement faibles voire nulles.

**Corrélations avec les prédispositions de santé**
Enfin, nous avons négativement corrélé le fait de présenter des facteurs à risques et de répondre à l'enquête simplement parce que ces personnes étaient plus susceptibles d'être malades et donc de ne pas être en mesure de répondre. Cependant, cette corrélation est faible puisqu'au contraire cela a pu éveiller leur curiosité.

### 5.2.3 Ajustement de la matrice de corrélations

Lors de la simulation des données sur R, nous avons rencontré le même problème que pour l'enquête TIC : la matrice de corrélations que nous avions établie n'était pas définie positive. Nous avons donc procédé de la même façon pour obtenir la matrice ci-dessous, dont les coefficients ont été arrondis au dixième :

$$Corr_{Y,X_{Age},X_{Sexe},X_{Revenu},X_{LieuVie},U_{Complotisme},U_{PredispositionsRisquesSante},Pi,Pt} = \begin{pmatrix} 1 & 0.4 & 0.3 & 0.4 & 0.4 & 0.3 & 0.7 & 0.4 & 0.3 \\ . & 1 & 0.2 & 0.5 & 0.2 & -0.2 & 0.7 & -0.6 & -0.6 \\ . & . & 1 & 0.6 & -0.1 & 0 & -0.1 & 0.1 & 0 \\ . & . & . & 1 & -0.5 & -0.3 & 0.4 & 0.1 & 0 \\ . & . & . & . & 1 & 0.3 & 0.3 & -0.1 & -0.1 \\ . & . & . & . & . & 1 & 0.1 & 0 & 0.1 \\ . & . & . & . & . & . & 1 & -0.1 & -0.2 \\ . & . & . & . & . & . & . & 1 & 0.6 \\ . & . & . & . & . & . & . & . & 1 \end{pmatrix}$$

Les signes des corrélations sont les mêmes que pour la matrice que nous avions créée dans la sous-section 5.2.2 précédente. Concernant les valeurs, trois corrélations diffèrent de plus de 0.2 point : la corrélation entre la variable d'intérêt et le revenu, celle entre la variable d'intérêt et le lieu de vie, et celle entre le revenu et le lieu de vie. Ce sont donc les trois mêmes variables qui sont concernées. Pour les deux premières corrélations, notre hypothèse de base surestime les coefficients par rapport à la matrice définie positive, et pour la denière nous étions en situation de sous-estimation. Cependant, les valeurs obtenues avec la nouvelle matrice coincident avec les hypothèses générales que nous avions émises.

## 5.3 Résultats

Nous avons testé treize méthodes différentes de redressement. Pour toutes ces méthodes, nous avons choisi la combinaison de variables explicatives suivante : $X_1 + X_3 + X_4$. Cette combinaison fait sens au regard de

l'objet d'étude et il s'agit des variables avec les plus fortes corrélations dans la matrice ajustée. Ci-dessous, le tableau avec les différentes méthodes et les résultats correspondants :

| Méthode | Biais | Variance | MSE | RMSE |
|---|---|---|---|---|
| **Heckman 1 étape** | 61 906.17 | 1 827 677 713 | 5 660 051 049 | 75 233.31 |
| **Heckman 2 étapes** | 79 174.11 | 1 867 067 683 | 8 135 606 781 | 90 197.60 |
| **Repondération que téléphone - Kmeans** | 119 959.54 | 468 723 449 | 14 859 014 655 | 121 897.56 |
| **Repondération que téléphone - Quantiles** | 121 278.73 | 500 839 420 | 15 209 369 657 | 123 326.27 |
| **Horvitz-Thompson sur répondants** | 131 669.17 | 516 326 060 | 17 853 097 624 | 133 615.48 |
| **Repondération deux modes - Kmeans** | 150 516.35 | 478 134 319 | 23 133 305 192 | 152 096.37 |
| **Repondération sans modes - Kmeans** | 151 622.02 | 460 994 421 | 23 450 231 260 | 153 134.68 |
| **Repondération deux modes - Quantiles** | 154 377.69 | 484 548 570 | 24 317 020 268 | 155 939.16 |
| **Repondération sans modes - Quantiles** | 154 748.30 | 484 733 446 | 24 431 770 884 | 156 306.66 |
| **Linéaire** | 158 328.53 | 353 889 509 | 25 421 812 828 | 159 442.19 |
| **Repondération séparée 2 modes - Kmeans** | 160 836.25 | 511 172 296 | 26 379 471 265 | 162 417.58 |
| **Repondération séparée 2 modes - Quantiles** | 165 407.28 | 521 082 680 | 27 880 649 589 | 166 975.00 |
| **Homogène par strates (2 strates)** | 192 465.48 | 476 871 863 | 37 519 833 221 | 193 700.37 |

Figure 4 – Résultats des méthodes de redressement pour l'exemple EpiCov par ordre croissant de l'erreur quadratique moyenne (RMSE)

Si les chiffres restent grands, les méthodes de redressement type Heckman, que ce soit en une ou deux étapes, sont bien plus efficaces que les autres. En effet, le biais est divisé par plus de trois et l'erreur quadratique moyenne également, comparativement à la méthode homogène par strates. Ces résultats corroborent avec les hypothèses présentées dans la partie 3 sur les Méthodes de redressement.

# 6 Enjeux liés à la définition des corrélations

Dans les deux exemples présentés dans les parties 4 et 5 (et plus globalement dans toutes nos simulations) se pose un problème pratique. En effet, le but de notre travail a été de concevoir un outil permettant à un chargé d'enquête de pouvoir identifier dans quel cas il se trouve vis-à-vis d'un potentiel effet de sélection, et de pouvoir alors adopter la méthode de redressement la plus adaptée à sa situation. La démarche de ce chargé d'enquête consiste donc en premier lieu à observer quelles sont les corrélations entre la variable d'intérêt de son enquête et les variables observables, de manière à pouvoir simuler une population permettant d'obtenir les mêmes corrélations. Cependant, plusieurs façons de générer les données pourraient permettre d'obtenir les corrélations qu'il observe *in fine*, et c'est précisément ici que se situe notre problème. Nous n'avons pas réussi à identifier des éléments permettant de faciliter la recherche du paramétrage menant à l'obtention d'un échantillon aux caractéristiques données.

Un autre problème que nous avons pu identifier dans la mise en oeuvre pratique de notre protocole est que lorsque nous construisons la matrice de corrélations, cette dernière est rarement définie positive, et ne peut donc pas être traitée en l'état. Bien que nous ayons observé que la correction apportée par la fonction *near_PD* du package *Matrix* ne change pas drastiquement les corrélations (plus particulièrement nous n'avons observé aucun changement de signe), la modification de la matrice de corrélations initiale pourrait ajouter de l'imprécision à des paramètres déjà choisis arbitrairement.

# Conclusion

Au cours de ce projet nous avons uniquement étudié le cas des enquêtes multimodes, mais il est important de noter que le biais de sélection intervient en fait dans la plupart des enquêtes. La spécificité du multimode n'est donc pas de présenter plus de biais de sélection, mais d'en présenter un dont les conséquences peuvent prendre plus d'ampleur.

A première vue et après avoir lu le présent rapport, on peut notamment remarquer qu'au global il y a moins de biais de sélection dans les enquêtes multimodes que dans les enquêtes monomodes, car en laissant plus de possibilités aux individus on accroît dans le même temps la représentativité des données. Mais cette affirmation passe sous silence le fait que les mécanismes de sélection sont différents au sein de chaque mode de collecte : par exemple, les personnes âgées répondent plus souvent par téléphone que les personnes jeunes (cf partie 1.2). Or pour pouvoir appréhender un biais de mesure éventuel il ne faut pas que les répondants internet aient un profil trop différent des répondants téléphone. Il n'est donc pas envisageable de traiter le biais de sélection dans les enquêtes multimodes avec le protocole habituel.

Ici nous avons laissé de côté le biais de mesure car il n'existe pas encore de méthode permettant de corriger simultanément les deux biais. Nous avons implémenté plusieurs corrections du biais de sélection (cf partie 3) et nous proposons finalement un outil permettant d'évaluer leur pertinence sur des données d'enquête fictives. A condition de simuler des données suffisamment proches des données d'enquête réelles, notre travail permet alors à l'utilisateur d'identifier les méthodes de redressement qui lui seront utiles.

Néanmoins, le choix des paramètres de la simulation (cf partie 2.3) et le choix des corrélations entre les variables n'est pas évident. Au delà des limites théoriques et computationnelles (présentées notamment dans la partie 6) cela nécessite aussi une bonne connaissance du sujet, d'où notre choix de présenter deux exemples concrets avec l'enquête TIC (cf partie 4) et l'enquête EpiCov (cf partie 5). Les résultats obtenus sont cohérents avec les conclusions externes, ce qui nous rassure sur la qualité de l'outil proposé et laisse la porte ouverte à des prolongements. En particulier, avant de combiner nos analyses avec l'étude individuelle du biais de mesure, il serait intéressant d'identifier les conditions à respecter dans les matrices de corrélation pour qu'une méthode soit adaptée.

# A Annexes

## A.1 Accès au code

Le code associé au projet est accessible via le lien suivant : https ://github.com/Cleo-BH/ENSAI-projet-methodologique-3A.

Les fichiers sont organisés en six catégories afin que l'utilisateur puisse personnaliser son étude et/ou ajouter de nouvelles fonctionnalités facilement. Il paraît important de les utiliser dans l'ordre :

- *Catégorie #0 :* Le fichier "Requirements" contient l'ensemble des packages utilisés par la suite et qu'il convient de charger dès le départ.
- *Catégorie #1* : Les fichiers définissent le contexte de la simulation (taille de la base de sondage, nombre de variables et lois associées, matrice de corrélation) et permettent quelques contrôles. Le code initial contient les points de départ pour les trois exemples théoriques ainsi que pour les démonstrations TIC et Epicov.
- *Catégorie #2* : Elle contient les fonctions permettant de définir un échantillon. Pour l'instant il n'y a qu'une seule option qui est le plan de sondage aléatoire simple sans remise.
- *Catégorie #3* : Elle contient les fonctions permettant de mettre en place la non-réponse. Nous fournissons le code associé à la partie 2.3 (non-réponse proportionnelle à $P_i$ et $P_t$) mais également le cas particulier de l'enquête TIC.
- *Catégorie #4* : Il y a un code pour chaque méthode présentée dans la partie 3.
- *Catégorie #5* : Le fichier "Evaluation d'une méthode.R" présente la méthode générique permettant de connaître le biais empirique, la variance empirique et l'erreur quadratique commise par une certaine fonction et dans un certain contexte. Il est accompagné de trois exemples théoriques et des codes utilisés pour les parties 4 et 5.

## A.2 Justification de la matrice de corrélation pour l'exemple sur l'enquête TIC ménages

Pour justifier les corrélations de la variable Y avec les autres variables, nous nous sommes basées sur deux articles publiés par l'Insee sur le sujet [10]. Ces références nous ont permis d'identifier le signe des corrélations entre les variables, cependant le choix des valeurs reste arbitraire. En revanche, nous avons été vigilantes au fait que ces dernières reflètent l'ordre d'importance des facteurs expliquant l'illectronisme. Ainsi, dans la mesure où les deux principaux facteurs explicatifs de cette variable sont l'âge (les individus âgés ont beaucoup plus tendance à être en situation d'illectronisme) et l'équipement (si un individu n'est pas équipé, il a de grandes chances d'être en situation d'illectronisme), nous avons choisi de fortes corrélations entre Y et ces deux variables. Un revenu et un niveau de diplôme élevés diminuent le risque d'être en situation d'illectronisme dans des proportions similaires, cependant ces facteurs sont moins décisifs que l'âge et le niveau d'équipement, d'où des valeurs plus faibles pour les corrélations.
Nous ne disposons pas réellement d'informations quant à l'effet de la défiance vis-à-vis d'internet sur le fait d'être en situation d'illectronisme ou non, cependant on peut légitimement penser que les individus méfiants quant à internet auraient tendance à ne pas l'utiliser et seraient ainsi plus souvent en situation d'illectronisme. Il faut en revanche tenir compte du fait que de nombreuses démarches se font désormais exclusivement en ligne, ce qui pourrait contraindre ces individus à utiliser internet au moins une fois dans l'année. Nous faisons donc le choix de fixer la corrélation entre $Y$ et $DefianceInternet$ à un niveau intermédiaire (0,5).
Enfin, il semble logique de supposer une forte corrélation négative entre $Y$ et la probabilité de répondre par internet. Le comportement de réponse par téléphone ne semble a priori pas avoir de raison d'être lié à la situation d'illectronisme. Nous choisissons donc de fixer cette corrélation à 0.

Plus globalement, on ne dispose que de peu d'informations concernant les corrélations entre la probabilité de réponse par téléphone et les autres variables. Aussi, on supposera qu'il existe uniquement deux corrélations non nulles entre $P_t$ et les autres variables. La première est une corrélation négative de $P_t$ avec l'âge, car il a été identifié que les personnes âgées avaient moins tendance à répondre aux enquêtes que les plus jeunes.

10. Cf. note précédente, et LEGLEYE Stéphane et ROLLAND Annaïck, "Une personne sur six n'utilise pas Internet, plus d'un usager sur trois manque de compétences numériques de base", *Insee Première*, No 1780, 30/10/2019.

Ceci devrait donc être le cas également si on considère la réponse par téléphone en particulier. La seconde corrélation non-nulle est celle entre les deux probabilités de réponse $P_i$ et $P_t$. Comme précédemment, nous la justifions par le fait que ces dernières sont vraisemblablement liées à un comportement de réponse plus global.

Les éléments permettant de définir les corrélations entre la probabilité de réponse par internet et les autres variables sont plus nombreux. Nous savons ainsi que les jeunes ont des taux de réponse par internet plutôt importants que les plus âgés, et qu'il en va de même pour les personnes diplômées du baccalauréat ou de l'enseignement supérieur par rapport au personnes ayant un niveau de diplôme inférieur ou pas de diplôme. Cette corrélation positive entre le niveau de diplôme et la probabilité de réponse par internet nous laisse penser que la corrélation entre cette dernière et le revenu devrait être assez similaire. En revanche, on suppose qu'il n'y a pas de corrélation avec le sexe. Enfin, concernant les corrélations entre $P_i$ et nos inobservables, il semble logique que la probabilité de réponse par internet soit d'autant plus faible que la défiance vis-à-vis d'internet est élevée, et que le niveau d'équipement soit positivement lié avec $P_i$.

Pour fixer les corrélations entre les variables socio-démographiques, nous nous sommes documentées sur la structure de la population.Nous ne possédons en revanche que de peu de documentation sur les liens existant entre le niveau de défiance vis-à-vis d'internet et nos variables socio-démographiques observées. Nous les avons toutes fixées à 0, à l'exception de celle existant avec la variable $Y$ que nous avons justifiée précédemment, et de celle avec l'âge que nous supposons être positive. En effet, les personnes les plus âgées n'ayant pas grandi dans une société ou le numérique était aussi présent qu'aujourd'hui, il est légitime de penser qu'elles soient plus réticentes à l'utiliser. La raisonnement inverse peut d'ailleurs s'appliquer aux individus les plus jeunes.

Enfin, les deux articles publiés par l'Insee en 2019 et 2023 que nous évoquions précédemment nous permettent d'établir les corrélations existant entre le niveau d'équipement et nos variables socio-démographiques (notons cependant ici encore que le choix précis des valeurs reste arbitraire). Selon ces articles, les personnes les plus âgées, les moins diplômées et les plus modestes ont moins souvent accès à internet.

# B Bibliographie

## B.1 Documents relatifs à la méthodologie

KOZLOWSKI Louise, "A la recherche de la bonne méthode de repondération / Un enjeu majeur pour les enquêtes multimodes à l'Insee", Mémoire de stage, 16/07/2022.

CASTELL Laura et SILLARD Patrick, "Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman", Document de travail, 03/2021.

## B.2 Documents relatifs à l'enquête TIC

BENDEKKICHE Hayet et VIARD-GUILLOT Louise, "15 % de la population est en situation d'illectronisme en 2021", *Insee Première*, No 1953, 22/06/2023.

LEGLEYE Stéphane et ROLLAND Annaïck, "Une personne sur six n'utilise pas Internet, plus d'un usager sur trois manque de compétences numériques de base", *Insee Première*, No 1780, 30/10/2019

Outil interactif en ligne produit par l'Insee, *Tableau de bord de l'économie française*, "Revenu-Niveau de vie-Pouvoir d'achat", `https://www.insee.fr/fr/outil-interactif/5367857/tableau/30_RPC/31_RNP#`, consulté le 10/02/2024

Insee, "France, portrait social" - édition 2019", *Insee références*, 19/11/2019, pp.212-213

Observatoire des inégalités, "A travail égal, salaire égal ?",
`https://www.inegalites.fr/femmes-hommes-salaires-inegalites`, consulté le 10/02/2024

Insee, DSDS, Présentation "Taux de réponse par internet", Séminaire DSDS, 2021, `https://intranet.insee.fr/jcms/1565885_DBFileDocument/fr/seminaire-dsds-taux-de-reponse-par-internet?fbclid=IwAR09UPsMvcBsmKto5U8OBJt-t3ItyZe-TIlkOYMGmosf1gEKaRFU_8G-njI`, consulté le 19/02/2024

LEGLEYE Stéphane, VIARD-GUILLOT Louise, NOUGARET Amandine, "Correction des effets de mode et biais de sélection : les apports d'une expérimentation de l'enquête TIC (Technologies de l'Information et de la Communication) en face-à-face", *Journées de méthodologie statistique de l'Insee*, 2022

## B.3 Documents relatifs à l'enquête EpiCov

Site de l'enquête `https://www.epicov.fr`, consulté le 17/02/2024

ALBOUY V., LEGLEYE S. 2020 « Conditions de vie pendant le confinement : des écarts selon le niveau de vie et la catégorie socioprofessionnelle », *Insee Focus* 197

HAZO J.-B., COSTEMALLE V. : « Confinement du printemps 2020 : une hausse des syndromes dépressifs, surtout chez les 15-24 ans. Résultats issus de la 1re vague de l'enquête EpiCov et comparaison avec les enquêtes de santé européennes (EHIS) de 2014 et 2019 » *Etudes et Résultats*, Drees, 2021

WARSARSZAWSKI J. et al, « En mai 2020, 4,5% de la population en France métropolitaine a développé des anticorps contre le Sars-Cov-2 » *Enquêtes et résultats*, Dress, 2020
Haute Autorité de Santé «Sexe, Genre et Santé» *Rapport d'analyse prospective 2020*

S DURIEUX, P ROUAUD (Insee), R BELLE (Drees) «Dans les quartiers les plus en difficulté, seulement un habitant sur trois en emploi», 2020, *Insee Analyses Provence-Alpes Côte-d'Azur* 82

BARKOVIC, CAVAN, «Des conditions de vie plus difficiles pour les mères isolées» 2022, *Insee Flash Hauts-de-France* 134