Méthodes d'Estimation du Retour sur les Investissements en Marketing

Essai

Chaimae Sriti

Maitrise statistique – Statistique

Québec, Canada

Méthodes d'estimation du retour sur les investissements en marketing

Essai

Chaimae Sriti

Sous la direction de :

Thierry Duchesne, directeur de recherche

Louis-Paul Rivest, codirecteur de recherche

Résumé

Dans l'objectif de commercialiser leurs produits d'assurance auprès des clients potentiels, les compagnies d'assurance investissent des sommes considérables en publicité de toute sorte : sollicitation par téléphone, envois postaux, publicités à la radio ou à la télévision, etc. Afin de guider leurs décisions stratégiques et d'optimiser leur politique d'investissement, il faut quantifier le retour de chaque investissement en terme de profit et par conséquent en soumissions reçues. Il serait ainsi intéressant de savoir combien de nouvelles soumissions d'assurance sont générées pour chaque dollar investi en publicité dans un certain type de média. Toutefois, comme les dépenses publicitaires dans les différents médias sont fortement corrélées entre elles, et comme il y a une forte saisonnalité au niveau des soumissions, il s'avère difficile d'isoler l'effet de chaque média, et de mesurer son retour sur investissement.

Dans le présent essai, on développe une méthodologie pour traiter cette problématique dans le contexte de données observationnelles. Dans une optique comparative, nous ajustons plusieurs modèles de régression à l'aide d'approches inférentielles statistique et causale, en vue d'estimer le retour sur investissement d'un média donné tel que mesuré par le nombre de soumissions reçues par la compagnie d'assurance « ASSURE » sur une période donnée.

Abstract

In order to market their insurance products to potential customers, insurance companies invest considerable amounts of money in advertising : phone solicitations, mailings, advertisements on radio or television, etc. In order to guide their strategic decisions and optimize their investment policy, it is necessary to quantify the return of each investment in terms of profit and consequently in terms of the requests for a quote received. It would be interesting to know how many new insurance quotes are generated for every dollar invested in advertising in every type of media. However, since the advertising expenses in the various media are strongly correlated with each other, and since there is a strong seasonality in the quote requests' cycle, it is difficult to isolate the effect of each media, and measure its return on investment.

In this essay, a methodology is developed to address this issue in the context of observational data. In a comparative perspective, we fit several regression models using statistical and causal inferential approaches, in order to estimate the return on a given media's investment on the number of quotes received by the insurance company « ASSURE » in a given period.

Table des matières

Liste des tableaux

Liste des figures

À Radia.

"The purpose of life is not to be happy. It is to be useful, to be honorable, to be compassionate, to have it make some difference that you have lived and lived well."

Ralph Waldo Emerson

Remerciements

Mes remerciements les plus sincères vont à mon directeur de recherche Thierry Duchesne, qui m'a guidée tout au long de mon parcours académique de maîtrise à l'Université Laval. Je le remercie pour sa grande disponibilité, son implication infaillible, et ses conseils précieux qui m'ont accompagnée tout au cours de ce projet de recherche. Je tiens à remercier aussi sincèrement mon co-directeur de recherche Louis-Paul Rivest, dont les directives et les suggestions ont été d'une aide énorme dans ce travail. Je le remercie du fond du cœur pour son dévouement, ses remarques, sa grande disponibilité, et son suivi tout le long de la rédaction de cet essai. Un grand merci à vous deux d'avoir accepté de diriger mon travail de recherche, et de m'avoir donné l'opportunité de travailler sur cet intéressant projet. Vous rencontrer a toujours été une opportunité d'apprentissage et d'évolution pour moi. La rédaction de ce document n'aurait pas été possible sans vos conseils. Ce fut un vrai plaisir de travailler avec vous. Je remercie également Denis Talbot pour ses conseils judicieux.

Par ailleurs, je tiens à remercier la compagnie qui m'a permis de réaliser ce travail de recherche. Ce fut un plaisir d'interagir et de collaborer avec vos consultants. Je remercie S.D de nous avoir accompagné tout au long de ce projet. Vos interventions ainsi que les échanges qu'on a eus avec vous ont été très enrichissants. Un grand merci également à MITACS qui a subventionné ce projet avec la compagnie.

Je tiens à remercier infiniment ma famille, ma raison d'être. Je remercie ma mère, la statisticienne qui m'inspire le plus, et la femme à qui je dois ce que je suis devenue aujourd'hui et ce que j'espère devenir demain, Merci d'être ma lumière . Je remercie mon père, l'amour de ma vie, je le remercie de m'avoir tout appris, et d'être ma source inépuisable de bonheur. Je remercie Yasser, ma plus grande bénédiction et le plus beau cadeau que la vie m'a offert, ton existence fait ma joie.

Je tiens à remercier toutes les personnes qui m'ont accompagnée dans cette aventure qui a commencé il y a deux ans. Ouassima, qui est beaucoup plus qu'une amie, une sœur. Merci d'avoir toujours été là pour le meilleur, mais surtout pour le pire. Merci d'avoir été la meilleure partenaire d'étude tout au long de ce parcours de sept ans. Merci pour tous ces beaux et drôles souvenirs passés, et vivement tous ceux qui sont à venir ! Un grand merci à ma petite sœur Majda pour toute la joie qu'elle me transmettait ! Merci à Souad, Hanaa, Imane et Ghizlane.

Grâce à vous toutes, ces deux années ont été exceptionnellement agréables.

Finalement, je remercie Dieu pour tout. Aucun mot ne saura combler l'expression de ma gratitude et de mon amour envers Toi.

Introduction

Le retour sur investissement est une mesure de performance qui permet d'évaluer l'impact d'un investissement ou de comparer l'effet de plusieurs investissements. Cette mesure indique directement le profit généré par un certain investissement et de juger par conséquent de sa pertinence. En marketing, le MROI (Marketing Return On Investment) permet de mesurer la rentabilité des efforts marketing, des actions publicitaires par exemple, et revient à mettre en relation les coûts de compagne et l'activité commerciale générée (par exemple 1 dollar investi rapporte 5 dollars de chiffre d'affaire).

Aujourd'hui, les compagnies d'assurance comptent sur la publicité pour faire connaitre leurs produits auprès de la clientèle potentielle. Ainsi, pour optimiser leur politique d'investissement (minimiser les coûts publicitaires tout en maximisant le profit), il serait judicieux pour toute compagnie de mesurer le retour sur investissement de chaque média publicitaire. Toutefois, cette tâche s'avère difficile et complexe vu que les données de marketing sont observationnelles et non pas une expérience randomisée, ceci implique la présense de corrélation qu'on retrouve entre les dépenses publicitaires dans les différents médias et avec les cycles du marché. Également, elles sont sujettes à de multiples variations saisonnières et souvent confondues avec les effets de variables non observables. Quand il est question d'isoler l'effet de l'investissement, il faut corriger pour toute autre variable confondante et donc construire des modèles qui contiennent à la fois des paramètres mesurant l'effet des dollars investis dans chaque type de publicité sur le nombre de soumissions tout en éliminant l'effet de la saisonnalité, et en prenant en compte les facteurs confondants, les variables instrumentales et l'interaction entre les deux.

Dans le présent essai, on traite la question de la mesure du retour sur l'investissement en marketing dans le cadre de données observationnelles et dans le contexte d'une compagnie d'assurance. L'objectif est de cerner le mieux possible l'effet causal d'un certain type d'investissement à travers l'ajustement de différents modèles d'inférence statistique et causale dans une optique comparative. Pour ce, nous disposons d'un jeu de données avec d'une part des soumissions et de l'autre des efforts marketing. On souhaite quantifier l'impact des efforts marketing sur les soumissions, mais plusieurs complexités surviennent : une forte saisonnalité des soumissions ; davantage d'efforts marketing lors des semaines d'achalandage, ce qui rend

difficile la tâche d'isoler l'effet de la saisonnalité naturelle de celui des efforts publicitaires ; une corrélation forte entre des efforts marketing qui agissent en même temps ; et la présence d'une corrélation avec des variables socio-démographiques qui influencent à la fois les soumissions et les efforts marketing.

Cet essai est divisé en deux parties. Le chapitre 1 présente les différentes notions statistiques qui sont utilisées dans le cadre de cette application. Il passe en revue quelques méthodes d'analyse exploratoires et les méthodes classiques de régression pour introduire l'inférence causale et le modèle du score de propension. Le chapitre 2 présente les résultats de l'analyse de données réalisée dans le cadre de ce projet. Ce chapitre commence par présenter la méthodologie suivie en vue de répondre à la question de recherche. On présente d'abord les résultats de l'analyse exploratoire descriptive des données, pour passer à l'ajustement des modèles qui permettent de conclure le retour sur investissement d'un certain média.

Chapitre 1

Notions statistiques préliminaires

Ce chapitre présente une brève introduction aux méthodes et concepts statistiques qui sont utilisés dans cet essai et qui seront nécessaires à la compréhension de l'application réalisée dans le cadre de ce projet. Il est à noter que dans ce chapitre, nous ne détaillons pas le développement mathématique/théorique des différentes notions traitées. L'objectif est de permettre au lecteur de se situer dans le contexte de l'étude et de comprendre l'objectif de chaque technique statistique utilisée ainsi que ses conditions d'utilisation. Pour le développement théorique, nous invitons le lecteur à se référer aux références bibliographiques. Nous supposons également que le lecteur est familier avec les notions d'analyse exploratoire et les méthodes de régression linéaire classiques. Nous en faisons ainsi un bref rappel pour mieux appréhender la notion d'inférence causale.

Dans la section 1.1, nous présentons une technique d'analyse exploratoire/de réduction de la dimensionnalité, à savoir l'analyse en composantes principales et nous introduisons brièvement la méthode de classification non supervisée K-moyennes dans le cadre de l'analyse de regroupement. Dans la section 1.2, nous faisons un bref rappel des méthodes classiques de régression ainsi que de quelques méthodes de sélection de variables. Dans la section 1.3, nous exposons la notion d'inférence causale. Nous présentons ensuite le modèle du score de propension et le modèle avec les poids d'appariement.

1.1 Rappel de quelques méthodes statistiques exploratoires

Dans cette section, nous faisons un bref rappel de deux techniques statistiques d'analyse exploratoire, la méthode de réduction de dimension *Analyse en composantes principales* et la méthode de classification non-supervisée *K-moyennes*. Nous expliquons l'objectif et le principe de la méthode, ainsi que les fonctions R permettant de la réaliser. Pour le développement théorique et les procédures algorithmiques, nous référons le lecteur aux références bibliographiques.

1.1.1 Analyse en composantes principales

L'analyse en composantes principales est une méthode d'analyse de données non supervisée dont l'objectif est de réduire le nombre de variables d'un ensemble de données tout en préservant le plus d'information possible. Cette méthode a été introduite par Pearson (1901) et développée par Hotelling (1933). Aujourd'hui, la méthode est utilisée dans une multitude d'applications où il est nécessaire de visualiser les jeux de données de grande dimension, réduire le nombre de variables pour faciliter l'ajustement de modèles d'apprentissage automatique, faire une rotation d'axes pour simplifier la structure de corrélation, etc.

L'idée centrale de l'analyse en composantes principales (ACP) est de réduire la dimensionnalité des données initiales (qui est p si l'on considère p variables quantitatives), en remplaçant les p variables initiales par q dimensions appelées composantes principales ($q < p$). Ainsi, pour former la première composante principale, on cherche la combinaison linéaire des variables d'origine qui aurait une variance maximale, pour former la deuxième composante principale, on cherche une deuxième combinaison linéaire des variables d'origine qui a une variance maximale, et qui est orthogonale à la première, et ainsi de suite. Pour le développement mathématique de l'analyse en composantes principales, on réfère le lecteur à Jolliffe (2016).

Il est à noter que parfois, la standardisation des données est nécessaire avant le passage à l'ACP (Jolliffe (2016)). Cette technique s'avère sensible à l'échelle de mesure des variables. En effet, comme l'ACP cherche une combinaison qui maximise la variance, une variable de grande variance aura beaucoup de poids dans les composantes principales. Par exemple, une mesure de masse en g plutôt qu'en kg multiplierait la variance de cette variable par un million, ce qui fait que cette mesure aurait un poids majeur sur toutes les composantes. Ainsi, lorsqu'on désire conduire une ACP sur des variables qui ont des unités différentes, il est recommandé de procéder à la standardisation des données.

Pour interpréter l'ACP, on dispose des coordonnées des observations et des variables sur chaque composante. Il faut pouvoir donner du sens à chacune des composantes principales retenues. Cela nécessite d'examiner le poids et la direction des coefficients des variables d'origine sur chaque axe. Plus la valeur absolue du coefficient est grande, plus la variable correspondante est importante dans le calcul de la composante.

En pratique, plusieurs fonctions sur le logiciel R permettent de réaliser une ACP, et de visualiser les résultats. En R, deux fonctions dans la librairie de base stats permettent de faire l'ACP : **princomp** et **prcomp**. Également, la librairie **FactoMineR** (Lê et al. (2008)) permet de réaliser une ACP complète grâce à la fonction **PCA**. La librairie **factoextra** (Kassambara and Mundt (2017)) permet l'extraction des résultats de l'ACP ainsi qu'une bonne visualisation des sorties de l'ACP. Nous disposons de trois graphiques principaux qui permettent d'interpréter les résultats de l'ACP. Un premier graphique qui représente le pourcentage de variabilité expliquée par chacune des composantes principales, ce graphique peut être généré par la fonc-

tion fviz_eig() du package **factoextra**. En vue d'extraire les résultats pour les variables et pour les individus, on peut utiliser la fonction get_pca_var() et get_pca_ind(), respectivement, du même package. Ces deux fonctions retournent une liste d'éléments contenant tous les résultats pour les variables et les individus (coordonnées, corrélation entre variables (i.e les individus) et les axes, cosinus-carré et contributions). Nous pouvons visualiser les variables et les individus et les colorer en fonction soit de leurs qualités de représentation (cos2) (Abdi and Lynne (2010)) ou de leurs contributions aux composantes principales (contrib).

1.1.2 Classification K-moyennes

La classification K-moyennes est une méthode d'apprentissage non supervisé qui consiste à identifier des groupes d'observations ayant des caractéristiques similaires. Le but de la méthode des K-moyennes (Lloyd (1982), Forgy (1965)) est de déterminer un partitionnement de ces individus en groupes homogènes, le nombre K de groupe étant choisi a priori.

L'idée principale est de choisir aléatoirement un ensemble de centroïdes fixé a priori et de chercher itérativement la partition optimale. Chaque individu est affecté au centroïde le plus proche, après l'affectation de toutes les données la moyenne de chaque groupe est calculée, elle constitue les nouveaux centroïdes des groupes à l'aide desquels on effectue une nouvelle allocation des individus. Lorsqu'on aboutit à un état stationnaire (aucune donnée ne change de groupe) l'algorithme est arrêté.

Pour mieux illustrer cet algorithme, on se réfère à la figure 1.1 ci-dessous.

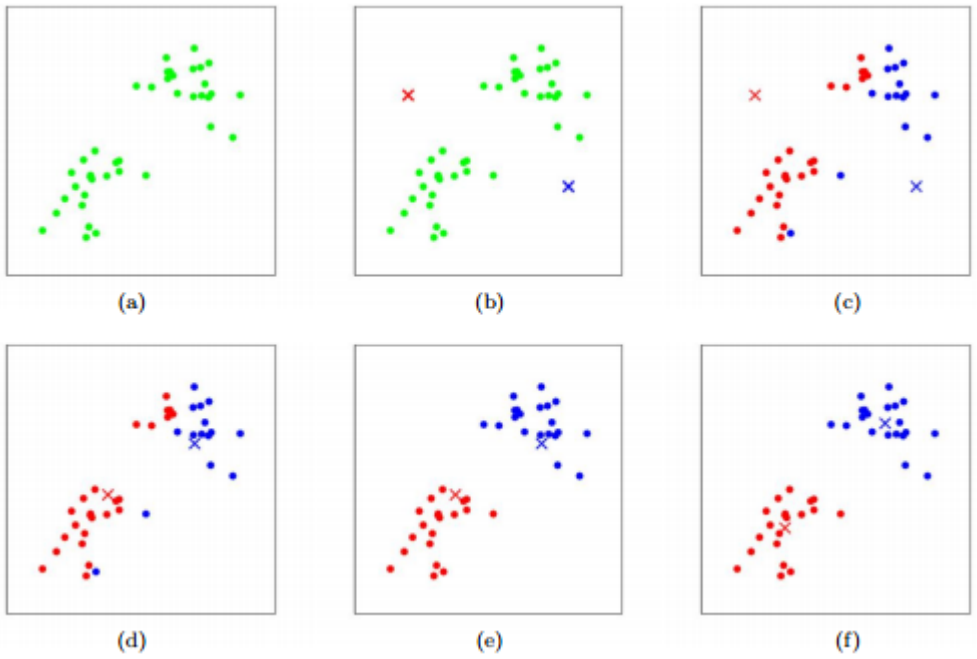


FIGURE 1.1 – Algorithme K-moyennes : Les observations sont représentées par des points et les centroïdes des groupes par des croix. (a) Jeu de données d'origine. (b) Centroïdes de groupe initiaux aléatoires. (c-f) Illustration de l'exécution de deux itérations de K-moyennes. Dans chaque itération, nous attribuons chaque donnée au centroïde du groupe le plus proche (indiqué en colorant les données de la même couleur que le centroïde du groupe auquel il est affecté) ; on déplace ensuite chaque centre du groupe vers la moyenne des points qui lui sont attribués. Source : Kulis and Jordan (2011)

Nous réfèrons le lecteur à Sanghamitra and Sriparna (2012) pour avoir les détails théoriques et pratiques sur cette méthode de classification.

Analyse de regroupement

Dans cette sous-section, nous voulons répondre à la question : Quel est le nombre optimal de groupes à choisir pour la réalisation de la classification ? La détermination du nombre optimal de groupes à créer à partir des données est une question importante dans les méthodes de classification. La classification K-moyennes nécessite la spécification du nombre de groupes à former à priori. Pour un même jeu de données, plusieurs regroupements sont possibles. La difficulté résidera donc à choisir un nombre de groupes K qui permettra de mettre en lumière des patterns intéressants.

Une des méthodes possibles pour choisir le nombre de groupes est de lancer K-moyennes avec différentes valeurs de K et de calculer la variance des différents groupes. La variance est la somme des distances carrées entre le centroïde d'un groupe et les différentes observations inclues dans le même groupe. Ainsi, on cherche à trouver un nombre de groupes K de telle

sorte que les groupes retenus minimisent la distance entre leurs centroïdes et les observations incluses dans le même groupe. On parle de minimisation de la variance intra-classe.

En pratique, la fonction NbClust du package portant le même nom (Charrad et al. (2014)) nous permet de déterminer le nombre optimal de groupes à retenir sur la base d'une trentaine d'indices de validation.

1.2 Rappels sur les méthodes de régression classiques

1.2.1 Modèles de régression linéaire

L'analyse de régression est une méthode statistique permettant de modéliser les relations entre une variable dépendante et une ou plusieurs variables explicatives (indépendantes). Dans les applications marketing , la variable dépendante est généralement une variable résultat (outcome) à laquelle on s'intéresse (par exemple, les ventes), tandis que les variables indépendantes sont les instruments avec lesquels nous parvenons à ce résultat (par exemple, les tarifs, les investissements ou encore la publicité). L'analyse de régression permet ainsi de quantifier l'effet de chaque variable indépendante x_i sur la variable Y que l'on cherche à expliquer. Ainsi, une analyse de régression (Sarstedt and Mooi (2014)) peut être utilisée pour indiquer s'il y a une relation significative entre les variables indépendantes et la variable dépendante, indiquer la force relative des effets de différentes variables indépendantes sur la variable dépendante ou encore pour faire des prédictions.

Les modèles de régression à adopter dans une analyse de données peuvent varier selon la nature de la variable dépendante. Dans le présent projet, nous avons recours au modèle de régression linéaire multiple (Kutner et al. (2005)) lorsque la variable dépendante est continue, et au modèle de régression logistique (Dunn and Gordon (2018)) lorsque la variable dépendante est dichotomique.

Dans un deuxième temps, on a recours au modèle de régression avec les observations pondérées (Chang and Lee (1996)). Une variable poids assigne une valeur (un poids) à chaque observation. Dans la majorité des applications, un poids prend une valeur non négative. Un poids égal à zéro signifie généralement l'exclusion de l'observation en question de l'analyse. Les observations ayant des poids relativement importants ont plus d'influence dans l'analyse que les observations ayant des poids plus petits. Une analyse non pondérée est identique à une analyse pondérée dans laquelle tous les poids sont égaux. En fait, on peut donner des poids pour donner plus d'importance aux observations plus précises ou corriger des biais dus à l'échantillonnage quand la représentativité des individus dans l'échantillon n'est pas la même que dans toute la population.

1.2.2 Méthodes de sélection de variables

Souvent dans les analyses statistiques, on se retrouve avec des bases de données contenant un grand nombre de variables exogènes. Comme l'objectif est d'obtenir des estimés de coefficients de régression non biaisés, il faut s'assurer que toutes les variables exogènes influant la variable à expliquer soient incluses dans le modèle. Toutefois, inclure un grand nombre de variables explicatives peut nuire à la robustesse et la précision des estimés (Li et al. (2016)). Ainsi, il faut procéder à la sélection d'un sous-ensemble de variables indépendantes à inclure dans le modèle de régression. Dans cette partie, on explique brièvement les différentes méthodes de sélection de variables qui ont été utilisées dans le présent projet.

- Sélection Stepwise

Pour chaque variable explicative, il faut savoir s'il est pertinent de l'inclure ou non dans le modèle. Pratiquement, on teste si le coefficient d'une variable explicative x_i est nul, ceci revient à tester l'hypothèse $H_0 : \beta_j = 0$. Plusieurs façons permettent de faire ce test, et ne conduisent pas forcément au même modèle final. Il existe trois méthodes classiques de sélection pas-à-pas de variables de régression (James et al. (2013), Bruce (2017)) :

1. Méthode descendante (backward elimination) : Nous partons du modèle contenant toutes les variables explicatives. Lors de chaque étape, la variable ayant la plus grande p-value du test de Student (ou de Fisher) est supprimée du modèle si cette p-value est supérieure à un seuil défini à l'avance (généralement 0.10 ou 0.05). La procédure s'arrête lorsque toutes les p-values sont inférieures au seuil.
2. Méthode ascendante (forward selection) : Nous partons du modèle nul avec uniquement l'ordonnée à l'origine. Ensuite, on introduit une à une et la variable la plus significative.Tant que le modèle continue de s'améliorer, nous continuons le processus en ajoutant une variable à la fois. On s'arrête dès qu'aucune des variables à ajouter n'est jugée significative.
3. Méthode pas-à-pas (stepwise selection) : La méthode pas-à-pas itère entre la méthode ascendante et la méthode descendante. On fixe un seuil d'entrée et un seuil de sortie. On commence avec le modèle ne contenant que l'ordonnée à l'origine. On procède à un pas de la méthode ascendante. Après ce pas, on effectue un pas de la méthode descendante afin que les variables présentes dans le modèle aient un seuil inférieur au seuil de sortie fixé. On fait ensuite un autre pas de la méthode ascendante, et ainsi de suite. L'algorithme cesse lorsqu'aucune variable additionnelle ne peut entrer dans le modèle et qu'aucune variable ne peut être sortie du modèle.

Dans la pratique, la fonction *step* du package R de base stats permet de faire la sélection de variables selon les méthodes précitées.

1.3 Inférence causale

Dans plusieurs domaines de recherche, l'objectif est souvent de quantifier l'effet d'une intervention particulière sur une variable d'intérêt. En santé, on peut s'intéresser à l'effet d'un nouveau traitement par rapport aux soins classiques. En marketing, on peut s'intéresser à quantifier l'impact d'une nouvelle campagne de publicité sur les ventes d'un certain produit. L'intérêt pour ces questions est motivé par des préoccupations de politiques ou de stratégies à suivre afin d'optimiser un certain processus.

De point de vue statistique, pour conduire une analyse causale, le défi serait d'isoler l'effet de l'intervention à laquelle on s'intéresse puisque, à moins que cette dernière agisse indépendamment de tout autre facteur, l'effet déduit peut être influencé par d'autres variables. L'analyse de régression permet d'inférer l'association entre différentes variables, d'estimer la probabilité d'événements passés et futurs, et de faire des prédictions. À moins que le modèle de régression ne soit parfait et tienne en compte toutes les variables influençant l'issue (la variable dépendante/outcome) et inclut toutes les variables confondantes corrélées à l'intervention et susceptibles d'avoir une influence sur l'issue d'intérêt, inférer la causalité à partir d'un modèle de régression peut s'avérer risqué. Pour s'assurer de mesurer correctement un effet causal, il serait mieux d'avoir recours à des méthodes d'inférence causale (Miguel and Robins (2019)).

Les techniques d'inférence causale sont construites pour prédire l'effet qu'aurait une intervention potentielle, comme un traitement, une campagne ou une politique, sur une issue. Les inférences causales sont le plus facilement effectuées à l'aide d'études randomisées, mais des approches pour inférer la causalité à partir de données observationnelles sont également établies. En effet, il est d'abord nécessaire d'identifier les variables confondantes pouvant biaiser l'association entre l'exposition (intervention/traitement) et l'issue d'intérêt. À moins que ces tâches ne soient accomplies adéquatement, les conclusions tirées peuvent être erronées, et peuvent ainsi induire en erreur les décideurs.

Dans la sous-section qui suit, nous présentons quelques notions permettant de mieux appréhender le concept d'inférence causale en partant d'un exemple introductif. Ensuite, nous présentons la méthode du score de propension dans le cas d'un traitement binaire, et la notion des poids d'appariement.

1.3.1 Quelques notions préliminaires

- **Exemple introductif**

Supposons que nous ayons mené une expérience randomisée dans laquelle nous mesurons certains résultats (par exemple, le taux de mortalité) au sein de deux groupes traité et non traité et dans lesquels des individus ont été attribués de manière aléatoire. Dans ce cas, l'affectation aléatoire du traitement signifie qu'il ne devrait y avoir en moyenne aucune différence signifi-

cative entre les groupes traités et non traités. Ainsi, toute différence de mortalité que nous observons entre les deux groupes serait due au traitement.

Supposons maintenant qu'on souhaite mesurer l'effet causal du tabagisme (Barter (2016)) sur la mortalité dans le cadre d'une étude observationnelle. Il est possible que la différence observée entre les groupes traité et non traité soit due à une variable autre que le traitement, soit le sexe par exemple. S'il existe un élément autre que le traitement qui diffère entre les groupes traité et non traité, nous ne pouvons pas affirmer de manière concluante que toute différence observée dans la mortalité entre les deux groupes est uniquement due au traitement. Une telle différence pourrait aussi être vraisemblablement due à ces autres variables qui diffèrent d'un groupe à l'autre.

Pour un exemple simplifié, on assume que les hommes sont plus susceptibles de fumer que les femmes, et qu'ils sont également plus susceptibles de mourir jeunes en raison d'une propension au risque. Dans ce cas, si nous constatons une différence de mortalité entre fumeurs et non-fumeurs, nous ne pourrions pas attribuer cette différence au tabagisme plutôt qu'au fait que le groupe de fumeurs contient plus d'hommes que le groupe contrôle et, par conséquent, à cause d'une aversion au risque, plutôt que dû au tabagisme, le taux de mortalité du groupe de traitement est supérieur à celui du groupe non traité. Dans ce cas, le taux de mortalité élevé chez les fumeurs n'est pas directement causé par le tabagisme seul, mais aussi par les disparités des sexes et la différence au niveau des comportements de prise de risque.

- **La randomisation et les études observationnelles**

L'effet causal peut être défini comme la différence entre le résultat obtenu sur le groupe traité et le résultat obtenu sur le groupe contrôle n'ayant pas reçu le traitement d'intérêt. Dans une étude randomisée, les participants sont aléatoirement assignés au groupe de contrôle et au groupe de traitement, la randomisation permet ainsi de réduire le biais lié à toute variable autre que le traitement. Dans une expérience randomisée, un échantillon de l'étude est divisé en un groupe qui recevra l'intervention étudiée (le groupe de traitement) et un autre groupe qui ne recevra pas l'intervention (le groupe de contrôle). Les études observationnelles sont celles dans lesquelles les chercheurs observent l'effet d'un facteur de risque, d'un traitement ou de toute autre intervention sans avoir aucun contrôle sur le choix de ces individus et de leurs caractéristiques. Dans le cadre de ces études, un défi majeur de l'inférence causale serait que l'exposition n'est pas nécessairement attribuée selon un mécanisme indépendant des autres variables. Par exemple, il se peut que l'exposition soit attribuée en fonction d'un ou plusieurs des prédicteurs mesurés ou pas.

- **La notion de variable confondante**

Une variable confondante est une variable aléatoire qui influence à la fois la variable dépendante et les variables indépendantes. L'existence de variables confondantes rend difficile de

conclure par rapport au lien de causalité à moins que les ajustements nécessaires prenant en compte ces variables soient faits. Dans le cas d'un traitement binaire et dans la présence d'un facteur confondant, le groupe d'individus traités et le groupe d'individus non traités ne sont pas directement comparables. L'effet d'une variable confondante C (Voir la figure 1.2) sur la variable dépendante serait potentiellement différent d'un groupe à l'autre. Dans cet exemple, une analyse naïve qui régresse Y sur X sans tenir compte de la valeur de C est susceptible de mener à des inférences biaisées.

FIGURE 1.2 – Absence d'un facteur confondant à gauche. Présence d'un facteur confondant C influençant à la fois X et Y à droite.

L'inférence causale consiste à comparer les groupes exposé et non exposé en partant des mêmes conditions, en isolant l'effet étudié de toute influence induite par d'autres variables indépendantes. Ceci serait possible avec une étude randomisée puisque celle-ci permettra d'éliminer l'association entre la variable indépendante principale X et une variable confondante C grâce à la randomisation. Dans le cas des données observationnelles, il faut procéder par des méthodes qui permettent d'ajuster pour les variables confondantes et d'en tenir compte.

- **La notion de contrefactuel**

En vue d'estimer l'effet causal, une des méthodes serait de définir le contrefactuel, c'est-à-dire le résultat qu'on aurait été obtenu si l'intervention n'avait pas été faite. En pratique, l'évaluation de l'effet causal nécessite de définir un groupe témoin pour estimer les résultats qu'auraient connus les unités d'observations si elles n'avaient pas subit l'intervention. Rubin (1977) propose un modèle causal contrefactuel où pour un même individu i il existe plusieurs résultats hypothétiques potentiels en fonction de l'exposition ou non de l'individu au traitement. L'effet causal d'un certain traitement serait la différence entre les deux résultats potentiels.

Dans ce qui suit, on présente la méthode du score de propension qui est adaptée au contexte d'études observationnelles, ainsi que la notion des poids d'appariement.

1.3.2 Score de propension avec traitement binaire

Dans cette sous-section, nous adoptons la notation suivante :

- i : l'indice des individus inclus dans l'étude,
- E : ensemble des individus inclus dans l'étude et exposés au traitement,
- N : ensemble des individus inclus dans l'étude et non exposés au traitement,
- Y_i : la variable réponse de l'individu i,
- Z_i : la variable dont on souhaite mesurer l'effet pour l'individu i, prend la valeur 0 si l'individu i n'est pas exposé au traitement, et la valeur 1 sinon,
- X_i : la valeur des facteurs confondants.

Pour formuler des questions d'inférence causale, il est utile de considérer certaines quantités de résultats hypothétiques qui représentent les résultats potentiels sous différentes alternatives d'exposition. Quand on considère le cas d'une exposition binaire, la paire des résultats potentiels $\{Y_i(0), Y_i(1)\}$ représente les résultats perçus par l'individu i si le sujet n'a pas été exposé, ou exposé, respectivement. Le résultat observé, Y_i, peut être écrit en termes de résultats potentiels et l'exposition observée, Z_i, comme $Y_i = (1 - Z_i) * Y_i(0) + Z_i * Y_i(1)$.

Pour chaque individu i, le traitement étudié cause l'effet suivant : $\delta_i = Y_i(1) - Y_i(0)$. Le problème fondamental de l'inférence causale est qu'il est impossible d'observer la valeur de $Y_i(1)$ et $Y_i(0)$ sur une même unité i. Par conséquent, il est impossible d'observer l'effet du traitement directement sur i. Comme nous n'avons pas la preuve contrefactuelle, c'est-à-dire que pour les individus traités, nous ne savons pas ce qui se serait passé en l'absence de traitement, nous ne pouvons pas inférer l'effet du traitement. Étant donné que l'effet causal pour un certain individu i ne peut pas être observé, nous cherchons à identifier l'effet causal moyen pour l'échantillon dans son ensemble. Pour cela, on mesure l'effet de traitement moyen (average treatment effect) « ATE ». Cet effet est estimé par :

$$\hat{\bar{\delta}} = \hat{\overline{Y_1}} - \hat{\overline{Y_0}} \tag{1.1}$$

où :

- $\overline{Y_1} = \frac{1}{\#E} \sum_{i \in E} Y_i$ la moyenne de la variable réponse Y dans le groupe exposé,
- $\overline{Y_0} = \frac{1}{\#N} \sum_{i \in N} Y_i$: la moyenne de la variable réponse Y dans le groupe non exposé.

Pour que cette estimation ne soit pas biaisée, il faut que les individus traités et non traités soient échangeables (l'attribution du traitement est indépendante des résultats potentiels), c'est-à-dire qu'il n'existe pas de facteurs confondants qui soient une cause commune du traitement et du résultat. Également, il faut que la probabilité de recevoir les deux niveaux de traitement soit positive pour chaque individu, appelée hypothèse de positivité. Cela signifie qu'il n'y a pas un individu pour qui recevoir le traitement est impossible.

Pour satisfaire ces conditions, il suffit que l'affectation au traitement ne soit pas corrélée aux distributions du résultat potentiel. La randomisation permet d'assurer la non-corrélation entre Z et Y. Lorsqu'il s'agit de données observationnelles, on peut détecter deux sources de biais dans les estimations des effets causaux faites à partir d'études observationnelles : le biais de sélection (lorsque les deux groupes sont différents à la base.), et l'hétérogénéité de l'effet du traitement (les deux groupes réagissent différemment au traitement).

L'analyse par score de propension (Labarère et al. (2008)) est une méthode d'ajustement qui consiste à calculer la probabilité conditionnelle d'un individu de recevoir le traitement d'intérêt, connaissant ses caractéristiques. C'est l'une des méthodes les plus utilisées pour étudier les effets causaux dans le cadre des études observationnelles et pour corriger pour l'effet des facteurs confondants (Li (2011)).

L'objectif de cette analyse est d'estimer l'effet causal d'une intervention ou d'un traitement particulier sur une certaine variable d'issue. En effet, l'appariement de chaque sujet traité à un sujet non traité ayant un score de propension identique ou proche aboutit à la constitution de deux groupes de sujets ayant des caractéristiques comparables et entre lesquels le critère de jugement peut être comparé. L'analyse par score de propension est une méthode d'ajustement a posteriori pertinente pour les situations cliniques où les essais randomisés sont difficilement réalisables du fait de contraintes économiques ou organisationnelles. Elle ne permet cependant qu'un ajustement sur les caractéristiques mesurées et une analyse de sensibilité des résultats peut être nécessaire.

Rosenbaum and Rubin (1983) montrent dans leur article qu'apparier ou stratifier des unités d'observation traitées et non-traitées avec le score de propension permet d'éliminer le biais dû aux différences de caractéristiques observables (variables confondantes). En effet, le score de propension permet d'équilibrer les distributions des covariables. Pour chaque valeur du score de propension, la distribution des covariables X dans le groupe traité devrait être identique à celle du groupe de contrôle, ce qui permet de les comparer en ayant ainsi ajusté les différences de composition. En pratique, le score de propension doit être estimé en fonction des covariables observées.

Le score de propension e_i peut être estimé par une régression logistique de Z_i sur X_i (Li (2011)) :

$$e_i = e(X_i, \beta) = Pr(Z_i = 1|X_i) = \frac{\exp\{X_i^T * \beta\}}{1 + \exp\{X_i^T * \beta\}} \tag{1.2}$$

On réfère au score de propension par e_i $(0 < e_i < 1)$ qui correspond à la probabilité qu'un individu soit exposé au traitement tenant compte des facteurs confondants représentés par X. Le score de propension ne peut pas prendre la valeur 0 ou 1, car ceci veut dire que l'individu n'a pas le potentiel d'être assigné aux deux traitements, et qu'on n'a aucune information sur

un des deux résultats potentiels.

Dans la sous-section suivante, nous présentons le score de propension avec les poids de pondération. Les poids de pondération sont utilisés dans le cas ou le score de propension (i.e la probabilité qu'une unité d'observation reçoive le traitement) est proche de 1 ou de 0. En effet, lorsque des observations ont des scores de propension proches de 0 ou de 1, cela veut dire que ces observations ne sont pas comparables et n'ont pas des potentiels comparables d'exposition. La pondération permet ainsi de donner plus de poids aux observations qui sont traitées mais qui ont un score de propension proche de 1 ou qui ne sont pas traitées mais qui ont un score de propension proche de 0 et de minimiser le biais dû au potentiel d'exposition.

1.3.3 Score de propension avec les poids de pondération

Le poids apparié de pondération (matching weight) est défini par :

$$W_i = \frac{\min(1 - e_i, e_i)}{Z_i * e_i + (1 - Z_i) * (1 - e_i)} \tag{1.3}$$

Le poids d'appariement est une modification du poids de probabilité inverse avec $\min(1-e_i, e_i)$, ceci contraint ainsi le poids à ne pas être excessivement grand lorsque e_i s'approche de 0 ou 1, stabilisant ainsi l'estimateur de l'effet causal, et améliore son efficacité (Li (2011)). La méthode de pondération d'appariement permet de pondérer les groupes exposés et non exposés afin de créer une population appariée par score de propension.

Une fois qu'on obtient les poids pour chaque observation à l'étude, on va s'en servir pour recalculer l'ATE qui peut être écrit comme suit :

$$\hat{\overline{\delta}}_{MW} = \hat{\overline{Y}}_{1\_MW} - \hat{\overline{Y}}_{0\_MW} \tag{1.4}$$

où :

- $\overline{Y}_{1\_MW} = \frac{\sum_{i \in E} W_i * Y_i}{\sum_{i \in E} W_i}$: la moyenne de la variable réponse Y dans le groupe exposé pondérée par les poids issus du modèle du score de propension et calculés dans l'équation 1.3,
- $\overline{Y}_{0\_MW} = \frac{\sum_{i \in N} W_i * Y_i}{\sum_{i \in N} W_i}$: la moyenne de la variable réponse Y dans le groupe non exposé pondérée par les poids issus du modèle du score de propension et calculés dans l'équation 1.3.

En pratique, le package PSW (Mao and Li (2018)) nous fournit des méthodes de pondération du score de propension pour contrôler pour les facteurs confondants dans l'inférence causale avec les traitements dichotomiques et les résultats continus et binaires.

Chapitre 2

Analyse de données

Après avoir introduit les différentes techniques statistiques employées dans ce projet, on tâchera, dans la présente section, de présenter le jeu de données utilisé, ainsi que toutes les analyses statistiques qui ont été réalisées dans l'objectif de quantifier l'impact des investissements en marketing sur le nombre de soumissions reçues. On veut ainsi définir une méthodologie permettant d'estimer l'effet causal de chaque véhicule publicitaire adopté par la compagnie sur le nombre de soumissions qu'elle reçoit. On va élaborer des modèles qui permettent de mesurer l'impact de l'investissement dans un véhicule publicitaire X sur le nombre de soumissions reçues par la compagnie ASSURE dans une période donnée.

2.1 Méthodologie statistique

En vue de construire nos modèles statistiques, nous adoptons une méthodologie qu'on peut résumer comme suit :

- Définition de la question de recherche et spécification de la problématique.
- Statistique descriptive et exploratoire en vue de différencier les particularités des données. (Mesures de corrélation, statistiques descriptives, visualisation de données, analyse en composantes principales, etc).
- Choix de la meilleure standardisation pour les variables d'intérêt : le nombre de soumissions ainsi que les dépenses dans les différents véhicules publicitaires.
- Analyse de regroupement : En vue de définir un groupe homogène de RTA (région de tri d'acheminement – les trois premiers caractères dans un code postal) en termes de caractéristiques sociodémographiques.
- Construction des modèles statistiques pour modéliser l'impact de l'investissement dans un véhicule publicitaire sur le nombre de soumissions reçues en tenant compte de variables :
    - Sélection de variables.

    - Ajustement des modèles linéaires
    - Ajustement de modèles selon une approche causale :
        - Modèle avec le score de propension
        - Modèle avec les poids d'appariement (Matching Weights).
- Comparaison des modèles

Dans ce qui suit, nous présentons la base de données à notre disposition, la problématique que nous traitons, et nous détaillons chaque point de la méthodologie adoptée.

2.2 Objectif de l'étude

Notre objectif est de cerner le mieux possible l'effet de l'investissement dans un certain média sur le nombre de soumissions reçues par la compagnie d'assurance ASSURE. Nous voulons ainsi pouvoir mesurer le nombre de soumissions générées par chaque dollar supplémentaire investi dans la publicité via ce média.

2.3 Présentation des données

2.3.1 Provenance des données

En vue de conduire notre analyse statistique, nous disposons de deux bases de données. La première base de données est en provenance de la compagnie ASSURE. L'unité d'observation est la RTA. Les RTA représentent un découpage géographique du marché de ASSURE en 417 unités. Nous disposons de données hebdomadaires concernant les dépenses publicitaires et des soumissions recueillies par RTA pour 3 années, de 2015 à 2017. Nous disposons de 156 semaines et de 21 variables, pour un total de 3276 observations par RTA. Pour avoir la signification de chaque variable, nous référons à l'annexe A.1.

La deuxieme base de données est en provenance de Statistique Canada (STATCAN). Elle décrit les caractéristiques sociodémographiques de chaque RTA. On dispose de 33 variables pour 417 RTA, et donc un total de 13761 observations. Ces variables décrivent la population de chaque RTA selon l'âge, le statut marital, le statut de citoyenneté, le niveau de vie, le statut du logement, le niveau d'éducation, et la langue maternelle. Pour avoir la signification de chaque variable, nous référons à l'annexe A.2

2.3.2 Variables d'intérêt

Dans la base de données ASSURE, on s'intéresse à la variable SOUM, qui décrit le nombre de soumissions reçues par semaine dans chaque RTA. On s'intéresse aux variables d'investissement, qui représentent les montants dépensés en efforts marketing en coûts et en volume dans

les 11 véhicules publicitaires que cette compagnie adopte pour commercialiser ses produits d'assurance. Par souci de confidentialité, on réfère à ces véhicules publicitaires par les lettres A à K.

Dans la base de données STATCAN, on s'intéresse à toute variable qui permet de bien cerner la condition sociodémographique des différentes RTA.

2.4 Présentation de la problématique

Afin d'illustrer notre problématique, on utilise l'exemple suivant décrivant le nombre de soumissions reçues dans deux RTA pendant la première semaine de 2017, ainsi que les montants investis respectivement dans les médias A, B et C. Pour des raisons de confidentialité, toutes les variables d'investissement ont été divisées par un nombre arbitraire :

TABLE 2.1 – Exemple de tableau de données

Semaine	Année	RTA	SOUM	A	B	C
1	2017	H1A	23	187.96	8.38	0
1	2017	H1B	17	139.38	8.38	0

Dans la première semaine de l'année 2017, la compagnie ASSURE a dépensé 187.96 dollars et 139.38 dollars dans le véhicule publicitaire A dans les RTA H1A et H1B, respectivement. La compagnie reçoit 23 soumissions depuis H1A et 17 soumissions depuis H1B. En supposant que toutes autres variables d'investissement sont égales par ailleurs et que les RTA H1A et H1B ont exactement la même composition sociodémographique, on pourrait conclure que l'effet causal de l'investissement en A, et par conséquent le retour sur cet investissement est de 6 soumissions supplémentaires pour une dépense additionnelle de 48 dollars. Néanmoins, dans la réalité, les autres variables d'investissement ne sont pas égales, et les RTA ne sont pas parfaitement semblables. Il s'avère donc plus difficile de cerner l'effet causal de chaque média sur les soumissions, vu la présence de plusieurs variables confondantes, et de corrélation entre les différents traitements publicitaires.

Dans les sections qui suivent, nous présentons les différentes analyses conduites dans le cadre de cette étude. Nous commençons d'abord par présenter la base de données de notre application.

2.4.1 Restriction de l'étude

Disposant d'un grand nombre de véhicules publicitaires et étant donné la présence d'une forte saisonnalité, il serait difficile d'étudier l'effet causal de tous les médias sur les trois années à la fois. Ainsi, dans le présent projet, on concentre notre étude sur un seul véhicule publicitaire,

qu'on appelle « A ». On veut ainsi quantifier l'effet causal du média A sur le nombre de soumissions que la compagnie reçoit. On restreint notre étude sur une période précise de l'année 2017, à savoir les semaines 8 à 16. Ce sont les semaines de l'année où l'on observe la plus grande évolution du nombre de soumissions. On considère qu'il serait intéressant de mesurer l'effet causal des coûts qu'engendre le média « A » en période intensive de ventes et d'investissement. La figure 2.2 ci-dessous montre l'évolution de la moyenne non standardisée des soumissions dans les 417 RTA entre 2015 et 2017.

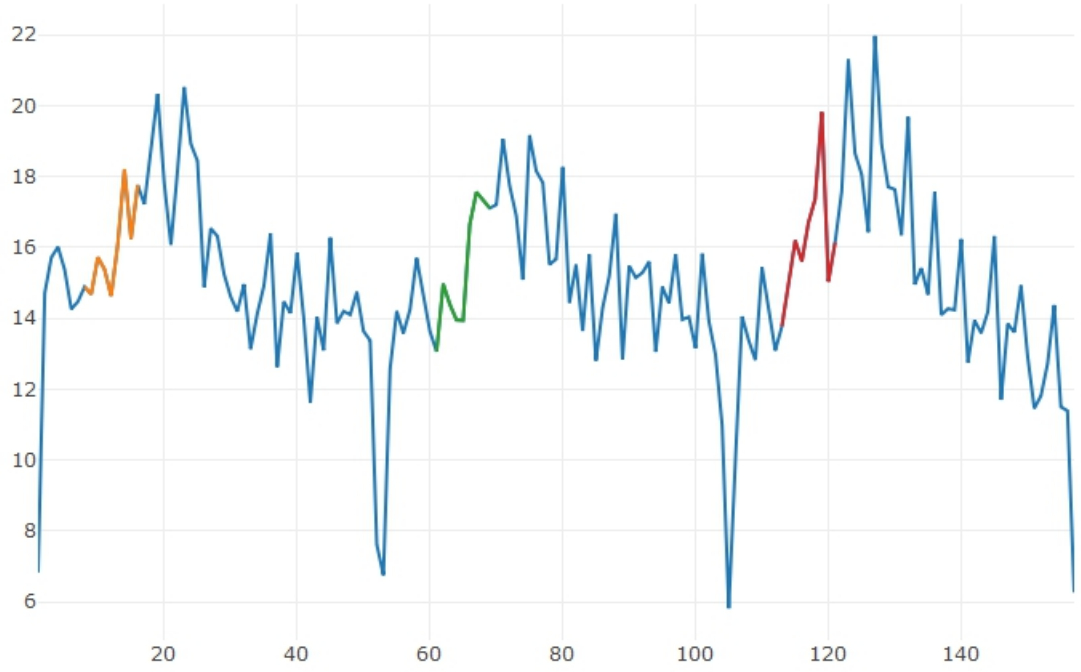


FIGURE 2.1 – Évolution du nombre de soumissions reçues entre 2015 et 2017 en distinguant les semaines 8 à 16 de chaque année.

2.5 Standardisation

La grande disparité entre les différentes RTA sur le plan sociodémographique nous incite à prendre en considération des facteurs confondants tenant en compte cet aspect dans l'analyse. Une première différence serait la différence sur le plan démographique. Par exemple, si dans une première RTA X avec une population de 37 000 habitants, on reçoit 35 soumissions dans une semaine donnée, et dans une deuxième RTA Y avec une population de 2 800 habitants, on reçoit également 35 soumissions, il ne serait pas judicieux de comparer ces deux RTA sur la base des soumissions ni sur la base des montants investis sans tenir compte de la taille de la population, d'où la nécessité de ramener toutes les localités à une même échelle. Ainsi, il nous faut standardiser toutes les variables d'investissement dans les véhicules publicitaires ainsi que

le nombre de soumissions par une quantité mesurant la taille de la clientèle potentielle de la RTA.

Deux variables de standardisation sont disponibles : GEO_POP, la variable décrivant le nombre d'habitants dans la RTA fournie par Statistique Canada, et GEO_NBF, la variable décrivant le nombre de dossiers de crédits ouverts dans la RTA fournie par la compagnie ASSURE. Nous désirons choisir la meilleure variable de standardisation. Dans la figure 2.2, on illustre les niveaux des deux variables de standardisation dans les RTA.

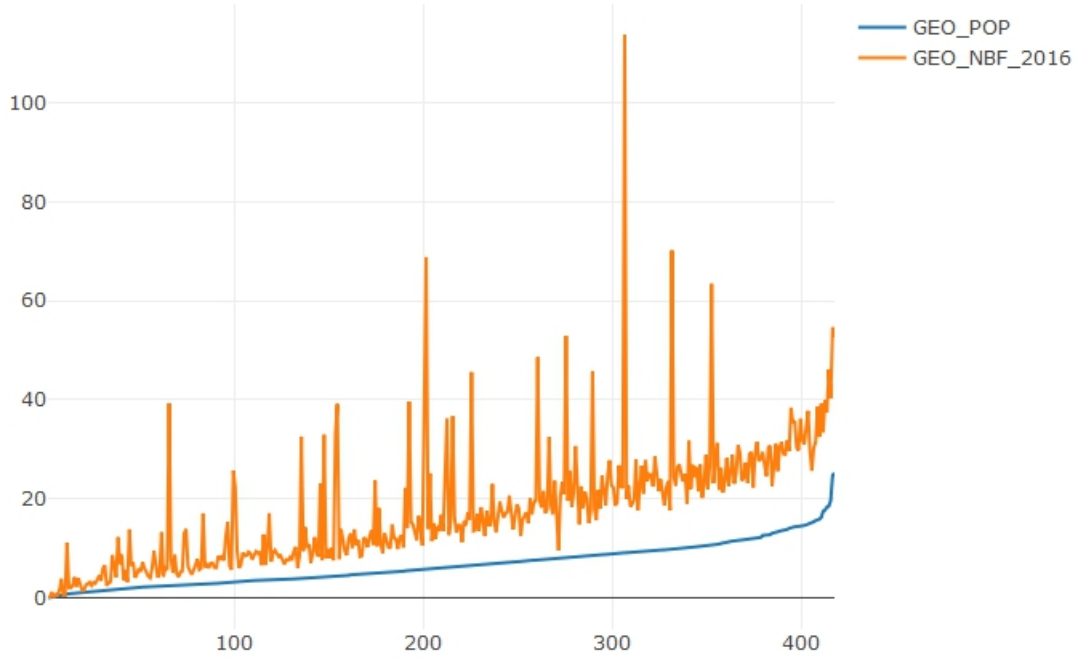
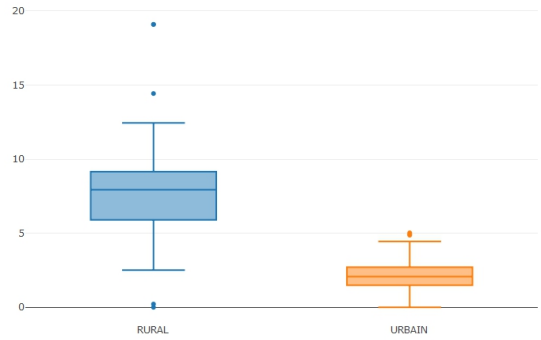


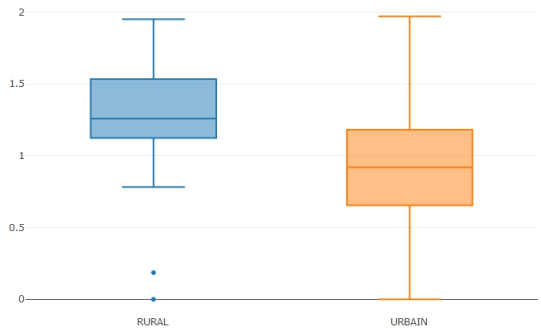
FIGURE 2.2 – La différences entre les deux variables de stardisation par RTA. L'axe des abscisse représente l'indice des 417 RTA classées par population croissante. La population et le nombre de dossiers de crédits sont donnés en milliers.

Dans un premier temps, on standardise intuitivement la base de données par la population dans la RTA ; c'est-à-dire qu'on divise toutes les variables d'intérêt par la population. Les RTA deviennent effectivement comparables à un certain point. Sauf qu'on soulève un autre point : la disparité entre les régions urbaines et rurales. On remarque que le nombre standardisé de soumissions reçues dans la zone urbaine est grandement inférieur à celui reçu dans la zone rurale. Néanmoins, la standardisation par la population ne permet pas d'absorber cette disparité. La figure 2.3a illustre le nombre de soumissions reçues dans chaque RTA divisé par la population dans la RTA (le nombre de millier d'habitants dans la RTA) , en distinguant le milieu rural du milieu urbain.

Cherchant une meilleure variable de standardisation, il s'est avéré que le nombre de dossiers de crédits ouverts dans la RTA et un meilleur indicateur de la taille de la clientèle potentielle de cette dernière dans le contexte de notre étude. Également les mesures de corrélation suggèrent que l'on choisisse le nombre de dossiers de crédits ouverts dans la RTA puisque la corrélation entre le nombre de soumissions et la population dans la RTA est de 0.47 et la corrélation entre le nombre de soumissions et le nombre de dossiers de crédits dans la RTA est de 0.82.



(a) Boîtes à moustache représentant le nombre de soumissions normalisé par **le nombre des habitants dans la RTA** dans le milieu **urbain** et **rural**.



(b) Boîtes à moustache représentant le nombre de soumissions normalisé par **le nombre de dossiers de crédits ouverts dans la la RTA** dans le milieu **urbain** et **rural**.

FIGURE 2.3 – Illustration de la différence entre les deux variables possibles de standardisation quant à la disparité urbain/rural.

À partir de ce point, toutes les variables d'investissement (nombre de soumissions, et tous les montants investis dans les différents véhicules publicitaires) qu'on utilise sont standardisées par le nombre de dossiers de crédit ouverts dans la RTA. Toutes les analyses qui viennent seront conduites sur les données d'investissement (nombre de soumissions, et les montants investis dans les différents véhicules publicitaires) standardisées par le nombre de dossiers de

crédits ouverts dans la RTA.

2.6 Analyse exploratoire

Notre question de recherche préalablement spécifiée peut être formulée ainsi : « Sur la période définie – entre les semaines 8 et 16 – quel est l'effet du média « A » sur le nombre de soumissions reçues par la compagnie d'assurance en question ? ».

Pour traiter cette question, dans un monde idéal, on aurait fait une expérience randomisée, où on prendrait deux groupes de RTA extrêmement similaires socio-démographiquement qui recevraient les mêmes niveaux d'investissement partout, sauf dans le média « A » : le groupe 1 ne recevra aucun investissement, et le groupe 2 recevrait un investissement. Ainsi, la différence observée au niveau du nombre de soumissions reçues dans chaque groupe serait due à l'investissement en « A » et on pourrait facilement quantifier combien de soumissions rapporte chaque dollar supplémentaire investi en « A ».

Dans le contexte de notre étude, comme il s'agit de données observationnelles, notre tâche primaire consiste en créer une pseudo-expérience randomisée à l'aide des données dont on dispose. On doit ainsi définir deux groupes de RTA qui ont reçu deux niveaux différents de traitement publicitaire.

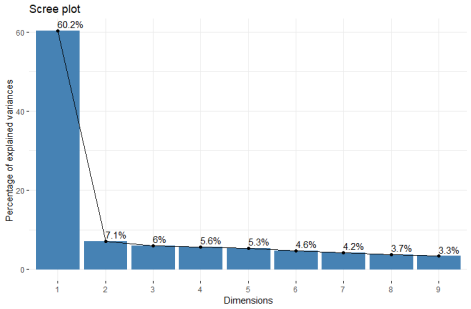
Dans ce qui suit, on opte pour une analyse en composantes principales afin d'explorer les données d'abord, et de définir les deux groupes de RTA ensuite.

2.6.1 Analyse en composantes principales

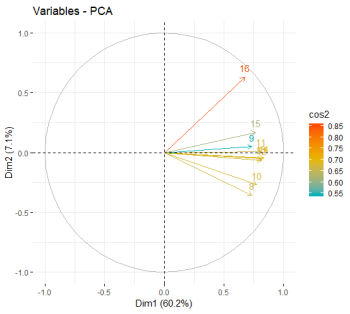
L'intuition derrière notre recours à l'ACP est de définir deux groupes de RTA qui ont reçu deux niveaux d'investissements différents : autrement dit, on veut avoir un groupe High où l'on investit beaucoup, et un groupe Low où l'on investit très peu. De cette manière, il y aura espoir de cerner l'effet causal de cet investissement sur le nombre de soumissions reçues.

Notre base de données contient les montants investis dans le média A dans 417 RTA et pendant 9 semaines. Comme nous cherchons à cerner le comportement général de l'investissement dans le média désigné, nous nous intéressons aux tendances à grande échelle, et non aux RTA qui reçoivent un traitement particulier. Nous supprimons ainsi les observations aberrantes, à savoir les RTA qui reçoivent un très grand investissement, avant de procéder à l'ACP. Notre analyse en composantes principales indique que 68% de la variabilité totale des données est expliquée par les deux premières dimensions.

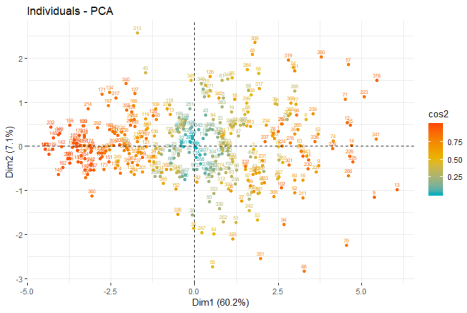
Dans la figure 2.4, on visualise les résultats issus de l'analyse en composantes principales :



(a) Pourcentage de variabilité expliquée par chaque composante principale.



(b) Graphique de corrélation entre les semaines et les deux premières composantes principales.



(c) Graphique des projections des observations/RTA sur les deux premières composantes principales.

FIGURE 2.4 – Illustration des graphiques résultants de l'ACP appliquée aux montants investis dans le média A entre les semaines 9 et 16 dans 417 RTA.

La figure 2.4a rapporte le pourcentage de variabilité expliqué par chaque composante principale, on constate que les deux premières dimensions expliquent 67.1% de la variabilité totale. Le figure 2.4b rapporte les résultats concernant les neuf semaines. En effet, elle représente la corrélation entre les semaines et les deux premières composantes principales. La figure 2.4c rapporte la projection de chacune des observations sur les deux premières dimensions. La couleur des points représente la qualité de la représentation. On constate que la première dimension représente un effet « taille de l'investissement », elle distingue les RTA où on dépense beaucoup (celles situées à l'extrême droite dans le graphique 2.4b) de celles où on dépense peu (celles situées à l'extrême gauche du graphique 2.4b). Et la deuxième représente « la stratégie d'investissement ». Là encore ; on peut faire 2 groupes différents selon la deuxième composante et comparer les variables dépendantes dans ces deux groupes. Si on représente les deux groupes sur une carte géographique selon la coordonnée sur la première dimension représentant l'effet taille, on obtient le graphique ci-dessous ; en rouge les localités ou on dépense beaucoup, et en mauve, celles où on dépense peu.

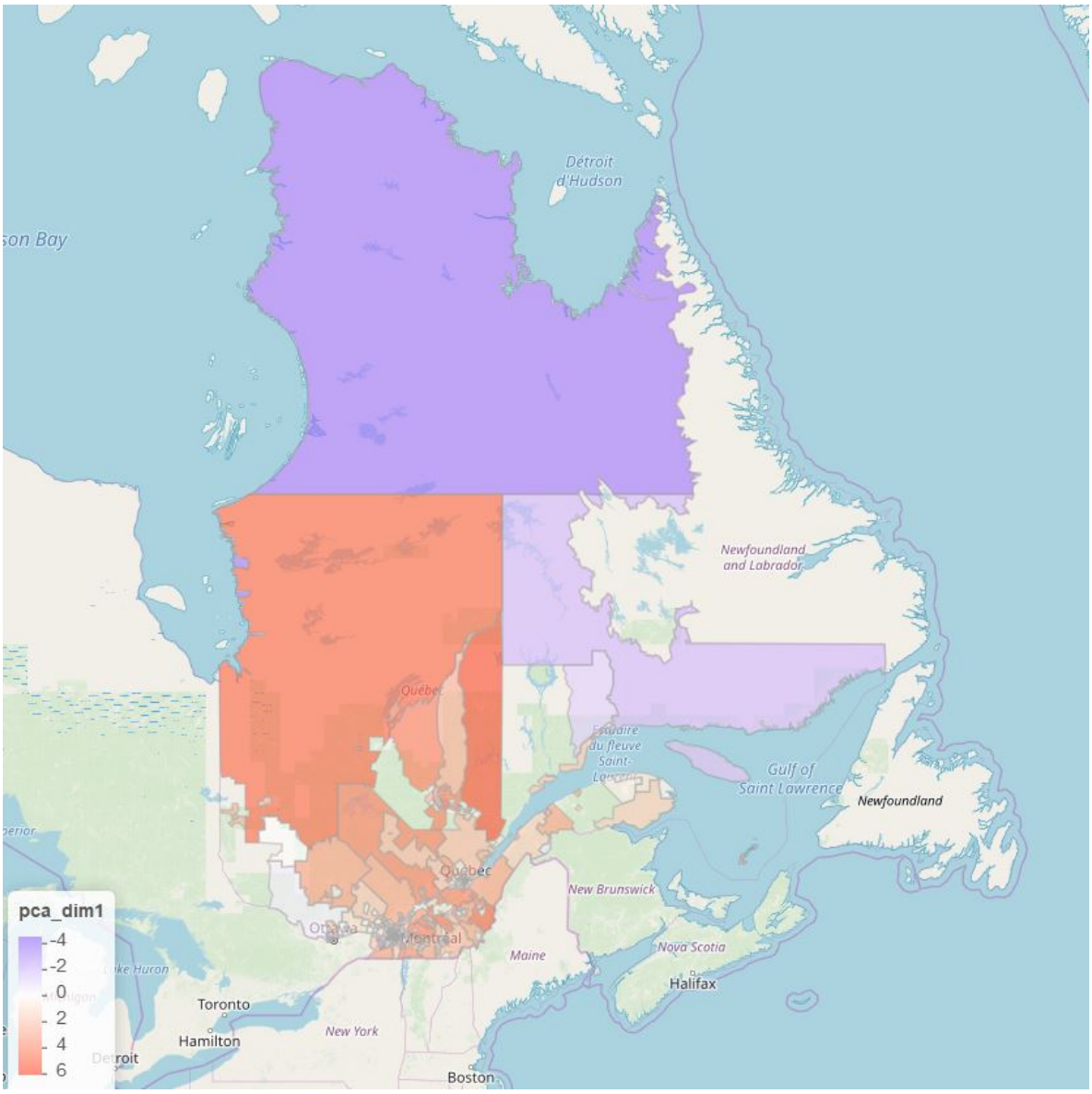


FIGURE 2.5 – Représentation de la taille d'investissement dans chaque RTA selon la première dimension de l'ACP sur une carte géographique.

En comparant le classement des RTA selon la coordonnée sur la première dimension, on obtient quasiment le même ordre que donnerait un classement des RTA selon l'investissement moyen sur les 9 semaines d'intérêt. Ainsi, le passage par l'analyse en composantes principales s'avère facultatif. On privilégie l'utilisation de la moyenne des investissements directement sans passer par l'ACP. Pour avoir les 2 groupes qui ont subit des traitements publicitaires différents, on peut simplement classer toutes les RTA par montants moyens investis dans le média « A » par ordre croissant, et on garde les 100 localités ayant reçu le plus d'efforts publicitaires contre les 100 localités ayant reçu le moins d'efforts publicitaires.

Toutefois, les disparités au niveau des autres investissements reçus par les deux groupes de RTA et leurs différences au niveau socio-démographique nous empêchent de conclure par rapport à la causalité - investissement en « A » => soumissions. Il faut prendre en considération la présence de multicollinéarité et de multiples variables confondantes.

À ce stade, on se focalise sur les localités et leurs particularités sociales et démographiques. La finalité est de trouver un groupe de RTA similaires et homogènes au niveau des propriétés socio-démographiques, et de définir à partir de ce groupe restreint, deux niveaux de traitement publicitaire (0 = Low, et 1= High).

2.7 Analyse de regroupement : identification d'un groupe de RTA ayant des caractéristiques socio-démographiques homogènes

L'objectif de la classification est de trouver des groupes de RTA homogènes sur le plan socio-démographique et sur le niveau d'investissement dans tout autre média que « A ».

On dispose d'une quarantaine de variables de classification au total, des variables d'investissement (Voir A.1), et les variables socio-démographiques (Voir A.2). Comme on ne peut pas les employer toutes dans cette classification car il serait impossible de trouver des RTA similaires au niveau de toutes ces variables, on choisit de classifier les RTA en se basant sur les variables qui influencent le plus l'investissement dans le média A. Pour ce faire, on ajuste un modèle de régression avec, comme variable réponse, l'investissement dans le média A, et comme variables indépendantes tous les facteurs confondants. Cette opération permet d'identifier 4 variables qui expliquent bien les investissements dans le média A, à savoir : Média_B, P_marie, P_immigr, P_mater_fr.

L'analyse de regroupement (Voir la section 1.1.2) nous suggère d'opter pour la formation de 5 groupes à former à partir de nos données. Les résultats de cette classification sont résumés au tableau 2.2 :

TABLE 2.2 – Résultats de la classification K-moyennes

Groupe	Taille	Variabilité intra-groupe
1	69	1.90
2	151	0.68
3	29	2.35
4	91	1.86
5	77	1.07

Quel groupe choisir ? Puisque nous cherchons le groupe le plus homogène, un critère de choix pertinent serait de prendre le groupe qui a la variance intra-groupe minimale. Ainsi, on retient le deuxième groupe qui contient 151 RTA. Finalement, nous divisons ces 151 RTA en trois sous-groupes à savoir : « Low », contenant les 50 RTA témoignant des efforts publicitaires les plus faibles en A, « Middle », contenant 51 RTA, et « High » contenant 50 RTA témoignant de grands efforts publicitaires.

À partir de ce moment, toute l'analyse statistique qui suit – consistant en l'ajustement des modèles servant à estimer l'effet de l'investissement dans le média « A » sur le nombre de soumissions - portera sur les 100 RTA des groupes « High » et « Low ». On crée une nouvelle variable qu'on nomme « Treated » et qui prend la valeur 1 si la RTA appartient au groupe « High », et 0 si la RTA appartient au groupe « Low ». Le tableau 2.3 récapitule les différences en moyenne entre les deux groupes en termes de soumissions, de dépenses (les soumissions et les dépenses sont exprimées par semaine et par 1000 dossiers de crédit ouverts) dans le média A, et dans le média B :

TABLE 2.3 – Tableau descriptif des moyennes des variables d'intérêt entre les groupes High et Low accompagnées des déviations standards entre parenthèses.

Groupe	RTA	Traitement 0/1	Soumissions	Média A	Média B
High	50	1	1.24(0.0280)	9.28(0.13)	2.20(0.0356)
Low	50	0	0.987(0.0268)	5.72(0.0916)	1.99(0.0329)

2.8 Comparaison entre les deux groupes High et Low : Ajustement des modèles de régression et les résultats

Dans cette partie, nous présentons dans une optique comparative, trois modèles de régression ajustés sur la base des 100 RTA des groupes "High" et "Low" avec une sélection de variables pour quantifier l'effet du Média A sur le nombre de soumissions reçues. En effet, on procède à une sélection stepwise (pas-à-pas) des variables indépendantes pour chacun des trois modèles ajustés. Les variables explicatives retenues pour chaque modèle sont données à l'annexe A.5.

- Le modèle 1 : Modèle de régression linéaire multiple avec la variable Média_A en continu :

Dans un premier temps, on ajuste un modèle linéaire avec le nombre de soumissions dans la RTA comme variable indépendante, et comme variables indépendantes, les dépenses dans le média « A » en continu, à savoir la moyenne des dépenses dans le média « A » sur 9 semaines, et les covariables d'investissement et sociodémographiques (A.5). On s'intéresse principalement au coefficient associé au média « A ». Le coefficient de régression correspondant au média « A »

est de 0.03549. On conclut que si l'on augmente l'investissement dans le média « A » d'une unité de mesure, le nombre de soumissions va augmenter de 3.54 soumissions par 100 dossiers de crédit ouverts, sachant que toute autre variable est égale par ailleurs.

- Le modèle 2 : Modèle de régression linéaire multiple avec la variable Média_A en catégorielle :

Dans un deuxième temps, on ajuste un modèle linéaire avec le nombre de soumissions dans la RTA comme variable indépendante, et comme variables indépendantes, les dépenses dans le média « A » en catégorielle : High VS Low. On s'intéresse au coefficient lié à la variable GROUP. Le coefficient de régression de 0.1261 correspondant à la différence au niveau de l'espérance du nombre de soumissions par dossier de crédits si on passe du groupe Low au groupe High.

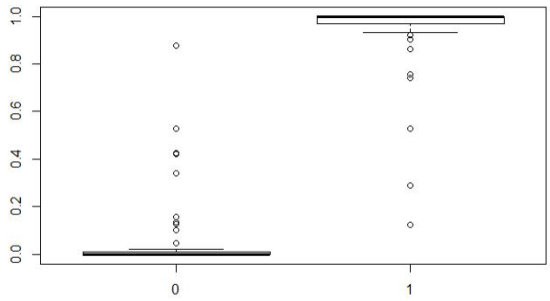
- Le modèle 3 : Modèle basé sur le score de propension et les poids d'appariement (Matching weights) :

Dans un troisième temps, on ajuste un modèle de régression linéaire en utilisant des pondérations pour les RTA. Ces pondérations sont issues du modèle du score de propension.

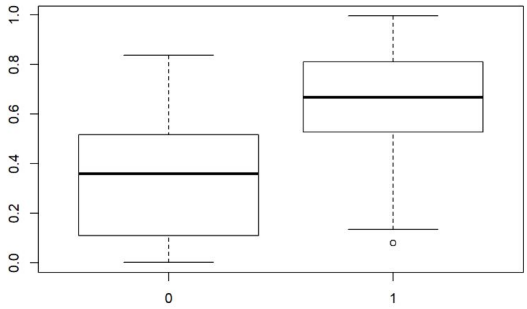
— Construction du score de propension :

Basé sur les 100 RTA, on construit un modèle avec les variables socio-démographiques et toutes les variables d'investissement autres que A pour prédire si la RTA sera traitée et recevra un investissement important (High) ou faible (Low) dans le média A. Pour cela, nous avons ajusté un modèle de régression logistique avec une variable dépendante $Treated$ qui prend la valeur 0 si la RTA appartient au groupe "Low" et 1, sinon, et des variables confondantes liées à la fois aux nombres de soumissions et aux efforts publicitaires fournis en A. Suite à une sélection de variable suivant la méthode descendante, le modèle a conservé les 5 variables suivantes : Média_B, Média_C, P_POP_4564, p_const2011_2016, CRP3_SCORE_STD, et présente un pouvoir prédictif représenté par une aire sous la courbe ROC de 81.36%. Sur la base du score de propension, on affecte à chaque observation un poids de pondération. Ces poids de pondération seront utilisés par la suite dans le modèle de régression (modèle 3) modélisant le nombre de soumissions par les variables d'investissement et les variables socio-démographiques.

La figure 2.6 montre la distribution du score de propension dans les groupes Low (0) et High (1) avant l'analyse de regroupement dans la figure 2.6a et après la création du groupe homogène de RTA dans la figure 2.6b. Ceci montre que l'analyse de regroupement a permis de rendre les groupes Low et High plus homogènes, et plus comparables au niveau de leur potentiel d'exposition.



(a) Score de propension avant l'identification du groupe de RTA homogènes.



(b) Score de propension après l'analyse de classification K-moyenne et la sélection de 100 RTA homogènes.

FIGURE 2.6 – Illustration de la distribution du score de propension dans les deux groupes High et Low. Le premier graphique représente la distribution du score de propension sur la base des groupes High et Low définis à partir de 417 RTA classés par ordre croissant de la taille d'investissement. Le deuxième graphique représente la distribution du score de propension des groupes High et Low définis à partir du regroupement retenu de la classification K-moyennes.

— Modèle basé sur les poids d'appariement :

Suite au calcul des scores de propension relatifs aux 100 observations à notre disposition, on affecte à chaque observation un poids représentant la probabilité que cette RTA soit traitée tel que décrit dans l'équation 1.2. Ces poids vont par la suite être pris en compte dans le modèle de régression linéaire multiple modélisant le nombre de soumissions. D'après cette modélisation (A.5), le coefficient de régression correspondant à 0.097 indique la différence de l'espérance du nombre de soumissions par dossier de crédit si on passe du groupe Low au groupe High.

Dans le tableau 2.4, on présente le récapitulatif des résultats des trois modèles.

Modèle	Traitement	Coefficient	Erreur-type	Valeur P
Modèle 1	Média_A	0.035	0.011	0.0017
Modèle 2	Group	0.126	0.041	0.003
Modèle 3	Group	0.097	NA	NA

TABLE 2.4 – Récapitulatif des modèles : la variable Traitement indique la variable des dépenses dans le média_A dont on veut déterminer l'effet causal sur le nombre de soumissions par dossier de crédit , cette variable est nommée Média_A en traitement continu, et Group en traitement binaire (Group = Low / High). Le tableau présente le coefficient de régression β associé à la variable traitement, ainsi que l'erreur-type et la valeur P qui y correspondent. Pour le modèle 3, on note que le calcul de l'erreur standard est plus difficile à obtenir pour la méthode de régression avec pondération et n'a pas été effectué dans le cadre de ce projet.

Dans le premier modèle, on traite les dépenses dans le média_A en continu, et on corrige pour toutes les variables confondantes. L'effet des dépenses dans le média_A est très significatif et on estime une augmentation de 0.035 au niveau des soumissions par dossier de crédit si on augmente les dépenses dans le média_A d'une unité.

Dans le deuxième modèle, on mesure l'effet du traitement en considérant les dépenses comme une variable catégorielle, en divisant les RTA en deux groupes traité et non traité et en estimant l'effet du traitement binaire (Low vs High) sur le nombre de soumissions par dossier de crédit. L'effet du traitement est très significatif dans ce modèle également, et l'effet inféré (0.126) représente l'effet sur les soumissions du passage du groupe contrôle au groupe traité. On remarque que l'effet du média_A est toujours positif. Dans une optique confirmatoire, si on multiplie la différence observée au niveau des dépenses entre les traités et les non traités (= 3.56) par le coefficient estimé par le modèle 1 (= 0.0354), on se retrouve avec 0.1260, ce qui indique que le modèle 1 et le modèle 2 suggèrent à peu près le même effet des dépenses publicitaires sur le nombre de soumissions.

Pour le troisième modèle, on procède exactement de la même manière que pour le modèle 2, mais cette fois-ci, en assignant des poids aux différentes RTA en se basant sur le modèle du score de propension. En affectant des poids aux observations, on accorde plus d'importance aux RTA qui ont un grand potentiel d'être traité mais qui ne le sont pas, et aux observations qui n'ont pas le potentiel d'être traitées, mais qui le sont. Cette méthode permet ainsi de diminuer le biais lié à l'échantillonnage. Le troisième modèle suggère une différence de 0.097 au niveau du nombre de soumissions par dossier de crédit si on passe du groupe Low au groupe High. Cette différence est inférieure à celle du modèle 2, ceci est du à la pondération qui a permis d'accorder plus d'importance à des RTA qui n'étaient pas assez bien représentées dans la première et deuxième modélisations. Cet estimé de l'effet est donc plus robuste que celui des modélisations précédentes. Il serait donc plus conservateur de se fier à cette estimation qui assure une meilleure représentativité des 417 RTA au sein du groupe de 100 RTA sur lequel

l'analyse a été conduite.

Compte tenu des résultats, les trois modèles pointent vers la même direction : on conclut que le média_A a un effet positif sur la variable réponse. On a réussi à converger vers une différence entre 0.09 et 0.12 au niveau des soumissions perçues par ASSURE entre les traités et non traités, ce qui correspond à une hausse d'environ 0.025 à 0.035 de centaine de soumissions reçues par dossier de crédit suite à une hausse d'une unité d'investissement dans le média_A en traitement continu, peu importe l'approche suivie.

Conclusion

Mesurer le retour sur investissement à partir de données observationnelles s'avère une tâche difficile. Il existe des méthodes d'inférence statistique pour quantifier l'effet d'un traitement sur une variable d'intérêt, mais ces méthodes présentent certaines limites quand il s'agit de conclure la causalité. En effet, les données observationnelles présentent de nombreux défis, surtout lorsque des variables non mesurables sont susceptibles d'influencer la variable résultat. Quand il s'agit de données marketing, celles-ci sont souvent sujettes à des variations saisonnières, et sont souvent confondues avec les effets de variables non observables.

Dans le présent essai, notre objectif a été de mesurer l'impact de l'investissement dans un véhicule publicitaire donné sur le nombre de soumissions perçues par la compagnie d'assurance ASSURE. Nous avons commencé par passer en revue les différentes notions statistiques utilisées, nous avons introduit quelques notions d'inférence causale ainsi que le modèle du score de propension. Ensuite, nous avons présenté les différents résultats de notre analyse de données en vue de répondre à la question de recherche et réussir à obtenir une estimation robuste de l'effet de l'investissement sur le nombre de soumissions reçues.

Ce projet nous présente certains défis majeurs. Le premier défi consiste en la présence d'une grande corrélation entre la variable intervention (dépense dans le véhicule publicitaire A) et une deuxième variable investissement (dépenses publicitaires dans le véhicule publicitaire B). Ce défi a été surmonté à travers une analyse de regroupement qui nous a permis de créer un groupe homogène d'observations avec des niveaux similaires au niveau de cette deuxième variable. Un deuxième défi concerne l'existence d'un grand nombre de variables confondantes observées qui affectent à la fois la variable d'intervention et la variable réponse. Ce défi a été surmonté en procédant à une sélection de variables qui permet de définir celles qui influencent le plus la variable d'intervention, et de procéder à une analyse de regroupement des RTA sur le base de ces variables. Ceci nous a permis de définir un groupe de RTA homogène et par conséquent d'identifier l'effet de la variable du traitement sur la variable réponse.

Le point fort de cette analyse consiste en la méthodologie suivie. Dans cette méthodologie, nous avons considéré l'aspect socio-démographique des RTA, aspect important mais qui n'a pas été pris en considération au préalable par ASSURE. L'idée de classifier les RTA selon les variables qui influencent le plus la variable d'investissement, dont on souhaite quantifier

l'effet, nous a permis de définir un groupe homogène et de définir une variable traitement à deux niveaux. Cette méthodologie a permis de réussir à inférer l'effet du traitement sur le nombre de soumissions. Toutefois, des limites demeurent présentes. Une limite de notre analyse réside dans la restriction qui a été faite en vue de rendre l'étude d'un effet causal possible, soit l'élimination de l'effet de la saisonnalité en se focalisant sur une période précise qui a fait que nous n'avons utilisé qu'une petite partie des données. Également, un problème de positivité est survenu lors de la construction du score de propension. En effet, certaines RTA avaient un potentiel négligeable d'être exposées au traitement (ou de ne pas être exposées au traitement), et ces RTA ont été exclues de l'analyse, rendant l'estimation d'un effet causal très difficile.

Pour les travaux futurs, il serait intéressant d'employer l'ensemble de la base de données longitudinale pour faire les estimations au lieu de se limiter à une période spécifique. Également, il serait peut être intéressant d'examiner si une année ou une période précise ne peut pas constituer une période de référence pour une année ultérieure sous une hypothèse d'indépendance temporelle : trouver un groupe de RTA qui subissent des traitements opposés au cours de la même période de deux années différentes peut donner une idée sur l'effet propre au média, puisque la période et la composition socio-démographique des RTA restent les mêmes. Ensuite, suivant la même méthodologie, on peut examiner dans un premier temps, l'effet du média_B (média présentant une grande corrélation aec le média_A), pour ensuite tenter d'estimer l'effet conjoint et la synergie de deux variables investissement (ou plus).

Annexe A

Dictionnaire des variables

A.1 Variables en provenance de la compagnie ASSURE

A.1.1 Variables d'investissement

Variable	Signification
SOUM	Nombre de soumissions moyen reçues dans une RTA par jour
Média_COST_i	Montant investi dans le média i = A à K dans une RTA pendant par jour
Compet	les investissemens (en $) dans le média_C de la compétition dans une RTA par jour

A.1.2 Autres variables relatives au crédit

Variable	Signification
GEO_NBF	total des dossiers de crédit ouverts dans une RTA
ROBP	Score de vol en 2011
BURP	Score d'introduction par effraction en 2011
MOTP	score de vol de véhicule à moteur en 2011
CRP3_SCORE_AVG	1- Risque de retard de paiement 90 jours - valeur moyenne dans le code postal
CRP3_SCORE_STD	1- Risque de retard de paiement 90 jours - écart type dans le code postal
INCP_SCORE_AVG	revenu prédit (avec utilisation du crédit) - valeur moyenne au niveau du code postal
INCP_SCORE_STD	revenu prédit (avec utilisation du crédit) - écart type au niveau du code postal

A.2 Variables en provenance de Statistique Canada

A.2.1 Variables démographiques

Variable	Signification
GEO_POP	total de la population dans une RTA
P_POP_0024	% de la population totale âgée de 0 à 24 ans
P_POP_2544	% de la population totale âgée de 25 à 44 ans
P_POP_4564	% de la population totale âgée de 45 à 64 ans
P_POP_65o	% de la population totale âgée de 65 ans et plus
P_immigr	% de la population totale des ménages privés immigrés
P_mino_visi	% de la population totale des ménages privés appartenant à une minorité visible
P_chom15P	Taux de chômage de la population totale âgée de 15 ans et plus
P_autoch	% de la population totale qui se sont identifiés comme autochtones
Ave_pers_menag	Taille moyenne des ménages
médiane_hh_doll	revenu total médian des ménages en 2015 en dollars

A.2.2 Variables liées au statut familial

Variable	Signification
P_marie	% de la population totale âgée de 15 ans et plus mariée
P_commlaw	% de la population totale âgée de 15 ans et plus vivant en union de fait
P_separated	% De la population âgée de 15 ans et plus séparée et divorcée

A.2.3 Variables liées au logement

Variable	Signification
nb_PO	% de ménages privés selon le mode d'occupation - Propriétaire
nb_LO	% de ménages privés selon le mode d'occupation - Locataire
P_condo	% des logements privés occupés ayant le statut de "copropriété"
P_rangee	% des logements privés du type de maison en rangée
P_unifam	% des logements privés du type de maison individuelle
P_5etaP	% des logements privés d'un appartement, immeuble de cinq étages ou plus, de type
P_movable	% des logements privés d'habitation mobile
P_nonoccupation	% de la population de 15 ans et plus avec occupation "non applicable"
P_const_av1960	% des logements privés par période de construction - avant 1960
P_const2011_2016	% des logements privés selon la période de construction - 2011 à 2016
cout_dwel_LO	Coûts d'habitation mensuels moyens des logements loués en dollars
cout_dwel_PO	Coûts d'habitation mensuels moyens des logements possédés en dollars
médian_dwel_doll	Valeur médiane des logements en dollars

A.2.4 Variables liées au niveau d'éducation

Variable	Signification
P_2564_nodiplo	% de la population âgée de 25 à 64 ans sans certificat, diplôme ou grade
P_2564_collcegep	% de la population âgée de 25 à 64 ans ayant un collège, Certificat ou diplôme d'un cégep ou d'un autre établissement d'enseignement non universitaire
P_2564_uni_bellow	% de la population âgée de 25 à 64 ans ayant un diplôme universitaire ou diplôme inférieur au baccalauréat
P_2564_uni_above	% de la population âgée de 25 à 64 ans ayant un certificat universitaire, diplôme ou grade de baccalauréat ou supérieur

A.2.5 Variables liées à la langue maternelle

Variable	Signification
P_mater_ang	% de la population totale dont la langue maternelle est l'anglais
P_mater_fr	% de la population totale dont la langue maternelle est le français

A.3 Variables de standardisation

Variable	Signification
GEO_POP	total de la population dans une RTA
GEO_NBF	total des dossiers de crédit ouverts dans une RTA

A.4 Nouvelles variables

Variable	Signification
Soum	Le nombre moyen de soumissions standarisées par le nombre de dossiers ouverts de crédits dans la RTA, sur les semaines 8 à 16.
Média_i	La moyenne des montants moyens investis dans le média_i standarisés par le nombre de dossiers de crédits ouverts dans la RTA dans le média i = A à K dans une RTA pendant les semaines 8 à 16
GROUP	Variable indiquant le groupe auquel appartient la RTA selon le niveau d'investissement dans le média A, elle a trois modalités : LOW, HIGH, et MIDDLE.
Treated	Variable dichotomique prenant la valeur 1 si la RTA appartient au groupe HIGH, et 0 si la RTA appartient au groupe LOW.

A.5 Covariables selon le modèle

Modèle	Covariables incluses dans le modèle
Modèle_1	Média_B ; Média_C ; COMPET ; P_POP_0024 ; P_POP_4564 ; P_POP_65o P_PO,P_5etaP ; P_nooccupation ; p_const2011_2016 P_mater_ang ; P_mater_fr ; P_autoch ; P_immigr P_chom15P ; P_2564_collcegep ; median_hh_doll ; cout_dwel_LO ; cout_dwel_PO ; median_dwel_doll ; BURC ; MOTC ; P_separated ; CRP3_SCORE_STD
Modèle_2	Média_B ; Média_C ; Média_D ; COMPET ; P_POP_0024 ; P_POP_4564 ; P_commlaw ; P_condo ; P_nooccupation ; p_const2011_2016 ; P_mater_ang ; P_mater_fr ; P_autoch ; P_immigr ; P_chom15P ; P_2564_nodiplo ; P_2564_collcegep ; P_2564_uni_below ; P_2564_uni_above ; cout_dwel_LO ; cout_dwel_PO ; CRP3_SCORE_STD
Modèle_3	Média_B ; Média_C ; COMPET ; P_POP_0024 ;P_POP_4564 ; P_POP_65o ; P_separated ; P_PO ; P_5etaP ; P_nooccupation ; p_const2011_2016 ; P_mater_ang ; P_mater_fr ; P_autoch ; P_immigr ; P_chom15P ; P_2564_collcegep ; median_hh_doll ; cout_dwel_LO ; cout_dwel_PO ; median_dwel_doll ; BURC ; MOTC ; CRP3_SCORE_STD

Annexe B

Fonctions ACP

- get_eigenvalue : cette fonction permet l'extraction des valeurs propres, variances des composantes principales et la variance cumulée. Les valeurs propres indique la quantité de variance expliquée par chaque composante principale. Les premières composantes correspondent aux directions portant la quantité maximale de variation contenue dans le jeu de données. Une valeur propre supérieure à 1 indique que la composante principale qui y correspond représente plus de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées. Pratiquement on retient les composantes principales qui expliquent 80% de la variabilité totale.

- fviz_eig : cette fonction permet de visualiser le pourcentage de variance expliquée par chaque axe.

- get_pca_var,get_pca_ind : cette fonction permet d'extraire les résultats relatifs aux variables et aux individus respectivement. Cette fonction retourne une liste d'éléments contenant tous les résultats pour les variables d'origine (et pour les individus respectivement) (coordonnées, corrélation entre variables et les axes, correlation entre individus et axes, et contributions)

- fviz_pca_var, fviz_pca_ind : cette fonction permet de visualiser les résultats relatifs aux variables et aux individus respectivement. La corrélation entre une variable et une composante principale (PC) est utilisée comme coordonnées de la variable sur la composante principale. La représentation des variables diffère de celle des observations : les observations sont représentées par leurs projections, mais les variables sont représentées par leurs corrélations (Abdi and Williams 2010).

Annexe C

Code R - Application

C.1 Téléchargement des librairies

```
library(MatchIt)
library(dplyr)
library(plotly)
library(psych)
library(FactoMineR)
library(factoextra)
library(DT)
library(sf)
library(gridExtra)
library(leaflet)
library(data.table)
library(broom)
library(magrittr)
library(tidyverse)
library(ggplot2)
library(PSW)
library(pROC)
```

C.2 ACP

C.2.1 Suppression des données aberrantes

```
data <- Media_A %>%
```

```
filter(`8` < min(boxplot.stats(Media_A$`8`)$out) ) %>%
filter(`9` < min(boxplot.stats(Media_A$`9`)$out) ) %>%
filter(`10` < min(boxplot.stats(Media_A$`10`)$out) ) %>%
filter(`11` < min(boxplot.stats(Media_A$`11`)$out) ) %>%
filter(`12` < min(boxplot.stats(Media_A$`12`)$out) ) %>%
filter(`13` < min(boxplot.stats(Media_A$`13`)$out) ) %>%
filter(`14` < min(boxplot.stats(Media_A$`14`)$out) ) %>%
filter(`15` < min(boxplot.stats(Media_A$`15`)$out) ) %>%
filter(`16` < min(boxplot.stats(Media_A$`16`)$out) )
```

C.2.2 Analyse en composantes principales

```
pca_Media_A <- PCA(
X=data[,4:12] ,
scale.unit = T,
graph      = F
)
```

C.2.3 Résultats de l'ACP

```
pca_eigval <- get_eigenvalue(pca_Media_A)
pca_vars <- get_pca_var(pca_Media_A)
pca_ind <- get_pca_ind(pca_Media_A)
```

C.2.4 Graphiques de l'ACP

```
fviz_eig(pca_Media_A, addlabels = TRUE)
fviz_pca_var(pca_Media_A, col.var = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),labelsize=4)
fviz_pca_ind (pca_Media_A,  col.ind = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
labelsize = 2)
```

C.3 K-Moyennes

```
K_means_vars <-c("MEDIA__B","P_marie","P_immigr","P_mater_fr")

Media_A_kmean <- kmeans(
x       = data_kmeans[,K_means_vars],
centers = 5,
nstart  = 20
)

Media_A_kmean$size  #264
Media_A_kmean$withinss/phone_kmean$size
```

C.4 Calcul des scores de propension

```
m_ps <- glm(formula = treated ~M dia_B +M dia_C+P_POP_4564+
p_const2011_2016+CRP3_SCORE_STD, family = "binomial", data =
data_var_statcan_Kmeans[data_var_statcan_Kmeans$treated %in%
c(1, 0), ])

summary(m_ps)


boxplot(m_ps$fitted~data_var_statcan_Kmeans$treat
[data_var_statcan_Kmeans$treated %in%
c(1,0)]); #Boxplot de la distribution des scores de propension


g <- roc(treated ~ predict(m_ps,type=c("response")),
data = data_var_statcan_Kmeans[data_var_statcan_Kmeans$treated %in%
c(1,0), ]) #calcul de la courbe ROC
```

C.5 Calcul des poids de pondération

```
prs_df_matching_2 <- data.frame(pr_score_2 = round(predict(m_ps,
type = "response"),10),treated = m_ps$model$treated)

prs_df_matching_2$matching_weight <- ifelse(prs_df_matching_2$
```

```
treated == 1,1/prs_df_matching_2$pr_score_2,1/(1
-prs_df_matching_2$pr_score_2))

prs_df_matching_2$GEO_RTA <- data_var_statcan_Kmeans$GEO_RTA[data_
var_statcan_Kmeans$treated %in% c(1,0)]
prs_df_matching_2$A_surveiller <- ifelse(prs_df_matching_2$
matching_weight>2,"***","␣")
```

C.6 Ajustement des modèles

C.6.1 Modèle 1

```
model_1 <- glm(formula = SOUM_MEANS ~ M dia_A + M dia_B + M dia C +
COMPET + P_POP_0024 + P_POP_4564 + P_POP_65o + P_separated +
P_PO + P_5etaP + P_nooccupation + p_const2011_2016 + P_mater_ang +
P_mater_fr + P_autoch + P_immigr + P_chom15P + P_2564_collcegep +
median_hh_doll + cout_dwel_LO + cout_dwel_PO + median_dwel_doll +
BURC + MOTC + CRP3_SCORE_STD, family = "gaussian", data =
data_var_statcan_Kmeans[data_var_statcan_Kmeans$treated %in% c(1, 0), ])

summary(model_1)
```

C.6.2 Modèle 2

```
model_2 <- glm(formula = SOUM_MEANS ~ M dia_B + M dia C + M dia_D+
COMPET + P_POP_0024 + P_POP_4564 + P_commlaw + P_condo + P_nooccupation
+ p_const2011_2016 + P_mater_ang + P_mater_fr + P_autoch + P_immigr +
P_chom15P + P_2564_nodiplo + P_2564_collcegep + P_2564_uni_below +
P_2564_uni_above + cout_dwel_LO + cout_dwel_PO + CRP3_SCORE_STD + GROUP,
family = "gaussian",data = data_var_statcan_Kmeans[data_var_statcan_
Kmeans$treated %in% c(1, 0), ])

summary(model_2)
```

C.6.3 Modèle 3

```
model_3 <- glm(formula = SOUM_MEANS ~ treated + M dia_B + M dia_C + COMPET
 + P_POP_0024 +P_POP_4564 + P_POP_65o + P_separated + P_PO + P_5etaP +
 P_nooccupation + p_const2011_2016 + P_mater_ang +  P_mater_fr + P_autoch +
 P_immigr + P_chom15P + P_2564_collcegep + median_hh_doll + cout_dwel_LO +
 cout_dwel_PO + median_dwel_doll +  BURC + MOTC + CRP3_SCORE_STD, data =
 data, weights = data$matching_weight)

summary(model_3)
```

Bibliographie

Hervé Abdi and Williams Lynne. Principal component analysis. *Wiley interdisciplinary reviews : computational statistics*, 2(4) :433–459, 2010.

Rebecca Barter. Confounding in causal inference : what is it, and what to do about it ? http://www.rebeccabarter.com/blog/2017-07-05-confounding/, 2016.

Bruce Andrew Bruce, Peter. *Practical statistics for data scientists : 50 essential concepts.* " O'Reilly Media, Inc.", 2017.

Ping-Teng Chang and E Stanley Lee. A generalized fuzzy weighted least-squares regression. *Fuzzy Sets and Systems*, 82(3) :289–298, 1996.

Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. NbClust : An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 2014.

Peter Dunn and Smyth Gordon. *Generalized Linear Models With Examples in R.* Springer-Verlag New York, 2018.

Edward W Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *biometrics*, 21 :768–769, 1965.

H. Hotelling. Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology*, 24 :417–441, 1933.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Jorge Jolliffe, Ian T et Cadima. Principal component analysis : a review and recent developments. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 2016.

Alboukadel Kassambara and Fabian Mundt. *factoextra : Extract and Visualize the Results of Multivariate Data Analyses*, 2017. URL https://CRAN.R-project.org/package=factoextra. R package version 1.0.5.

Brian Kulis and Michael I Jordan. Revisiting k-means : New algorithms via bayesian nonparametrics. *arXiv preprint arXiv :1111.0352*, 2011.

Michael H Kutner, Christopher J Nachtsheim, John Neter, William Li, et al. *Applied linear statistical models*, volume 5. McGraw-Hill Irwin Boston, 2005.

José Labarère, Jean-Luc Bosson, Patrice François, and MJ Fine. Propensity score analysis in observational research : application to a study of prophylaxis against venous thromboembolism. *La Revue de medecine interne*, 29(3) :255–258, 2008.

Sébastien Lê, Julie Josse, and François Husson. FactoMineR : A package for multivariate analysis. *Journal of Statistical Software*, 2008.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection : A data perspective. *ACM Computing Surveys*, 50, 01 2016.

Liang Li. Propensity score analysis with matching weights. *arXiv preprint arXiv :1105.2917*, 2011.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2) :129–137, 1982.

Huzhang Mao and Liang Li. *PSW : Propensity Score Weighting Methods for Dichotomous Treatments*, 2018. URL `https://CRAN.R-project.org/package=PSW`. R package version 1.1-3.

Hernan Miguel and Jamie Robins. *Causal Inference*. Boca Raton : Chapman Hall/CRC, 2019.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 1901.

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

Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1) :1–26, 1977.

Bandyopadhyay Sanghamitra and Saha Sriparna. *Unsupervised Classification : Similarity Measures, Classical and Metaheuristic Approaches, and Applications*. Springer Publishing Company, Incorporated, 2012.

Marko Sarstedt and Erik Mooi. *Regression Analysis*, pages 193–233. 03 2014.