Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

The Effect of Multiple Contact Patterns During the COVID-19 Lockdown on Older People

A Doubly Robust Approach to Propensity Score Weighting with Gradient Boosted Regression

Relatore: Prof. Bruno Arpino
Dipartimento di Scienze Statistiche

Laureando: Gaetano Tedesco
Matricola n. 2006942

Anno Accademico 2023/2024

To my beloved and amazing Family

Contents

Preface

Given its adverse impacts on health, longevity, and well-being, loneliness represents a significant public health concern. Characterised as the perceived deficit between the actual and desired quality or quantity of social relationships, loneliness is prevalent among older adults and the feeling is known to increase during aging. Previous research has linked loneliness and social isolation with an elevated risk of various mental health conditions, including depression, anxiety, cognitive decline, and Alzheimer's disease (Lim *et al.*, 2020), as well as cardiovascular conditions and coronary heart disease (Cacioppo *et al.*, 2002).

The COVID-19 pandemic, with its stringent restrictions on physical contact and enforced lockdowns, significantly reduced in-person interactions, increasing feelings of loneliness among people. However, the increase in remote interactions may helped to counteract the negative repercussions of diminished face-to-face social engagement.

This research explores the causal relationship between modified contact patterns and perceived loneliness among older Italian adults. Data were taken from the *Intergen-COVID* online survey (Arpino *et al.*, 2020) which aim was to assess the impact of lockdown restrictions on people's live posing particular attention on social relationship and mental health.

Focusing on the influence of non-physical contacts in alleviating perceived loneliness, the study covers different non-physical contact patterns through various treatment formulations:

- A binary comparison between individuals who increased their non-physical contact and those who did not.
- A multi-treatment analysis exploring mixed pattern of individuals who increased non-physical interactions with some contacts while decreasing or maintaining unchanged levels with others.

Within the *Neyman-Rubin* causal framework the study employs a doubly robust method to estimate the Average Treatment Effect on the Treated (ATT) for changes in non-physical interactions, combining an outcome model with a weighting procedure to derive consistent causal effect estimates if either the outcome model or the *propensity score* (Rosenbaum and Rubin, 1983) model are correct, but not necessarily both. Following McCaffrey *et al.* (2013), gradient boosted regression is implemented for propensity score weighting to effectively balance key pre-treatment covariates and address potential confounders in the treatment-outcome relationship. The outcome is modelled using the g-formula proposed by Robins (1986).

Our findings suggest that enhanced non-physical contact significantly alleviates the perception of loneliness feeling during the COVID-19 lockdown across all examined frameworks.

The work unfolds as follows: in Chapter 1 we explore the problem of loneliness among older individuals with a particular focus to the context of COVID-19 lockdown. Chapter 2 shifts the focus to causal inference, defining the quantities of interest, explaining the *Rubin Causal Model*, and the usage of propensity scores in observational studies. Chapter 3 delves deeper into propensity score weighting, exploring Generalized Boosted Models (GBM) as a tool for *nonparametric* propensity scores estimation. We then describe our approach to the outcome analysis through *G-computation* aiming to a *doubly robust* estimation of the causal effect. In Chapter 4 we finally apply these theoretical and methodological frameworks in order to estimate the causal relationship between non-physical contact patterns during the COVID-19 lockdown and perceived *severe* loneliness among older Italian adults.

Chapter 1

Loneliness, Older People and COVID-19

1.1 The Definition of Loneliness

Life in 21^{st} is different from past periods of the human history. People are living longer and the rise of Internet has transformed the way we interact with each other and with the environment that surrounds us. People are increasingly connected through digital technology that allow them to communicate in new ways but the prevalence of loneliness also appears to be rising. Over the past 40 years loneliness has become more widespread and example of this have been documented in previous study. For instance in U.K. the research of Victor and Yang (2012) have estimated a prevalence of *severe* loneliness between 5%-6% and people who reported to feel lonely "sometimes" between 21%-31%.

Loneliness is characterised as the perceived deficit between actual and desired quantity or quality of social relationship. The definition underline the fact that *feeling* lonely does not necessarily mean *being* alone and viceversa (Cacioppo *et al.*, 2016), emphasizing that social being require the presence of *significant* others and the need of *feeling* connected to not feel lonely.

Because of the strong correlation between loneliness and depression, varying from .31 to .71 (Cacioppo *et al.*, 2002), many believed that the loneliness was simply a dimension of depression. There is now considerable evidence showing that loneliness and depression are separable and that the former increase the risk of the latter (Cacioppo *et al.* (2002); Heinrich and Gullone (2006)), making loneliness increasingly recognised as the next critical public health issue.

1.2 A Conceptual Model for Loneliness

Recently, new conceptual model of loneliness has been introduced illustrating the complexity of loneliness and the importance of prevention and intervention to solve or alleviate the problem (Lim *et al.*, 2020).

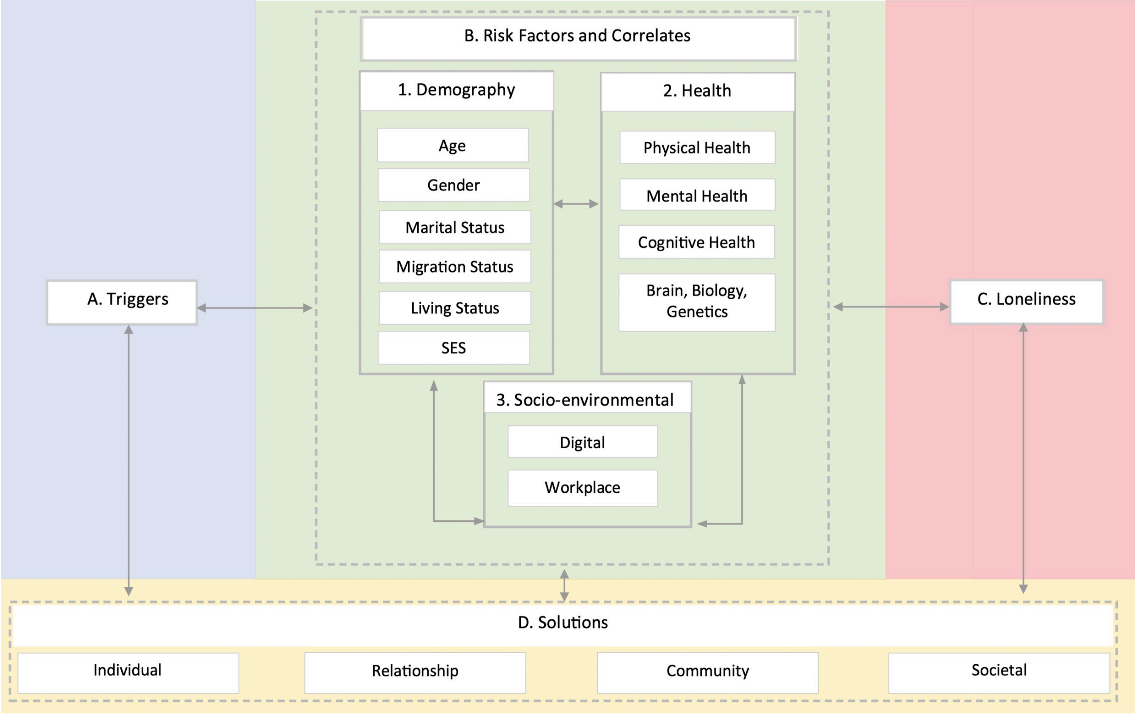


Figure 1.1: Conceptual model of loneliness (Lim *et al.*, 2020)

The conceptual model proposed by Lim is made up of four section, namely:

1. Triggers,
2. Loneliness correlates and risk factors,
3. Loneliness as a consequence,
4. Solutions.

1.2.1 Triggers

Are considered *triggers* all the significant life event or stage transitions that precedes and, possibly, initiate the development of problematic loneliness in a person. These events may include a change of residence, change in marital status, illness or death of members of the individual's inner social network.

1.2.2 Loneliness Correlates and Risk Factors

The model examines a long list of possibles risk factors and correlates of loneliness assuming that every individual holds a level of risk to experiencing problematic loneliness and that this state might be maintained by multiple factors (from demographics to modifiable factors). For the sake of brevity we outline only those factors that have been useful to this study, for a more detailed discussion, see Lim *et al.* (2020) and Heinrich and Gullone (2006).

Demography

When considering age, loneliness have been found to be higher among adults older than 65 and in young individuals aged between 18 and 25 years old (Victor and Yang, 2012). Women have been showed to feel more lonely then men in older adults and the model does not exclude the presence of some unique predictor of loneliness to women compared with opposite sex (Thurston and Kubzansky, 2009). Other risks factors related to demographic variable are:

- marital status: where people who are unmarried typically reported being lonelier than those who are married (with higher levels for widowed or divorced individuals),
- Co-living status: adults living alone were associated with higher loneliness,
- Socioeconomic status: in particular people with lower income, lower education level and who suffered economic problems showed higher levels of loneliness.

Physical Health

Multiple health indicators are related to loneliness. In particular a number of studies have found that the feeling is associated with increase incidence of coronary disease and increase risk of stroke (Valtorta *et al.*, 2016). A part from that, loneliness exhibits an exposure-response relationship with various cardiovascular health indicators, such as blood pressure, higher cortisol and poor sleep (Cacioppo *et al.*, 2002). Furthermore, Thurston and Kubzansky (2009) found that loneliness was a risk factor for coronary heart disease in women.

Mental Health

Loneliness is associated with poorer mental health outcomes, including higher anxiety, depression and psychotic symptoms (Lim *et al.*, 2020).

Cognitive Health

Loneliness is also harmful to cognitive health. In particular, there is evidence that loneliness is a risk factor for Alzheimer's disease and dementia. Specifically, people who reported higher loneliness were shown to be 64% more likely to suffer from dementia, and were associated with twice the risk to develop Alzheimer's disease compared with their less lonely counterpart (Wilson *et al.*, 2007).

Socio-environmental factors

When considering digital communication Lim *et al.* (2020) sustain that more research is needed in order to understand how digital technology shapes the way people interact with each others. Another aspect to consider is the connection between loneliness and workplace, as a matter of fact loneliness had been found to have a negative effect on employee well being, work satisfaction and employment itself, suggesting a bidirectional relationship between the latter and transitory or sustained feeling of loneliness (Morrish *et al.*, 2022).

1.2.3 Loneliness as Consequence

In this section of the model Lim *et al.* (2020) assume that all individuals have different degrees of experiencing one or more of the above risk factors and triggers. In particular, the model assumes that even in presence of triggering events these will interact with one or more of the risk factors. Hence, loneliness as a consequence can differ from person to person and, given its subjectivity, assessing *how* and *when* it became a problem might be difficult.

1.2.4 Solutions

Because, as previously said, loneliness may be consequence of a multitude of factors, it follows naturally, that loneliness may not be easy to solve. The conceptual model provide that solutions can be applied either for preventing loneliness or to address its severity. The model suggests that solutions should be relevant to the unique experience of the person in other to be effective and that, depending on the particular case, solutions might be delivered in four different levels: individual, relationship, community and societal.

1.3 The Effects of Loneliness

Loneliness, a complex and universally experienced human condition, has emerged as a critical area of interest in health research due to its profound implications on physical and psychological well-being. The subjective sense of social isolation transcends mere physical solitude, impacting individuals across various life stages and demographics. By examining the intersection of loneliness with socio-demographic factors and health outcomes, the studies here presented shed light on the pervasive and multifaceted nature of loneliness.

Studies identify significant correlations between loneliness and increased total peripheral resistance along with poorer sleep quality, especially pronounced in older demographics, suggesting age-related increases in blood pressure and sleep disturbances due to loneliness (Cacioppo *et al.*, 2002).

Thurston and Kubzansky (2009) revealed that loneliness increases coronary heart disease (CHD) risk in women, even when controlling for various confounders including depressive symptoms, establishing a clear link between emotional isolation and cardiovascular health. This association was not observed in men indicating a gender-specific vulnerability that underscores the necessity for individual-tailored health interventions exhorted by Lim *et al.* (2020). Building upon health-related issues, further studies have shown that elevated levels of loneliness and social isolation augment the risk of stroke for both genders by 32% (Valtorta *et al.*, 2016).

In a longitudinal study, based on a sample of 13,752 U.S. adults aged 50 from the Health and Retirement Study (HRS), grater loneliness was significantly associated with all-cause mortality, with a 43% increased risk for individual with the highest (versus lowest) level of loneliness (measure based on an 11-item revised UCLA Loneliness Scale) even when adjusting for a wide range of covariates. Furthermore, loneliness was linked, as risk factor, to a range of health outcomes, including lung disease (25% increased risk) , chronic pain (13% increased risk), sleep problems, and various aspects of psychological well-being (Hong *et al.*, 2023).

Collectively, these studies underline the necessity of integrated approaches in research and public health to address this issue.

1.4 Loneliness among Older People

When examining the relationship between loneliness and age groups an important contribute was made by Victor and Yang (2012). Based on a sample of individuals aged 15+ from the European Social Survey, their studies have highlighted that loneliness is a prevalent issue among older populations, demonstrating a U-shaped distribution across the lifespan, where individuals under 25 and those over 65 experience higher levelf of it. Among the key factors that contribute to increased feelings of loneliness in this population, their study identified:

- **depression**: associated with loneliness across all age groups,
- **quality of social engagement**: shows protective effects against loneliness for those in mid and later life,
- **marital Status**: being married or in a committed relationship appears to be more important for mitigating loneliness in older people adults,
- **socioeconomic factors**: lower socioeconomic status, lack of transport, and living in rural areas have been identified as predictors of loneliness among older adults.

Other studies underlined that for this demographic, the influence of both subjective and objective isolation on risk for mortality is comparable with other well established risk factor for mortality like obesity or substance abuse (Holt-Lunstad *et al.*, 2015).

1.5 COVID-19 Lowdown and Remote Contacts

During the recent COVID-19 pandemic, measures introduced to slow down the spread of the virus, such as mandatory *stay-at-home* or minimizing physical contact, may have exacerbated loneliness, particularly among older adults (Dahlberg, 2021).

This demographic have been identified with higher risk of poor outcome if infected and underwent a greater restriction on physical contacts. Findings have shown a worsening of the general well-being and loneliness feeling on older individuals, with a significant increase of 9 percentage point in perceived loneliness (Macdonald and Hülür, 2021).

Recent studies also provided new insights on the heterogeneity of COVID-19 lockdowns effect on loneliness. Specifically, childless and unpartnered older individuals had a higher risk of loneliness compared to those with family ties. Such conditions were significantly associated with a 5.4 and 6.7 percentage points higher probability of feeling lonelier than before the pandemic (Arpino *et al.*, 2022), highlighting the role of social ties in influencing loneliness among older adults.

On the other hand, remote contact substantially rose in Italy, Spain and France, counteracting the lack of in-person relations (Arpino *et al.*, 2020). This might have helped buffer the adverse effects of the pandemic on the emotional well-being and loneliness of older adults with studies providing evidence that, when face-to-face relationship are not possible, technology-based interactions provides essential emotional support helping maintain social ties and mitigate loneliness (Liddle *et al.*, 2020). Such claims were further supported by the work of Macdonald and Hülür (2021), showing that adults who were more satisfied with their communication were also related to lower level of loneliness.

It is, therefore, important to continue exploring and supporting the use of technology to maintain social connections among older adults, especially during times of crisis that necessitate physical distancing.

Chapter 2

Causal Inference

2.1 The Definitions of Cause and Effects

2.1.1 Philosophical Point of View

The definition of *cause* finds its roots in the work of many Greek's and later philosopher. Aristotle, in his work about *Physics*, gave us four different definitions of the "causes of a thing": the *material cause* (that *out of which* the thing is made), the *formal cause* (that *into which* the thing is made), the *efficient cause* (that *which makes* the thing) and the *final cause* (that *for which* the thing is made). It is evident, that only the definition of efficient cause is relevant for our discussions about causation. A broader definition was first introduced by Locke (1690): "that which produces any simple or complex idea, we denote by the general name 'cause', and that which is produced, 'effect'." Both these definitions, while successfully conveying the idea of a cause, give us very little information about what conditions have to be satisfied in order to define something as a cause or an effect.

An important contribution to the subject comes from the Hume's analysis of causality (1740,1748) in which the philosopher argued that an experienced event called *cause* is always followed by the experienced event called *effect*. In his analysis, Hume recognized the basics criteria for causation, namely the spatial/temporal contiguity and the temporal succession of cause and effect. What is still lacking is the notion that a cause may have been different from what it was and that this difference defines the effect.

John Stuart Mill, on the other side, firstly introduced the notion of experiment as crucial to causation. In 1843 he stated that experimentation was the only way to establish causal relationship and distinguish them from mere correlations. In his work, Mill, also defined what we call causal effect, or more generally, the effect of a cause:

> *If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring in the former; the circumstances in*

which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause of the phenomenon.

2.1.2 From Philosophy to Probability

A first attempt to build a mathematical framework for causal-effect relationship was proposed by Patrick Suppes in his *Probabilistic Theory of Causality.* His goal was to improve Hume's analysis using the language of stochastic processes. Given that both, cause and effect, are expressible as events, he defined:

1. the **prima facie cause** of an event as an event that temporally precedes it and that is positively associated with it, if $r < s$ the event C_r is a *prima facie cause* of the event E_s if
$$\mathrm{Pr}(E_s|C_r) > Pr(E_s),$$
2. the **spurious cause** of an effect as a prima facie cause that is conditionally independent of the effect given another event, prior to the prima facie cause. C_r is a *spurious cause* of E_s if, for some $q < r < s$, exist an event D_q such that:
$$\mathrm{Pr}(E_s|C_r, D_q) = Pr(E_s|D_q),$$
and
$$\mathrm{Pr}(E_s|C_r, D_q) \geq Pr(E_s|C_r),$$
3. the **genuine cause** as a prima facie cause that is not spurious. C_r is a *genuine cause* of E_s if C_r is prima face cause of E_s but is not a spurious cause of E_s.

The major lack of the Suppes' theory is that it does not provide the machinery to express the effect of a cause in a particular case (or at unit level).

2.1.3 Statistically Speaking

Cause

Given the definition of Aristotle, Hume and Mill, anything can be a cause or, at least, a *potential* cause. In modern causal inference, we consider **causes** only those things that could, in principle, be considered as treatments in experiments or in a *hypothetical experiments.* Under this notion, the cause in an experiment is defined in the same way we define a cause in an *observational study* with the only difference being the degree of control the researcher has over the investigated phenomena. In other words, an attribute of a particular unit cannot be a cause since the notion of *potential exposability* cannot be applied to it (Holland, 1986). The definition of cause provided might looks loosely at first sight but it is already narrowing the set of events or elements that we can define as such, giving a more specific meaning to the word *cause* and brought Holland and Rubin to made up the famous motto:

> *"No causation without manipulation."*

underlying the importance of this restriction.

Effect as Causal Effect

The formal statistical definition of causal effect was first introduced by Rubin (1974). His work laid the foundation for understanding the impact of treatments in a causal framework, marking a significant milestone in the field of statistics and causal inference.

Considering an experiment of $2N$ units, half exposed to a treatment (T) and the other half exposed to a control treatment (C), in a way that ensured that each unit was equally likely to be exposed to T or to C (a *randomized experiment*). The objective of the study is to assess, for some population units, the effect of treatment T versus treatment C on a dependent variable Y. Let a unit be associated with a pair of times, t_1 and t_2, where t_1 denotes the time of initiation of a treatment and t_2 denote the time of measurement of a dependent variable, Y, with $t_1 < t_2$. Assuming that:

1. a time of initiation of treatment can be ascertained for each unit,

2. T and C are exclusive of each other, thus a unit cannot receive both T and C.

The **causal effect** of the treatment T versus the treatment C is defined as

$$Y(T) - Y(C) \tag{2.1}$$

for a particular unit and the times t_1 and t_2. $Y(T)$ denote the value of Y measured (assuming negligible "technical errors") at t_2, given that the unit received T at time t_1 and $Y(C)$ is the value of Y measured at t_2, given that the unit received C at time t_1. As a result of the experimental setup in which the definition above is given, it is clear that the effects of causes are always relative to other causes.

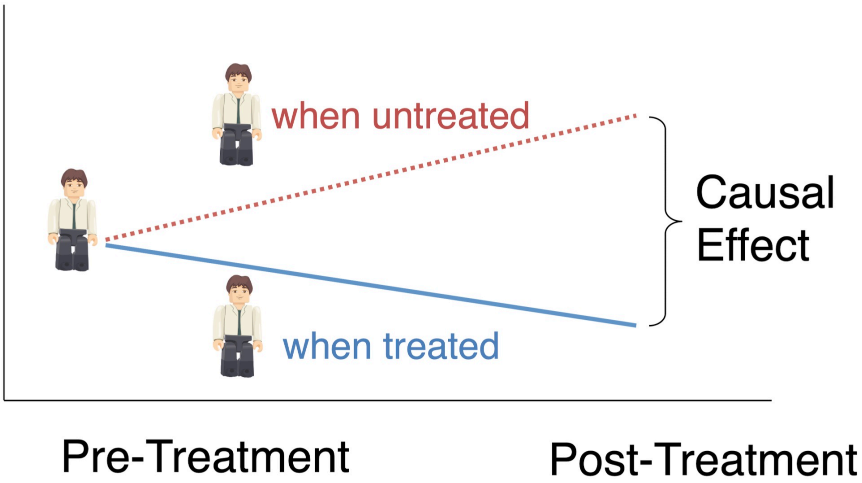


Figure 2.1: Causal Effect.

Much of the basics ideas on which modern causal inference is built upon, lie in the R. A. Fisher and Neyman's foundation of experimental design for causal inference in 1926 and 1923, respectively. As described by Holland (1986), both papers (in a more or less explicit way) introduced the idea of multiple versions of the response as *potential outcomes* of a treatment (in both cases the outcome was the yield from an experimental plot of land under a particular treatment condition in an agricultural experiment).

2.2 A Formal Framework for Causal Inference

2.2.1 Potential Outcomes

The first *explicit* introduction of the notion of potential outcomes is given by Neyman (1923) who, talking about his agricultural experiment, stated: "... U_{ik} is the yield of the ith variety on the kth plot." In this case U_{ik} is a "*potential* yield" and not an observed yield, since each plot is exposed to only one variety of crops. Such notation will later allow causal effects ("Neyman-Rubin model", Pearl (1996)) and causal *estimand* (such as ATE or ATT) to be defined without reference to any particular probability model for the data in both randomized (such the one proposed by Neyman) and observational studies.

Definition 1 (Potential Outcome) *The potential outcome for each unit i and each treatment condition T is defined as the outcome that would be observed if the unit were exposed to that treatment condition.*

In presence of two causes (Holland, 1986), two treatments or more commonly a binary treatment T, where $T = 1$ indicates the treatment and $T = 0$ indicates the control the potential outcomes for unit i are $Y_i(1)$ and $Y_i(0)$, respectively.

This definition of potential outcomes extend naturally to situation in which we have more than two treatments or a multinomial treatment. In this case, the treatment $T = 1, 2, ..., M$, with M number of different treatments, define the potential outcomes $Y_i(t)$, where $t = 1, 2, ..., M$.

2.2.2 The Fundamental Problem of Causal Inference

Following from Equation (2.1), in order to asses the causal effect of a treatment versus another, we would like to compare the potential outcome $Y_i(1)$ and the potential outcome $Y_i(0)$ of the ith unit.

From his definition of the causal effect (now considered as unit-level causal effect) arise what Holland (1986) define as the *Fundamental Problem of Causal Inference*, namely the impossibility to *observe* both value $Y_i(1)$ and $Y_i(0)$ for the same unit and , therefore, the impossibility to *observe* the effect of the treatment T on the unit i. This inability to directly observe the quantity of interest is what distinguishes causal inference from standard statistical inference.

Although the *definition* of causal effect does not require more than one unit, in order to learn relevant knowledge about such effect we will require multiple units.

From the definition of potential outcome it is clear that on a unit-level we will, most likely, observe only one potential outcome. Therefore, to make causal inference possible again, we must rely on multiple units.

2.2.3 Rubin Model for Causal Inference

We will now give a formal introduction of the *Rubin Causal Model* (also referred to as *Neyman-Rubin Causal Model*), one of the principal framework for causal inference, summarizing all the notation used so far and define the quantities of interest in the causal analysis. We will then show how one can address the Fundamental Problem of Causal Inference under this framework.

The Formal Model

The elements that constitute the Rubin Causal Model are the following (Please note that some notation might differ from the papers in which such model was first introduced, but the structure is the same):

- a population of units, U, of size N indexed by $i = 1, 2, ..., N$,
- a set of T defined treatment or causal agents (Holland and Rubin, 1980) to which each unit can potentially be exposed (for the sake of notation we will consider a binary setup with $T = \{1, 0\}$ but the results naturally extends to more than two treatments setup),
- a response variable Y that represents our outcome of interest, such variable can be measured for each unit after the exposure to *a* causal agent T.

As we saw before, the intuitive definition of the causal effect described by the model is the difference between the response variable that would be measured after exposing a unit to treatment ($t = 1$) and the response that would be measured on the *same* unit had it been exposed to the control ($t = 0$). Without lost of generality we will consider our response, Y, as a dichotomous variable, since it is the interest of our study. The extension to a general Y is straightforward.

The model assumes that each unit has T potential outcome one for each treatment to which the unit could have been exposed. Formally, for a binary treatment setup, we associate to each unit in U the vector

$$(Y_i(1), Y_i(0)), \tag{2.2}$$

where $Y_i(t)$ is the potential outcome of the ith unit if it would be exposed to cause $t \in T$.

Assumption 1 (SUTVA) *The potential outcomes of any unit do not vary with the treatments assigned to other units (no interference), and, for each unit, there are no different forms or versions of each treatment level (consistancy), which lead to different potential outcomes.*

Under *SUTVA* (Rubin, 1978) the vector given by Equation (2.2) is fully representative of all the possible values of Y under all pairing of $t \in T$ with $i \in U$. This simplify both definition and computation of the causal effect, allowing us to finally overcome the problem of causal inference at *population level*.

Treatments and Outcome

Further developing the model notation, we define T_i as the treatment indicator that takes values 0 for the control treatment and 1 for the active treatment. For each unit $i = 1, 2, ..., N$ the model assume one realized (and possible observed) potential outcome denoted by:

$$Y_i^{\text{obs}} = Y_i(T_i) = \begin{cases} Y_i(0) & \text{if } T_i = 0, \\ Y_i(1) & \text{if } T_i = 1. \end{cases}, \tag{2.3}$$

and one missing potential outcome, denoted by:

$$Y_i^{\text{mis}} = Y_i(1 - T_i) = \begin{cases} Y_i(1) & \text{if } T_i = 0, \\ Y_i(0) & \text{if } T_i = 1. \end{cases}. \tag{2.4}$$

It is clear, from the notation used, that causal inference is, at its root, a *missing data problem* (Rubin, 1974). Given any treatment assigned to an individual, the potential outcome associated with any alternative treatment is missing.

Attributes, Pre-treatment Variables or Covariates

Following from Equation (2.3) and Equation (2.4), to estimate the causal effect of any particular unit in the form given by Equation (2.1) we will generally need to estimate the missing potential outcomes. Generating such estimates may result difficult, since they involve assumptions about the assignment mechanism and the comparisons between different units exposed to only one treatment. In this case unit-specific background can assist in making such predictions.

Let $\mathbf{X}$ denote the *pre-treatment* variables of each unit, the key characteristic of such variables is that they are known, or assumed, to be *unaffected* by the treatment assignment and, therefore, can be used to make our prediction task somewhat easier.

Such information commonly serve to:

- make our estimates more precise, by explaining some of the variation of the outcome,
- might be used to select a sub-group of the population of interest on which we would like to know *typical* casual effect of the treatment,
- they might have an effect on both the assignment mechanism and the potential outcomes.

In the last role, covariates are used to make assumptions about the relationship between outcome and treatment which are more plausible within homogeneous subpopulation (with respect to these variable) rather than unconditionally.

2.2.4 Assignment Mechanism

Building upon the structure explained so far, we now introduce the setup in which the model was firstly developed, randomized experiments, and then extend it to more complicated cases concerning *observational studies.*

Definition 2 (Assignment Mechanism) *Given a population of N units, the assignment mechanism is a function* $\Pr(\mathbf{T}|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0))$*, taking values in* $[0, 1]$*, satisfying*

$$\sum_{\mathbf{T} \in \{0,1\}^N} \Pr(\mathbf{T}|\mathbf{X}, Y(0), Y(1)) = 1,$$

for all $\mathbf{X}, \mathbf{Y}(0)$ *and* $\mathbf{Y}(1)$*.*

A *Randomized Control Trial* or, generally speaking, a randomized experiment is an experiment in which the assignment mechanism is *known* (i.e. coin flip, which would give to each unit in the population the same probability of receiving the treatment). Such an assignment mechanism:

1. is probabilistic,
2. has a known functional form that is controlled by the researcher.

They produce, so to speak, a "fair lottery" (Rosenbaum, 2023). Let T_i indicate the treatment assignment for unit i, again taking values 1 or 0. The *assignment mechanism* (namely the methods that assign treatments to units) gives us the probability of the vector $\mathbf{T} = (T_1, T_2, ..., T_N)$ given the fixed values of $\mathbf{X}, \mathbf{Y}(1)$ and $\mathbf{Y}(0)$ (that being the arrays for all units).

In a randomized trial the assignment mechanism can be written as:

$$\Pr(\mathbf{T}|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) = Pr(\mathbf{T}|\mathbf{X}). \tag{2.5}$$

Since randomized experiments have the property of balancing both measured and *unmeasured* covariates they are, as defined by Rubin (1976), *ignorable.* Practically speaking, the distribution of observed and unobserved covariates are equal in expectation. Thus, the outcome of the exposed group is expected to be representative of the outcome of the unexposed group (or all individuals in the study) had they all been exposed, and vice versa (property often referred to as *"exchangeability"*). This allows us to compute the *typical* causal effect of the treatment on our sample.

2.2.5 Causal Estimands

Expanding on the definition of causal effect given in Equation (2.1) we introduce a more formal description of the ultimate object of interest of our analysis, the *causal estimand.*
Again, we consider a *finite* population of N individuals, indexed by $i = 1, ..., N$. Each unit in the population can be exposed to one of a set of treatment that we considered (without loss of generality) $T = \{0, 1\}$. We define the treatment indicator as $T_i \in T, \forall i$, and the potential outcomes $Y_i(0), Y_i(1)$ as in Equations (2.3), (2.4).

Unit-level causal effect

For each unit i and for each treatment $T = \{0, 1\}$, the *unit-level causal effect* is defined as the comparison between the two potential outcomes $Y_i(0), Y_i(1)$. In this case, as in the rest of the work, it is assumed that such a comparison would be expressed as the difference:

$$Y_i(1) - Y_i(0). \tag{2.6}$$

As said before such an effect it is difficult, if not impossible, to estimate due to the fundamental problem of causal inference (Holland, 1986). The *statistical* solution will make use of our population U and will allow us to gain information about the causal effect using information *observed* on *different* units.

Typical causal effect (population level)

If assignment to treatment is randomized, causal inference at level population is straightforward. Since treatment assignment is independent from all the *observed* or *unobserved* pre-treatment characteristics, observed and unobserved confounder balance across both groups and the treatment assignment is independent of the potential outcomes, formally $\{Y(0), Y(1) \perp\!\!\!\perp W\}$ (where $\perp\!\!\!\perp$ denote independence), we define the *Average Treatment Effect* as:

$$\begin{aligned} \tau &= E(Y(1)|T = 1) - E(Y(0)|T = 0) \\ &= E(Y|T = 1) - E(Y|T = 0). \end{aligned} \tag{2.7}$$

Equation (2.7) can be consistently estimated, in an experimental setting, by:

$$\hat{\tau} = \bar{y_1} - \bar{y_0}, \tag{2.8}$$

with $\bar{y_1}, \bar{y_0}$ being the sample mean of the treated and control group, respectively (Neyman, 1923).

2.2.6 Observational Study

In some situation randomized experimentation is unethical or unfeasible. A person cannot be imposed to undergo an harmful treatment or the costs of a randomized experiment might be prohibitive for the researchers. For instance, people cannot be forced to smoke in order to understand the damages caused by smoking or emotional trauma cannot be inflicted to people to understand post-traumatic stress syndrome. In such situation, one might use the evidence from various sources, including non-randomized or *observational study.*

Cochran (1965) identified two principal characteristic of observational studies, namely:

- is an "empiric investigation" in which the objective is to elucidate cause-and-effect relationship,
- it is not feasible to use controlled experimentation.

Formally speaking, are referred to as observational studies those for which the assignment mechanisms of the treatments are unknown.

In an observational setting, unless specific actions are taken, the treatment and control groups are almost never balanced, and the direct comparison that defines the causal effect may be biased. In such studies, individuals exposed to one treatment systematically differ from those exposed to another, making the units *incomparable.* This incomparability stems from the imbalance between the treatment and control groups in their covariates ($\mathbf{X}$) and the presence of heterogeneous treatment effects. Consequently, it becomes important and interesting to distinguish between different types of average treatment effects. The common quantities of interest are the Average Treatment Effect (ATE), which is the expected effect of the treatment across the entire population; the Average Treatment Effect *among the Treated* (ATT), which is the expected effect of the treatment on those who have been treated:

$$\begin{aligned}\tau^T &= E(Y(1) - Y(0)|T = 1)\\ &= E(Y(1)|T = 1) - E(Y(0)|T = 1)\\ &= E(Y|T = 1) - E(Y(0)|T = 1),\end{aligned} \tag{2.9}$$

and the Average Treatment Effect *among the Control* group (ATC), which is the expected effect on the control units:

$$\begin{aligned}\tau^C &= E(Y(1) - Y(0)|T = 0)\\ &= E(Y(1)|T = 0) - E(Y(0)|T = 0)\\ &= E(Y(1)|T = 0) - E(Y|T = 0).\end{aligned} \tag{2.10}$$

.

Due to the imbalance and heterogeneity, ATE (2.7), ATT (2.9), and ATC (2.10) can be significantly different, highlighting the need for a nuanced analysis to understand the causal effects accurately.

The equations expressed in (2.9) and (2.10) cannot be estimated directly from data. Since $Y(0)$ and $Y(1)$ are not observed for treated and comparison group, the quantities $E(Y(0)|T = 1)$ and $E(Y(1)|T = 0)$ are unknown. Such quantities are *counterfactuals*, expressing means of the potential outcomes corresponding to the treatment level that is the opposite of the actual received treatment.

As we saw in randomized setup, in order to compute the *typical* causal effect of the treatment we require:

- the assignment mechanism to be independent from the potential outcomes,
- balance across both group in terms of observed and unobserved covariates,

in order to use information about *different* units to gain information about the causal effects.

To achieve our objective, we can use the background information about the units and assume that the assignment mechanism for the treatment is independent to the potenital outcomes *conditionally* to the *observable* covariates. Producing what Rubin (1974) defined as a "carefully controlled nonrandomized" experiment.

Following Rosenbaum and Rubin (1983) we assume *strong ignorability.*

Assumption 2 (Strong Ignorability) *The strong ignorability assumtion requires that the potential outcomes for each unit is independent from the treatment assignment mechanism given the covariates. In particular:*

1. ***Sufficeint overlap (Positivity):*** $0 < Pr(T = 1|X) < 1$*, every subject has non-zero probability of receiving each treatment.*
   There are no patterns of values for the covariates which preclude the unit from receiving one treatment or another.
2. ***No unmeasured confounders (conditional exchangeability):***
   $(Y(1), Y(0)) \perp\!\!\!\perp T|X$

Under such assumption, that *mimic* the properties of an RCT (with respect to the observed covariates) making units comparable, we obtain:

$$E[Y(0)|T = 1, X] = E[Y(0)|T = 0, X] = E[Y|T = 0, X],$$
$$E[Y(1)|T = 1, X] = E[Y(1)|T = 0, X] = E[Y|T = 1, X].$$

and the causal estimands (2.9) and (2.10) can be computed as:

$$\tau^T = E\ \{E(Y|T = 1, X) - E(Y|T = 0, X)|\ T = 1\}, \tag{2.11}$$

and

$$\tau^C = E\ \{E(Y|T = 1, X) - E(Y|T = 0, X)|\ T = 0\}, \tag{2.12}$$

where the outer expectation is taken over the distribution of $X|T = 1$ or $X|T = 0$, which represents the distribution of the pre-treatment variables for treated and control group, respectively.

2.3 Propensity Scores

In order to derive knowledge about the causal effect in an observational study we need to make units receiving different treatments *look alike.* Using the background information at our disposal, we want to *balance* the distributions of the pre-treatments covariates **X** for the two (or multiple) groups. Often with hundreds of covariates for a single unit achieve balance between the two multivariate distributions might be complex or unfeasible. In one of the most cited paper in statistics, Rosenbaum and Rubin (1983) proposed a key concept that might help us with this task.

Definition 3 (Propensity Score) *For each unit i with covariates X_i, a binary treatment indicator T_i, and the potential oucome $Y_i(0), Y_i(1)$, the* ***propensity score*** *is defined as:*

$$e(X_i, Y_i(1), Y_i(0)) = \Pr(T_i = 1|X_i, Y_i(1), Y_i(0)) \tag{2.13}$$

Under the hypothesys of ***strong ignorability****, we have:*

$$\begin{aligned} e(X_i, Y_i(1), Y_i(0)) &= \Pr(T_i = 1|X_i, Y_i(1), Y_i(0)) \\ &= \Pr(T_i = 1|X_i), \end{aligned}$$

So the propensity score reduces to:

$$e(X_i) = \Pr(T_i = 1|X_i), \tag{2.14}$$

namely the conditional probability of assignment of a particular treatment, given a vector of observed covariates.

2.3.1 Propensity Score as a Dimensional Reduction Tool

Theorem 1 *If $T \perp\!\!\!\perp \{Y(1), Y(0)\}|\ X,$ then $T \perp\!\!\!\perp \{Y(1), Y(0)\}|\ e(X)$.*

Theorem (1) states that if *strong ignorability* holds conditional on covariates X, then it also holds conditional on the scalar propensity score $e(X)$. Implying that conditioning on the propensity score $e(X)$ removes all confounding induced by covariates X. While the original covariates **X** can be general and have many dimensions, the propensity score $e(X)$ is a one-dimensional scalar variable bounded between 0 and 1 that reduces the dimension of the original pre-treatment information but still maintains the ignorability. Therefore, make us able to compute the causal estimands (2.11) and (2.12) in observational studies as:

$$\tau^T = E\ \{E(Y|T = 1, e(X)) - E(Y|T = 0, e(X))|\ T = 1\}, \tag{2.15}$$

and

$$\tau^C = E\ \{E(Y|T = 1, e(X)) - E(Y|T = 0, e(X))|\ T = 1\}. \tag{2.16}$$

Chapter 3

Propensity Score Weighting

3.1 Propensity Score in Observational Studies

The main challenge of non-experimental studies is *confounding* (VanderWeele and Shpitser, 2013). In presence of confounding, exposed and unexposed groups differ in term of distribution of covariates that predict both the exposure to the treatment and the outcome (we will refer to such covariates as *confounders*), loosing the property of *exchangeability* and impeding us to obtain unbiased estimates of the causal effects of interest. Thus, in non-experimental studies, it is crucial that the researchers implement some strategies to *adjust for potential confounders.* Such strategies can be categorized in:

- "*Analysis* - based": approach involving direct modeling of the outcome and its relationship with treatment and potential conforunders,
- "*Design* - based": methods attempting to create or adjust a sample in order to balance the distribution of the *observed* covariates between the groups (Rubin, 2007).

While comparing the two types of strategies we would like to underline few key benefits that design-based approaches have over the direct outcome modeling. First, the *design stage*, which we will refer as *balancing*, is done without the use of the outcome data, showing a clear separation between design and effect estimation as would naturally happen in an experimental study. A second key benefits of the design approach is that there are clear diagnostic tools to assess its success. After a *matched* or *weighted* sample is created, covariates balance of this sample can be checked before using such data to estimate the causal effect. Finally, such strategy reduce the impact of potential biases introduced by extrapolation of the outcome model when treated and control group differ substantially. For instance, if the treated group is much younger than the untreated group, then the implicit imputation of the missing potential outcomes under exposure for the control units would be based on a model of the potential outcome fitted among much younger people than those for whom the prediction is made.

In this work we will combine analysis- and design-based approach and we will use a *doubly robust* estimator of the causal effect. In particular, we will implement regression adjustment in order to reduce any remaining imbalance in the sample after the design stage (Chang and Stuart, 2022).

3.1.1 Propensity Score as Metrics for Adjustment

The first step is to define a measure of *similarity* between two individuals. At this stage, follows from Definition (3) that *propensity score* is a suitable tool as measure of similarity because of its balancing properties. Firstly, propensity score *summarize* the background information about the individual in unique number between 0 and 1. Furthermore, under the hypothesis of *strong ignorability*, conditionally to the propensity score the distribution of the covariates, on which the score is estimated, are balanced between the groups.

3.1.2 Design Strategies

Focusing on the design-based approach, we will briefly describe and compare the principal adjustment methods and then outline the particular one used in this study.

- **Matching**: the idea behind matching methods is to select a subset of the sample units such that the observed covariates are balanced across treatment and comparison *(sub)*groups. In literature, there are many different techniques of matching that differs mainly regarding the number of matches in the control group for one treated unit, what is considered to be a match (namely, the *closeness* of the matches) and the matching algorithm used. Whatever the choices on the aspects described above, matching will most likely result in many controls individuals to be discarded and a consequently lost in precision of the treatment effect estimate. Further description of matching methods is out of the scope of this work, we refer the reader to Greifer and Stuart (2021) for more details.

- **Stratification**: in this case the sample is divided into *mutually exclusive* strata of individuals who are similar in pre-treatment characteristics based on *quantiles* of the propensity scores. The idea is that, within each strata, the distribution of the observed covariates is expected to be roughly identical. Problems with such techniques arises from the choice of the optimal number of strata that should be chosen in order to achieve strata small enough to provide adequate balance, but large enough to get stable within-stratum estimates of the causal effect. Poor choices of the number of strata might lead to many strata with few treated or control units, decreasing the precision of the final estimate.

- **Weighting**: adjustments through weighting involve direct use of the estimated propensity scores to assign weight to each individual. The main idea is to weight all the units in order that the *adjusted* sample shows balanced distributions of the observed covariates between treated and untreated groups. Differently from the other methods presented above the weights for each units may vary depending on the causal estimand of interest and the technique make a more efficient use of all the data available.

3.2 Propensity Score Weighting

We will now outline the main characteristics of the *propensity score weighting* (Rosenbaum, 1987) methodology, posing particular attention on the advantages that such procedure has over the other methods and explaining how the methods works in both the binary and the multinomial treatment setups. The choice of weighting rely on the following advantages when compared with other adjustment techniques:

- **Efficient use of the available data**: unlike matching, which may discard unmatched units, weighting can include all the observation in the analysis. This is particular important when many units are unmatched since the procedure will result in great loss of sample size and increased variance of the effect estimate. Such gain in efficiency is true, for other reasons, even when comparing weighting and stratification methods. As matter of fact, stratification might suffer from loss of information due to the grouping continuous scores into discrete categories while weighting can take advantage of the entire continuous distribution of the propensity scores.
- **Flexibility**: Weighting methods, particularly IPTW, can be adapted (changing the way in which units are weighted) to balance the treatment groups for different causal estimand (ATE, ATT, ATC). Matching estimates, on the other side, are more specific to the matched sample and can be considered to estimate average treatment effects for the treated if only matched pairs are analyzed.
- **Robustness**: Matching is sensitive to the choice of matching algorithm and parameters (e.g., caliper width, matching ratio). Poor choices can lead to poor balance and biased estimates. Weighting is less sensitive to these choices.

On the other side weighting make direct use of the estimated propensity score and this underline the importance of precise estimates of the scores. Moreover, if extreme weight are present, these might increase variability and introduce numerical instability into the causal effect estimation.

3.2.1 Balancing Groups through Weighting

Propensity score weighting, specifically inverse probability of treatment weighting (IPTW), adjust the contribution of each unit to the analysis based on their propensity score. The aim of such method is to create a *synthetic* sample where the distribution of covariates is equal, or as close as possible, across treatment groups reducing or eliminating confounding.

Given a finite sample of N individuals, indexed by $i = 1, 2, ..., N$, let T_i denote the *observed* treatment indicator for the ith individual. For sake of generality we take $T_i = t$ if individual i was observed under treatment t, where $t \in 1, 2, ..., M$ and M is the number of all the possible treatments.

Under the potential outcomes framework defined in Chapter 2 each unit has M potential outcomes, but only one corresponding to the observed treatment indicator (T_i). For a given unit i, $T_i = t$, then $Y_i = Y_i[t]$ is the observed outcome. Let $\mathbf{X}_i$ denote the vector of K observed pre-treatment covariates of unit i.

Following from Definition (3), provided that N is large enough and that Assumption (2) holds, propensity score weighing allows us to estimates a nearly unbiased treatment effect.

As stated above, depending on causal estimand of interest, the method allows us to estimate different set of weights and adjust the sample consequently.

Weights for ATE

Assuming a binary treatment $T = \{t', t''\}$ and the interest of the study being the comparison of the mean outcomes had then entire population been observed under one treatment t', versus had the entire population had been observed under another treatment t'', we would estimate the ATE (Equation (2.7)) as

$$\begin{aligned} \hat{E}(Y[t'] - Y[t'']) &= \hat{E}(Y[t']) - \hat{E}(Y[t'']) \\ &= \hat{\mu_{t'}} - \hat{\mu_{t''}}, \end{aligned} \tag{3.1}$$

whether the expectation is over the entire population at hand.

In this case we would weight the exposed (with treatment $T = t'$) units with weights equal to

$$w_i[t] = \frac{1}{\hat{e}_{t'}(\mathbf{X_i})}, \tag{3.2}$$

where $\hat{e}_{t'}(\mathbf{X_i}) = \hat{\Pr}(T_i = t'|\mathbf{X}_i)$ are the *estimated* propensity scores for unit i.

From the notation used we see how this procedure can be naturally extended to multiple treatment scenario where $t = 1, 2, ..., M$ allowing us to produce an unbiased estimate of what McCaffrey *et al.* (2013) defined *pairwise ATEs.*

Weighting for ATT

When the research focus is the effectiveness of one treatment on the population that its targets, we are interested the ATT. Again, starting with a binary setup we consider $T = \{t', t''\}$ and we aim to estimate the ATT of t' relative to t''. The idea is to compare the mean outcome of units treated with t' when they received the treatment they actually received (t') with the mean outcome they would have had if treated with t''. Formally

$$\begin{aligned} \hat{E}(Y[t'] - Y[t'']|T = t') &= \hat{E}(Y[t']|T = t') - \hat{E}(Y[t'']|T = t') \\ &= \hat{\mu}_{t',t'} - \hat{\mu}_{t',t''}, \end{aligned} \tag{3.3}$$

where $\hat{\mu}_{t',t''}$ for $t' \neq t''$ estimates the mean outcomes of units for the treatments they did not receive, that being, the *conterfactual* mean.

It is clear that the definition of ATT depends on what we define as "the treated". This group would indeed receive weight 1, while the "comparison" group is weighted with $\frac{\hat{e}_i}{1-\hat{e}_i}$. This choice underlines that in this case the objective of the adjusting procedure is to make the comparison group *look alike* the treated group.

The multiple treatment setting, on the other side, requires more careful explanation. The main difference lie in the definition of the quantity $1 - \hat{e}_i$, namely the probability of *non* receiving the treatment, that would be expressed in terms of probability of receiving *another* treatment, allowing us to compute the *pairwise ATTs* (McCaffrey *et al.*, 2013).

In this case, the appropriate weight for individuals receiving treatment t'' is the ration between the probability of receiving treatment t' and the probability of receiving the treatment t'' for $t'' \neq t'$ and $t', t'' = 1, 2, ..., M$

$$w_i[t', t''] = \frac{\hat{e}_{t'}(\mathbf{X}_i)}{\hat{e}_{t''}(\mathbf{X}_i)}. \tag{3.4}$$

Intuitively, individuals with pre-treatment characteristic that are much more common in their own treatment group than in the *target* group have small weights since they are less representative of the target group and vice versa. In this way we give more importance to units that closely resemble the units of interest making them *heavier* in the general comparison between groups.

3.3 Propensity Scores Estimation for Weighting

In practice, propensity scores are unknown and must be estimate from the data. Provided that Assumption (2) holds, the estimated propensity scores, and the respective weights, can produce better effect estimate even compared with situation in which $e(\mathbf{X}_i)$ are known. As discussed by Rosenbaum (1987) this occur because weighing by estimated propensity scores "compensates for both systematic and chance imbalances."

In this study we aim to understand the causal effect that different pattern of non-physical contacts had on the perceived loneliness of people who actually show such patterns, therefore, we will now outline the technique implemented to estimates the propensity scores relative to the ATT estimand in a multinomial setup. The connection with the more simple case of a binary treatment is straightforward.

3.3.1 Standard Approach

In the *standard* approach to propensity score weighting *multinomial logistic regression* with the treatment indicator as outcome have been widely used to modeling the relationship between the covariates and the assignment mechanism. Multinomial logistic regression models the probability that an outcome equals each of its possible values as a function of a linear combination of the covariates, their products and cross product (Salvan *et al.*, 2020). In a model for the treatment assignment T_i this would be:

$$\begin{aligned}\Pr(T_i = t|X_i) &= \frac{e^{\beta_t' X_i}}{1+\sum_{t'=1}^{M-1} e^{\beta_{t'}' X_i}}, \quad t = 1, \ldots, M-1,\\ \Pr(T_i = M|X_i) &= \frac{1}{1+\sum_{t'=1}^{M-1} e^{\beta_{t'}' X_i}},\end{aligned}$$

where $\boldsymbol{\beta_t},\ t = 1, 2, ..., M$ are unknown and estimated from the data and $\Pr(T_i = M|X_i)$ is assumed as baseline. Once we estimate the unknown parameter from the data we can generate $\hat{\Pr}(T_i = t|X_i) = \hat{e}_t(\mathbf{X_i})$ and use it to weight our sample as described by Equation (3.4). The challenge with this approach is choosing the correct set of interaction and polynomial terms among the covariates to capture any *nonlinearities* in their relationship with the treatment assignment.

Based on the method proposed by Zanutto *et al.* (2005) for selecting the correct form of the propensity score model in the context of stratification, we outline a modified approach that might be used in the context of IPWT for multiple treatments (McCaffrey *et al.*, 2013). The iterative approach involves the following 6 steps:

1. fit a simple model with only main effects for all the possible confounders,
2. for $T = 1, 2, ..., M$ find the areas of common support for $\hat{\boldsymbol{\beta_t}}' \mathbf{X}_i$ among all the treatment groups and retain only the observation in those areas (ensuring overlap or positivity),
3. test whether treatment indicator is a significant predictor of each covariate running a one-way ANOVA with each observation weighted by its IPTW. If the $M-1$ degree-of-freedom F-test is significant the covariate is not balanced,
4. Add to the model polynomial and interactions terms for covariates that do not balance,
5. fit the new model,
6. repeat steps 3-5 until all covariates are balanced.

Challenges and Drawbacks

One of the key challenge of the method proposed above is the level of variable selection required by step 4 in order to successfully obtain the correct propensity scores model. With multiple covariates and a model for each treatment level, there are many possible interactions and polynomial terms to add any time covariates do not balance that would be very time consuming, or even prohibitive, for the researcher to try all possible variation. On the other side, because to the modification to the significance test due to the weighting, the F-test might fail in finding group differences significant.

Precision-wise, logistic regression has been shown to yield high MSE for the causal effect estimates (Lee *et al.*, 2010). Moreover, because of the underling linearity assumptions, logistic regression can lead to very small probabilities and extreme large weights that would exacerbate the variability and the numerical instability of the weighting procedure.

3.3.2 Machine Learning Approach

Due to the direct incorporation of the actual propensity scores in the weighting procedure, precision and numerical stability of such estimates are crucial for the weightings methods. Furthermore, accurate treatment effect estimates require the propensity score model to account for all the potential confounders and, as demonstrated by Drake (1993), propensity score model misspecification can lead to substantially biased causal effect estimates.

Given the clear drawbacks and limitation of the method described above, in the past decades much research has been done in estimating the propensity scores through *machine learning* techniques. Such techniques do not rely on a specific *data model* (as logistic regression and its extensions) but try to extract the relationship between outcome and predictors form the data itself.

This *automatic* selection procedure make such algorithms better suited to avoid propensity scores model misspecification, enabling them to *learn* from the data the correct set of interaction and polynomial terms to capture the functional form of the assignment mechanism given the covariates. As result, machine learning algorithms proved to produce more numerically stable propensity scores and subsequent weights, achieve better balance between treatment and comparison groups and, therefore, reduce the bias in the treatment effect estimates (McCaffrey *et al.* (2004); Cannas and Arpino (2019)).

As demonstrated by the work of Lee *et al.* (2010), in the context of machine learning methods one algorithm showing promising results is *Generalizes Boosted Model* (GBM, Ridgeway (1999)).

3.4 Gradient Boosted Regression

Boosting is a general, *automated*, *data adaptive* modeling algorithm that can be used with a large number of covariates to fit a nonlinear predictive surface for the outcome. Many *flavours* of boosting have appeared in the statistical and machine learning literature including the original AdaBoost algorithm (Freund and Schapire, 1997), generalized boosted models (Ridgeway, 1999) and the gradient boosting machine (Friedman, 2001). Such methods are famous since they generally outperform alternative algorithms in term of predictive error (Friedman (2001); Ridgeway (1999)). Boosting algorithms are particularly effective when the model involves a large set of covariates and, between the others, generalized boosted models is tuned to produce models yielding well-calibrated probability estimates matching the empirical probabilities.

GBM adds together many simple functions to estimate a smooth function of a large number of covariates of mixed types. Each individual simple function lacks smoothness and is a poor approximation to the function of interest, but together they can approximate a smooth function. In the implementation of GBM, each simple function is a *regression tree* with limited depth.

We will now provide a brief outline of the algorithm starting from the description of his constituent block (referred to as *weak learner*), the regression tree (CART, Breiman *et al.* (1984)), followed by an in depth explanation of the GBM algorithm and its implementation in propensity scores estiamtion.

3.4.1 Regression Trees (CART)

Tree-based methods partition the covariate space into distinct regions, fitting a very simple model on each one, to obtain regions of constant prediction. For simplicity and ease of interpretation, *"splits"* parallel to the axes are used to form p-dimensional rectangular regions (Figure 3.1).

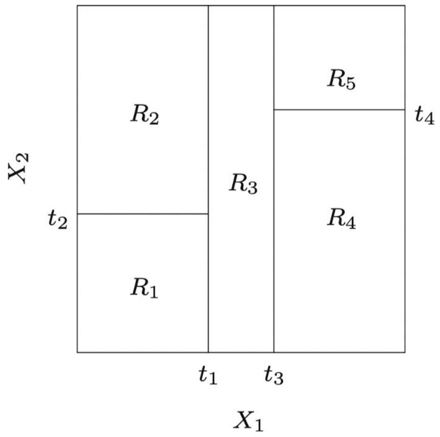


Figure 3.1: Partition of the feature space by CART (Hastie *et al.*, 2009).

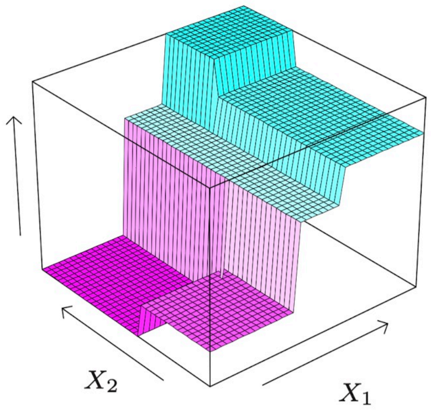


Figure 3.2: Prediction surface (Hastie *et al.*, 2009).

The estimation algorithm for regression trees is known as *recursive binary splitting* and consists of two steps:

1. Divide the covariate space into J non-overlapping regions $R_1, R_2, ..., R_J$ such that the *residual sums of squares* (referred to as RSS) given by:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{3.5}$$

is minimized.

2. For every observation that falls into the region R_j, assign the same prediction, which is the average response of the observations within the jth partition:

$$\hat{f}(x) = \hat{y}_j = \text{ave}(y_i | x_i \in R_m). \tag{3.6}$$

In most cases, it is computationally infeasible to calculate every possible partition of the space into J *rectangles*, and the algorithm select the best *split* using a *greedy* top-down approach.

Starting with all observations belonging to a single region, a variable X_j is considered as an axis and a value s from its support as a split point, such that the partition into two regions

$$R_1(j,s) = \{X | X_j \leq s\} \text{ and } R_2(j,s) = \{X | X_j > s\}$$

leads to the greatest possible reduction in terms of RSS at the current step rather than at a future step.

Among all predictors $X_1, X_2, ..., X_p$ and all possible values s for each of them, the predictor and the respective value that minimize the equation:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{3.7}$$

is selected.

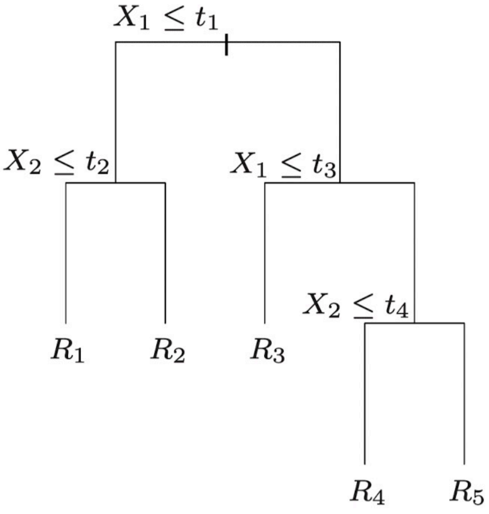


Figure 3.3: Resulting tree (Hastie *et al.*, 2009).

The operation is repeated to further divide the data by minimizing the RSS across all resulting regions until a certain stopping condition is reached.

Such a method is particularly handy when we have to deal with a large number of covariates even when these are correlated between them or with the outcome. On the other side, such algorithm is particularly prone to *overfitting* and this is especially true for large tree which fit the data too close producing highly variable estimates and lack smoothness, often resulting in poor propensity scores' estimates. For further details, see Breiman *et al.* (1984), Hastie *et al.* (2009).

3.4.2 Generalized Boosted Models

In this work we implemented generalized boosted models, a *multivariate*, *non-parametric* regression techniques to estimate propensity scores. The main idea behind GBM is to build a piece-wise constant function flexible enough to approximate a smooth curve by iteratively combining simple regression trees, overcoming many of the problem of the regression tree algorithm.

We will now describe the algorithm in the case of the propensity score estimation. Let t_i denote the treatment indicator and $p(\mathbf{x}_i)$ the propensity score for unit i. Theoretically speaking, following the likelihood principles the best estimate of the propensity score $p(\mathbf{x})$ is the one that maximize the expected Bernoulli log-likelihood function

$$E(\ell(p)) = E(t_i \log p(\mathbf{x}_i) + (1-t_i)\log(1-p(\mathbf{x}_i))|\mathbf{x}_i). \tag{3.8}$$

For computational simplicity and to ensure that $p(\mathbf{x})$ will vary inside the range $[0,1]$ the algorithm models the log-odds of treatment assignment

$$g(\mathbf{x_i}) = \log\left(\frac{p(\mathbf{x_i})}{1-p(\mathbf{x_i})}\right),$$

that substituted into Equation (3.8) gives us the log-likelihood in terms of the regression function $g(\mathbf{x})$

$$E(\ell(g)) = E(t_i\mathbf{g}(\mathbf{x}_i) - \log(1+\exp(\mathbf{g}(\mathbf{x}_i)))|\mathbf{x}_i). \tag{3.9}$$

Equation (3.9) will yield relatively large values when we have agreement between $g(\mathbf{x}_i)$ and t_i, that is for $t_i = 0$ when $g(\mathbf{x})$ is negative or $t_i = 1$ when $g(\mathbf{x})$ is positive.

The algorithm assumes as initial state of Equation (3.9) $\hat{g}(\mathbf{x}) = \log(\bar{t}/(1-\bar{t})$, where $\bar{t}$ is the average treatment assignment indicator for the entire sample, and try to iteratively search for a small adjustment, $h(\mathbf{x})$, that can improve the propensity score model's fit to the data. When it finds $h(\mathbf{x})$ such that

$$E(\ell(\hat{g}+\lambda h)) > E(\ell(\hat{(g)}))$$

it updates the current guess as

$$\hat{g}(\mathbf{x}) \leftarrow \hat{g}(\mathbf{x}) + \lambda h(\mathbf{x}) \tag{3.10}$$

for some step size λ.

The *gradient* of Equation (3.9) with respect to $g(\mathbf{x})$ indicates the local "direction" providing the greatest increase in the expected log-likelihood. Therefore, Friedman (2001) suggests to take

$$\begin{aligned} h(\mathbf{x}) &= \frac{\partial}{\partial g(\mathbf{x})} E(L(g)) \\ &= E\left(t - \frac{1}{1+\exp(-g(\mathbf{x}))} \middle| \mathbf{x} \right) \\ &= E(t - p(\mathbf{x}) | \mathbf{x}). \end{aligned} \tag{3.11}$$

as adjustment to the current g.

Equation (3.11) is actually the difference between the treatment indicator and the propensity score, we can think about it as some kind of "residual" between the actual data and the estimates achieved so far. This residual cannot be computed without knowing the probability of assignment to the treatment but can be estimated, from our sample, with a flexible (enough) least squares regression procedure.

To find $h(\mathbf{x}_i)$ we use a regression tree algorithm (CART, as explained in the previous section) and obtain a nonparametric estimates of the adjustment. Now, one approach would be to predict the quantity $(t - p(\mathbf{x}))$, from $\mathbf{x}$, as the mean of each partition residuals and then find the value of λ that offer the greatest increase in the log-likelihood. Friedman (2001) suggest, instead, to solve for the best update separately for each region of the residuals feature space (identified by the terminal nodes of the regression tree).

In this way the tree will partition the individuals of the sample, based on the values of their covariates, and the optimal adjustment to $\hat{g}(\mathbf{x})$ can be found separately in each region solving

$$\begin{aligned} h(\mathbf{x}) &= \underset{\theta}{\operatorname{argmax}} \sum_{x_i \in T_k} t_i \log(\hat{g}(\mathbf{x}_i + \theta)) - \log(1 + \exp(\hat{g}(\mathbf{x}_i) + \theta)) \\ &\approx \frac{\sum_{x_i \in T_k} t_i - p(\mathbf{x}_i)}{\sum_{x_i \in T_k} p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))}. \end{aligned} \tag{3.12}$$

for all units that fall into the kth partition, $\mathbf{x} \in T_k$.

The variability of $\hat{g}(\mathbf{x})$ can also be reduced using a *shrinkage* coefficient $\alpha \in (0, 1]$. Smaller values of α allow the algorithm to *slowly* fit the data, increasing the number of iteration needed to produce good propensity score's estimates but, as empirical evidence shows (Friedman, 2001), improving the overall fit of the model. This shrinkage strategy reduces the variance (problematic for other ML methods such as regression trees) without necessarily increasing the bias. The number of iterations determines the model's complexity and must be determined from the data. At each iteration the model becomes more complex, fitting additional features of the data and, with enough iteration, $\hat{g}(\mathbf{x})$ is flexible enough to *overfit* the data.

Since balance between pretreatment characteristic is our goal, we will then select the optimal iteration to be the one that maximize an external balance criterion (such as ASAM or Kolmogorv-Smirnov statistic).

Algorithm 1 Gradient Boosted Logistic Regression

```
Initialize $\hat{g}_0(\mathbf{x}) = \log \frac{\bar{t}}{1-\bar{t}}$
for $m = 1 \rightarrow M$ do
    Let $r_i = t_i - \frac{1}{1+\exp(-\hat{g}_{m-1}(\mathbf{x}_i))}$
    Construct a tree structured predictor of $r_i$ to partition the features into terminal nodes $T_1, \dots, T_k$.
    Compute the updates for each terminal node:
        $\theta_k = \frac{\sum_{x_i \in T_k} t_i - p(\mathbf{x}_i)}{\sum_{x_i \in T_k} p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))}$
    Update the logistic regression model as:
        $\hat{g}_m(\mathbf{x}) \leftarrow \hat{g}_{m-1}(\mathbf{x}) + \alpha\theta_{k(\mathbf{x})}$
    where $k(\mathbf{x})$ determines to which terminal node an observation with features $\mathbf{x}$ belongs.
end for
```

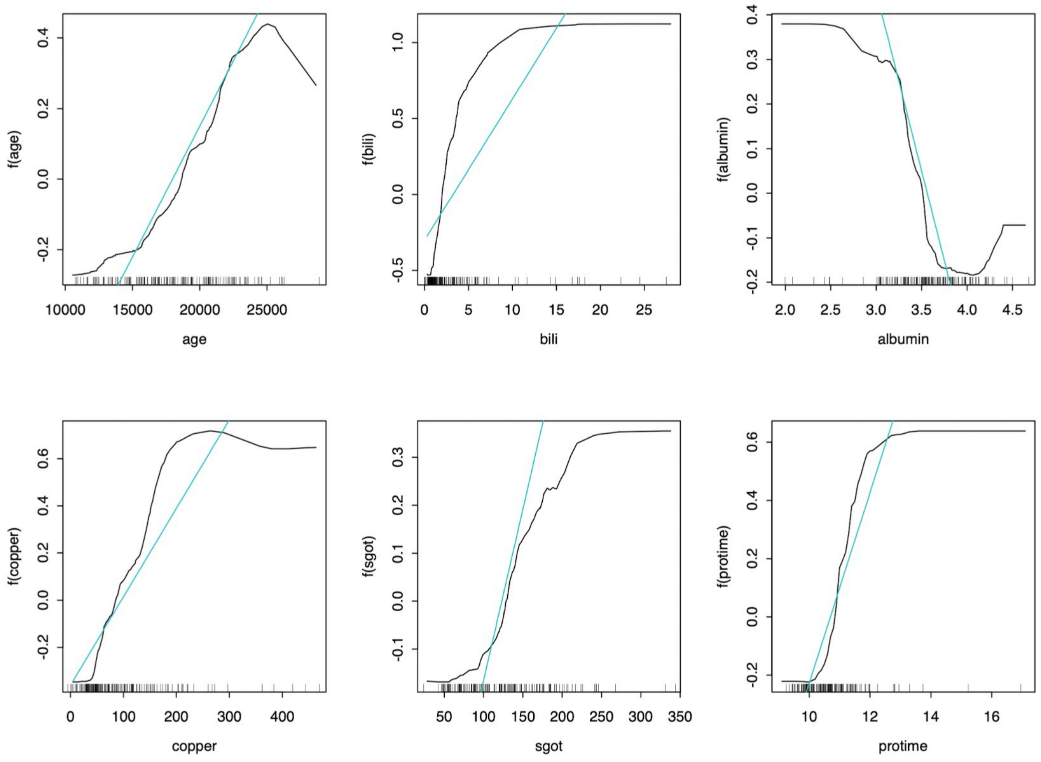


Figure 3.4: Boosted estimate (black curve) versus estimate from a linear model (blu line) of the main effects on sample data (Ridgeway, 1999)

3.4.3 GBM for Balancing

Following closely the recommendations provided by McCaffrey *et al.* (2013) we implemented GBM in order to estimate the propensity scores which weights achieved the best balance across groups. After a review of the GBM method we will now discuss how measures of imbalance can be optimized as *stopping rules* of the algorithm in order to eliminate or reduce confounding.

Binary treatment

The key to achieve unbiased estimate of the ATT in a binary setup is to use GBM in a iterative fashion. Ensure that the model can fit the data closely and select the optimal number of iteration (and, therefore, the optimal number of trees) in order to *minimize* a criterion based on the difference between weighted distributions of the pretreatment covariates in the two groups. In practice, we have two main stopping rules:

- **Absolute standardized mean difference (ASD)**: the absolute value of the difference between the weighted mean for treatment and control group, divided by the unweighted standard deviation of the *treatment* group (for ATT)

$$ASD_k = \frac{\left|\bar{X}_{1k} - \bar{X}_{0k}\right|}{\hat{\sigma}_{k1}}, \tag{3.13}$$

  where $\bar{X}_{kt}$ is the weighted mean of the kth covariate for treatment ($t = 1$) or control ($t = 0$) and $\hat{\sigma}_{k1}$ is the standard deviation for the treatment group of the covariate. Note that, when computing ATT, units of the treatment group have weight equal to one, therefore $\bar{X}_{kt}$ is the actual mean of the covariate for that group.

- **Kolmogorov-Smirnov statistic (KS)**: defined as the upper bound of the absolute difference between the empirical distribution functions for the treatment and control samples

$$KS_k = \sup_x |EDF_{1k}(x) - EDF_{0k}(x)|. \tag{3.14}$$

  For covariate k,

$$EDF_{tk}(x) = \frac{\sum_{i=1}^{n} w_i[t]T_i[t]I(X_{ik} \leq x)}{\sum_{i=1}^{n} w_i[t]T_i[t]},$$

  for $t \in \{0, 1\}$.

Generally speaking, values equal or above 0.2, are considered "large". In this work will take a more conservative approach, considering all values greater than 0.1 as symptoms of imbalance,for both ASD an KS

Multinomial Treatment

McCaffrey *et al.* (2013) suggest to tuning the GBM fits specifically for the causal estimand of interest yield better balance and subsequent treatment estimation. Therefore, for estimating the weights for the pairwise ATTs of t' relative to t'' the method is to fit the treatment indicator for $T = t'$ using only the subsample with $T \in \{t', t''\}$ using the stopping rules for estimating ATT with binary treatment and then weight the unit in the comparison group (t'') with the ATT weights defined by Equation (3.4) for binary treatment. The procedure is repeated for all $t'' \neq t'$.

Such approach follows from the fact that our target population is a particular treatment group (t') so we would like to find weight that make the covariate distributions for each other treatment group *match* the distributions of the t' group. Differently from the ATE procedure, it is not necessary to have all conditions in the same estimation model for all the treatments and a pointed comparison between pairwise treatments is sufficient to yield ATT weights with superior balance for each comparison.

Effective Sample Size

Another important factor to take into account is the increase in sampling variance due to the weighting procedure. The *Effective Sample Size* (ESS) is a conservative way to capture such increase and its impact on precision.

Specifically,

$$ESS_t = \frac{\left(\sum_{i=1}^{N} T_i[t]w_i\right)^2}{\sum_{i=1}^{N} T_i[t]w_i^2}, \tag{3.15}$$

for treatment t and w_i defined by Equation (3.4), provide a useful measure of disparity between the weights of the treatment group's sample and the comparison one. Very small values of ESS_t relative to the actual sample size indicate that a small number of units receive very high weights compared to the majority of the sample.

3.4.4 GBM hyper-parameter tuning

Like all the other *ensemble* methods, GBM is particularly sensible to *hyper-parameter* tuning. Unlike *statistical* parameters, that can be *learned* from the data when fitting the model, hyper-parameters are set prior to the training, cannot be estimated from the data and dictate the *behavior* of the learning algorithm itself.

The choice of hyperparameters can greatly impact the performance of a machine learning model. Good hyperparameter settings can make the algorithm converge faster and achieve better accuracy, while poor choices can lead to *overfitting* or *underfitting.* Therefore, careful tuning is important to optimize these values for the best performance on a given task.

In the GBM algorithm the parameters that controls learning process are:

- **number of iterations** (or regression trees): such value is crucial for governing the overfitting - underfitting trade-off. Values that are too small result in too general estimate of the dependant variable. The modeling algorithm did not had *time* to learn the correct functional relation between the response and the covariates and does not capture important features of the data. On the other side, with too many iterations we generally end up with overfitting the data.

- **shrinking parameter** α: as explained in Algorithm (1) smaller values of α improve the overall fit of the model and the smoothness of the predictive surface at cost of more iteration. There is a clear relationship with the shrinkage parameter and the number of iteration.
- **interaction depth**: this parameter dictate the maximum *depth* (or number of *splits*) of every tree added to the boosted regression model. Basically allowing for different level of interaction between the covariates.

  If we allow each tree to have only one split the update defined by Equation (3.10) will look like:

  $$\hat{g}(x) = g_0 + g_1(X_1) + g_2(X_2) + g_3(X_2) + g_4(X_3) + g_5(X_1) + \dots \tag{3.16}$$

  . Since the algorithm adds terms sequentially, different variables could appear many times. Grouping together those tree that split on the same variable will result in an additive model allowing the algorithm to approximate any kind of curve.

  If we allow each tree to have two or more splits then we are fitting an additive model with two or more interaction, respectively.
- **bag fraction**: following from Friedman (2002) GBM select at each iteration a *different* random subsample of the data and use only that portion to estimate the adjustemnt from Equation (3.12). The optimization of this stochastic component may result in more robust estimates with lower bias and variance.

Griffin *et al.* (2017) have shown empirically that when using ML methods for the task of estimating the propensity scores, optimizing the algorithm while "chasing balance" rather than model fit provide lower bias and variance of the treatment effect estimates, compare with the classical *ML-workflow*.

Practically speaking, instead of select the hyper-parameters that yield the best prediction out-of-sample (using cross-validation or out-of-bag estimates) we focus on the set of parameters that achieve the optimal balance with respect to causal estimand of interest (ATT in our case) selecting the values that minimize either Equation (3.13), either Equation (3.14) across all the covariates considered.

3.5 Outcome modeling

In this work we implemented both a *design*- and a *analysis*-based approach. In particular, we used propensity scores weighting to balance the covariate distributions in the different treatment groups and we computed the actual causal effect (ATT) through *G-computation* (also referred to as *G-formula*, Robins (1986)).

Effect Measure

A fundamental step of analysis stage is to decide how to measure the actual causal effect. Given the binary nature of the outcome of interest we could compute the effect as *risk-difference* (RD), *risk-ratio* (RR) or *odd-ratio* (OR).

In this work, the effect is measured using the risk-difference that closely resemble the form of the ATT expressed by Equation (2.15).

3.5.1 Doubly Robust Approach

The *doubly robust* (DR for short) approach aim to produce an unbiased estimator of the treatment effect if either the model for the treatment assignment mechanism or the model for the counterfactual data were correct, but not necessarily both. Doubly robust estimation can also be more efficient than the simple weighted estimator, such as IPWT (Neugebauer and van der Laan, 2005) and it can be used even if all the covariates balance to provide protection against possible errors in the propensity score model (McCaffrey *et al.*, 2013).

Doubly Robust Estimator

The approach implemented in this work is trying to mimic the property of the *doubly robust estimator* for ATT while maintaining a certain degree of separation between the design stage and the effect estimation.

For the sake of completeness we will briefly introduce the doubly robust estimator. Starting from Equation (2.9) we know that $E(Y|T = 1)$ is identifiable directly from the data while the second term $E(Y(0)|T = 1)$ is counterfactual. Under the strong ignorability (Assumption (2)) we can write the counterfactual mean as a function of the propensity score model $e(X, \alpha)$, the treatment assignment T and the outcome model $\mu_0(X, \beta_0)$ (in this case assumed to be a parametric regression model of the covariates and the outcome)

$$\tilde{\mu}_{0T}^{dr} = \frac{E\left[o(X,\alpha)(1-T)\{Y-\mu_0(X,\beta_0)\} + T\mu_0(X,\beta_0)\right]}{e}, \tag{3.17}$$

where $e = pr(T = 1)$ is the marginal probability of the treatment and $o(X, \alpha) = e(X, \alpha)/(1 - e(X, \alpha))$.

Theorem 2 *Under Assumption (2), if either the propensity score model $e(X, \alpha) = e(X)$ or the model for the potential outcome $\mu_0(X, \beta_0) = \mu_0(X)$ are well specified, then $\tilde{\mu}_{0T}^{dr} = E\{Y(0)|T = 1\}$.*

From Theorem (2) we can obtain, based on the data $(X_i, T_i, Y_i)_{i=1}^n$, a doubly robust estimator for τ^T with the following procedure:

1. obtain the fitted values of the propensity scores $e(X_i, \hat{\alpha})$ and then obtain the fitted values of the odds $o(X_i, \hat{\alpha}) = e(X_i, \hat{\alpha})/(1 - e(X_i, \hat{\alpha}))$;

2. obtain the fitted values of the outcome mean under control $\mu_0(X_i, \hat{\beta}_0)$;
3. construct the doubly robust estimator: $\hat{\tau}_T^{dr} = \bar{Y}(1) - \hat{\mu}_{0T}^{dr}$, where

$$\hat{\mu}_{0T}^{dr} = \frac{1}{n_1} \sum_{i=1}^{n} \left[o(X_i, \hat{\alpha})(1 - T_i)(Y_i - \mu_0(X_i, \hat{\beta}_0)) + T_i \mu_0(X_i, \hat{\beta}_0) \right]. \tag{3.18}$$

In particular, we estimates the weights in the design stage using the GBM algorithm as $o(X, \hat{\alpha})$ and the potential outcomes using g-computation in the analysis stage as $\mu_0(X_i, \hat{\beta}_0)$.

3.5.2 G - computation

We implemented G-computation as our framework of choice due to its generality. Such method works the same regardless the form of the outcome model or the type of outcome variable (whether continuous or binary) and can work with sample adjusted by both weighting and matching methods. Moreover, G-computations enable us to compute an analytic approximation of the variance of our effect estimate. Nevertheless, this approach overlooks the variability inherent in the estimated propensity scores, which is a critical drawback. Despite this limitation, it has been established that the analytic approximation yielded by G-computations acts as an upper bound for the actual variability of the procedure (McCaffrey *et al.*, 2004). This provides a conservative estimate of the variability, circumventing the necessity for computationally demanding techniques such as bootstrapping.

Implementation

The first step of G-computation is to fit a regression of the outcome on the exposure and relevant covariates, using the observed data set. This regression model is frequently called the *Q-model* and this not different than a traditional regression of a dependent variable on a set of covariates. Such model is used to predict the counterfactual outcomes for each observation under each of the treatment level plugging in $T = t$, with $t = 1, 2, ..., M$ and M the number of treatments, into the regression model. Once the counterfactual outcomes Y_t are estimated we have a full data set free of confounding (given the correct specification of the Q-model) and we have finally resolved the *missing data problem* described in Chapter 2.

Having generated the full data we subsequently fit a marginal structural model (MSM) on the outcome Y_t for each of the treatment t:

$$E(Y_t) = \beta_0 + \beta_1 t. \tag{3.19}$$

The final step is to compute the risk difference $E(Y_{t'} - Y_{t''})$ for every $t' \neq t''$ considered. Since G-computation is thought to estimate marginal causal effect and our causal estimand of interest is ATT, we only estimate potential outcomes (under both treatment and control) using a sub-sample with only the *treated* units. For further details, see Snowden *et al.* (2011).

3.5.3 Weighted Logistic Regression as Q-model

When doing g-computation after weighting, the outcome model should be fit incorporating the estimated weights in this study we implemented, as our Q-model, a weighted linear regression.

The model assumes a sample of n units drawn by a finite population of N units, a given set of observation $\{(y_i, \mathbf{x}_i);\ i \in S\}$ where y is the response variable and $\mathbf{x}$ is a p-dimensional vector of possible explanatory variables. Such observation are associated with a set of weights $\{w_i;\ i \in S\}$. In our case the weights are the propensity score weight estimated by Equation (3.4).

The population is assumed to be a realisation of some probability model with density $f(Y|X;\beta)$, in which

$$g(E[Y|X = x]) = g(\mu) = \eta = \mathbf{x}'\beta, \quad \beta \in R^p \tag{3.20}$$

where $g(\cdot)$ is the logistic function and marginal variance given by

$$Var[Y|X = x] = \sigma^2 V(\mu). \tag{3.21}$$

The model assume β_0 as the true parameter value of the *super*population, namely the entire population from which the sample have been drawn, $_N$ for the maximum quasi-likelihood estimator of β_0 that would be obtained from the full population (the *census*) and β^* is the true parameter for the limit in probability of $\hat{\beta}_n$ the estimator of the census parameter estimated from the sample.

Weighted Estimation

The classical *design-based* estimator solves

$$\begin{aligned}\hat{U}(\beta) &= \sum_{i=1}^{N} w_i R_i U_i(\beta) \\ &= \sum_{i=1}^{N} R_i w_i x_i \left(\frac{1}{g'(\mu_i)V(\mu_i)}\right)(y_i - \mu_i(\beta)) = 0,\end{aligned} \tag{3.22}$$

which are unbiased estimating equation for $\tilde{\beta}_N$ and the resulting estimate $\hat{\beta}_n$ is asymptotically normal and a consistent estimate for β_0.

The variance of $\hat{\beta}_n$ is the sum of two components: the finite-population sampling variance of $\hat{\beta}_n$ around $\hat{\beta}_N$, of order n^{-1}, and the model-based sampling variance of $\hat{\beta}_N$ around β_0, of order N^{-1}. When $n \ll N$, the latter term is often ignored.

In this setting, the variance of $\sum_{i=1}^{N} w_i R_i U_i(\beta)$ can be computed at $\beta = \hat{\beta}_n$ by the *Horvitz-Thompson* variance estimator (Horvitz and Thompson, 1952). A standard delta-method argument, gives the "sandwich" form $A^{-1}BA^{-1}$ for the estimated variance of $\hat{\beta}_n$, with

$$A = \sum_{i=1}^{N} w_i R_i \left.\frac{\partial U_i(\beta)}{\partial \beta}\right|_{\beta = \hat{\beta}_n} \tag{3.23}$$

and

$$B = \hat{Var}\left(\sum_{i=1}^{N} w_i R_i U_i(\beta)\right). \tag{3.24}$$

The relationship with our propensity score weighting problem is drawn by $R_i = T_i$ with T_i being our treatment assignment indicator and w_i given by Equation (3.4). A more depth discussion on the topic is out of the scope of this work, the reader is referred to Lumley and Scott (2017).

Chapter 4

Application

4.1 Research Question

Given its detrimental effects on health and well-being of the individual, loneliness represents a significant public health issue (Holt-Lunstad *et al.*, 2015). During the COVID-19 pandemic enforced isolation measures have exacerbated this condition, especially among older individuals (Macdonald and Hülür, 2021). Conversely, remote contact has increased, potentially helping to counteract the negative effects of reduce face-to-face interactions (Liddle *et al.*, 2020).

Through the lens of the *Intergen-COVID* survey this study explores the impact of non-physical contact patterns on perceived loneliness among older Italians, providing insights into potential strategies for public health interventions in similar future scenarios. In particular, our research question was:

> How would people *like* those incrementing non-physical contact experience the change in frequency of loneliness feelings, had they not increased such contacts?

To answer such question we estimated the *causal effect* of the increase in non-physical contacts on *severe* loneliness.

4.2 Intergen-COVID

Data are based on *Intergen-COVID* online survey carried out in Italy, France and Spain from 14th to the 24th of April 2020 through the market survey platform Lucid. Data were collected through *quota sampling* targeting people aged 18+ for a total sample of 9,186 individuals.

The survey aimed to investigate the indirect consequences of COVID-19 lockdown measures on people's lives, posing particular attention on four key domains:

1. relationship (inter-generational and other type, physical, non-physical),
2. living arrangements,

3. mental health
4. events experienced during lockdown (income or job loss, death of a friend/relative due to COVID-19) and future intentions (i.e., fertility, changes in marital status, retirement).

Furthermore, the survey examined whether the effect of the restrictive measure had an unequal impact on groups of population defined by key socio-demographic characteristics.

Among the other results, the survey pointed out a general increase in non-physical contact in all countries considered and about 35% of the respondents declared to have felt lonely more often than usual during pandemic (Arpino *et al.*, 2020).

4.3 Exploratory Data Analysis

In this analysis we focused on a sub-sample of $N = 1,573$ Italian individuals aged 50+. We will now describe the data used in the study posing particular attention on the distributions of the outcome, the treatment (in both binary and multiple setup) and some descriptive statistics on the potential confounders included in the analysis.

4.3.1 Outcome: Severe Loneliness

The dependent variable of the study is the measure of changes in perceived loneliness in the week previous to the survey.

Question 1 *During the past week, have you felt lonely:*

- *Rarely or none of the time (e.g., less than 1 day)*
- *Some or a little of the time (e.g., 1-2 days)*
- *Occasionally or a moderate amount of time (e.g., 3-4 days)*
- *All of the time (e.g., 5-7 days)*

In particular we focus on *severe* perceived loneliness and, therefore, the variable has been dichotomized taking value 1 if respondent reported to have felt lonely "all the time" and 0 otherwise.

From Figure 4.1 we notice that only 5% of the sample experienced severe loneliness in the period previous to the survey.

4.3.2 Treatment: Increase in non-physical contacts

The treatment variable is represented by the *increase* in non-physical contact. People were asked about their change in non-physical contacts during the COVID-19 lockdown with non-coresident individual in two different questions.

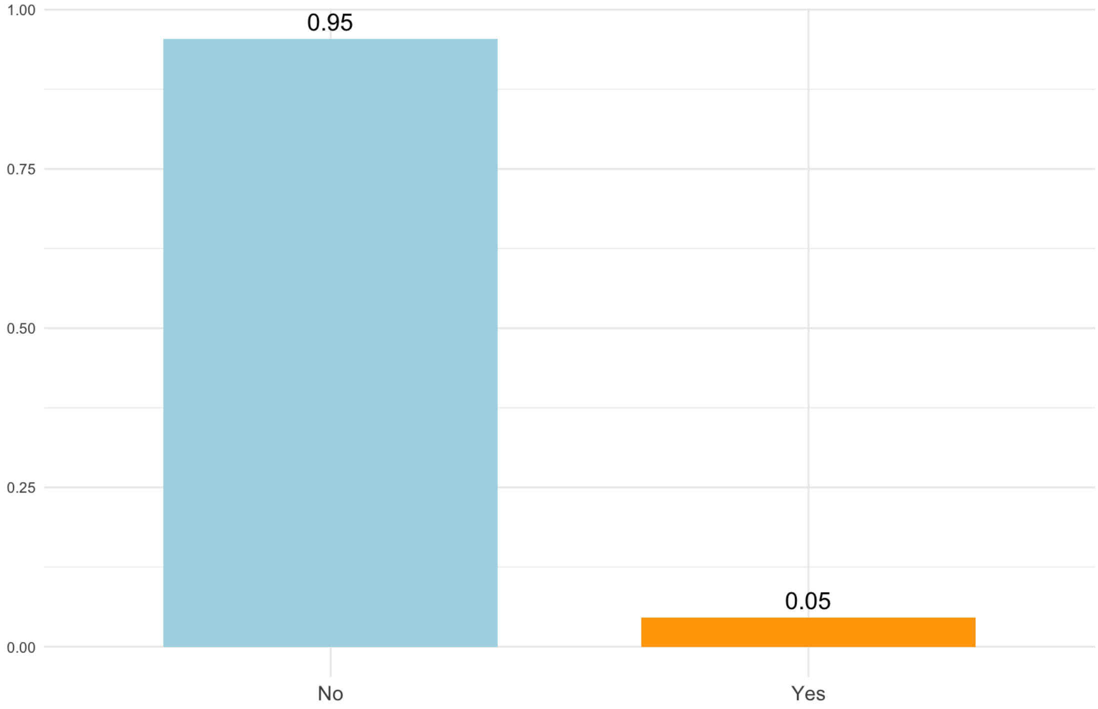


Figure 4.1: Distribution of *severe* loneliness among respondents.

Question 2 *Considering non-resident persons only, **since the entry into force** of the first nationwide restrictions due to the Coronavirus in your country, with whom did you **increase** the frequency of non-physical contacts (e.g. on the phone)?*

- *children (1 or more)*
- *grandchildren (1 or more)*
- *mother and/or father*
- *grandparents (1 or more)*
- *other relatives (e.g. siblings)*
- *friends (1 or more)*
- *I did not experience an increase in non-physical contacts with non-resident persons*

In the first part of our work we posed our attention on people who answered they had increased their remote contact with at least one non-coresident. We dichotomized the variable to take value 0 if the respondent answered "I did not experience an increase in non-physical contacts with non-resident persons" and 1 otherwise.

Question 3 *Considering non-resident persons only,* ***since the entry into force*** *of the first nationwide restrictions due to the Coronavirus in your country, with whom did you* ***decrease*** *the frequency of non-physical contacts (e.g. on the phone)?*

- *children (1 or more)*
- *grandchildren (1 or more)*
- *mother and/or father*
- *grandparents (1 or more)*
- *other relatives (e.g. siblings)*
- *friends (1 or more)*
- *I did not experience an increase in non-physical contacts with non-resident persons*

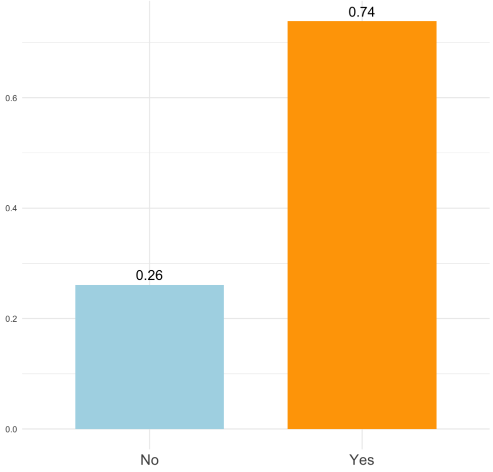


Figure 4.2: Distribution of changes in non-physical contacts among respondents (binary).

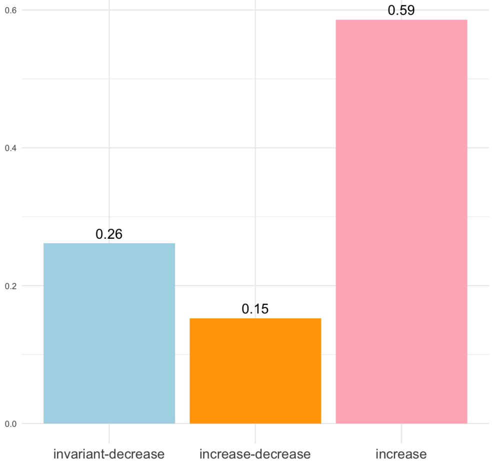


Figure 4.3: Distribution of changes in non-physical contacts among respondents (multinomial).

In the second setup we considered *mixed* contact pattern for which people might have increased their remote relations with someone and decreased or let them unchanged with someone else, therefore we considered a *multinomial* treatment of levels: *"increase"*, *"increase-decrease"*, *"invariant-decrease"*.

From both figures we notice how more than half of the individual in our sample increased their non-physical contacts during lockdown in both the setups. In particular for the multinomial treatment (Figure 4.3) we have 59% of individuals increasing their contacts against 15% and 26%, experiencing "increase-decrease" and "invariant-decrease" respectively. In a causal lens we can say that in both setups at least half of the sample underwent the treatment. Moreover, the time-frames in which the question regarding the outcome and the question regarding the treatment are referring deserve particular attention. In our data the treatment precede the cause, making clear the temporal relationship between cause and effect as required by the Rubin model. Such distinction is crucial for the observational studies and respect the design guideline exhorted by Cochran (1965).

4.3.3 Control Variables

Control variables include **socio-demographic variables** such as age (divided into three age groups: *50-59*, *60-69*, *70+*), gender, level of education (taking value *low* is defined as below secondary education, *medium* as up to high school, and *high* refers to a university education or above), employment status, economic situation, marital status, cohabiting situation (taking value 1 if the respondent answered of living with someone, 0 otherwise), neighborhood self-description, kin availability (taking value 1 if the respondent reported to have kin alive, 0 otherwise).

We considered **health-related** variables measured retrospectively to COVID-19 pandemic: respondent self-perceived health status (expressed by "very good", "good", "acceptable" or "poor"), a dummy variable taking value 1 if respondent reported suffering from any chronic diseases and 0 otherwise and baseline self-reported levels of both depression and loneliness *pre-pandemic.*

To improve the *robustness* of our results, we include a set of dummy variables accounting for whether, during lockdown, respondent experienced: change in residence, death of relative/ friend due to COVID, relative or friend infected, income loss or job loss. Descriptive statistics of the variable are presented in Table 4.1.

Table 4.1: Descriptive Statistics of Control Variables (%)

Variables	Categories	Total
Age	"50-59"	36.5
	"60-69"	35.8
	"70+"	27.7
Gender	Female	52.5
	Male	47.5
Educational level	High	25.4
	Medium	57.6
	Low	17
Employment status before lockdown	Employed	39.2
	Unemployed	8.14
	Not able to work	0.7
	Retired	0.39
	Home care	9.73
	Other	2.7
Marital Status	Single	8.52
	Married	65.90
	Domestic partnership	8.58

Continued on next page

Table 4.1 Continued from previous page

Variables	Categories	Total
	Not cohabiting partnership	2.73
	Separated/ divorce	8.77
	Widowed	5.47
Income before COVID-19 pandemic	Living comfortably	17.36
	Coping	47.07
	Finding it difficult	27.92
	Finding it very difficult	7.63
Self-rated health	Very good	10.55
	Good	45.48
	Acceptable	38.10
	Poor	5.85
Chronic condition	Yes	60.75
	No	39.25
Kin alive	Yes	97.39
	No	2.61
Cohabiting	Yes	88.23
	No	11.77
Percieved depression pre-Lockdown	More often than usual	43.76
	Same as usual	21.50
	Less often than usual	2.67
	Never been	32.06
Percieved loneliness pre-Lockdown	More often than usual	31.93
	Same as usual	26.97
	Less often than usual	4.07
	Never been	37.02
Neighborhood self-description	Big city	18.82
	Suburb	11.06
	Town	39.63
	Village	26.39
	Countryside	4.10
Experiences during pandemic	Suffered income loss	31.42
	Lost job	4.26
	Death of relative/friend due to COVID	8.01
	A relative or friend was infected	12.08

Note: N = 1,573.
Source: Intergen-covid online survey (Arpino *et al.*, 2020).

4.3.4 Is there Selection?

Due to the *missingess* of our data about the unobserved potential outcome we cannot study the relationship between outcome and covariates as we would do in a standard statistical approach. Instead we proposed a detailde examination of the *selection bias* due to the *observed* confounders.

Definition 4 (Selection Bias) *Selection bias due to confounders occurs when in-*

dividuals who are treated are substantially different from those who are untreated in terms of covariates that are associated with the outcome (Parast and Griffin, 2020).

In order to estimate the selection bias due to each of the observed confounder we compare an unadjusted (naive) treatment effect estimate with the propensity score adjusted estimate. The aim of the procedure is to identify the observed covariate that explains the largest portion of the estimated selection bias and include them in the propensity score model in order to remove the selection bias from the estimated treatment effect. Pragmatically speaking, we can see the selection bias as measure of the *confoundess* of our data, reducing or remove the selection bias from our data will result in a unconfounded causal effect estimate.

For this analysis we will use two different approaches, namely single confounder *removal* and single confounder *inclusion*. In the *removal* approach we will estimate the bias comparing the fully adjusted estimate of the treatment effect with the adjusted estimate if one were to remove the confounder of interest. Both estimates are obtained by adjusting the sample through the propensity score weighting procedure mentioned in Chapter 3. The fully adjusted estimate is given by

$$\hat{\Delta}_{ps} = \frac{\sum_{i:T_i=1} Y_i^{obs} W_i}{\sum_{i:T_i=1} W_i} - \frac{\sum_{i:T_i=0} Y_i^{obs} W_i}{\sum_{i:T_i=0} W_i}, \tag{4.1}$$

where Y_i^{obs} is the observed outcome of unit i, $T_i \in \{0,1\}$ is the treatment indicator and W_i are the propensity score weights described in Equation (3.2).

While the estimate that exclude the confounder X_j is defined as

$$\hat{\Delta}_{ps}(-X_j) = \frac{\sum_{i:T_i=1} Y_i^{obs} W_i(-X_j)}{\sum_{i:T_i=1} W_i(-X_j)} - \frac{\sum_{i:T_i=0} Y_i^{obs} W_i(-X_j)}{\sum_{i:T_i=0} W_i(-X_j)} \tag{4.2}$$

where $p(-X_j) = \Pr(T_i = 1 | X_i\{X_{ji}\}, X_i\{X_{ji}\})$ denote the propensity score estimated after removing the confounder X_j and weights for unit i are given by

$$W_i(-X_j) = \begin{cases} \frac{1}{p_i(-X_j)} & \text{if } T_i = 1 \\ \frac{1}{1-p_i(-X_j)} & \text{if } T_i = 0. \end{cases}$$

We define

$$\hat{\lambda}(-X_j) = \hat{\Delta}_{ps} - \hat{\Delta}_{ps}(-X_j) \tag{4.3}$$

such that $\hat{\lambda}(-X_j)$ estimate the *shift* due to the removal of confounder X_j from the fully adjusted treatment effect and estimate the proportion of *observed* bias explained by such confounder as

$$\hat{B}(-X_j) = \frac{\hat{\lambda}(-X_j)}{\beta^R} \tag{4.4}$$

with $\beta^R = \sum_{j=1}^{k} \left| \hat{\lambda}(-X_j) \right|$ being the the total bias capture by all the included covariates.

In order to estimate the adjusted treatment effect we use the non-parametric procedure explained by Chapter 3 implemented in the `SBDecomp` R package (Parast, 2022). One important detail is that the selection bias estimated are relative to the ATE (and not the ATT, our estimand of interest), their only function is to help us understand better, on a sample level, which are the potential confounders of our analysis that have to be included in the propensity score model in order to obtain an unbiased treatment effect estimate.

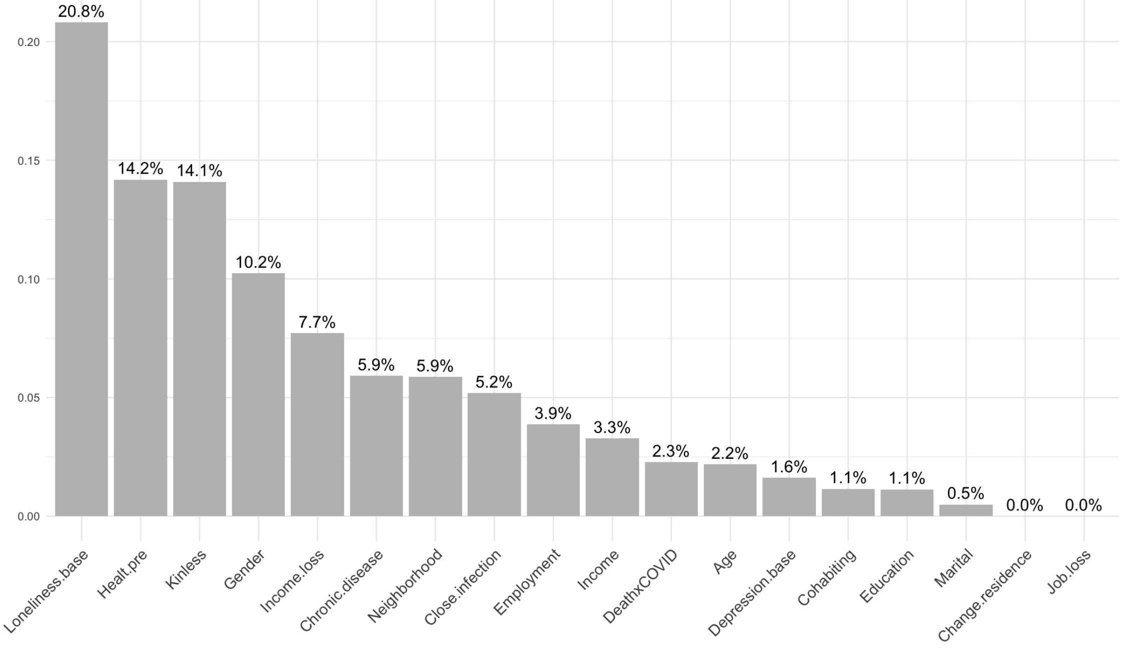


Figure 4.4: Proportion of Estimated Selection Bias explained by each confounder using the *Removal* approach (non optimized).

From Figure 4.4 we see how the loneliness pre-lockdown level, the self-reported level of health, the availability of kin and the gender of the individual explain for more than 50% of the entire selection bias ($\approx$ 59.3%) while the two dummy variables accounting for change in residence of the respondent and job loss during the lockdown did not contributed to the selection bias at all.

A slightly different picture emerge when the propensity score are estimated with an *optimized* version of the GBM algorithm (more on that later on) as we can see in Figure 4.5.

In this case the marital status of the respondents is the confounder that justify most of the bias, around 17%, followed by the self reported health condition, the baseline loneliness pre-pandemic, the availability of kin and the dummy variable accounting for whether the individual suffered an income loss during the lockdown. Together these covariate explain around the 50% of the selection bias while the change in residence have not contributed to the bias at all. Is important to notice that, even if the the estimate for such variable changed numerically speaking,

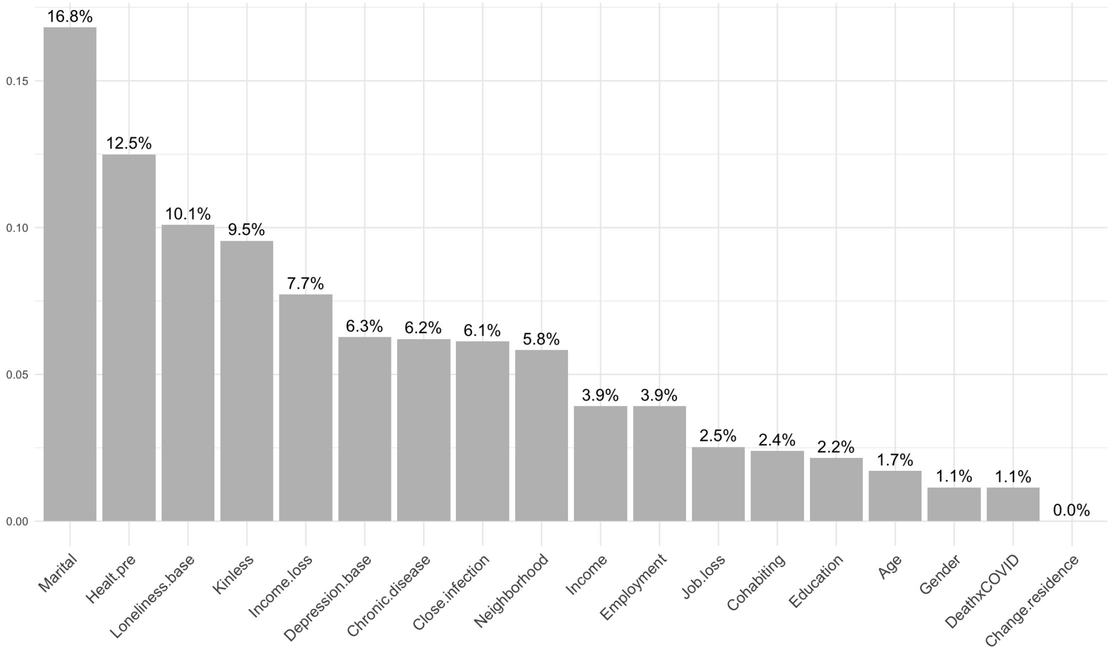


Figure 4.5: Proportion of Estimated Selection Bias explained by each confounder using the *Removal* approach (optimized).

the most important variables in defining the shift between the treatment and control group (namely `loneliness.base, health.pre, kinless, income.loss`) are present in both version of the estimates.

The reason why we see such difference in the two estimated biases has to do with the assumption made by the removal approach. In this method we are comparing the fully adjusted estimate against the estimated without the jth covariate, therefore, assuming that the full adjusted estimate is unbiased and taking it as reference.

In order to overcome such limitation we will now introduce the single confounder *inclusion* approach. Such approach aims to provide a better understanding of the value of each single convariate included in the propensity score model in reducing the selection bias beyond all the other observed confounders. In this case we estimate the amount of bias *removed* from the *naive* treatment effect estimate

$$\hat{\Delta}^{\text{naive}} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i^{\text{obs}} - \frac{1}{n_0} \sum_{i:T_i=0} Y_i^{\text{obs}}, \tag{4.5}$$

if one were to use a propensity score weighted estimate of the effect adjusting for a single pre-treatment covariate.

The adjusted effect is given by

$$\hat{\Delta}^{ps} = \frac{\sum_{i:T_i=1} Y_i^{\text{obs}} W_i}{\sum_{i:T_i=1} W_i} - \frac{\sum_{i:T_i=0} Y_i^{\text{obs}} W_i}{\sum_{i:T_i=0} W_i}, \tag{4.6}$$

where the weights are those of Equation (3.2) using the propensity score $p_i(X_j) =$

$\Pr(T_i = 1|X_{ji})$.

We define the shift in the naive estimate captured by the observed confounder X_j as

$$\hat{\lambda} = \hat{\Delta}^{\text{naive}} - \hat{\Delta}^{ps}. \tag{4.7}$$

and estimate the proportion of bias explained by X_j with

$$\hat{B}(X_j) = \frac{\hat{\lambda}(X_j)}{\beta^I} \tag{4.8}$$

where $\beta^I = \sum_{j=1}^{k} \left|\hat{\lambda}(X_j)\right|$ is the total bias due to the inclusion of the single covariate.

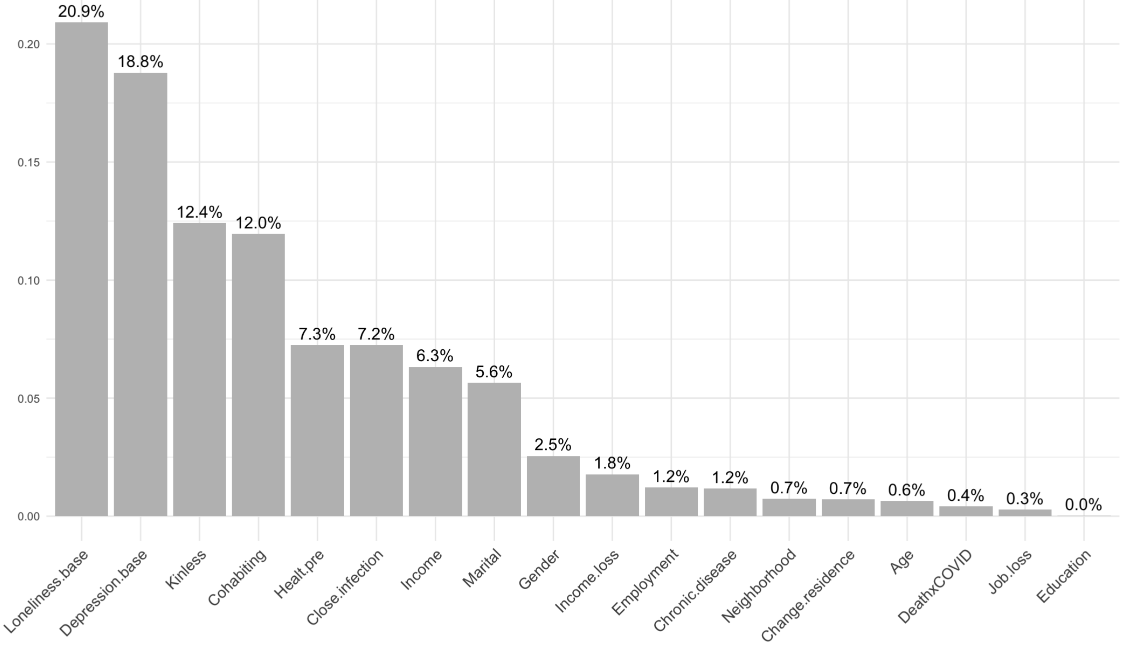


Figure 4.6: Proportion of Estimated Selection Bias explained by each confounder using the *Inclusion* approach (same for both optimized and non-optimized GBM).

Once again, we find that, considered singularly, the variable accounting for the biggest proportion of the selection bias are, as expected, the baseline loneliness pre-lockdown of the respondent, the baseline depression, the availability of kin and the self-assessed health condition pre-pandemic. These variables were proven already as significantly correlated with the outcome of interest (Arpino *et al.*, 2022), validating our hypothesis regarding their fundamental inclusion in the propensity score model. On the other side, variables such as the age, the education level and the dummy variables accounting for change in residence and job loss never accounted for more than $\approx 7\%$ of the total observed bias.

Curiously the variable accounting for the cohabiting situation of the respondent explain the 12% of the selection when considered alone (Figure 4.6) and less than 3% when considered in the both the removal approaches tested, indicating that such information might be strongly correlated with some other variable that express the majority of the information about the co-living situation.

We will now briefly explore the relationship between the treatment and the covariates that account for the largest proportion of the selection bias in order to asses how these characteristics impact the change in non-physical contact (for both dichotomous and multinomial treatments).

Treatment vs Loneliness Pre-Lockdown

As expected we strongly refuse the hypothesis of independence between the treatment and the loneliness baseline (Table 4.2)

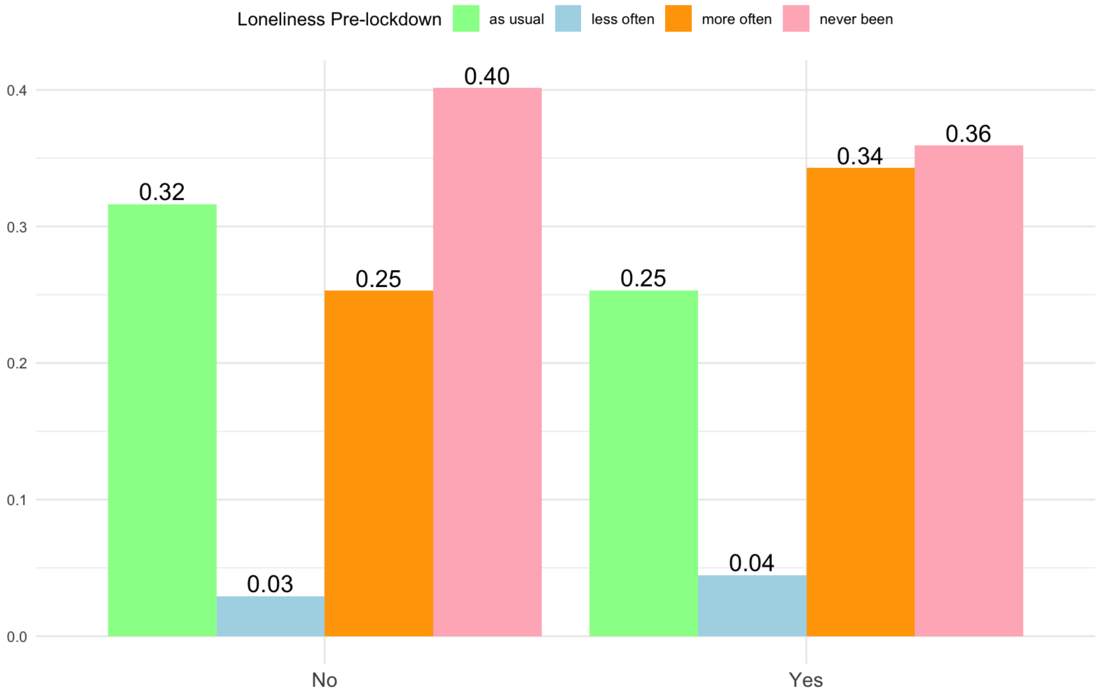


Figure 4.7: Proportion of pre-lockdown levels of loneliness with respect to the treatment (binary)

Test	Statistic	Degrees of freedom	p-value
Pearson's χ^2 test (binary)	15.415	3	0.001494
Monte Carlo test (binary)	15.415	-	0.0014
Pearson's χ^2 test (multinomial)	38.361	6	9.549×10^{-7}
Monte Carlo test (multinomial)	38.361	-	2×10^{-4}

Table 4.2: Independence test: Binary Treatment vs Loneliness Pre-lockdown
Number of Monte Carlo replicates: 5000.

A deeper exploration of the relationship between the treatment and the perceived loneliness pre-lockdown underline that, regardless the treatment setup, people who increased their non-physical contact are those who felt more lonely from the beginning of the enforced "stay-at-home" while the "non-increasing" group felt lonely "as usual" or have never felt lonely.

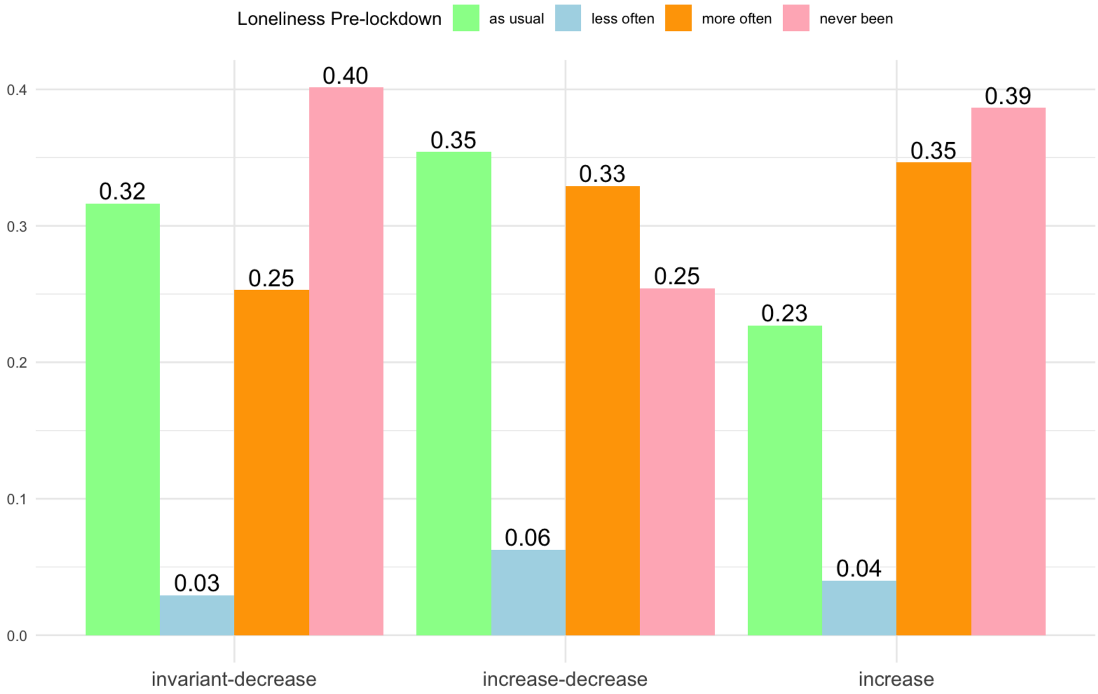


Figure 4.8: Proportion of pre-lockdown levels of loneliness with respect to the treatment (multinomial)

Treatment vs Depression Pre-lockdown

Further confirmation of the importance of the baseline depression levels when addressing the selection bias can be found in the comparison between treatment indicator. Such comparison show clearly how almost half of the people who increased their non-physicial contact had a higher baseline depression compared with the "invariant-decrease" group and such relationship does not differ in the multinomial setup. Furthermore, we see how the "non-increased" group is composed by almost 40% of people who have never suffer of depression.

Test	Statistic	Degrees of freedom	p-value
Pearson's χ^2 test (binary)	21.88	3	6.91×10^{-5}
Monte Carlo test (binary)	21.88	-	0.0003999
Pearson's χ^2 test (multinomial)	30.386	6	3.319×10^{-5}
Monte Carlo test (multinomial)	30.386	-	0.0002

Table 4.3: Independence test: Treatment vs Depression Pre-lockdown
Number of Monte Carlo replicates: 5000.

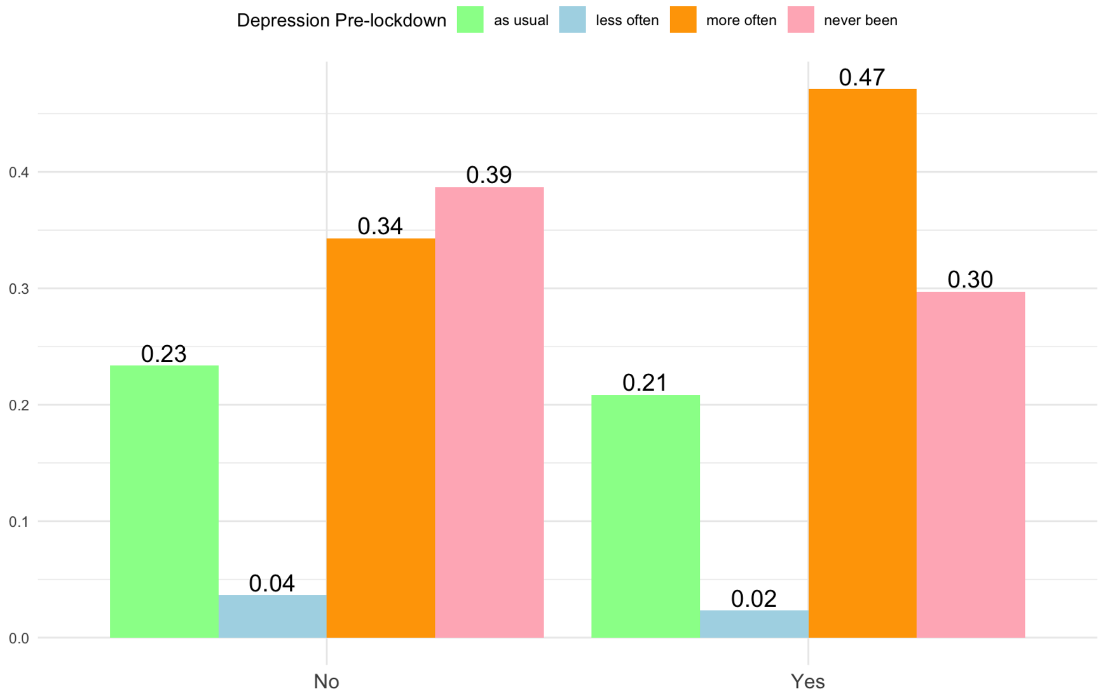


Figure 4.9: Proportion of pre-lockdown levels of depression with respect to the treatment (binary)

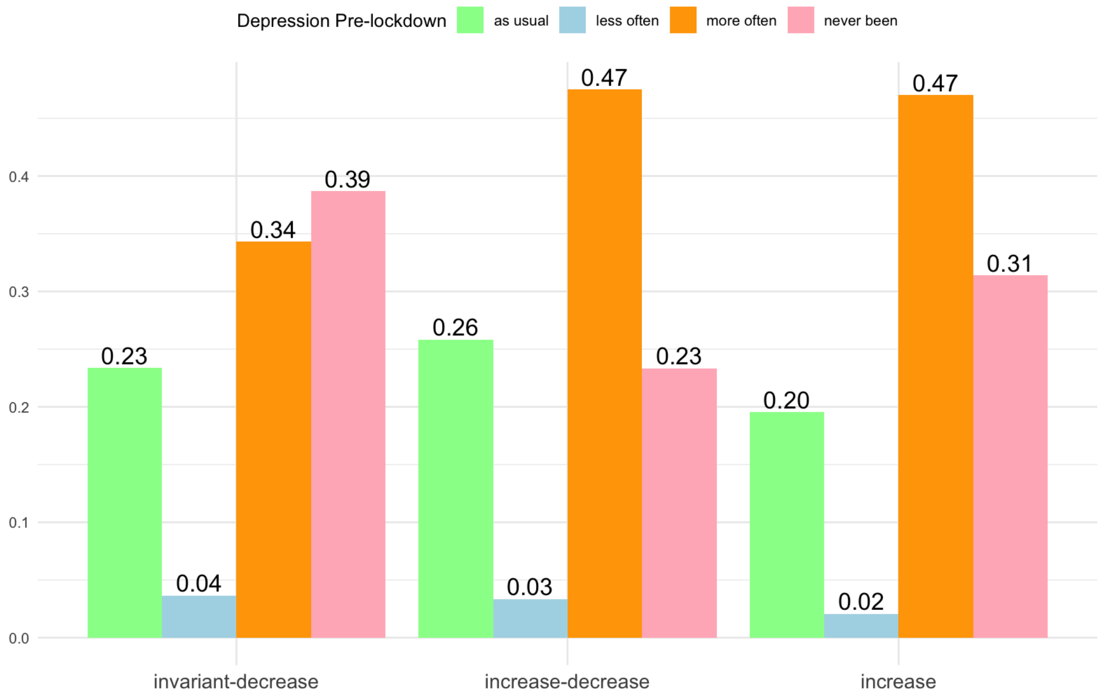


Figure 4.10: Proportion of pre-lockdown levels of depression with respect to the treatment (multinomial)

Treatment vs Availability of Kin

The analysis of the availability of kin with the respect to the treatment show a clear trend in which people with no kin were less prone to not increase their non-physical contacts. This might suggest that having kin available is associated with greater flexibility or variability in nonphysical contact changes, whereas those without kin show less variation and are less likely to increase their nonphysical contact.

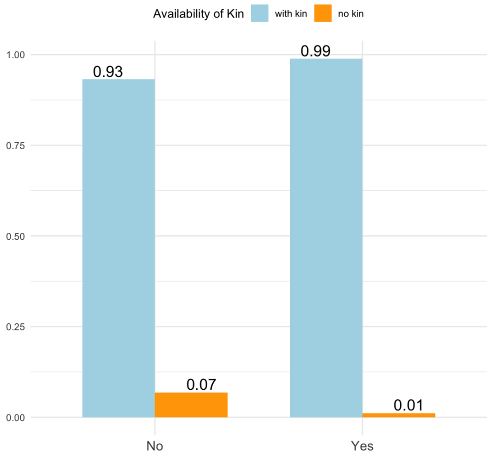


Figure 4.11: Proportion of kinless people with respect to the treatment (binary)

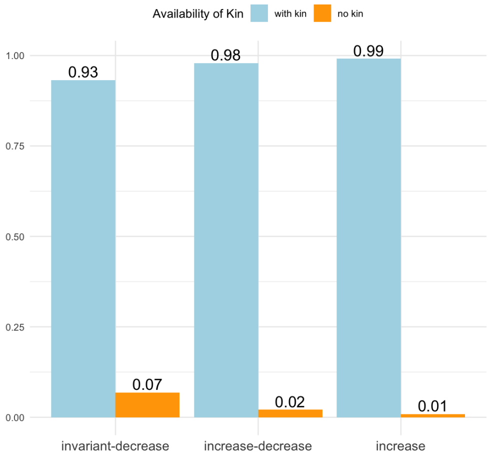


Figure 4.12: Proportion of kinless people with respect to the treatment(multinomial)

Test	Statistic	Degrees of freedom	p-value
Pearson's χ^2 test (binary)	36.52	1	1.511×10^{-9}
Monte Carlo test (binary)	38.729	-	0.0002
Pearson's χ^2 test (multinomial)	39.835	2	2.238×10^{-9}
Monte Carlo test (multinomial)	39.835	-	0.0002

Table 4.4: Independence test: Treatment vs Availability of Kin
Number of Monte Carlo replicates: 5000.

Treatment vs Health Condition Pre-pandemic

Curiously, the self-reported health condition, even if accounting for 7.7% - 14.2% (Figures 4.4, 4.6) of the selection bias, result independent from the treatment assignment mechanism (Table 4.5).

Test	Statistic	Degrees of freedom	p-value
Pearson's χ^2 test (binary)	7.3383	3	0.0618
Monte Carlo test (binary)	7.3383	-	0.0549
Pearson's χ^2 test (multinomial)	11.16	6	0.0835
Monte Carlo test (multinomial)	11.16	-	0.0803

Table 4.5: Independence test: Treatment vs Self-Reported Health Pre-Lockdown
Number of Monte Carlo replicates: 5000.

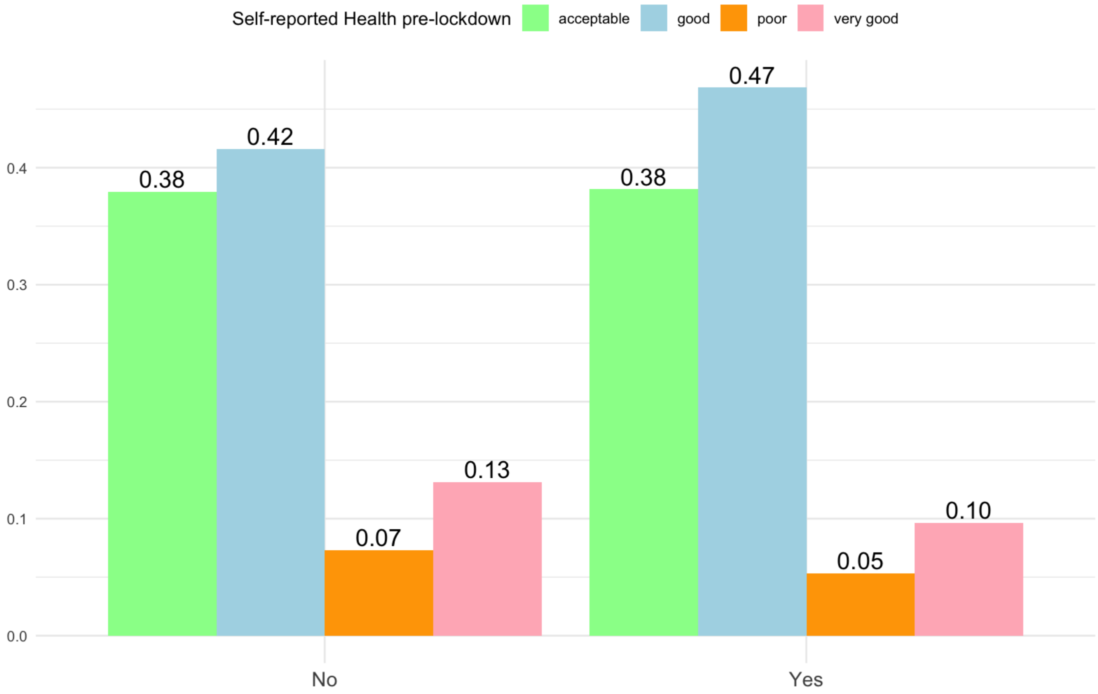


Figure 4.13: Proportion of self-reported health condition pre-lockdown with respect to the treatment (binary)

Figure 4.13 and Figure 4.14 enforce the idea that the health status does not cause treatment assignment. However, it still affects the outcome, thereby inducing selection bias in the estimation of the treatment effect. The independence from the treatment indicator make the variable a perfect instrument to balance the treatments group and would be considered in the propensity score model.

In any case, our analysis is to considered limited to our sample and relatively to the treatment and outcome of choice. Therefore, given the uncertainty about the real *model* and relationship that might affect the perceived loneliness we will continue to consider all the background information as potential confounders and include them in the final model for the propensity scores.

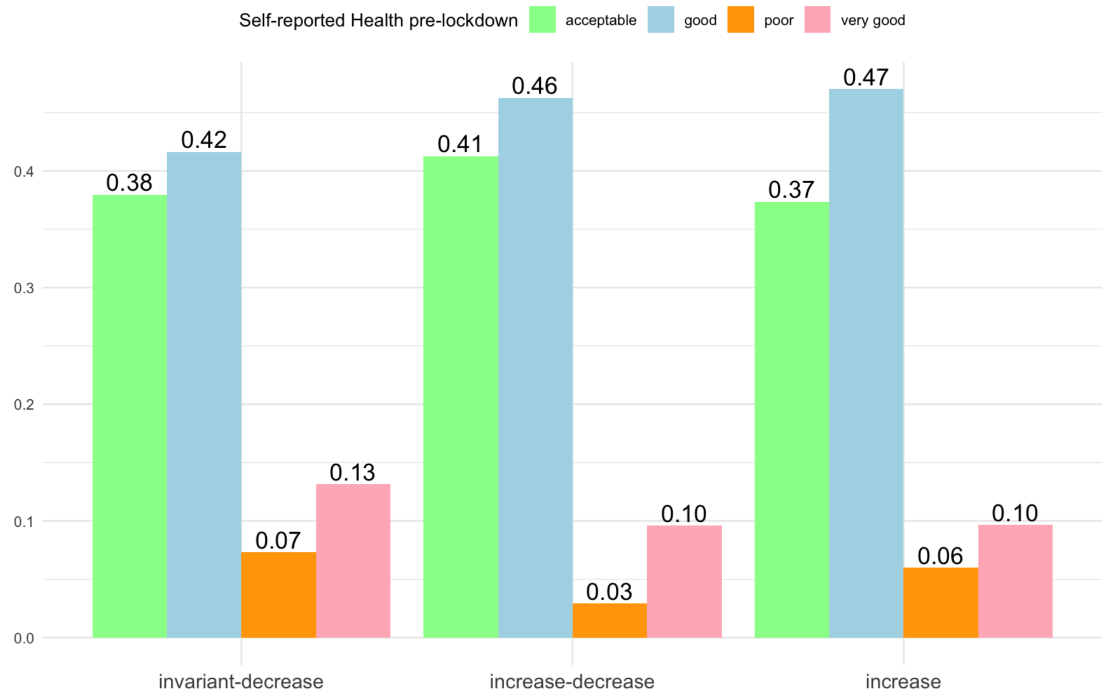


Figure 4.14: Proportion of self-reported health condition pre-lockdown with respect to the treatment (multinomial)

4.4 Balancing

After a detailed analysis of the sample we will now proceed explaining the necessary steps we made in order to apply propensity score weighting to our data and the results of such procedure.

As said before, due to its great flexibility in handling a large number of covariates and its capability in capture any nonlinearities between the treatment assignment mechanism and the potential counfounders, we will estimate the propensity scores using GBM in the iterative fashion described in Chapter 3 (McCaffrey *et al.*, 2013).

The necessary step in order to implement propensity score weighting are:

1. Estimating the propensity scores for each treatment.

2. Compute the propensity score weights based on the causal estimand of interest (in our case we will use ATT: Equation (3.4)).

3. Assess balance among groups using the desired balance metrics (Equation (3.13); Equation (3.14)),

4. Assess overlap among treatment groups (this step might be considered as a *positivity* check)

5. *Analysis*-stage and causal effect estimation.

The entire *workflow* for propensity scores weighting is implemented in the `twang` R package (Cefalu *et al.*, 2022).

4.4.1 Binary Setup

We first start by defining our propensity score model including all the variables mentioned earlier as covariates and the *binary* treatment assignment indicator T_i as response variable

$$T_i = \begin{cases} 1 & \text{if unit } i \text{ \textit{increased} the non-physical contacts} \\ 0 & \text{otherwise} \end{cases}$$

Due to the implementation of the GBM, Table 4.6 represents categorical variables assigning a dummy to each level of the variable. The balancing statistics, namely the absolute standardized difference and the Kolmogorov-Smirnov statistic, are then computed for each dummy. Such procedure aim to minimize the difference (or maximize the balance) between each level of the single factor.

From the balance table (Table 4.6), we notice that the majority of the variable result balanced between treatment and comparison group with the exception of: employment status, income status, marital status, cohabiting situation (or `coliving`), kinless, self-reported health status, depression baseline, loneliness baseline, self-reported neighborhood description, the level of education, the dummy variable regarding the infection of a close connection experience and the one regarding income loss (marked by red). In particular, the most unbalanced variables are the loneliness baseline and the availability of kin.

Table 4.6: Unweighted Balance Table

Variable	**Treatment**		**Control**		*ASD*	*KS*
	Mean	SE	Mean	SE		
female	0.487	0.500	0.443	0.497	0.088	0.044
agecat: 50-59	0.368	0.482	0.355	0.479	0.026	0.013
agecat: 60-69	0.352	0.478	0.375	0.484	-0.047	0.022
agecat: 70+	0.280	0.449	0.270	0.444	0.022	0.010
edu: low	0.164	0.370	0.187	0.390	-0.064	0.024
edu: medium	0.590	0.492	0.535	0.499	0.111	0.055
edu: high	0.246	0.431	0.277	0.448	-0.072	0.031
empstatus: Employed	0.393	0.488	0.389	0.488	0.007	0.003
empstatus: Unemployed	0.077	0.266	0.095	0.293	-0.069	0.018
empstatus: Not Able to work	0.005	0.072	0.012	0.110	-0.098	0.007
empstatus: Retired	0.400	0.490	0.382	0.486	0.036	0.018

Continued on next page

Table 4.6 – continued from previous page

Variable	Treatment		Control		*ASD*	*KS*
	Mean	SE	Mean	SE		
empstatus: Home care	0.099	0.299	0.092	0.290	0.022	0.007
empstatus: Other	0.027	0.161	0.029	0.168	0.015	0.002
income: comfortable	0.171	0.377	0.180	0.384	0.023	0.009
income: sufficient	0.492	0.500	0.411	0.492	0.161	0.081
income: struggling	0.267	0.442	0.314	0.464	0.106	0.047
income: desperate	0.070	0.255	0.095	0.293	0.099	0.025
marital: Single	0.069	0.253	0.131	0.338	0.247	0.062
marital: Married	0.670	0.470	0.628	0.483	0.090	0.042
marital: Domestic partnership	0.087	0.282	0.083	0.275	0.015	0.004
marital: Not Cohabiting	0.030	0.171	0.019	0.138	0.062	0.011
marital: Separated/ divorced	0.084	0.278	0.097	0.296	0.046	0.013
marital: Widowed	0.059	0.236	0.041	0.199	0.076	0.018
coliving	0.898	0.302	0.837	0.370	0.203	0.061
changeres	0.010	0.101	0.012	0.110	0.018	0.002
kinless	0.011	0.105	0.068	0.252	0.541	0.057
healthpre: very good	0.096	0.295	0.131	0.338	0.118	0.035
healthpre: good	0.469	0.499	0.416	0.493	0.105	0.053
healthpre: acceptable	0.382	0.486	0.380	0.485	0.004	0.002
healthpre: poor	0.053	0.225	0.073	0.260	0.087	0.020
chronic	0.408	0.492	0.348	0.477	0.123	0.060
DDC	0.087	0.282	0.061	0.239	0.093	0.026
RFI	0.137	0.344	0.075	0.264	0.179	0.062
incomeloss	0.331	0.471	0.268	0.443	0.134	0.063
jobloss	0.041	0.199	0.046	0.210	0.025	0.005
depr.base: more often	0.471	0.499	0.343	0.475	0.257	0.128
depr.base: as usual	0.208	0.406	0.234	0.423	0.062	0.025
depr.base: less often	0.023	0.151	0.036	0.188	0.088	0.013
depr.base: never been	0.297	0.457	0.387	0.487	0.196	0.090
lone.base: more often	0.343	0.475	0.253	0.435	0.921	0.090
lone.base: as usual	0.253	0.435	0.316	0.465	0.145	0.063
lone.base: less often	0.045	0.207	0.029	0.168	0.075	0.016
lone.base: never been	0.359	0.480	0.401	0.490	0.088	0.042
neighbor: big city	0.185	0.388	0.197	0.398	0.031	0.012
neighbor: suburb	0.107	0.309	0.122	0.327	0.048	0.015
neighbor: town	0.402	0.490	0.380	0.485	0.046	0.023
neighbor: village	0.273	0.446	0.238	0.426	0.078	0.035
neighbor: countryside	0.033	0.178	0.063	0.243	0.172	0.031

The first step of the procedure is to fit the GBM model and assessing the convergence of the statistic. In a first moment we used the default parameters of the algorithm. To assess the balance procedure we used the average over all the covariates of both the absolute standardized difference and the Kolmogorov-Smirnov statistics (referred as ASD, KS from now on).

Parameter	**Value**
n.trees	10000
interaction.depth	3
shrinkage	0.01
bag.fraction	1
n.minobsinnode	10

Table 4.7: Default Parameters

Assessing Convergence

As explained in Chapter 3 the objective is to let GBM *overfit* the data and than select the optimal iteration to be the one that minimize the balancing criterion.

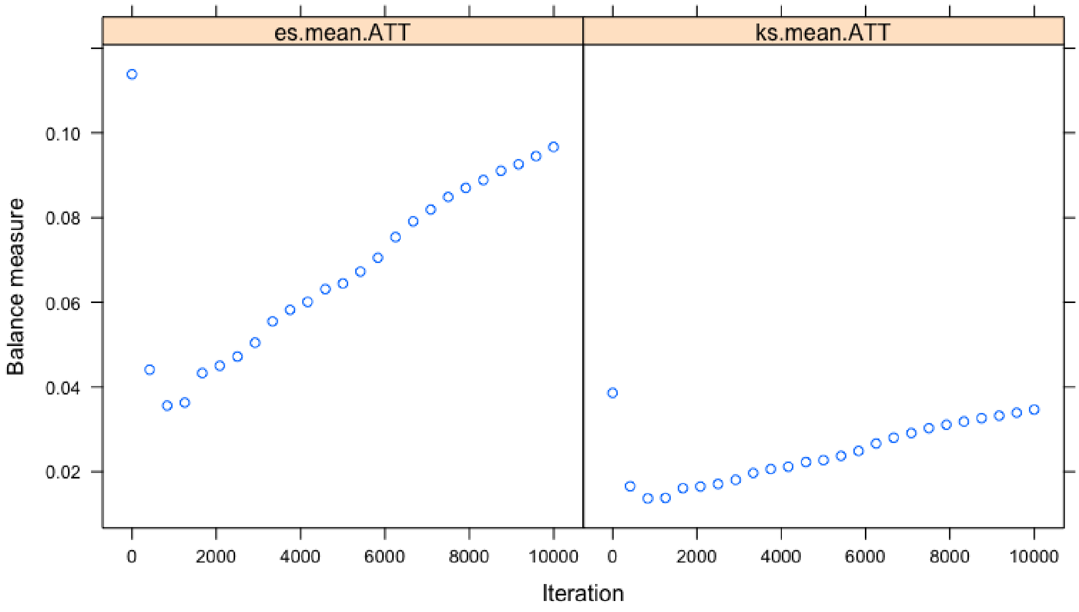


Figure 4.15: Convergence plot of the GBM algorithm

The convergence plot in Figure 4.15 show the balance measures as a function of the number of iterations of the GBM algorithm, where higher iterations correspond to more complicated models. For both the statistics considered we can notice how the optimal iteration is achieved between the first thousands iterations. This is partially due to the fact that our two groups were already similar. On the other side, we notice how as the number of iterations and the complexity of the model increase, manly due to the inclusion of the interaction terms, the balancing does not improve, instead it get worse.

	n.treat	n.ctrl	ess.treat	ess.ctrl	mean.es	mean.ks	iter
unw	1161	411	1161.00	411.00	0.114860	0.312611	-
es.mean	1161	411	1161	258.6691	0.034840	0.01336	1053
ks.mean	1161	411	1161	259.2129	0.034849	0.01335	1037

Table 4.8: Weighting summary statistics for binary treatment (no optimization)

From a numerical perspective Table 4.8 shows us different aspects of the weighting procedure. At first glance, we see the comparison between the actual number of units, in the treatment and comparison group, before and after the weighting procedure capture by the effective sample size (`ess`). It is important to notice that since we are interested in the ATT the treatment units are not weighted and the value of units remain the same after the weighting procedure. For the control group, instead, we see a reduction due to the increase variability added by the weighting procedure.

Even if the difference between the two statistics is quantitatively very small, the bigger effective sample size and the smoother converge curve Figure 4.15 brought us to prefer the KS statistic as the main balancing criterion for the rest of the analysis. Furthermore, such statistics consider the differences along the entire empirical cumulative distribution of the covariates and not the punctual difference in mean (as for ASD) favoring a balance among the entire distribution of the variable at hand.

Assessing Overlap

In order to control if the positivity assumption is confirmed by the data we evaluate the overlap between the distributions of the estimated propensity scores for both groups. Contrarily to other balancing method, which require almost perfect overlap (at the cost of discarding units), covariate balance can be achieved with weighting even when propensity scores distributions show little overlap. In such cases tho, to balance units that are very *far apart* in term of propensity scores, the weights must increase in magnitude, potentially bringing numerical instability and degrading the overall balance.

Evaluating Figure 4.16 we see few areas of non-overlap in the propensity scores distributions, namely for value smaller than 0.3 or bigger then 0.9. While the distribution of the weights shown by Figure 4.17 does not underline any particularly big weight.

Assessing Balance

Figure 4.18 illustrates the results of the balancing procedure comparing the magnitude of differences between groups on each pretreatment covariate before and after

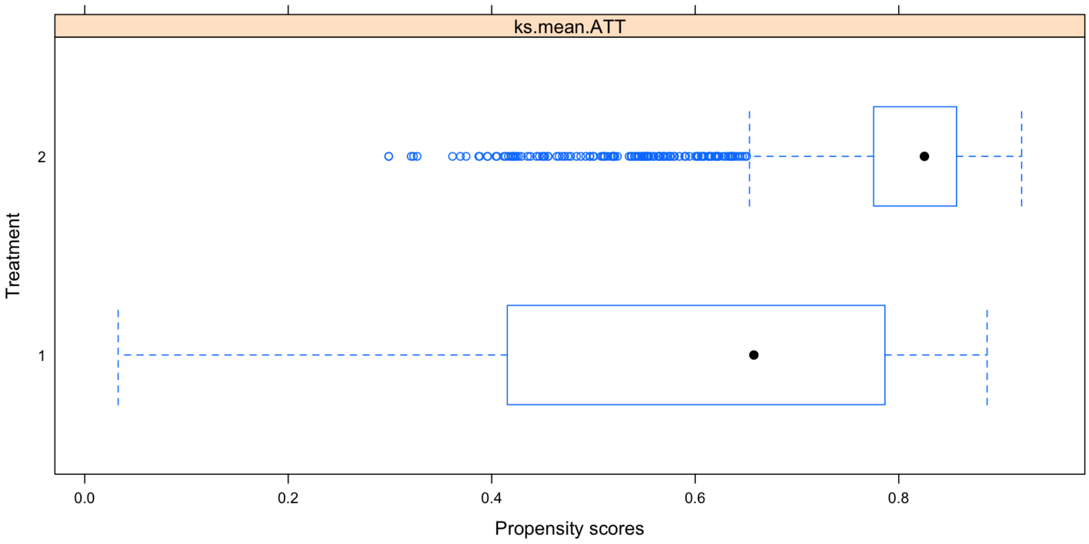


Figure 4.16: Boxplot of the estimated propensity scores for treatment (2) and control (1) group

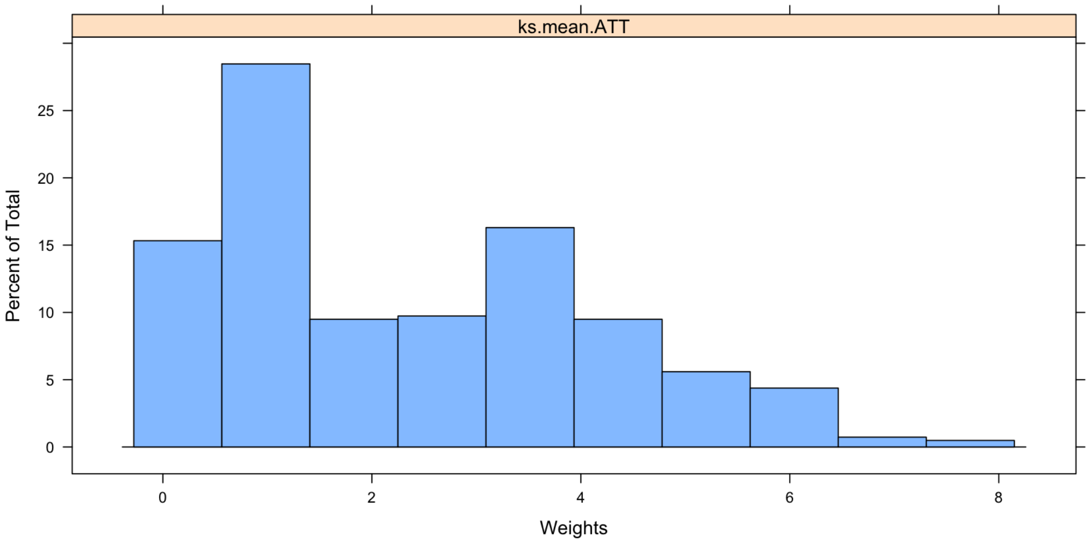


Figure 4.17: Histogram of propensity score weights

weighting. The plot shows a substantial reduction in ASD for most of the covariates (blue lines), with few variable showing a trivial increase (red lines), such increase is mostly due to the covariate `lone.base` and `kinless` that had very different values between groups. In order to balance these two variables the algorithm slightly increase the imbalance for other covariates that were closer before the weighting. After the weighting we only have two covariates that register an absolut standard difference greater than 0.1, namely `chronic`, dummy variable accounting for the presence of chronic disease and `agecat` accounting for the age of the respondent, with 0.137 and 0.104 respectively. Only the variable accounting for chronic disease results statistically significant after weighting (closed red circle).

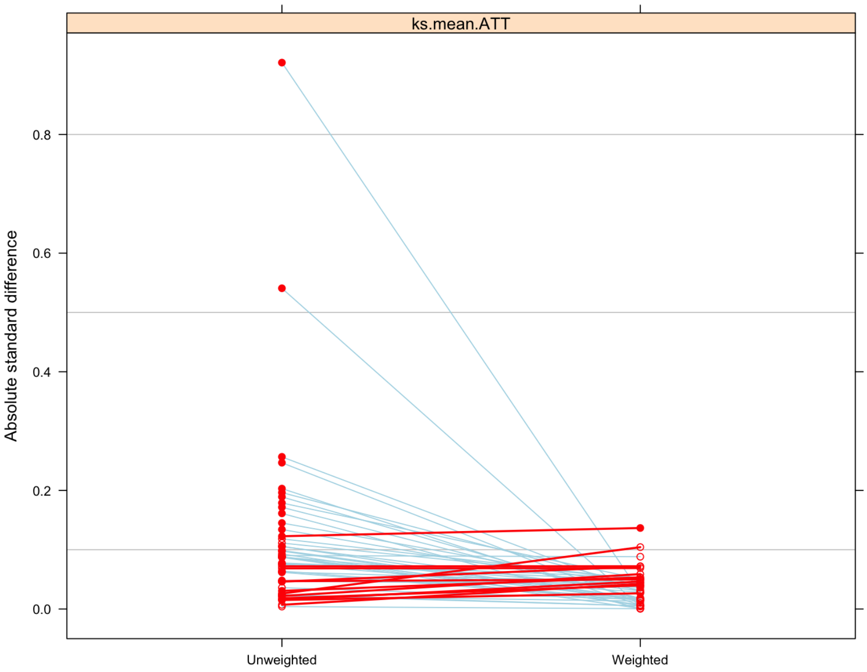


Figure 4.18: Absolute standard differences before and after weighting (non optimized GBM)

Hyper-parameters Optimizations

After this first result we tested whether the optimization of the hyper-parameters of the GBM algorithm would yield a better balance between groups. As we anticipated we found that, following Griffin *et al.* (2017), pursuing covariate balance when optimizing the algorithm, indeed, yielded better results than the default setup.

We, therefore, defined the `opt.balance.gbm` function that, taking as input the data and a grid of possible parameters for the algorithm (explained in Chapter 3), train a GBM algorithm for each possible combination of the grid. For each instance of the GBM the function proceeds to the numerical optimization of the KS statistic over the predicted values of each iteration of the algorithm and find the one which predictions minimize the balance criterion.

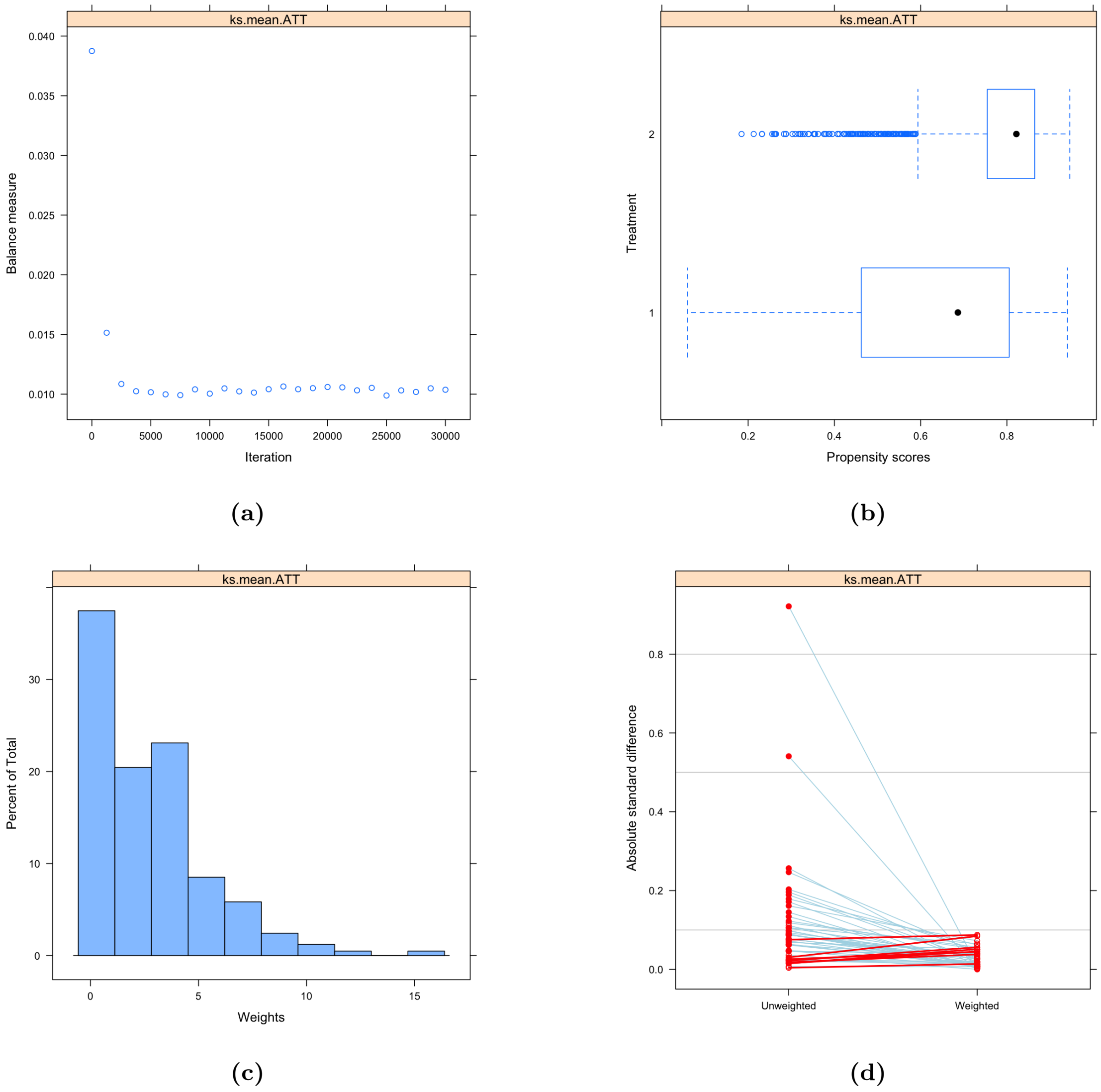
Parameter	Value
n.trees	30000
interaction.depth	1
shrinkage	0.005
bag.fraction	0.5
n.minobsinnode	15

Table 4.9: Optimized GBM Parameters

Afterwards, the set of parameters that yield the best value of the statistic (at the optimal iteration) is selected as the best combination and returned to the user. Such function provide an *in-sample* optimization of the algorithm and should be re-applied any time the sample is modified.

In our case, we found that the suspicious about the worsening effect of the increasing complexity of the model were correct and that simpler model (basically without any interaction) would allow for better balance between groups. Also, our findings suggested that increasing the number of iteration and decreasing the shrinkage parameter resulted in better performance overall.

Using the parameters presented in Table 4.9 we were able to address, at the cost of a higher number of iterations (Table 4.10), the problems of the first balancing procedure, namely increasing the overlapped areas of the estimated propensity scores



(a)

(b)

(c)

(d)

Figure 4.19: (a) Converge plot, (b) Overlap of estimated propensity scores, (c) Distribution of propensity score weights, (d) Absolute standard difference reductions, for optimized GBM

distributions and removing any significant difference between the groups.

	n.treat	n.ctrl	ess.treat	ess.ctrl	mean.es	mean.ks	iter
unw	1161	411	1161.00	411.00	0.114860	0.312611	-
ks.mean	1161	411	1161	259.2129	0.034849	0.01335	1037
ks.mean.opt	1161	411	1161	226.103	0.028653	0.00986	25003

Table 4.10: Numerical comparison with optimized GBM

4.4.2 Multinomial Setup

For a multiple (or multinomial) treatment setup McCaffrey *et al.* (2013) suggest to repeat the same procedure of the binary case using only the subsample defined by the treatment level of interest and only one other level, taken as comparison. We, therefore, repeated the procedure twice, adjusting first the sample units identified by the level "invariant-decrease" and then those who underwent the "increase-decrease" treatment, in order to match the pretreatment characteristics of those individual who "increased" their non-physical contacts.

Due to the differences between the two sub-groups we will now report the table with the maximum values for the balancing descriptive statistics collapsed by covariate. For the sake of brevity, we focus on the variables that shows unweighted ASD greater than 0.1 in one of the two pairwise comparisons.

Table 4.11: Summary of balance statistics for covariate

Variable	Max ASD	Max KS
female	0.26	0.13
agecat:50-59	0.16	0.08
edu:low	0.10	0.04
edu:high	0.18	0.08
empstatus:Employed	0.16	0.08
empstatus:Retired	0.18	0.09
income:comfortable	0.10	0.04
income:sufficient	0.18	0.09
income:struggling	0.10	0.04
income:desperate	0.13	0.03
marital:Single	0.26	0.07
marital:Married	0.15	0.07
marital:Domestic partnership	0.11	0.03
marital:Not Cohabiting with partner	0.16	0.03
marital:Separated or divorced	0.11	0.03
marital:Widowed	0.17	0.04

Continued on next page

Table 4.11 – continued from previous page

Variable	Max ASD	Max KS
coliving	0.16	0.05
kinless	0.64	0.06
healthpre:very good	0.12	0.03
healthpre:good	0.11	0.05
healthpre:poor	0.13	0.03
chronic	0.16	0.08
DDC	0.11	0.03
changeres	0.15	0.01
RFI	0.21	0.08
incomeloss	0.12	0.06
lone.base:more often	0.20	0.09
lone.base:as usual	0.30	0.13
lone.base:less often	0.11	0.02
lone.base:never been	0.27	0.13
depr.base:more often	0.25	0.13
depr.base:as usual	0.16	0.06
depr.base:less often	0.11	0.02
depr.base:never been	0.17	0.08
neighbor:suburb	0.11	0.03
neighbor:countryside	0.14	0.03

Given the computational complexity of the optimization procedure described above, we decided to start with the optimized parameters selected, using the entire sample, for the binary setup and to adjust them with empirical trials. We selected the parameters presented in Table 4.12 as the set of hyper-parameters that showed the best result in balancing *both* the sub-sample.

Parameter	Value
n.trees	30000
interaction.depth	1
shrinkage	0.005
bag.fraction	0.75
n.minobsinnode	10

Table 4.12: Optimized parameter (multinomial GBM)

The only difference with the parameters used in the binary setup is the value of the `bag.fraction` parameter that set the proportion of sample randomly selected by the algorithm at each new iteration (Friedman, 2002). In this case, due to the cardinality of the sub-samples (particularly the one relative to the comparison treatment "increase-decrease", see Figure 4.3), we need more units in order to achieve good precision of the estimates.

The low number of units available for the comparison between treatment "increase" and comparison "increase-decrease" is notable from the convergence plot of the GBM algorithm that show a clear degradation of the balancing performance after few thousands iterations.

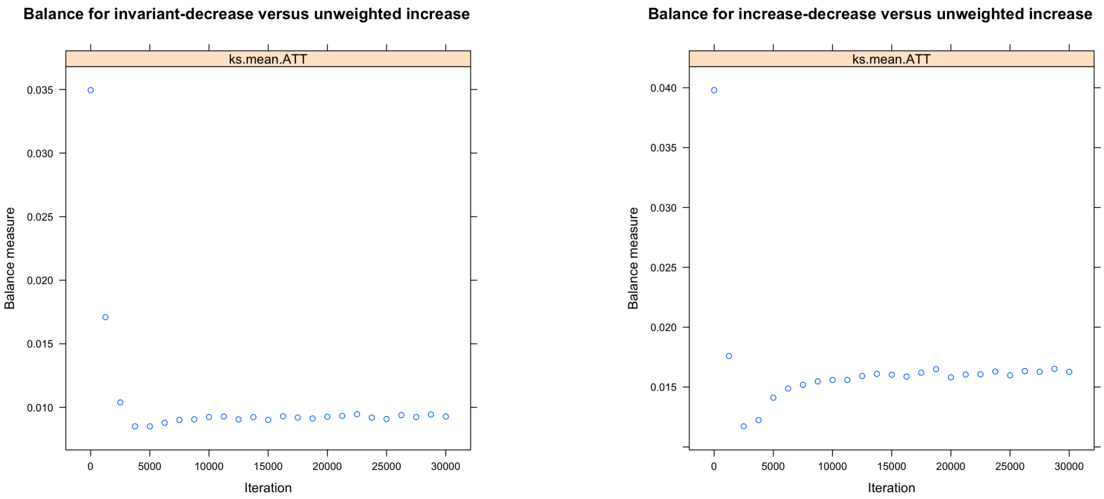


Figure 4.20: Convergence plot of the GBM algorithm for pairwise ATT.

Nevertheless, overlap is achieved in both the pairwise comparisons.

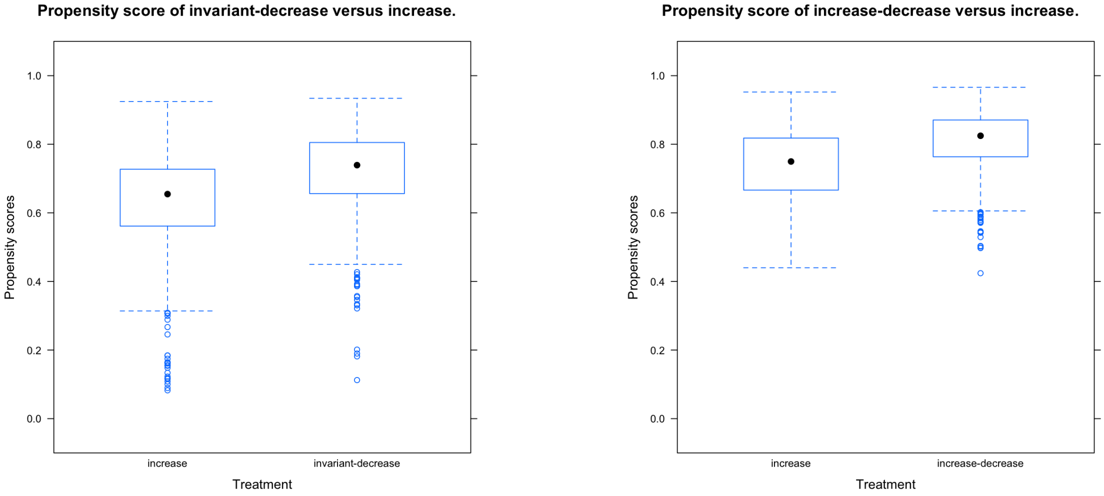


Figure 4.21: Boxplot of the estimated propensity scores for pairwise ATT.

Balancing is achieved for both pairwise comparisons, with the only problematic variable being the self-reported neighborhood description showing an absolute standard difference of 0.1 in the "increase" versus "increase-decrease" weighted subsample.

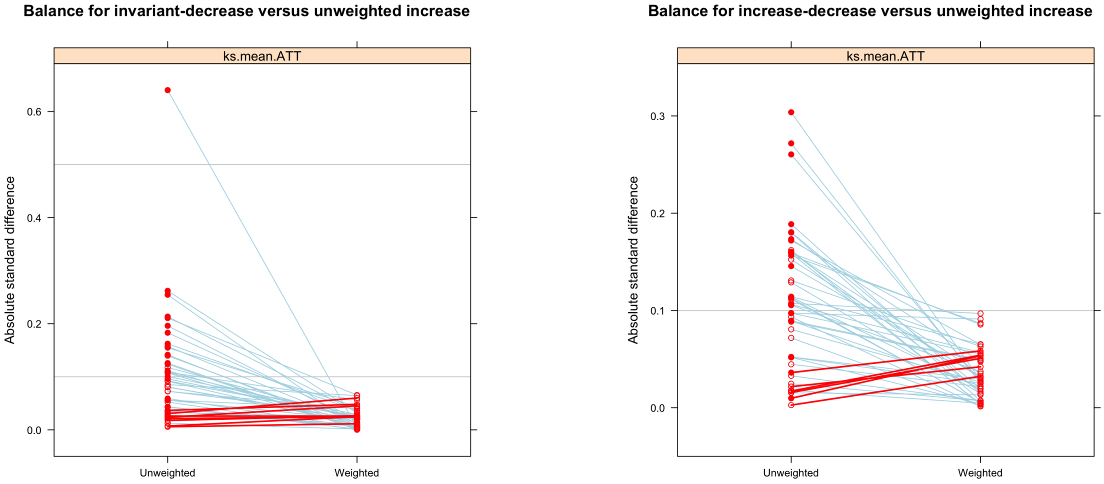


Figure 4.22: Absolute standard differences before and after weighting for pairwise ATT.

4.5 Outcome Modeling

The last step of the procedure is the estimation of the actual effect that different contact patterns had on the perceived severe loneliness. As explained in Chapter 3, we implemented a doubly robust estimator using G-computation with missing potential outcomes estimated through a weighted logistic regression.

For the G-computation step we considered four different Q-models, each based on a different set of covariates:

- **Model 1**: only the treatment indicator.
- **Model 2**: treatment indicator and baseline level of perceived loneliness pre-lockdown.
- **Model 3**: treatment indicator, baseline level of loneliness and the pre-lockdown level of depression (considered as a strong predictor of the outcome).
- **Model 4**: considering all the covariates included in the propensity score model.

Each model was fitted using the `survey` R package (Lumley, 2023) and the respectively ATT was computed with the `marginaleffects` library (Arel-Bundock, 2023). We will now report the results of such fitting procedure of each model. We will indicate only the estimate for the treatment indicators since all the other coefficients might be strongly biased, due to the weighting procedure, and they do not represent any causal relationship with the response variable.

4.5.1 Binary Setup

We quantified the effect of the increase in non-physical contacts as the *risk-difference* in terms of the probability of experiencing *severe* loneliness.

Specifically, we compare the probability of experiencing severe loneliness among those who increased their non-physical contacts (denoted by 1) against the *conterfactual* probability, of these same individuals, in experiencing loneliness had they not increased their remote contacts (denoted by 0).

Table 4.13: Treatment coefficients and ATT estimates for each Q-model (binary treatment)

Treatment	Model 1		Model 2		Model 3		Model 4	
	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)
Increase	-0.7557**	(0.2848)	-0.8214**	(0.3001)	-0.8267**	(0.3014)	-0.9611**	(0.3329)
ATT: 1-0	-0.0395*	(0.018)	-0.0392*	(0.0167)	-0.0395*	(0.0168)	-0.0362**	(0.0125)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We find that in result to the increment of such contacts people experienced, on average, a reduction in the probability of experiencing severe loneliness by ≈ 4 percentage points.

As shown in Table 4.13, such a reduction was constant across all the models adapted. In particular, we obtained very similar results for Model 1, Model 2 and Model 3 , while for Model 4 we have small quantitative reduction of the effect and an increase in its significance, suggesting that some of the variables included in the model still account for some confoundness even after the weighting procedure.

Our results suggest that the treatment had a statistically significant and potentially meaningful impact on reducing severe loneliness among the treated individuals.

4.5.2 Multinomial Setup

In the multinomial setup we fitted the weighted regression model using "invariant-decrease" as reference level of the treatment factor, therefore, the estimates of the regression coefficient are intended with respect to such level.

For the sake of clarity we will now denote the level "increase" as "1", the level "increase-decrease" as "2" and the level "invariant-decrease" as "3".

Table 4.14: Treatment coefficients and ATT estimates for each Q-model (multinomial treatment)

Treatment	Model 1		Model 2		Model 3		Model 4	
	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)
1	-0.8251*	(0.3223)	-0.8903**	(0.3396)	-0.8940**	(0.3411)	-1.1178**	(0.3961)
2	-0.4236	(0.4970)	-0.5365	(0.5106)	-0.5294	(0.5094)	-0.5599	(0.4666)
ATT: 1-3	-0.0459*	(0.0222)	-0.0441*	(0.0198)	-0.0442*	(0.0198)	-0.0411**	(0.0148)
ATT: 1-2	-0.0182	(0.0233)	-0.0143	(0.0204)	-0.0148	(0.0205)	-0.0177	(0.0164)
ATT: 3-2	0.0277	(0.0309)	0.0298	(0.0271)	0.0295	(0.0271)	0.0234	(0.0186)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In the multiple treatment analysis we found that the people who only increased their non-physical contact during COVID-19 lockdown had a significant reduction in the probability of experiencing sever loneliness compared to the probability they would had if they had let them unchanged or decrease. In this setup, such effect, denoted as "ATT: 1-3" (see Table 4.14), is greater than 4 pp for all model tested.

Our suspicious is that considering only the individuals who only increased their remote contacts allowed us to see clearly the effect of such an increase against what would happened if they had not, after accounting for the impact of mixed pattern of contacts.

In particular, we found the "increase" treatment diminished, from a numerical perspective, the probability of perceived severe loneliness even when compared with the mixed pattern "increase-decrease", but the comparison did not yield a significant effect. Such result might be due in first place to the lower number of units undergoing the mixed-pattern treatment. Secondly, it might be the case that, when considering the severe loneliness as response, any pattern concerning an increase of connections is helpful in lowering the probability of such outcome.

Regarding the comparison of the ATTs of the "increase-decrease" and "invariant-decrease" treatments among people who increased their non-physicial contacts (denoted by "ATT: 3-2" in Table 4.14) we observe a positive numerical effect in all model. This result is aligned with what we said so far, suggesting that the probability of severe loneliness for individuals who increased their remote contacts (we are always referring to the treated population) would have been higher if they did not increase such connection even compared with a mixed situation in which, aside for increasing non-physical contacts with someone, they also decreased their relations with someone else. The lack of statistical significance of the effect suggests, anyway, that there is insufficient evidence to do any claim regarding such comparison.

Finally, when examining the results of the different models we have that Model 4, accounting for all the covariates, produce a slightly more accurate estimate of the ATTs of interests (lower SE) making the effect even more evident (p-value = 0.00542).

4.6 Limitations and Further Development

As in all statistical works, the key limitations of this study are *assumptions.* In Chapter 3 we discussed that the results here presented are based on the *strong ignorability* assumption and, therefore, they are valid only if the assumption holds. Now, if *positivity* was measured and assessed checking the overlap of the distributions of the propensity scores, *unconfoundness* remain an open question. In this study we considered a set of variables accounting for potential confounders as robustness check in order to account for all the possible *observed* confounders but we do not have control over *unmeasured* characteristic that might influence the relationship between non-physical contacts and severe loneliness (treatment and outcome, respectively). Future studies might address the problem implementing *sensitivity analysis* and assess the robustness of the results against potential unmeasured confounders or increase, based on new research, the set of covariates considered in the analysis.

Another limitation of our approach concerns with the variability of the treatment effect. As we saw, the standard error of the Q-model account for the weighting procedure through the Thomson-Horvitz estimator of the variance (Horvitz and Thompson, 1952), such estimator assumes that the weights are given and, therefore, does not take into account the uncertainty due to the propensity score estimation. McCaffrey *et al.* (2004) suggest that ignoring the uncertainty of the estimation procedure and computing the variance using analytical estimator results in an upper bound for the actual sampling variability of the estimated treatment effect for the observed sample. This critique, while not posing any threat to the validity of our results, could be easily addressed whit the implementation of bootstrap procedures that empirically factor in such variability. In our work, due to the computational complexity of the GBM, such procedures resulted prohibitive, so other solutions might be explored to obtain the precise estimate of the effect variability.

Further research might also explore the treatment heterogeneity using the information (available in the Intergen-COVID survey data) on the specific communication technology used by the individual in order to understand which of them (video-calls, instant messaging or social networks) have been more helpful in counteract the loneliness feeling during the lockdown.

Chapter 5

Conclusions

This study investigated the impact of different patterns of non-physical contacts on mitigating the feelings of severe loneliness, especially among older Italian adults during the COVID-19 lockdown.

While previous research have examined the efficacy of non-physical interactions this work represents, to the best of our knowledge, the first application of a *doubly robust* method combining a *machine learning* based approach to propensity score weighting and *G-computation* for estimating the Average Treatment Effect on the Treated (ATT). The integration of this methodological approach in our analysis provides a novel perspective for future research in social health and well-being.

Utilizing data from the Intergen-COVID online survey, we demonstrated that enhanced non-physical contact during the pandemic alleviated feelings of loneliness among the study population. The analysis revealed that individuals who increased their non-physical contacts saw, on average, a reduction in the probability of experiencing severe loneliness by approximately 4 percentage points. This effect was consistent and statistically significant across all models tested, highlighting the potential of remote interactions in reducing the adverse effects of social isolation. Although the increase treatment group showed a numerically lower probability of severe loneliness compared to the 'mixed pattern' group, the statistical analysis did not reveal a significant difference. Our results also suggest that scenarios where contacts are both increased and decreased might be beneficial in reducing the risk of severe loneliness when compared with situation of complete reduction in connections. However, the absence of statistical significance in this comparison calls for cautious interpretation of these findings.

For policymakers, these insights suggest the importance of promoting and facilitating non-physical interactions among older populations as a possible public health intervention, particularly in times of enforced physical distancing.

In conclusion, our research has highlighted the potential of non-physical interactions to mitigate loneliness. It is a clarion call for a collective and informed response to one of the silent epidemics of our time – loneliness.

Bibliography

Arel-Bundock V. (2023). *marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests.* R package version 0.17.0.

Arpino B.; Pasqualini M.; Bordone V.; Solé-Auró A. (2020). Indirect consequences of covid-19 on people's lives: Findings from an online survey in france, italy and spain.

Arpino B.; A. C. M.; Quashie N. T.; Antczak R. (2022). Loneliness before and during the covid-19 pandemic—are unpartnered and childless older adults at higher risk? *European Journal of Ageing*, **19**(1), 19–32.

Breiman L.; Friedman J.; Olshen R. A.; Stone C. J. (1984). *Classification and Regression Trees.* Chapman and Hall/CRC, 1 edition.

Cacioppo J. T.; Hawkley L. C.; Crawford L. E.; Ernst J. M.; Burleson M. H.; Kowalewski R. B.; Malarkey W. B.; Van Cauter E.; Berntson G. G. (2002). Loneliness and health: Potential mechanisms. *Psychosomatic Medicine*, **64**, 407–417.

Cacioppo S.; Grippo A. J.; London S.; Goossens L.; Cacioppo J. T. (2016). Loneliness: Clinical import and interventions. *Perspect Psychol Sci*, **10**(2), 238–249. PMCID: PMC2016 March 01.

Cannas M.; Arpino B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, **61**(6), 1049–1072.

Cefalu M.; Ridgeway G.; McCaffrey D.; Morral A.; Griffin B. A.; Burgette L. (2022). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups.* R package version 2.5.

Chang T.-H.; Stuart E. A. (2022). Propensity score methods for observational studies with clustered data: A review. *Statistics in Medicine*, **41**(18), 3612–3626.

Cochran W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A*, **128**, 234–255.

Dahlberg L. (2021). Loneliness during the covid-19 pandemic. *Aging & Mental Health*, **25**(7), 1161–1164.

Drake C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, **49**, 1231–1236.

Freund Y.; Schapire R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**, 119–139.

Friedman J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**(5), 1189–1232.

Friedman J. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, **38**(4), 367–378.

Greifer N.; Stuart E. A. (2021). Matching methods for confounder adjustment: An addition to the epidemiologist's toolbox. *Epidemiologic Reviews*, **43**(1), 118–129.

Griffin B. A.; McCaffrey D. F.; Almirall D.; Burgette L. F.; Setodji C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of Causal Inference*.

Hastie T.; Tibshirani R.; Friedman J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pp. 259–335. Springer Series in Statistics. Springer New York, NY, 2 edition.

Heinrich L. M.; Gullone E. (2006). The clinical significance of loneliness: a literature review. *Clin Psychol Rev*, **26**(6), 695–718.

Holland P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**(396), 945–960.

Holland P. W.; Rubin D. B. (1980). Causal inference in prospective and retrospective studies. In *Proceedings of the American Statistical Association Annual Meeting, Jerome Cornfield Memorial Session*, American Statistical Association Annual Meeting.

Holt-Lunstad J.; Smith T. B.; Baker M.; Harris T.; Stephenson D. (2015). Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science*, **10**(2), 227–237.

Hong J. H.; Nakamura J. S.; Berkman L. F.; Chen F. S.; Shiba K.; Chen Y.; Kim E. S.; VanderWeele T. J. (2023). Are loneliness and social isolation equal threats to health and well-being? an outcome-wide longitudinal approach. *Perspectives on Psychological Science*.

Horvitz D. G.; Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Lee B. K.; Lessler J.; Stuart E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, **29**(3), 337–346.

Liddle J.; Stuart A.; Worthy P.; et al. (2020). Building the threads of connection that we already have: The nature of connections via technology for older people. *Clinical Gerontologist*.

Lim M. H.; Eres R.; Vasan S. (2020). Understanding loneliness in the twenty-first century: an update on correlates, risk factors, and potential solutions. *Social Psychiatry and Psychiatric Epidemiology*, **55**(7), 793–810.

Lumley T. (2023). *survey: Analysis of Complex Survey Samples*. R package version 4.2-1.

Lumley T.; Scott A. (2017). Fitting regression models to survey data. *Statistical Science*, **32**(2), 265–278.

Macdonald B.; Hülür G. (2021). Well-being and loneliness in swiss older adults during the covid-19 pandemic: The role of social relationships. *The Gerontologist*, **61**(2), 240–250.

McCaffrey D. F.; Ridgeway G.; Morral A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, **9**(4), 403–425.

McCaffrey D. F.; Griffin B. A.; Almirall D.; Slaughter M. E.; Ramchand R.; Burgette L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, **32**(19), 3388–3414.

Morrish N.; Mujica-Mota R.; Medina-Lara A. (2022). Understanding the effect of loneliness on unemployment: propensity score matching. *BMC Public Health*, **22**, 740.

Neugebauer R.; van der Laan M. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, **129**, 405–426.

Neyman J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Roczniki Nauk Rolniczych Tom X [in Polish]; translated in Statistical Science*, **5**, 465–480.

Parast L. (2022). *SBdecomp: Estimation of the Proportion of SB Explained by Confounders*. R package version 1.2.

Parast L.; Griffin B. A. (2020). Quantifying the bias due to observed individual confounders in causal treatment effect estimates. *Statistics in Medicine*, **39**(18), 2447–2476. PMID: 32388870; PMCID: PMC8162899.

Pearl J. (1996). Causation, action and counterfactuals In *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge.* Edited by Shoham Y., pp. 57–73, San Francisco, CA. Morgan Kaufmann.

Ridgeway G. (1999). The state of boosting. *Computing Science and Statistics*, **31**, 172–181.

Robins J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Modelling*, **7**(9-12), 1393–1512.

Rosenbaum P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, **82**(398), 387–394.

Rosenbaum P. R. (2023). *CAUSAL INFERENCE*, p. 42.

Rosenbaum P. R.; Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5), 688–701.

Rubin D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, **6**, 34–58.

Rubin D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, **26**(1), 20–36.

Salvan A.; Sartori N.; Pace L. (2020). *Modelli Lineari Generalizzati*, pp. 161–189. UNITEXT. Springer Milano.

Snowden J. M.; Rose S.; Mortimer K. M. (2011). Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, **173**(7), 731–738.

Thurston R. C.; Kubzansky L. D. (2009). Women, loneliness, and incident coronary heart disease. *Psychosom Med*, **71**(8), 836–842.

Valtorta N. K.; Kanaan M.; Gilbody S.; Ronzi S.; Hanratty B. (2016). Loneliness and social isolation as risk factors for coronary heart disease and stroke: systematic review and meta-analysis of longitudinal observational studies. *Heart*, **102**(13), 1009–1016.

VanderWeele T. J.; Shpitser I. (2013). On the definition of a confounder. *Annals of Statistics*, **41**(1), 196–220.

Victor C. R.; Yang K. (2012). The prevalence of loneliness among adults: A case study of the united kingdom. *Journal of Psychology*, **146**, 85–104. PubMed: 22303614.

Wilson R. S.; Krueger K. R.; Arnold S. E.; Schneider J. A.; Kelly J. F.; Barnes L. L.; Tang Y.; Bennett D. A. (2007). Loneliness and risk of alzheimer disease. *Arch Gen Psychiatry*, **64**(2), 234–240.

Zanutto E.; Lu B.; Hornik R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, **30**(1), 59–73.