Université de Montréal

Modélisation incrémentale par méthode bayésienne

par

Kevin Rosamont

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistiques

16 mars 2016

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Modélisation incrémentale par méthode bayésienne

présenté par

Kevin Rosamont

a été évalué par un jury composé des personnes suivantes :

David Haziza

(président-rapporteur)

Jean-François Angers

(directeur de recherche)

François Bellavance

(codirecteur)

Mylène Bédard

(membre du jury)

Mémoire accepté le

16 mars 2016

SOMMAIRE

Les modèles incrémentaux sont des modèles statistiques qui ont été développés initialement dans le domaine du marketing. Ils sont composés de deux groupes, un groupe contrôle et un groupe traitement, tous deux comparés par rapport à une variable réponse binaire (le choix de réponses est « oui » ou « non »). Ces modèles ont pour but de détecter l'effet du traitement sur les individus à l'étude. Ces individus n'étant pas tous des clients, nous les appellerons : « prospects ». Cet effet peut être négatif, nul ou positif selon les caractéristiques des individus composants les différents groupes.

Ce mémoire a pour objectif de comparer des modèles incrémentaux d'un point de vue bayésien et d'un point de vue fréquentiste. Les modèles incrémentaux utilisés en pratique sont ceux de Lo (2002) et de Lai (2004). Ils sont initialement réalisés d'un point de vue fréquentiste. Ainsi, dans ce mémoire, l'approche bayésienne est utilisée et comparée à l'approche fréquentiste. Les simulations sont effectuées sur des données générées avec des régressions logistiques. Puis, les paramètres de ces régressions sont estimés avec des simulations Monte-Carlo dans l'approche bayésienne et comparés à ceux obtenus dans l'approche fréquentiste. L'estimation des paramètres a une influence directe sur la capacité du modèle à bien prédire l'effet du traitement sur les individus.
Nous considérons l'utilisation de trois lois *a priori* pour l'estimation des paramètres de façon bayésienne. Elles sont choisies de manière à ce que les lois *a priori* soient non informatives. Les trois lois utilisées sont les suivantes : la loi bêta transformée, la loi Cauchy et la loi normale.
Au cours de l'étude, nous remarquerons que les méthodes bayésiennes ont un réel impact positif sur le ciblage des individus composant les échantillons de petite taille.

Mots clefs : Modélisation incrémentale, simulation Monte-Carlo, régression logistique bayésienne, densité *a priori*, ciblage, marketing direct.

SUMMARY

Uplift modelling is a statistical method initially developed in marketing. It has two groups (a control group and a treatment group) that are compared using a binary response variable (the response can be « yes » or « no »). The goal of this model is to detect the treatment effect on prospects. This effect can be either negative, null or positive. It depends on characteristics of each individual in each group.

The purpose of this master thesis is to compare the Bayesian point of view with the frequentist one on uplift modelling. The uplift models used in this thesis are Lo model (2002) and Lai model (2004). Both of them are originally modeled using the frequentist point of view. Therefore, the Bayesian approach is modeled and compared to the frequentist one. Simulations are done on generated data from logistic regressions. Then regression parameters are estimated with Monte-Carlo simulations for Bayesian approach. They are then compared to parameter estimations from the frequentist approach. Parameter estimations have direct influences on the ability of the modelling to predict treatment effect on individual. Three priors are considered for the Bayesian estimation of the parameters. These densities are chosen such that they are non-informative. They are the following : transformed beta, Cauchy and normal.
In the course of the study, we will notice the Bayesian method has a real positive impact on targeting individual from the small size sample.

Key words : Uplift modelling, Monte-Carlo simulation, Bayesian logistic regression, *a priori* density, targeting, direct marketing.

TABLE DES MATIÈRES

Sommaire v

Summary vii

Liste des figures xi

Liste des tableaux xv

Remerciements xvii

Introduction 1

Chapitre 1. Rappels Statistiques 3

1.1. La régression logistique 4
1.1.1. Maximum de vraisemblance 5
1.1.2. Méthode itérative de Newton-Raphson 9
1.1.3. Exemple 10

1.2. Modèle bayésien 12
1.2.1. Paradigme bayésien 12
1.2.2. Loi *a priori* et *a posteriori* 13
1.2.3. Estimation ponctuelle 14

1.3. Simulation Monte-Carlo 15

1.4. La régression logistique bayésienne 17
1.4.1. Exemple (suite) 18

1.5. Les lois *a priori* utilisées 19

Chapitre 2. Modélisation Incrémentale 23

2.1. Qu'est-ce qu'un modèle incrémental ? 23

2.2. Modèles de Lo (2002) 27

2.3. Modèles de Lai (2004) 30

Chapitre 3. Simulation 35

3.1. Les paramètres de simulation 35

3.2. Génération des données 36

3.3. Exemple avec une variable explicative 37

3.3.1. Modèle de Lo 39

3.3.2. Modèle de Lai 45

3.3.3. Comparaison des modèles 47

3.4. Exemple avec deux variables explicatives 47

3.4.1. Modèle de Lo 49

3.4.2. Modèle de Lai 54

3.4.3. Comparaison des modèles 56

3.5. Simulations plus exhaustives 57

3.5.1. Description générale des jeux de données 58

3.5.2. Modèle de Lo 59

3.5.3. Modèle de Lai 69

3.5.4. Comparaison des modèles 73

Conclusion 77

Bibliographie 79

Annexes 81

LISTE DES FIGURES

1.1 Les points observés et leur estimation 12

2.1 Disposition des classes 26
2.2 Méthodologie proposée par Lo (2002) 27
2.3 Valeur ajoutée du ciblage du modèle. 32

3.1 Réponses et probabilités de réponses des prospects en fonction de X_1. 39
3.2 Incréments moyens observés par décile avec le modèle de Lo. 43
3.3 Différence des incréments moyens prédits et observés par décile. 44
3.4 Incréments moyens observés par décile avec le modèle de Lai. 46
3.5 Réponses et probabilités de réponses des prospects en fonction de X_1 et X_2. 49
3.6 Incréments moyens observés par décile avec le modèle de Lo. 52
3.7 Différence des incréments moyens prédits et observés par déciles. 53
3.8 Incréments moyens observés par décile avec le modèle de Lai. 55
3.9 Modèle de Lo avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 65
3.11 Comparaison entre les incréments moyens prédits et observés avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 67
3.12 Modèle de Lai avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 72
3.13 Modèle de Lo avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 81

3.14 Modèle de Lai avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 82

3.15 Modèle de Lo avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 83

3.16 Modèle de Lai avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 84

3.17 Modèle de Lo avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 85

3.18 Modèle de Lai avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 86

3.19 Modèle de Lo avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 87

3.20 Modèle de Lai avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 88

3.21 Modèle de Lo avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 89

3.22 Modèle de Lai avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 90

3.23 Modèle de Lo avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 91

3.24 Modèle de Lai avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 92

3.25 Modèle de Lo avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 93

3.26 Modèle de Lai avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 94

3.27 Modèle de Lo avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 95

3.28 Modèle de Lai avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 96

3.29 Modèle de Lo avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 97

3.30 Modèle de Lai avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 98

3.31 Modèle de Lo avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 99

3.32 Modèle de Lai avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 100

3.33 Modèle de Lo avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 101

3.34 Modèle de Lai avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 102

3.35 Modèle de Lo avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 103

3.36 Modèle de Lai avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement 104

3.37 Modèle de Lo avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 105

3.38 Modèle de Lai avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement 106

3.39 Modèle de Lo avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 107

3.40 Modèle de Lai avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement 108

LISTE DES TABLEAUX

2.1 Mesure de performance d'une campagne 25
2.2 Les réponses observées 30

3.1 Description des données 38
3.2 Paramètres estimés selon les différentes méthodes 40
3.3 Matrice de confusion avec le GLM 41
3.4 Matrice de confusion avec π_1 (loi bêta transformée *a priori*) 41
3.5 Matrice de confusion avec π_2 (loi Cauchy *a priori*) 42
3.6 Matrice de confusion avec π_3 (loi normale *a priori*) 42
3.7 Comparaison des incréments moyens prédits et observés avec le modèle de Lo. 45
3.8 Paramètres estimés selon les différentes méthodes 45
3.9 Récapitulatif des incréments moyens observés 47
3.10 Description des données 48
3.11 Paramètres estimés selon les différentes méthodes 50
3.12 Matrice de confusion avec le GLM 50
3.13 Matrice de confusion avec π_1 (loi bêta transformée *a priori*) 51
3.14 Matrice de confusion avec π_2 (loi Cauchy *a priori*) 51
3.15 Matrice de confusion avec π_3 (loi normale *a priori*) 51
3.16 Comparaison des incréments moyens prédits et observés avec le modèle de Lo 53
3.17 Paramètres estimés selon les différentes méthodes 54
3.18 Récapitulatif des incréments moyens observés 56
3.19 Valeur des paramètres selon le taux de réponse 57
3.20 Moyenne (écart-type) des jeux de données 58
3.21 Paramètres estimés selon les différentes méthodes 60

3.22 Les écart-types des paramètres estimés 61
3.23 Moyennes (écart-types) des classifications des matrices de confusion avec le GLM ($N = 500$) 62
3.24 Moyennes (écart-types) des classifications des matrices de confusion avec π_1 (loi bêta transformée *a priori*) ($N = 500$) 62
3.25 Moyennes (écart-types) des classifications des matrices de confusion avec π_2 (loi Cauchy *a priori*) ($N = 500$) 63
3.26 Moyennes (écart-types) des classifications des matrices de confusion avec π_3 (loi normale *a priori*) ($N = 500$) 63
3.27 Moyennes (écart-types) des classifications des matrices de confusion avec le GLM ($N = 10\,000$) 63
3.28 Moyennes (écart-types) des classifications des matrices de confusion avec π_1 (loi bêta transformée *a priori*) ($N = 10\,000$) 64
3.29 Moyennes (écart-types) des classifications des matrices de confusion avec π_2 (loi Cauchy *a priori*) ($N = 10\,000$) 64
3.30 Moyennes (écart-types) des classifications des matrices de confusion avec π_3 (loi normale *a priori*) ($N = 10\,000$) 64
3.31 Récapitulatif des incréments moyens prédits et observés 68
3.32 Paramètres estimés selon les différentes méthodes 69
3.33 Les écart-types des paramètres estimés 70
3.34 Récapitulatif des incréments moyens observés 74

REMERCIEMENTS

L'écriture du mémoire fut un travail de longue haleine qui m'aurait été impossible de réaliser sans le soutien de ma famille et de mes amis. Je tiens à remercier monsieur Angers pour sa disponibilité, son implication, son sens de la pédagogie et la confiance qu'il m'a accordée lors de l'écriture de ce mémoire. Je remercie également Monsieur Bellavance pour cette idée de mémoire et pour avoir accepté la codirection et m'avoir aidé à me poser les bonnes questions lorsque je rencontrais des difficultés.
En dernier lieu, je tiens plus spécialement à remercier le Département de mathématiques et de statistique de l'Université de Montréal, qui m'a offert un environnement où l'apprentissage est facilité par des laboratoires aux ordinateurs puissants, des professeurs toujours disponibles et un personnel administratif toujours réactif.

INTRODUCTION

Lorsqu'une entreprise met un nouveau produit ou une nouvelle offre sur le marché, elle cherche à cibler les prospects qui peuvent être intéressés. En effet, parmi les clients qui achèteront le produit, il y a deux catégories : les « décidés », qui sont les prospects répondant favorablement à l'offre puisqu'ils étaient intéressés avant que l'entreprise ne les contacte, et les « non-décidés », qui sont les prospects répondant favorablement à l'offre uniquement parce que l'entreprise les a contactés via des techniques de marketing telles que l'envoie de courriels, les appels téléphoniques, etc. S'il n'y avait pas eu de prise de contact avec ces prospects, ils ne répondraient pas positivement à l'offre. Cela peut s'illustrer par la mise sur le marché d'un nouvel abonnement dans le domaine de la télécommunication. Les « non-décidés » seront ceux qui n'avaient pas l'intention de changer d'abonnement mais qui le feront après que l'entreprise les ait contactés (via une technique marketing), alors que les « décidés » contacteront l'entreprise de leur propre gré. Par conséquent, les entreprises cherchent à utiliser leurs capitaux pour cibler les « non-décidés » afin de maximiser leur retour sur investissement.
Cependant, il y a des situations dans lesquelles il est difficile de savoir si le client était un prospect « décidé » ou « non-décidé » une fois qu'il a répondu positivement, c'est la raison pour laquelle il est important de collecter des informations sur les clients pour identifier et différentier ces deux types de clients. Ainsi, les modèles incrémentaux sont des modèles statistiques de prévision qui ont pour but de détecter les différences de probabilités de réponse positive lorsque le prospect est contacté par l'entreprise et lorsqu'il ne l'est pas. Les modèles incrémentaux rompent avec la vision traditionnelle du marketing directe qui voit la réponse positive d'un prospect comme une conséquence de la prise de contact effectuée par l'entreprise mais, oubliant que dans les clients qui répondent positivement, il y a des « décidés » qui auraient répondu positivement indépendamment de la prise de contact. Les modèles traditionnels ont aussi pour défaut de perdre des clients potentiels qui auraient répondu positivement s'ils n'avaient pas été contactés. La prise de contact peut donc avoir un effet négatif sur certains clients (Allen, 1997).

La modélisation incrémentale dans le domaine du marketing a été approchée de manières différentes par Radcliffe (1999), Radcliffe et Surry (2007), Chickering et Heckerman (2000), Hansotia *et al.* (2002), Lo (2002) et Lai (2004). Dans le cadre de ce mémoire, nous nous sommes intéressés à modéliser d'un point de vue bayésien les méthodes des deux derniers auteurs qui peuvent être modélisées à partir d'une régression logistique bayésienne. Trois lois *a priori* ont été utilisées pour les régressions logistiques bayésiennes : la loi normale, une transformation de la loi bêta et une loi Cauchy comme le suggère l'article de Gelman *et al.* (2008). Les résultats obtenus ont été analysés et comparés à ceux obtenus avec des régressions logistiques classiques.

Dans le premier chapitre sont présentés les différents concepts statistiques qui permettent la modélisation incrémentale. Dans un premier temps, nous nous intéressons au point de vue fréquentiste, la régression logistique est présentée ainsi que la méthode de maximum de vraisemblance pour estimer ses paramètres et la méthode de Newton-Raphson pour obtenir une solution numérique. Dans un second temps, nous nous intéressons au point de vue bayésien, la régression logistique bayésienne est présentée ainsi que l'estimation des paramètres et l'approche numérique par simulation Monte-Carlo. Dans le second chapitre, la modélisation incrémentale est présentée dans son ensemble. Puis, les modèles de Lo (2002) et de Lai (2004) sont présentés au même titre que leur mesure de performance respective. Dans le troisième chapitre, les simulations sont réalisées et analysées. Dans un premier temps, les régressions sont effectuées avec une variable explicative puis deux. Pour finir, les simulations sont réalisées avec trois variables explicatives dans le même cadre que les simulations réalisées par Lo (2002). Les jeux de données sont simulés plusieurs fois afin d'obtenir des résultats généralisables. Ces derniers sont présentés sur des graphiques en barres et comparés en fonction des méthodes utilisées et des paramètres de simulations. Enfin, dans un tableau sont résumés les résultats obtenus selon les différentes caractéristiques des échantillons simulés.

Chapitre 1

RAPPELS STATISTIQUES

Dans ce chapitre sont rappelés certains concepts statistiques qu'il est important d'introduire avant d'évoquer la régression logistique bayésienne. Ce chapitre est essentiellement concentré sur l'estimation des paramètres.
Dans un premier temps, la régression logistique est rappelée ainsi que la méthode de maximum de vraisemblance pour estimer les paramètres. Sachant que les paramètres ne peuvent être estimés de façon analytique, la méthode itérative de Newton-Raphson est évoquée. L'estimation des paramètres est illustrée par un exemple.
Dans un second temps, nous nous intéressons à l'estimation de paramètres via l'inférence bayésienne : le théorème de Bayes, le modèle bayésien avec la loi *a priori* et *a posteriori*. L'estimateur ponctuel de Bayes est une valeur théorique que nous allons approcher numériquement par des simulations de Monte-Carlo. L'utilisation de cette méthode est justifiée par la loi faible des grands nombres.
Pour finir, la régression logistique bayésienne est abordée ainsi que l'estimation de ses paramètres. Puis, dans l'exemple en fin de chapitre, les estimateurs des paramètres obtenus avec les estimateurs bayésiens et la méthode de maximum de vraisemblance sont comparés.
Différentes notations vont être utilisées dans l'ensemble du mémoire, les variables aléatoires sont écrites en lettre majuscule comme « X » alors que les valeurs observées sont écrites en minuscule comme « x ». De plus, les lettres écrites en gras font référence à des vecteurs comme « $\mathbf{X}$ » qui est un vecteur aléatoire composé d'un certain nombre p de variables aléatoires (qui sera toujours précisé) : $\mathbf{X} = (X_1, X_2, \ldots, X_p)^t$. Il y aura aussi des vecteurs de valeurs observées comme « $\mathbf{x}$ ».

Définition 1.0.1. *Modèle paramétrique statistique.*
Un modèle paramétrique statistique consiste en l'observation d'une variable aléatoire Y_i distribuée selon $f(y_i|\theta)$, où seulement le paramètre θ est inconnu et appartient à un espace de dimension finie.

Dans l'analyse statistique fréquentiste, θ est vue comme une valeur fixe qui est inconnue et que l'on veut estimer alors que dans l'analyse statistique bayésienne, le paramètre θ est vu comme une variable aléatoire telle que $\theta \in \Theta$, un espace de dimension finie ou non. De plus, si les Y_i sont des variables aléatoires indépendantes et identiquement distribuées (*i.i.d.*), alors on peut aussi écrire :

$$f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

L'information fournie par le vecteur $\mathbf{y}$ est contenue dans la densité conjointe $f(\mathbf{y}|\theta)$ qui désigne aussi la fonction de vraisemblance L :

$$L(\theta|\mathbf{y}) = f(\mathbf{y}|\theta).$$

Les paramètres sont inversés pour faire comprendre que θ est inconnu et qu'il s'estime à partir des observations $\mathbf{y}$. Cette inversion reflète un des objectifs premiers de cette approche qui est d'estimer θ avec un certain degré de précision. Elle est naturellement liée au théorème de Bayes qui formalise l'inversion des conditionnements dans les probabilités.

1.1. La régression logistique

La régression logistique est un modèle multidimensionnel où Y est la variable dépendante et $\mathbf{X} = (X_1, X_2, \ldots, X_p)^t$, est un ensemble de variables indépendantes qui peuvent être qualitatives ou quantitatives. La variable dépendante est habituellement la survenue ou non d'un événement et les variables indépendantes sont celles susceptibles d'avoir une incidence dans la survenue de cet événement.
L'intérêt de la régression réside dans le fait de pouvoir modéliser la variable dépendante Y en fonction des variables indépendantes $\mathbf{X}$ et des paramètres $\boldsymbol{\beta}$ qui leurs sont attribués, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$. Dans le cadre de l'étude, Y est toujours une variable binaire, ainsi, la régression la plus adaptée à ce genre de variable est la régression logistique.

Le modèle fondamental de la régression logistique popularisée par Good (1950), Jaynes (1956) et Tribus (1969), est :

$$\log\left(\frac{\mathbb{P}(\mathbf{X}|Y=1,\boldsymbol{\beta})}{\mathbb{P}(\mathbf{X}|Y=0,\boldsymbol{\beta})}\right)=\beta_0^*+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p,$$

où β_0^* est l'ordonnée à l'origine du modèle ci-dessus.

En posant $\nu(\mathbf{X},\boldsymbol{\beta})=\mathbb{P}(Y=1|\mathbf{X},\boldsymbol{\beta})$, on obtient l'égalité suivante :

$$\begin{aligned}\log\left(\frac{\nu(\mathbf{X},\boldsymbol{\beta})}{1-\nu(\mathbf{X},\boldsymbol{\beta})}\right)&=\log\left(\frac{\mathbb{P}(Y=1|\mathbf{X},\boldsymbol{\beta})}{\mathbb{P}(Y=0|\mathbf{X},\boldsymbol{\beta})}\right)\\&=\log\left(\frac{\mathbb{P}(Y=1)\mathbb{P}(\mathbf{X}|Y=1,\boldsymbol{\beta})}{\mathbb{P}(Y=0)\mathbb{P}(\mathbf{X}|Y=0,\boldsymbol{\beta})}\right)\\&=\log\left(\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}\right)+\log\left(\frac{\mathbb{P}(\mathbf{X}|Y=1,\boldsymbol{\beta})}{\mathbb{P}(\mathbf{X}|Y=0,\boldsymbol{\beta})}\right)\\&=\log\left(\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}\right)+\beta_0^*+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p\\&=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p,\end{aligned}\tag{1.1.1}$$

où

$$\beta_0=\log\left(\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}\right)+\beta_0^*.$$

Par conséquent, on déduit d'après (1.1.1) que :

$$\nu(\mathbf{X},\boldsymbol{\beta})=\frac{\mathrm{e}^{\beta_0+\sum_{j=1}^p\beta_jX_j}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^p\beta_jX_j}}.$$

Il est important de souligner que dans la littérature, $\log\left(\frac{\nu(\mathbf{X},\boldsymbol{\beta})}{1-\nu(\mathbf{X},\boldsymbol{\beta})}\right)$ s'écrit comme le logit de $\nu(\mathbf{X},\boldsymbol{\beta})$:

$$\mathrm{logit}(\nu(\mathbf{X},\boldsymbol{\beta}))=\log\left(\frac{\nu(\mathbf{X},\boldsymbol{\beta})}{1-\nu(\mathbf{X},\boldsymbol{\beta})}\right).$$

1.1.1. Maximum de vraisemblance

Dans un cadre fréquentiste, les $p+1$ paramètres $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_p)^t$ s'estiment à l'aide du maximum de vraisemblance. Pour chaque individu i, est associée une variable dépendante Y_i. Cette variable suit une loi de Bernoulli car l'événement peut survenir comme il peut ne pas survenir. Si la population est composée de

N individus, il y aura une réponse y_i par individu et le vecteur $\mathbf{y}$ s'écrit donc $\mathbf{y} = (y_1, y_2, \ldots, y_N)^t$, tel que $y_i \in \{0,1\}$ et $\mathbf{Y} \sim f(\mathbf{y}|\boldsymbol{\beta})$. Les variables indépendantes et les valeurs observées de l'individu i s'écrivent respectivement : $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^t$ et $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^t$. Ainsi,

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}) &= \prod_{i=1}^{N} f(y_i|\boldsymbol{\beta}) \\ &= \prod_{i=1}^{N} \left(\frac{\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} \right)^{y_i} \left(1 - \frac{\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} \right)^{1-y_i} \\ &= \prod_{i=1}^{N} \left(\frac{\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} \right)^{y_i} \left(\frac{1}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} \right)^{1-y_i}. \end{aligned}$$

On va noter :

$$\begin{aligned} \nu_i &= \nu(\mathbf{x}_i, \boldsymbol{\beta}) \\ &= \frac{\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}, \end{aligned}$$

La densité conjointe est aussi la fonction de vraisemblance L :

$$f(\mathbf{y}|\boldsymbol{\beta}) = L(\boldsymbol{\beta}|\mathbf{y}),$$

et par conséquent :

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{N} (\nu_i)^{y_i} (1-\nu_i)^{1-y_i}. \tag{1.1.2}$$

Dans un premier temps, les estimateurs de $\boldsymbol{\beta}$ qui vont nous intéresser sont les estimateurs de maximum de vraisemblance, notés $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)^t$. Ces estimateurs sont les valeurs des paramètres $\beta_0, \beta_1, \ldots, \beta_p$ qui maximisent la fonction de vraisemblance L. Ainsi, pour trouver ces valeurs, il faut dans un premier temps passer par la dérivée et résoudre le système d'équations suivant : $\dfrac{\partial}{\partial \beta_k} L(\boldsymbol{\beta}|y) = 0$ avec $k \in \{0,1,2,3,\ldots,p\}$. Dans un second temps, il faut vérifier que la matrice hessienne soit définie négative afin que l'estimateur obtenu soit bien un maximum. Ces deux étapes sont détaillées dans les lignes qui suivent.
Pour estimer $\boldsymbol{\beta}$, il est souvent plus pratique d'utiliser la transformation de la log-vraisemblance. Calculer le maximum de la fonction de vraisemblance revient à calculer celui de la log-vraisemblance. Nous avons :

$$\begin{aligned} l(\boldsymbol{\beta}|\mathbf{y}) &= \log(L(\boldsymbol{\beta}|\mathbf{y})) \\ &= \sum_{i=1}^{N} y_i \log(\nu_i) + \sum_{i=1}^{N} (1-y_i) \log(1-\nu_i) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \log(1-\nu_i) + \sum_{i=1}^{N} y_i \log\left(\frac{\nu_i}{1-\nu_i}\right) \\
&= \sum_{i=1}^{N} \log\left(\frac{1}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}\right) + \sum_{i=1}^{N} y_i \log\left(\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}\right) \\
&= -\sum_{i=1}^{N} \log\left(1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}\right) + \sum_{i=1}^{N} y_i \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right),
\end{aligned}
$$

cela implique que :

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_0} &= -\sum_{i=1}^{N} \frac{\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} + \sum_{i=1}^{N} y_i \\
&= \sum_{i=1}^{N} \left(y_i - \nu_i\right),
\end{aligned}
$$

et pour $k > 0$, nous avons :

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_k} &= -\sum_{i=1}^{N} \frac{x_{ik}\,\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} + \sum_{i=1}^{N} y_i x_{ik} \\
&= \sum_{i=1}^{N} \left(y_i - \nu_i\right) x_{ik}.
\end{aligned}
$$

Ainsi, pour trouver l'estimateur de vraisemblance maximale, on pose le système suivant :

$$
\begin{cases}
\dfrac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_0} &= 0 \\
\dfrac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_1} &= 0 \\
\vdots & \\
\vdots & \\
\dfrac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_p} &= 0,
\end{cases}
\tag{1.1.3}
$$

qui revient à :

$$
\begin{cases}
\sum_{i=1}^{N} \left(y_i - \nu_i\right) &= 0 \\
\sum_{i=1}^{N} \left(y_i - \nu_i\right) x_{i1} &= 0 \\
\vdots & \\
\vdots & \\
\sum_{i=1}^{N} \left(y_i - \nu_i\right) x_{ip} &= 0.
\end{cases}
$$

Ensuite, il faut que la matrice hessienne soit définie négative aux valeurs estimées $\widehat{\boldsymbol{\beta}}$ pour que ces estimations soient bien des maximums de vraisemblance :

$$\mathrm{H}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_0^2} & \cdots & \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_0 \partial \beta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_n \partial \beta_0} & \cdots & \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_n^2} \end{pmatrix},$$

où

$$\begin{aligned} \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_0^2} &= -\sum_{i=1}^{N} \left(\frac{\mathrm{e}^{\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ik}} (1 + \mathrm{e}^{\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ik}}) - \left(\mathrm{e}^{\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ik}} \right)^2}{(1 + \mathrm{e}^{\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ik}})^2} \right) \\ &= -\sum_{i=1}^{N} \widehat{\nu}_i (1 - \widehat{\nu}_i), \end{aligned} \tag{1.1.4}$$

avec $\widehat{\nu}_i = \dfrac{\mathrm{e}^{\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ik}}}{1 + \mathrm{e}^{\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ik}}}$,

et pour $k > 0$, nous avons :

$$\begin{aligned} \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_k \beta_0} &= -\sum_{i=1}^{N} \left(\frac{\nu_i}{1 + \mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}} \right) x_{ik} \\ &= -\sum_{i=1}^{N} \widehat{\nu}_i (1 - \widehat{\nu}_i) x_{ik}, \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_k^2} &= -\sum_{i=1}^{N} \left(\frac{x_{ik} \mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} (1 + \mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}) - x_{ik} \left(\mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} \right)^2}{(1 + \mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}})^2} \right) x_{ik} \\ &= -\sum_{i=1}^{N} \widehat{\nu}_i (1 - \widehat{\nu}_i) x_{ik}^2 \end{aligned} \tag{1.1.5}$$

et pour $h \neq k$, $h, k > 0$

$$\begin{aligned} \frac{\partial^2 l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_k \beta_h} &= -\sum_{i=1}^{N} \left(\frac{\nu_i}{1 + \mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}} \right) x_{ih} x_{ik} \\ &= -\sum_{i=1}^{N} \widehat{\nu}_i (1 - \widehat{\nu}_i) x_{ih} x_{ik}. \end{aligned} \tag{1.1.6}$$

Il n'est pas possible de résoudre le système d'équations (1.1.3) analytiquement, il faut procéder par une méthode numérique. Le logiciel R, qui est utilisé pour

l'ensemble de l'étude, estime les paramètres en procédant par la méthode itérative de Newton-Raphson.

1.1.2. Méthode itérative de Newton-Raphson

Cette méthode est un algorithme itératif qui consiste à trouver la meilleure approximation d'un zéro (ou racine) de l'équation de la tangente.
On rappelle que l'équation de la tangente s'écrit :

$$y = f(x_0) + f'(x_0)(x - x_0).$$

Puisque nous voulons approcher une racine de cette équation, nous posons $y = 0$, ainsi :

$$f(x_0) + f'(x_0)(x - x_0) = 0, \tag{1.1.7}$$

puis en résolvant l'équation (1.1.7), on trouve x :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Une fois la solution x_1 trouvée, de nouvelles itérations de la racine cherchée sont effectuées pour $k \geq 0$:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Ainsi, on calcule $x_1, x_2, x_3, \ldots$, jusqu'à ce que x_k converge q-quadratiquement vers x^* tel que $f(x^*) = 0$.

Définition 1.1.1. *Convergence q-quadratique.*
On dit que $(x_k)_{k\geq 1}$ converge q-quadratiquement vers x^* s'il existe une constante $\varsigma \geq 0$ telle que $\forall k \geq 1$ on ait :

$$\| x_{k+1} - x \| \leq \varsigma \| x_k - x \|^2 .$$

Nous avons vu comment utiliser l'algorithme Newton-Raphson dans un cas unidimensionnel. Maintenant, généralisons l'algorithme à un cas multidimensionnel. Pour que l'algorithme converge q-quadratiquement, il faut que f vérifie les conditions suivantes :

— f est deux fois continument dérivable et

$$\| \nabla^2 f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}) \| \leq \varsigma \| \mathbf{x}_k - \mathbf{x} \|^2,$$

où ς est une constante telle que $\varsigma \geq 0$.

— ∇f est la matrice des dérivées premières, ici $\nabla f(\mathbf{x}_k) = 0$.

— $\nabla^2 f$ est la matrice des dérivées secondes, ici $\nabla^2 f(\mathbf{x}_k)$ est positivement définie.

Dans le cas de l'estimation de $\boldsymbol{\beta}$, il faut généraliser la méthode, au système d'équations (1.1.3) qui contient $(p+1)$ équations et $(p+1)$ inconnus. Dans l'équation (1.1.7), on voulait approcher une racine de f, ici nous voulons plutôt approcher les racines de ∇l qui est le gradient de l. L'opérateur ∇ est tel que :

$$\nabla = \left(\frac{\partial}{\partial \beta_0}, \frac{\partial}{\partial \beta_1}, \ldots, \frac{\partial}{\partial \beta_p} \right)^t.$$

La matrice hessienne $\mathrm{H}(\boldsymbol{\beta}^{(k)})$ représente les dérivées secondes de l qui sont données aux équations (1.1.4) à (1.1.6). La généralisation de la méthode de Newton-Raphson s'écrit donc :

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left[-\mathrm{H}(\boldsymbol{\beta}^{(k)}) \right]^{-1} \nabla l(\boldsymbol{\beta}^{(k)}), \tag{1.1.8}$$

où $\mathrm{H}(\boldsymbol{\beta}^{(k)}) = \nabla^2 l(\boldsymbol{\beta}^{(k)})$.

Cette méthode converge assez rapidement, pour plus de détails, voir Kelly (1999). Dans la sous-section suivante, un exemple est effectué afin d'illustrer cette méthode.

1.1.3. Exemple

Soient huit individus dont nous connaissons les valeurs des variables X et Y. La variable Y correspond à une variable dichotomique qui prend les valeurs 0 ou 1 et X est une variable continue telle que $X \in [0, 1]$. Nous voulons modéliser une régression logistique avec les valeurs observées (x, y), c'est-à-dire, exprimer y en fonction de x :

x	0,20	0,35	0,50	0,60	0,65	0,70	0,80	0,95
y	0	0	0	1	1	0	1	1

Le modèle logistique s'écrit :

$$\nu(x, \boldsymbol{\beta}) = \frac{\mathrm{e}^{\beta_0 + \beta_1 x}}{1 + \mathrm{e}^{\beta_0 + \beta_1 x}},$$

où β_0 et β_1 sont les paramètres à estimer. Il faut donc résoudre le système d'équations (1.1.3) :

$$\begin{cases} \dfrac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_0} &= 0 \\ \dfrac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_1} &= 0, \end{cases}$$

où la fonction de log-vraisemblance l s'écrit :

$$l(\boldsymbol{\beta}|\mathbf{y}) = -\sum_{i=1}^{8} \log\left(1 + \mathrm{e}^{\beta_0+\beta_1 x_i}\right) + \sum_{i=1}^{8} y_i \left(\beta_0 + \beta_1 x_i\right).$$

Pour résoudre ce système d'équations non linéaires, la méthode de Newton-Raphson est utilisée :

$$\begin{pmatrix} \beta_0^{(k+1)} \\ \beta_1^{(k+1)} \end{pmatrix} = \begin{pmatrix} \beta_0^{(k)} \\ \beta_1^{(k)} \end{pmatrix} - \begin{pmatrix} \dfrac{\partial^2 l(\boldsymbol{\beta}^{(k)}|\mathbf{y})}{\partial \beta_0^2} & \dfrac{\partial^2 l(\boldsymbol{\beta}^{(k)}|\mathbf{y})}{\partial \beta_0 \partial \beta_1} \\ \dfrac{\partial^2 l(\boldsymbol{\beta}^{(k)}|\mathbf{y})}{\partial \beta_0 \partial \beta_1} & \dfrac{\partial^2 l(\boldsymbol{\beta}^{(k)}|\mathbf{y})}{\partial \beta_1^2} \end{pmatrix}^{-1} \begin{pmatrix} \dfrac{\partial l(\boldsymbol{\beta}^{(k)}|\mathbf{y})}{\partial \beta_1} \\ \dfrac{\partial l(\boldsymbol{\beta}^{(k)}|\mathbf{y})}{\partial \beta_1} \end{pmatrix},$$

où $\boldsymbol{\beta}^{(k)} = (\beta_0^{(k)}, \beta_1^{(k)})$.
En partant initialement de $k = 0$, on choisit $\boldsymbol{\beta}^{(0)} = (0, 0)$. Numériquement, nous obtenons :

$$\begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 2{,}000 & 11{,}88 \\ 11{,}88 & 80{,}69 \end{pmatrix}^{-1} \begin{pmatrix} 0{,}00 \\ 6{,}25 \end{pmatrix} = \begin{pmatrix} -3{,}65 \\ 6{,}14 \end{pmatrix}$$

Puis $\boldsymbol{\beta}^{(k)}$ est calculé itérativement jusqu'à ce que $\nabla l(\boldsymbol{\beta}^{(k)}|\mathbf{y}) \approx 0$.
Dans cet exemple, on atteint la convergence à $k = 7$ avec $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(7)}$:

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} -7,46 \\ 12{,}25 \end{pmatrix}.$$

En remplaçant $\boldsymbol{\beta}$ par $\widehat{\boldsymbol{\beta}}$ dans le système d'équations (1.1.3), nous obtenons :

$$\begin{cases} \dfrac{\partial l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_0} &= -1{,}95 \times 10^{-6} \\ \dfrac{\partial l(\widehat{\boldsymbol{\beta}}|\mathbf{y})}{\partial \beta_1} &= 8{,}39 \times 10^{-7}. \end{cases}$$

Ainsi, nous avons bien $\nabla l(\boldsymbol{\beta}^{(k)}|\mathbf{y}) \approx 0$.

Ainsi, la valeur y est prédite par la fonction $\hat{\nu}$ qui s'écrit encore :

$$\hat{\nu}(x) = \nu(x, \hat{\boldsymbol{\beta}}).$$

La règle suivante est appliquée : lorsque $\hat{\nu} > 0{,}5$, la valeur prédite de y sera 1 sinon ce sera 0.
Les points observés formés des coordonnées (x, y) et leur estimation $(x, \hat{\nu}(x))$ sont représentés sur la figure 1.1

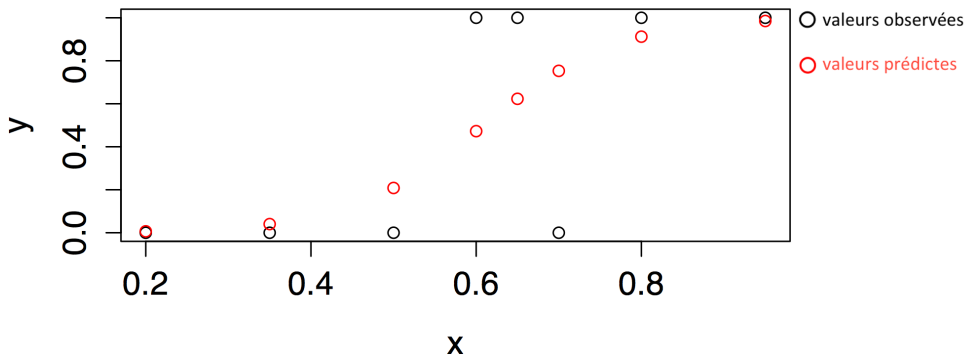


FIGURE 1.1. Les points observés et leur estimation

1.2. MODÈLE BAYÉSIEN

Les principaux concepts de la modélisation bayésienne sont rappelés dans cette section. Pour plus de détails ou approfondir le sujet, voir Robert (2006).

1.2.1. Paradigme bayésien

Soit $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^t$ un vecteur aléatoire qui a pour densité f et comme paramètre θ, tel que $\mathbf{Y} \sim f(\mathbf{y}|\theta)$. Puis, introduisons les différents espaces intervenant dans la mise en place du modèle bayésien :
$\mathscr{X}$: l'espace des observations,
Θ : l'espace des paramètres,
$\mathscr{A}$: l'espace des actions ou des décisions, dont les éléments sont des images des observations par une application d appelée règle de décision. Dans le cas de l'estimation ponctuelle, $\mathscr{A} = \Theta$.

Théorème 1.2.1. *Théorème de Bayes.*
Si A et B sont des événements tels que $\mathbb{P}(B) \neq 0$, alors $\mathbb{P}(A|B)$ et $\mathbb{P}(B|A)$ sont reliées par :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}.$$

Si v et w sont deux variables aléatoires continues, la version continue de ce théorème est la suivante :

$$g(w|v) = \frac{f(v|w)g(w)}{f(v)},$$

avec $f(v) \neq 0$. De plus, $f(v) = \int_{\mathbf{W}} f(v|w)g(w)dw$, où $\mathbf{W}$ est le support de la variable aléatoire.

1.2.2. Loi *a priori* et *a posteriori*

Dans l'analyse statistique bayésienne, le paramètre θ est considéré comme une variable aléatoire à valeur dans Θ. Cet espace est de dimension finie ou non et il est muni d'une distribution π tel que $\theta \sim \pi$. La distribution $\pi(\theta)$ est appelée loi *a priori*. Ainsi, elle détermine ce que l'on sait sur θ avant d'observer les $\mathbf{y}$.

Définition 1.2.1. *Modèle statistique bayésien.*
Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique, $f(\boldsymbol{y}|\theta)$, et d'une loi a priori pour le paramètre $\pi(\theta)$.

La difficulté du choix de la loi *a priori* réside dans le fait que l'information *a priori* n'est pas suffisamment précise pour qu'une seule loi de probabilité soit compatible avec l'information à disposition. En effet, il y a souvent plusieurs lois qui semblent compatibles et qu'il faut considérer. La loi *a priori* peut avoir un grand impact sur la loi *a posteriori* qui en découle puisque cette dernière est obtenue en utilisant la version continue du théorème de Bayes, elle s'écrit $\pi(\theta|\mathbf{y})$. Elle est donnée par :

$$\begin{aligned} \pi(\theta|\mathbf{y}) &= \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} \\ &= \frac{L(\theta|\mathbf{y})\pi(\theta)}{f(\mathbf{y})} \\ &= \frac{L(\theta|\mathbf{y})\pi(\theta)}{\int_\Theta L(\theta|\mathbf{y})\pi(\theta)d\theta}. \end{aligned} \tag{1.2.1}$$

La loi *a posteriori* représente une mise à jour de l'information après avoir observé les $\mathbf{y}$. On utilise souvent la notation de proportionnalité $\pi(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$ signifiant que la loi *a posteriori* de $\theta : \pi(\theta|\mathbf{y})$, est égale à $f(\mathbf{y}|\theta)\pi(\theta)$ à une constante près, cette constante est $\dfrac{1}{\int_\Theta L(\theta|\mathbf{y})\pi(\theta)d\theta}$.

Une des critiques faite à l'encontre de l'approche bayésienne vient du fait que l'on choisisse la loi *a priori*. Par conséquent, il est possible de choisir une loi *a*

priori qui donnera la réponse que l'on souhaite obtenir puisque le choix de la loi *a priori* est une étape cruciale dans la détermination de la loi *a posteriori*. Mais il est important de rappeler qu'une loi *a priori* non fondée donnera assez souvent une loi *a posteriori* non justifiée.

1.2.3. Estimation ponctuelle

Pour estimer θ, il est nécessaire d'utiliser une règle de décision d et une fonction de coût C.

Définition 1.2.2. *Fonction de coût.*
On appelle fonction de coût, toute fonction C de $\mathscr{A} \times \Theta$ dans $\mathbb{R}$.

La fonction de coût quantifie les conséquences de l'erreur commise en estimant le paramètre θ par la règle de décision d. L'estimation a un coût moyen égal à $\mathbb{E}_\pi[C(d,\theta)|\mathbf{y}]$ où cette notation signifie que l'espérance est prise sous $\pi(\theta|\mathbf{y})$. L'estimation optimale est celle qui minimise cette erreur. On définit donc un estimateur bayésien $\delta(\mathbf{y})$ comme la règle qui à chaque échantillon observé associe la solution du problème de minimisation :

$$\delta(\mathbf{y}) = \underset{d\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_\pi[C(d,\theta)|\mathbf{y}].$$

La fonction de perte par défaut est la fonction de perte quadratique définie par :

$$C(d,\theta) = (d-\theta)^2,$$

mais dans le cas multidimensionnel, nous utiliserons la norme quadratique. L'estimateur bayésien $\delta(\mathbf{y})$ correspondant est donné par :

$$\begin{aligned}\delta(\mathbf{y}) &= \mathbb{E}_\pi[\theta|\mathbf{y}] \\ &= \int_\Theta \theta\pi(\theta|\mathbf{y})d\theta. \end{aligned} \tag{1.2.2}$$

Preuve

$$\begin{aligned}\mathbb{E}_\pi[C(d,\theta)|\mathbf{y})] &= \int_\Theta L(d,\theta)|\mathbf{y})\pi(\theta|\mathbf{y})d\theta \\ &= \int_\Theta (d-\theta)^2\pi(\theta|\mathbf{y})d\theta \\ &= d^2\int_\Theta \pi(\theta|\mathbf{y})d\theta + \int_\Theta \theta^2\pi(\theta|\mathbf{y})d\theta - 2d\int_\Theta \theta\pi(\theta|\mathbf{y})d\theta.\end{aligned}$$

Or $\int_\Theta \pi(\theta|\mathbf{y})d\theta = 1$, par conséquent :

$$\mathbb{E}_\pi[C(d,\theta)|\mathbf{y})] = d^2 + \int_\Theta \theta^2\pi(\theta|\mathbf{y})d\theta - 2d\int_\Theta \theta\pi(\theta|\mathbf{y})d\theta$$

$$
\begin{aligned}
&= d^2 + \mathbb{E}_\pi[\theta^2|\mathbf{y}] - 2d\mathbb{E}_\pi[\theta|\mathbf{y}] + \mathbb{E}_\pi[\theta|\mathbf{y}]^2 - \mathbb{E}_\pi[\theta|\mathbf{y}]^2 \\
&= (d - \mathbb{E}_\pi[\theta|\mathbf{y}])^2 + \mathrm{var}(\theta|\mathbf{y}) \\
&\geq \mathrm{var}(\theta|\mathbf{y}).
\end{aligned}
$$

Donc l'espérance $\mathbb{E}_\pi[C(d,\theta)|\mathbf{y})]$ est minimisée par :

$$d = \mathbb{E}_\pi[\theta|\mathbf{y}].$$

□

1.3. Simulation Monte-Carlo

Pour approximer numériquement des espérances, il y a plusieurs méthodes, mais l'une des approches les plus simples et les plus efficaces est d'utiliser la méthode Monte-Carlo. Cette méthode est explicitée dans cette section mais elle est plus détaillée dans Gentle (1998).
Soit g une fonction intégrable définie sur un intervalle fermé de $\mathbb{R}$ et U une variable aléatoire telle que $U \sim f$ où f est une fonction de densité sur $\mathbb{R}$, alors $\mathbb{E}[g(U)]$ s'écrit :

$$\mathbb{E}[g(U)] = \int_{\mathbb{R}} g(u)f(u)du.$$

Théorème 1.3.1. ***(Théorème de la loi faible des grands nombres)*** *Soit $(U_m)_{m\in\mathbb{N}}$ une suite de variables aléatoires intégrables indépendantes et identiquement distribuées (i.i.d.), alors, lorsque $m \to \infty$,*

$$\frac{1}{m}\sum_{i=1}^{m} U_i \xrightarrow{P} \mu = \mathbb{E}[U_1],$$

pour une certaine constante μ, si et seulement si $\mathbb{E}[|U_1|] < \infty$.
(Une preuve de ce théorème est donnée dans Durrett [2010, p.53-55].)

La méthode Monte-Carlo peut être vue comme une mise en application de la loi faible des grands nombres à condition que g soit une fonction intégrable par rapport à f. L'approximation de $\mathbb{E}[g(U)]$ passe donc par la génération de M observations simulées $u_{(i)}$ provenant de la loi de U, Ainsi :

$$
\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M} g(u_{(i)}) &\simeq \mathbb{E}[g(U)] \\
&= \int_{\mathbb{R}} g(u)f(u)du.
\end{aligned}
$$

Cependant, la précision de l'estimé dépend de plusieurs paramètres : la fonction g, le nombre M de simulations générées et de la loi de U.

Par la suite, pour obtenir l'intervalle de confiance de $\mathbb{E}[g(U)]$, on utilise le théorème de la limite centrale.

Théorème 1.3.2. ***(Théorème central limite)***
Soit $(U_m)_{m\in\mathbb{N}}$ une suite de variables aléatoires réelles i.i.d. de carré intégrable, c'est-à-dire que $\mathbb{E}[|U_1|^2] < \infty, \mu = \mathbb{E}[U_1]$ et $\sigma^2 = Var(U_1)$.
Alors la suite $\frac{\sqrt{m}}{\sigma}\left(\overline{U}_m - \mu\right)$ converge en loi vers une variable Z de loi gaussienne centrée réduite, notée $\mathscr{N}(0,1)$.
(Une preuve ce théorème est donnée dans Durrett [2010, p.106-107].)

Du théorème central limite est directement déduit le corollaire suivant.

Corollaire 1.3.1. *Soit $(U_m, m \geqslant 1)$ une suite de variables aléatoires intégrables indépendantes et identiquement distribuées, de carré intégrable et d'espérance μ et de variance σ^2. Alors pour toute fonction $g : \mathbb{R} \to \mathbb{R}$ continue bornée, si Z désigne une variable aléatoire gaussienne $\mathscr{N}(0,1)$:*

$$\lim_{n\to\infty} \mathbb{E}\left[g\left(\frac{\sqrt{m}}{\sigma}\left(\overline{U}_m - \mu\right)\right)\right] = \mathbb{E}\left[g(Z)\right] = \int_{-\infty}^{\infty} g(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\,dz.$$

De plus, pour tout couple de nombre réels $a < b$, on a :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sigma}{\sqrt{m}}a \leq \overline{U}_m - \mu \leq \frac{\sigma}{\sqrt{m}}b\right) = \int_a^b \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\,dz.$$

Une table de fonction de répartition d'une loi gaussienne centrée réduite montre que si Z est $\mathscr{N}(0,1), \mathbb{P}(|Z| \leq 1,96) = 0,95$, on en déduit que pour m assez grand,

$$\mathbb{P}\left(|\overline{U}_m - \mathbb{E}[U_1]| \leq 1,96\frac{\sigma}{\sqrt{m}}\right) \simeq 0{,}95;$$

c'est-à-dire que l'on a un intervalle de confiance de $\mathbb{E}[U_1]$ à 95% en posant :

$$\left[\overline{U}_m - 1,96\frac{\sigma}{\sqrt{m}}, \overline{U}_m + 1,96\frac{\sigma}{\sqrt{m}}\right]. \tag{1.3.1}$$

Cependant, il faut procéder à une estimation de σ, le théorème suivant en donne la formule.

Théorème 1.3.3. *Soit $(U_i, 1 \leq i \leq m)$ un échantillon de taille m d'une variable aléatoire U de carré intégrable. Notons $\overline{U}_m$ la moyenne empirique de cet échantillon, la variance empirique de l'échantillon est :*

$$\begin{aligned}\widehat{\sigma}_m^2 &= \frac{1}{m-1}\sum_{i=1}^{m}(U_i - \overline{U}_m)^2 \\ &= \frac{m}{m-1}\left(\frac{1}{m}\sum_{i=1}^{m}U_i^2 - \overline{U}_m^2\right).\end{aligned}$$

Alors $\widehat{\sigma}_m^2$ est un estimateur sans biais convergeant de σ^2, c'est-à-dire que $\mathbb{E}[\widehat{\sigma}_m^2] = \sigma_m^2$ et que la suite $\widehat{\sigma}_m^2$ converge presque sûrement vers σ^2 quand $m \to \infty$.

Ainsi, l'intervalle de confiance de $\mathbb{E}[U_1]$ à 95% devient :

$$\left[\overline{U}_m - 1,96\frac{\widehat{\sigma}_m}{\sqrt{m}}, \overline{U}_m + 1,96\frac{\widehat{\sigma}_m}{\sqrt{m}}\right],$$

et il est approximativement égal à (1.3.1) lorsque m est suffisament grand.

1.4. La régression logistique bayésienne

Comme pour la régression logistique vue dans la section 1.1, les paramètres β_i vont être estimés. Cependant, l'estimation se fait désormais avec l'estimateur bayésien (voir l'équation (1.2.2)). En effet, les β_i sont maintenant des variables aléatoires ayant pour loi *a priori* $\pi_i(\beta_i)$ et pour loi *a posteriori* $\pi_i(\beta_i|\mathbf{y})$. De plus, ils sont définis sur $\mathbb{R}$ et supposés indépendants de loi conjointe $\pi(\beta_0, ..., \beta_p)$. On rappelle que $\pi(\beta_0, ..., \beta_p)$ s'écrit :

$$\pi(\beta_0, ..., \beta_p) = \prod_{i=0}^{p}\pi_i(\beta_i).$$

L'estimateur bayésien $\tilde{\beta}_i$ s'écrit donc :

$$\begin{aligned}\tilde{\beta}_i &= \mathbb{E}_{\pi_i}[\beta|\mathbf{y}] \\ &= \int_{\mathbb{R}}\beta_i\pi_i(\beta_i|\mathbf{y})d\beta_i, \qquad (1.4.1)\end{aligned}$$

et la loi *a posteriori* $\pi_i(\beta_i|\mathbf{y})$ s'écrit encore comme :

$$\pi_i(\beta_i|\mathbf{y}) = \int_{D^p}\pi(\beta_0, ..., \beta_p|\mathbf{y})d\beta_0...d\beta_{i-1}d\beta_{i+1}...d\beta_p.$$

Soit f la fonction de densité marginale de $\mathbf{y}$. Alors, d'après l'équation (1.2.1), on déduit que :

$$\begin{aligned}\pi(\beta_0, ..., \beta_p|\mathbf{y}) &= \frac{L(\beta_0, ..., \beta_p|\mathbf{y})\pi(\beta_0, ..., \beta_p)}{f(\mathbf{y})} \\ &= \frac{L(\beta_0, ..., \beta_p|\mathbf{y})\pi(\beta_0, ..., \beta_p)}{\int_{D^{p+1}}L(\beta_0, ..., \beta_p|\mathbf{y})\pi(\beta_0, ..., \beta_p)d\beta_0...d\beta_p},\end{aligned}$$

où $L(\beta_0, ..., \beta_p|\mathbf{y})$ est la fonction de vraisemblance, $\mathbf{y}$ étant un vecteur de valeurs observées et par conséquent, $f(\mathbf{y})$ ne dépend pas des paramètres $\beta_0, \beta_1, ..., \beta_p$.

Ainsi, nous avons :

$$\pi_i(\beta_i|\mathbf{y}) = \frac{1}{f(\mathbf{y})} \int_{D^p} L(\beta_0, ..., \beta_p|\mathbf{y})\pi(\beta_0, ..., \beta_p)d\beta_0...d\beta_{i-1}d\beta_{i+1}...d\beta_p,$$

où il est rappelé que :

$$L(\beta_0, ...\beta_p|\mathbf{y}) = \prod_{i=1}^{N} \left(\frac{\mathrm{e}^{\beta_0+\sum_{j=1}^{p} \beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p} \beta_j x_{ij}}} \right)^{y_i} \left(\frac{1}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p} \beta_j x_{ij}}} \right)^{1-y_i}.$$

Donc finalement, l'estimateur bayésien $\tilde{\beta}_i$ s'écrit :

$$\begin{aligned}\tilde{\beta}_i &= \frac{1}{f(\mathbf{y})} \int_D \beta_i \left(\int_{D^p} L(\beta_0, ..., \beta_p|\mathbf{y})\pi(\beta_0, ...\beta_p)d\beta_0...d\beta_{i-1}d\beta_{i+1}...d\beta_p \right) d\beta_i \\ &= \frac{1}{f(\mathbf{y})} \int_{D^{p+1}} \beta_i L(\beta_0, ..., \beta_p|\mathbf{y})\pi(\beta_0, ...\beta_p)d\beta_0...d\beta_{i-1}d\beta_{i+1}...d\beta_p d\beta_i.\end{aligned}$$

L'estimateur $\tilde{\beta}_i$ est un rapport d'intégrales qui ne peut être résolu analytiquement. Pour trouver l'estimateur $\tilde{\beta}_i$, la méthode Monte-Carlo est utilisée. Les paramètres β_i sont générés selon leur loi *a priori* respective. Chacun des $p+1$ paramètres est généré m fois de manière à obtenir une approximation numérique :

$$\tilde{\beta}_i \approx \frac{\sum_{k=1}^{m} \beta_{i_{(k)}} L(\beta_{0_{(k)}}, ..., \beta_{p_{(k)}}|\mathbf{y})}{\sum_{k=1}^{m} L(\beta_{0_{(k)}}, ..., \beta_{p_{(k)}}|\mathbf{y})}. \tag{1.4.2}$$

Il est important de prendre une valeur de m assez grande pour que les paramètres obtenus soient convergents.

1.4.1. Exemple (suite)

Dans la sous-section 1.1.3, les paramètres $\boldsymbol{\beta}$ ont été estimés en utilisant la méthode de maximum de vraisemblance, les valeurs des paramètres sont les suivantes : $\boldsymbol{\beta} = (-5{,}50; 9{,}50)^t$ et nous avions obtenu $\widehat{\boldsymbol{\beta}} = (-7{,}46; 12{,}25)^t$. Maintenant, en reprenant le même exemple, les paramètres $\boldsymbol{\beta}$ vont être estimés en utilisant l'estimation bayésienne. On rappelle l'équation (1.4.1) :

$$\tilde{\beta}_i = \int_{\mathbb{R}} \beta_i \pi_i(\beta_i|\mathbf{y})d\beta_i.$$

Puis, il a été vu qu'il fallait passer par l'approximation numérique (1.4.2) pour approcher cette intégrale. On rappelle que L s'écrit :

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{8} \left(\frac{\mathrm{e}^{\beta_0+\beta_1 x_i}}{1+\mathrm{e}^{\beta_0+\beta_1 x_i}} \right)^{y_i} \left(\frac{\mathrm{e}^{\beta_0+\beta_1 x_i}}{1+\mathrm{e}^{\beta_0+\beta_1 x_i}} \right)^{1-y_i}.$$

Une loi *a priori* est choisie pour générer les observations $\beta_{0_{(k)}}$ et $\beta_{1_{(k)}}$. Nous décidons de prendre les lois suivantes : $\beta_0 \sim \mathscr{N}(0;25)$ et $\beta_1 \sim \mathscr{N}(0;25)$. Ces lois sont dites non informatives car le poids de l'information que porte la loi sur le paramètre à estimer est réduit dans l'inférence. Pour chacune de ces lois, 1000 observations sont générées en utilisant le logiciel R. En utilisant l'approximation (1.4.2), nous obtenons les valeurs estimées :

$$\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix} = \begin{pmatrix} -4{,}94 \\ 8{,}54 \end{pmatrix}.$$

Dans le but de savoir quelle est la meilleure méthode d'estimation, nous chercons à prédire cinq nouveaux points. La variable y représente la variable dépendante alors que $y_{\widehat{\boldsymbol{\beta}}}$ et $y_{\tilde{\boldsymbol{\beta}}}$ représentent respectivement les prédictions selon la régression logistique classique et la régression logistique bayésienne. La règle utilisée est la même pour $y_{\widehat{\boldsymbol{\beta}}}$ et $y_{\tilde{\boldsymbol{\beta}}}$. Prenons l'exemple de $y_{\widehat{\boldsymbol{\beta}}}$, si $\nu(x, \widehat{\boldsymbol{\beta}}) > 0{,}5$ alors $y_{\widehat{\boldsymbol{\beta}}} = 1$, sinon $y_{\widehat{\boldsymbol{\beta}}} = 0$. Ainsi, nous avons les prédictions suivantes :

x	0,44	0,47	0,58	0,62	0,77
y	0	0	1	1	1
$y_{\widehat{\boldsymbol{\beta}}}$	0	0	0	1	1
$y_{\tilde{\boldsymbol{\beta}}}$	0	0	1	1	1

Dans cet exemple, nous pouvons voir que la meilleure estimation de $\boldsymbol{\beta}$ est $\tilde{\boldsymbol{\beta}}$.

1.5. Les lois *a priori* utilisées

Dans le cadre de ce mémoire, trois lois *a priori* sont utilisées de manière non informative, c'est-à-dire que quel que soit le jeu de données en notre possession, la loi *a priori* va être utilisée sans changer ses paramètres. Une transformation de la loi bêta, la loi Cauchy et la loi normale sont utilisées pour estimer les paramètres des régressions logistiques bayésiennes.

Pour la loi normale, la densité *a priori* correspond à la loi normale centrée : $\mathscr{N}(0;25)$. Une grande variance est prise pour que les valeurs des paramètres générées lors de l'intégration par la méthode de Monte-Carlo puissent parcourir un large ensemble de valeurs.

Pour la transformation de la loi bêta, on prend la variable Ψ telle que $\Psi = g(\chi)$ où g est la fonction logit et χ suit une loi bêta $\mathscr{B}(a;b)$ avec $a, b > 0$. Les paramètres a et b sont choisis de manière à ce que la loi de Ψ soit d'espérance 0 et de variance 3. On rappelle que l'espérance et la variance de χ s'écrivent respectivement μ_χ et ${\sigma^2}_\chi$ avec :

$$\begin{cases} \mu_\chi &= \dfrac{a}{a+b} \\ {\sigma^2}_\chi &= \dfrac{ab}{(a+b)^2(a+b+1)}, \end{cases}$$

et :

$$\begin{aligned} \Psi &= g(\chi) \\ &= \text{logit}(\chi) \\ &= \log\left(\frac{\chi}{1-\chi}\right). \end{aligned}$$

Par la méthode delta, les moments d'ordre 1 et d'ordre 2 de Ψ sont estimés :

$$\begin{aligned} \mathbb{E}[\Psi] &= \mathbb{E}\left[g(\chi)\right] \\ &= \mathbb{E}\left[g(\mu_\chi + (\chi - \mu_\chi))\right] \\ &\approx \mathbb{E}\left[g(\mu_\chi) + g'(\mu_\chi)(\chi - \mu_\chi) + \frac{1}{2}g''(\mu_\chi)(\chi - \mu_\chi)^2\right]. \end{aligned}$$

Comme $\mathbb{E}[(\chi - \mu_\chi)] = 0$, nous obtenons :

$$\mathbb{E}[\Psi] \approx g(\mu_\chi) + \frac{1}{2}g''(\mu_\chi){\sigma^2}_\chi.$$

et

$$\begin{aligned} \text{var}[g(\chi)] &\approx (g'(\mathbb{E}[\chi])^2 \text{var}(\chi) \\ &\approx (g'(\mu_\chi))^2 {\sigma^2}_\chi. \end{aligned}$$

Sachant que $g(\chi) = \log\left(\dfrac{\chi}{1-\chi}\right)$, les dérivées première et seconde s'écrivent :

$$g'(\chi) = \frac{1}{\chi(1-\chi)}$$

et

$$g''(\chi) = \frac{2\,\chi - 1}{\chi^2(1-\chi)^2}.$$

Ainsi :

$$\begin{aligned}\mathbb{E}[g(\chi)] &\approx \log\left(\frac{\mu_\chi}{1-\mu_\chi}\right) + \frac{2\,\mu_\chi - 1}{2{\mu^2}_\chi(1-\mu_\chi)^2}{\sigma^2}_\chi \\ &\approx \log\left(\frac{a}{b}\right) + \frac{a^2-b^2}{2ab(a+b+1)},\end{aligned}$$

et

$$\begin{aligned}\mathrm{var}(g(\chi)) &\approx \frac{{\sigma^2}_\chi}{{\mu^2}_\chi(1-\mu_\chi)^2} \\ &\approx \frac{(a+b)^2}{ab(a+b+1)}.\end{aligned}$$

Sachant que :

$$\begin{cases} \mathbb{E}[g(\chi)] &= 0 \\ \mathrm{var}(g(\chi)) &= 3. \end{cases}$$

Ainsi, on pose le système suivant :

$$\begin{cases} \log\left(\frac{a}{b}\right) + \frac{a^2-b^2}{2ab(a+b+1)} = 0 \\ \frac{(a+b)^2}{ab(a+b+1)} = 3. \end{cases}$$

Puis, ne pouvant trouver de solution explicite à ce système d'équation (en utilisant la fonction *solver* du logiciel Maple), la première équation du système est simplifiée en retirant $\frac{1}{2}g''(\mu_\chi){\sigma^2}_\chi$ de l'approximation :

$$\begin{cases} \log\left(\frac{a}{b}\right) = 0 \\ \frac{(a+b)^2}{ab(a+b+1)} = 3. \end{cases}$$

Puis, en résolvant ce système, nous obtenons :

$$\begin{cases} a &= b \\ b &= \frac{1}{6}. \end{cases}$$

Pour la loi de Ψ, les observations sont générées avec la loi bêta $\mathscr{B}\left(\frac{1}{6};\frac{1}{6}\right)$ puis transformées via la fonction g.

La loi Cauchy s'écrit : $\mathscr{C}\,(a;b)$ où a est le paramètre de positon et b est le paramètre d'échelle. La loi *a priori* pour l'ordonnée à l'origine est une Cauchy avec $a = 0$

et $b = 10$ tandis que pour les autres coefficients $a = 0$ et $b = \frac{5}{2}$, tel qu'il est recommandé dans Gelman *et al.* (2008).
Ainsi, pour chaque simulation, les résultats obtenus à partir de ces trois différentes lois *a priori* vont être comparés.

Dans ce chapitre, nous avons présenté deux méthodes pour estimer les paramètres d'une régression. Du point de vue fréquentiste, nous avons présenté la méthode du maximum de vraisemblance qui a été approchée numériquement par la méthode de Newton-Raphson alors que du point de vue bayésien, les estimateurs bayésiens ponctuels ont été estimés avec une fonction de perte quadratique. Puis, les simulations Monte-Carlo ont été présentées pour approcher numériquement cette seconde méthode. Une fois les méthodes d'estimation de paramètres présentées, la régression logistique bayésienne a été introduite en fin de chapitre. Un exemple a été présenté afin d'illustrer l'estimation des paramètres d'un modèle de régression logistique. Dans le chapitre suivant, les modèles incrémentaux seront présentés, ils seront définis puis leur fonctionnement sera présenté.

Chapitre 2

MODÉLISATION INCRÉMENTALE

Dans ce chapitre est abordée la modélisation incrémentale. Cette modélisation statistique est surtout utilisée dans le domaine de la vente, de la relation client ou du marketing où l'étude des clients peut optimiser le retour sur investissement (voir Berry et Linoff (2011) et Chickering et Heckerman (2000)). Lorsqu'un nouveau produit est mis en vente, le service marketing d'une entreprise s'intéresse à l'efficacité de la publicité faite autour de ce produit. Autrement dit, la modélisation incrémentale s'intéresse à la différence S entre la probabilité d'achat du produit lorsque le prospect a été soumis à la publicité (groupe traitement T) et la probabilité d'achat lorsque le prospect n'a pas été soumis à la publicité (groupe contrôle C). Cette différence S est appelée l'incrément. Plus l'incrément S est grand, plus grand sera l'impact de la publicité sur le prospect. Soit Y_i la variable indépendante (achat ou non du produit) et $\mathbf{X}_i$ les variables indépendantes du prospect i. L'incrément S du prospect i s'écrit S_i :

$$S_i = \mathbb{P}_T\left(Y_i = 1|\mathbf{X}_i\right) - \mathbb{P}_C\left(Y_i = 1|\mathbf{X}_i\right). \tag{2.0.1}$$

Une fois la modélisation incrémentale présentée dans son ensemble, deux modèles incrémentaux spécifiques seront introduits : le modèle de Lo (2002) et le modèle de Lai (2004). Ces deux modèles cherchent à estimer l'incrément S de manière différente. Le modèle de Lo estime l'incrément S pour chaque prospect alors que le modèle de Lai estime l'incrément S pour des sous-groupes.

2.1. Qu'est-ce qu'un modèle incrémental ?

Dans le cadre d'une campagne de marketing, une entreprise commercialise un nouveau produit. Elle souhaite identifier les prospects qui pourraient acheter ce nouveau produit seulement s'ils sont contactés par l'entreprise.
Pour réaliser cette étude, un échantillon représentatif est extrait d'une population,

c'est-à-dire qu'un tirage aléatoire est effectué sur la liste complète des clients de la compagnie. Cet échantillon est scindé en deux : le groupe traitement T que l'entreprise va contacter via des courriels, des annonces par la poste ou par voie téléphonique et le groupe contrôle C avec lequel l'entreprise ne va pas prendre contact. Chaque prospect i appartient à un groupe de manière exclusive. Nous connaissons aussi ses caractéristiques $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^t$ qui peuvent être l'âge, le salaire annuel, la ville ou d'autres variables que l'entreprise possède sur le prospect et nous observons une variable réponse Y_i telle que :

$$Y_i = \begin{cases} 0, & \text{si le prospect } i \text{ ne répond positivement pas à l'offre ;} \\ 1, & \text{si le prospect } i \text{ répond positivement à l'offre.} \end{cases}$$

La modélisation incrémentale consiste à utiliser un modèle statistique pour identifier les prospects qui sont susceptibles de répondre favorablement à une offre commerciale seulement s'ils sont contactés par l'entreprise.
De l'équation (2.0.1), quatre classes de prospects sont distinguées selon l'incrément S et la probabilité $\mathbb{P}_C\left(Y_i = 1|\mathbf{X}_i\right)$:

- Les « décidés » sont les prospects qui répondent positivement sans avoir besoin d'être contactés par une offre marketing. Ils se distinguent par une probabilité $\mathbb{P}_C\left(Y_i = 1|\mathbf{X}_i\right)$ élevée et un incrément S faible.
- Les « non-décidés » sont les prospects qui répondent positivement uniquement quand ils sont contactés par une offre marketing. Ils se distinguent par une probabilité $\mathbb{P}_C\left(Y_i = 1|\mathbf{X}_i\right)$ faible et un incrément S positif élevé.
- Les causes perdues sont les prospects qui ne sont pas intéressés (qu'ils soient contactés ou non par l'offre marketing). Il se distinguent par une probabilité $\mathbb{P}_C\left(Y_i = 1|\mathbf{X}_i\right)$ et un incrément S faibles.
- Les prospects à ne pas déranger sont ceux qui ont une probabilité plus grande de répondre positivement lorsqu'ils ne sont pas contactés par l'offre marketing que quand ils le sont. Ils se distinguent par un incrément S négatif.

Les notions « élevé » ou « faible », dépendent de l'étude et de l'échantillon.

Tableau 2.1. Mesure de performance d'une campagne

	Traitement (exemple : offre par courriel)	Contrôle	Différence incrémentale
Échantillon ciblé par le modèle prédictif	A	B	A − B
Échantillon aléatoire	C	D	C − D
Modèle prédictif − échantillon aléatoire	A − C	B − D	(A − B) − (C − D)

Le tableau 2.1 illustre la performance d'un modèle lorsque l'offre est envoyée à un certain nombre de prospects. Pour maximiser le retour sur investissement, il serait préférable que le taux de réponse soit le plus élevé possible sur les prospects ciblés par l'offre (cas traitement). Les éléments A, B, C et D sont les taux de réponse positive associés à chacune des cellules. La première colonne correspond aux taux de réponse positive des prospects qui sont soumis à l'offre marketing (le cas traitement), la deuxième colonne correspond aux taux de réponse positives des prospects qui ne reçoivent pas l'offre (le cas contrôle) et la dernière colonne correspond à la différence des deux colonnes précédentes. À l'horizontale, il y a deux scénarios, le premier où l'offre est envoyée à des prospects ciblés par la modélisation prédictive et le second où l'offre est envoyée à l'échantillon aléatoire. La dernière ligne correspond à la différence entre les deux scénarios.
Dans la modélisation incrémentale, la présence du groupe contrôle est nécessaire pour mesurer l'impact de l'offre (ou de traitement) sur l'échantillon. En effet, les prospects qui reçoivent l'offre et ceux qui ne la reçoivent pas sont issus de la même population, par conséquent l'impact de l'offre sur les prospects est directement quantifiable.
Si l'offre est bonne, les prospects du groupe traitement devraient avoir un taux de réponse supérieur à celui des personnes n'ayant pas reçu d'offre :

$$A - B > 0$$

et

$$C - D > 0.$$

Si la modélisation est bien effectuée, son taux de réponse doit être supérieur lorsque les prospects sont ciblés, c'est-à-dire :

$$A - C > 0$$

et

$$B - D > 0.$$

Il est important de souligner que B et D doivent être supérieur à 0 pour que la modélisation incrémentale soit plus avantageuse que la modélisation prédictive traditionnelle. Sinon, il n'y a pas d'acheteurs volontaire, $B = 0$ et $D = 0$, le modèle incrémental revient à un modèle prédictif traditionnel.

Le modèle sera évalué comme bon si la différence entre les taux de réponse du groupe contrôle et du groupe traitement est supérieur lorsqu'il y a un ciblage (échantillon ciblé par le modèle) que lorsqu'il n'y en a pas (échantillon aléatoire), c'est-à-dire si $(A - B) - (C - D) > 0$. Plus grande sera cette différence, meilleur sera le modèle.

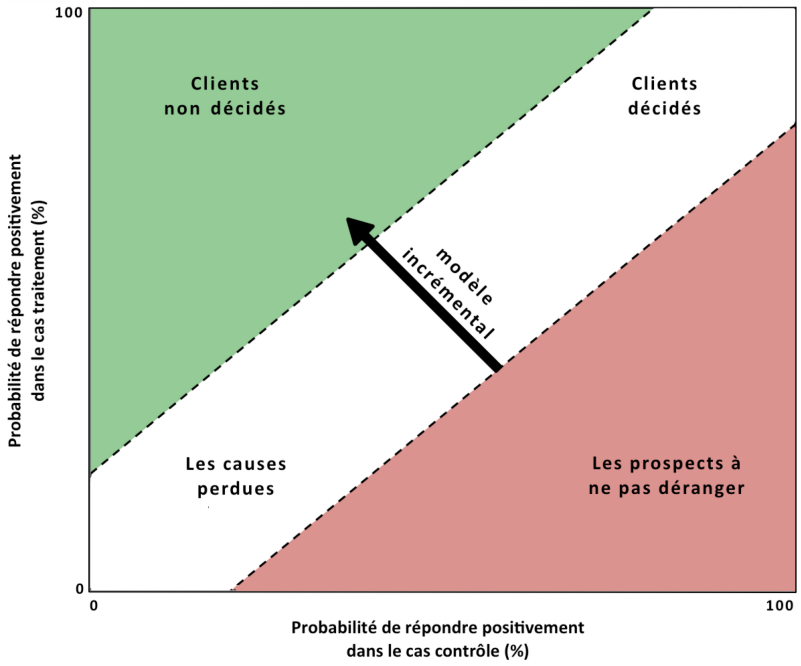


FIGURE 2.1. Disposition des classes

La figure 2.1 provenant de Radcliffe (2007), montre dans quelle zone se situent les différents types de prospects en comparant leur probabilité de répondre positivement lorsqu'ils sont supposés être dans le groupe traitement et lorsqu'ils sont supposés être dans le groupe contrôle. Seule la zone verte est intéressante car elle indique les prospects qui ont leur taux de réponse amélioré positivement lorsqu'ils

reçoivent l'offre marketing.

Ce mémoire va se concentrer plus particulièrement sur deux modélisations incrémentales qui permettent de repérer les prospects qui sont potentiellement des clients « non décidés ». La première est le modèle de Lo (2002) et la seconde est celle de Lai (2004).

2.2. Modèles de Lo (2002)

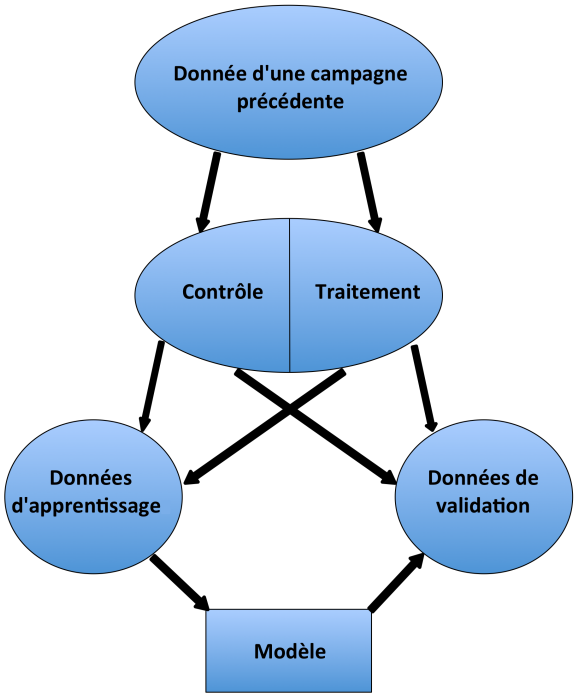


Figure 2.2. Méthodologie proposée par Lo (2002)

La figure 2.2 montre les différentes étapes de cette méthodologie. Lo s'est inspiré des modèles utilisés dans les essais cliniques (voir Almquist *et al.* (2001), Bellamy *et al.* (2007), Goetghebeur *et al.* (1997) et Van Belle *et al.* (2004)). Dans un premier temps, les données sont récupérées d'une campagne similaire à celle qui s'apprête à être menée, d'un côté, il y a les données représentant les prospects ayant été soumis à l'offre (traitement) et de le l'autre côté, ceux qui ne l'ont pas reçue (contrôle). À partir de ces deux groupes, deux nouveaux groupes de données sont formés : les données d'apprentissage et les données de validation. Ils sont tous les deux composés de données du groupe traitement et du groupe contrôle. Les données d'apprentissage et les données de validation doivent avoir un taux non significativement différent de données venant du groupe traitement et de

données venant du groupe contrôle. Dans la littérature se référant à l'exploration de données, il est souvent conseillé de prendre 70% de l'ensemble des données pour constituer les données d'apprentissage et les 30% restant pour les données de validation. On note N_a le nombre de prospects dans les données d'apprentissage et N_v le nombre de prospects dans les données de validation. N est le nombre total de prospects : $N = N_a + N_v$. Une fois ces deux sous-jeux de données constitués, la modélisation incrémentale est faite à partir des données d'apprentissage, puis son efficacité est testée sur les données de validation. Connaissant la réponse observée y des données de validation, le taux de bonne classification peut être connu et évalué sur une partie ou l'ensemble des données de validation. Les probabilités $\mathbb{P}_T(Y_i = 1|\mathbf{X}_i)$ et $\mathbb{P}_C(Y_i = 1|\mathbf{X}_i)$ peuvent encore s'écrire $\mathbb{E}_T(Y_i|\mathbf{X}_i)$ et $\mathbb{E}_C(Y_i|\mathbf{X}_i)$. On rappelle que :

$$\begin{aligned}\mathbb{E}(Y_i|\mathbf{X}_i) &= 1 \times \mathbb{P}(Y_i = 1|\mathbf{X}_i) + 0 \times \mathbb{P}(Y_i = 0|\mathbf{X}_i) \\ &= \mathbb{P}(Y_i = 1|\mathbf{X}_i).\end{aligned}$$

Ainsi, en se basant sur l'équation (2.0.1), l'incrément S_i du prospect i s'écrit encore :

$$S_i = \mathbb{E}_T(Y_i|\mathbf{X}_i) - \mathbb{E}_C(Y_i|\mathbf{X}_i), \tag{2.2.1}$$

et l'incrément estimé $\widehat{S}_i$ s'écrit :

$$\widehat{S}_i = \widehat{\mathbb{E}}_T(Y_i|\mathbf{X}_i) - \widehat{\mathbb{E}}_C(Y_i|\mathbf{X}_i),$$

où pour passer respectivement de $\mathbb{E}_T(Y_i|\mathbf{X}_i)$ et $\mathbb{E}_C(Y_i|\mathbf{X}_i)$ à $\widehat{\mathbb{E}}_T(Y_i|\mathbf{X}_i)$ et $\widehat{\mathbb{E}}_C(Y_i|\mathbf{X}_i)$, il suffit de remplacer les paramètres par leurs estimateurs.

Comme il a été vu dans le chapitre précédent, les espérances $\widehat{\mathbb{E}}_T(Y_i|\mathbf{X}_i)$ et $\widehat{\mathbb{E}}_C(Y_i|\mathbf{X}_i)$ sont calculées en modélisant une régression logistique :

$$\widehat{\mathbb{E}}_C(Y_i|\mathbf{X}_i) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \ldots + \widehat{\beta}_p X_{ip}}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \ldots + \widehat{\beta}_p X_{ip}}},$$

et

$$\widehat{\mathbb{E}}_T(Y_i|\mathbf{X}_i) = \frac{e^{(\widehat{\beta}_0 + \widehat{\gamma}_0) + (\widehat{\beta}_1 + \widehat{\gamma}_1) X_{i1} + \ldots + (\widehat{\beta}_p + \widehat{\gamma}_p) X_{ip}}}{1 + e^{(\widehat{\beta}_0 + \widehat{\gamma}_0) + (\widehat{\beta}_1 + \widehat{\gamma}_1) X_{i1} + \ldots + (\widehat{\beta}_p + \widehat{\gamma}_p) X_{ip}}},$$

où $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^t$ et $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0, \ldots, \widehat{\gamma}_p)^t$ sont des paramètres de régression logistique estimés.

Pour valider le modèle, les données de validation sont ordonnées par rapport à l'incrément estimé $\widehat{S}_i$ puis divisées en déciles cumulés k, pour $k \in \{1, \ldots, 10\}$. Plus $\widehat{S}_i$ est grand, plus le prospect sera situé dans les premiers déciles pour que son ciblage soit prioritaire. La notation c_k est utilisée pour représenter les k premiers déciles, autrement dit, le k^{e} décile cumulé.
Notons $\widehat{S}_{c_k}$, l'incrément moyen estimé pour les prospects du décile cumulé c_k :

$$\widehat{S}_{c_k} = \frac{1}{\text{Card}(c_k)} \sum_{i=1}^{N_v} \widehat{S}_i \mathbb{1}_{c_k}(y_i),$$

où $\text{Card}(c_k)$ est le nombre de prospects dans le décile cumulé c_k, et :

$$\mathbb{1}_{c_k}(y_i) = \begin{cases} 1, & \text{si le prospect } i \text{ est dans le décile cumulé } c_k; \\ 0, & \text{sinon.} \end{cases}$$

Au sein du modèle, le ciblage des prospects est donné par l'incrément estimé $\widehat{S}_i$. Ainsi, plus $\widehat{S}_i$ est proche de 1, plus le prospect a une probabilité élevée de répondre favorablement lorsqu'une offre lui est proposée (groupe traitement). Il en est de même pour l'incrément moyen estimé $\widehat{S}_{c_k}$, plus $\widehat{S}_{c_k}$ est proche de 1, plus grande sera la probabilité que les prospects qui composent le décile cumulé, répondent positivement à l'offre.

Dans chaque décile cumulé, la moyenne des réponses observées des prospects originellement du groupe contrôle $\overline{y}_{c_k,C}$ et celle des prospects originellement du groupe traitement $\overline{y}_{c_k,T}$ sont calculées séparément :

$$\overline{y}_{c_k,C} = \frac{1}{\text{Card}(c_k)} \sum_{i=1}^{N_v} y_i \mathbb{1}_{C \cap c_k}(y_i),$$

et

$$\overline{y}_{c_k,T} = \frac{1}{\text{Card}(c_k)} \sum_{i=1}^{N_v} y_i \mathbb{1}_{T \cap c_k}(y_i),$$

où

$$\mathbb{1}_{C \cap c_k}(y_i) = \begin{cases} 1, & \text{si le prospect } i \text{ est dans le groupe contrôle } C \\ & \text{et dans le décile cumulé } c_k\,; \\ 0, & \text{sinon.} \end{cases}$$

et

$$\mathbb{1}_{T\cap c_k}(y_i) = \begin{cases} 1, & \text{si le prospect } i \text{ est dans le groupe traitement } T \\ & \text{et dans le décile cumulé } c_k\,; \\ 0, & \text{sinon.} \end{cases}$$

Ainsi, $\bar{S}_{c_k}$, l'incrément moyen observé du décile cumulé c_k s'écrit :

$$\bar{S}_{c_k} = \overline{y}_{c_k,T} - \overline{y}_{c_k,C}.$$

Puis $\bar{S}_{c_k}$ et $\widehat{S}_{c_k}$ sont comparés sur un histogramme pour chacun des déciles cumulés. Si le modèle est bien choisi, $\bar{S}_{c_k}$ et $\widehat{S}_{c_k}$ devraient avoir des valeurs assez proches pour chaque décile.
Le ciblage aléatoire correspond à l'incrément moyen observé sur l'ensemble de l'échantillon, c'est-à-dire, $\bar{S}_{c_{10}}$. Pour évaluer la performance du modèle, l'incrément estimé cumulé $\widehat{S}_{c_k}$ est comparé au ciblage aléatoire. Si le modèle est bien constitué, l'incrément des deux ou trois premiers déciles cumulés devrait être bien supérieur à celui de l'ensemble de l'échantillon de validation.

2.3. Modèles de Lai (2004)

Lai (2004) s'intéresse aussi au fait que les réponses ne soient pas équilibrées, c'est-à-dire, qu'il y a beaucoup plus de réponses négatives que de réponse positive. C'est un sujet courant dans la plupart des problèmes d'apprentissage supervisé, voir Hansotia et Rukstales (2002). Cependant, Lai propose une méthodologie qui s'inspire de celle de Lo (2002). Comme dans la méthodologie précédente, les données d'apprentissage et les données de validation sont constituées de prospects provenant du groupe traitement et de prospects provenant du groupe contrôle. Seulement, un réarrangement des réponses observées est opéré sur les données d'apprentissage.

Tableau 2.2. Les réponses observées

	Réponses	
	Oui	**Non**
Traitement	Décidés + Non-décidés (1)	Pas intéressés (2)
Contrôle	Décidés (3)	Pas intéressés + Non-décidés (4)

Dans le but de regrouper les « non-décidés », les groupes (1) et (4) du tableau 2.2 sont réunis dans une nouvelle classe appelée « classe positive ». Les groupes (2) et

(3), qui rassemblent les prospects avec lesquels il n'est pas utile de communiquer, forment la « classe négative ». Ainsi, la variable V représente l'appartenance à la classe positive ou la classe négative, $\boldsymbol{v} = (v_1, \ldots, v_{N_a})$ sont les observations associées (on rappelle que N_a est le nombre de prospects dans les données d'apprentissage). L'observation $v_i,$ pour $i \in \{1, \ldots, N_a\}$ est telle que :

$$\begin{aligned} v_i &= \mathbb{1}_{\text{classe positive}}(y_i) \\ &= \begin{cases} 1, & \text{si le prospect } i \text{ est dans la classe positive ;} \\ 0, & \text{si le prospect } i \text{ est dans la classe negative.} \end{cases} \end{aligned}$$

Puis, sur les données d'apprentissage, est modélisée une régression logistique qui a pour but d'identifier les prospects de la classe positive. La probabilité estimée $\widehat{\mathbb{E}}(V_i|\mathbf{X}_i)$ s'écrit encore :

$$\widehat{\mathbb{E}}(V_i|\mathbf{X}_i) = \frac{\mathrm{e}^{\widehat{\alpha}_0+\widehat{\alpha}_1 X_{i1}+\ldots+\widehat{\alpha}_p X_{ip}}}{1+\mathrm{e}^{\widehat{\alpha}_0+\widehat{\alpha}_1 X_{i1}+\ldots+\widehat{\alpha}_p X_{ip}}}.$$

Une fois la régression modélisée, l'espérance $\mathbb{E}(V_i|\mathbf{X}_i)$ est calculée pour l'ensemble des données de validation. Ensuite, les données de validation sont divisées en décile selon l'espérance $\widehat{\mathbb{E}}(V_i|\mathbf{X}_i)$. Plus l'espérance est grande, plus le prospect est classé dans les premiers déciles.
Pour évaluer le modèle, l'incrément S va être estimé pour chaque décile c_k, S_{c_k} représente l'incrément sur la population du k-ième décile cumulé. Le taux de réponse observé $\overline{y}_{c_k,C}$ des prospects venant du groupe contrôle et le taux de réponse observées $\overline{y}_{c_k,T}$ des prospects venant du groupe traitement sont calculés pour le décile cumulé c_k. Ainsi, l'estimateur $\tilde{S}_{c_k}$ de S_{c_k} est estimé de la manière suivante :

$$\tilde{S}_{c_k} = \overline{y}_{c_k,T} - \overline{y}_{c_k,C},$$

où $\overline{y}_{c_k,T}$ et $\overline{y}_{c_k,C}$ estiment respectivement $\mathbb{P}_T(Y_i = 1|\mathbf{X}_i)$ et $\mathbb{P}_C(Y_i = 1|\mathbf{X}_i)$ pour le décile c_k. Sur l'ensemble de la population, nous avons :

$$\tilde{S} = \overline{y}_T - \overline{y}_C,$$

où $\tilde{S}$ revient encore à écrire $\tilde{S}_{c_{10}}$.
Il est important de souligner que $\tilde{S}_{c_k}$ et $\bar{S}_{c_k}$ (vu dans le modèle de Lo) s'estiment de la même manière mais ils ont généralement des valeurs différentes car l'ordre du ciblage et les prospects qui composent les déciles ne sont pas les mêmes.

Les prospects du groupe traitement et ceux du groupe contrôle sont issus de la même population, ainsi la différence entre le taux de réponse du groupe traitement $\overline{y}_T$ et celui du groupe contrôle $\overline{y}_C$ donne l'estimateur $\tilde{S}$ qui peut aussi être

interprété comme le taux de non-décidés qui répondent positivement dans la population.
Un bon modèle doit pouvoir cibler uniquement les prospects qui répondent positivement dans le cas traitement et qui ne répondraient pas dans le cas contrôle c'est-à-dire sans traitement. Les premiers déciles doivent avoir un incrément estimé $\tilde{S}_{c_k}$ nettement supérieur à l'incrément estimé $\tilde{S}$ sur l'échantillon. Dans l'idéal, lorsque le modèle cible bien les prospects, plus les individus se situent dans les premiers déciles, plus élevé sera l'incrément estimé $\tilde{S}_{c_k}$. On rappelle qu'un ciblage aléatoire sur l'ensemble de la population revient à avoir un incrément estimé $\tilde{S}$.

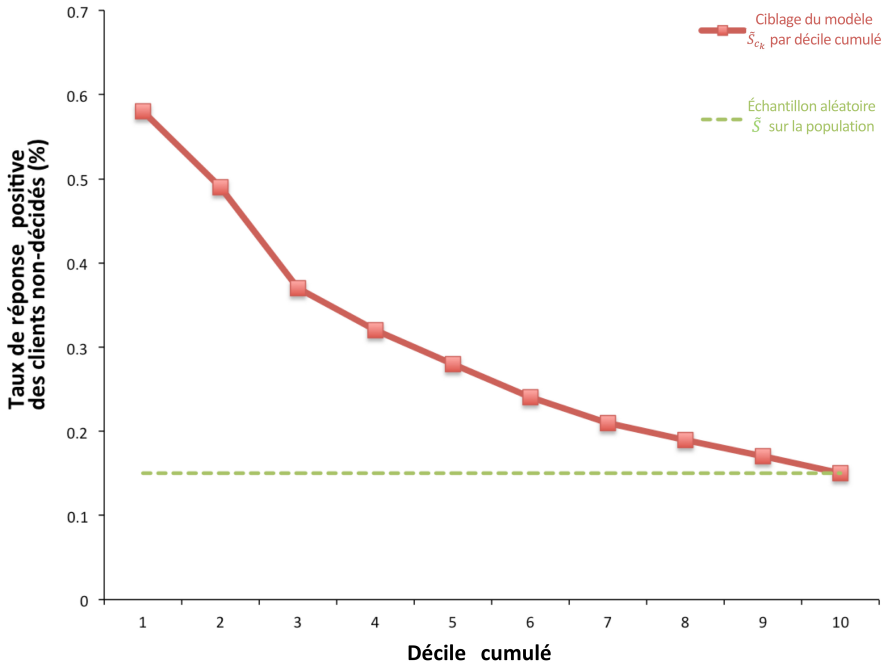


FIGURE 2.3. Valeur ajoutée du ciblage du modèle.

Ainsi, sur la figure 2.3 est résumée l'évaluation du modèle. Elle indique à quel point le modèle est pertinent pour cibler les « non-décidés ». En rouge est représenté l'incrément moyen estimé $\tilde{S}_{c_k}$ par décile cumulé. Il peut encore être vu comme le taux de réponse positive estimé des « non-décidés » par décile lorsqu'ils sont ciblés par le modèle (correspond aux taux de réponse A-B dans le tableau 2.1). Alors qu'en pointillé vert est représenté l'incrément moyen estimé $\tilde{S}$, il peut encore être vu comme le taux de réponse des « non-décidés » lorsque le ciblage est aléatoire (correspond aux taux de réponse C-D dans le tableau 2.1).

Dans ce chapitre, la modélisation incrémentale a été introduite ainsi que les deux modèles qui seront étudiés tout au long de ce mémoire, le modèle de Lo (2002) et le modèle de Lai (2004). Nous avons vu que le premier modèle estime directement S pour chaque prospect alors que le second passe par la modélisation d'une variable V appelée « classe positive » pour identifier les prospects « non-décidés », puis S est estimé par décile.
Dans le chapitre suivant vont être présentés les résultats obtenus pour chacun des deux modèles pour différentes simulations. Les modélisations seront effectuées avec une régression logistique classique (le GLM) ou une régression logistique bayésienne avec trois lois *a priori* différentes : la loi de bêta transformée, la loi Cauchy et la loi normale, respectivement appelées π_1, π_2 et π_3.

Chapitre 3

SIMULATION

Dans ce chapitre, cinq sections sont présentées, les deux premières sont consacrées aux paramètres des simulations et à la génération des données alors que les trois suivantes sont des simulations où les incréments moyens observés de Lo (2002) et de Lai (2004) sont comparés. Les paramètres de ces deux modèles sont estimés avec quatre méthodes (GLM, π_1, π_2 et π_3). Les simulations sont réalisées avec une puis deux variables indépendantes à titre d'exemples. Dans la dernière section, un total de 136 simulations sont réalisées avec des paramètres différents dans le but d'avoir des simulations plus exhaustives. Pour chacune des simulations présentées, les paramètres des modèles de régression utilisés vont être estimés et la capacité à bien prédire la variable réponse sera analysée. Enfin, les incréments moyens vont être estimés afin de distinguer les prospects ayant une plus grande probabilité de répondre favorablement à l'offre après avoir été contactés par l'entreprise.

3.1. Les paramètres de simulation

Les simulations sont faites selon un modèle (celui de Lo (2002) ou celui de Lai (2004)). Celui-ci est estimé d'un point de vu fréquentiste avec la méthode de maximum de vraisemblance (qui sera appelé GLM dans les graphiques) et d'un point de vue bayésien avec les trois lois *a priori* utilisées : loi bêta transformée, loi Cauchy et la loi normale. Elles seront respectivement appelées π_1, π_2 et π_3 dans les graphiques. Les simulations vont être effectuées en faisant varier le nombre total de prospects, le pourcentage de prospects et les taux de réponse positive dans le groupe contrôle et dans le groupe traitement.

— Le nombre total de prospects dans la simulation va varier dans le but de voir si de meilleurs résultats sont obtenus pour les différentes méthodes bayésiennes utilisées. Les résultats seront comparés à ceux obtenus avec la méthode fréquentiste. Les tailles d'échantillons N retenues sont les suivantes : $250, 500, 1\,000, 5\,000$ et $10\,000$.

— Les modèles incrémentaux ont nécessairement un groupe contrôle, cependant, dans la pratique, la taille du groupe contrôle est égale ou plus petite à celle du groupe traitement car l'entreprise veut envoyer son offre (traitement) au plus grand nombre possible de prospects ciblés. Le taux de prospects dans chacun des groupes va être modifié dans les cinq scénarios suivants :
  — Contrôle : 50% Traitement : 50% ;
  — Contrôle : 40% Traitement : 60% ;
  — Contrôle : 30% Traitement : 70% ;
  — Contrôle : 20% Traitement : 80% ;
  — Contrôle : 10% Traitement : 90%.

— Le taux de réponse positive dans le groupe contrôle et le groupe traitement est aussi modifié. Trois scénarios vont être distingués, le premier où la différence de taux de réponse est faible, un second où elle est moyenne et le dernier où elle est élevée :
  — Contrôle : 5% Traitement : 8% ;
  — Contrôle : 5% Traitement : 10% ;
  — Contrôle : 5% Traitement : 13%.

  Les taux de réponse sont pris assez faibles pour se rapprocher le plus possible des taux de réponse de direct marketing dans la réalité, voir Desmet (2005).

3.2. Génération des données

Les données du groupe contrôle et du groupe traitement sont générées comme dans Lo (2002). Les variables indépendantes sont générées de la manière suivante : $X_1 \sim \mathscr{N}(45; 13^2)$, $X_2 \sim \mathscr{N}(800 + 3\,X_1; 150^2)$ et $X_3 \sim \mathscr{N}(400 + 0{,}3\,X_2; 150^2)$. Puis la probabilité de répondre positivement est associée au prospect en fonction des variables X_1, X_2 et X_3. Si le prospect i est dans le groupe contrôle, la probabilité de répondre positivement est $\mathbb{E}_C(Y_i|\mathbf{X}_i)$, celle-ci est calculée comme il suit :

$$\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right) = \frac{\mathrm{e}^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}{1+\mathrm{e}^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}. \tag{3.2.1}$$

Si le prospect i est plutôt dans le groupe traitement, alors sa probabilité de répondre positivement est $\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right)$ et celle-ci s'écrit :

$$\begin{aligned}\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right) &= \frac{\mathrm{e}^{\beta_0+\gamma_0+(\beta_1+\gamma_1)X_{i1}+(\beta_2+\gamma_2)X_{i2}+(\beta_3+\gamma_3)X_{i3}}}{1+\mathrm{e}^{\beta_0+\gamma_0+(\beta_1+\gamma_1)X_{i1}+(\beta_2+\gamma_2)X_{i2}+(\beta_3+\gamma_3)X_{i3}}} \\ &= \frac{\mathrm{e}^{\zeta_0+\zeta_1 X_{i1}+\zeta_2 X_{i2}+\zeta_3 X_{i3}}}{1+\mathrm{e}^{\zeta_0+\zeta_1+X_{i1}+\zeta_2 X_{i2}+\zeta_3 X_{i3}}},\end{aligned} \tag{3.2.2}$$

où $\zeta_i = \beta_i + \gamma_i$. La valeur de ces paramètres sera spécifiée lors des exemples.

Les probabilités de répondre positivement $\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right)$ et $\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right)$ permettent de générer les valeurs de la variable dépendante Y :

$$Y \sim \begin{cases} Ber(\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right)), & \text{si le prospect } i \text{ est dans le groupe } C\,; \\ Ber(\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right)), & \text{si le prospect } i \text{ est dans le groupe } T. \end{cases}$$

Une fois les données générées, 30% des données sont sélectionnées aléatoirement afin de constituer les données de validation, le reste des données sert à constituer les données d'apprentissage qui permettent la modélisation des différentes méthodes d'estimation.

Deux exemples sont effectués dans les deux prochaines sections de ce chapitre. Le premier est modélisé avec une variable explicative et le second avec deux variables explicatives. Ces exemples permettent une meilleure visualisation et compréhension des méthodes utilisées.

3.3. Exemple avec une variable explicative

Dans un premier temps, une simulation est effectuée avec uniquement X_1 comme variable explicative. La variable Y est par conséquent générée avec une loi de Bernoulli dépendant de $\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right)$ ou de $\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right)$ (selon le groupe d'appartenance du prospect i) telles que :

$$\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right) = \frac{\mathrm{e}^{\beta_0+\beta_1 X_{i1}}}{1+\mathrm{e}^{\beta_0+\beta_1 X_{i1}}}, \tag{3.3.1}$$

et

$$\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right) = \frac{\mathrm{e}^{\zeta_0+\zeta_1 X_{i1}}}{1+\mathrm{e}^{\zeta_0+\zeta_1 X_{i1}}}. \tag{3.3.2}$$

Les données sont au nombre de 5 000, 70% de ces données (3 505) composent les données d'apprentissage et 30% (1 495) composent les données de validation. Les paramètres (β_0, β_1) d'une part et (ζ_0, ζ_1) d'autre part sont fixés de telle manière à ce que le taux de réponse des données dans le groupe contrôle soit de l'ordre de 5% et celui du groupe traitements soit de l'ordre de 8%. Ainsi $(\beta_0, \beta_1) = (-2{,}1; 2{,}7)$ et $(\zeta_0, \zeta_1) = (-2{,}1; 3{,}0)$, il est important de préciser que ces combinaisons ne sont pas uniques. D'autres combinaisons auraient pu mener au même taux de réponse. De plus, nous parlons « d'ordre » et non pas de valeur exacte car il y a du bruit dans la réponse des prospects.

Tableau 3.1. Description des données

	Données d'apprentissage	
	Nombre de prospects	Taux de réponses (%)
Groupe contrôle	1 087	5,02
Groupe traitement	2 418	8,04
	Données de validation	
	Nombre de prospects	Taux de réponses (%)
Groupe contrôle	443	4,98
Groupe traitement	1 052	8,02

Dans cet exemple, les données d'apprentissage qui servent aux modélisations des différentes méthodes de ciblage sont séparées, c'est-à-dire qu'il y a une valeur des variables explicatives qui scinde les réponses des prospects. En dessous de cette valeur les prospects répondent négativement et au-dessus, ils répondent positivement. Cette valeur est appelée le seuil, il existe dans cet exemple un seuil pour le groupe contrôle et un autre seuil pour le groupe traitement. Ces seuils peuvent être proches comme sur les figures 3.1. Sur la figure de droite sont représentées les réponses et les probabilités de répondre du groupe contrôle en fonction de X_1 et sur la figure de gauche, les réponses et les probabilités de répondre du groupe traitement en fonction de X_1. Les probabilités de répondre prennent l'allure de courbes croissantes alors que les réponses observées y prennent deux valeurs, 0 ou 1. La probabilité de répondre positivement est représentée en bleu clair pour le groupe contrôle et en rose pour le groupe traitement. Puis la réponse y des groupes contrôle et traitement sont respectivement représentées en bleu et en violet. La probabilité de répondre positivement qui correspond à

$\mathbb{E}_C(Y_i|\mathbf{X}_i)$ ou à $\mathbb{E}_T(Y_i|\mathbf{X}_i)$ (selon le groupe du prospect i) et la réponse observée y sont illustrées en fonction de la variable indépendante X_1 qui est centrée. Nous pouvons constater que les espérances $\mathbb{E}_C(Y_i|\mathbf{X}_i)$ et $\mathbb{E}_T(Y_i|\mathbf{X}_i)$ sont positivement corrélées à la variable X_1. Plus X_1 prend une valeur élevée, plus les prospects ont des chances de répondre positivement.

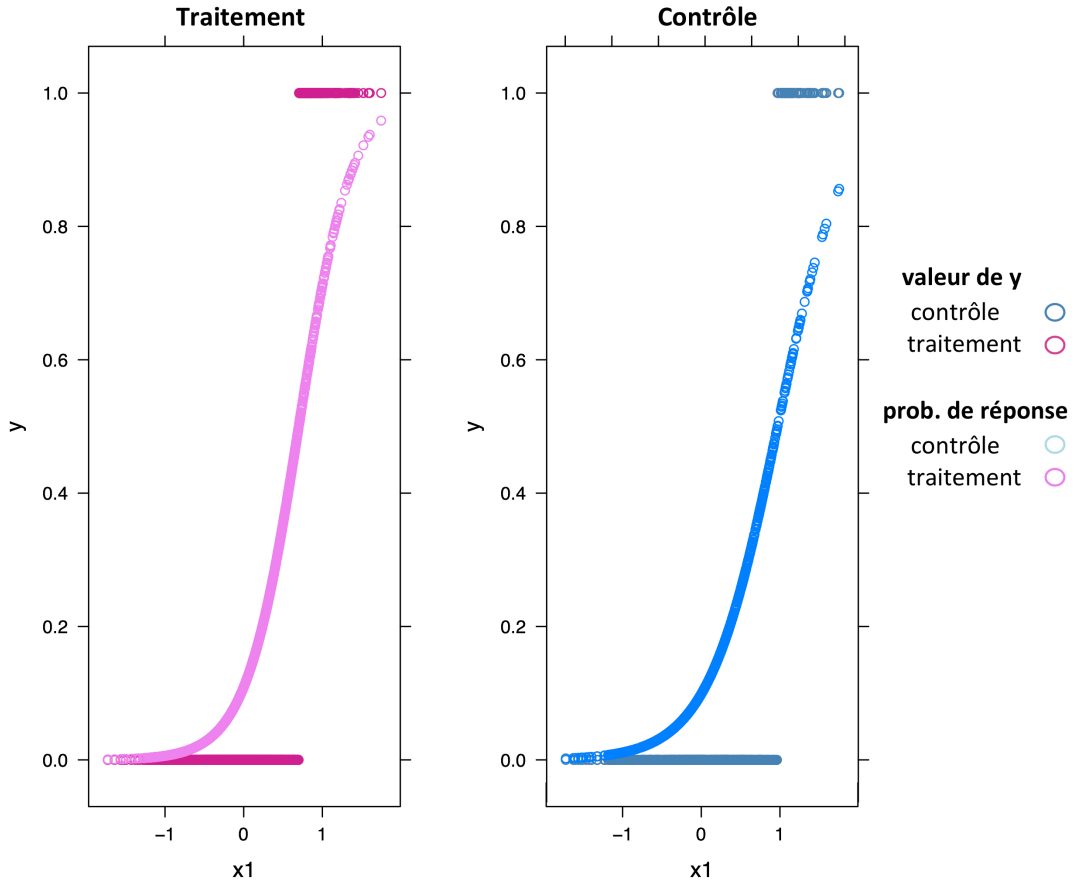


FIGURE 3.1. Réponses et probabilités de réponses des prospects en fonction de X_1.

3.3.1. Modèle de Lo

Dans le cas où les données sont séparées, il existe un problème de convergence de l'estimateur de maximum de vraisemblance. Ce problème de convergence est moins présent dans le cas où des estimateurs bayésiens sont utilisés. L'information apportée par la loi *a priori* permet une meilleure estimation des paramètres.

Cependant, comme la loi *a priori* joue un rôle plus important, cela a une conséquence sur le biais des estimateurs. Ce problème est étudié dans Albert et Anderson (1984). Les estimateurs obtenus par les quatre méthodes d'estimation utilisées sont biaisés mais les méthodes bayésiennes sont plus proches des vraies valeurs que le maximum de vraisemblance (GLM) :

TABLEAU 3.2. Paramètres estimés selon les différentes méthodes

L'estimation des paramètres

paramètres	vraies valeurs	GLM	π_1	π_2	π_3
β_0	−2,10	−2 812,00	−8,15	−9,59	−6,63
β_1	2,70	3 630,00	7,07	11,99	6,67
ζ_0	−2,10	−2 480,54	−8,83	−8,67	−6,90
ζ_1	3,00	3 539,48	8,38	12,41	7,09

Les écart-types des estimateurs

paramètres	GLM	π_1	π_2	π_3
β_0	3 968,82	0,41	1,23	1,13
β_1	5 125,59	0,34	1,96	1,70
ζ_0	4 711,36	1,23	3,59	4,22
ζ_1	6 276,64	1,62	3,69	4,16

On remarque dans le tableau 3.2 que les valeurs estimées sont toutes éloignées des vraies valeurs, cependant, pour chacune des méthodes, les estimateurs de β_0 et de ζ_0 sont toujours proches et l'estimateur de ζ_1 est légèrement supérieur à celui de β_1 pour toutes les méthodes bayésiennes. Les paramètres et les écart-types du GLM n'ont pas de sens cependant ils sont nécessaires pour l'évaluation des données de validation. Ils permettent de calculer des matrices de confusion, l'estimation des paramètres permet de prédire la réponse y des prospects. Les probabilités de réponses prédites du prospect i sont $\widehat{\mathbb{E}}_C\left(Y_i|\mathbf{X}_i\right)$ et $\widehat{\mathbb{E}}_T\left(Y_i|\mathbf{X}_i\right)$ respectivement pour le groupe contrôle et le groupe traitement. Il suffit de remplacer les paramètres par leurs estimateurs dans les équations (3.3.1) et (3.3.2) :

$$\widehat{\mathbb{E}}_C\left(Y_i|\mathbf{X}_i\right) = \frac{\mathrm{e}^{\widehat{\beta}_0+\widehat{\beta}_1 X_{i1}}}{1+\mathrm{e}^{\widehat{\beta}_0+\widehat{\beta}_1 X_{i1}}},$$

et

$$\widehat{\mathbb{E}}_T\left(Y_i|\mathbf{X}_i\right) = \frac{\mathrm{e}^{\widehat{\zeta}_0+\widehat{\zeta}_1 X_{i1}}}{1+\mathrm{e}^{\widehat{\zeta}_0+\widehat{\zeta}_1 X_{i1}}}.$$

Puis la réponse prédite $\widehat{y}_i$ du prospect i est telle que :

$$\widehat{y}_i = \begin{cases} 1, & \text{si } \widehat{\mathbb{E}}_C\left(Y_i|\mathbf{X}_i\right) \text{ ou } \widehat{\mathbb{E}}_T\left(Y_i|\mathbf{X}_i\right) \text{ est supérieur à } 0{,}5\,; \\ 0, & \text{sinon.} \end{cases}$$

Les réponses prédites $\widehat{y}$ par les modèles sont comparées aux réponses observées y. Les matrices de confusion sont calculées avec chacune des méthodes pour les données de validation. Les matrices de confusion sont présentées de la manière suivante :

		Val. réelle	
		0	1
Val. prédite	0	$a1$	$a2$
	1	$a3$	$a4$

où $a1$ et $a4$ représentent les prédictions qui coïncident avec les valeurs réelles alors que $a2$ et $a3$ représentent les mauvaises prédictions. La précision du modèle peut être calculée afin de connaître le pourcentage de bonne prédiction :

$$\text{Précision } := \frac{a1 + a4}{a1 + a2 + a3 + a4}.$$

Ainsi, pour chacune des méthodes nous obtenons les matrices de confusion suivantes. Pour chacune des méthodes, les prospects du groupe traitement sont séparés des prospects du groupe contrôle.

TABLEAU 3.3. Matrice de confusion avec le GLM

Groupe contrôle

		Val. réelle	
		0	1
Val. prédite	0	419	0
	1	0	24

Précision : 100%

Groupe traitement

		Val. réelle	
		0	1
Val. prédite	0	963	0
	1	0	89

Précision : 100%

TABLEAU 3.4. Matrice de confusion avec π_1 (loi bêta transformée *a priori*)

Groupe contrôle

		Val. réelle	
		0	1
Val. prédite	0	419	2
	1	0	22

Précision : 99,5%

Groupe traitement

		Val. réelle	
		0	1
Val. prédite	0	951	0
	1	12	89

Précision : $98,9\%$

TABLEAU 3.5. Matrice de confusion avec π_2 (loi Cauchy *a priori*)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	419	0
1	0	24

Précision :100%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	963	5
1	0	84

Précision : 99,5%

TABLEAU 3.6. Matrice de confusion avec π_3 (loi normale *a priori*)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	419	2
1	0	22

Précision : 99,5%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	963	0
1	0	89

Précision : 100%

On remarque que les quatre méthodes prédisent presque parfaitement la réponse observée y des prospects avec une prédiction bonne à plus de $98,9\%$ que ce soit pour le groupe traitement ou le groupe contrôle. Seulement, les matrices de confusion permettent de comparer les prédictions $\widehat{y}$ et les observations y mais elles ne permettent pas de s'intéresser directement aux taux de réponse des « non-décidés » et de cibler les prospects qui semblent intéressants. L'estimation des paramètres effectuée dans le tableau 3.2 permet de calculer l'incrément estimé $\widehat{S}$ et de classer les prospects en ordre décroissant par rapport à la valeur de l'incrément estimé $\widehat{S}$ pour chacune des quatre méthodes. Nous rappelons que $\widehat{S}$ s'écrit :

$$\widehat{S} = \widehat{\mathbb{E}}_T\left(Y_i|\mathbf{X}_i\right) - \widehat{\mathbb{E}}_C\left(Y_i|\mathbf{X}_i\right)$$

Dans un premier temps, l'incrément $\widehat{S}$ est estimé pour chacun des i prospects, avec les différentes méthodes et densités *a priori*. Par la suite, les données de validation sont divisées en déciles cumulés qui représentent l'incrément moyen estimé $\bar{S}_{c_k}$ où $k \in 1, 2, 3, \ldots, 10$. Pour le modèle de Lo, deux incréments moyens estimés sont distingués, $\bar{S}_{c_k,\text{obs}}$ et $\bar{S}_{c_k,\text{pred}}$. Le premier est basé sur y, la réponse observée du prospect alors que le second est basé sur $\widehat{y}$:

$$\bar{S}_{c_k,\text{obs}} = \overline{y}_{c_k,T} - \overline{y}_{c_k,C}$$

et

$$\bar{S}_{c_k,\text{pred}} = \overline{\widehat{y}}_{c_k,T} - \overline{\widehat{y}}_{c_k,C}.$$

Par abus de langage, nous allons tout simplement parler d'incrément moyen observé et d'incrément moyen prédit. La figure 3.2 représente l'incrément moyen observé par décile, pour chacun des modèles. Le premier décile est composé des prospects ayant les plus forts incréments, c'est-à-dire, les prospects que le modèle va cibler en premier lieu. Alors que par opposition, le dernier décile cumulé est composé de l'ensemble de la population, son incrément moyen $\bar{S}_{c_{10}}$, correspond au taux de réponse positive observé des « non-décidés » de l'échantillon.
Dans les études marketing, les offres sont envoyées à 30% des prospects ou moins, ce qui se traduit ici par le troisième décile.

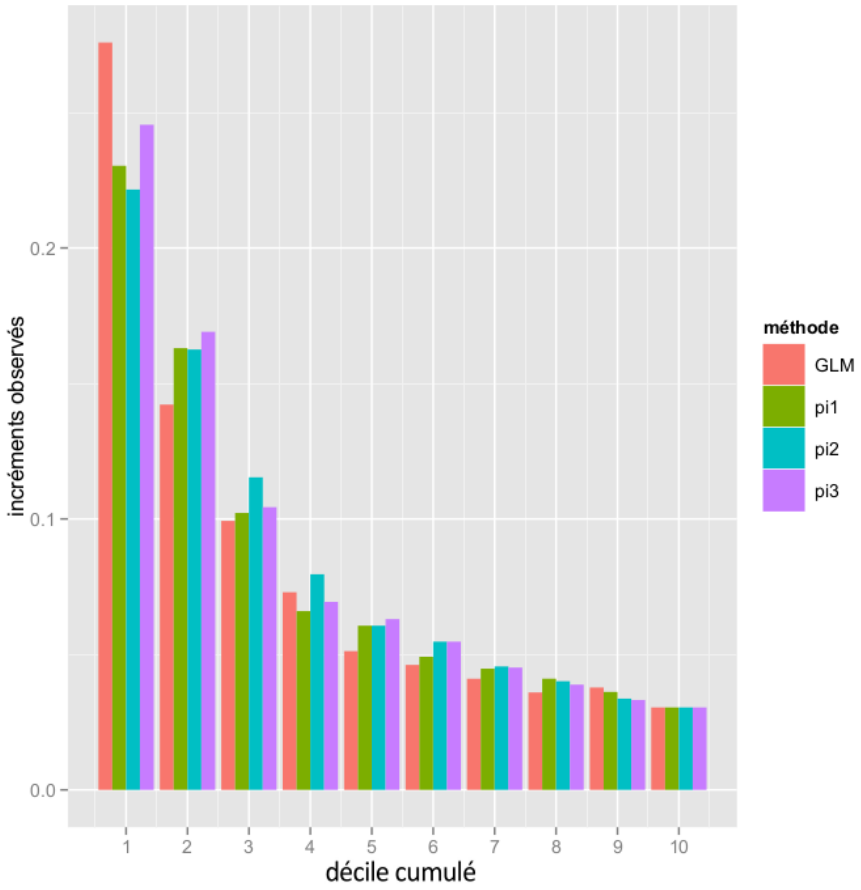


FIGURE 3.2. Incréments moyens observés par décile avec le modèle de Lo.

Nous observons que les quatre méthodes donnent des incréments qui sont assez proches, bien que les estimateurs obtenus soient assez différents (voir le tableau 3.7). Si le modèle choisi doit cibler uniquement 30% des prospects, le modèle π_2, c'est-à-dire, la méthode bayésienne avec la loi Cauchy comme loi *a priori* devrait

être choisi. Sur la figure 3.3 sont comparés les incréments moyens prédits et observés par décile.

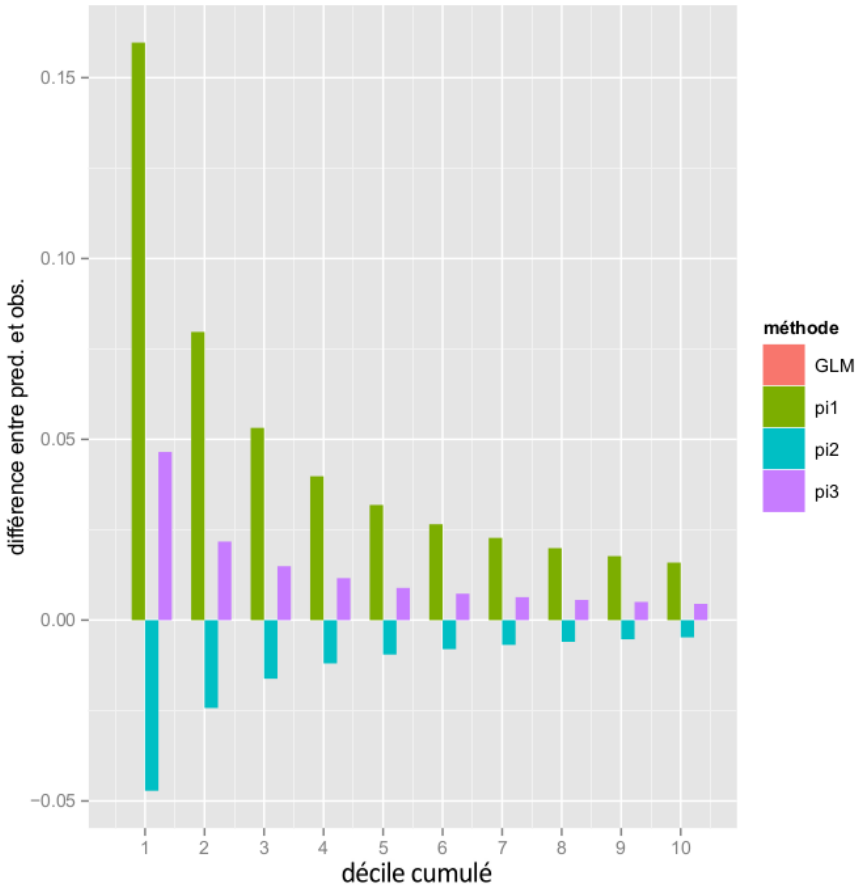


FIGURE 3.3. Différence des incréments moyens prédits et observés par décile.

Nous observons que π_1 et π_3 ont tendance à surévaluer les taux moyens d'incréments pour tous les déciles alors qu'inversement π_2 a tendance à sous-évaluer les taux moyens d'incréments pour tous les déciles. Le tableau 3.7 nous montre bien que le GLM donne une prédiction parfaite des incréments moyens pour cette simulation alors que les méthodes bayésiennes sont moins précises.

TABLEAU 3.7. Comparaison des incréments moyens prédits et observés avec le modèle de Lo.

	Prédiction				Observation			
	GLM	π_1	π_2	π_3	GLM	π_1	π_2	π_3
1	27,59	39,01	17,44	29,20	27,59	23,04	22,16	24,55
2	14,22	24,28	13,83	19,08	14,22	16,30	16,25	16,91
3	9,93	15,54	9,92	11,93	9,93	10,22	11,54	10,43
4	7,30	10,58	6,76	8,10	7,30	6,60	7,96	6,94
5	5,12	9,25	5,11	7,20	5,12	6,06	6,06	6,30
6	4,61	7,57	4,67	6,20	4,61	4,91	5,47	5,47
7	4,10	6,75	3,87	5,15	4,10	4,47	4,56	4,52
8	3,60	6,10	3,41	4,45	3,60	4,10	4,01	3,89
9	3,78	5,39	2,84	3,82	3,78	3,62	3,37	3,32
10	3,04	4,63	2,57	3,49	3,04	3,04	3,04	3,04

3.3.2. Modèle de Lai

Pour le modèle de Lai, la variable dépendante n'est plus Y mais V qui est telle que :

$$V = \begin{cases} 1, & \text{si le prospect est dans la classe positive ;} \\ 0, & \text{si le prospect est dans la classe negative.} \end{cases}$$

Par conséquent, la régression logistique modélisée a la même forme mais pas les mêmes paramètres que (3.2.1) et (3.2.2). Les paramètres $({\beta_0}^*, {\beta_1}^*)$ et $({\zeta_0}^*, {\zeta_1}^*)$ remplacent (β_0, β_1) et (ζ_0, ζ_1). Seulement, la valeur des vrais paramètres n'est pas connue puisque la probabilité de répondre positivement est générée avec les équations (3.2.1) et (3.2.2). Les paramètres obtenus sont donnés au tableau 3.8 :

TABLEAU 3.8. Paramètres estimés selon les différentes méthodes

L'estimation des paramètres

paramètres	GLM	π_1	π_2	π_3
β_0^*	2,23	2,64	1,35	2,21
β_1^*	−2,93	−0,15	−1,18	−2,89
ζ_0^*	−1,99	−0,22	−1,87	−1,99
ζ_1^*	2,91	0,86	1,08	2,89

Les écart-types des estimateurs

paramètres	GLM	π_1	π_2	π_3
β_0^*	0,12	1,02	0,13	0,12
β_1^*	0,15	1,24	0,54	0,15
ζ_0^*	0,14	1,19	0,37	0,14
ζ_1^*	0,22	0,64	0,12	0,22

Puis pour ce modèle, il n'y a pas d'intérêt à calculer les matrices de confusion ou comparer les incréments moyens prédits et observés puisque seules les réponses observées y peuvent être connues. Le seul critère comparable au modèle de Lo, est l'incrément moyen estimé par décile $\bar{S}_{c_k}$. Comme pour le modèle de Lo, l'estimation des paramètres permet d'estimer l'incrément par prospect puis l'incrément par décile, comme il est représenté sur la figure 3.4 :

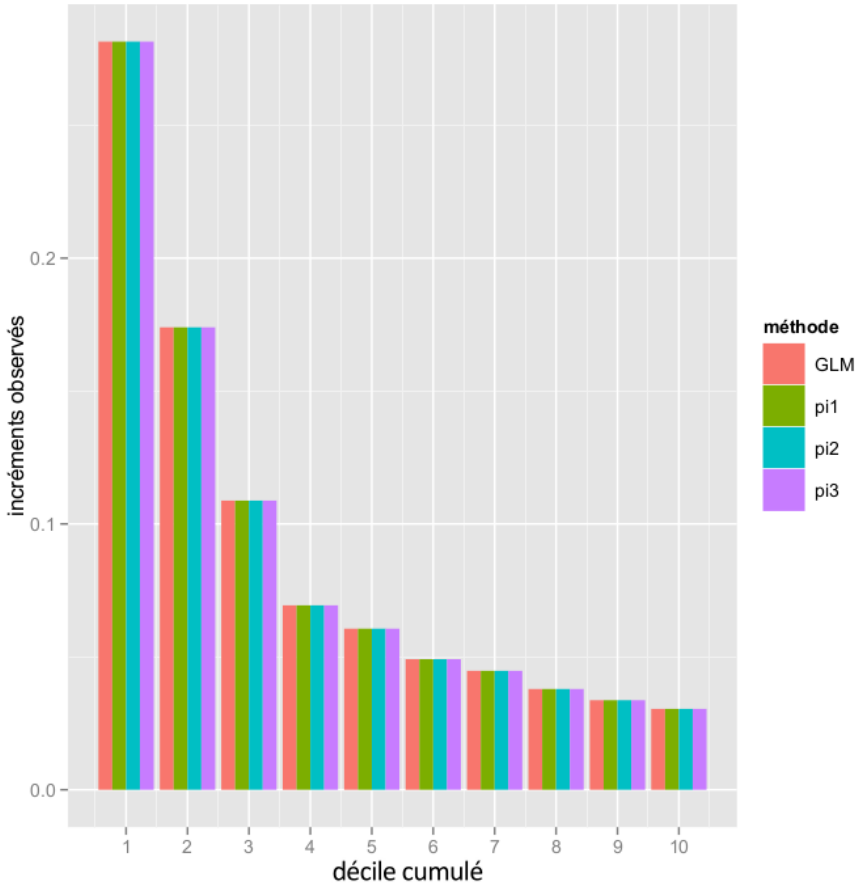


FIGURE 3.4. Incréments moyens observés par décile avec le modèle de Lai.

Sur la figure 3.4, nous observons que contrairement au modèle précédent, les taux de réponse sont les mêmes pour les quatre méthodes d'estimation pour cet exemple à une variable indépendante, ce qui n'est pas étonnant puisque les valeurs obtenues pour chacun des paramètres sont très proches avec chacune des méthodes.

3.3.3. Comparaison des modèles

Lorsque les taux de réponse estimés pour chacun des déciles sont comparés, nous obtenons le tableau 3.9 qui regroupe les incréments moyens obtenus avec le modèle de Lai et le modèle de Lo. Pour cette simulation, les quatre méthodes donnent les mêmes incréments pour le modèle de Lai. Par conséquent, les résultats suivant cette méthodologie sont tous regroupés sous la mention "Lai" :

TABLEAU 3.9. Récapitulatif des incréments moyens observés

		Lo			Lai
déciles	GLM	π_1	π_2	π_3	
1	27,59	23,04	22,16	24,55	28,15
2	14,22	16,30	16,25	16,91	17,40
3	9,93	10,22	11,54	10,43	10,88
4	7,30	6,60	7,96	6,94	6,94
5	5,12	6,06	6,06	6,30	6,06
6	4,61	4,91	5,47	5,47	4,91
7	4,10	4,47	4,56	4,52	4,47
8	3,60	4,10	4,01	3,89	3,79
9	3,78	3,62	3,37	3,32	3,37
10	3,04	3,04	3,04	3,04	3,04

On remarque qu'avec le modèle de Lai, les incréments sont meilleurs sur le premier et le second décile mais qu'au troisième décile, l'estimation effectuée avec π_2 (modèle de Lo) donne un meilleur incrément moyen. Comme nous nous concentrons sur le 3^{e} décile cumulé, le modèle de Lo avec la loi *a priori* Cauchy est choisi.

Un exemple avec les mêmes caractéristiques sur la population est effectué avec deux variables indépendantes.

3.4. Exemple avec deux variables explicatives

À présent, les simulations sont effectuées avec X_1 et X_2 comme variables explicatives. Les données sont générées avec les paramètres $\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right)$ et $\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right)$ tels que :

$$\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right) = \frac{\mathrm{e}^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}}}{1+\mathrm{e}^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}}} \tag{3.4.1}$$

et

$$\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right) = \frac{\mathrm{e}^{\zeta_0+\zeta_1 X_{i1}+\zeta_2 X_{i2}}}{1+\mathrm{e}^{\zeta_0+\zeta_1 X_{i1}+\zeta_2 X_{i2}}}, \tag{3.4.2}$$

où $(\beta_0, \beta_1, \beta_2) = (-3{,}2; 2{,}5; 2{,}6)$ et $(\zeta_0, \zeta_1, \zeta_2) = (-3{,}1; 2{,}9; 2{,}8)$.

Les données sont simulées avec les mêmes caractéristiques et les mêmes taux de réponse positive que dans l'exemple précédent.

Tableau 3.10. Description des données

	Données d'apprentissage	
	Nombre de prospects	Taux de réponses (%)
Groupe contrôle	1 054	4,96
Groupe traitement	2 460	8,03
	Données de validation	
	Nombre de prospects	Taux de réponses (%)
Groupe contrôle	424	4,95
Groupe traitement	1 062	8,10

Dans un premier temps, les données sont visualisées sur la figure 3.5, la variable Y est représentée en fonction des variables X_1 et X_2 qui sont centrées. Comme sur la figure 3.1 vue dans l'exemple précédent, la probabilité de répondre positivement des données d'apprentissage est représentée en rose pour le groupe traitement et en bleu clair pour le groupe contrôle, . On rappelle que la probabilité de répondre positivement correspond à $\mathbb{E}_C\left(Y_i|\mathbf{X}_i\right)$ ou à $\mathbb{E}_T\left(Y_i|\mathbf{X}_i\right)$ (selon le groupe du prospect i). Puis la réponse observée y du groupe traitement et du groupe contrôle sont respectivement représentées en violet et en bleu foncé.

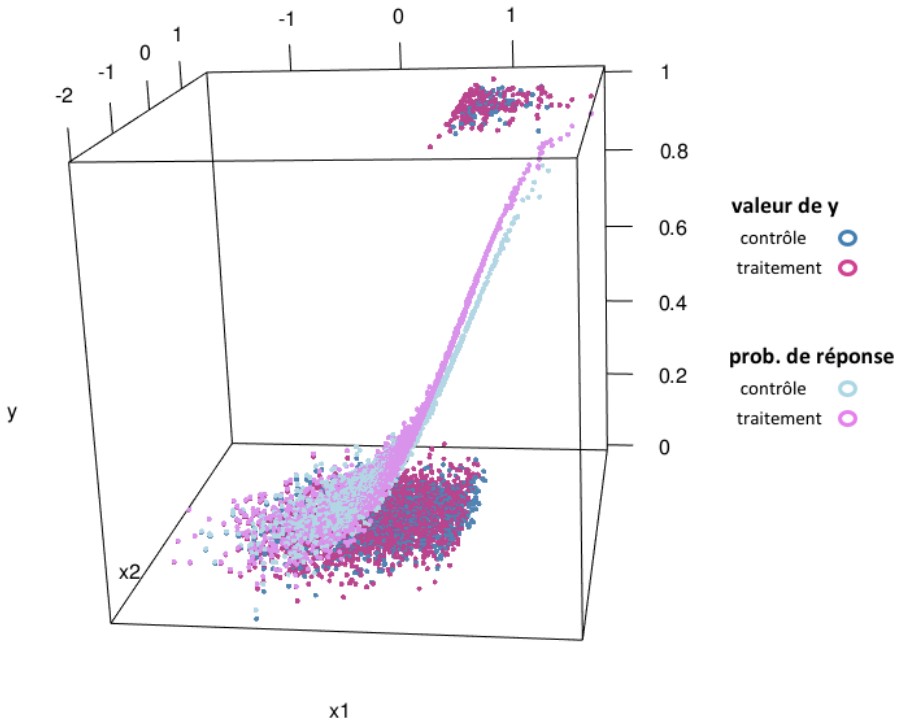


FIGURE 3.5. Réponses et probabilités de réponses des prospects en fonction de X_1 et X_2.

Nous observons sur la figure 3.5 que les données sont séparées par rapport à Y. Le seuil qui sépare les réponses 0 et 1 ne se modélise plus par une ligne mais par un hyperplan à partir duquel les valeurs observées y passent de 0 à 1. On remarque que ce plan est très proche pour les données contrôle et les données traitement mais que l'effet du traitement permet à des prospects de répondre positivement pour des valeurs x_1 et x_2 plus basses que celles des prospects du groupe contrôle.

3.4.1. Modèle de Lo

Comme dans l'exemple précédent, les données sont séparées. Par conséquent, les estimateurs obtenus sont biaisés pour les quatre méthodes, ils sont donnés dans le tableau 3.11 :

TABLEAU 3.11. Paramètres estimés selon les différentes méthodes

	L'estimation des paramètres				
paramètres	vraies val.	GLM	π_1	π_2	π_3
β_0	$-2{,}10$	$-2\,812{,}00$	$-8{,}15$	$-9{,}59$	$-3{,}36$
β_1	2,70	3 630,00	1,07	11,99	3,87
β_2	2,70	3 630,00	1,07	11,99	3,87
ζ_0	$-2{,}10$	$-2\,480{,}54$	$-8{,}83$	$-8{,}67$	$-3{,}90$
ζ_1	3,00	3 539,48	1,38	12,41	4,09
ζ_2	3,00	3 539,48	1,38	12,41	4,09

	Les écart-types des estimateurs			
paramètres	GLM	π_1	π_2	π_3
β_0	3 968,82	1,41	1,23	1,13
β_1	5 125,59	0,34	0,96	1,70
β_2	5 125,59	0,34	0,96	1,70
ζ_0	4 711,36	1,23	1,59	2,92
ζ_1	6 276,64	3,62	2,69	3,16
ζ_2	6 276,64	3,62	2,69	3,16

Une fois les paramètres estimés pour chacune des méthodes avec les données d'apprentissage, les matrices de confusion sont calculées pour chacune des méthodes avec les données de validation. Les matrices de confusion pour les prospects du groupe traitement sont séparées des matrices de confusion pour les prospects du groupe contrôle.

TABLEAU 3.12. Matrice de confusion avec le GLM

Groupe contrôle

		Val. réelle	
		0	1
Val. prédite	0	402	0
	1	1	21

Précision : 99,7%

Groupe traitement

		Val. réelle	
		0	1
Val. prédite	0	976	0
	1	0	86

Précision : 100%

TABLEAU 3.13. Matrice de confusion avec π_1 (loi bêta transformée *a priori*)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	401	5
1	2	16

Précision : $98,3\%$

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	973	2
1	3	84

Précision : $99,5\%$

TABLEAU 3.14. Matrice de confusion avec π_2 (loi Cauchy *a priori*)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	401	1
1	2	20

Précision :$99,3\%$

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	976	14
1	0	72

Précision : $98,7\%$

TABLEAU 3.15. Matrice de confusion avec π_3 (loi normale *a priori*)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	403	0
1	0	21

Précision : 100%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	976	0
1	0	86

Précision : 100%

Les matrices de confusion obtenues montrent que les modèles obtenus prédisent bien la réponse observée y à plus de $98,3\%$. Le modèle qui a la meilleure prédiction est π_3 avec 100% de bonnes prédictions pour les groupes traitement et contrôle.

Les incréments moyens observés sont par la suite estimés puis représentés sur un diagramme en barres en fonction des déciles sur la figure 3.6. Plus le prospect se situe dans le premier décile, plus ses chances d'être un « non-décidés » sont grandes.

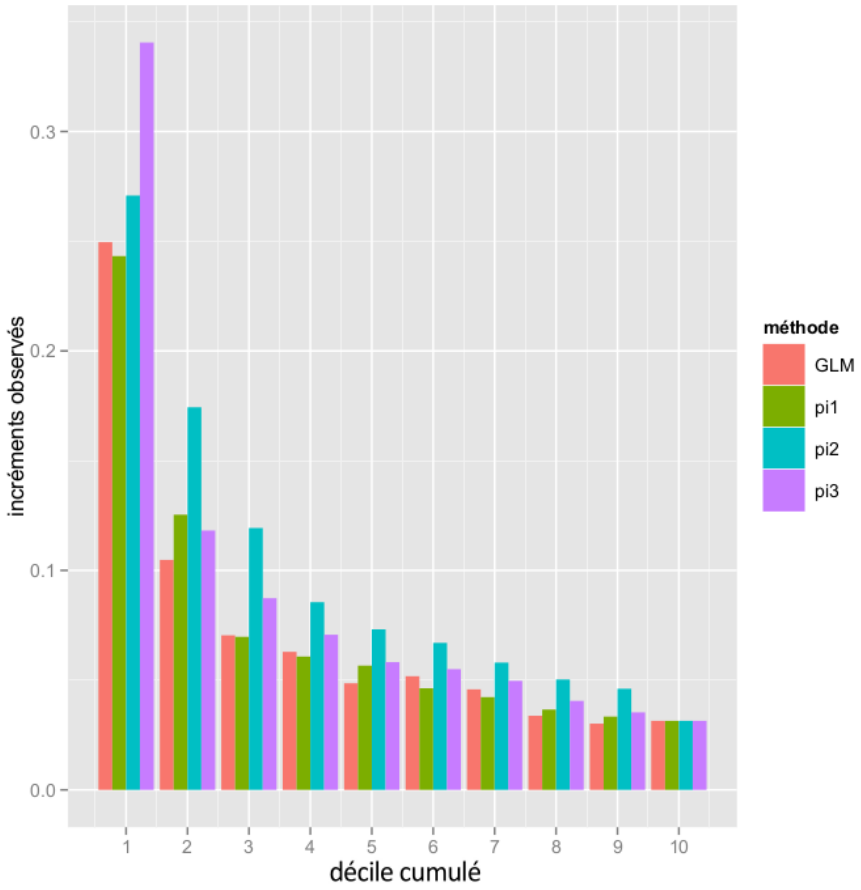


FIGURE 3.6. Incréments moyens observés par décile avec le modèle de Lo.

On remarque que sur les trois premiers déciles, les méthodes d'estimation bayésienne sont meilleures ou très proches de l'incrément obtenu avec le GLM. Sur le premier décile, π_3 est nettement meilleur mais pour les huit déciles suivants, π_2 est meilleur. La méthode π_2 devrait être choisie pour cibler cette population, au troisième décile, π_2 est la seule méthode à avoir un incrément moyen supérieur à 10%.

La différence entre l'incrément moyen prédit et l'incrément moyen observé est visible sur la figure 3.7. comme sur la figure 3.3 vue dans l'exemple précédent, nous remarquons que π_1 a tendance à surévaluer alors que π_2 a tendance à sous-évaluer la prédiction des incréments. Par contre le GLM et π_3 sont parfaitement prédits sur les neuf premiers déciles, il n'y a pas de différence entre les incréments moyens prédits et observés.

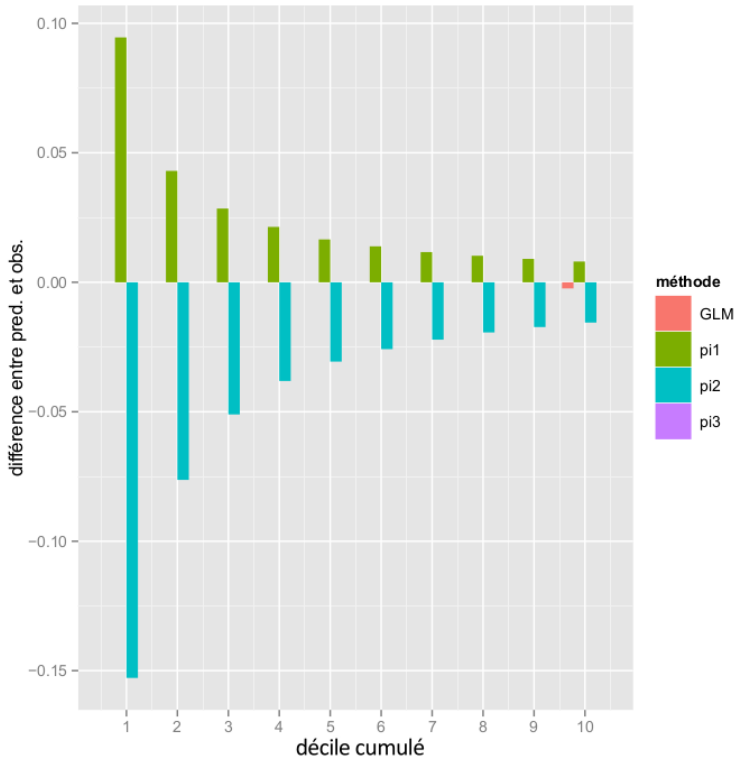


FIGURE 3.7. Différence des incréments moyens prédits et observés par déciles.

Puis le tableau 3.16 permet de confirmer ce qui a été aperçu sur la figure 3.7. Le GLM et π_3 donnent des prédictions fiables alors que π_1 et π_2 le sont beaucoup moins.

TABLEAU 3.16. Comparaison des incréments moyens prédits et observés avec le modèle de Lo

	Prédiction				Observation			
	GLM	π_1	π_2	π_3	GLM	π_1	π_2	π_3
1	24,96	33,78	11,81	34,06	24,96	24,32	27,08	34,06
2	10,48	16,84	9,80	11,82	10,48	12,54	17,43	11,82
3	7,04	9,82	6,83	8,73	7,04	6,97	11,93	8,73
4	6,29	8,21	4,74	7,07	6,29	6,08	8,55	7,07
5	4,86	7,32	4,25	5,82	4,86	5,66	7,31	5,82
6	5,18	6,02	4,12	5,50	5,18	4,63	6,70	5,50
7	4,58	5,39	3,59	4,97	4,58	4,23	5,80	4,97
8	3,38	4,69	3,09	4,06	3,38	3,66	5,03	4,06
9	3,02	4,24	2,88	3,53	3,02	3,34	4,61	3,53
10	2,91	3,95	1,59	3,15	3,15	3,15	3,15	3,15

3.4.2. Modèle de Lai

Pour le modèle de Lai, les paramètres estimés sont ceux de $({\beta_0}^*, {\beta_1}^*, {\beta_2}^*)$ et de $({\zeta_0}^*, {\zeta_1}^*, {\zeta_2}^*)$. Comme précédemment, les valeurs des vrais paramètres ne sont pas connues. Cependant, les valeurs des paramètres sont estimées dans le tableau 3.17.

TABLEAU 3.17. Paramètres estimés selon les différentes méthodes

L'estimation des paramètres

paramètres	GLM	π_1	π_2	π_3
β_0^*	$3,34$	$3,43$	$4,48$	$3,22$
β_1^*	$-2,44$	$-0,96$	$-1,25$	$-2,32$
β_2^*	$-2,90$	$-2,25$	$-1,73$	$-2,75$
ζ_0^*	$-3,09$	$-3,22$	$-3,06$	$-3,05$
ζ_1^*	$2,90$	$1,93$	$1,25$	$2,86$
ζ_2^*	$2,98$	$2,45$	$2,70$	$2,94$

Les écart-types des estimateurs

paramètres	GLM	π_1	π_2	π_3
β_0^*	0,21	1,08	0,54	0,20
β_1^*	0,29	0,93	0,87	0,27
β_2^*	0,31	0,62	1,01	0,30
ζ_{0^*}	0,24	1,72	0,53	0,23
ζ_1^*	0,34	1,01	0,36	0,33
ζ_2^*	0,37	0,78	0,54	0,35

Les paramètres estimés prennent encore une fois des valeurs qui sont proches les unes des autres. Ces paramètres estimés permettent de classer les prospects selon leur incrément estimé $\widehat{S}$ puis de les regrouper en déciles cumulés. Sur la figure 3.8, les incréments moyens observés sont représentés pour chacun des déciles et chacune des méthodes d'estimation.

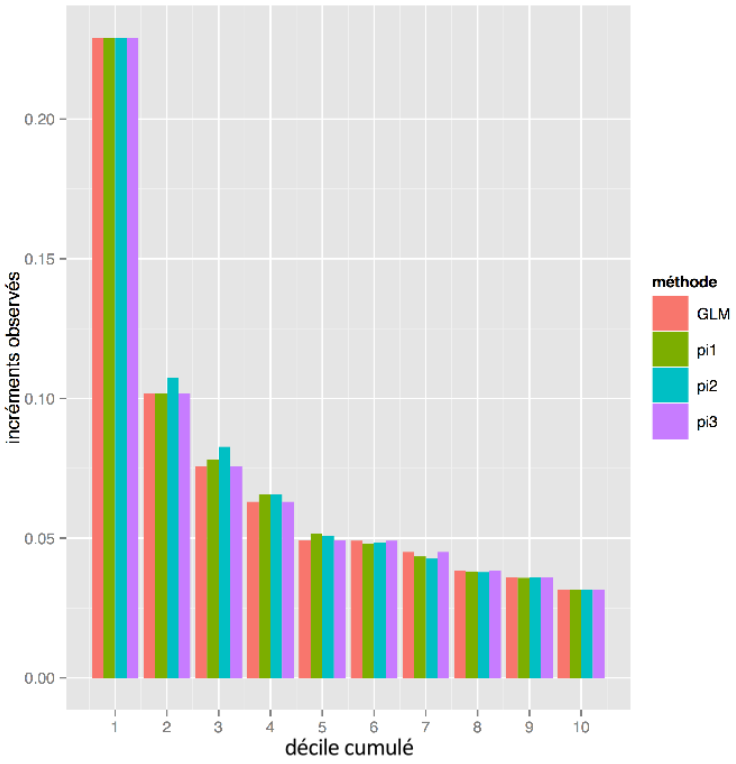


FIGURE 3.8. Incréments moyens observés par décile avec le modèle de Lai.

Nous remarquons que les quatre méthodes ont le même incrément sur le premier décile. Cependant, sur le second et troisième décile, la méthode π_2 se détache légèrement des autres méthodes bien qu'elles aient des incréments assez proches. Dans la section suivante les incréments obtenus avec cette méthode et celle de Lo sont comparés afin de désigner la méthode d'estimation qui permet d'avoir le meilleur incrément sur le troisième décile. Autrement dit, nous voulons savoir quelle est la méthode qui permet d'avoir les incréments les plus élevés sur 30% de l'échantillon.

3.4.3. Comparaison des modèles

Lorsque les incréments estimés sont comparés par rapport aux méthodes d'estimation de Lo et de Lai, nous obtenons le tableau 3.18.

TABLEAU 3.18. Récapitulatif des incréments moyens observés

	Lo				Lai			
	GLM	π_1	π_2	π_3	GLM	π_1	π_2	π_3
1	24,96	24,32	27,08	34,06	22,92	22,92	22,92	22,92
2	10,48	12,54	17,43	11,82	10,19	10,19	10,74	10,19
3	7,04	6,97	11,93	8,73	7,56	7,80	8,27	7,56
4	6,29	6,08	8,55	7,07	6,30	6,56	6,56	6,30
5	4,86	5,66	7,31	5,82	4,92	5,17	5,09	4,92
6	5,18	4,63	6,70	5,50	4,91	4,80	4,85	4,91
7	4,58	4,23	5,80	4,97	4,50	4,35	4,27	4,50
8	3,38	3,66	5,03	4,06	3,84	3,81	3,78	3,84
9	3,02	3,34	4,61	3,53	3,60	3,58	3,60	3,60
10	3,15	3,15	3,15	3,15	3,15	3,15	3,15	3,15

Nous observons sur le tableau 3.18 que le modèle de Lo est plus efficace sur cette simulation. Sur les trois premiers déciles, π_2 (modèle de Lo) a au moins 3 points de plus que l'estimation permettant d'obtenir les meilleurs incréments avec le modèle de Lai, c'est-à-dire, π_2 (modèle de Lai). Sur ce jeu de données, la méthode π_2 (modèle de Lo) permet au troisième décile d'espérer des incréments presque quatre fois supérieurs aux incréments de l'ensemble de la population.

3.5. Simulations plus exhaustives

Dans cette section, pour chaque simulation sont générés 100 jeux de données alors que dans les exemples présentés dans les sections 3.3 et 3.4, seul un jeu de données était généré. Par conséquent, les résultats présentés sont les moyennes des résultats obtenus pour les 100 jeux de données. L'avantage de travailler sur plusieurs jeux de données est que les moyennes des résultats sont moins variables. De plus, pour chaque loi *a priori* utilisée, 250 000 observations sont générées dans le but d'obtenir des estimateurs plus précis.

Dans un premier temps, les moyennes et les écart-types associés sont présentés pour le nombre de prospects et le taux de réponse des différents groupes. Dans un second temps, les résultats des méthodes Lo et Lai seront présentés puis dans un troisième temps ces résultats seront comparés.

Un ensemble de 136 simulations sont réalisées. Trois variables explicatives sont utilisées pour générer les probabilités de réponse positive des prospects. Ainsi, ces probabilités correspondent à l'équation (3.2.1) si le prospect est dans le groupe contrôle ou à l'équation (3.2.2) si le prospect est dans le groupe traitement.

Les paramètres utilisés pour générer les probabilités de réponse positive pour l'ensemble des simulations sont données au tableau 3.19.

Tableau 3.19. Valeur des paramètres selon le taux de réponse

paramètres	5 et 8%	5 et 10%	5 et 13%
β_0	$-3,8$	$-3,8$	$-3,8$
β_1	$2,7$	$2,7$	$2,7$
β_2	$2,7$	$2,7$	$2,7$
β_3	$1,0$	$1,0$	$1,0$
ζ_0	$-3,3$	$-3,0$	$-3,0$
ζ_1	$2,8$	$2,8$	$3,3$
ζ_2	$2,8$	$2,8$	$3,2$
ζ_3	$1,0$	$1,0$	$1,0$

N'étant pas en mesure de présenter les 136 simulations réalisées, seules celles ayant une taille du groupe contrôle de 10% et un taux de réponse de 5 et 8% respectivement dans les groupes contrôle et traitement sont présentées dans les sous-sections suivantes. Ces simulations sont représentatives du phénomène qu'on peut apercevoir sur l'ensemble des simulations. Les résultats des autres simulations sont présentés en annexes.

3.5.1. Description générale des jeux de données

Les jeux de données sont générés initialement avec 10 000 prospects qui vont être répartis dans les données d'apprentissage et les données de validation. Puis, les simulations réalisées avec moins de prospects correspondent à des simulations réalisées avec des échantillons imbriqués. Ainsi, les prospects présents dans l'étude à 250 prospects, se retrouvent dans celle à 500 prospects. De même que ceux présents dans l'étude à 500 prospects, se retrouvent dans celle à 1 000 prospects et ainsi de suite.

Le tableau 3.20 décrit les taux de réponse positive moyens et le nombre moyen de prospects par groupe pour les 100 jeux de données. Les écart-types associés aux moyennes sont mis entre parenthèses.
Nous observons que le taux de réponse positive pour les données d'apprentissage ou les données de validation varient faiblement d'une simulation à une autre.

Tableau 3.20. Moyenne (écart-type) des jeux de données

	Données d'apprentissage	
	Nombre moyen de prospects	taux de réponse positive moyen (%)
Groupe contrôle	699,3 (52,47)	5,05 (0,82)
Groupe traitement	6 293,5 (102,47)	8,30 (0,24)
	Données de validation	
	Nombre moyen de prospects	taux de réponse positive moyen (%)
Groupe contrôle	300,7 (61,42)	5,04 (0,98)
Groupe traitement	2 706,4 (138,06)	8,06 (0,32)

Une fois les taux de réponse moyens observés, nous nous intéressons aux deux modèles étudiés : le modèle de Lo et le modèle de Lai. Puis pour chacun des deux modèles, les simulations sont effectuées pour l'échantillon de taille 10 000 et pour les sous-échantillons imbriqués de taille 5 000, 1 000 et 500. En annexes, des résultats avec des échantillons de taille 250 sont présentés pour des simulations avec un groupe contrôle et des taux réponse positive pour le groupe traitement plus grands.

3.5.2. Modèle de Lo

Dans un premier temps, les moyennes et les écart-types des estimateurs des paramètres sont présentés dans les tableaux 3.21 et 3.22.
Nous observons qu'encore une fois, nous sommes en présence de données séparées. Les paramètres estimés par le GLM prennent des valeurs très éloignées des vrais paramètres. Les estimateurs bayésiens sont moins éloignés. On remarque que certains d'entre eux ont la vraie valeur du paramètre dans leur intervalle de confiance asymptotique à 95%. Comme la loi asymptotique est normale alors calculer l'intervalle de crédibilité revient à calculer l'intervalle de confiance à 95%, il est obtenu en posant le calcul suivant :

$$\widehat{\beta}_0 \pm 1{,}96\,\widehat{\sigma}_{\widehat{\beta}_0},$$

où $\widehat{\beta}_0$ est l'estimateur de β_0 et $\widehat{\sigma}_{\widehat{\beta}_0}$ est l'écart-type estimé de $\widehat{\beta}_0$.
Le calcul est le même pour les autres paramètres. Les estimateurs bayésiens donnent des estimateurs plus petits que ceux du GLM mais lorsque les matrices de confusion sont observées, nous remarquons que le GLM est tout aussi précis que les trois autres méthodes.

TABLEAU 3.21. Paramètres estimés selon les différentes méthodes

		L'estimation des paramètres			
	paramètres	vraies val.	GLM	π_1	π_2	π_3
$N = 500$	β_0	−3,8	−122,27	−12,88	−8,42	−6,04
	β_1	2,7	83,23	9,16	7,76	6,96
	β_2	2,7	87,42	9,77	7,70	7,49
	β_3	1,0	37,56	2,96	2,03	3,04
	ζ_0	−3,3	−880,49	−12,98	−8,14	−6,08
	ζ_1	2,8	751,94	10,37	8,97	7,48
	ζ_2	2,8	743,76	10,20	8,17	7,07
	ζ_3	1,0	274,84	5,33	4,38	4,30
$N = 1\,000$	β_0	−3,8	236,94	−7,39	−5,17	−4,28
	β_1	2,7	173,18	8,77	6,46	5,71
	β_2	2,7	147,85	8,95	6,57	5,07
	β_3	1,0	70,30	4,63	3,67	3,86
	ζ_0	−3,3	−1 616,28	−7,01	−6,29	−5,41
	ζ_1	2,8	1 362,69	6,52	5,75	5,38
	ζ_2	2,8	1 378,76	6,67	5,45	5,56
	ζ_3	1,0	478,89	3,60	3,96	3,49
$N = 5\,000$	β_0	−3,8	236,94	−5,39	−4,87	−4,28
	β_1	2,7	173,18	5,77	4,46	3,71
	β_2	2,7	147,85	5,95	4,57	3,73
	β_3	1,0	70,30	2,63	2,67	3,86
	ζ_0	−3,3	−1 616,28	−5,01	−5,29	−4,41
	ζ_1	2,8	3 899,64	5,01	4,32	4,79
	ζ_2	2,8	3 909,20	5,65	4,19	4,08
	ζ_3	1,0	1 393,11	3,22	3,19	2,44
$N = 10\,000$	β_0	−3,8	−613,54	−5,08	−4,39	−4,05
	β_1	2,7	442,48	4,14	3,67	3,34
	β_2	2,7	420,86	4,95	3,42	3,44
	β_3	1,0	172,08	2,83	2,75	3,69
	ζ_0	−3,3	−4 601,41	−4,85	−4,07	−4,04
	ζ_1	2,8	9 278,31	3,16	3,87	3,49
	ζ_2	2,8	9 279,42	3,43	3,42	3,54
	ζ_3	1,0	3 313,29	2,97	3,19	2,28

Puis les résultats des 100 matrices de confusion réalisées (chacune étant issue d'un jeu de données) sont résumés dans une matrice de confusion moyenne où une classification moyenne et son écart-type sont présentés dans chacune des cellules.

TABLEAU 3.22. Les écart-types des paramètres estimés

		Les écart-types des estimateurs			
	paramètres	GLM	π_1	π_2	π_3
$N = 500$	β_0	485,83	4,07	3,46	3,49
	β_1	344,60	5,16	4,60	3,93
	β_2	348,41	5,46	3,54	4,02
	β_3	135,78	5,85	3,20	2,02
	ζ_0	1 688,21	3,95	3,78	3,07
	ζ_1	5 589,36	6,57	3,04	4,76
	ζ_2	5 484,07	4,68	2,95	4,73
	ζ_3	1 608,40	3,47	3,83	1,67
$N = 1\,000$	β_0	185,52	3,55	1,55	2,86
	β_1	130,06	2,06	3,37	2,67
	β_2	144,58	3,81	2,97	3,15
	β_3	77,02	3,08	2,10	1,24
	ζ_0	746,62	2,01	2,01	1,62
	ζ_1	1 619,98	2,48	2,18	2,76
	ζ_2	1 617,74	2,67	1,77	3,06
	ζ_3	579,66	2,06	1,87	1,08
$N = 5\,000$	β_0	180,48	1,90	1,22	0,76
	β_1	135,90	1,81	1,11	1,01
	β_2	128,65	1,98	1,48	1,14
	β_3	46,83	1,74	1,61	0,42
	ζ_0	706,55	1,65	1,84	0,71
	ζ_1	582,22	1,87	1,91	0,72
	ζ_2	610,33	1,87	1,69	0,68
	ζ_3	213,53	1,73	1,56	0,26
$N = 10\,000$	β_0	78,00	0,78	1,18	0,65
	β_1	58,07	0,86	0,87	0,79
	β_2	68,44	0,99	1,09	0,74
	β_3	49,59	1,31	1,07	0,36
	ζ_0	408,11	1,64	1,62	0,86
	ζ_1	356,75	1,21	1,14	0,76
	ζ_2	335,68	1,22	1,64	0,68
	ζ_3	139,09	1,66	1,46	0,23

Les matrices de confusion moyennes présentées sont effectuées pour les simulations avec 500 et 10 000 prospects. Les matrices de confusion pour les échantillons de taille 1 000 et 5 000 sont mises en annexes. Le nombre de prospects représenté dans

les matrices de confusion correspond au nombre de prospects dans les données de validation.

Dans un premier temps les matrices de confusion pour les simulations de 500 prospects sont présentées dans les tableaux 3.23 à 3.26.

TABLEAU 3.23. Moyennes (écart-types) des classifications des matrices de confusion avec le GLM ($N = 500$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	14,56 (0,62)	0,11 (0,32)
1	0,11 (0,32)	0,89 (0,32)

Précision : 98,60%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	123,56 (0,70)	0,28 (0,57)
1	0,28 (0,57)	10,72 (0,57)

Précision : 99,58%

TABLEAU 3.24. Moyennes (écart-types) des classifications des matrices de confusion avec π_1 (loi bêta transformée *a priori*) ($N = 500$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	14,83 (0,38)	0,06 (0,24)
1	0,06 (0,24)	0,94 (0,24)

Précision : 99,24%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	123,39 (1,04)	0,28 (0,99)
1	0,28 (0,99)	9,83 (0,99)

Précision : 98,27%

TABLEAU 3.25. Moyennes (écart-types) des classifications des matrices de confusion avec π_2 (loi Cauchy *a priori*) ($N = 500$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	14,83 (0,51)	0,17 (0,38)
1	0,17 (0,38)	0,83 (0,38)

Précision :97,87%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	123,22 (1,06)	1,17 (1,04)
1	1,17 (1,04)	9,83 (1,04)

Précision : 98,27%

TABLEAU 3.26. Moyennes (écart-types) des classifications des matrices de confusion avec π_3 (loi normale *a priori*) ($N = 500$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	14,89 (0,32)	0,11 (0,32)
1	0,11 (0,32)	0,89 (0,32)

Précision : 98,62%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	123,50 (0,86)	0,72 (0,96)
1	0,72 (0,96)	10,28 (0,96)

Précision : 98,93%

Dans un second temps les matrices de confusion pour les simulations de 10 000 prospects sont présentées dans les tableaux 3.27 à 3.30.

TABLEAU 3.27. Moyennes (écart-types) des classifications des matrices de confusion avec le GLM ($N = 10\,000$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	280,81 (17,84)	0,24 (0,44)
1	0,24 (0,44)	14,81 (0,87)

Précision : 99,84%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	2 497,62 (47,66)	0,19 (0,51)
1	0,19 (6,57)	219,29 (0,51)

Précision : 99,99%

TABLEAU 3.28. Moyennes (écart-types) des classifications des matrices de confusion avec π_1 (loi bêta transformée *a priori*) ($N = 10\,000$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	277,29 (18,30)	4,57 (3,67)
1	4,57 (3,67)	10,48 (4,11)

Précision : 96,92%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	2488,52 (46,48)	201,29 (16,40)
1	18,19 (13,53)	201,29 (16,40)

Précision : 98,67%

TABLEAU 3.29. Moyennes (écart-types) des classifications des matrices de confusion avec π_2 (loi Cauchy *a priori*) ($N = 10\,000$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	275,33 (15,91)	4,81 (4,88)
1	4,81 (4,88)	10,24 (5,17)

Précision :96,74%

Groupe traitement

Val. prédite \ Val. réelle	0	1
0	2488,76 (52,29)	26,57 (21,22)
1	26,57 (21,22)	192,90 (20,14)

Précision : 98,05%

TABLEAU 3.30. Moyennes (écart-types) des classifications des matrices de confusion avec π_3 (loi normale *a priori*) ($N = 10\,000$)

Groupe contrôle

Val. prédite \ Val. réelle	0	1
0	281,10 (17,78)	0,86 (1,01)
1	0,86 (1,01)	14,19 (1,50)

Précision : 99,42%

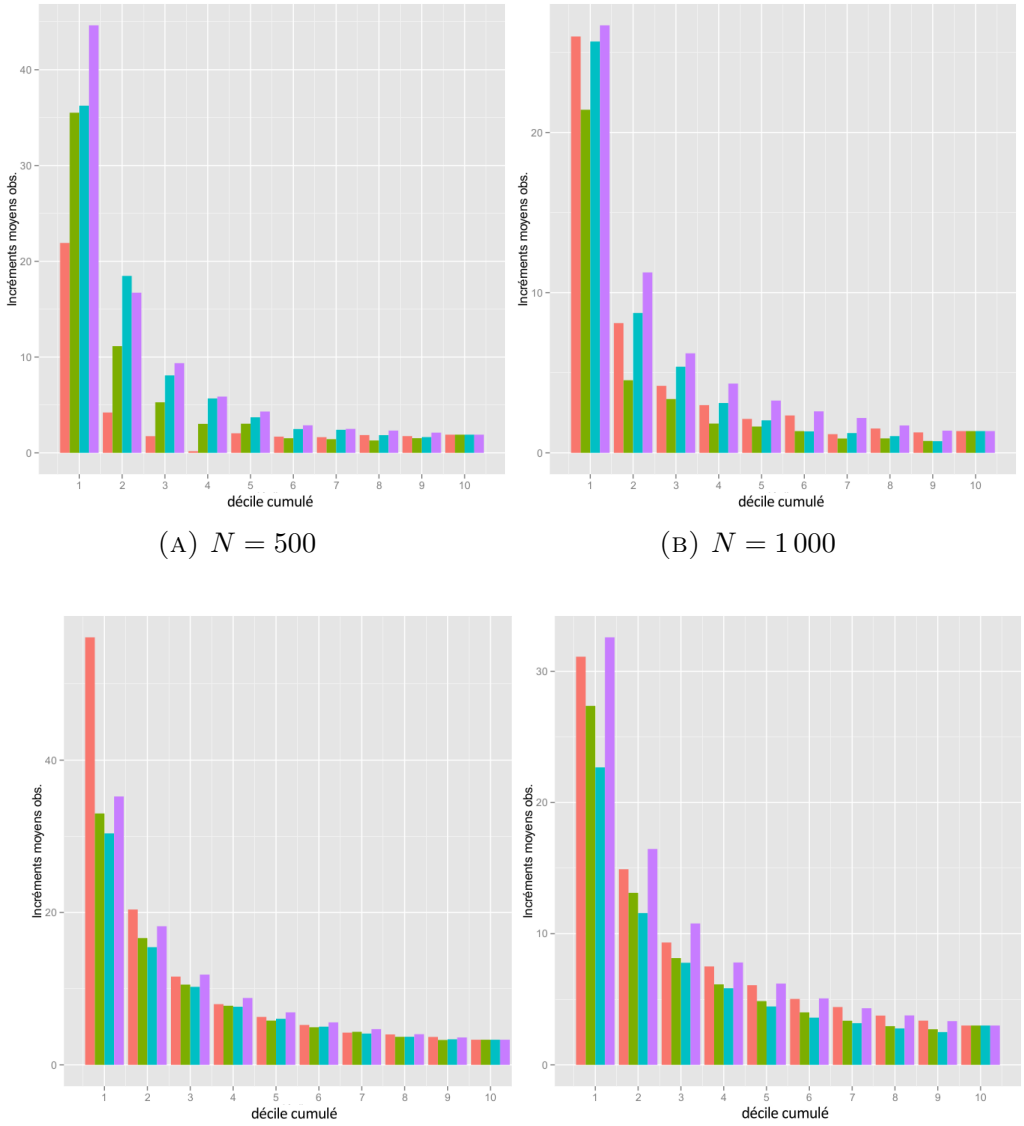
Groupe traitement

Val. prédite \ Val. réelle	0	1
0	2497,14 (47,59)	1,52 (1,57)
1	1,52 (1,57)	217,95 (6,89)

Précision : 99,89%

Nous observons que les matrices de confusion ont un taux de précision élevé pour toutes les méthodes. Pour l'échantillon de 500 prospects, la précision est supérieure à 97%. Les méthodes ont des taux de précision très proches. Pour l'échantillon de 10 000 prospects, la précision est supérieure à 96% pour l'ensemble des méthodes. Cependant, le GLM et π_3 ont des taux de précision supérieurs à

99, 4%. Dans l'ensemble des matrices de confusion observées, nous remarquons que les différents modèles ont une bonne capacité à détecter la classe rare, c'est-à-dire, les prospects qui répondent positivement. Ensuite, les incréments moyens sont estimés par décile pour chacune des méthodes utilisées. Ils sont présentés sur l'ensemble de la figure 3.9. Les figures associées aux simulations sont présentées du plus petit échantillon au plus grand.



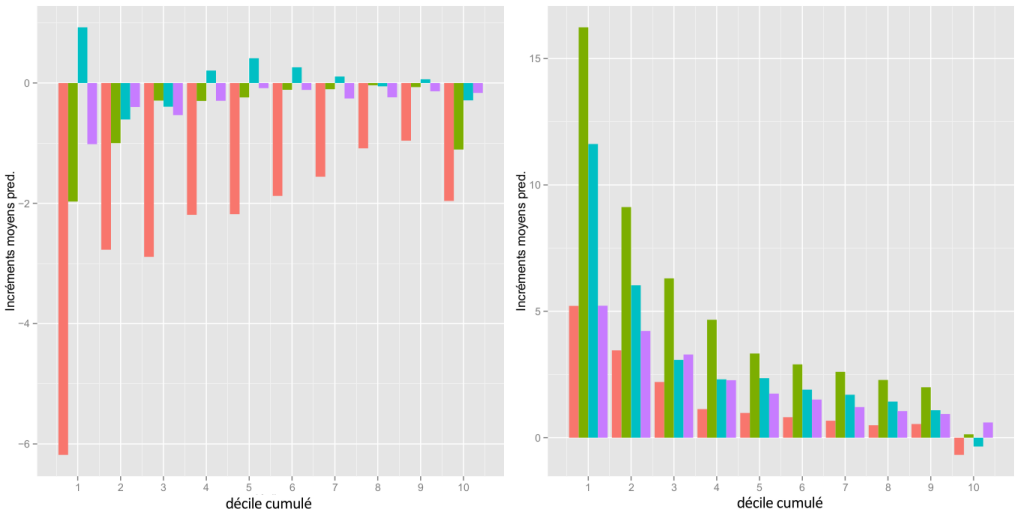
(A) $N = 500$ (B) $N = 1\,000$

(C) $N = 5\,000$ (D) $N = 10\,000$

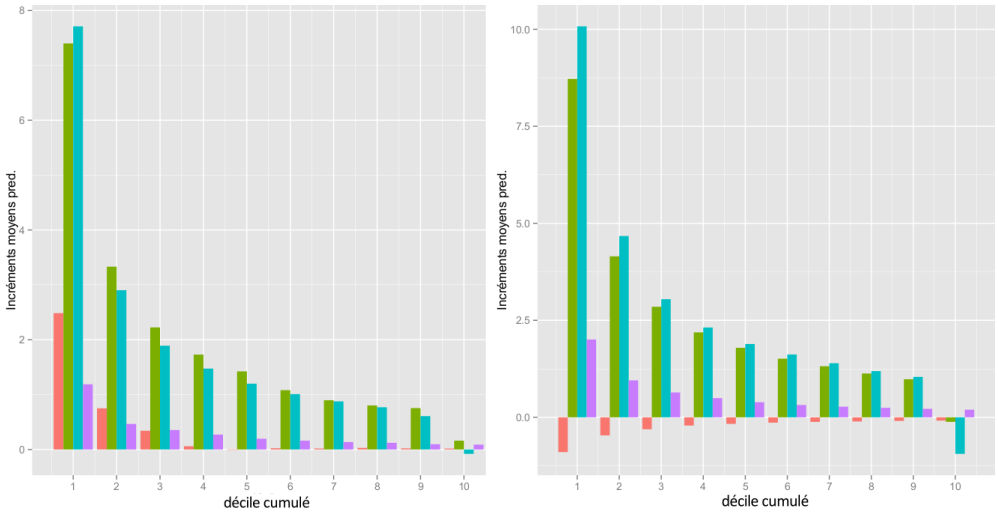
FIGURE 3.9. Modèle de Lo avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement

Nous observons sur les trois premiers déciles de la figure (A) que les incréments sont nettement plus élevés pour les méthodes bayésiennes que pour le GLM. La méthode π_3 a des incréments plus élevés sur les trois premiers déciles, elle est deux fois supérieure au GLM sur le premier décile et respectivement quatre fois et cinq fois supérieure sur le second et le troisième décile. Sur la figure (B), le GLM est meilleur que sur la simulation précédente. Son incrément moyen est aussi bon que ceux des méthodes bayésiennes sur le premier décile. Mais dès le second décile, il devient nettement moins performant que π_2 et π_3. Par contre, sur la figure (C), le GLM est le meilleur sur les deux premiers déciles puis il est proche de π_3 sur le troisième. Enfin, sur la figure (D), On remarque que pour chaque décile les méthodes ont des incréments qui se rapprochent de plus en plus. Le GLM et π_3 ont des valeurs très proches encore une fois mais π_3 est légèrement meilleur. Lorsque les échantillons sont assez grands, le GLM produit des incréments moyens élevés sur les trois premiers déciles. L'avantage de π_3 sur les autres méthodes est le fait qu'elle soit efficace sur les échantillons de toutes tailles.

Dans le but de savoir si les incréments prédits sont fiables, nous comparons les incréments moyens observés présentés précédemment avec les incréments moyens prédits. La figure 3.11 représente les différences entre les incréments moyens prédits et les incréments moyens observés.



(A) $N = 500$ (B) $N = 1\,000$



(A) $N = 5\,000$ (B) $N = 10\,000$

FIGURE 3.11. Comparaison entre les incréments moyens prédits et observés avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement

Nous remarquons que la figure (A) est assez différente des autres figures. Lorsqu'il n'y a que 500 prospects dans la simulation, les prédictions ont tendances à sous évaluer les incréments. Le GLM est le modèle où la différence est la plus grande entre la prédiction et l'observation. Cependant les prédictions des méthodes bayésiennes sont rarement à plus de 1 point des valeurs observées. Sur la figure (B), on remarque que les prédictions des méthodes π_1 et π_2 ont tendance à surévaluer les incréments de plus de 10 points sur le premier décile et plus de 5 sur le second. Le GLM et π_3 ont les prédictions les plus fiables bien qu'il y ait 5 points de différences sur le premier décile. Sur la figure (C), on remarque que les prédictions sont plus proches des valeurs observées. Cependant les prédictions de π_1 et π_2 restent éloignées des incréments observés. Les prédictions de π_3 sont les plus proches des valeurs observées avec moins de 1 point de différence sur chacun des déciles. Puis sur la figure (D), nous remarquons qu'à 10 000 prospects, les prédictions du GLM sont les plus précises et les plus fiables des simulations. Les prédictions du GLM ont moins de 0,9 point de différence. De plus, nous observons que les prédictions de π_3 sont bonnes avec moins de 2 points de différence.

Les incréments moyens prédits et observés sont indiqués dans le tableau 3.31 pour les différentes méthodes.

TABLEAU 3.31. Récapitulatif des incréments moyens prédits et observés

		Prédiction				Observation			
		GLM	π_1	π_2	π_3	GLM	π_1	π_2	π_3
$N = 500$	1	15,74	33,55	37,18	43,62	21,92	35,52	36,25	44,64
	2	1,44	10,14	17,87	16,33	4,21	11,14	18,48	16,72
	3	-1,15	4,99	7,70	8,83	1,74	5,28	8,09	9,37
	4	-2,01	2,73	5,89	5,58	0,18	3,03	5,68	5,88
	5	-0,14	2,81	4,13	4,23	2,04	3,05	3,71	4,32
	6	-0,18	1,42	2,75	2,76	1,70	1,53	2,49	2,88
	7	0,08	1,32	2,52	2,25	1,64	1,43	2,41	2,50
	8	0,77	1,26	1,80	2,09	1,86	1,30	1,85	2,33
	9	0,79	1,46	1,71	1,97	1,75	1,53	1,65	2,11
	10	-0,06	0,79	1,61	1,74	1,90	1,90	1,90	1,90
$N = 1\,000$	1	31,21	37,66	37,30	31,91	26	21,42	25,68	26,69
	2	11,56	13,66	14,76	15,49	8,10	4,53	8,73	11,26
	3	6,39	9,66	8,46	9,51	4,19	3,36	5,38	6,21
	4	4,11	6,50	5,42	6,60	2,98	1,83	3,11	4,33
	5	3,10	4,98	4,39	5,01	2,12	1,64	2,03	3,26
	6	3,14	4,26	3,24	4,10	2,33	1,36	1,34	2,59
	7	1,84	3,50	2,93	3,39	1,17	0,90	1,23	2,18
	8	2,01	3,19	2,47	2,76	1,52	0,90	1,05	1,71
	9	1,82	2,74	1,82	2,32	1,28	0,74	0,73	1,39
	10	0,68	1,49	1,01	1,96	1,36	1,36	1,36	1,36
$N = 5\,000$	1	58,61	40,40	38,10	36,42	56,12	33	30,39	35,23
	2	21,14	19,97	18,35	18,66	20,39	16,64	15,45	18,20
	3	11,93	12,76	12,13	12,20	11,59	10,54	10,24	11,85
	4	8,03	9,49	9,10	9,05	7,97	7,76	7,63	8,78
	5	6,28	7,23	7,25	7,08	6,29	5,81	6,05	6,89
	6	5,26	6,02	6,02	5,74	5,24	4,94	5,01	5,58
	7	4,25	5,24	4,98	4,83	4,23	4,34	4,11	4,69
	8	4,02	4,48	4,46	4,15	4,00	3,68	3,69	4,02
	9	3,68	4,02	3,96	3,69	3,66	3,26	3,36	3,60
	10	3,32	3,46	3,22	3,39	3,30	3,30	3,30	3,30
$N = 10\,000$	1	30,23	36,10	32,76	34,61	31,12	27,37	22,69	32,60
	2	14,45	17,27	16,25	17,41	14,92	13,12	11,57	16,46
	3	9,02	10,99	10,83	11,42	9,33	8,14	7,79	10,78
	4	7,29	8,33	8,16	8,30	7,50	6,14	5,84	7,81
	5	5,91	6,65	6,34	6,59	6,07	4,86	4,45	6,20
	6	4,89	5,51	5,22	5,39	5,03	4,00	3,60	5,07
	7	4,29	4,68	4,57	4,59	4,41	3,37	3,18	4,32
	8	3,65	4,08	3,97	4,01	3,76	2,95	2,78	3,77
	9	3,28	3,70	3,54	3,55	3,37	2,71	2,50	3,33
	10	2,92	2,88	2,06	3,20	3,00	3,00	3,00	3,00

3.5.3. Modèle de Lai

Dans cette sous section, les simulations sont effectuées sur les mêmes jeux de données que ceux utilisés avec le modèle de Lo dans le but de comparer les différents résultats obtenus.
Dans un premier temps, les estimateurs des paramètres vont être présentés dans le tableau 3.32 ainsi que leur écart-types dans le tableau 3.33. Dans un second, les incréments moyens observés vont être présentés sur des graphiques.

TABLEAU 3.32. Paramètres estimés selon les différentes méthodes

		L'estimation des paramètres			
	paramètres	GLM	π_1	π_2	π_3
$N = 500$	β_0^*	30,62	3,65	3,46	2,93
	β_1^*	−17,52	−1,11	−0,47	−1,15
	β_2^*	−27,65	−1,13	−0,85	−1,73
	β_3	0,06	−0,67	0,19	−0,32
	ζ_0^*	−3,39	−3,69	−3,64	−3,24
	ζ_1^*	2,69	2,78	2,96	2,52
	ζ_2^*	2,91	3,12	3,23	2,75
	ζ_3^*	0,98	0,69	0,94	0,91
$N = 1\,000$	β_0^*	100,25	4,36	4,84	3,21
	β_1^*	−85,40	−1,42	−0,17	−1,65
	β_2^*	−71,60	−0,85	−0,26	−1,97
	β_3	−0,87	−0,60	−0,38	−0,64
	ζ_0^*	−3,26	−3,57	−3,25	−3,19
	ζ_1^*	2,68	2,82	2,92	2,60
	ζ_2^*	2,80	2,90	2,75	2,72
	ζ_3^*	0,92	0,83	1,05	0,89
$N = 5\,000$	β_0^*	4,06	3,86	3,37	3,74
	β_1^*	−2,87	−0,53	−0,92	−2,53
	β_2^*	−2,86	−2,03	−1,13	−2,54
	β_3	−0,96	−0,41	−0,02	−0,82
	ζ_0^*	−3,32	−3,54	−3,23	−3,30
	ζ_1^*	2,81	3,07	2,58	2,80
	ζ_2^*	2,86	2,70	2,87	2,84
	ζ_3^*	0,96	1,15	1,04	0,95

	L'estimation des paramètres			
paramètres	GLM	π_1	π_2	π_3
$N = 10\,000$				
β_0^*	3,93	3,94	3,40	3,78
β_1^*	−2,75	−1,31	−0,69	−2,59
β_2^*	−2,81	−1,36	−0,37	−2,66
β_3	−1,08	−0,71	−0,03	−0,99
ζ_0^*	−3,30	−3,38	−3,28	−3,29
ζ_1^*	2,82	3,23	2,59	2,82
ζ_2^*	2,79	2,83	2,50	2,79
ζ_3^*	0,99	0,99	0,69	0,99

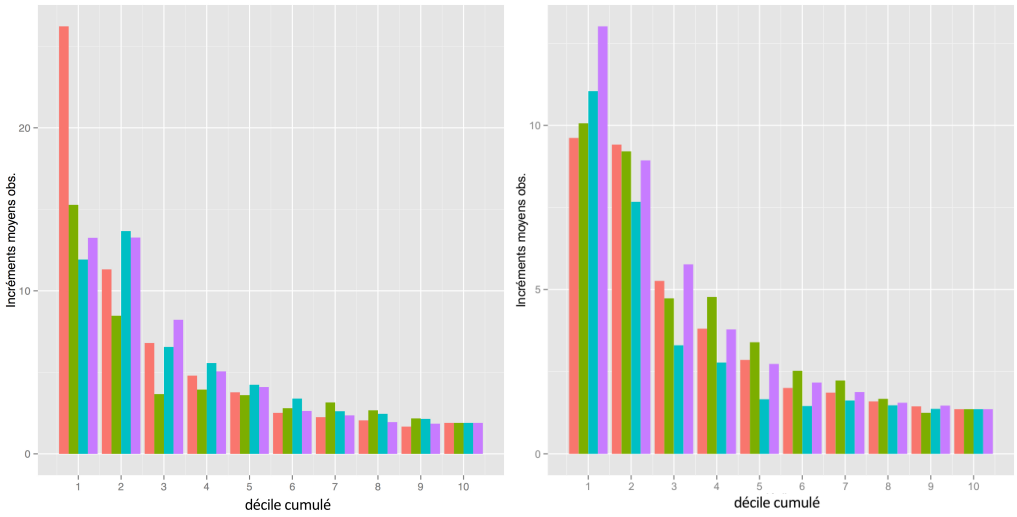
TABLEAU 3.33. Les écart-types des paramètres estimés

	Les écart-types des estimateurs			
paramètres	GLM	π_1	π_2	π_3
$N = 500$				
β_0^*	36,51	1,69	1,05	0,55
β_1^*	33,36	2,18	1,66	0,80
β_2^*	38,74	3,01	2,09	1,24
β_3	32,90	2,64	1,34	0,81
ζ_0^*	0,40	0,85	0,63	0,36
ζ_1^*	0,65	1,28	1,26	0,61
ζ_2^*	0,77	1,44	1,44	0,73
ζ_3^*	0,43	1,02	0,83	0,41
$N = 1\,000$				
β_0^*	401,76	2,49	4,18	0,50
β_1^*	346,87	2,40	1,99	0,82
β_2^*	286,59	2,27	1,53	1,21
β_3	7,18	2,00	2,02	0,85
ζ_0^*	0,24	0,58	0,63	0,23
ζ_1^*	0,33	0,80	1,11	0,32
ζ_2^*	0,41	1,06	1,05	0,41
ζ_3^*	0,27	1,07	0,95	0,26

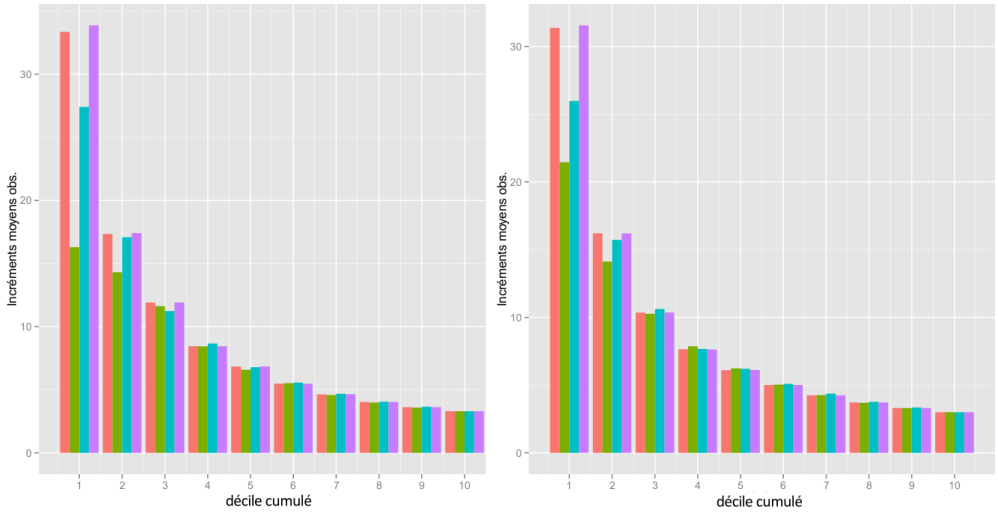
	Les écart-types des estimateurs			
paramètres	GLM	π_1	π_2	π_3
$N = 5\,000$				
β_0^*	0,42	1,62	1,79	0,34
β_1^*	0,70	2,14	1,55	0,61
β_2^*	0,81	2,59	3,06	0,70
β_3	0,84	1,98	1,70	0,71
ζ_0^*	0,09	0,72	0,41	0,09
ζ_1^*	0,16	0,94	0,66	0,16
ζ_2^*	0,13	1,05	1,09	0,13
ζ_3^*	0,16	0,94	0,84	0,16
$N = 10\,000$				
β_0^*	0,28	1,55	1,58	0,25
β_1^*	0,45	2,20	2,34	0,42
β_2^*	0,32	2,60	1,81	0,30
β_3	0,41	1,87	1,22	0,39
ζ_0^*	0,07	0,78	0,70	0,07
ζ_1^*	0,12	1,09	1,03	0,12
ζ_2^*	0,08	0,98	0,64	0,08
ζ_3^*	0,11	1,05	0,72	0,11

Comme nous l'avons vu dans les sections 3.3 et 3.4, les paramètres estimés par ce modèle sont assez différents de ceux obtenus par le modèle précédent car les prospects sont réarrangés en fonction de la variable V qui prend la valeur 1 pour la classe positive et 0 pour la classe négative.

Nous observons que pour les différentes méthodes (GLM, π_1, π_2 et π_3), les estimateurs prennent des valeurs qui sont beaucoup plus proches entre elles lorsque le modèle de Lai est utilisé (comme nous pouvons le voir sur les valeurs obtenues dans les tableaux 3.21 et 3.32). Pour l'échantillon de 500 prospects, les estimateurs du GLM sont éloignés des estimateurs des trois méthodes bayésiennes. Cependant, plus la taille de l'échantillon est grande, plus le GLM a des estimateurs qui se rapprochent de ceux des méthodes bayésiennes. Lorsque l'échantillon est de taille 10 000, le GLM et π_3 ont les estimateurs les plus proches.
Une fois les estimateurs calculés, les incréments moyens observés sont représentés dans la figure 3.12.



(A) $N = 500$ (B) $N = 1\,000$



(C) $N = 5\,000$ (D) $N = 10\,000$

FIGURE 3.12. Modèle de Lai avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement

Les incréments obtenus avec cette méthode sont assez différents de ceux obtenus avec la méthode précédente. Nous observons sur la figure (A) (où l'échantillon est de taille 500) que le GLM a un incrément plus élevé sur le premier décile mais que π_2 et π_3 performent mieux sur les autres déciles cumulés. Puis, sur la figure (B), π_3 a des incréments plus élevés sur le premier et troisième décile cumulé et il reste très proche du GLM et de π_1 sur le second décile cumulé. À partir de la figure (C), nous observons que les quatre méthodes permettent d'obtenir des incréments assez proches avec un léger avantage pour π_3. Enfin sur la figure (D),

les incréments du GLM et de π_3 sont très proches sur les deux premiers déciles cumulés. Puis sur les suivants les incréments des quatre méthodes sont presque égaux. Contrairement à la méthode précédente, le GLM fonctionne aussi bien que les autres méthodes indépendamment de la taille de l'échantillon.

3.5.4. Comparaison des modèles

Pour finir, les incréments obtenus sur les figures 3.9 et 3.12 sont comparés dans le tableau 3.34 en fonction du nombre de prospects de la simulation.
On observe que le modèle de Lo est bien plus efficace que le modèle de Lai sur les simulations où les échantillons sont de taille 500 et 1 000. Cependant, le GLM du modèle de Lai est meilleur que celui du modèle de Lo sur l'échantillon de taille 500. Sur les échantillons de taille inférieure ou égale à 5 000, les méthodes bayésiennes issues du modèle Lo ont des incréments moyens nettement plus élevés que ceux issues du modèle Lai. On remarque que plus le nombre de prospects est élevé, plus les incréments issus du modèle de Lai se rapprochent de ceux du modèle de Lo. Pour l'échantillon de taille 10 000, les modèle de Lo et de Lai produisent des incréments sensiblement proches (voir le tableau 3.34).
Il est important de souligner que pour l'ensemble des simulations analysées dans cette sous-section, la meilleure méthode sur les trois premiers déciles cumulés est toujours issue du modèle de Lo.
Lorsque le GLM et les méthodes bayésiennes sont comparés, nous observons qu'avec un échantillon de taille 500, π_3 permet d'obtenir des incréments moyens de $44,64\%$, $16,72\%$ et $9,37\%$ sur les trois premiers déciles cumulés. Il faut avoir un échantillon 10 fois plus grand pour observer des incréments aussi élevés avec le GLM.

TABLEAU 3.34. Récapitulatif des incréments moyens observés

		Lo				Lai			
		GLM	π_1	π_2	π_3	GLM	π_1	π_2	π_3
$N = 500$	1	21,92	35,52	36,25	44,64	26,22	15,27	11,92	13,25
	2	4,21	11,14	18,48	16,72	11,32	8,47	13,66	13,27
	3	1,74	5,28	8,09	9,37	6,80	3,67	6,56	8,22
	4	0,18	3,03	5,68	5,88	4,80	3,94	5,57	5,06
	5	2,04	3,05	3,71	4,32	3,78	3,60	4,24	4,10
	6	1,70	1,53	2,49	2,88	2,52	2,80	3,39	2,63
	7	1,64	1,43	2,41	2,50	2,25	3,15	2,62	2,36
	8	1,86	1,30	1,85	2,33	2,05	2,67	2,46	1,95
	9	1,75	1,53	1,65	2,11	1,67	2,17	2,15	1,85
	10	1,90	1,90	1,90	1,90	1,90	1,90	1,90	1,90
$N = 1\,000$	1	26,00	21,42	25,68	26,69	9,62	10,06	11,04	13,01
	2	8,10	4,53	8,73	11,26	9,41	9,21	7,67	8,93
	3	4,19	3,36	5,38	6,21	5,26	4,73	3,30	5,77
	4	2,98	1,83	3,11	4,33	3,81	4,78	2,78	3,79
	5	2,12	1,64	2,03	3,26	2,86	3,40	1,66	2,74
	6	2,33	1,36	1,34	2,59	2,01	2,53	1,46	2,17
	7	1,17	0,90	1,23	2,18	1,86	2,23	1,62	1,88
	8	1,52	0,90	1,05	1,71	1,60	1,67	1,48	1,56
	9	1,28	0,74	0,73	1,39	1,45	1,25	1,37	1,47
	10	1,36	1,36	1,36	1,36	1,36	1,36	1,36	1,36
$N = 5\,000$	1	56,12	33,00	30,39	35,23	33,34	16,29	27,40	33,87
	2	20,39	16,64	15,45	18,20	17,34	14,31	17,08	17,41
	3	11,59	10,54	10,24	11,85	11,91	11,63	11,24	11,91
	4	7,97	7,76	7,63	8,78	8,45	8,44	8,66	8,45
	5	6,29	5,81	6,05	6,89	6,84	6,58	6,78	6,84
	6	5,24	4,94	5,01	5,58	5,48	5,52	5,56	5,47
	7	4,23	4,34	4,11	4,69	4,63	4,58	4,68	4,64
	8	4,00	3,68	3,69	4,02	4,03	3,99	4,05	4,03
	9	3,66	3,26	3,36	3,60	3,62	3,58	3,65	3,62
	10	3,30	3,30	3,30	3,30	3,30	3,30	3,30	3,30
$N = 10\,000$	1	31,12	27,37	22,69	32,60	31,37	21,45	25,97	31,55
	2	14,92	13,12	11,57	16,46	16,20	14,12	15,72	16,20
	3	9,33	8,14	7,79	10,78	10,35	10,26	10,62	10,36
	4	7,50	6,14	5,84	7,81	7,65	7,86	7,67	7,63
	5	6,07	4,86	4,45	6,20	6,10	6,24	6,20	6,12
	6	5,03	4,00	3,60	5,07	5,01	5,04	5,09	5,01
	7	4,41	3,37	3,18	4,32	4,24	4,26	4,37	4,25
	8	3,76	2,95	2,78	3,77	3,72	3,70	3,77	3,72
	9	3,37	2,71	2,50	3,33	3,31	3,30	3,35	3,31
	10	3,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00

Dans ce chapitre ont été présentés les résultats des différentes simulations réalisées. La variable réponse a d'abord été exprimée en fonction d'une seule variable explicative et avec un certain nombre de paramètres fixés (taille de l'échantillon, taux de réponse). Puis on a observé que le meilleur modèle était celui de Lo avec la méthode π_2 si l'on s'arrête au troisième décile cumulé. Par contre, si l'étude se contentait du premier ou second décile cumulé, le modèle Lai serait meilleur quelle que soit la méthode utilisée. Lorsque la variable réponse a été exprimée en fonction de deux variables explicatives, le modèle de Lo avec la méthode π_3 a eu les incréments les plus élevés sur les trois premiers déciles cumulés. Dans la dernière section, la variable réponse est exprimée en fonction de trois variables explicatives comme dans les simulations réalisées dans l'étude de Lo (2002). Les nombreuses simulations réalisées montrent que plus l'échantillon est grand, plus les incréments observés des différentes méthodes sont proches. La méthode π_3 du modèle de Lo est la méthode la plus avantageuse, ses incréments font toujours partis des plus élevés indépendamment de la taille de l'échantillon.

CONCLUSION

Les modèles incrémentaux analysent et prédisent l'impact qu'une action comme une campagne marketing pourrait avoir sur la décision d'achat d'un prospect. Le but de ces méthodes est de déceler puis cibler les prospects qui changent de comportement dû à cette action. Plus le taux de réponse des individus ciblés est élevé, plus la campagne de marketing est performant. Les deux modèles incrémentaux utilisés au cours de ce mémoire sont le modèle de Lo (2002) et celui de Lai (2004). Ils sont tous les deux modélisés sur la base de deux régressions logistiques, une pour le groupe contrôle et l'autre pour le groupe traitement.

Dans ce mémoire, les modélisations ont été effectuées d'un point de vue bayésien et d'un point de vue fréquentiste afin que leur ciblage soit comparé. Dans l'approche fréquentiste, des régressions logistiques sont utilisées pour la modélisation alors que dans l'approche bayésienne, la modélisation a été effectuée avec des régressions logistiques bayésiennes. Cette seconde approche nécessite plus d'outils statistiques. Dans le but de mieux comprendre le comportement des prospects, ces modèles ont dans un premier temps été testés sur des simulations avec une puis deux variables explicatives. Ces simulations ont permis de visualiser le comportement des prospects en deux dimensions puis en trois dimensions.
Puis, les simulations avec trois variables explicatives ont été effectuées sur la base des simulations menées par Lo (2002). Le nombre de prospects dans chaque simulation a oscillé entre 250 et 10 000. Nous avons observé plusieurs avantages en faveur des méthodes bayésiennes. Dans un premier temps, nous avons remarqué que plus le nombre de prospects était petit, plus élevés étaient les incréments des méthodes bayésiennes par rapport à ceux du GLM. Les méthodes bayésiennes permettaient d'avoir de meilleures estimations des paramètres dû à l'information de la loi *a priori*. Par contre, lorsque le nombre de prospects était élevé, les méthodes avaient tendance à rapidement devenir équivalentes en terme de ciblage. Dans un second temps, nous avons remarqué qu'en présence de données séparées,

les méthodes bayésiennes estimaient mieux les paramètres des régressions logistiques que le GLM. C'est un avantage considérable si par la suite nous voulons étudier les variables et leurs influences sur la variable réponse. Dans un troisième temps, il a été remarqué que la méthode du GLM est bonne sur le premier décile mais sur les déciles cumulés suivants, les méthodes bayésiennes (dans leur ensemble), sont meilleures ou aussi efficaces que le GLM. Sur le second décile cumulé, la meilleure méthode bayésienne peut avoir un incrément jusqu'à 4 fois plus élevé que le GLM.
Au vu des résultats obtenus, la méthode la plus stable et donnant les meilleurs résultats sur l'ensemble de l'étude est la méthode bayesienne ayant la loi normale en loi *a priori*.

Il a été intéressant de réaliser des modélisations bayésiennes dans ce domaine puisqu'aucun article trouvé ne s'y intéressait. Cependant, dans le but de pousser plus loin la recherche, il serait intéressant d'appliquer ces modélisations sur de véritables données d'entreprise et d'y ajouter des contraintes de coût afin de mieux déterminer le nombre de prospects qu'il faut cibler pour optimiser le retour sur investissement. De plus, dans le but de rendre les modélisations plus réalistes, il serait intéressant de travailler avec plusieurs groupes traitement, chacun pourrait représenter une manière de communiquer avec le prospect (courriel, téléphone déplacement) afin de comparer, en fonction de leur coût, les différents impacts qu'ils peuvent avoir sur la décision du prospect.

Bibliographie

Allen, M. (1997), *Direct Marketing (Marketing in Action)*, Kogan Page, 160 p.

Almquist, E. et Wyner, G. (2001), *Boost your marketing roi with experimental design*, Harvard Business Review, 135-141.

Bellamy, S., Lin, J. et Have, T. T. (2007), *An introduction to causal modeling in clinical trials*, Clinical Trials, **4**(1), 58-73.

Berry, M. J.A. et Linoff, G.S. (2011), *Data Mining Techniques : For Marketing, Sales and Customer Support*, Wiley, 888 p.

Chickering, D. M. et Heckerman, D. (2000), *A decision theoretic approach to targeted advertising*, UAI, Stanford, 82-88.

Desmet, P. (2005), *Marketing direct,Concepts et méthodes*, Dunod, 380 p.

Durret, R. (2010), *Probability : theory and examples*, Cambridge University Press, 4 ème édition, 440 p.

Gelman, A., Jakulin, A., Pittau, M.G. et Su, Y.S. (2008), *A weakly informative default prior distribution for logistic and other regression models*, Annals of Applied Statistics, **2**, 1360-1383.

Gentle, J.E. (1998), *Random Number Generation and Monte Carlo Methods, Statistics and Computing*, Springer, 382 p.

Goetghebeur, E. et Lapp, K. (1997), *The effect of treatment compliance in a placebo controlled trial : Regression with unpaired data*, Applied Statistics, **46**(3), 351-364.

Hansotia, B. et Rukstales, B. (2002), *Incremental value modeling*, Journal of Interactive Marketing, **16**, 35-46.

Kelly, C.T. (1999), *Iterative Methods for Optimizations*, Siam, 180 p.

Lai, Y.T. (2004), *Influential marketing : a new direct marketing strategy addressing the existence of voluntary buyers*, *In* University of British Columbia, 60 p.

Lai, Y.T., Wang, K., Ling, D., Shi, H. et Zhang, J. (2006), *Direct marketing when there are voluntary buyers*, *In* 2013 IEEE 13th international conference on data mining, IEEE computer society, 922-927.

Lo, V. (2002), *The true lift model, a novel data mining approach response modeling in database marketing*, SIGKDD Explorations, **4**, 78-86.

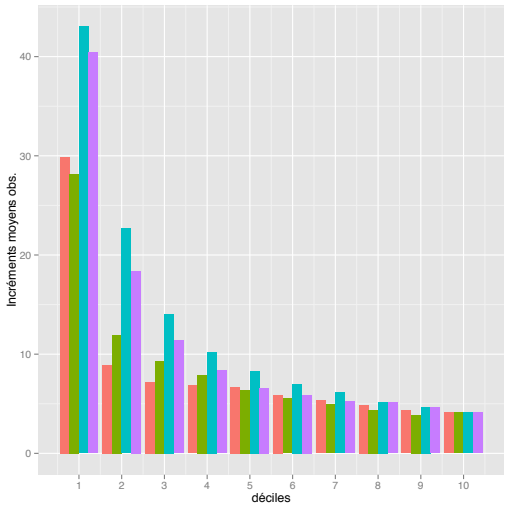
Radcliffe, N. J. (2007), *Using control groups to target on predicted lift : Building and assessing uplift models*, Direct Marketing Journal, Direct Marketing Association Analytics Council, **1**, 14-21.

Radcliffe, N. J. et Surry, P. D. (1999), *Differential response analysis : Modeling true response by isolating the effect of a single action*, Proceedings of Credit Scoring and Credit Control VI. Credit Research Centre, University of Edinburgh Management School.

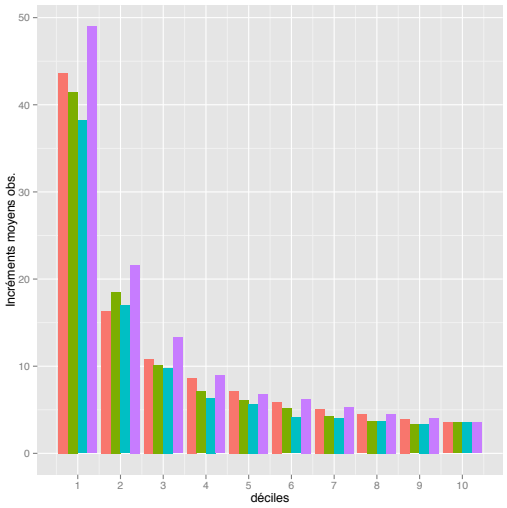
Robert, C. P. (2006), *Le choix bayésien Principes et pratique*, Springer, 638 p.

Van Belle, G., Fisher, L. D., Heagerty, P. J. et Lumley, T. (2004), *Biostatistics : a methodology for the health sciences*, Wiley, 896 p.

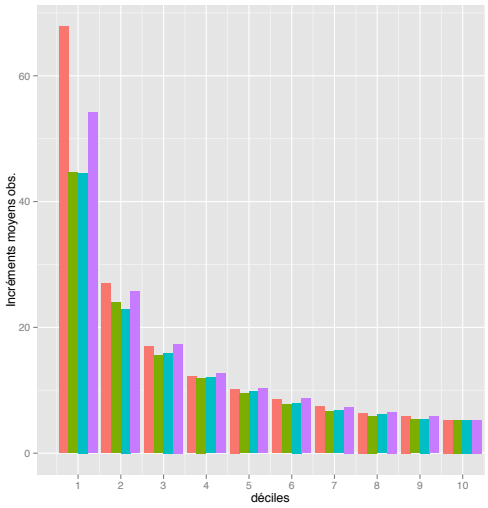
ANNEXES



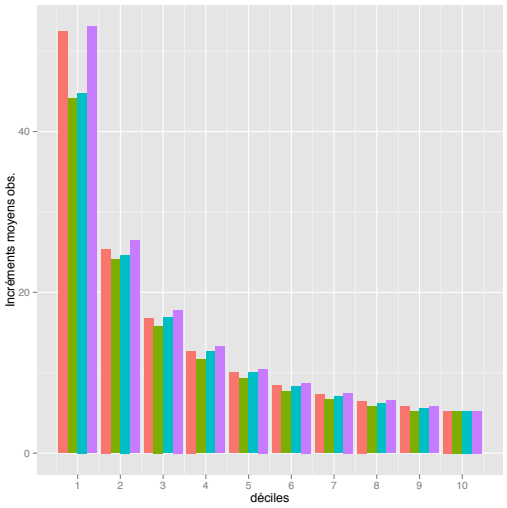
(A) $N = 500$



(B) $N = 1\,000$

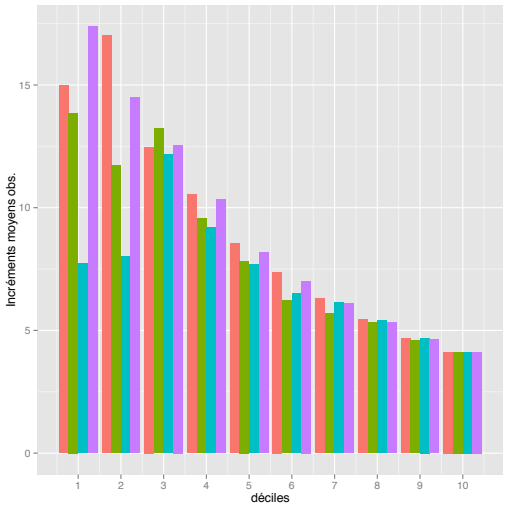


(C) $N = 5\,000$

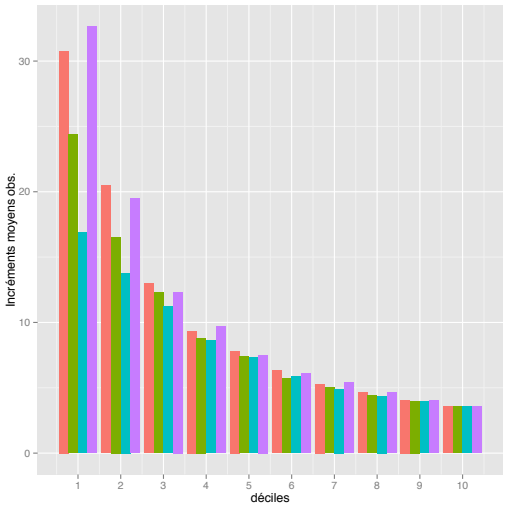


(D) $N = 10\,000$

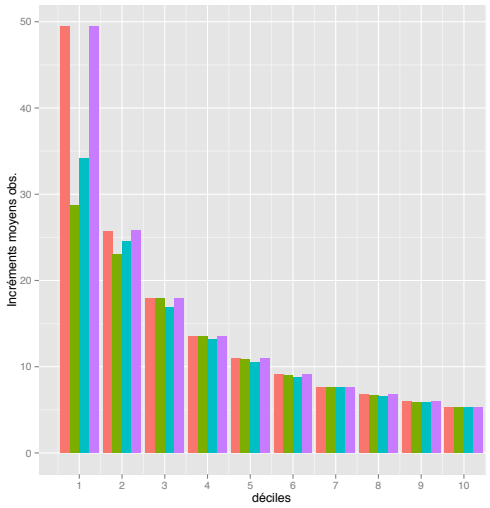
FIGURE 3.13. Modèle de Lo avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



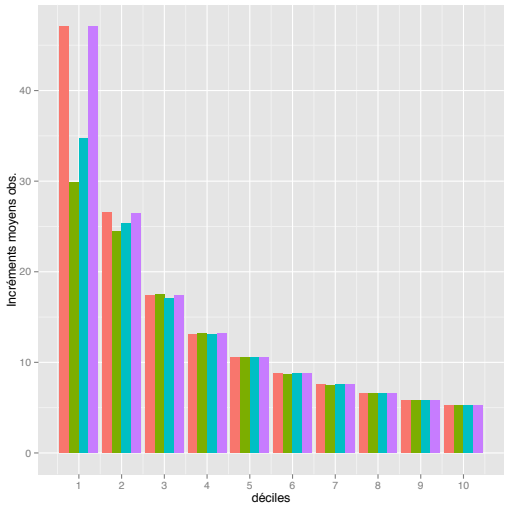
(A) $N = 500$



(B) $N = 1\,000$

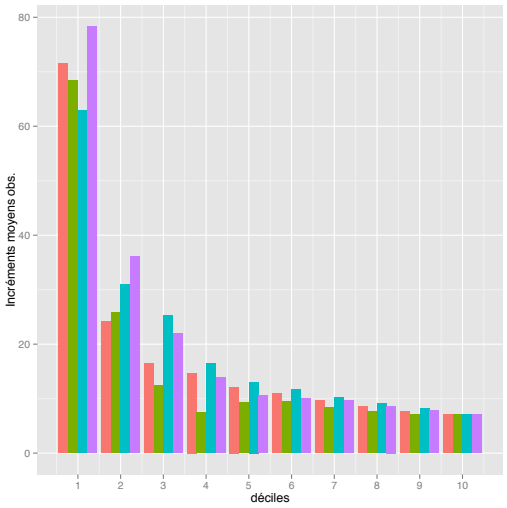


(C) $N = 5\,000$

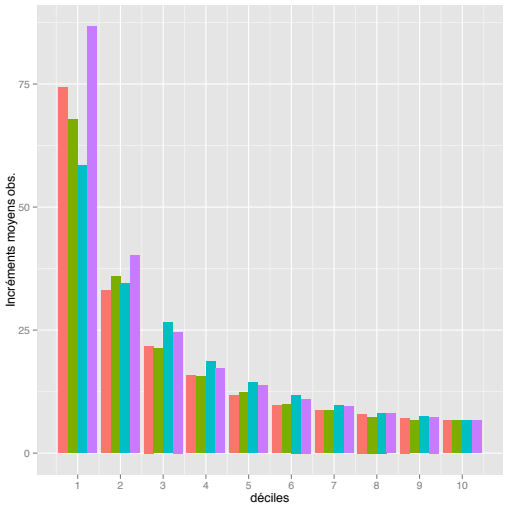


(D) $N = 10\,000$

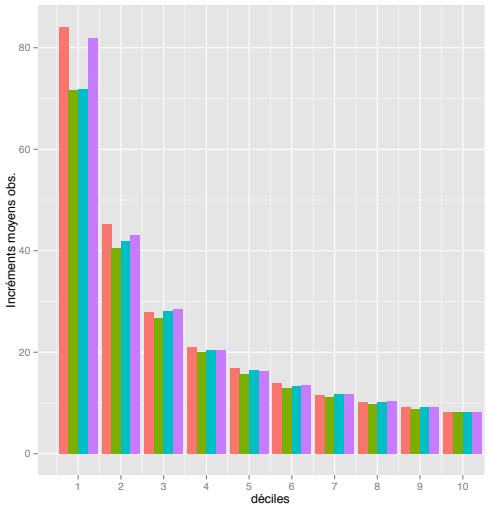
FIGURE 3.14. Modèle de Lai avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



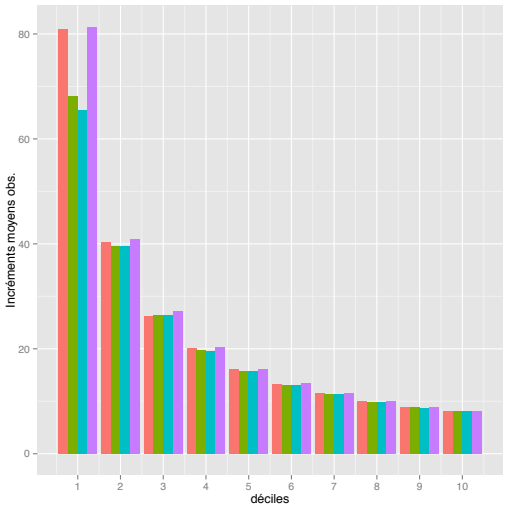
(A) $N = 500$



(B) $N = 1\,000$

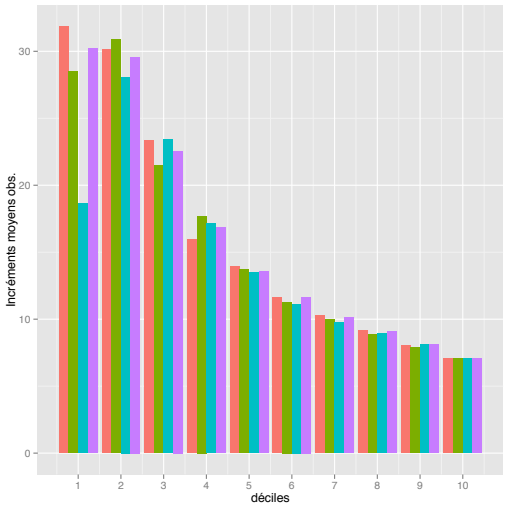


(C) $N = 5\,000$

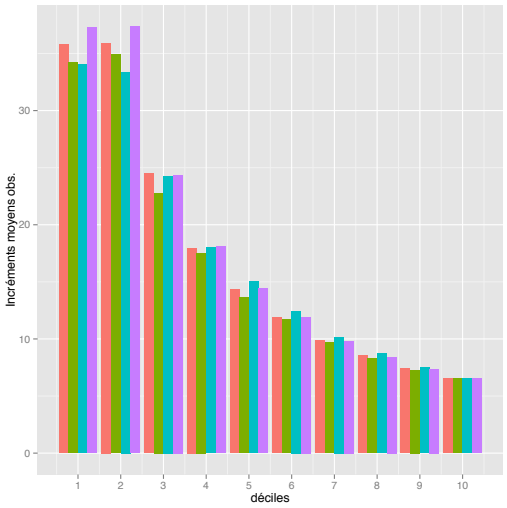


(D) $N = 10\,000$

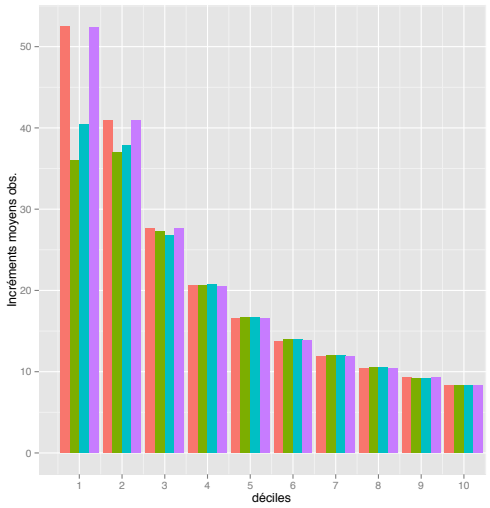
FIGURE 3.15. Modèle de Lo avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



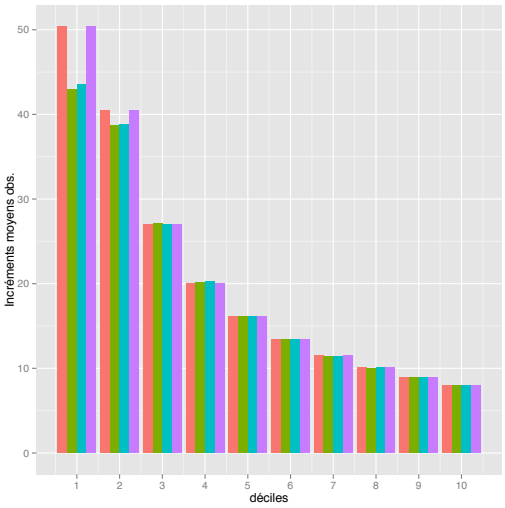
(A) $N = 500$



(B) $N = 1\,000$

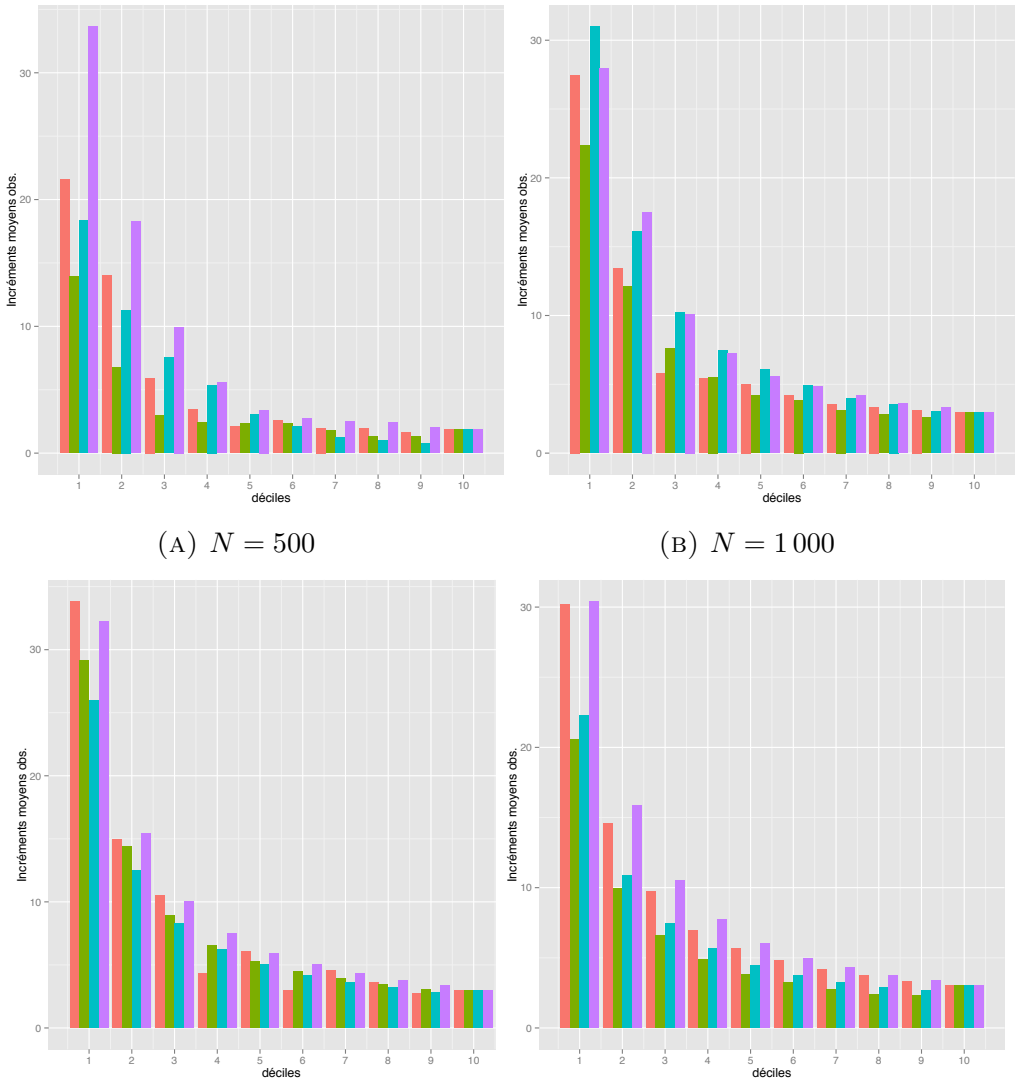


(C) $N = 5\,000$



(D) $N = 10\,000$

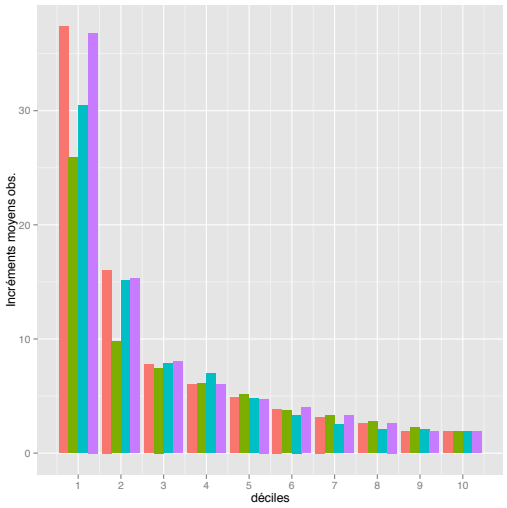
FIGURE 3.16. Modèle de Lai avec un groupe contrôle de 10% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



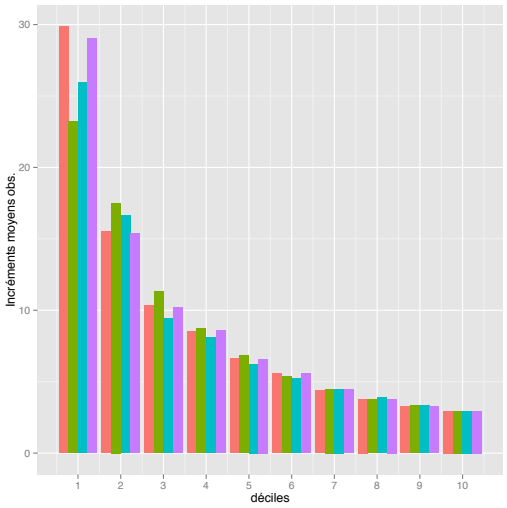
(A) $N = 500$ (B) $N = 1\,000$

(C) $N = 5\,000$ (D) $N = 10\,000$

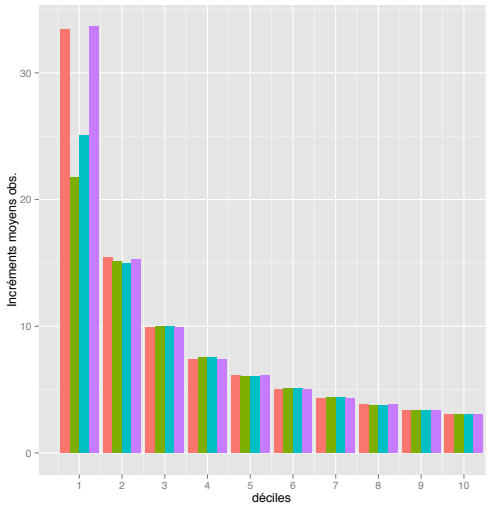
FIGURE 3.17. Modèle de Lo avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



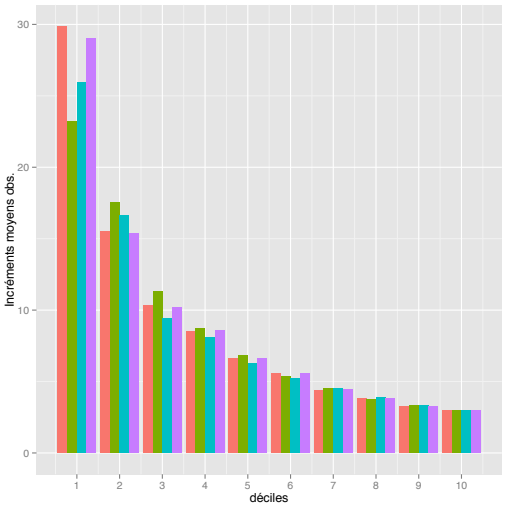
(A) $N = 500$



(B) $N = 1\,000$

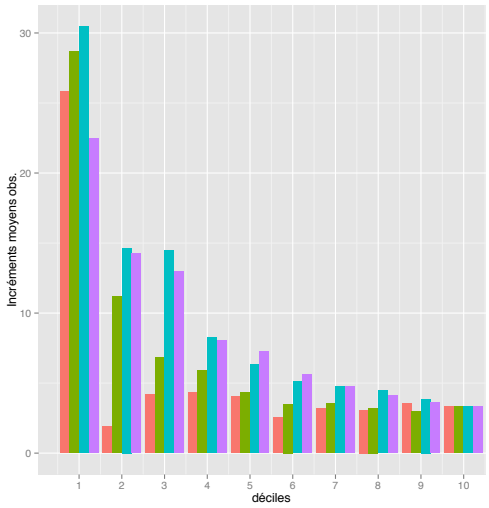


(C) $N = 5\,000$

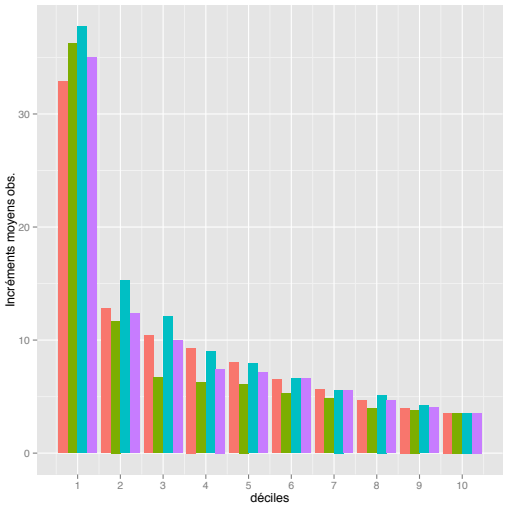


(D) $N = 10\,000$

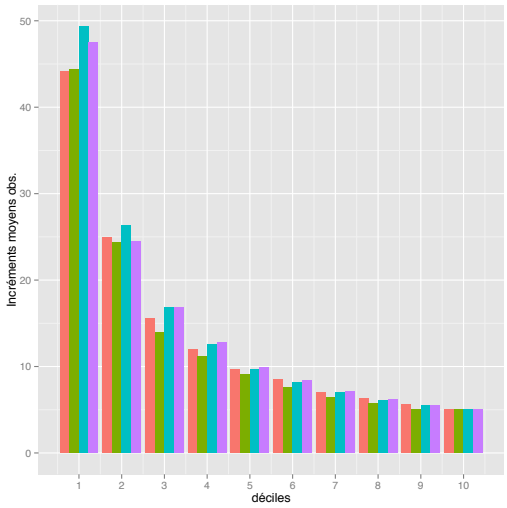
FIGURE 3.18. Modèle de Lai avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



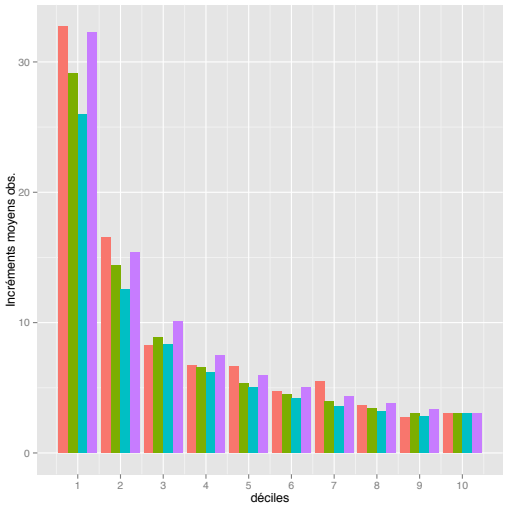
(A) $N = 250$



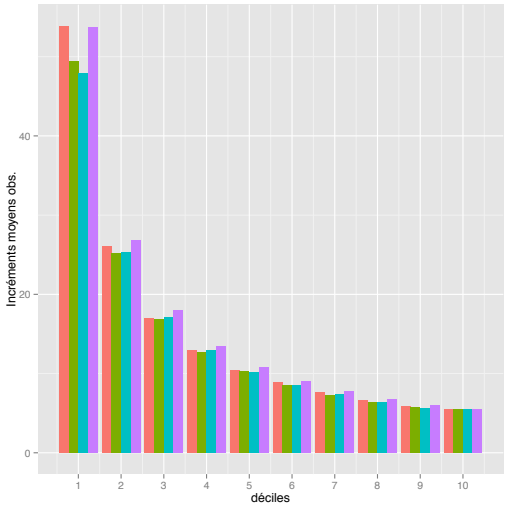
(B) $N = 500$



(C) $N = 1\,000$

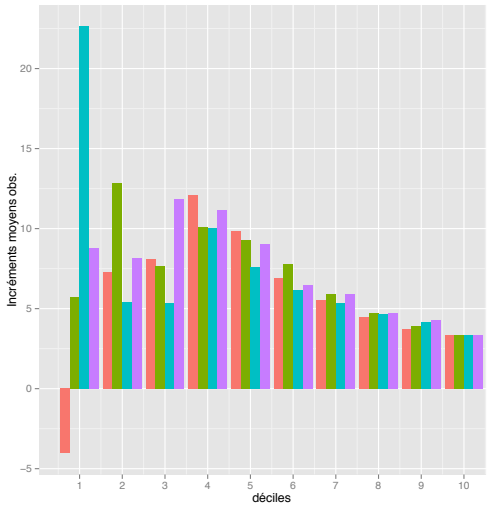


(D) $N = 5\,000$

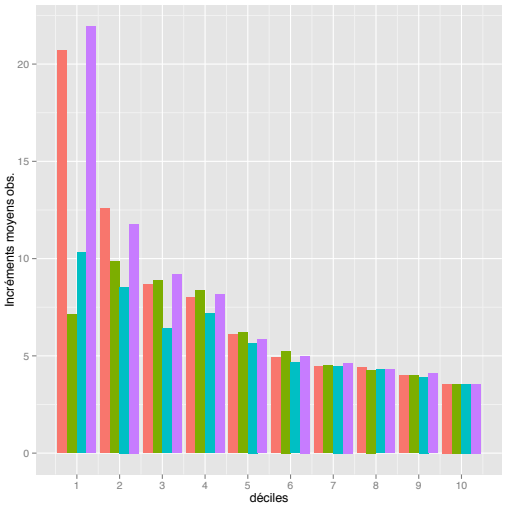


(E) $N = 10\,000$

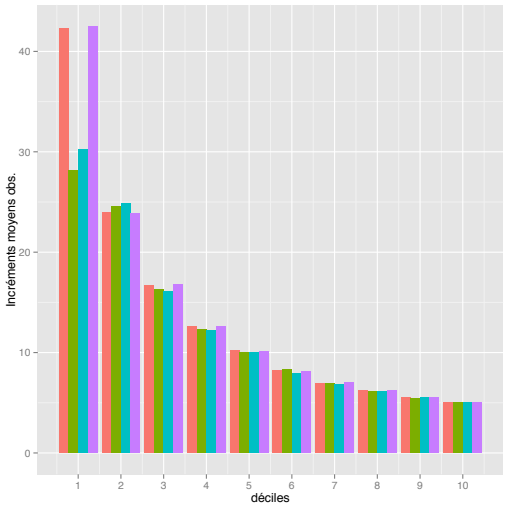
FIGURE 3.19. Modèle de Lo avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



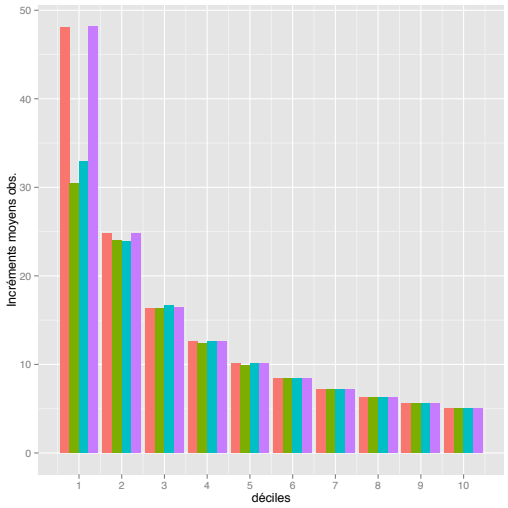
(A) $N = 250$



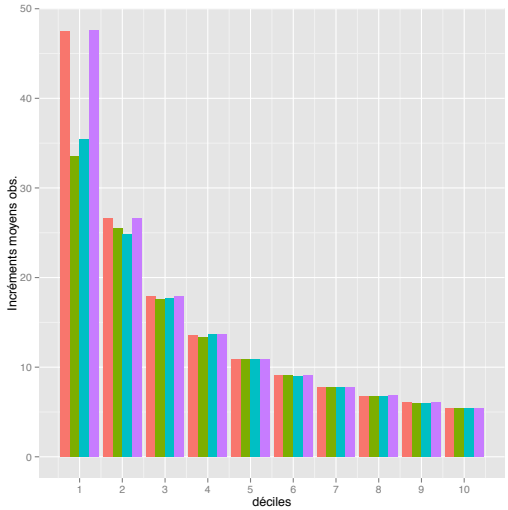
(B) $N = 500$



(C) $N = 1\,000$

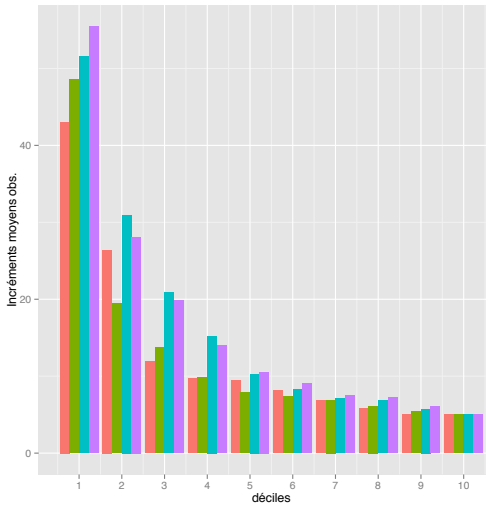


(D) $N = 5\,000$

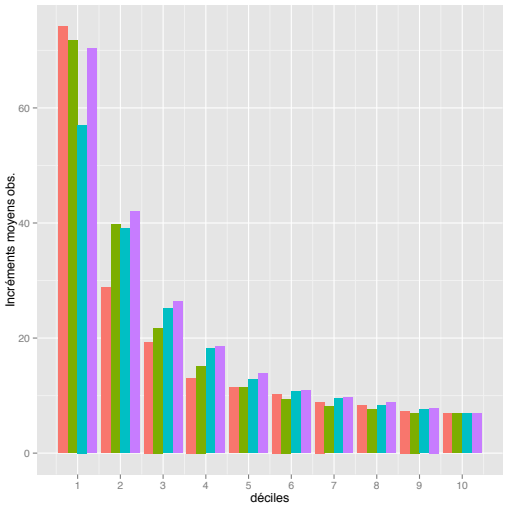


(E) $N = 10\,000$

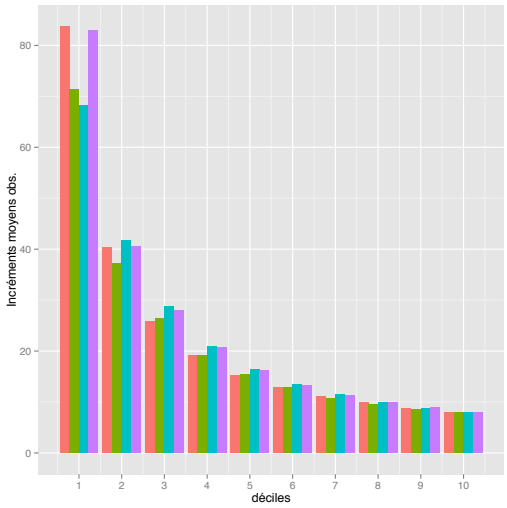
FIGURE 3.20. Modèle de Lai avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



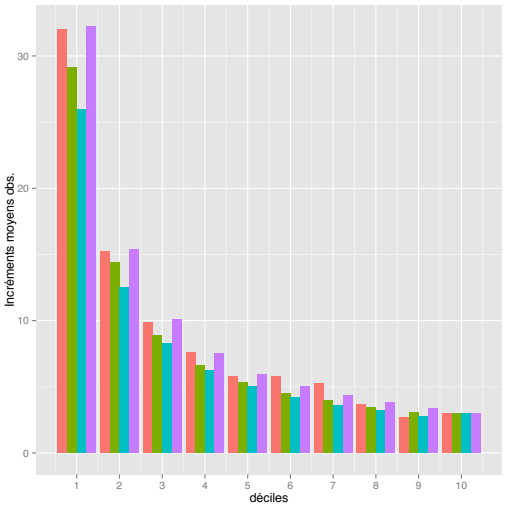
(A) $N = 250$



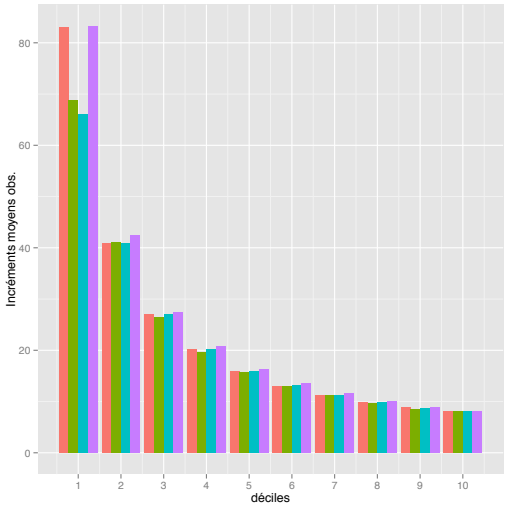
(B) $N = 500$



(C) $N = 1\,000$

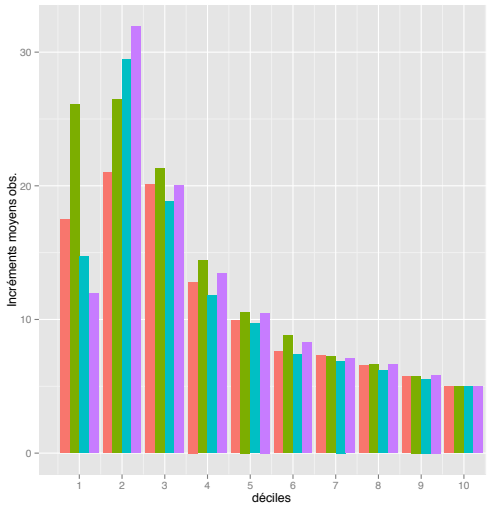


(D) $N = 5\,000$

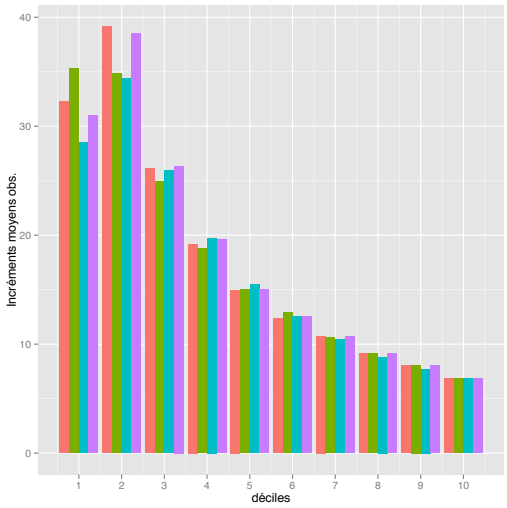


(E) $N = 10\,000$

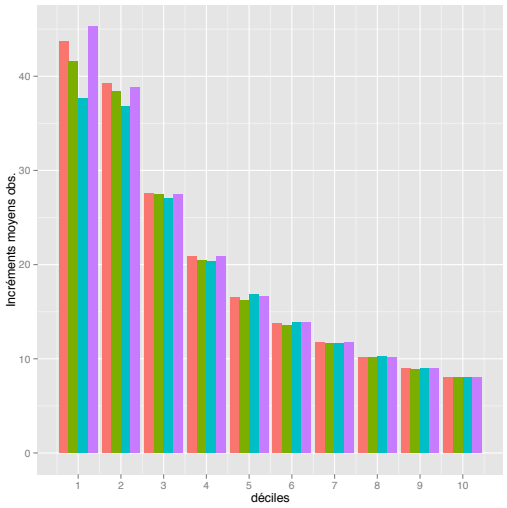
FIGURE 3.21. Modèle de Lo avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



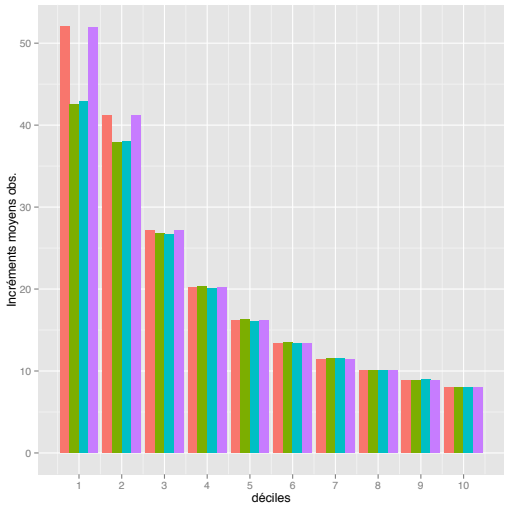
(A) $N = 250$



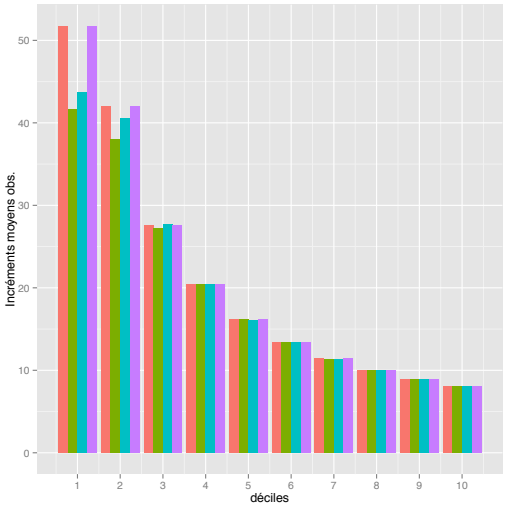
(B) $N = 500$



(C) $N = 1\,000$

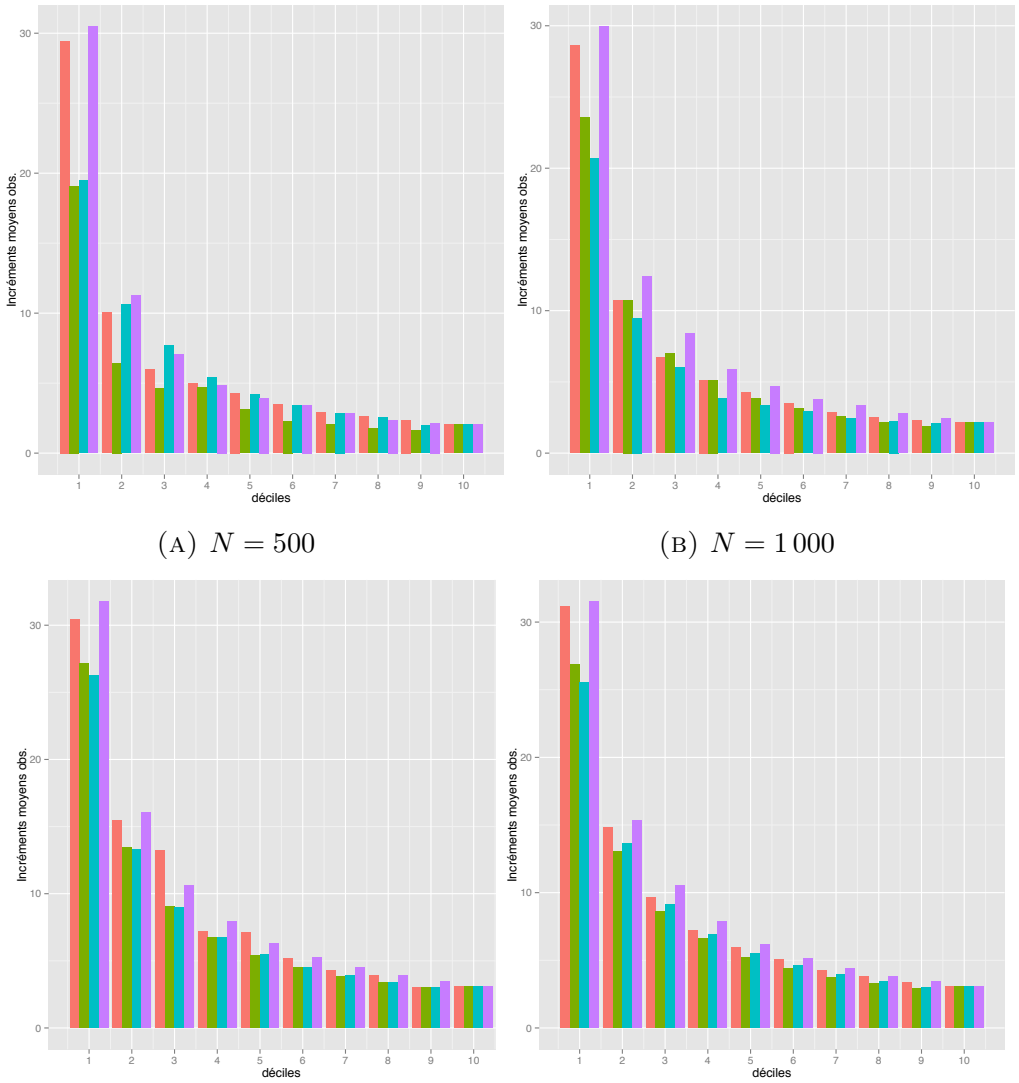


(D) $N = 5\,000$



(E) $N = 10\,000$

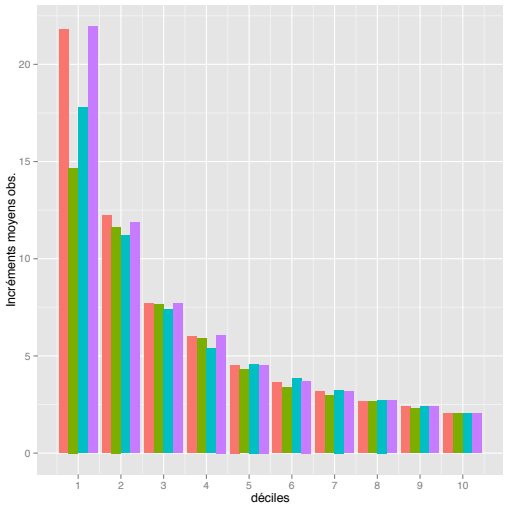
FIGURE 3.22. Modèle de Lai avec un groupe contrôle de 20% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



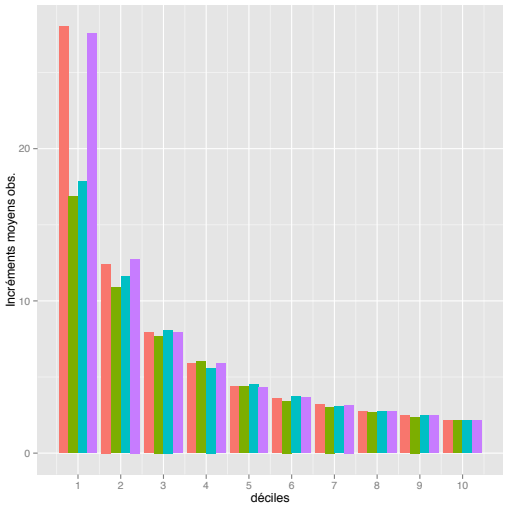
(A) $N = 500$ (B) $N = 1\,000$

(C) $N = 5\,000$ (D) $N = 10\,000$

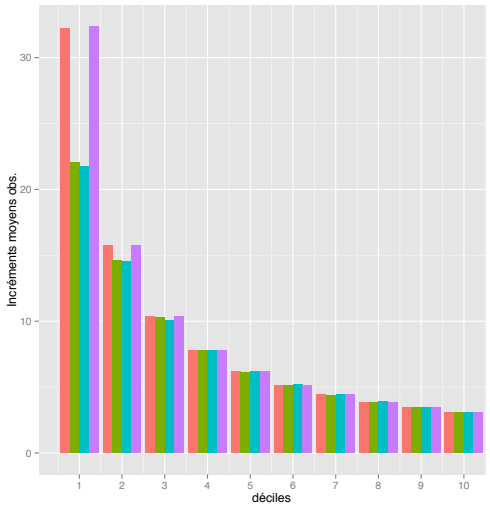
FIGURE 3.23. Modèle de Lo avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



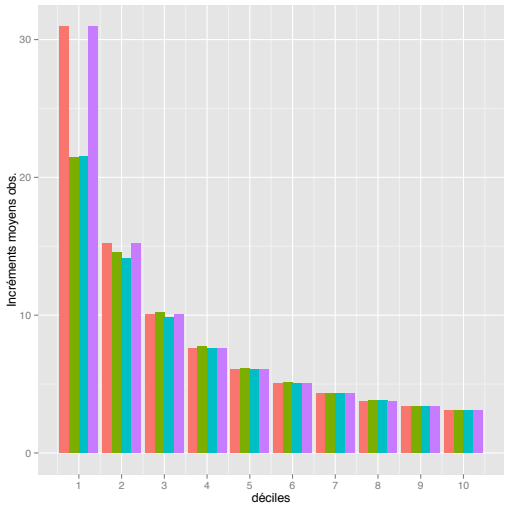
(A) $N = 500$



(B) $N = 1\,000$

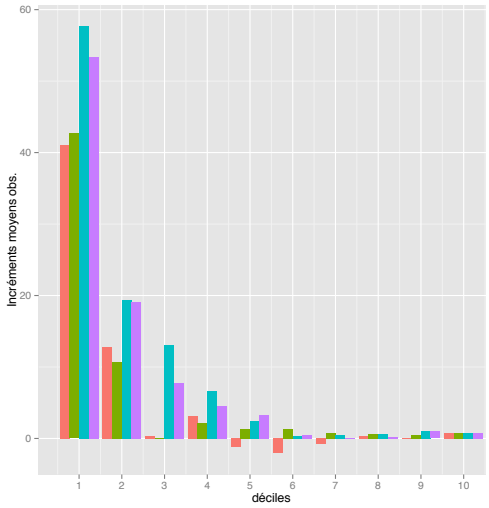


(C) $N = 5\,000$

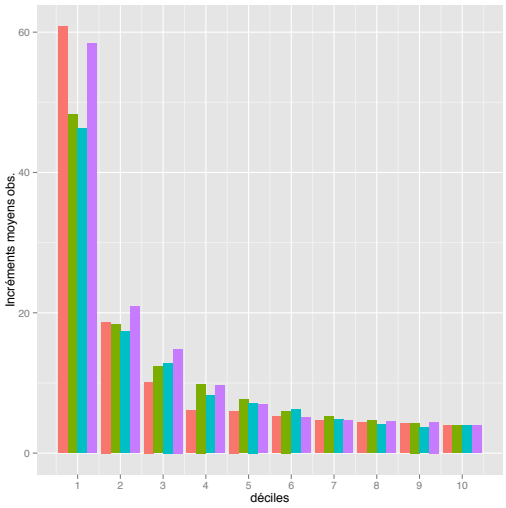


(D) $N = 10\,000$

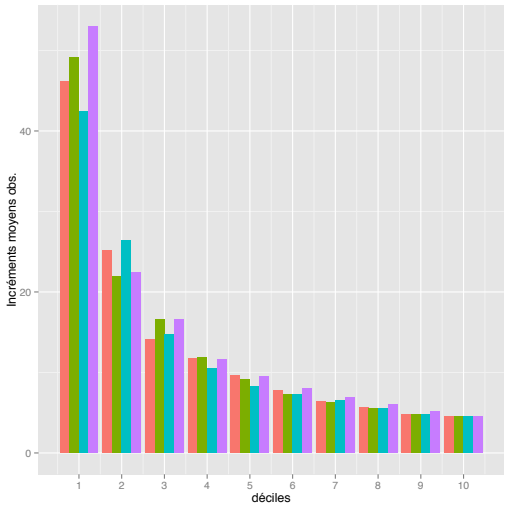
FIGURE 3.24. Modèle de Lai avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



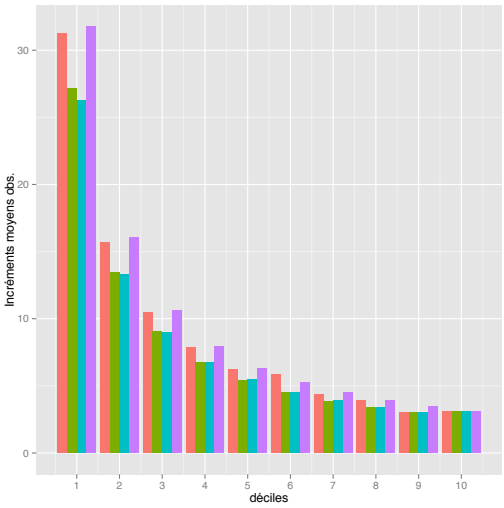
(A) $N = 250$



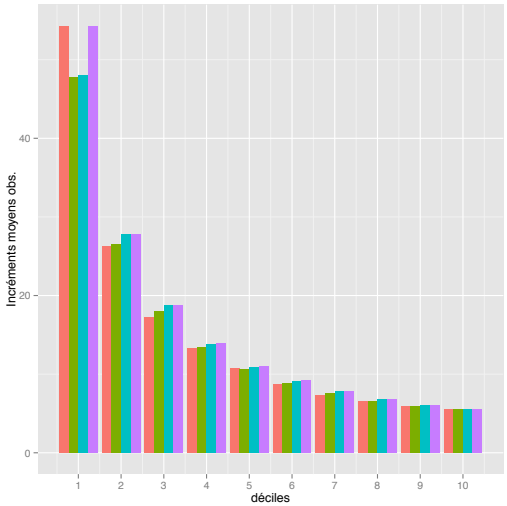
(B) $N = 500$



(C) $N = 1\,000$

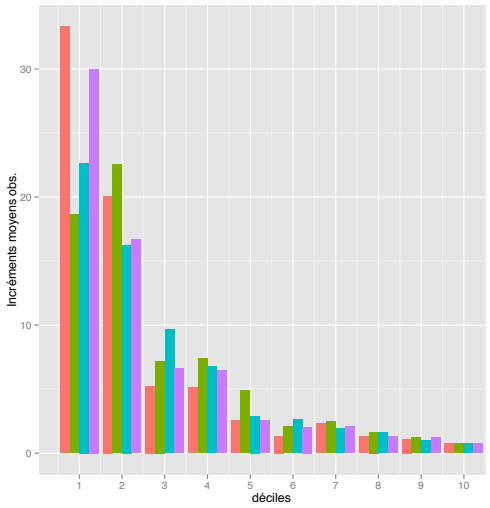


(D) $N = 5\,000$

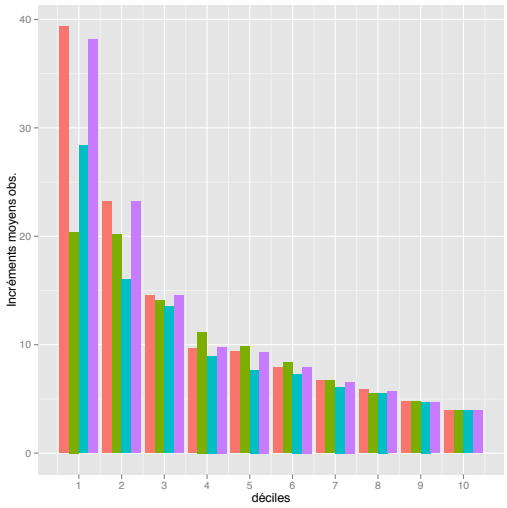


(E) $N = 10\,000$

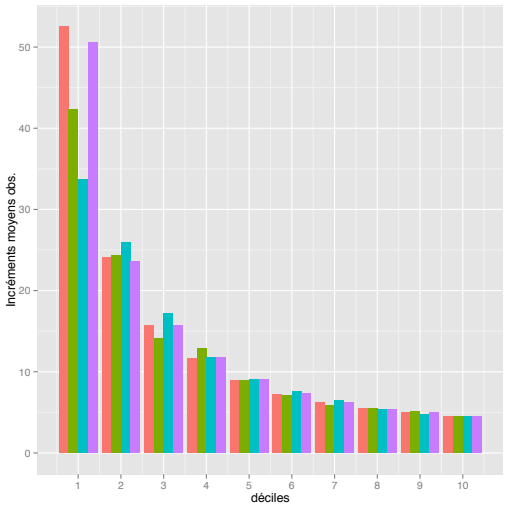
FIGURE 3.25. Modèle de Lo avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



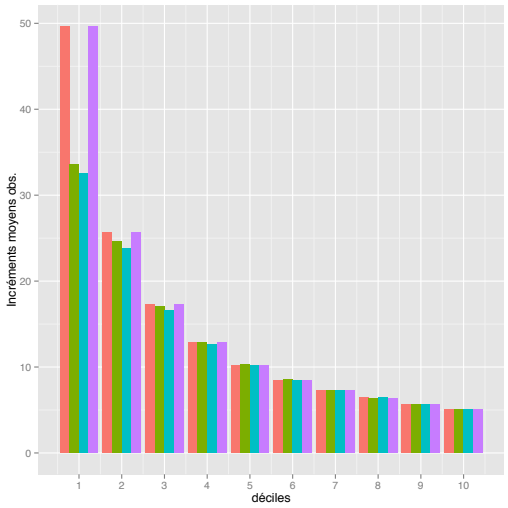
(A) $N = 250$



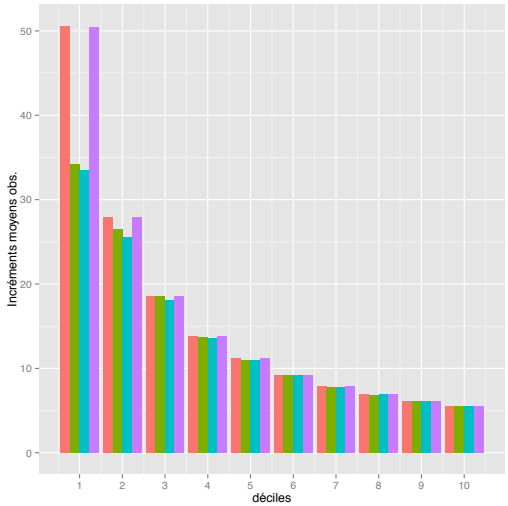
(B) $N = 500$



(C) $N = 1\,000$

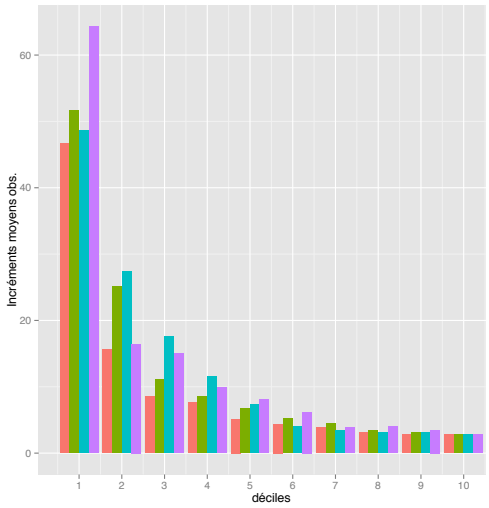


(D) $N = 5\,000$

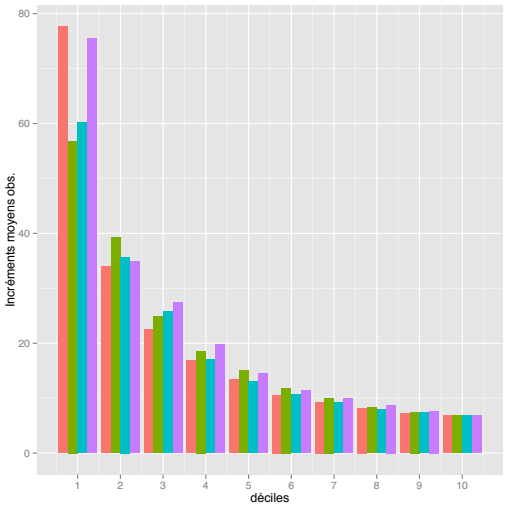


(E) $N = 10\,000$

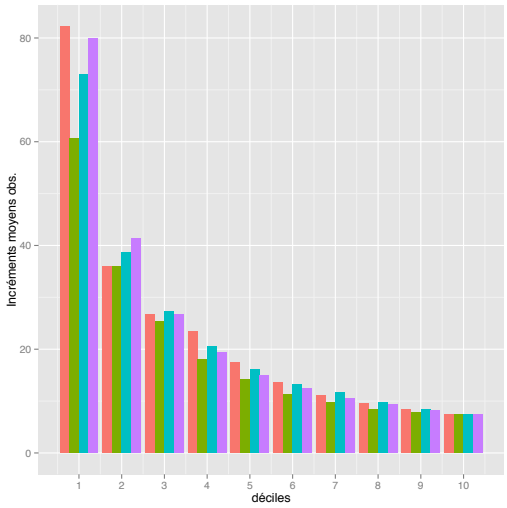
FIGURE 3.26. Modèle de Lai avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



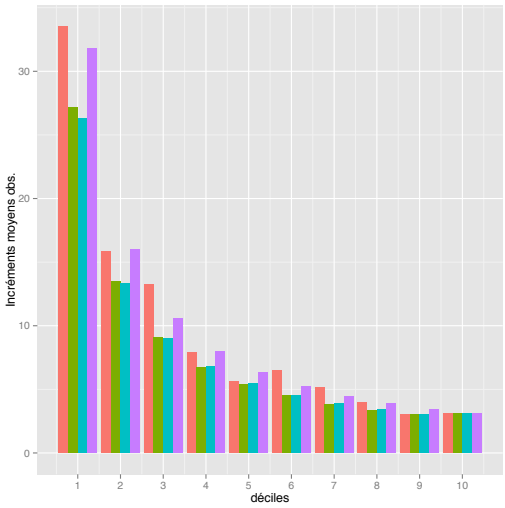
(A) $N = 250$



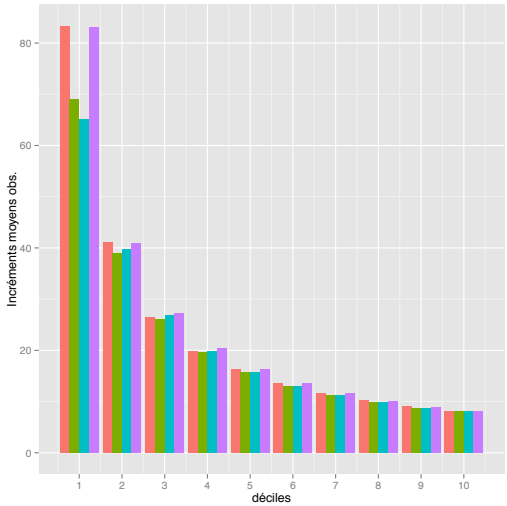
(B) $N = 500$



(C) $N = 1\,000$

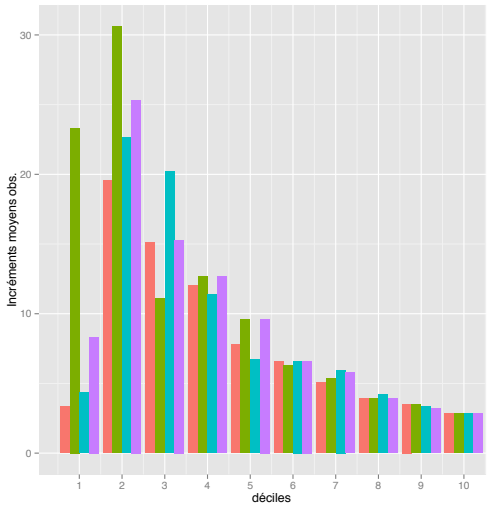


(D) $N = 5\,000$

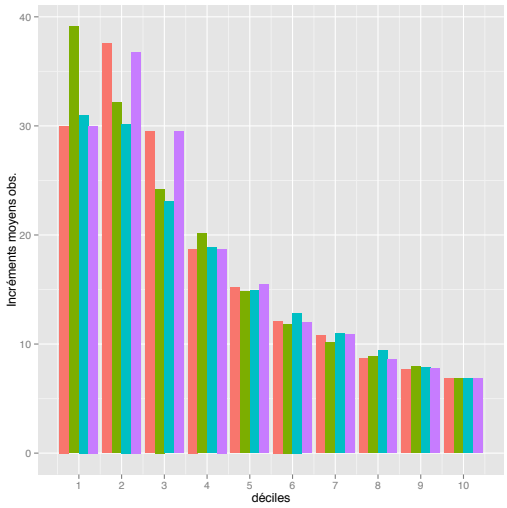


(E) $N = 10\,000$

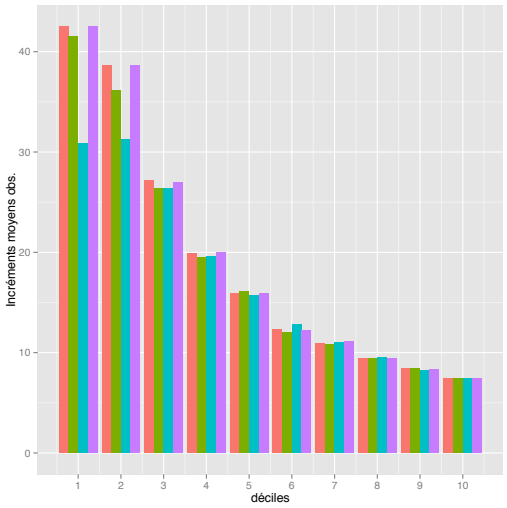
FIGURE 3.27. Modèle de Lo avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



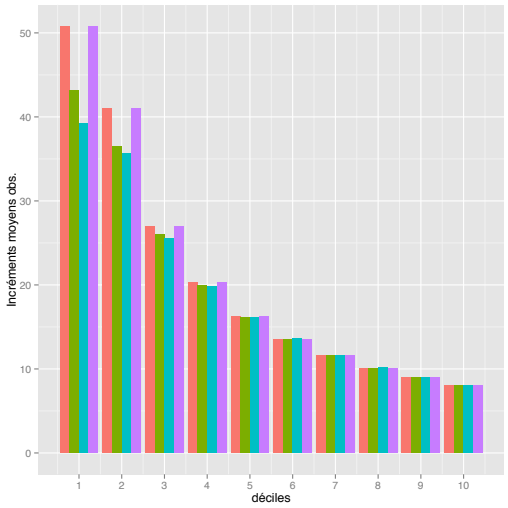
(A) $N = 250$



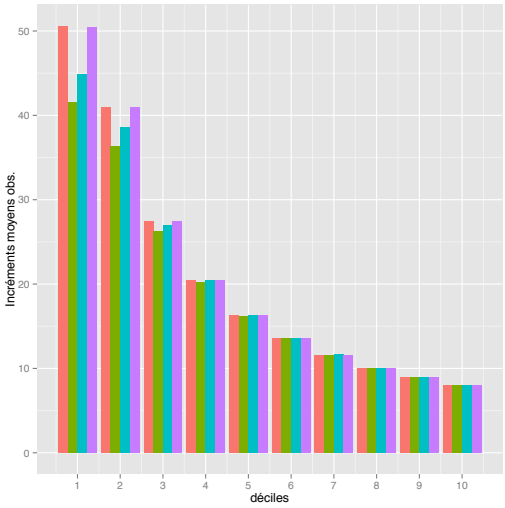
(B) $N = 500$



(C) $N = 1\,000$

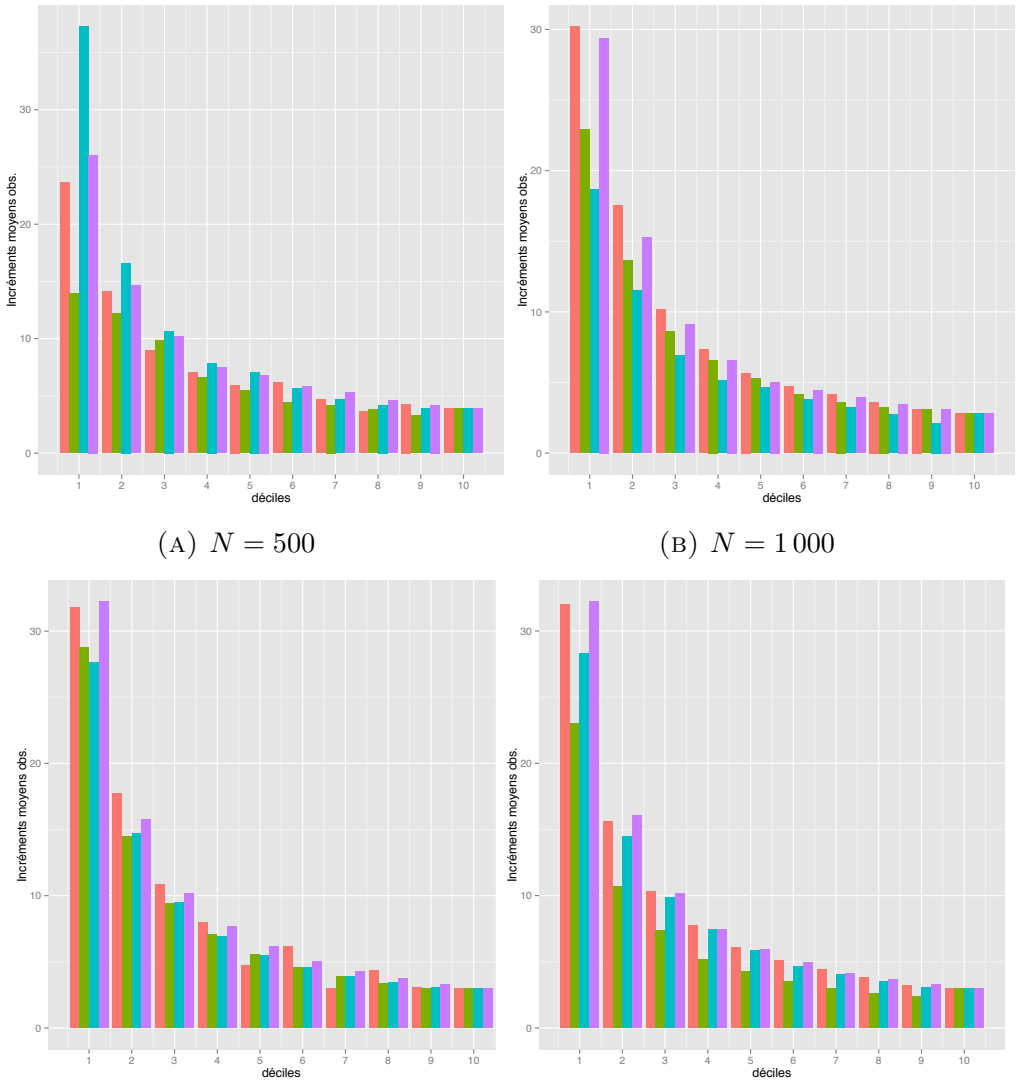


(D) $N = 5\,000$



(E) $N = 10\,000$

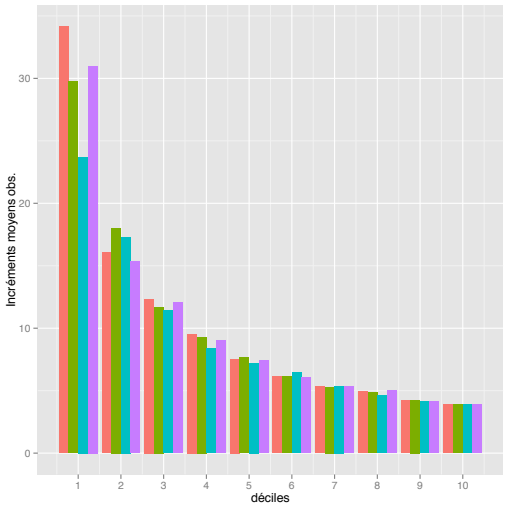
FIGURE 3.28. Modèle de Lai avec un groupe contrôle de 30% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



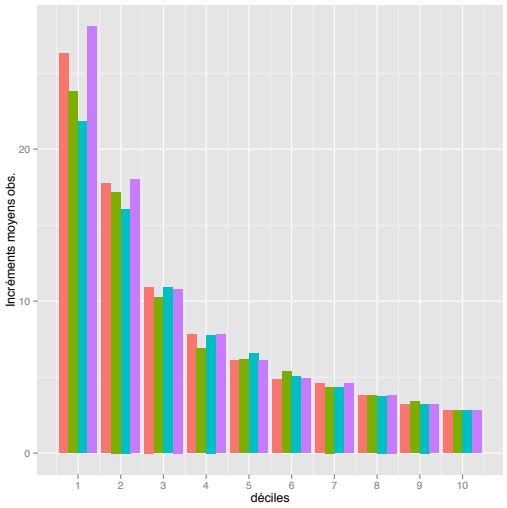
(A) $N = 500$ (B) $N = 1\,000$

(C) $N = 5\,000$ (D) $N = 10\,000$

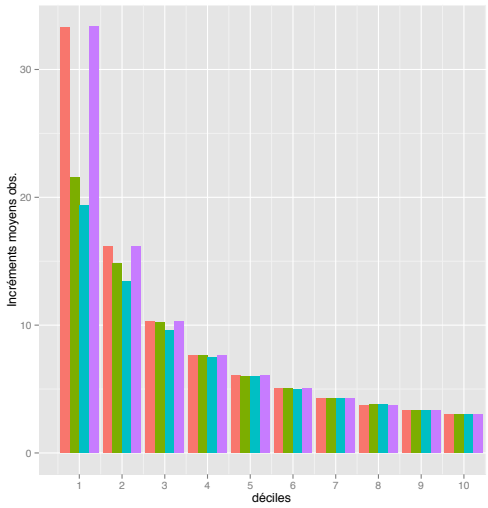
FIGURE 3.29. Modèle de Lo avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



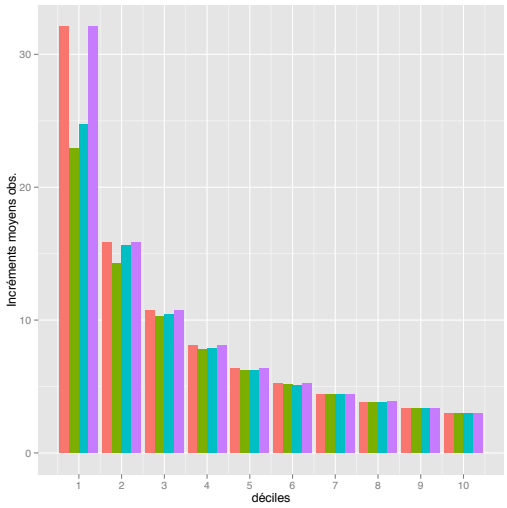
(A) $N = 500$



(B) $N = 1\,000$

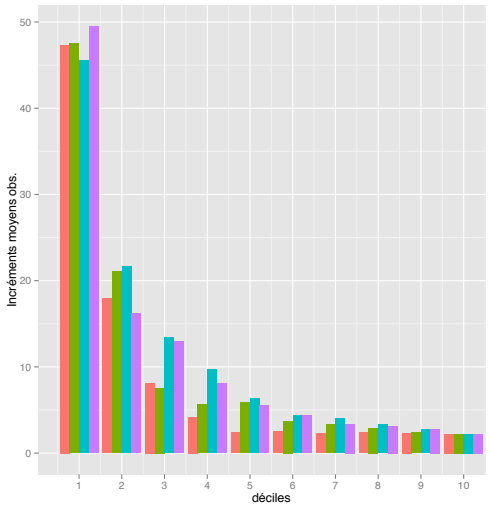


(C) $N = 5\,000$

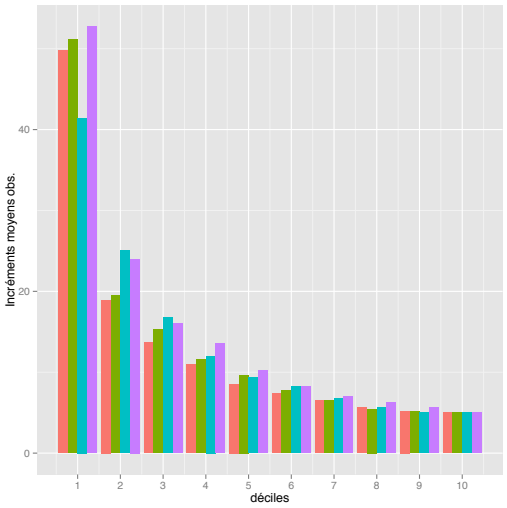


(D) $N = 10\,000$

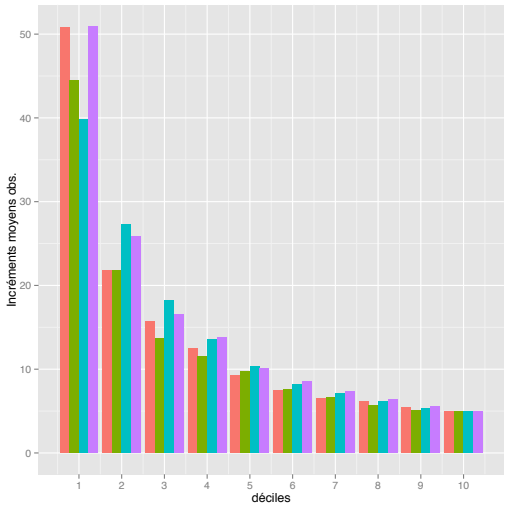
FIGURE 3.30. Modèle de Lai avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



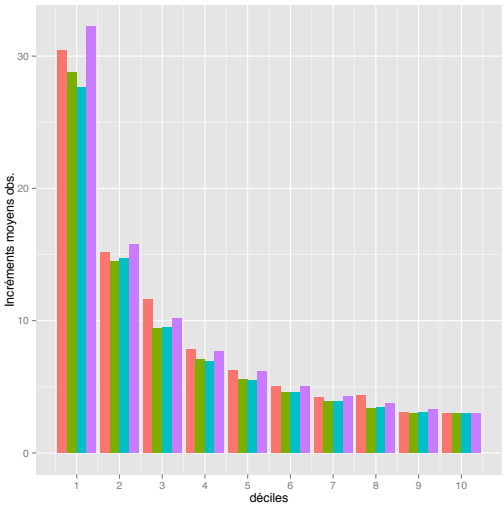
(A) $N = 250$



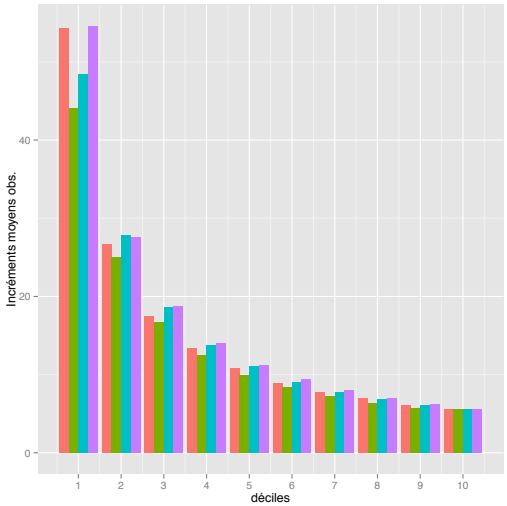
(B) $N = 500$



(C) $N = 1\,000$

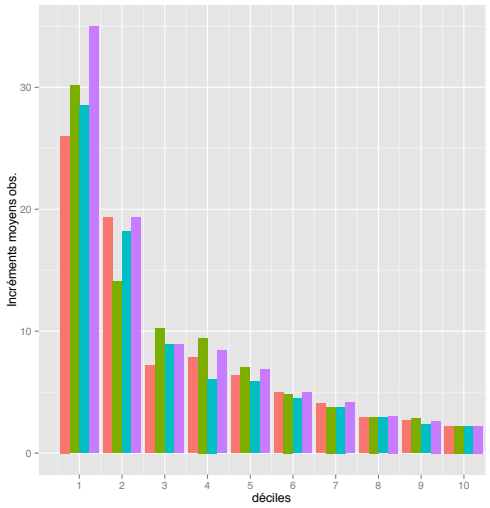


(D) $N = 5\,000$

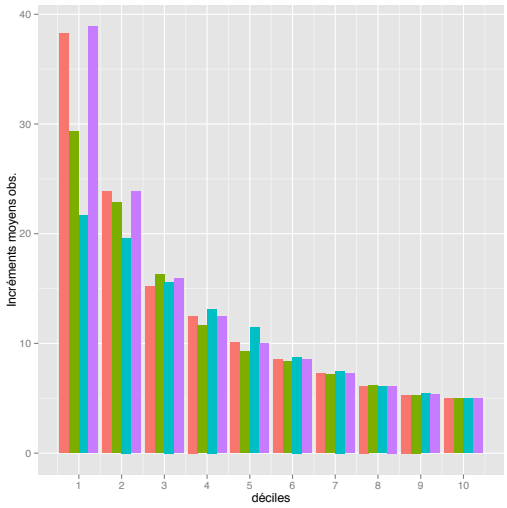


(E) $N = 10\,000$

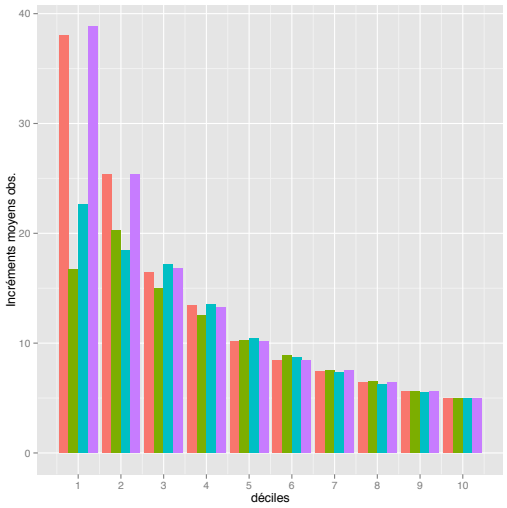
FIGURE 3.31. Modèle de Lo avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



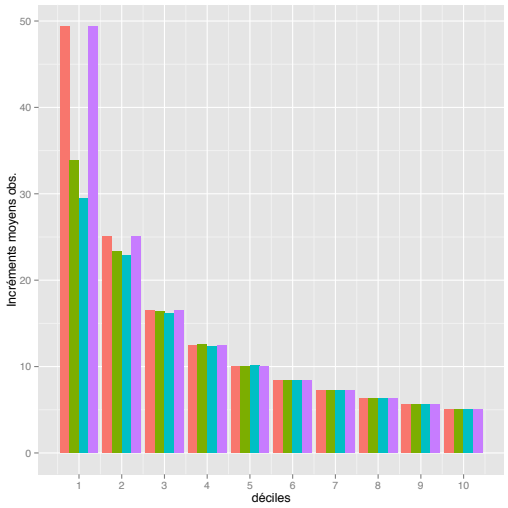
(A) $N = 250$



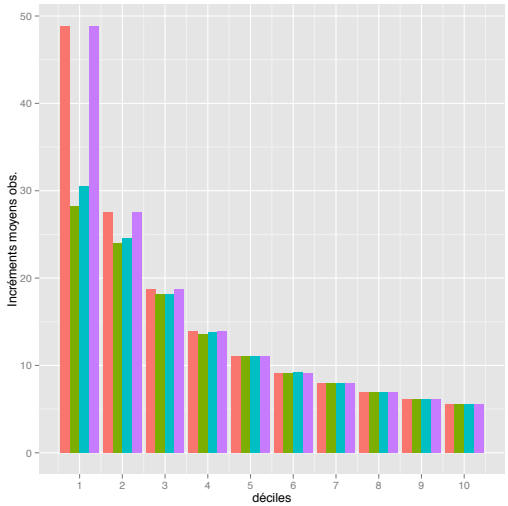
(B) $N = 500$



(C) $N = 1\,000$

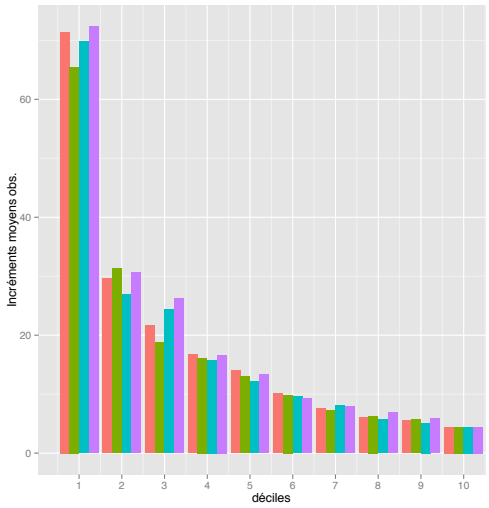


(D) $N = 5\,000$

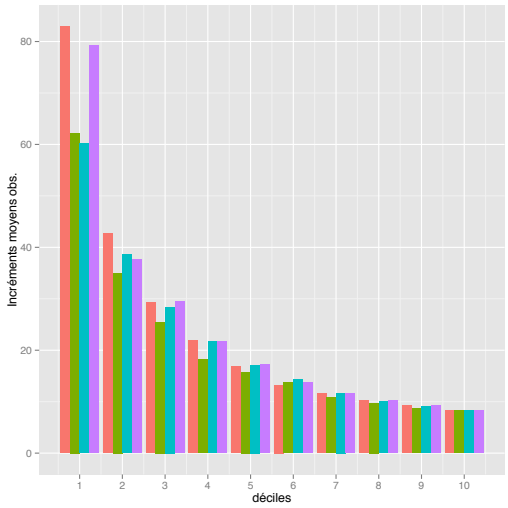


(E) $N = 10\,000$

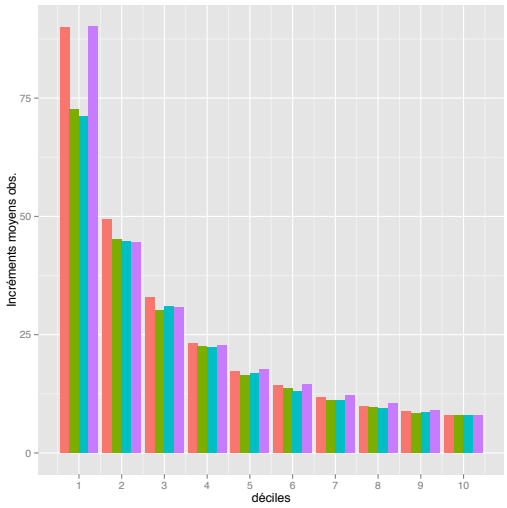
FIGURE 3.32. Modèle de Lai avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



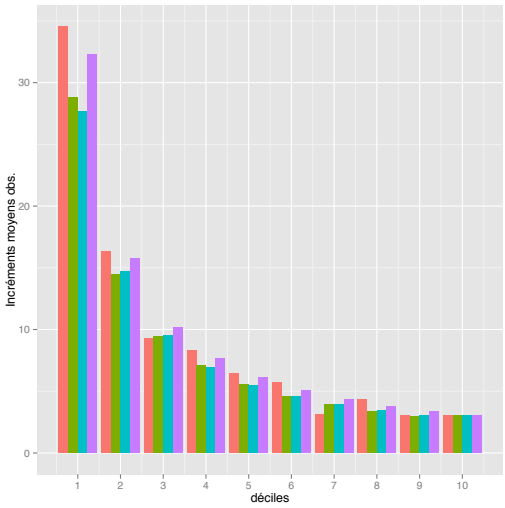
(A) $N = 250$



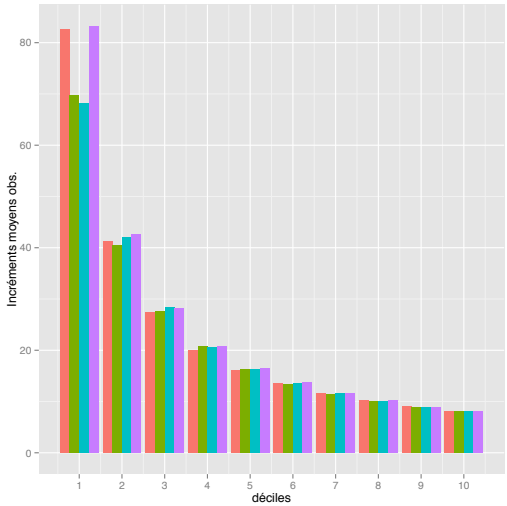
(B) $N = 500$



(C) $N = 1\,000$

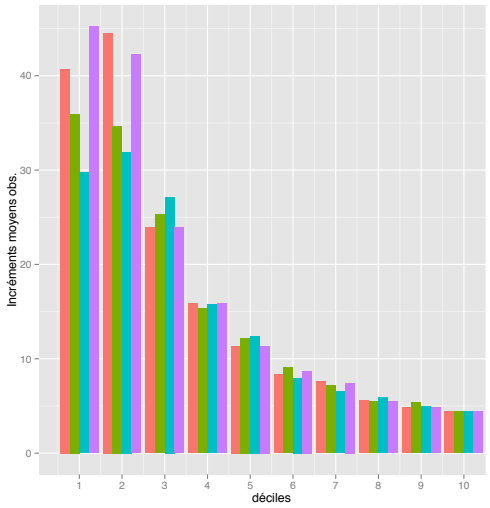


(D) $N = 5\,000$

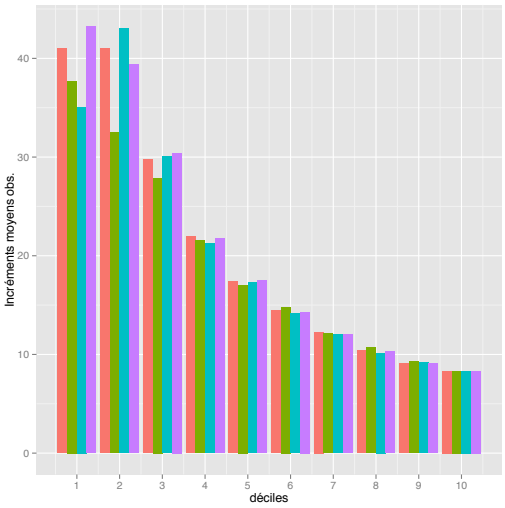


(E) $N = 10\,000$

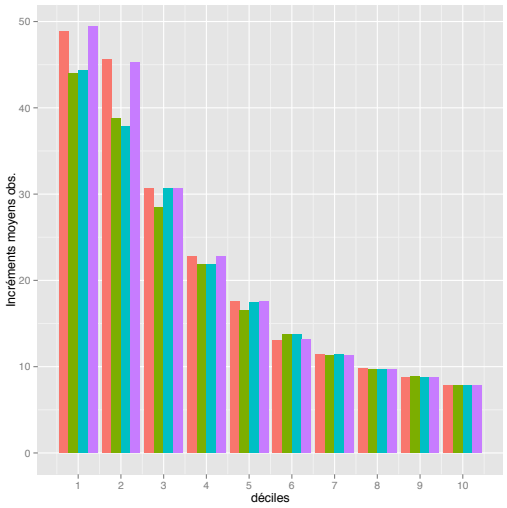
FIGURE 3.33. Modèle de Lo avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



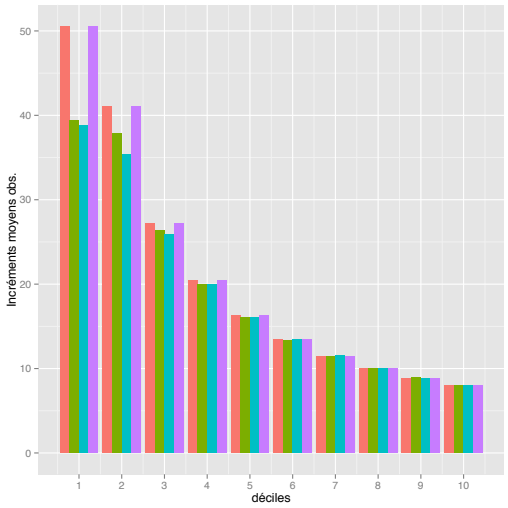
(A) $N = 250$



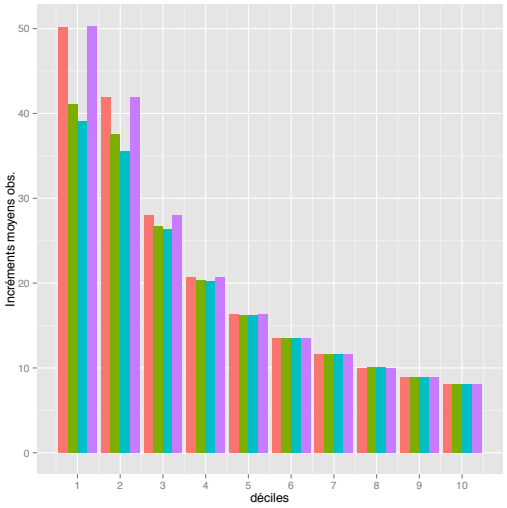
(B) $N = 500$



(C) $N = 1\,000$

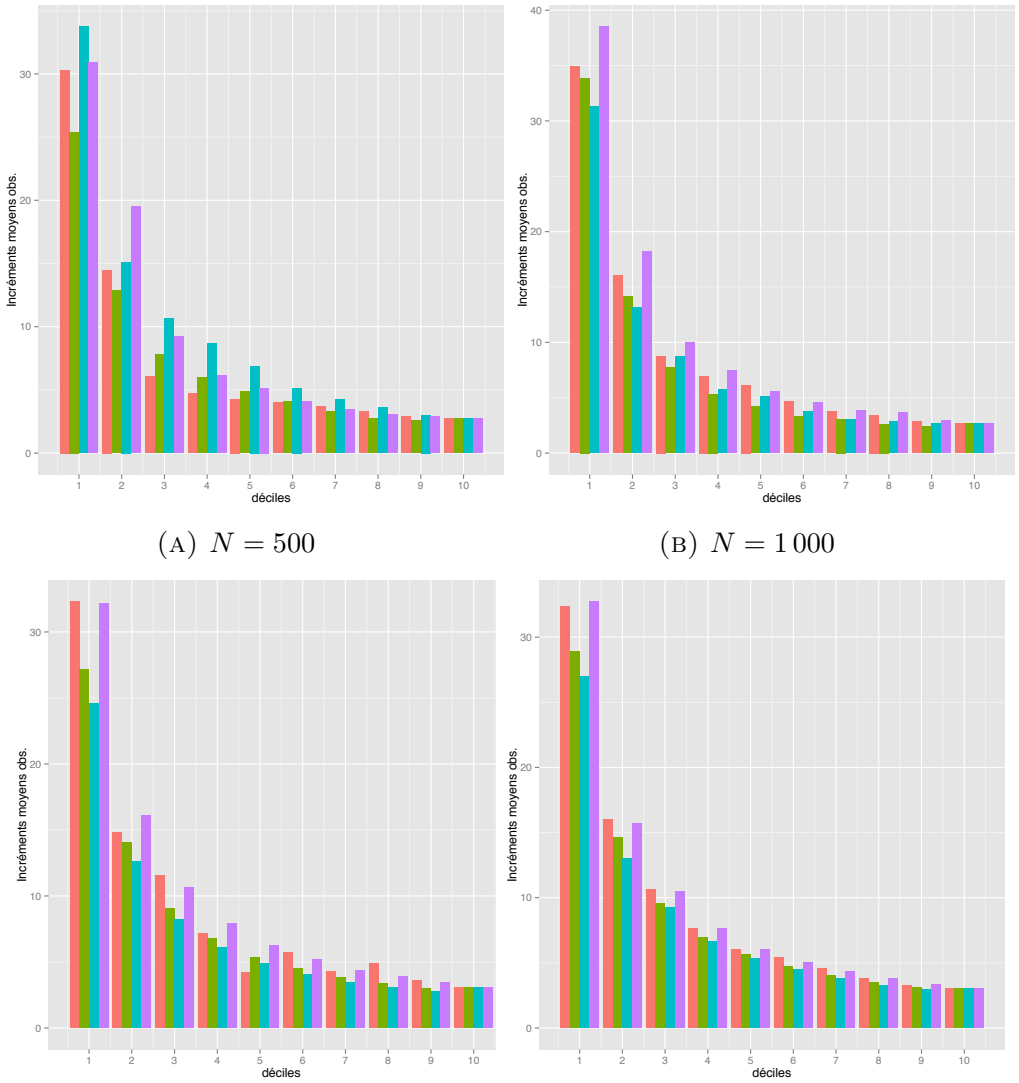


(D) $N = 5\,000$



(E) $N = 10\,000$

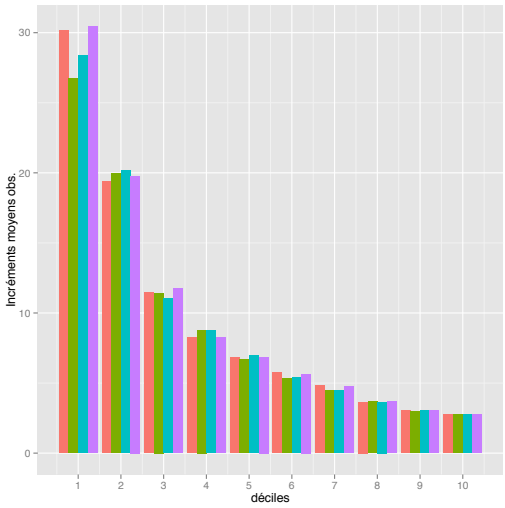
FIGURE 3.34. Modèle de Lai avec un groupe contrôle de 40% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



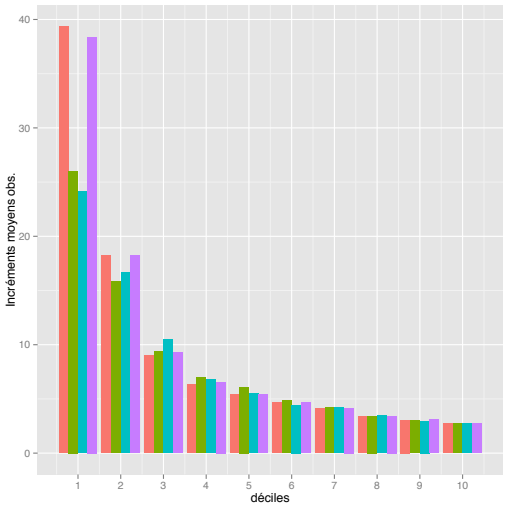
(A) $N = 500$ (B) $N = 1\,000$

(C) $N = 5\,000$ (D) $N = 10\,000$

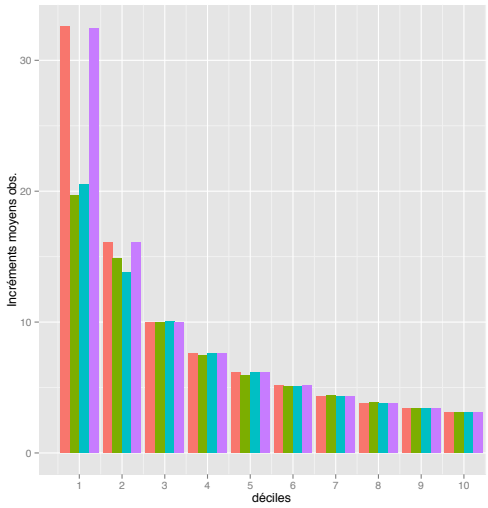
FIGURE 3.35. Modèle de Lo avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



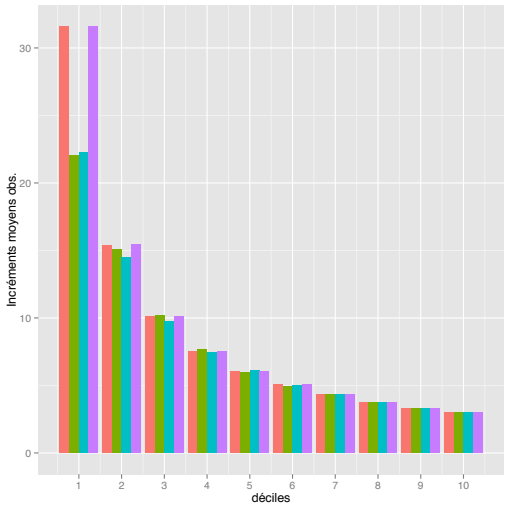
(A) $N = 500$



(B) $N = 1\,000$

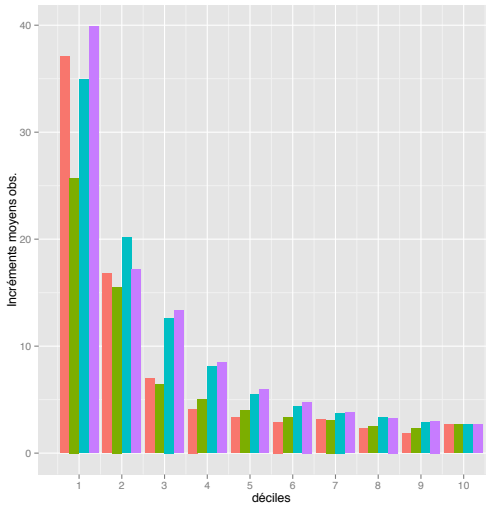


(C) $N = 5\,000$

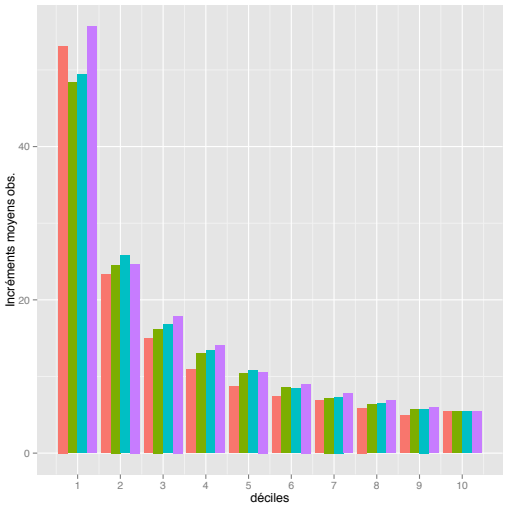


(D) $N = 10\,000$

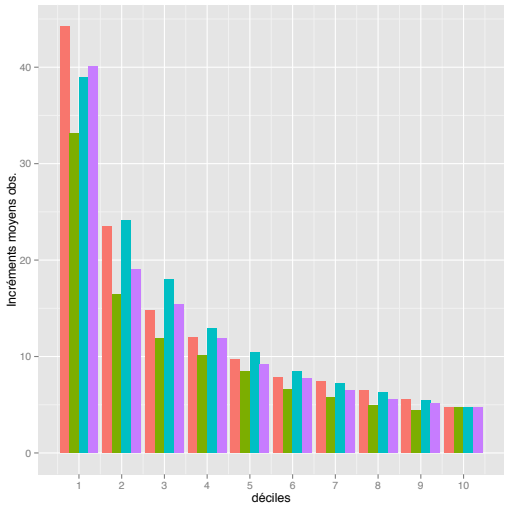
FIGURE 3.36. Modèle de Lai avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 8% pour les groupes contrôle et traitement respectivement



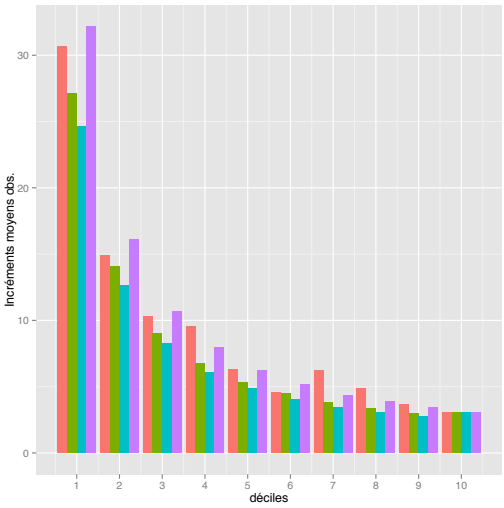
(A) $N = 250$



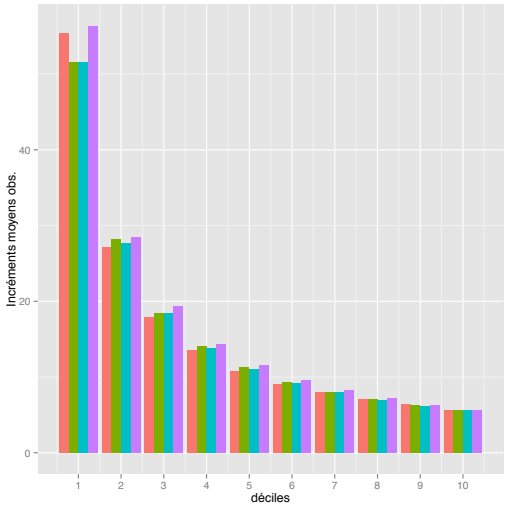
(B) $N = 500$



(C) $N = 1\,000$

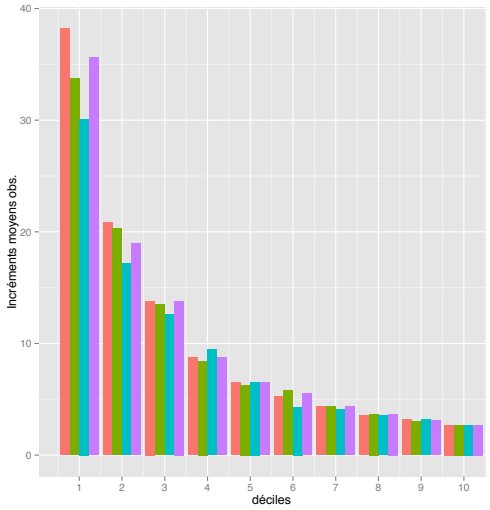


(D) $N = 5\,000$

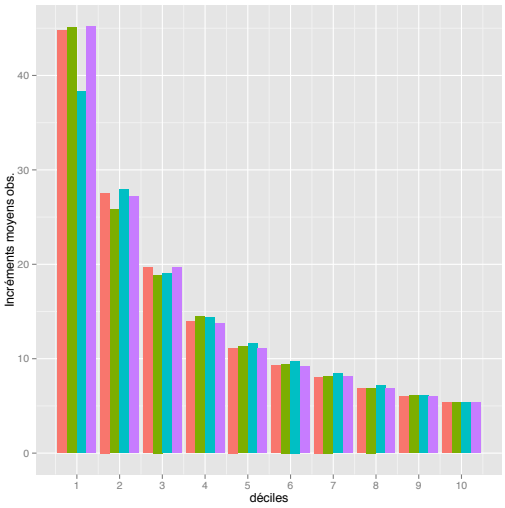


(E) $N = 10\,000$

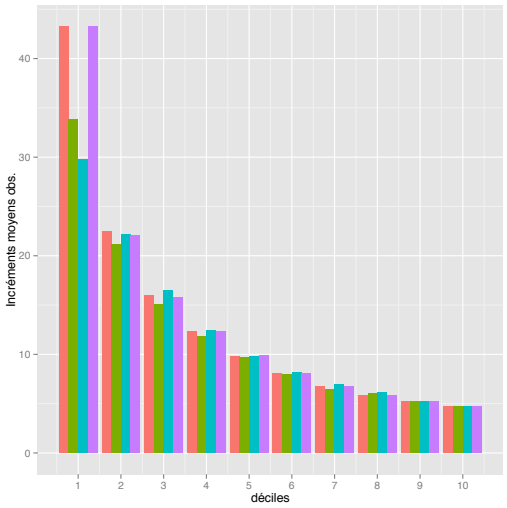
FIGURE 3.37. Modèle de Lo avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



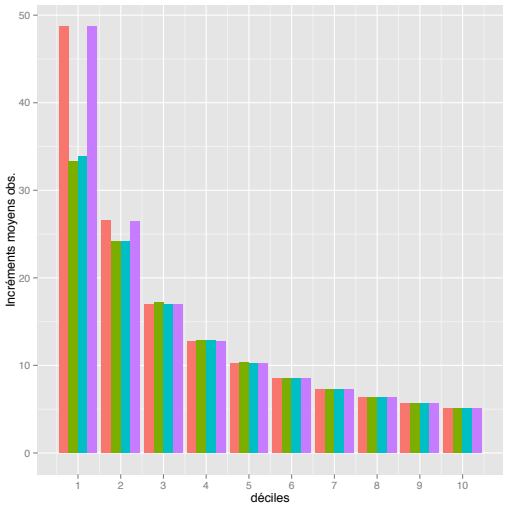
(A) $N = 250$



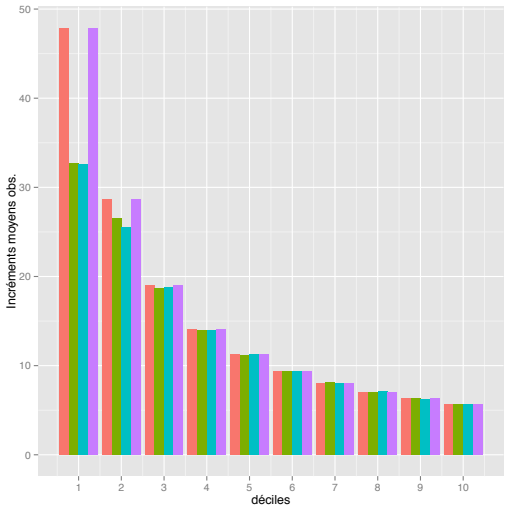
(B) $N = 500$



(C) $N = 1\,000$

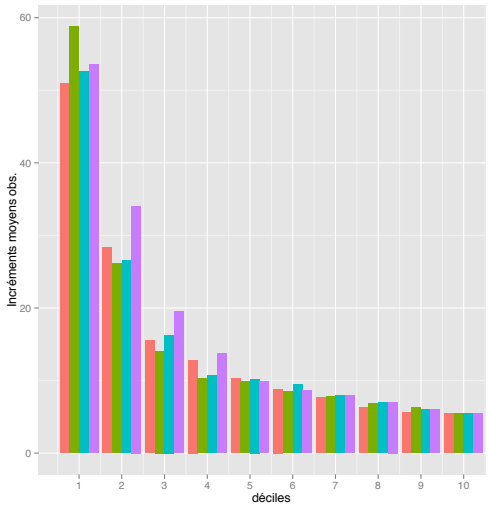


(D) $N = 5\,000$

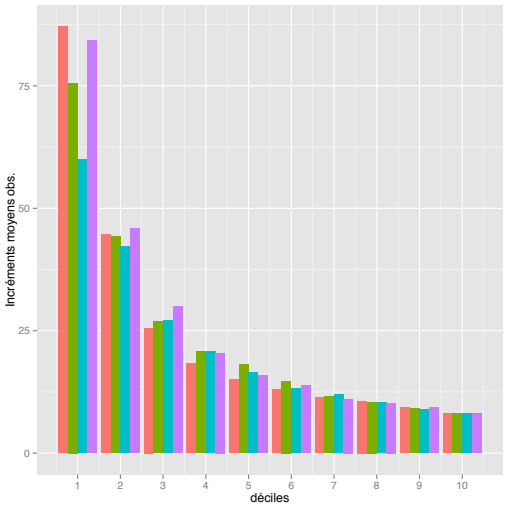


(E) $N = 10\,000$

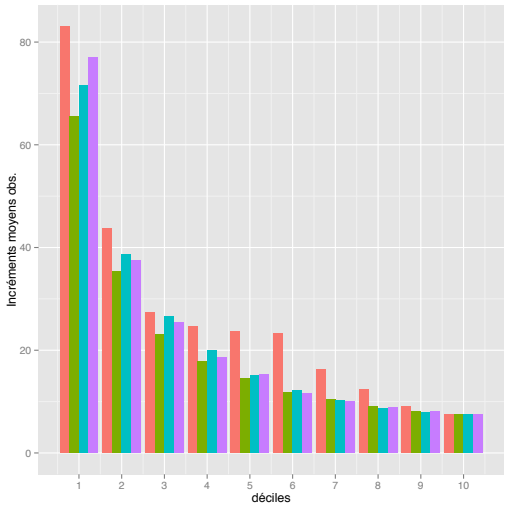
FIGURE 3.38. Modèle de Lai avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 10% pour les groupes contrôle et traitement respectivement



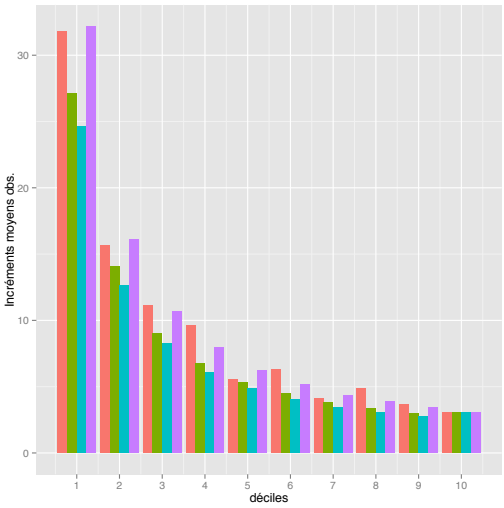
(A) $N = 250$



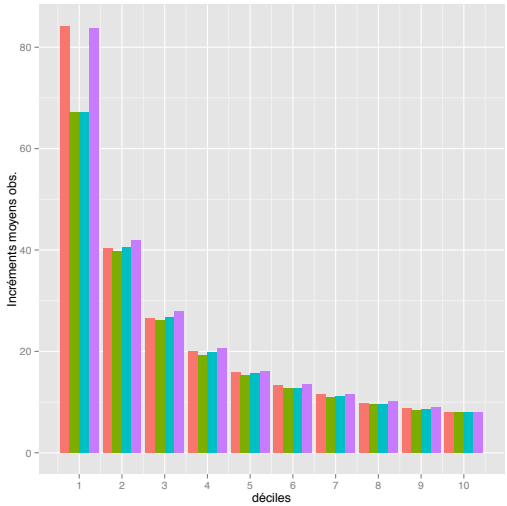
(B) $N = 500$



(C) $N = 1\,000$

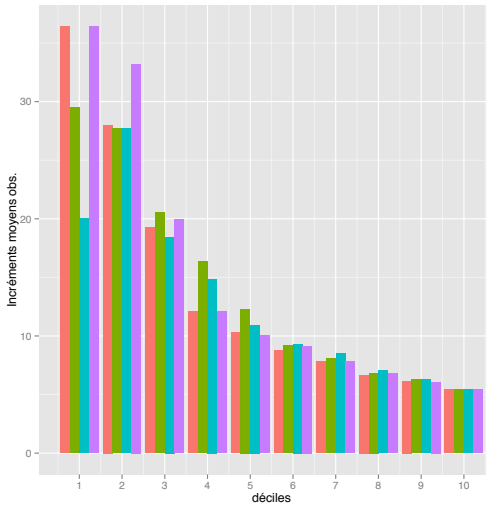


(D) $N = 5\,000$

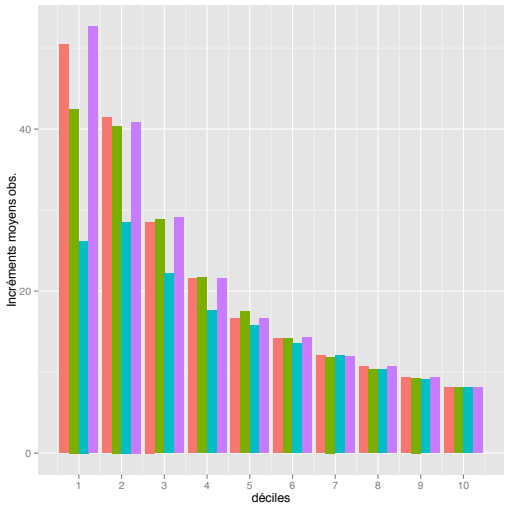


(E) $N = 10\,000$

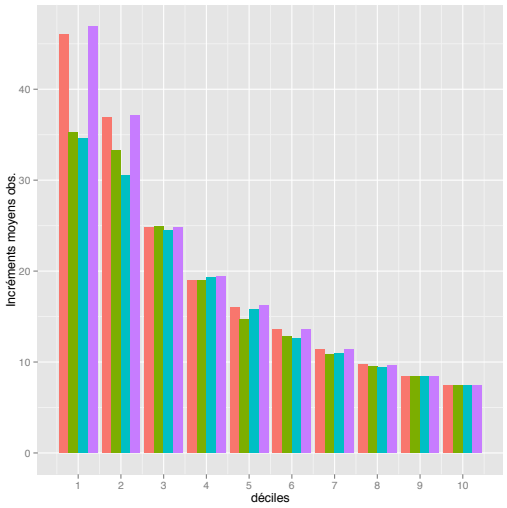
FIGURE 3.39. Modèle de Lo avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement



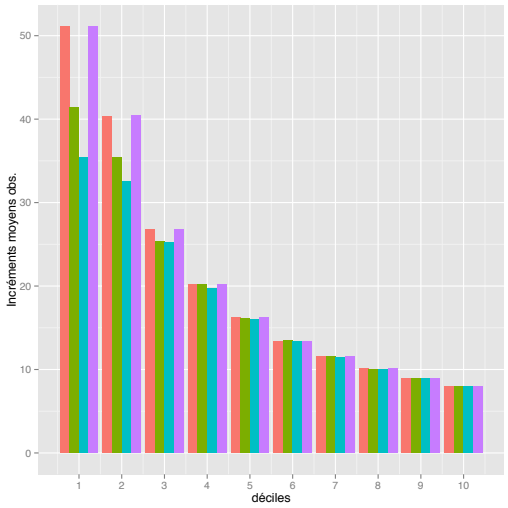
(A) $N = 250$



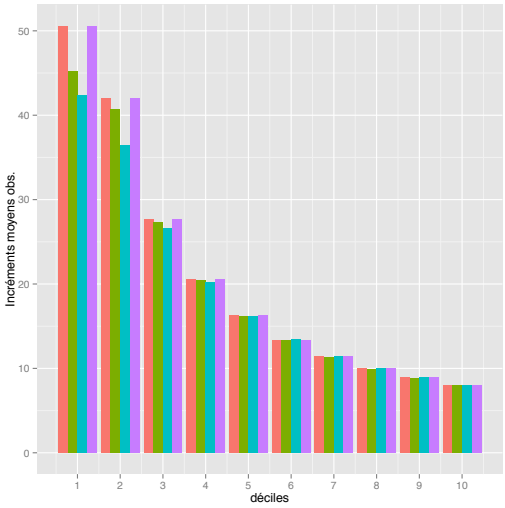
(B) $N = 500$



(C) $N = 1\,000$



(D) $N = 5\,000$



(E) $N = 10\,000$

FIGURE 3.40. Modèle de Lai avec un groupe contrôle de 50% et un taux de réponse positive de 5 et 13% pour les groupes contrôle et traitement respectivement