

Analyse multidimensionnelle appliquée

(c) Denis Larocque, Léo Belzile

version du 2021-12-05

Table des matières

1	Introduction	7
1.1	Survol du cours	7
1.2	Analyse exploratoire de données	10
2	Analyse factorielle exploratoire	21
2.1	Introduction	21
2.2	Rappels sur le coefficient de corrélation linéaire	22
2.3	Exemple de questionnaire	23
2.4	Description du modèle d'analyse factorielle	25
2.5	Estimation des facteurs	27
2.6	Choix du nombre de facteurs	29
2.7	Construction d'échelles à partir des facteurs	32
2.8	Compléments d'information	37
3	Analyse de regroupements	45
3.1	Introduction	45
3.2	Segmentation de seniors en voyage organisé	47
3.3	Exploration graphique préalable et analyse en composantes principales	48
3.4	Méthodes hiérarchiques	51
3.5	Calcul alternatif des distances pour le regroupement hiérarchique	61
3.6	Standardisation des variables	61
3.7	Autres mesures de dissemblance	64

3.8 Méthodes non hiérarchiques	65
3.9 Considérations pratiques	69
4 Sélection de variables et de modèles	71
4.1 Introduction	71
4.2 Sélection de variables et de modèles selon les buts de l'étude	72
4.3 Mieux vaut plus que moins	72
4.4 Trop beau pour être vrai	73
4.5 Principes généraux	73
4.6 Pénalisation et critères d'information	77
4.7 Division de l'échantillon et validation croisée	80
4.8 Cibler les clients pour l'envoi d'un catalogue	84
4.9 Recherche automatique du meilleur modèle	90
4.10 Méthodes classiques de sélection	93
4.11 Recherche séquentielle automatique limitée	97
4.12 Méthodes de régression avec régularisation	100
4.13 Moyenne de modèles	101
4.14 Évaluation de la performance	105
5 Régression logistique	109
5.1 Introduction	109
5.2 Modèle de régression logistique	109
5.3 Estimation des paramètres	112
5.4 Exemple du <i>Professional Rodeo Cowboys Association</i>	114
5.5 Classification et prédiction à l'aide de la régression logistique	125
5.6 Classification avec une matrice de gain	135
5.7 Sélection de variables en régression logistique	139
5.8 Performance des différents modèles pour l'exemple des clients cibles	143
5.9 Extensions du modèle de régression logistique à plus de deux catégories	147

6 Analyse de survie	155
6.1 Introduction	155
6.2 Fonctions de survie et de risque	157
6.3 Estimation d'une courbe de survie et de risque	158
6.4 Comparaison de deux courbes de survie	161
6.5 Modèle à risques proportionnels de Cox	165
6.6 Extensions du modèle de Cox	170
6.7 Risques non proportionnels	177
7 Données manquantes	181
7.1 Terminologie	181
7.2 Quelques méthodes	182
7.3 Example d'application de l'imputation	185
7.4 Valeurs manquantes dans un contexte de prédiction	190

Chapitre 1

Introduction

1.1 Survol du cours

1.1.1 Analyse factorielle exploratoire

On dispose de p variables X_1, \dots, X_p . Peut-on expliquer les interrelations (la structure de corrélation) entre ces variables à l'aide d'un certain nombre (moins de p) de facteurs latents (non observés) ?

L'analyse factorielle est souvent utilisée pour analyser des questionnaires (construction d'échelles) comme dans l'exemple suivant.

Exemple 1.1. Pour les besoins d'une enquête, on a demandé à 200 consommateurs adultes de répondre aux questions suivantes par rapport à un certain type de magasin :

Sur une échelle de 1 à 5,

1. pas important
2. peu important
3. moyennement important
4. assez important
5. très important

Pour vous, à quel point est-ce important...

1. que le magasin offre de bons prix tous les jours?
2. que le magasin accepte les cartes de crédit majeures (Visa, Mastercard)?
3. que le magasin offre des produits de qualité?
4. que les vendeurs connaissent bien les produits?

5. qu'il y ait des ventes spéciales régulièrement?
6. que les marques connues soient disponibles?
7. que le magasin ait sa propre carte de crédit?
8. que le service soit rapide?
9. qu'il y ait une vaste sélection de produits?
10. que le magasin accepte le paiement par carte de débit?
11. que le personnel soit courtois?
12. que le magasin ait en stock les produits annoncés?

Pouvons-nous identifier un nombre restreint de facteurs (concepts, dimensions) qui pourraient bien rendre compte de la structure de corrélation entre ces 12 variables?

Buts :

- Décrire et comprendre la structure de corrélation d'un ensemble de variables à l'aide d'un nombre restreint de concepts (appelés facteurs).
- Réduire le nombre de variables en créant une nouvelle variable par facteur. Ces nouvelles variables pourront par la suite être utilisées dans d'autres analyses (régression linéaire multiple par exemple).

1.1.2 Analyse de regroupements

On cherche à créer des groupes (« *clusters* ») d'individus homogènes en utilisant p variables X_1, \dots, X_p .

Exemple 1.2. Cette méthode est utilisée en marketing pour la **segmentation de marché**, qui consiste en

...définir des sous-groupes réunissant des consommateurs qui partagent les mêmes préférences ou qui réagissent de façon semblable à des variables de marketing¹

But :

- Combiner des sujets en groupes (interprétables) de telle sorte que les individus d'un même groupe soient les plus semblables possible par rapport à certaines caractéristiques et que les groupes soient les plus différents possible.

1.1.3 Sélection de variables et de modèles

Dans plusieurs situations, on doit développer un modèle de prévision. Par exemple, on pourrait devoir développer un modèle pour :

1. d'Astous, A. (2000). *Le projet de recherche en marketing*, 2e édition. Chenelière/McGraw-Hill.

- Déetecter les faillites des clients (ou des entreprises)
- Cibler les clients qui seront intéressés par une offre promotionnelle
- Déetecter les fraudes (par carte de crédit ou dans les rapports de revenus)
- Prévoir si un client va nous quitter.

Il y a en général plusieurs variables explicatives potentielles, et aussi plusieurs types de modèles possibles (régression linéaire, réseaux de neurones, arbres de régression ou de classification, etc.). Dans ce chapitre, nous verrons des principes généraux et des outils afin de sélectionner des modèles performants, ou bien un sous-ensemble de variables avec un bon pouvoir prévisionnel.

1.1.4 Régression logistique

On cherche à expliquer le comportement d'une variable binaire Y ($0 - 1$), à l'aide de p variables quelconques X_1, \dots, X_p .

Exemple 1.3. Une banque offre aux gens la possibilité de faire une demande de carte de crédit en ligne en promettant une approbation (conditionnelle) en quelques minutes seulement. Le tout est basé sur un modèle automatique de classification qui décide d'accorder ou non la carte ($Y = 1$ ou $Y = 0$) en fonction des réponses fournies par les clients potentiels à différentes questions comme : quel est votre revenu annuel brut (X_1), avez-vous d'autres cartes de crédit (X_2), êtes-vous locataire ou propriétaire (X_3), etc...

Buts :

- Comprendre comment et dans quelle mesure les variables X influencent la catégorie d'appartenance de Y .
- Développer un modèle pour faire de la classification, c'est-à-dire, prévoir la catégorie d'appartenance de Y pour un nouveau sujet à partir des variables X .

1.1.5 Analyse de survie

On s'intéresse au temps avant qu'un événement survienne. Par exemple :

- Temps qu'un client demeure abonné à un service offert par notre compagnie.
- Temps de survie d'un individu après avoir été diagnostiqué avec un certain type de cancer.
- Temps qu'un employé demeure au service de la compagnie.
- Temps qu'une franchise demeure en activité.
- Temps avant la faillite d'une entreprise (ou d'un particulier).
- Temps avant le prochain achat d'un client.

On observe chaque sujet jusqu'à ce que l'une des deux choses suivantes se produise : l'événement survient avant la fin de la période d'observation ou bien l'étude se termine et l'événement n'est

toujours pas survenu. Dans le premier exemple, l'événement correspond au fait d'interrompre son abonnement. On dispose donc d'une variable temps T pour chaque individu qui est soit censurée, soit non censurée. Si l'individu a expérimenté l'événement avant la fin de la période d'observation, la valeur de T est non censurée. Si l'événement n'est toujours pas survenu à la fin de la période d'observation, la valeur de T est censurée. Pour chaque individu, on dispose également d'un ensemble de variables explicatives X_1, \dots, X_p .

But :

- Étudier les effets des variables explicatives sur le temps de survie et obtenir des prévisions du temps de survie.

1.1.6 Données manquantes

Il arrive fréquemment d'avoir des valeurs manquantes dans notre échantillon.

Simplement ignorer les sujets avec des valeurs manquantes et faire l'analyse avec les autres sujets conduit généralement à des estimations biaisées et à de l'inférence invalide.

Dans ce chapitre, nous verrons une méthode très générale afin de traiter les données manquantes, l'imputation multiple. Nous verrons comment elle peut être utilisée dans un contexte d'inférence et dans un contexte de prévision.

1.2 Analyse exploratoire de données

L'analyse exploratoire, comme son nom l'indique, est une étape préliminaire à la modélisation servant à l'acquisition d'une meilleure compréhension des données. L'analyse exploratoire sert à nous assurer que notre analyse ou notre traitement de ces dernières est cohérent. Le but de l'analyse exploratoire graphique est d'extraire des informations utiles, le plus souvent par le biais d'une série de questions qui sont raffinées au fur et à mesure que progresse l'analyse. On s'intéresse particulièrement aux relations et interactions entre différentes variables et la distribution empirique de chaque variable. Les étapes majeures sont :

1. Formuler des questions sur les données
2. Chercher des réponses à ces questions à l'aide de statistiques descriptives, de tableaux de fréquence ou de contingence et de graphiques.
3. Raffiner nos questions, et utiliser les trouvailles pour peaufiner notre analyse

Dans un rapport, un résumé des caractéristiques les plus importantes devrait être inclus pour que le lecteur ou la lectrice puisse valider son interprétation des données.

1.2.1 Types de variables

- Une **variable** représente une caractéristique de la population d'intérêt, par exemple le sexe d'un individu, le prix d'un article, etc.
- une **observation**, parfois appelée donnée, est un ensemble de mesures collectées sous des conditions identiques, par exemple pour un individu ou à un instant donné.

Le choix de modèle statistique ou de test dépend souvent du type de variables collectées. Les variables peuvent être de plusieurs types : quantitatives (discrètes ou continues) si elles prennent des valeurs numériques, qualitatives (binaires, nominales ou ordinales) si elles sont décrites par un adjectif; je préfère le terme catégorielles, plus évocateur.



FIGURE 1.1 – Illustration par Allison Horst de variables continues (gauche) et discrètes (droite).

On distingue deux types de variables quantitatives :

- une variable discrète prend un nombre dénombrable de valeurs; ce sont souvent des variables de dénombrement ou des variables dichotomiques.
- une variable continue peut prendre (en théorie) une infinité de valeurs, même si les valeurs mesurées sont arrondies ou mesurées avec une précision limitée (temps, taille, masse, vitesse, salaire). Dans bien des cas, nous pouvons considérer comme continues des variables discrètes si elles prennent un assez grand nombre de valeurs.

Les variables catégorielles représentent un ensemble fini de possibilités. On les regroupe en deux types, pour lesquels on ne fera pas de distinction :

- nominales s'il n'y a pas d'ordre entre les modalités (sexe, couleur, pays d'origine) ou
- ordinaire (échelle de Likert, tranche salariale).

La codification des modalités des variables catégorielles est arbitraire; en revanche, on préservera l'ordre lorsqu'on représentera graphiquement les variables ordinaires. Lors de l'estimation, chaque variable catégorielle doit être transformée en un ensemble d'indicateurs binaires : il est donc essentiel de déclarer ces dernières dans votre logiciel statistique, surtout si elles sont encodées dans la base de données à l'aide de valeurs entières.

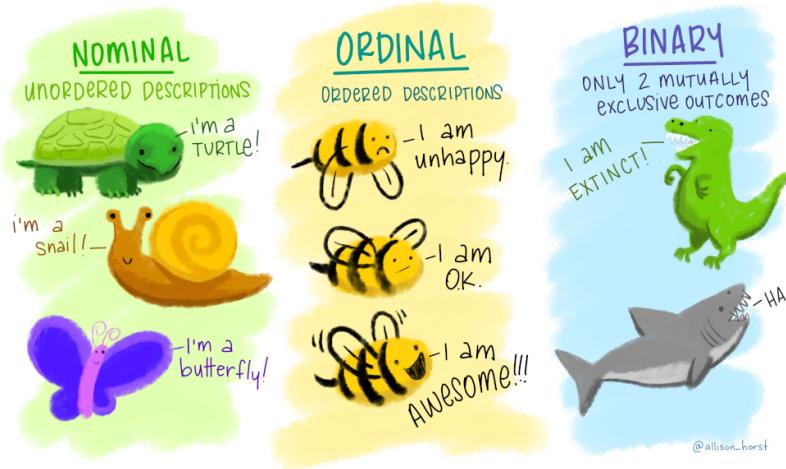


FIGURE 1.2 – Illustration par Allison Horst de variables catégorielles nominales (gauche), ordinaires (centre) et binaires (droite).

1.2.2 Validation des données.

Avant de regarder les données, il est souvent utile de se plonger dans la description de la base de données. Il n'est pas rare que cette dernière contienne des informations pertinentes sur la codification des données, par exemple

- telle variable catégorielle est stockée avec des valeurs entières et les étiquettes ne sont disponibles que dans la description.
- des valeurs manquantes sont encodées avec -1 (pour les variables positives) ou 999.
- une variable est une fonction, transformation ou combinaison d'autres variables.

1.2.3 Graphiques

Le principal type de graphique pour représenter la distribution d'une variable catégorielle est le diagramme en bâtons, dans lequel la fréquence de chaque catégorie est présentée sur l'axe des ordonnées (y) en fonction de la modalité, sur l'axe des abscisses (x), et ordonnées pour des variables ordinaires. Cette représentation est en tout point supérieur au diagramme en camembert, une emprise répandue qui devrait être honnie (notamment parce que l'humain juge mal les différences d'aires, qu'une simple rotation change la perception du graphique et qu'il est difficile de mesurer les proportions) — ce n'est pas de la tarte!

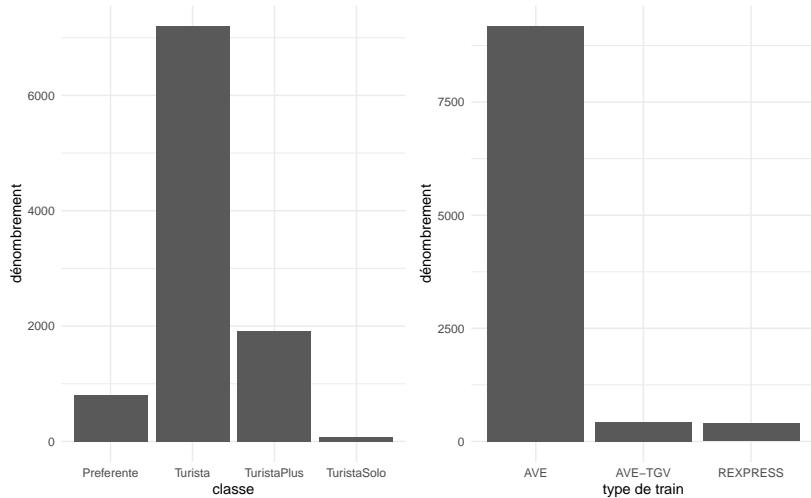


FIGURE 1.3 – Diagramme en bâtons pour la classe des billets de trains du jeu de données Renfe.

Puisque les variables continues peuvent prendre autant de valeurs distinctes qu'il y a d'observations, on ne peut simplement compter le nombre d'occurrence par valeur unique. On regroupera plutôt dans un certain nombre d'intervalle, en discrépitant l'ensemble des valeurs en classes pour obtenir un histogramme. Le nombre de classes dépendra du nombre d'observations si on veut que l'estimation ne soit pas impactée par le faible nombre d'observations par classe : règle générale, le nombre de classes ne devrait pas dépasser \sqrt{n} , où n est le nombre d'observations de l'échantillon. On obtiendra la fréquence de chaque classe, mais si on normalise l'histogramme (de façon à ce que l'aire sous les bandes verticales égale un), on obtient une approximation discrète de la fonction de densité. Faire varier le nombre de classes permet parfois de faire apparaître des caractéristiques de la variable (notamment la multimodalité, l'asymétrie et les arrondis).

Puisque qu'on groupe les observations en classe pour tracer l'histogramme, il est difficile de voir l'étendue des valeurs que prenne la variable : on peut rajouter des traits sous l'histogramme pour représenter les valeurs uniques prises par la variable, tandis que la hauteur de l'histogramme nous renseigne sur leur fréquence relative.

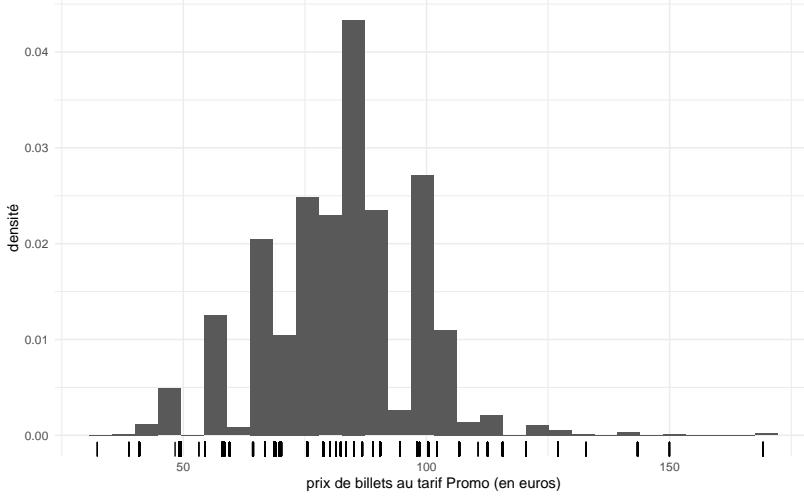


FIGURE 1.4 – Histogramme du prix des billets au tarif Promo de trains du jeu de données Renfe

Une boîte à moustaches représente graphiquement cinq statistiques descriptives.

- La boîte donne les 1e, 2e et 3e quartiles q_1, q_2, q_3 . Il y a donc 50% des observations sont au-dessus/en-dessous de la médiane q_2 qui sépare en deux la boîte.
- La longueur des moustaches est moins de 1.5 fois l'écart interquartile $q_3 - q_1$ (tracée entre 3e quartile et le dernier point plus petit que $q_3 + 1.5(q_3 - q_1)$, etc.)
- Les observations au-delà des moustaches sont encerclées. Notez que plus le nombre d'observations est élevé, plus le nombres de valeurs extrême augmente. C'est un défaut de la boîte à moustache, qui a été conçue pour des jeux de données qui passeraient pour petits selon les standards actuels.

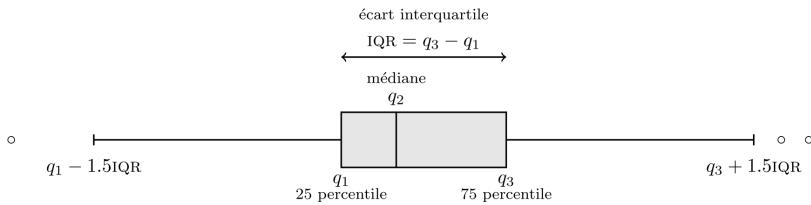


FIGURE 1.5 – Boîte à moustache.

On peut représenter la distribution d'une variable réponse continue en fonction d'une variable catégorielle en traçant une boîte à moustaches pour chaque catégorie et en les disposant côté-à-côte. Une troisième variable catégorielle peut être ajoutée par le biais de couleurs, comme dans la Figure 1.6.

Si on veut représenter la covariabilité de deux variables continues, on utilise un nuage de points où

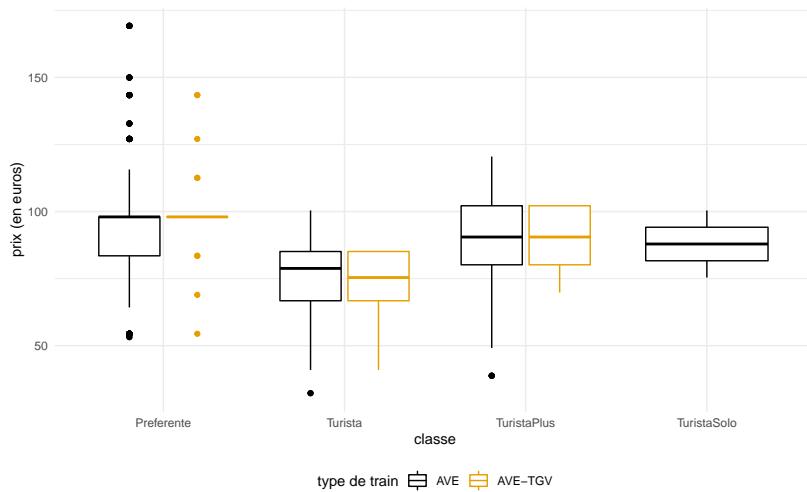


FIGURE 1.6 – Boîte à moustaches du prix des billets au tarif Promo en fonction de la classe pour le jeu de données Renfe.

chaque variable est représentée sur un axe et chaque observation donne la coordonnée des points. Si la représentation graphique est dominée par quelques valeurs très grandes, une transformation des données peut être utile : vous verrez souvent des données positives à l'échelle logarithmique.

Plutôt que de décrire plus en détail le processus de l'analyse exploratoire, on présente un exemple qui illustre le cheminement habituel sur les données de trains de la Renfe introduites précédemment.

1.2.4 Analyse exploratoire des billets de trains Renfe

La première étape consisterait à lire la description de la base de données. Le jeu de données `renfe` contient les variables suivantes :

- `prix` : prix du billet (en euros);
- `dest` : indicateur binaire du trajet, soit de Barcelone vers Madrid (0) ou de Madrid vers Barcelone (1);
- `tarif` : variable catégorielle indiquant le tarif du billet, un parmi `AdultoIda`, `Promo` et `Flexible`;
- `classe` : classe du billet, soit `Preferente`, `Turista`, `TuristaPlus` ou `TuristaSolo`;
- `type` : variable catégorielle indiquant le type de train, soit Alta Velocidad Española (AVE), soit Alta Velocidad Española conjointement avec TGV (un partenariat entre la SNCF et Renfe pour les trains à destination ou en provenance de Toulouse) AVE-TGV, soit les trains régionaux REXPRESS; seuls les trains étiquetés AVE ou AVE-TGV sont des trains à grande

vitesse.

- duree : longueur annoncée du trajet (en minutes) ;
- jour entier indiquant le jour de la semaine du départ allant de dimanche (1) à samedi (7).

Il n'y a pas de valeurs manquantes et un aperçu des données (`head(renfe)`) montre qu'elles sont en format long, ce qui veut dire que chaque ligne contient une seule valeur pour la variable réponse, ici le prix d'un billet de train. On entame l'analyse exploratoire avec des questions plutôt vagues, par exemple

1. Quels sont les facteurs déterminant le prix et le temps de parcours ?
2. Est-ce que le temps de parcours est le même pour tous les types de train ?
3. Quelles sont les caractéristiques distinctives des types de train ?
4. Quelles sont les principales différences entre les tarifs ?

À l'exception de `prix` et de `duree`, toutes les variables explicatives sont catégorielles. La variable `jour` prend des valeurs entre 1 et 7; s'en souvenir pour éviter les mauvaises surprises ultérieures. En analysant le nombre de trains dans les catégories, on remarque qu'il y a autant de billets de type `REXPRESS` que le nombre de billets au tarif `AdultoIda`. On peut faire le décompte par catégorie avec un tableau de contingence, qui compte le nombre respectif dans chaque sous-catégorie. Dans la base de données Renfe, tous les billets pour les `RegioExpress` sont vendus au tarif `AdultoIda` en classe `Turista`. Le nombre de billets est minime, à peine 397 sur 10000. Cela suggère une nouvelle question : pourquoi ces trains sont-ils si peu populaires ?

On remarque également que seulement 17 temps de parcours sont affichés sur les billets. On peut donc penser que la durée affichée sur le billet (en minutes) est le temps de trajet annoncé. La majeure partie (15 sur 17) des temps de parcours sont sous la barre des 3h15, hormis deux qui dépassent les 9h ! Selon Google Maps, les deux villes sont distantes de 615km par la route, 500km à vol d'oiseau. Cela implique que, vraisemblablement, certains trains dépassent les 200km/h, tandis que d'autres vont plutôt à 70km/h. Quels sont ces trains plus lents ? La variable `type` codifie probablement ce fait, et permet de voir que ce sont les trains `RegioExpress` qui sont dans cette catégorie.

Aller de Madrid à Barcelone à l'aide d'un train régulier prend 18 minutes de plus. Avec plus de 9h de trajet, pas étonnant donc que ces billets soient peu courus. Encore plus frappant, on note que le prix des billets est fixe : 43.25 euros peu importe que le trajet soit aller ou retour. C'est probablement la trouvaille la plus importante jusqu'à maintenant, car les billets de train de type `RegioExpress` ne forment pas un échantillon : il n'y a aucune variabilité ! On aurait pu également découvrir cette anomalie en traçant une boîte à moustaches du prix en fonction du type de train.

On pourrait soupçonner que les trains étiquetés `AVE` soient plus rapides, sachant que c'est l'acronyme de *Alta Velocidad Española*, littéralement haute vitesse espagnole. Qu'en est-il des distinctions entre les deux types de trains étiquetés `AVE` ? Selon le site de la SNCF, les trains `AVE-TGV` sont des partenariats entre la Renfe et la SNCF et effectuent des liaisons entre la France et l'Espagne.

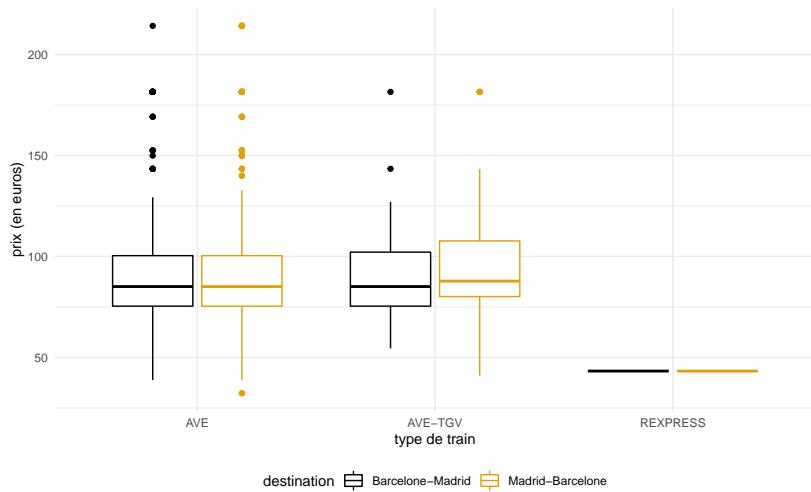


FIGURE 1.7 – Boîte à moustaches du prix de billets de train de Renfe en fonction de la destination et du type de train.

Les prix sont beaucoup plus élevés, en moyenne plus de deux fois plus que les trains régionaux. Les écarts de prix importants (l'écart type est de 20 euros) indique qu'il y a peut-être d'autres sources d'hétérogénéité, mais on pourrait soupçonner que la Renfe pratique la tarification dynamique. Il y a un seul temps de parcours prévu pour les trains AVE-TGV. On ne note pas de différence de prix notable selon la direction ou le type de train grande vitesse, mais peut-être que les tarifs ou la classe disponibles diffèrent selon que le train ou non est en partenariat avec la compagnie française.

On n'a pas encore considéré le tarif et la classe des billets, hormis pour les trains RegioExpress. On voit dans la Figure 1.9 une forte différence dans l'hétérogénéité des prix selon le tarif; le tarif Promo prend plusieurs valeurs distinctes, tandis que les tarifs AdultoIda et Flexible semblent ne prendre que quelques valeurs. La première classe (Preferente) est plus chère et il y a moins d'observations dans ce groupe. La classe Turista est la classe la moins dispendieuse et la plus populaire. TuristaPlus offre plus de confort, tandis que TuristaSolo permet d'obtenir un siège individuel.

Côté tarif, Promo et PromoPlus permettent d'obtenir des rabais pouvant aller jusqu'à respectivement 70% et 65%. Les annulations et changements ne sont pas possibles avec Promo, mais disponibles avec PromoPlus moyennant une pénalité équivalente à 30-20% du prix du billet. Le tarif Flexible est disponible au même prix que les billets réguliers, avec des bénéfices additionnels.

On note que la répartition des prix pour les billets de classe Flexible est inhabituelle. Notre boîte à moustaches est écrasée et l'écart interquartile semble nul, même si quelques valeurs inexplicables sont aussi présentes. L'écrasante majorité des billets Flexibles sont en classe Turista, donc ça pourrait être dû à un (trop) faible nombre de billets dans chaque catégorie. On peut rejeter cette hypothèse.

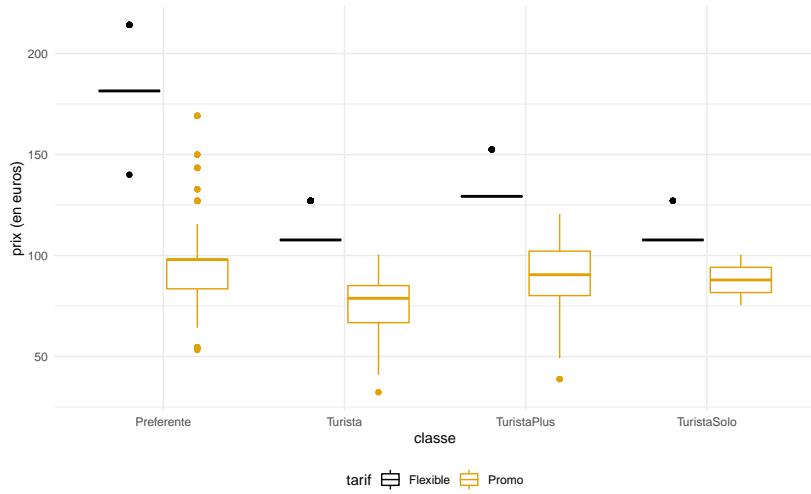


FIGURE 1.8 – Boîte à moustaches du prix en fonction du tarif et de la classe de billets de trains à haute vitesse de la Renfe.

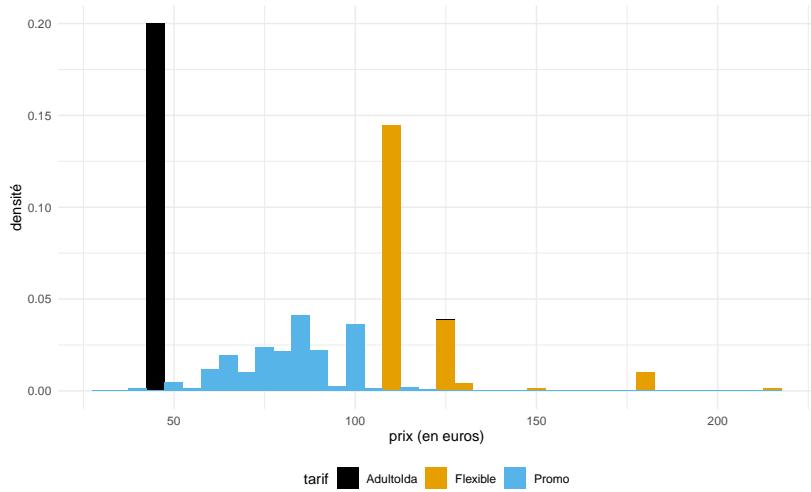


FIGURE 1.9 – Histogrammes du prix en fonction du tarif de billets de trains de la Renfe.

TABLE 1.1 – Nombre de billets au tarif Flexible selon le prix de vente.

prix	classe	n
108	Turista	1050
108	TuristaSolo	67
127	Turista	285
127	TuristaSolo	9
129	TuristaPlus	31
140	Preferente	2
152	TuristaPlus	10
182	Preferente	78
214	Preferente	12

thèse en calculant le nombre de trains au tarif Flexible pour les différents types de billets, comme dans le Tableau 1.1. Ni la durée, ni le type de train, ni la destination n'expliquent pas pourquoi le prix de certains billets Flexibles est plus faible ou élevés. Le prix des billets Promo est plus faible, et les billets au tarif Preferente (la première classe) sont plus élevés.

On peut résumer notre brève analyse exploratoire :

- plus de 91% des trains sont des trains à grande vitesse AVE.
- le temps de trajet dépend du type de train : les trains à grande vitesse mettent 3h20 au maximum pour relier Madrid et Barcelone.
- les temps de trajets sont ceux annoncés (variable discrète avec 17 valeurs uniques, dont 13 pour les trains AVE)
- le prix de trains RegioExpress est fixe (43.25€) ; tous ces billets sont dans la classe Turista et au tarif Adulato Ida. 57% de ces trains vont de Barcelone à Madrid. La durée du trajet pour les RegioExpress est de 9h22 de Barcelona à Madrid, 18 minutes de plus que dans l'autre direction.
- les billets en classe Preferente sont plus chers et moins fréquents. La classe Turista est la classe la moins dispendieuse et la plus populaire. TuristaPlus offre plus de confort, tandis que TuristaSolo permet d'obtenir un siège individuel.
- selon le site web de la Renfe, les billets au tarif Flexible « viennent avec des offres additionnelles qui permettent aux passagers d'échanger leurs billets ou d'annuler si les trains manquent leurs trains. » ; en contrepartie, ces billets sont plus chers et leur tarif est fixe sauf une poignée de billets dont le prix reste inexpliqué.
- la distribution des prix des billets de TGV au tarif Promo est plus ou moins symétrique, tandis que les billets au tarif Flexible apparaissent tronqués à gauche (le prix minimum pour ces billets est 107.7€ dans l'échantillon).
- la Renfe pratique la tarification dynamique pour les billets au tarif promotionnel Promo : ces

derniers peuvent être jusqu'à 70% moins chers que les billets à prix régulier lorsqu'achetés via l'agence officielle ou le site de Renfe. Ces billets ne peuvent être ni remboursés, ni échangés.

- il n'y a pas d'indication à effet de quoi les prix varient selon la direction du trajet.

1.2.5 Commentaire sur les graphiques

Si vous incluez un graphique (ou un tableau), il est important d'ajouter une légende qui décrit le graphique et le résume, les noms de variables (avec les unités) sur les axes, mais aussi de soigner le rendu et le formatage pour obtenir un produit fini propre, lisible et cohérent : en particulier, votre description devrait coïncider avec le rendu. Votre graphique raconte une histoire, aussi prenez-soin que cette dernière soit nécessaire et attrayante.

Chapitre 2

Analyse factorielle exploratoire

2.1 Introduction

On dispose de p variables X_1, \dots, X_p .

- Y a-t-il des groupements de variables?
- Est-ce que les variables faisant partie d'un groupement semblent mesurer certains aspects d'un facteur commun (non observé)?

Un tel groupement peut être détecté si plusieurs variables sont très corrélées entre elles. Est-ce que la structure de corrélation entre les p variables peut être expliquée à l'aide d'un nombre restreint de facteurs?

Exemple de facteurs : Habiléty quantitative, habileté sociale, importance accordée à la qualité du service, importance accordée à la loyauté, habileté de leader, etc...

L'analyse factorielle est aussi une méthode de réduction du nombre de variables. En effet, une fois qu'on a identifié les facteurs, on peut remplacer les variables individuelles par un résumé pour chaque facteur (qui est souvent la moyenne des variables qui font partie du facteur).

Pour faire une analyse factorielle, la taille d'échantillon devrait être conséquente : le nombre d'entrées dans la base de données est np , et le nombre de paramètres de la matrice de covariance à estimer est $O(p^2)$. Plusieurs références, essentiellement arbitraires, suggèrent d'avoir une taille d'échantillon entre cinq et 20 fois le fois le nombre de variables, ou bien un nombre minimal de 100 à 1000 observations. Des études de simulations suggèrent que la taille critique dépend des paramètres, communalités, distribution des données, etc.

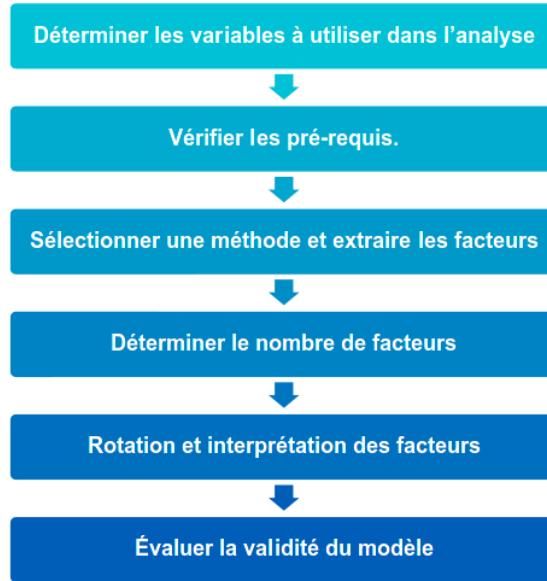


FIGURE 2.1 – Schéma des différents étapes de l'analyse factorielle exploratoire.

2.2 Rappels sur le coefficient de corrélation linéaire

On veut examiner la relation entre deux variables X_j et X_k et on dispose de n couples d'observations, où $x_{i,j}$ (respectivement $x_{i,k}$) est la valeur de la variable X_j (X_k) pour le i e individu.

Le coefficient de corrélation linéaire entre X_j et X_k , que l'on note $r_{j,k}$, cherche à mesurer la force de la relation linéaire entre deux variables, c'est-à-dire à quantifier à quel point les observations sont alignées autour d'une droite. Le coefficient de corrélation est

$$r_{j,k} = \frac{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k)}{\{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 \sum_{i=1}^n (x_{i,k} - \bar{x}_k)^2\}^{1/2}}$$

Les propriétés les plus importantes du coefficient de corrélation linéaire r sont les suivantes :

- 1) $-1 \leq r \leq 1$;
- 2) $r = 1$ (respectivement $r = -1$) si et seulement si les n observations sont exactement alignées sur une droite de pente positive (négative). C'est-à-dire, s'il existe deux constantes a et $b > 0$ ($b < 0$) telles que $y_i = a + bx_i$ pour tout $i = 1, \dots, n$.

Règle générale,

- Plus la corrélation est près de 1, plus les points auront tendance à être alignés autour d'une droite de pente positive. Par conséquent, plus la valeur de X augmente, plus celle de Y aura tendance à augmenter et vice-versa.
- Plus la corrélation est près de -1 , plus les points auront tendance à être alignés autour d'une droite de pente négative. Par conséquent, plus la valeur de X augmente, plus celle de Y aura tendance à diminuer et vice-versa.
- Lorsque la corrélation est presque nulle, les points n'auront pas tendance à être alignés autour d'une droite. Il est très important de noter que cela n'implique pas qu'il n'y a pas de relation entre les deux variables. Cela implique seulement qu'il n'y a pas de **relation linéaire** entre les deux variables.

2.3 Exemple de questionnaire

Le questionnaire suivant porte sur une étude dans un magasin. Pour les besoins d'une enquête, on a demandé à 200 consommateurs adultes de répondre aux questions suivantes par rapport à un certain type de magasin sur une échelle de 1 à 5, où

1. pas important
2. peu important
3. moyennement important
4. assez important
5. très important

Pour vous, à quel point est-ce important...

1. que le magasin offre de bons prix tous les jours?
2. que le magasin accepte les cartes de crédit majeures (Visa, Mastercard)?
3. que le magasin offre des produits de qualité?
4. que les vendeurs connaissent bien les produits?
5. qu'il y ait des ventes spéciales régulièrement?
6. que les marques connues soient disponibles?
7. que le magasin ait sa propre carte de crédit?
8. que le service soit rapide?
9. qu'il y ait une vaste sélection de produits?
10. que le magasin accepte le paiement par carte de débit?
11. que le personnel soit courtois?
12. que le magasin ait en stock les produits annoncés?

Une analyse factorielle cherchera à identifier automatiquement des groupes de variables qui sont fortement corrélées entre elles.

Les statistiques descriptives ainsi que la matrice des corrélations sont obtenues en exécutant les lignes suivantes :

```
proc corr data=multi.factor;
var x1-x12;
run;
```

Statistiques simples							
Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum	Libellé
x1	200	2.26	1.13	451	1	5	x1
x2	200	2.51	1.24	502	1	5	x2
x3	200	3.01	1.19	601	1	5	x3
x4	200	2.91	1.33	582	1	5	x4
x5	200	3.55	1.17	710	1	5	x5
x6	200	2.14	1.14	428	1	5	x6
x7	200	1.82	1.06	364	1	5	x7
x8	200	2.92	1.32	583	1	5	x8
x9	200	3.04	1.12	608	1	5	x9
x10	200	2.59	1.32	518	1	5	x10
x11	200	2.99	1.33	597	1	5	x11
x12	200	3.45	1.16	690	1	5	x12

Coefficients de corrélation de Pearson, N = 200 Proba > r sous H0: Rho=0												
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
x1	1.00	-0.08	-0.14	-0.07	0.38	-0.01	-0.10	-0.13	-0.03	-0.11	-0.12	-0.01
x1		0.25	0.06	0.34	0.00	0.91	0.18	0.06	0.70	0.12	0.09	0.87
x2	-0.08	1.00	0.04	-0.02	-0.08	0.06	0.50	0.01	-0.01	0.43	-0.12	0.07
x2	0.25		0.55	0.80	0.26	0.43	0.00	0.84	0.92	0.00	0.09	0.35
x3	-0.14	0.04	1.00	0.10	-0.06	0.39	0.00	0.05	0.47	0.08	0.13	0.46
x3	0.06	0.55		0.15	0.43	0.00	0.99	0.53	0.00	0.29	0.07	0.00
x4	-0.07	-0.02	0.10	1.00	-0.05	0.06	0.08	0.57	0.01	0.09	0.50	0.09
x4	0.34	0.80	0.15		0.52	0.39	0.25	0.00	0.86	0.22	0.00	0.22
x5	0.38	-0.08	-0.06	-0.05	1.00	-0.04	-0.04	-0.02	0.03	-0.07	-0.06	-0.07
x5	0.00	0.26	0.43	0.52		0.58	0.60	0.83	0.64	0.34	0.43	0.34
x6	-0.01	0.06	0.39	0.06	-0.04	1.00	0.07	0.04	0.32	0.07	-0.04	0.32
x6	0.91	0.43	0.00	0.39	0.58		0.32	0.56	0.00	0.36	0.56	0.00
x7	-0.10	0.50	0.00	0.08	-0.04	0.07	1.00	0.09	-0.02	0.51	-0.03	0.02
x7	0.18	0.00	0.99	0.25	0.60	0.32		0.22	0.74	0.00	0.63	0.76
x8	-0.13	0.01	0.05	0.57	-0.02	0.04	0.09	1.00	-0.03	0.16	0.55	0.04
x8	0.06	0.84	0.53	0.00	0.83	0.56	0.22		0.62	0.02	0.00	0.53
x9	-0.03	-0.01	0.47	0.01	0.03	0.32	-0.02	-0.03	1.00	0.01	0.02	0.39
x9	0.70	0.92	0.00	0.86	0.64	0.00	0.74	0.62		0.91	0.77	0.00
x10	-0.11	0.43	0.08	0.09	-0.07	0.07	0.51	0.16	0.01	1.00	0.01	0.02
x10	0.12	0.00	0.29	0.22	0.34	0.36	0.00	0.02	0.91		0.91	0.75
x11	-0.12	-0.12	0.13	0.50	-0.06	-0.04	-0.03	0.55	0.02	0.01	1.00	0.05
x11	0.09	0.09	0.07	0.00	0.43	0.56	0.63	0.00	0.77	0.91		0.48
x12	-0.01	0.07	0.46	0.09	-0.07	0.32	0.02	0.04	0.39	0.02	0.05	1.00
x12	0.87	0.35	0.00	0.22	0.34	0.00	0.76	0.53	0.00	0.75	0.48	

2.4 Description du modèle d'analyse factorielle

On dispose d'observations sur p variables X_1, \dots, X_p . Le modèle d'analyse factorielle fait l'hypothèse que ces variables dépendent linéairement d'un plus petit nombre m de variables aléatoires, F_1, \dots, F_m , appelées facteurs communs et de p termes d'erreurs (ou facteurs spécifiques) $\varepsilon_1, \dots, \varepsilon_p$, de moyenne $E(\varepsilon_i) = 0$ et de variance $\text{Var}(\varepsilon_i) = \psi_i$ pour $i = 1, \dots, p$. Les facteurs F_1, \dots, F_m ont une moyenne nulle $E(F_i) = 0$, et une variance unitaire, $\text{Var}(F_i) = 1$ ($i = 1, \dots, p$) .

Spécifiquement, le modèle est

$$\begin{aligned} X_1 &= \mu_1 + \gamma_{11}F_1 + \gamma_{12}F_2 + \cdots + \gamma_{1m}F_m + \varepsilon_1 \\ X_2 &= \mu_2 + \gamma_{21}F_1 + \gamma_{22}F_2 + \cdots + \gamma_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p &= \mu_p + \gamma_{p1}F_1 + \gamma_{p2}F_2 + \cdots + \gamma_{pm}F_m + \varepsilon_p, \end{aligned}$$

où μ_i est l'espérance de la variable aléatoire X_i ($i = 1, \dots, p$) et où γ_{ij} est le chargement de la variable X_i sur le facteur F_j ($i = 1, \dots, p$; $j = 1, \dots, m$).

Les espérances (μ_i), les chargements (γ_{ij}) et les variances (ψ_i) sont des quantités fixes, mais inconnues, tandis que les facteurs communs (F_i) et les aléas (ε_i) sont des variables aléatoires non observables; on suppose que les aléas ne sont pas corrélées aux facteurs F et entre elles.

Le modèle d'analyse factorielle est un modèle pour la matrice de covariance/corrélation des variables explicatives X_1, \dots, X_p , avec $\text{Va}(X_j) = \sum_{l=1}^k \gamma_{jl}^2 + \psi_j$. On appelle **communalité** le terme $h_j = \sum_{l=1}^k \gamma_{jl}^2$, qui représente la proportion de variance totale de X_j due à la corrélation entre les facteurs. Le terme ψ_j est connu sous le vocable **unicité**.

De plus, si les variables ont été préalablement standardisées de telle sorte que $E(X_i) = 0$ et $\text{Var}(X_i) = 1$ (note : ceci revient à utiliser la matrice de corrélation des observations dans l'analyse ce qui est fait par défaut dans **SAS**), alors $\text{Cor}(X_i, F_j) = \gamma_{ij}$, c'est-à-dire, le chargement de la variable X_i sur le chargement F_j est le coefficient de corrélation entre cette variable et ce facteur.

Sans aucune contrainte sur le modèle, la matrice de covariance de X_1, \dots, X_p possède $p(p+1)/2$ paramètres, soit p variances et $p(p-1)/2$ termes de corrélation. Avec le modèle d'analyse factorielle, on suppose que l'on peut décrire cette structure en utilisant seulement $p(m+1) - m(m-1)/2$ paramètres (p variances spécifiques et pm chargements, moins les contraintes de diagonalisation). Par exemple, avec $p = 50$ variables et $m = 6$ facteurs, on essaie de décrire la structure de covariance à l'aide de 350 paramètres au lieu de 1275.

Il existe plusieurs méthodes pour extraire les facteurs, c'est-à-dire pour estimer les paramètres du modèle (les ψ_i et les γ_{ij}). Nous allons discuter de deux d'entre elles : la méthode du maximum de vraisemblance et la méthode des composantes principales. L'avantage de l'estimation par maximum de vraisemblance est qu'elle permet l'utilisation de critères d'information et de statistiques de tests pour guider le choix du nombre de facteurs, en supposant toutefois la normalité des facteurs et des aléas. L'estimation des paramètres requiert une optimisation numérique qui peut être délicate selon les cas de figure et qui mène parfois à des solutions paradoxales : un cas dit de quasi-Heywood survient quand $h_j = 1$ pour une variable j , (on parle de cas de Heywood si $h_j > 1$). Si on modélise des variables explicatives centrées réduites, $\text{Va}(X_j) = 1$, d'où un problème d'interprétation car le terme ψ_j serait nul (cas de quasi-Heywood) ou négatif (cas de Heywood) alors même que ce terme représente la variance du *je* aléa. Les cas de quasi-Heywood ont plusieurs causes,

lesquelles sont listées dans la documentation **SAS**. Souvent, c'est dû à l'utilisation d'un trop petit ou trop grand nombre de facteurs ou une taille d'échantillon trop petite, etc. Cela complique l'interprétation et nous amène à questionner la validité du modèle d'analyse factorielle comme simplification de la structure de covariance.

2.4.1 Rotation des facteurs

Dans le modèle d'analyse factorielle, on peut montrer que, lorsqu'il y a deux facteurs ou plus, il existe plusieurs configurations de facteurs qui donnent la même structure de covariance. En fait, les chargements peuvent seulement être déterminés à une transformation orthogonale près (note : une transformation orthogonale est une transformation qui préserve le produit scalaire; elle préserve ainsi toutes les distances et les angles entre deux vecteurs). Si les chargements provenant d'une méthode d'extraction des facteurs ne sont pas uniques, la matrice de corrélation estimée par le modèle est par contre unique.

Il existe plusieurs techniques de rotation de facteurs. Le but de ces techniques est d'essayer de trouver une solution qui fera en sorte que les facteurs seront facilement interprétables. La méthode la plus utilisée est la méthode **varimax** : elle produit une configuration de chargement en maximisant la variance de la somme des carrés des chargements pour les m facteurs. La méthode varimax tend à produire une configuration de facteurs tel que les chargements de chaque variable sont dispersés (des chargements élevés positifs ou négatifs et d'autres presque nuls).

Je vous suggère de toujours tenter d'interpréter la solution avec une rotation varimax. Si ce n'est pas suffisamment clair, il existe d'autres méthodes de rotation dont certaines (les rotations de type oblique) permettent la présence de corrélation entre les facteurs.

2.5 Estimation des facteurs

Les chargements estimés pour la solution à quatre facteurs, suite à la rotation varimax, sont obtenus avec le code SAS suivant :

```
proc factor data=multi.factor
  method=ml rotate=varimax nfact=4
  maxiter=500 flag=.3 hey;
  var x1-x12;
run;
```

Caractéristique du facteur de rotation											
		Factor1		Factor2		Factor3		Factor4			
x1	x1		-8		-2		-6		99	*	
x2	x2		-8		5		67	*	-5		
x3	x3		8		75	*	1		-11		
x4	x4		71	*	7		6		0		
x5	x5		-2		-3		-5		37	*	
x6	x6		1		51	*	8		1		
x7	x7		3		0		75	*	-5		
x8	x8		79	*	-1		10		-6		
x9	x9		-3		63	*	-4		-2		
x10	x10		10		5		66	*	-6		
x11	x11		71	*	5		-10		-7		
x12	x12		5		61	*	3		1		

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.

Les valeurs supérieures à 0.3 sont indiquées par un signe **.

En général, on associe une variable à un groupe (facteur) si son chargement est supérieur à 0, 3 (en valeur absolue), ce qui donne

- Facteur 1 : X_4 , X_8 et X_{11}
- Facteur 2 : X_3 , X_6 , X_9 et X_{12}
- Facteur 3 : X_2 , X_7 et X_{10}
- Facteur 4 : X_1 et X_5 .

Ces facteurs sont interprétables :

- Le facteur 1 représente l'importance accordée au service.
- Le facteur 2 représente l'importance accordée aux produits.
- Le facteur 3 représente l'importance accordée à la facilité de paiement.
- Le facteur 4 représente l'importance accordée aux prix.

Dans cet exemple, les choses se sont bien passées et le nombre de facteurs que nous avons spécifié (4) semble être adéquat, mais ce n'est pas toujours aussi évident. Il est utile d'avoir des outils pour guider le choix du nombre de facteurs.

2.6 Choix du nombre de facteurs

Il existe différentes méthodes pour se guider dans le nombre de facteurs, m , à utiliser. Cependant, le point important à retenir est que, peu importe le nombre choisi, il faut que les facteurs soient **interprétables**. Par conséquent, les méthodes qui suivent ne devraient servir que de guide et non pas être suivies aveuglément. La méthode du maximum de vraisemblance que nous avons utilisée dans l'exemple possède l'avantage de fournir trois critères pour choisir le nombre de facteurs appropriés. Ces critères sont :

- AIC (critère d'information d'Akaike)
- BIC (critère d'information bayésien de Schwarz)
- Le test du rapport de vraisemblance pour l'hypothèse nulle que le modèle de corrélation décrit le modèle factoriel avec m facteurs est adéquat, contre l'alternative qu'il n'est pas adéquat.

Les critères d'information servent à la sélection de modèles ; ils seront traités plus en détail dans les chapitres qui suivent. Pour l'instant, il est suffisant de savoir que le modèle avec la valeur du critère AIC (ou BIC) la plus petite est considéré le « meilleur » (selon ce critère).

Les sorties suivantes proviennent du même programme SAS et correspondent au modèle factoriel avec quatre facteurs estimé par maximum de vraisemblance.

Tests de significativité basés sur 200 observations			
Test	DDL	Pr >	Khi-2 khi-2
H0: Aucun facteur commun	66	503.4490	<.0001
HA: Au moins un facteur commun			
H0: 4 facteurs suffisants	24	12.5708	0.9727
HA: davantage de facteurs sont requis			
<hr/>			
Khi-2 sans correction de Bartlett		13.06317	
Critère d'information d'Akaike		-34.93683	
Critère bayésien de Schwarz		-114.09645	
Coefficient de fiabilité de Tucker et Lewis		1.07185	

Pour choisir le nombre de facteurs avec les critères d'information, il faut ajuster le modèle en faisant varier le nombre de facteurs (option `nfact`) et extraire la valeur numérique correspondante. Si `nfact` dépasse le nombre de valeurs propres estimées supérieures à 1 (critère `mineigen`), la

spécification `nfact` sera ignorée à moins que vous ne spécifiez un autre critère, par exemple `priors=one`. Cette option revient à démarrer l'algorithme d'optimisation avec la contrainte $\sum_{l=1}^k \gamma_{jl}^2 = 1, \psi_j = 0$. En revanche, méfiez-vous : cela mène à des maximums locaux, n'employez l'option que si nécessaire.

Il faut garder en tête que l'estimation par maximum de vraisemblance du modèle d'analyse factorielle est très sensible à l'initialisation : on peut aussi parfois obtenir des valeurs différentes selon les logiciels. Cette fragilité, couplée à la haute fréquence de cas de Heywood (la solution numérique donne des communalités qui excèdent un, or ces dernières représentent le carré d'un coefficient de corrélation linéaire, ce qui rend l'interprétation du modèle caduque), fait en sorte que je préfère utiliser la méthode des composantes principales pour l'estimation.

Le tableau 2.1 présente les valeurs estimées des critères d'information et des valeurs-*p* pour le test du rapport de vraisemblance pour cinq modèles. Le critère AIC suggère quatre facteurs, tandis que les deux autres critères, BIC et test du rapport de vraisemblance, suggèrent plutôt trois facteurs.

TABLE 2.1: Critères d'information et valeurs-*p* pour le modèle factoriel à *m* facteurs

<i>m</i>	AIC	BIC	valeur- <i>p</i>
1	228.0	49.9	<0.001
2	99.5	-42.3	<0.001
3	-20.5	-129.3	0.096
4	-34.9	-114.1	0.973
5	-24.8	-77.6	0.975

On peut considérer le modèle avec trois facteurs : les chargements (après rotation varimax) sont données dans la figure 2.2.

Cette solution récupère les trois facteurs *service*, *produits* et *paiement* de la solution précédente à quatre facteurs. Le facteur *prix* (qui était formé de X_1 et X_5) n'est plus présent : que faire avec ce dernier? Cela dépend du but de l'analyse et nous y reviendrons plus tard.

Pour terminer cette section, voici la description de deux autres critères *classiques* pour choisir le nombre de facteurs si l'on ajuste le modèle avec la méthodes des composantes principales (plutôt qu'avec le maximum de vraisemblance). Ces deux critères sont :

- Critère de Kaiser, un critère basé sur les valeurs propres. Avec une analyse en composantes principales basée sur la matrice des corrélations, la valeur propre associée à un facteur représente la partie de la variance totale qui est expliquée par ce facteur. Chaque variable compte pour un dans la variance totale. Le nombre de facteurs choisis est le nombre de valeurs propres supérieures à 1. L'idée est de garder seulement les facteurs qui expliquent plus de variance qu'une variable individuelle.

		Factor1	Factor2	Factor3
x1	x1	-15	-9	-14
x2	x2	-9	3	67 *
x3	x3	10	76 *	4
x4	x4	71 *	5	7
x5	x5	-5	-6	-10
x6	x6	1	50 *	9
x7	x7	2	-3	75 *
x8	x8	79 *	-3	12
x9	x9	-2	63 *	-2
x10	x10	9	3	67 *
x11	x11	72 *	5	-8
x12	x12	6	60 *	5

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.

Les valeurs supérieures à 0.3 sont indiquées par un signe **.

FIGURE 2.2 – Estimés des chargements pour trois facteurs avec rotation varimax

- le diagramme d'éboulis : un graphique des valeurs propres ordonnées de la plus grande à la plus petite en fonction de $1, \dots, p$. Habituellement, ce graphe prendra la forme d'une chute assez importante suivie d'une stabilisation des valeurs propres. Avec ce critère, le nombre de facteurs est déterminé par le nombre de valeurs propres avant le début du coude où il a stabilisation apparente. L'idée est de choisir l'endroit où l'ajout d'un facteur supplémentaire n'apporte qu'un gain marginal faible. Ce critère est par contre subjectif et dépend de l'analyste. En ajoutant `scree` comme option à `proc factor`, on obtient le diagramme d'éboulis mais il est facile de le créer manuellement et le résultat est esthétiquement plus réussi.

Les sorties qui suivent proviennent du programme :

```
proc factor data=multi.factor method=principal
  scree rotate=varimax flag=.3;
  ods output Eigenvalues=eigen;
  var x1-x12;
run;

proc sgplot data=eigen;
  scatter x=number y=eigenvalue;
  yaxis label="valeurs propres";
  xaxis label='nombre';
run;
```

Cette fois-ci, c'est la méthode des composantes principales qui est utilisée ; cette dernière consiste à estimer les chargements en utilisant les m premières valeurs propres et vecteurs propres de la matrice de corrélation. En ne spécifiant pas l'option `nfact`, **SAS** choisit le nombre de facteurs en utilisant par défaut le critère de Kaiser (valeurs propres supérieures à 1). Quatre facteurs sont retenus, tel qu'indiqué par la sortie au bas du tableau 2.3. Pour le diagramme d'éboulis de la figure 2.4, le choix est assez subjectif : il semble raisonnable de choisir trois ou quatre facteurs.

On suggère d'utiliser *de facto* les trois critères découlant de l'utilisation de la vraisemblance et de déterminer le nombre de facteurs à extraire selon différents critères avant d'examiner les modèles avec ce nombre de facteurs et ceux avec un facteur de moins ou de plus. Au final, le plus important est de pouvoir interpréter raisonnablement les facteurs et donc le modèle retenu est souvent choisi selon le critère **Wow!**. On veut dire par là que la configuration de facteurs choisie est compréhensible.

2.7 Construction d'échelles à partir des facteurs

Si le seul but de l'analyse factorielle est de comprendre la structure de corrélation entre les variables, alors se limiter à l'interprétation des facteurs est suffisant.

Valeurs propres de la matrice de corrélation:
Total = 12 Moyenne = 1

	Valeur propre	Définition	Proportion	Cumulé
1	2.42730801	0.42845477	0.2023	0.2023
2	1.99885324	0.05649946	0.1666	0.3688
3	1.94235378	0.64106103	0.1619	0.5307
4	1.30129275	0.56297464	0.1084	0.6392
5	0.73831811	0.04657350	0.0615	0.7007
6	0.69174461	0.12512145	0.0576	0.7583
7	0.56662316	0.02697168	0.0472	0.8055
8	0.53965148	0.03002176	0.0450	0.8505
9	0.50962972	0.03638282	0.0425	0.8930
10	0.47324690	0.01804874	0.0394	0.9324
11	0.45519816	0.09941808	0.0379	0.9704
12	0.35578007		0.0296	1.0000

4 facteurs seront retenus par le critère MINEIGEN.

FIGURE 2.3 – Valeurs propres et proportion de variance

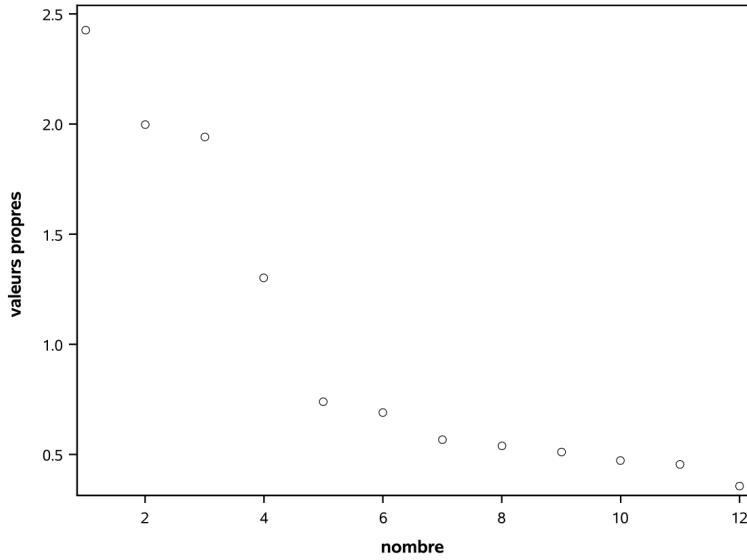


FIGURE 2.4 – Diagramme d'éboulis

Si par contre, le but est de réduire le nombre de variables pour pouvoir par la suite procéder à d'autres analyses statistiques, l'analyse factorielle peut alors servir de guide pour construire de nouvelles variables (échelles). En supposant que l'analyse factorielle a produit des facteurs qui sont interprétables et satisfaisants, la méthode de construction d'échelles la plus couramment utilisée consiste à construire m nouvelles variables, une par facteur. Pour un facteur donné, la nouvelle variable est simplement la moyenne des variables ayant des chargements élevés sur ce facteur (positifs ou négatifs, mais de même signe). Une autre méthode, les scores factoriels, sera présentée plus loin.

Lorsqu'on construit une échelle, il est important d'examiner sa cohérence interne. Ceci peut être fait à l'aide du coefficient alpha de Cronbach. Ce coefficient mesure à quel point chaque variable faisant partie d'une échelle est corrélée avec le total de toutes les variables pour cette échelle. Plus le coefficient est élevé, plus les variables ont tendance à être corrélées entre elles. L'alpha de Cronbach est

$$\alpha = \frac{k}{k-1} \frac{S^2 - \sum_{i=1}^k S_i^2}{S^2},$$

où k est le nombre de variables dans l'échelle, S^2 est la variance empirique de la somme des variables et S_i^2 est la variance empirique de la i e variable. En pratique, on voudra que ce coefficient soit au moins égal à 0,6 pour être satisfait de la cohérence interne de l'échelle.

Avec **SAS**, la procédure **corr** permet de calculer α .

```

/* pour le facteur service */
proc corr data=multi.factor alpha;
var x4 x8 x11;
run;
/* pour le facteur produits */
proc corr data=multi.factor alpha;
var x3 x6 x9 x12;
run;
/* pour le facteur paiement */
proc corr data=multi.factor alpha;
var x2 x7 x10;
run;
/* pour le facteur prix */
proc corr data=multi.factor alpha;
var x1 x5;
run;

```

Coefficient Alpha de Cronbach	
Variables	Alpha
Brut	0.780524
Normalisé	0.780611

Coefficient Alpha de Cronbach avec variable supprimée					
Variable supprimée	Variables brutes		Variables standardisées		
	Corrélation avec total	Alpha	Corrélation avec total	Alpha	Libellé
x4	0.609256	0.712551	0.609409	0.712565	x4
x8	0.649038	0.668928	0.649028	0.668929	x8
x11	0.595499	0.727412	0.595641	0.727434	x11

FIGURE 2.5 – Alpha de Cronbach pour le facteur *service*.

Il faut utiliser le alpha brut. Ainsi, les alphas de Cronbach sont tous satisfaisants (plus grand que 0, 6) sauf pour le facteur *prix* ($\alpha = 0,546$). SAS fournit également la matrice des corrélations des variables de l'échelle ainsi que la valeur du alpha de Cronbach si on retire une variable à la fois de l'échelle. Tout est donc cohérent. Les échelles provenant des facteurs *service*, *produits* et *paiement*,

Coefficient Alpha de Cronbach	
Variables	Alpha
Brut	0.718253
Normalisé	0.717602

Coefficient Alpha de Cronbach avec variable supprimée					
	Variables brutes	Variables standardisées			
Variable supprimée	Corrélation avec total	Alpha	Corrélation avec total	Alpha	Libellé
x3	0.584387	0.606777	0.584428	0.606820	x3
x6	0.430389	0.699956	0.429825	0.699753	x6
x9	0.509722	0.654325	0.508628	0.653545	x9
x12	0.501808	0.658864	0.500831	0.658223	x12

FIGURE 2.6 – Alpha de Cronbach pour le facteur *produits*.

Coefficient Alpha de Cronbach					
	Variables	Alpha			
Variable supprimée	Corrélation avec total	Alpha	Corrélation avec total	Alpha	Libellé

Coefficient Alpha de Cronbach avec variable supprimée					
	Variables brutes	Variables standardisées			
Variable supprimée	Corrélation avec total	Alpha	Corrélation avec total	Alpha	Libellé
x2	0.532476	0.661213	0.538157	0.672224	x2
x7	0.596466	0.601795	0.596509	0.602521	x7
x10	0.536537	0.663250	0.540698	0.669254	x10

FIGURE 2.7 – Alpha de Cronbach pour le facteur *paiement*.

Coefficient Alpha de Cronbach	
Variables	Alpha
Brut	0.545634
Normalisé	0.545805

FIGURE 2.8 – Alpha de Cronbach pour le facteur *prix*.

sont satisfaisantes. Ces facteurs sont identifiés à la fois dans la solution à quatre, mais aussi dans la solution à trois facteurs. Le facteur *prix* est celui qui apparaît en plus dans la solution à quatre facteurs. Il a une interprétation claire, mais son faible alpha ferait en sorte qu'il serait discutable de travailler avec l'échelle *prix* dans d'autres analyses (du moins avec selon l'usage habituel du alpha).

2.8 Compléments d'information

2.8.1 Variables ordinaires

Théoriquement, une analyse factorielle ne devrait être faite qu'avec des variables continues. Par contre, en pratique, on l'utilise souvent aussi avec des variables ordinaires (comme pour l'exemple portant sur le questionnaire) et même avec des variables binaires (0-1).

Dans ce genre de situation, on peut aussi utiliser d'autres mesures d'associations au lieu du coefficient de corrélation linéaire. Par exemple, on peut utiliser la corrélation polychorique, qui est une mesure de corrélation entre deux variables ordinaires. La corrélation tétrachorique correspond au cas spécial de deux variables binaires.

Ma suggestion est d'utiliser la corrélation linéaire ordinaire avec des variables ordinaires (même binaires). Si les résultats ne sont pas satisfaisants, on peut alors essayer avec d'autres mesures d'associations.

On peut refaire l'analyse des données portant sur le magasin dans **SAS** en utilisant la corrélation polychorique calculées par la procédure `corr` et en passant la sortie à la procédure `factor`.

```
proc corr data=multi.factor polychoric out=poly_corr;
var x1-x12;
run;

proc factor data=poly_corr
method=ml rotate=varimax nfact=4
```

```
maxiter=500 flag=.3 hey;
var x1-x12;
run;
```

Les chargements sont donnés dans la figure 2.9. Les facteurs obtenus sont les mêmes qu'en utilisant les corrélations linéaires.

		Caractéristique du facteur de rotation			
		Factor1	Factor2	Factor3	Factor4
x1	x1	-8	-2	-6	99 *
x2	x2	-8	5	67 *	-5
x3	x3	8	75 *	1	-11
x4	x4	71 *	7	6	0
x5	x5	-2	-3	-5	37 *
x6	x6	1	51 *	8	1
x7	x7	3	0	75 *	-5
x8	x8	79 *	-1	10	-6
x9	x9	-3	63 *	-4	-2
x10	x10	10	5	66 *	-6
x11	x11	71 *	5	-10	-7
x12	x12	5	61 *	3	1

Les valeurs imprimées sont multipliées par 100 et

arrondies au nombre entier le plus proche.

Les valeurs supérieures à 0.3 sont indiquées par un signe *.

FIGURE 2.9 – Chargements estimés pour la corrélation polychorique

2.8.2 Autres méthodes d'extractions de facteurs

Il n'y a pas de formule explicites pour l'estimation des paramètres avec la méthode du maximum de vraisemblance et un algorithme d'optimisation est nécessaire pour l'option des paramètres. Dans certains cas, l'algorithme peut terminer sans solution ou retourner un cas limite (où la variance est négative ou nulle). C'est le cas dans notre exemple avec quatre facteurs (solution de

Heywood), bien que ce ne soit pas indiqué. La sortie **SAS** contient des informations sur la convergence de l'estimé : idéalement, on obtient la mention Critère de convergence respecté; autrement, essayez de varier le nombre. Un autre signe que l'algorithme n'a pas convergé est la présence de degrés de libertés négatifs pour le test du rapport de vraisemblance.

La méthode par les composantes principales (mentionnée lors de la présentation des valeurs propres et du diagramme d'éboulis) a une solution explicite et peut donc dépanner si on n'arrive pas à obtenir le maximum de vraisemblance.

D'autres méthodes sont aussi disponibles dans **SAS** (voir la rubrique d'aide du logiciel) mais les deux méthodes mentionnées devraient être suffisantes pour la grande majorité des applications.

2.8.3 Autres méthodes de rotation des facteurs

Jusqu'à présent, nous avons utilisé la méthode de rotation orthogonale varimax. Il existe de nombreuses autres méthodes de rotations orthogonales telles, orthomax, quartimax, parsimax et equimax (voir la rubrique d'aide de **SAS**). Rappelez-vous que le modèle d'analyse factorielle de base suppose que les facteurs sont non corrélés. Les rotations de type obliques quant à elles permettent d'introduire de la corrélation entre les facteurs. Quelquefois, une telle rotation facilitera davantage l'interprétation des facteurs qu'une rotation orthogonale. **SAS** permet l'utilisation de plusieurs méthodes de rotation obliques qui sont documentées dans la rubrique d'aide. Notez qu'il faut être prudent lorsqu'on utilise une méthode de rotation oblique car il y aura trois matrices de chargements après rotation (coefficients de régression normalisés, corrélations semi-partielles ou corrélations). On suggère l'utilisation de la première, soit la représentation avec **coefficients de régression normalisés**. Il s'agit des coefficients de régression si on voulait prédire les variables à l'aide des facteurs. Ils indiquent donc à quel point chaque facteur est associé à chaque variable. Dans le cas d'une rotation orthogonale, ces trois matrices sont les mêmes et il s'agit de trois interprétations valides des chargements.

Le programme suivant fait une analyse factorielle avec quatre facteurs, mais en utilisant une rotation varimax oblique (option `rotate=obvarimax`).

```
proc factor data=multi.factor
maxiter=500 flag=.3 hey;
var x1-x12;
run;
```

La matrice des corrélations entre facteurs est donnée dans la figure 2.10 et les chargements sont présentés dans la figure 2.11. On voit ici qu'on obtient les mêmes quatre facteurs qu'avec une rotation varimax orthogonale.

Corrélations inter-facteurs				
	Factor1	Factor2	Factor3	Factor4
Factor1	100 *	7	4	-11
Factor2	7	100 *	6	-7
Factor3	4	6	100 *	-13
Factor4	-11	-7	-13	100 *

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.
Les valeurs supérieures à 0.3 sont indiquées par un signe '*'.

FIGURE 2.10 – Corrélation interfacteurs pour rotation varimax oblique

2.8.4 Scores factoriels

Avec les données de l'exemple, en nous basant sur les résultats de l'analyse factorielle, nous avons créé quatre nouvelles échelles (une par facteur) que l'on peut calculer pour chaque individu :

- $service = (X_4 + X_8 + X_{11})/3$,
- $produit = (X_3 + X_6 + X_9 + X_{12})/4$,
- $paiement = (X_2 + X_7 + X_{10})/3$,
- $prix = (X_1 + X_5)/2$.

Par exemple, la variable $prix$ peut donc être vu comme une combinaison linéaire des 12 variables où seulement X_1 et X_5 reçoivent un poids (égal) différent de zéro. Une autre façon de créer de nouvelles variables consiste à calculer des scores factoriels (un pour chaque facteur) pour chaque individu. Par exemple, pour un individu i donné, le score factoriel pour le facteur k peut être prédit à l'aide de la formule

$$\begin{aligned}\hat{F}_{ik} &= \hat{\Gamma}^\top \mathbf{R}^{-1} \mathbf{z} \\ &= \hat{\beta}_{1,k} z_{i,1} + \cdots + \hat{\beta}_{12,k} z_{i,12},\end{aligned}$$

où $z_{i,1}, \dots, z_{i,12}$ sont les valeurs centrées et réduites des observations correspondant à l'individu et où $\hat{\beta}_{1,k}, \dots, \hat{\beta}_{12,k}$ sont des coefficients estimés à partir des chargements γ_{ij} (après rotation) et de la matrice de corrélation des variables \mathbf{R} , avec $\hat{\beta}_{i,k} = \sum_{j=1}^p \hat{\gamma}_{kj} r_{jk}$.

Ainsi, chacune des 12 variables originales contribue au calcul du score factoriel. Les variables ayant des chargements plus élevés sur ce facteur auront tendance à avoir des poids ($\hat{\gamma}$) plus élevés. Par contre, les scores factoriels ne sont pas uniques car ils dépendent des chargements utilisés (et

Représentation du facteur avec rotation (Coefficients de régression normalisés)						
		Factor1	Factor2	Factor3	Factor4	
x1	x1	-2	3	3	100	*
x2	x2	-9	2	67 *	-2	
x3	x3	6	75 *	-2	-10	
x4	x4	72 *	4	4	3	
x5	x5	0	-1	-2	37 *	
x6	x6	0	51 *	7	2	
x7	x7	2	-3	75 *	-1	
x8	x8	79 *	-5	8	-3	
x9	x9	-4	63 *	-5	-1	
x10	x10	9	2	66 *	-3	
x11	x11	71 *	2	-11	-4	
x12	x12	4	61 *	2	2	

**Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.
Les valeurs supérieures à 0.3 sont indiquées par un signe **.**

FIGURE 2.11 – Chargements avec rotation oblique varimax

donc à la fois de la méthode d'estimation et de la méthode de rotation). On peut également utiliser les scores factoriels au lieu des 12 variables originales dans des analyses subséquentes. Il est suggéré d'utiliser les nouvelles variables (échelles) obtenues en faisant les moyennes des variables identifiées comme faisant partie de chaque facteur pour les raisons suivantes :

- l'interprétation des scores factoriels est moins claire (chaque facteur dépend de toutes les variables)
- les scores factoriels ne sont pas uniques (ils dépendent de la méthode d'estimation et de rotation).
- les coefficients servant au calcul seront différents d'une étude à l'autre.

Coefficients du score normalisés						
		Factor1	Factor2	Factor3	Factor4	
x1	x1	0.03059	0.04866	0.04199	1.01161	
x2	x2	-0.04524	0.01276	0.30701	0.01431	
x3	x3	0.00898	0.45246	-0.01635	0.00975	
x4	x4	0.30245	0.01130	0.01197	0.02657	
x5	x5	0.00310	-0.00605	-0.00925	-0.00041	
x6	x6	-0.00837	0.17490	0.02305	0.00447	
x7	x7	-0.00542	-0.01815	0.44283	0.02490	
x8	x8	0.45256	-0.04535	0.04422	0.03993	
x9	x9	-0.02516	0.26576	-0.02469	0.00227	
x10	x10	0.02061	0.00819	0.29821	0.01927	
x11	x11	0.30260	0.00462	-0.06951	0.02170	
x12	x12	0.00289	0.24472	0.00325	0.00581	

FIGURE 2.12 – Coefficients du score normalisés

Pour obtenir les scores avec **SAS**, il suffit d'insérer l'option **score** à la procédure **factor**. L'option **out=...** permet de créer un fichier de données **SAS** qui contient la valeur des m scores pour chaque individu. Les scores factoriels pour l'exemples sont rapportés dans la figure 2.12. On remarque que :

- pour le premier facteur, trois variables ont des poids importants (X_4 , X_8 et X_{11}). Il s'agit donc d'un facteur très proche du facteur *service*.

- pour le deuxième facteur, les variables X_3 , X_6 , X_9 et X_{12} ont des poids importants. Il s'agit donc d'un facteur très proche du facteur *produits*.
- pour le troisième facteur, les variables X_2 , X_7 , X_{10} ont des poids importants. Il s'agit donc d'un facteur très proche du facteur *paiement*.
- pour le quatrième facteur, seule la variable X_1 a un poids important. On aurait pu s'attendre à ce que ce soit également le cas pour X_5 , en lien avec le facteur *prix* — ce facteur était moins clair selon le alpha de Cronbach.

Les corrélations entre les échelles (construites avec les moyennes) et les scores factoriels sont données dans la figure 2.13. On remarque la forte corrélation entre le score factoriel et les échelles construites avec les moyennes pour les facteurs *service*, *produits* et *paiement*. Cela veut dire qu'utiliser les échelles ou les scores factoriels ne devrait pas faire de différence dans des analyses subséquentes. Par contre, cette corrélation est plus faible (0.82) pour le facteur *prix*.

Coefficients de corrélation de Pearson, N = 200				
	Factor1	Factor2	Factor3	Factor4
service	0.99397	0.04659	0.02748	-0.05223
produit	0.04598	0.98350	0.03233	-0.03972
paiement	0.02640	0.04496	0.98615	-0.06790
prix	-0.07126	-0.03562	-0.07746	0.81920

FIGURE 2.13 – Corrélation entre scores et échelles

Chapitre 3

Analyse de regroupements

3.1 Introduction

On cherche à créer des groupes (*clusters*) d'individus homogènes en utilisant p variables X_1, \dots, X_p . Plus précisément, on veut combiner des sujets en groupes (interprétables) de telle sorte que les individus d'un même groupe soient le plus semblables possible par rapport à certaines caractéristiques et que les groupes soient le plus différent possible.

Nous disposons des observations pour n individus et X_{ij} dénote la valeur de la j e variable explicative pour le i e sujet : les variables correspondant au sujet S_i sont donc X_{i1}, \dots, X_{ip} .

Il y a une certaine analogie avec l'analyse factorielle. En analyse factorielle, on cherche à déterminer s'il y a des groupes de **variables** corrélées entre elles. On cherche donc à regrouper des variables. En analyse de regroupements, on cherche plutôt à créer des groupes de **sujets** similaires. Les deux méthodes servent pour l'analyse exploratoire : en particulier, on ne peut pas regrouper les données pour ensuite comparer la moyenne des segments obtenus à l'aide de statistiques sans ajustement préalable.

Étapes d'une analyse de regroupements

- 1) Choisir les variables pertinentes à l'analyse.
- 2) Décider comment seront mesurées les « distances » entre les sujets. Cela revient à choisir une mesure de dissemblance.
- 3) Décider quelle méthode sera utilisée (méthode hiérarchique, méthode non hiérarchique).
- 4) Choisir le nombre de groupes, soit à partir de connaissances à priori, soit en se basant sur l'analyse de regroupements elle-même.
- 5) Interpréter les groupes obtenus.
- 6) Utiliser ces groupes dans d'autres analyses, le cas échéant.

3.1.1 Mesures de dissemblance

Une mesure de dissemblance sert à quantifier la distance entre deux sujets S_i et S_j en se basant sur les p variables X_1, \dots, X_p . Plus cette mesure est petite, plus les sujets S_i et S_j sont similaires. Même s'il y a des exceptions, la plupart des mesures de dissemblances d ont les propriétés suivantes :

- 1) $d(S_i, S_j) \geq 0$ (positivité);
- 2) $d(S_i, S_i) = 0$;
- 3) $d(S_i, S_j) = d(S_j, S_i)$ (symmétrie);
- 4) $d(S_i, S_j)$ augmente au fur et à mesure que les deux sujets deviennent plus différents.

Lorsque toutes les variables sont continues, une mesure de dissemblance souvent utilisée est la distance euclidienne entre sujets, soit

$$d(S_i, S_j) = \sqrt{(X_{i1} - X_{j1})^2 + \dots + (X_{ip} - X_{jp})^2}.$$

La distance euclidienne est tout simplement la longueur du segment qui relie les deux points dans l'espace.

Nous verrons plus tard d'autres mesures de dissemblances incluant des mesures qui peuvent être utilisées avec des variables binaires, nominales et ordinaires.

3.1.2 Méthodes hiérarchiques et non hiérarchiques

Les **méthodes hiérarchiques** assignent les individus aux groupes à l'aide d'un algorithme glouton en partant du cas à n groupes où chaque sujet est un groupe à part entière. La distance entre chaque paire de groupe est calculée. Les deux groupes ayant la distance la plus petite sont regroupés pour ne laisser que $n - 1$ groupes. La distance entre chaque paire de groupe est à nouveau calculée (pour les groupes). Les deux groupes ayant la distance la plus petite sont regroupés pour ne former qu'un seul groupe et ainsi de suite. Le processus se continue ainsi jusqu'à ce que tous les sujets soient regroupés en un seul groupe.

Avec une méthode hiérarchique, on n'a pas besoin de spécifier le nombre de groupes à priori. Cependant, une fois qu'un sujet est assigné à un groupe, il ne peut le quitter pour être réassigné à un autre groupe plus tard. Ce qui différencie les différentes méthodes hiérarchiques est la manière dont est calculée la distance entre deux groupes.

Pour les **méthodes non hiérarchiques**, le nombre de groupe est spécifié au départ et un algorithme cherche, à partir d'une solution initiale, la meilleure distribution des sujets à travers ce nombre de groupe d'une manière itérative. Avec ces méthodes, l'assignation d'un sujet peut être modifiée d'une itération à l'autre. Il faut cependant spécifier le nombre de groupe et les « centres » de ces groupes au départ. La solution peut être très sensible au choix des centres initiaux.

3.2 Segmentation de seniors en voyage organisé

Les données sont inspirées de

Hsu, C. H. C. et Lee E.-J. (2002). Segmentation of Senior Motorcoach Travelers. *Journal of Travel Research*, **40**, 364-373.

Les buts de l'étude étaient

- 1) Regrouper les gens de 55 et plus qui participent à des voyages organisés en autobus en groupes homogènes selon des caractéristiques reliées au choix de l'opérateur et du voyage.
- 2) Examiner les caractéristiques de ces groupes.
- 3) Examiner les caractéristiques démographiques de ces groupes.

Nous allons nous intéresser principalement aux deux premiers points ici. Un questionnaire a été élaboré afin d'évaluer l'importance de 55 caractéristiques des opérateurs de voyages organisés en autobus et des voyages eux-mêmes à l'aide d'une échelle de Likert à cinq points, allant de extrêmement important (5) à pas important du tout (1). Des données sont disponibles pour 150 sujets (il y en avait 817 dans l'article). Elles se trouvent dans le fichier `cluster.sas7bdat`.

Au lieu de faire une analyse de regroupements avec les 55 items du questionnaire, les auteurs ont choisi de faire une analyse factorielle avec rotation varimax au préalable afin de réduire le nombre de variables à six facteurs interprétables :

- Activités sociales (X_1) : formé de cinq items
- Politiques de l'opérateur et références (X_2) : formé de six items.
- Horaires flexibles (X_3) : formé de trois items.
- Santé et sécurité (X_4) : formé de quatre items.
- Matériel publicitaire (X_5) : formé de deux items.
- Réputation (X_6) : formé de deux items.

On voit donc que 22 items, parmi les 55, sont utilisés dans la définition de ces six facteurs. Dans l'article, les auteurs ont décidé d'inclure ces 22 items dans l'analyse de regroupements. Pour notre part et afin de simplifier l'exemple, nous allons plutôt créer six nouvelles échelles en faisant la moyenne des items de chaque facteur ci-haut et utiliser seulement ces six échelles dans l'analyse de regroupements. Les valeurs de ces six variables pour les 150 sujets se trouvent dans le fichier `cluster.sas7bdat` et sont toutes dans l'intervalle [1,5], puisqu'elles représentent la moyenne d'échelles de Likert.

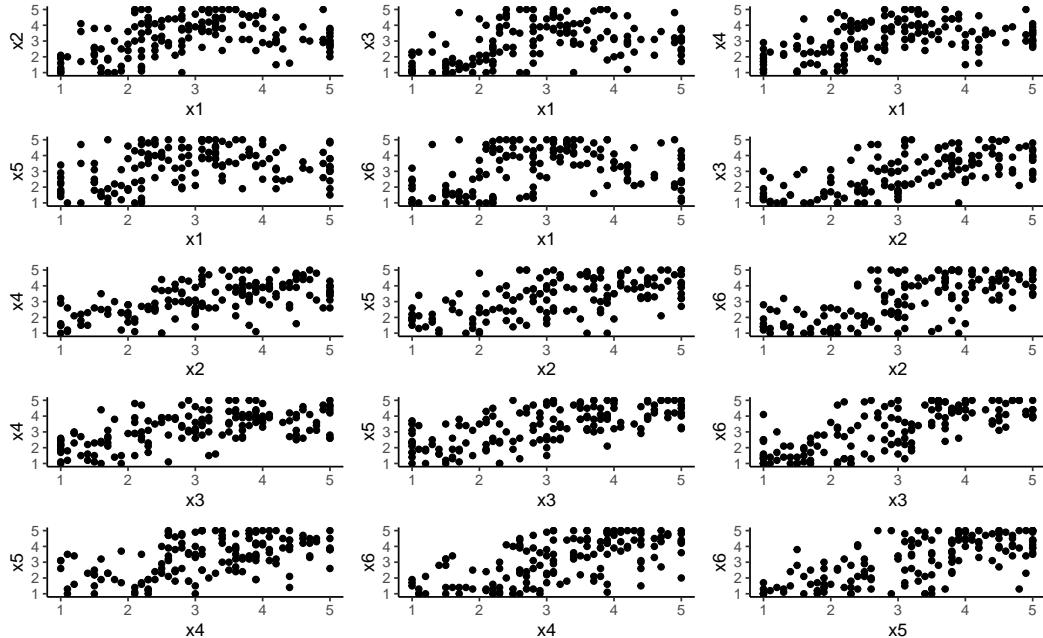
Variable	Libellé	Moyenne	Ec-type	Minimum	Maximum
x1	x1	2.93	1.14	1.00	5.00
x2	x2	3.22	1.16	1.00	5.00
x3	x3	2.96	1.22	1.00	5.00
x4	x4	3.20	1.05	1.00	5.00
x5	x5	3.32	1.18	1.00	5.00
x6	x6	3.17	1.36	1.00	5.00

3.3 Exploration graphique préalable et analyse en composantes principales

Comme c'est le cas avec n'importe quelle analyse statistique, il est nécessaire de tenter d'explorer les données graphiquement. On peut parfois réussir à visualiser les groupes d'observations, ce qui nous permettra une fois l'analyse de regroupement complétée de vérifier la qualité de cette dernière.

Une première idée consiste à faire le graphe de toutes les paires de variables mais ceci possède deux limites,

- i) il y aura beaucoup de graphes si le nombre de variables est grand et ii) on examine seulement les relations bivariées.



Il n'est pas nécessairement évident de détecter des groupes d'observations ainsi, alors qu'on n'a déjà que six variables.

3.3.1 Analyse en composantes principales

L'analyse en composantes principales peut être vue comme une méthode de réduction de la dimensionnalité. En fait, elle peut servir à faire de l'analyse factorielle et nous en avons déjà parlé dans le chapitre correspondant. En partant de p variables X_1, \dots, X_p , on forme de nouvelles variables qui sont des combinaisons linéaires des variables originales,

$$C_j = w_{j1}X_1 + w_{j2}X_2 + \dots + w_{jp}X_p, \quad (j = 1, \dots, p),$$

de telle sorte que

- La première variable formée, C_1 , appelée première composante principale, possède la variance maximale parmi toutes les combinaisons linéaires sous la contrainte $w_{11}^2 + \dots + w_{1p}^2 = 1$.
- La deuxième composante principale C_2 possède la variance maximale parmi toutes les combinaisons linéaires qui sont non corrélées avec C_1 sous la contrainte $w_{21}^2 + \dots + w_{2p}^2 = 1$.
- La troisième composante principale C_3 possède la variance maximale parmi toutes les combinaisons linéaires qui sont non corrélées avec C_1 et C_2 sous la contrainte $w_{31}^2 + \dots + w_{3p}^2 = 1$.

et ainsi de suite. Les contraintes sont nécessaires afin de standardiser le problème car il serait possible d'avoir des variances infinies sinon. Ainsi, les composantes principales forment un ensemble de variables non corrélées entre elles, qui récupèrent en ordre décroissant le plus possible de la variance des variables originales. La somme des variances des p composantes principales est égale à la somme des variances des p variables originales.

Mathématiquement, les composantes principales correspondent aux vecteurs propres de la matrice de covariance. On peut également utiliser la matrice de corrélation, et cette dernière est sélectionnée par défaut dans la méthode `princomp` à moins de spécifier `cov` dans l'appel. L'avantage de la matrice de corrélation (ou de la standardisation des variables) est que l'unité de mesure n'importe pas le résultat; autrement, un point plus important est attribué aux variables qui ont la plus forte hétérogénéité.

Si on conserve toutes les composantes principales, cela revient à changer le système de coordonnées dans lequel sont exprimées nos observations en effectuant une rotation : on trouve la direction dans le système 2D dans lequel l'étendue est la plus grande.

La figure 3.1 démontre cette décomposition sur des données simulées. La variance des données dans le premier panneau est 13.51 pour l'axe des abscisses et 6.43 pour l'axe des ordonnées avec une corrélation de 0.86, à comparer avec des variances de 18.65 et 1.21 et une corrélation nulle entre les deux composantes principales.

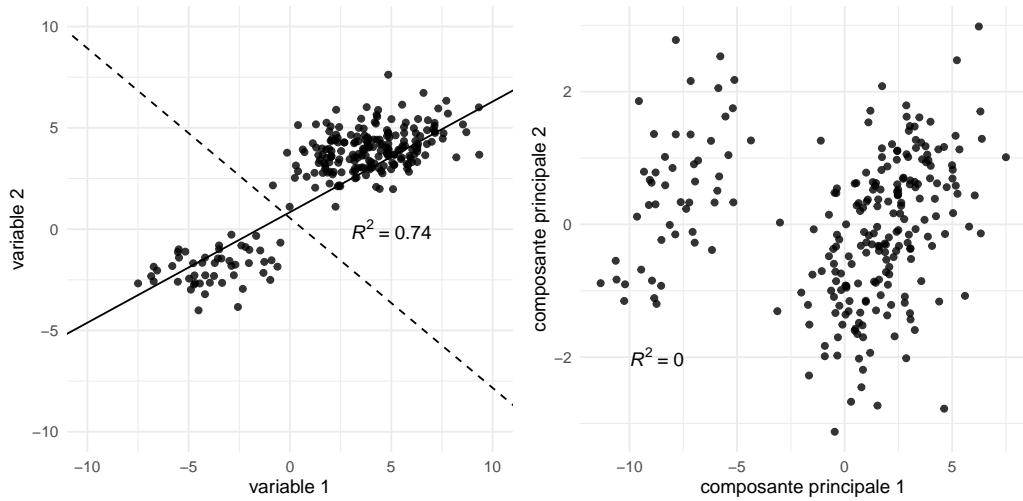
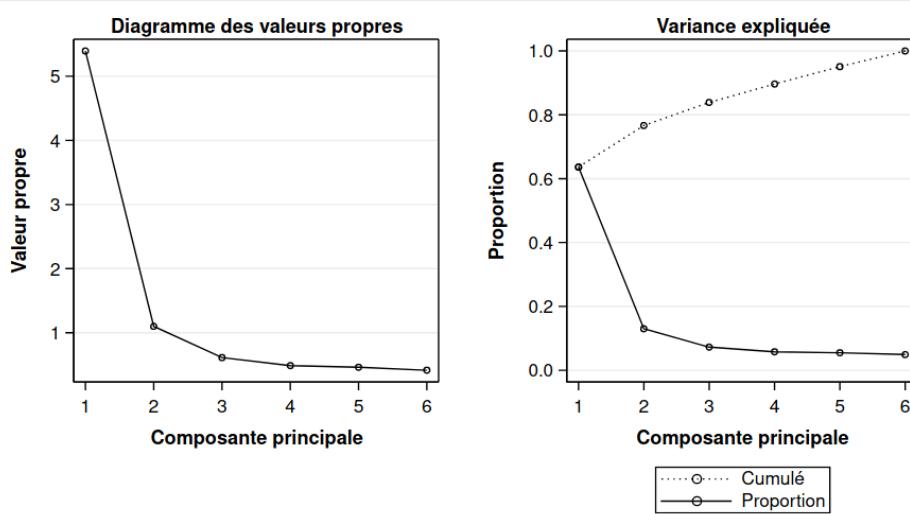


FIGURE 3.1 – Nuage de points avant (gauche) et après (droite) analyse en composantes principales. Les directions des composantes principales (lignes pleines et traitillées), qui forment un angle droit, sont ajoutées au nuage de points à gauche. On peut constater que la corrélation entre les deux composantes principales est nulle.

Si une simple rotation peut sembler inutile, la méthode ne dévoile son utilité qu'en haute dimension. On espère en général qu'un petit nombre de composantes principales réussira à expliquer la plus grande partie de la variance totale. Ces composantes pourraient servir de variables pour l'analyse de regroupement : une fois les étiquettes obtenues, on pourrait alors calculer les statistiques descriptives sur les variables originales. Dans notre exemple, on va seulement s'en servir comme outil graphique pour une analyse de regroupements en réduisant la dimension afin de permettre la visualisation des regroupements obtenus.



Cette étape est généralement effectuée avant l'analyse de regroupement. Le fichier `cluster7_acp.sas` contient les commandes pour faire une analyse en composantes principales et sauvegarder les composantes principales afin d'en faire des graphiques. La sortie inclut une mesure de la variance cumulée des K premières valeurs propres. La colonne Proportion donne la proportion de la variance totale qui est expliquée par la composante correspondante. La colonne Cumulé donne le cumul de variance totale expliquée par les composantes jusqu'à là. Ainsi, les deux premières composantes principales reproduisent 76,7% de la variance totale originale.

Même si on ne connaît pas l'appartenance des observations au regroupement, on distingue assez clairement trois groupes. Le panneau droit du graphique 3.2 montre les deux composantes principales, mais avec l'identification des groupes obtenus suite à l'analyse de regroupement avec la méthode des K -moyennes couverte plus tard.

3.4 Méthodes hiérarchiques

Cette méthode débute avec n groupes, un par sujet, et procède en regroupant des groupes formés au préalable d'une manière hiérarchique jusqu'à ce que tous les sujets ne forment qu'un seul

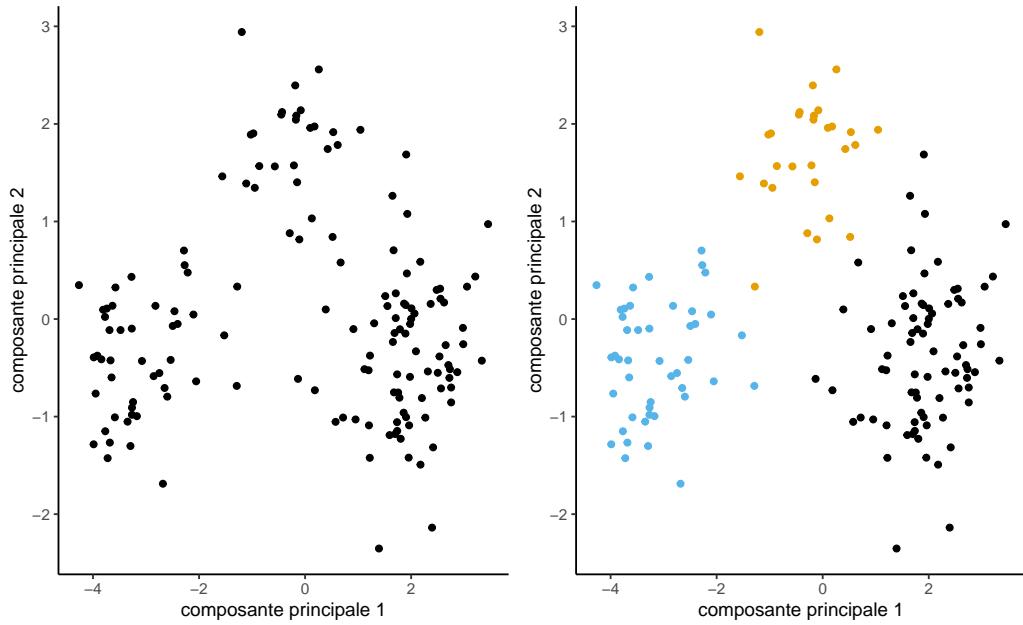


FIGURE 3.2 – Projection des observations sur les composantes principales avec les regroupements finaux créés à la fin du chapitre.

groupe. Le nombre de groupe retenu pourra être sélectionné à l'aide de certains critères que nous verrons plus tard.

À une étape donnée, il faut choisir quels groupes seront combinés. Les deux groupes dont la distance est la plus faible seront combinés. Il faut donc être en mesure de calculer la distance entre deux groupes. Nous allons décrire la méthode de Ward, qui compte parmi les plus populaires. Nous reviendrons brièvement sur d'autres méthodes plus loin.

3.4.1 Méthode de Ward

Cette méthode est basée sur un critère d'homogénéité global des groupes. Pour un groupe donné, cette homogénéité est mesurée par la somme des carrés des observations par rapport à la moyenne du groupe. L'homogénéité globale est alors la somme des homogénéités de tous les groupes. Plus l'homogénéité globale est petite, plus les groupes sont homogènes. À une étape donnée, les deux groupes qui causent la plus petite hausse de l'homogénéité globale (la plus petite perte d'information) sont regroupés. La méthode de Ward donne des groupes compacts d'apparence sphérique.

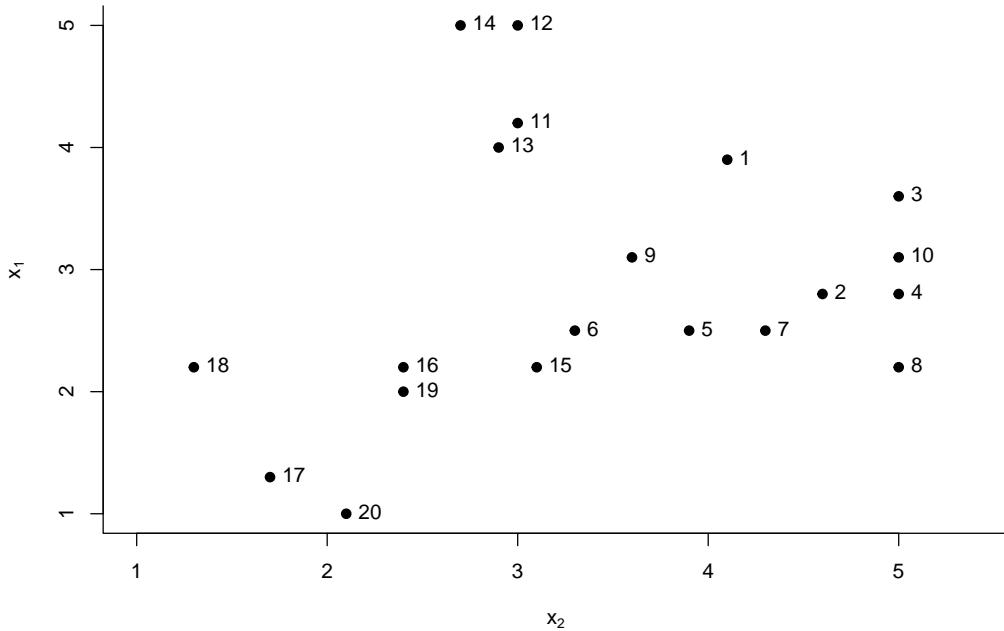
Plus précisément, supposons qu'à une étape du processus hiérarchique, nous avons M groupes et que nous voulons passer à $M - 1$ groupes. Pour un groupe K (parmi $1, 2, \dots, M$), définissons la somme des carrés des distances par rapport à la moyenne du groupe, SCD_k . Plus SCD_k est petite, plus le groupe est compact et homogène.

On peut calculer cette distance pour tous les M groupes et définir l'**homogénéité globale** comme la somme de l'homogénéité de tous les groupes,

$$SCD_G = SCD_1 + \dots + SCD_M.$$

Plus l'homogénéité globale SCD_G est petite, mieux c'est. Pour passer de M à $M - 1$ groupes, la méthode de Ward va regrouper les deux groupes qui feront que SCD_G sera la plus petite possible.

On procède à une analyse simplifiée des données pour le voyage organisé avec deux variables et vingt observations afin d'être en mesure de visualiser l'algorithme de regroupement.



La première analyse utilise la méthode de Ward. Les commandes **SAS** se trouvent dans **cluster1_simplifie.sas**; la présentation de la procédure et de la syntaxe est différée. L'historique de regroupement est décrit dans la sortie **SAS**. La première colonne donne le nombre de groupes. Au départ, les observations 16 et 19 sont regroupées, il y a maintenant 19 groupes. Ensuite, les observations 11 et 13 sont regroupées, il y a maintenant 18 groupes. Au moment de passer de 14 à 13 groupes, c'est le groupe formé à l'étape 16 qui est fusionné avec l'observation 2 et ainsi de suite. La colonne Fréq donne le nombre d'observations dans le groupe qui vient d'être formé.

Historique des classifications						
Nombre de clusters	R carré semi-partiel			Somme des carrés entre clusters	Lien	
	Clusters jointes	Fréq	R carré			
19	OB16	OB19	2	0.0004	1.00	0.02
18	OB11	OB13	2	0.0005	.999	0.025
17	OB12	OB14	2	0.0009	.998	0.045 T
16	OB4	OB10	2	0.0009	.997	0.045
15	OB6	OB15	2	0.0014	.996	0.065
14	OB5	OB7	2	0.0017	.994	0.08
13	OB2	CL16	3	0.0025	.992	0.1217
12	OB17	OB20	2	0.0026	.989	0.125
11	CL13	OB8	4	0.0079	.981	0.3808 T
10	CL14	OB9	3	0.0084	.973	0.4067
9	OB1	OB3	2	0.0093	.963	0.45
8	CL15	CL19	4	0.0146	.949	0.7025
7	CL18	CL17	4	0.0170	.932	0.82
6	CL12	OB18	3	0.0203	.911	0.975
5	CL9	CL11	6	0.0325	.879	1.5642
4	CL5	CL10	9	0.0356	.843	1.7139
3	CL8	CL6	7	0.0618	.782	2.9754
2	CL4	CL7	13	0.2955	.486	14.228
1	CL2	CL3	20	0.4860	.000	23.399

Les quantités `sprsq` et `rsq` sont des statistiques qui peuvent servir de guide pour choisir le nombre de groupes. Le RSQ est une mesure similaire au R^2 régression linéaire qui mesure globalement à quel point les groupes sont homogènes. Elle prend une valeur entre 0 et 1 où 0 et plus le RSQ est élevé, meilleur le regroupement. On définit le RSQ comme la proportion de la variabilité expliquée par les groupes. C'est une version standardisée de la somme des homogénéités, SCD_G ,

$$RSQ = 1 - \frac{SCD_G}{SCD_T},$$

où SCD_T est la somme des carrés des distances par rapport à la moyenne lorsque toutes les observations sont dans un même groupe. Le graphique 3.3 montre l'évolution du RSQ en fonction du nombre de groupes.

L'idée est généralement de choisir un petit nombre de groupe avec un RSQ assez élevé. Ici, on voit que le RSQ chute brutalement en passant de trois à deux groupes (il passe de 78,2% de variabilité expliquée à 48,6%). Ainsi, choisir trois groupes semble raisonnable.

L'autre mesure, le SPRSQ ou *R carré semi-partiel*, mesure la perte d'homogénéité résultant du fait que l'on vient de former un nouveau groupe. Comme on veut des groupes homogènes, on veut

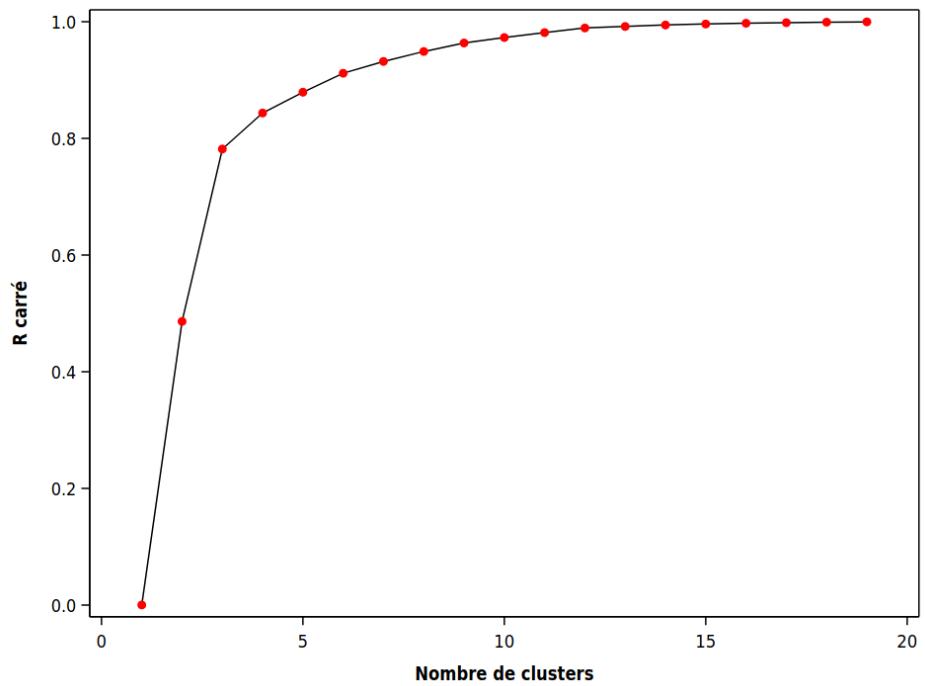


FIGURE 3.3 – R carré en fonction du nombre de groupes

qu'elle soit petite. Plus précisément, supposons que les groupes k_1 et k_2 viennent d'être regroupés à une étape donnée. Soient SCD_{k_1} et SCD_{k_2} les homogénéités de ces deux groupes et SCD_k l'homogénéité du nouveau groupe formé en fusionnant les deux. On définit la perte d'homogénéité (relative) en combinant ces deux groupes

$$SPRSQ = \frac{SCD_k - SCD_{k_1} - SCD_{k_2}}{SCD_T}$$

On peut ainsi tracer une courbe pour le SPRSQ en fonction du nombre de groupes, comme dans le graphique 3.4.

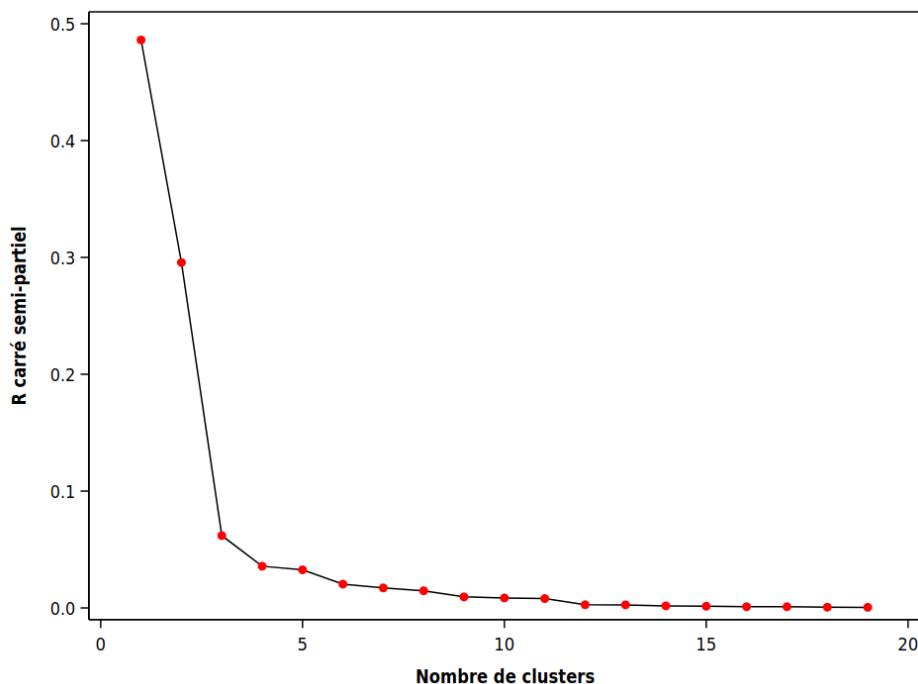


FIGURE 3.4 – Courbe du R^2 semi-partiel en fonction des groupements hiérarchiques

La procédure **SAS** qui permet d'effectuer une analyse de regroupements hiérarchique est `cluster`. Le fichier `cluster2_complet.sas` explique les différentes options disponibles.

```
proc cluster data=temp method=ward outtree=temp1 nonorm rsquare;
var x1-x6;
copy id cluster_vrai x1-x6;
ods output stat.cluster.ClusterHistory=criteres;
run;
```

On peut représenter graphique le R carré (Figure 3.5), le R carré semi partiel (Figure 3.6) en fonction du nombre de groupes.

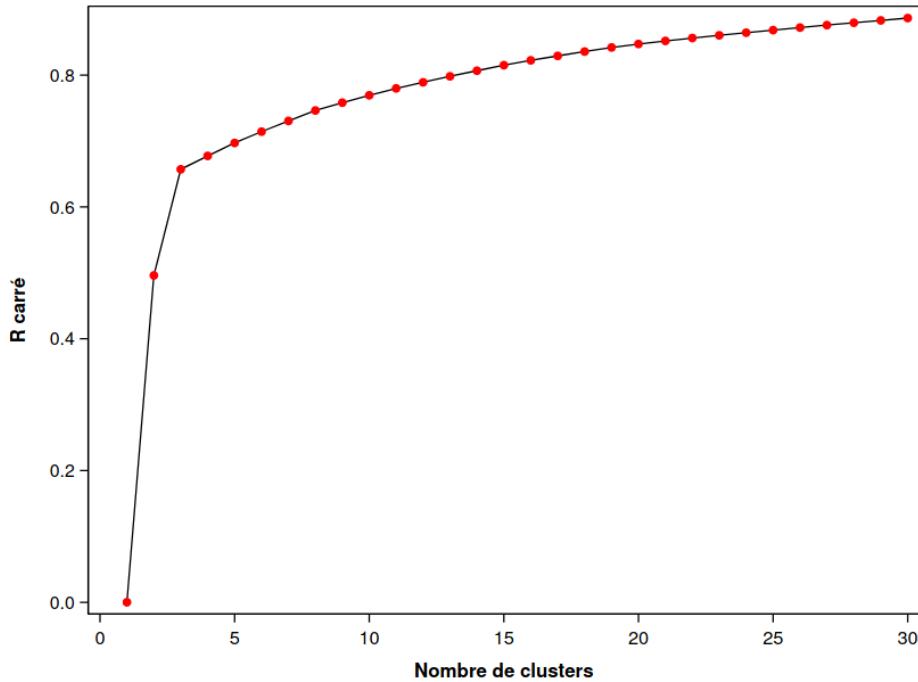


FIGURE 3.5 – R carré en fonction des regroupements hiérarchiques

Parfois, l'information est présentée sous forme de dendrogramme, qui trace l'arbre et la fusion des groupes. On peut ainsi retracer l'historique de la procédure hiérarchique. Celui produit par **SAS** donne, à un facteur multiplicatif près, le SPRSQ. Il n'y a donc pas de nouvelles informations ici. On voit que c'est lorsqu'on passe de trois à deux groupes, qu'il y a une bonne perte d'homogénéité.

En pratique, on ne peut jamais savoir si on a bel et bien regroupé ensemble les bons sujets. Mais ici, comme il s'agit de données artificielles qui ont été générées, nous connaissons la composition des vrais groupes. Il s'avère qu'il y en a effectivement trois. De plus, la solution à trois groupes obtenue avec la méthode de Ward a bien réussi à retrouver les groupes. Ceci est un exemple facile où les observations sont bien séparées : ce ne sera pas toujours aussi simple en pratique.

Interprétation des groupes : la méthode la plus simple consiste à inspecter les moyennes des variables de chaque groupe et de voir s'il en découle une interprétation raisonnable. La procédure **tree** permet d'extraire la solution avec un nombre spécifié de groupes et il est ensuite facile (avec la procédure **means**) d'obtenir ces moyennes (voir le fichier **cluster2_complet.sas**).

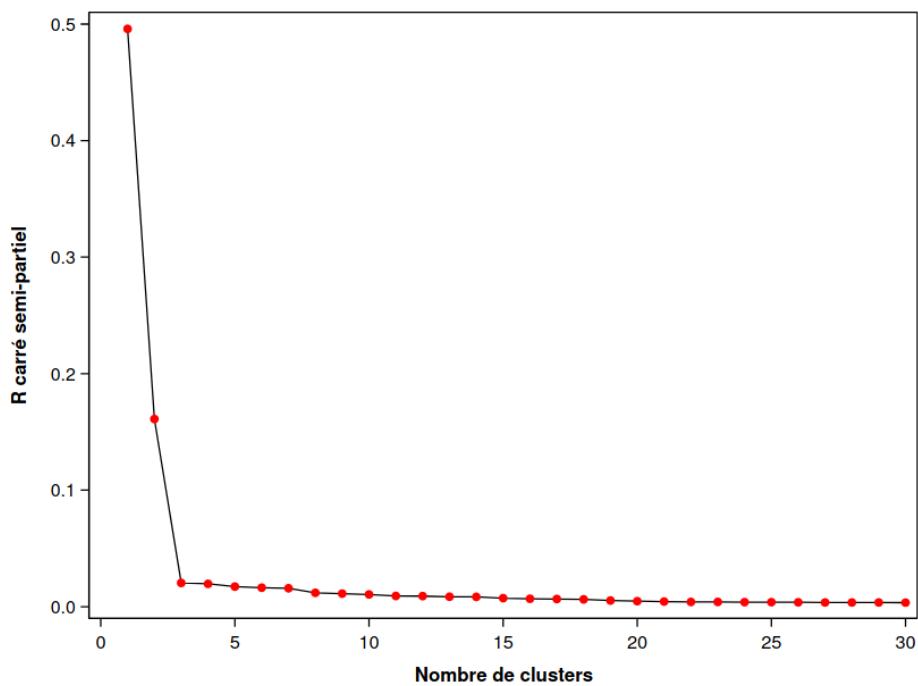


FIGURE 3.6 – R carré semi-partiel en fonction des regroupements hiérarchiques

CLUSTER=1

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
x1	x1	43	1.70	0.53	1.00	2.80
x2	x2	43	2.01	0.79	1.00	3.90
x3	x3	43	1.55	0.50	1.00	2.80
x4	x4	43	2.06	0.72	1.00	4.40
x5	x5	43	2.06	0.77	1.00	3.70
x6	x6	43	1.57	0.59	1.00	3.20

CLUSTER=2

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
x1	x1	75	3.00	0.66	1.30	4.90
x2	x2	75	4.05	0.71	2.40	5.00
x3	x3	75	3.86	0.80	2.10	5.00
x4	x4	75	3.85	0.71	2.50	5.00
x5	x5	75	4.19	0.65	2.10	5.00
x6	x6	75	4.27	0.64	2.60	5.00

CLUSTER=3

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
x1	x1	32	4.43	0.67	2.40	5.00
x2	x2	32	2.88	0.76	1.00	4.50
x3	x3	32	2.72	0.70	1.00	3.80
x4	x4	32	3.19	0.71	1.50	5.00
x5	x5	32	2.97	0.79	1.50	4.90
x6	x6	32	2.72	0.89	1.10	4.80

Le groupe 1 est le groupe où les sujets ont les valeurs, en moyenne, les plus faibles pour les six variables. Le groupe 2 est celui où les sujets ont les valeurs, en moyenne, les plus élevées pour les 6 variables sauf pour la variable X_1 (activité sociale). Le groupe 3 est celui où les sujets ont, en moyenne, la valeur la plus élevée de la variable X_1 et des valeurs moyennes inférieures au groupe 3 mais supérieures au groupe 2 pour les cinq autres variables.

Dans l'article, les auteurs ont baptisé les sujets du groupe 1, les « indépendants », ceux du groupe 2, les « dépendants » et ceux du groupe 3, les « sociables ». Notez qu'on **ne peut pas** tester l'égalité des moyennes des variables pour les différents groupes avec une ANOVA; la sélection des groupes est faite à l'aide d'un algorithme glouton pour maximiser la distance entre les groupes, aussi cela invalide l'inférence. On peut aussi explorer les groupes en modélisant les effets des variables en ce qui a trait à l'appartenance aux groupes. Traditionnellement, l'analyse discriminante est utilisée à cette fin. Il est aussi possible d'utiliser un arbre de classification ou une autre méthode prévisionnelle, telle la régression multinomiale logistique. La variable identifiant le groupe d'appartenance obtenu avec l'analyse de regroupement sert alors de variable dépendante Y . Ce type d'analyse permet de creuser un peu plus pour essayer de comprendre la structure des groupes formés.

3.5 Calcul alternatif des distances pour le regroupement hiérarchique

Nous avons utilisé la méthode de Ward afin de calculer la distance entre les groupes et procéder au passage de n groupes à un groupe, avec l'approche hiérarchique. Supposons que nous avons choisi une mesure de dissemblance $d(S_i, S_j)$ quelconque (distance euclidienne par exemple) pour mesurer la distance entre deux sujets. Voici comment sont choisis les regroupements avec ces méthodes.

- Méthode du plus proche voisin ou méthode de liaison simple (*nearest neighbor, single linkage*) : utilise la distance minimale entre chaque paire de sujets (un pour chaque groupe) provenant des deux groupes. Cette méthode fonctionne bien si l'écart entre deux regroupements est suffisamment grand. À l'inverse, s'il y a des observations bruitées entre deux regroupements, la qualité des regroupements en sera affectée.
- Méthode du voisin le plus éloigné ou méthode de liaison complète (*complete linkage*) : utilise la distance maximale entre toutes les paires de sujets (un pour chaque groupe) provenant des deux groupes. Cette méthode est moins sensible au bruit et aux faibles écarts entre regroupements, mais a tendance à casser les regroupements globulaires.
- Méthode de liaison moyenne (*average linkage*) : utilise la moyenne des distances entre toutes les paires de sujets (un pour chaque groupe) provenant des deux groupes.
- Méthode du barycentre (*centroid*) : utilise la distance entre les représentants moyens de chaque groupe où le représentant moyen d'un groupe est le barycentre, soit la moyenne variable par variable, des sujets formant le groupe.

Le fichier `cluster3_voisin_eloigne.sas` contient les commandes pour utiliser la méthode du voisin le plus éloigné (avec l'option `method=complete`). Le graphe plus bas donne cette distance pour les deux groupes qui viennent d'être fusionnés. Il s'agit donc du maximum des distances entre chaque paire de sujets (un pour chaque groupe) provenant des deux groupes fusionnés.

Comme on veut que cette distance soit petite pour les groupes fusionnés, on pourrait être tenté d'arrêter à trois groupes ici sur la base de la Figure 3.8. L'interprétation des groupes ne change pas comparativement aux analyses précédentes. La taille des groupes, (44, 71, 35), change un peu par rapport à la solution avec la méthode de Ward qui donnait des tailles de (43, 75, 32).

On peut comparer les performances des regroupements hiérarchiques selon la méthode de regroupement. La page web de scikit-learn developers montre la performance sur des exemples jouets, qui montre que selon les hypothèses et la structure, aucune ne performe mieux que les autres dans tous les exemples.

3.6 Standardisation des variables

Dans les exemples précédents, nous avons utilisé les variables X_1 à X_6 telles quelles. En général, plus une variable a une grande variance, plus elle aura de l'influence sur le calcul de la distance

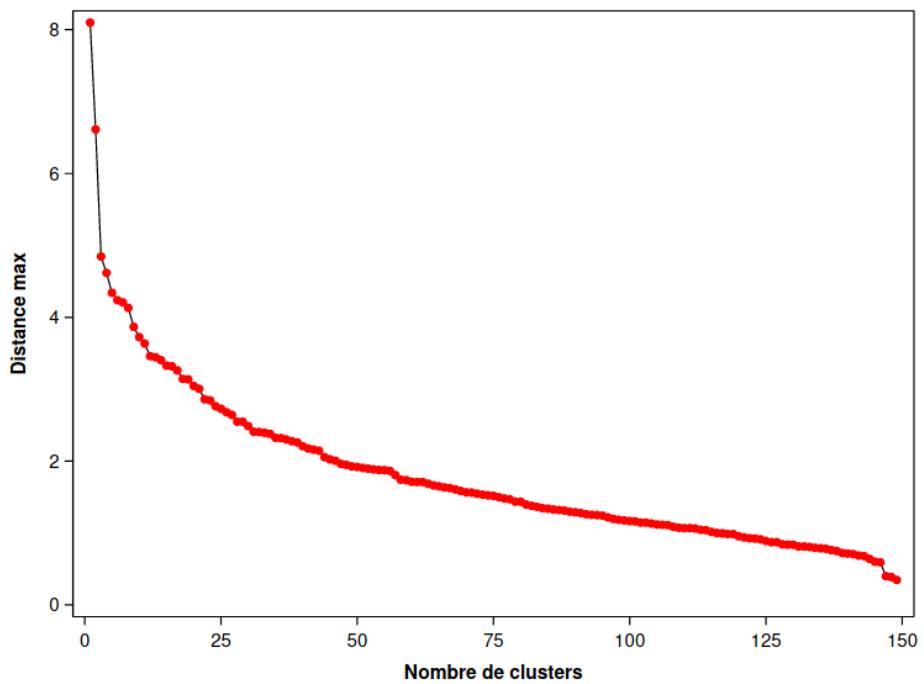


FIGURE 3.7 – Distance maximale entre groupes en fonction des regroupements hiérarchiques pour la méthode du voisin le plus éloigné.

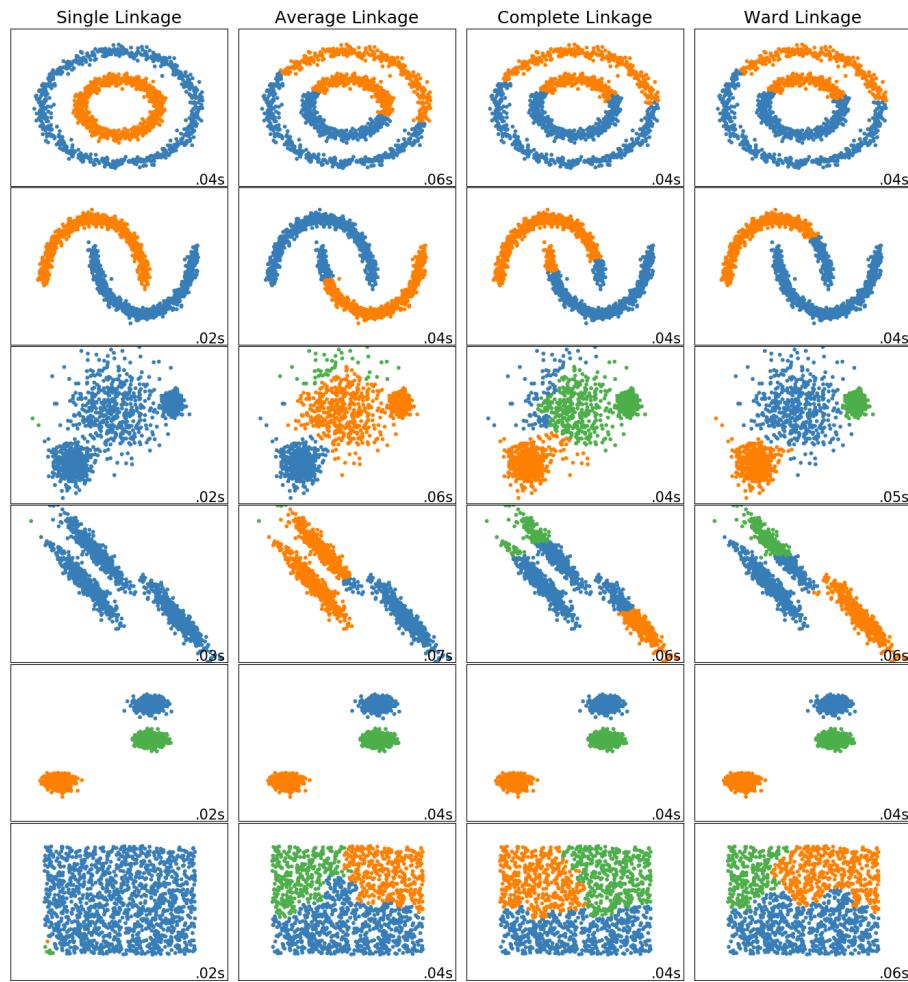


FIGURE 3.8 – Comparaison des méthodes de groupement sur des données test.

euclidienne. Ainsi, en utilisant les variables telles quelles, nous accordons plus de poids aux variables avec de grandes variances, ce qui peut être bon ou mauvais selon la structure des groupes. Règle générale, il est préférable d'éviter qu'une variable domine dans la segmentation.

Avec la procédure `stdize`, on peut standardiser au préalable les variables avant de faire l'analyse. Par défaut, les variables seront standardisées afin d'avoir une moyenne de zéro et une variance de un. On peut ensuite faire les analyses comme précédemment. Le fichier `cluster4_standardisation.sas` contient les commandes pour standardiser les variables et ensuite refaire l'analyse de regroupements avec la méthode de Ward. Notez que les six variables ont des variances semblables, donc les résultats ne devraient donc pas être trop affectés en standardisant les variables. Il s'avère effectivement que les résultats changent très peu si on standardise les variables. Les tailles des trois groupes de la solution sont (44, 75, 31), comparativement à (43, 75, 32) sans la standardisation.

3.7 Autres mesures de dissemblance

Nous avons déjà mentionné que lorsque toutes les variables sont continues, la mesure de dissemblance la plus utilisée est la distance euclidienne. Dans la procédure `cluster`, SAS utilise la distance euclidienne au carré par défaut. Pour utiliser la distance euclidienne elle-même, il faut mettre le mot clé `nosquare` dans les options suivant l'appel à `proc cluster` (première ligne). Pour utiliser une mesure de dissemblance autre que la distance euclidienne (au carré ou pas), on peut utiliser la procédure `distance` au préalable pour calculer les distances, et ensuite fournir la matrice des distances directement à `cluster` en lieu et place des observations.

Il existe un très grand nombre d'autres mesures de dissemblance pour variables quantitatives, ordinaires, nominales et binaires. Voici une brève description de certaines d'entre elles, qui sont disponibles dans `proc distance`.

Mesures de dissemblance pour variables quantitatives :

- 1) Distance euclidienne ou distance euclidienne au carré
- 2) Distance de Manhattan, ou taxi-distance :

$$d(S_i, S_j) = |X_{i1} - X_{j1}| + \dots + |X_{ip} - X_{jp}|$$

Mesure de dissemblance pour variables nominales :

Le plus simple est d'utiliser la mesure d'appariement simple (*simple matching*). Cette mesure est simplement de la proportion des variables pour lesquelles les deux sujets ont des valeurs différentes.

Mesures de dissemblances pour variables ordinaires :

- 1) Une manière simple consiste à les traiter comme des variables continues, et utiliser la distance euclidienne ou la distance de Manhattan. On peut aussi les standardiser au préalable si elles ne sont pas sur la même échelle (par exemple, si certaines vont de 1 à 5 et d'autres de 1 à 7).
- 2) On peut aussi les traiter comme des variables nominales avec la mesure d'appariement simple; ce faisant, on n'utilise pas l'ordre entre les modalités.

Le fichier `cluster4_cityblock.sas` contient les commandes pour refaire l'analyse des regroupements du voyage organisé avec la taxi-distance et la méthode de liaison moyenne (`method=average`) dans `proc cluster`.

Encore une fois, l'interprétation des groupes ne change pas comparativement aux analyses précédentes. La taille des groupes, (45, 75, 30) change un peu par rapport à la solution avec la méthode de Ward qui donnait des tailles de (43, 75, 32).

Règle générale, les différentes étapes des méthodes agglomératives hiérarchiques nécessitent $O(n^3)$ opérations, bien qu'une version plus parsimonieuse existe avec complexité $O(n^2 \ln n)$ ou $O(n^2)$ pour les méthodes de liaison simple et complexe. La formule de Lance–Williams permet de mettre à jour récursivement les distances entre regroupements pour la plupart des méthodes considérées. Le coût élevé de la méthode de regroupement hiérarchique, qui dépend de la taille de l'échantillon, devient prohibitif avec des mégadonnées. Il nécessite aussi le calcul d'une mesure de dissemblance et l'évaluation de la qualité de l'agglomération, autre que graphique, n'est pas évidente. Ces méthodes sont largement discontinuées de nos jours par des alternatives modernes (regroupement spectraux, mélanges de modèles, K -médoïdes par itérations de Voronoï) qui ne sont pas implémentées en SAS.

3.8 Méthodes non hiérarchiques

Contrairement aux méthodes hiérarchiques, il faut spécifier le nombre de groupe désiré dès le départ pour les méthodes non hiérarchiques. La méthode des K moyennes (*K means*) sera la seule décrite ici. Cette méthode utilise la distance euclidienne et est donc seulement applicable avec des variables quantitatives. Supposons que l'on veuille K groupes. La méthode des K moyennes peut être décrite en trois étapes :

- 1) On sélectionne K germes (*seeds*) qui vont agir comme centres préliminaires des groupes.
 - i) Ces germes peuvent être les centres des groupes obtenus à partir d'une autre méthode comme une méthode hiérarchique.
 - ii) Ces germes peuvent être choisis à même l'ensemble de données. Par exemple, on peut choisir K sujets au hasard.
- 2) On assigne dans l'ordre les observations au groupe le plus proche en utilisant la distance euclidienne par rapport au germe. Soit on assigne toutes les observations avant de mettre

à jour les germes, soit on met à jour les germes après chaque assignation d'un sujet. Le nouveau germe d'un groupe est la moyenne des observations faisant partie du groupe.

- 3) On peut répéter le processus jusqu'à ce que les changements des germes des groupes deviennent négligeables ou nuls. Les groupes finaux sont formés en assignant les sujets au groupe le plus proche.

À chaque étape, l'algorithme des K -moyennes minimise le critère

$$\sum_{i=1}^n \sum_{k=1}^K I(x_i \in g_k) \|x_i - \mu_k\|^2 = \sum_{i=1}^n \sum_{k=1}^K I(x_i \in g_k) \sum_{j=1}^p (x_{ij} - \mu_{kj})^2$$

où $I(x_i \in g_k)$ est une variable binaire qui vaut un si l'observation i appartient au groupe g_k et μ_k est le centroïde du regroupement k . On minimise ainsi la variance intra-groupe. On voit ainsi dans la Figure 3.9 que l'algorithme détecte bien les vrais regroupements si les groupes sont sphériques et de même variance, et que les regroupements sont bien séparés (c'est-à-dire, quand la distance euclidienne entre les regroupements est élevée).

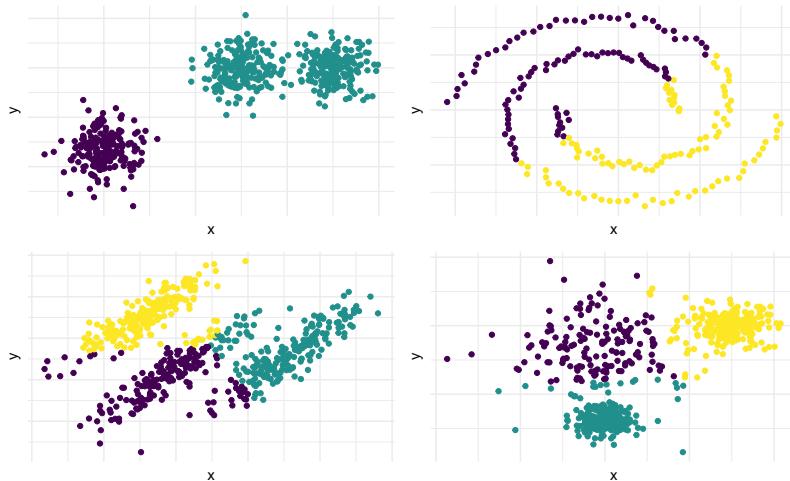


FIGURE 3.9 – Performance de l'algorithme des K -moyennes en fonction de différents scénarios. En haut à gauche : nombre incorrect de classe et données normales de même variance, bien séparées. En haut à droite : spirales : les K -moyennes ignorent la topologie des regroupements, et ne segmentent pas adéquatement les regroupements connectés. En bas à gauche : données elliptiques de même variance, mais fortement corrélées. Comme le critère minimise la distance intra-groupe sans pondération, les points regroupés appartiennent à différentes classes. En bas à droite : données sphériques de variances différentes. L'algorithme des K -moyennes réussit une bonne segmentation si les groupements sont compacts et bien séparés.

Nous allons utiliser cette procédure pour raffiner la solution obtenue précédemment avec la méthode de Ward en utilisant les moyennes des groupes comme centres préliminaires. Le fichier

`cluster5_non-hierarchique.sas` explique les différentes options. La syntaxe de la procédure **SAS** fastclus est la suivante :

```
proc fastclus data=temp seed=initial distance maxclusters=3 out=temp3 maxiter=30;
var x1 x2 x3 x4 x5 x6;
run;
```

Voici une partie de la sortie **SAS** :

La procédure FASTCLUS
Replace=FULL Radius=0 Maxclusters=3 Maxiter=30 Converge=0.02

Cluster	Valeurs initiales					
	x1	x2	x3	x4	x5	x6
1	1.702325581	2.011627907	1.548837209	2.058139535	2.058139535	1.574418605
2	3.000000000	4.053333333	3.864000000	3.853333333	4.194666667	4.274666667
3	4.428125000	2.881250000	2.721875000	3.187500000	2.968750000	2.725000000

Distance minimale entre les valeurs init 3.010887

Historique des itérations

Changement relatif dans les valeurs initiales de clusters

Itération	Critère	1	2	3
1	0.6899	0.0290	0.0197	0.0583
2	0.6887	0	0	0

Répond au critère de convergence.

Synthèse des clusters						
Cluster	Fréquence	Ecart type RMS	Distance max.		Distance entre	
			de la valeur initiale	Rayon à l'obs. dépassé	Cluster le + proche	centoïdes des clusters
1	45	0.6719		2.5779	3	3.5804
2	77	0.7047		2.7997	3	3.0661
3	28	0.7079		2.7087	2	3.0661

Statistiques pour variables				
Variable	STD total	Intra-STD	R-cartré	RSQ/(1-RSQ)
x1	1.14215	0.62954	0.700267	2.336300
x2	1.15857	0.73099	0.607256	1.546186
x3	1.22032	0.71005	0.665985	1.993881
x4	1.04781	0.71807	0.536659	1.158239
x5	1.17555	0.72650	0.623191	1.653864
x6	1.36307	0.65221	0.774121	3.427156
OVER-ALL	1.18840	0.69566	0.661934	1.958005

Dist. entre centroïdes de classe				
Cluster le + proche	1	2	3	
1	.	5.066445139	3.580358275	
2	5.066445139	.	3.066072576	
3	3.580358275	3.066072576	.	

Évidemment, comme la solution obtenue avec la méthode de Ward est déjà excellente, on ne pourra pas avoir une amélioration notable. Il y a peu de changements par rapport à la solution de la méthode de Ward. Les tailles des groupes étaient de (43, 75, 32) avant. Elles sont maintenant (45, 77, 28). Le R^2 passe de 65,7% (avec Ward) à 66.2%.

L'interprétation des groupes est la même que précédemment.

Moyennes des classes						
Cluster	x1	x2	x3	x4	x5	x6
1	1.742222222	1.988888889	1.604444444	2.100000000	2.068888889	1.597777778
2	3.048051948	4.035064935	3.859740260	3.857142857	4.166233766	4.281818182
3	4.528571429	2.946428571	2.646428571	3.142857143	3.007142857	2.639285714

Ecarts types des classes						
Cluster	x1	x2	x3	x4	x5	x6
1	0.5508212418	0.7886185338	0.5530722276	0.7351561368	0.7600305682	0.5994273361
2	0.7114866551	0.7102198017	0.7921090668	0.7049502729	0.6687794046	0.6305234402
3	0.4882752281	0.6898642609	0.6930780208	0.7264467411	0.8205366678	0.7818966900

Le champ des applications des K -moyennes est parfois surprenant. Par exemple, cet article de FiveThirtyEight propose une segmentation des électeurs démocrates new-yorkais. Un autre exemple incongru est la compression d'images : la Figure 3.10 montre une image du bâtiment Decelles (coin supérieur gauche) et la reconstruction avec trois, quatre et 10 couleurs obtenues en appliquant l'algorithme des K -moyennes sur la matrice formée par les valeurs des canaux (rouge, vert, bleu) de l'image.



FIGURE 3.10 – Compression d'image avec l'algorithme des K -moyennes : image originale (en haut à gauche), compression avec trois (en haut à droite), quatre (en bas à gauche) et 10 (en bas à droite) couleurs.

3.9 Considérations pratiques

Il peut être intéressant de comparer les résultats provenant d'une même méthode avec des nombres différents de groupes et aussi comparer ceux provenant de plusieurs méthodes (voir plus loin pour la description de certaines autres méthodes). Le choix de la méthode et du nombre de groupe n'est pas facile et devrait être basé sur des considérations pratiques et d'interprétation (comme en analyse factorielle). Il n'est pas rare qu'on obtienne des résultats très différents d'une méthode à l'autre pour un même ensemble de données.

Avec une méthode non hiérarchique, il est préférable de fournir des germes de départ « raisonnablement bon » (provenant d'une méthode hiérarchique par exemple) plutôt que de laisser l'algorithme les choisir au hasard.

Le choix des variables est important. En général on veut créer des groupes d'individus qui sont homogènes par rapport à un certain aspect de leur comportement ou de leur situation. On ne doit alors inclure que les variables pertinentes à cet aspect. Par exemple, si le but de l'analyse est de segmenter nos clients selon leurs habitudes de consommation (genre de boutiques fréquenté, fréquence, etc.), on ne voudra peut-être pas inclure des variables démographiques. En fait, souvent l'analyse de regroupements servira justement à créer des groupes qui seront comparés par rapport à d'autres variables qui n'ont pas été utilisées pour créer les groupes. Dans notre exemple sur les voyages organisés, on a segmenté les voyageurs en trois groupes (indépendants, dépendants et sociables). Les auteurs de l'article (voir page 369 de l'article) ont comparé les trois groupes selon l'expérience de voyage, la taille de la communauté où ils habitent (avec des ANOVA), selon leur âge, leur revenu et leur éducation (avec des tests d'indépendance du khi-deux). Notez que ces tests sont réalisés sur des variables qui ne sont pas utilisées lors de la segmentation, mais demeurent invalides si les données sont corrélées avec celle utilisées pour la segmentation.

Le problème majeur avec l'analyse de regroupements est qu'il n'y a pas de façon claire de quantifier la performance de notre analyse. Lorsqu'on développe un modèle de prédiction (régression linéaire ou logistique par exemple), on peut estimer la performance de notre modèle d'une manière objective à l'aide de l'erreur quadratique de généralisation (régression linéaire) ou du taux de bonne classification (régression logistique). Ces quantités peuvent être estimées d'une manière objective en utilisant une méthode telle la validation croisée ou la division de l'échantillon. On ne peut faire de même avec l'analyse de regroupements car on n'a pas de variable réponse à prédire. Tout comme pour l'analyse factorielle, les connaissances à priori, le jugement, et les considérations pratiques font partie d'une analyse de regroupements.

Chapitre 4

Sélection de variables et de modèles

4.1 Introduction

Ce chapitre présente des principes, outils et méthodes très généraux pour choisir un « bon » modèle. Nous allons principalement utiliser la régression linéaire pour illustrer les méthodes en supposant que tout le monde connaît ce modèle de base. Les méthodes présentées sont en revanche très générales et peuvent être appliquées avec n'importe quel autre modèle (régression logistique, arbres de classification et régression, réseaux de neurones, analyse de survie, etc.)

L'expression « sélection de variables » fait référence à la situation où l'on cherche à sélectionner un sous-ensemble de variables à inclure dans notre modèle à partir d'un ensemble de variables X_1, \dots, X_p . Le terme variable ici inclut autant des variables distinctes que des transformations d'une ou plusieurs variables.

Par exemple, supposons que les variables `age`, `sex` et `revenu` sont trois variables explicatives disponibles. Nous pourrions alors considérer choisir entre ces trois variables. Mais aussi, nous pourrions considérer inclure age^2 , age^3 , $\log(\text{age})$, etc. Nous pourrions aussi considérer des termes d'interactions entre les variables, comme $\text{age} \cdot \text{revenu}$ ou $\text{age} \cdot \text{revenu} \cdot \text{sex}$. Le problème est alors de trouver un bon sous-ensemble de variables parmi toutes celles considérées.

L'expression « sélection de modèle » est un peu plus générale. D'une part, elle inclut la sélection de variables car, pour une famille de modèles spécifiques (régression linéaire par exemple), choisir un sous-ensemble de variables revient à choisir un modèle. D'autre part, elle est plus générale car elle peut aussi faire référence à la situation où l'on cherche à trouver le meilleur modèle parmi des modèles de natures différentes. Par exemple, on pourrait choisir entre une régression linéaire, un arbre de régression, une forêt aléatoire, un réseau de neurones, etc.

4.2 Sélection de variables et de modèles selon les buts de l'étude

Nous disposons d'une variable réponse Y et d'un ensemble de variables explicatives X_1, \dots, X_p . L'attitude à adopter dépend des buts de l'étude.

1e situation : On veut développer un modèle pour faire des prédictions sans qu'il soit important de tester formellement les effets des paramètres individuels.

Dans ce cas, on désire seulement que notre modèle soit performant pour prédire des valeurs futures de Y . On peut alors baser notre choix de variable (et de modèle) en utilisant des outils qui nous guideront quant aux performances prédictives futures du modèle (voir AIC, BIC et validation croisée plus loin). On pourra enlever ou rajouter des variables et des transformations de variables au besoin afin d'améliorer les performances prédictives. Les méthodes que nous allons voir concernent essentiellement ce contexte.

2e situation : On veut développer un modèle pour estimer les effets de certaines variables sur notre Y et tester des hypothèses de recherche spécifiques concernant certaines variables.

Dans ce cas, il est préférable de spécifier le modèle dès le départ selon des considérations scientifiques et de s'en tenir à lui. Faire une sélection de variables dans ce cas est dangereux car on ne peut pas utiliser directement les valeurs- p des tests d'hypothèses (ou les intervalles de confiance sur les paramètres) concernant les paramètres du modèle final car elles ne tiennent pas compte de la variabilité due au processus de sélection de variables.

Une bonne planification de l'étude est alors cruciale afin de collecter les bonnes variables, de spécifier le ou les bons modèles, et de s'assurer d'avoir suffisamment d'observations pour ajuster le ou les modèles désirés.

Si procéder à une sélection de variables est quand même nécessaire dans ce contexte, il est quand même possible de le faire en divisant l'échantillon en deux. La sélection de variables pourrait être alors effectuée avec le premier échantillon. Une fois qu'un modèle est retenu, on pourrait alors réajuster ce modèle avec le deuxième échantillon (sans faire de sélection de variables cette fois-ci). L'inférence sur les paramètres (valeurs- p , etc.) sera alors valide. Le désavantage ici qu'il faut avoir une très grande taille d'échantillon au départ afin d'être en mesure de le diviser en deux.

4.3 Mieux vaut plus que moins

Il est préférable d'avoir un modèle un peu trop complexe qu'un modèle trop simple. Plaçons-nous dans le contexte de la régression linéaire et supposons que le vrai modèle est inclus dans le modèle qui a été ajusté. Il y a donc des variables en trop dans le modèle qui a été ajusté : ce dernier est dit surspécifié.

Par exemple, supposons que le vrai modèle est $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ mais que c'est le modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ qui a été ajusté. Dans ce cas, règle générale, les estimateurs des paramètres et

les prédictions provenant du modèle sont sans biais. Mais leurs variances estimées seront un peu plus élevées car on estime des paramètres pour des variables superflues.

Supposons à l'inverse qu'il manque des variables dans le modèle ajusté et que le modèle ajusté est sous-spécifié. Par exemple, supposons que le vrai modèle est $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, mais que c'est le modèle $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ qui est ajusté. Dans ce cas, généralement, les estimateurs des paramètres et les prédictions sont biaisés.

Ainsi, il est généralement préférable d'avoir un modèle légèrement surspécifié qu'un modèle sous-spécifié. Plus généralement, il est préférable d'avoir un peu trop de variables dans le modèle que de prendre le risque d'omettre une ou plusieurs variables importantes. Il faut faire attention et ne pas tomber dans l'excès et avoir un modèle trop complexe (avec trop de variables inutiles) car il pourrait souffrir de surajustement (*over-fitting*). Les exemples qui suivent illustreront ce fait.

4.4 Trop beau pour être vrai

Cette section traite de l'optimisme de l'évaluation d'un modèle lorsqu'on utilise les mêmes données qui ont servies à l'ajuster pour évaluer sa performance. Un principe fondamental lorsque vient le temps d'évaluer la performance prédictive d'un modèle est le suivant : si on utilise les mêmes observations pour évaluer la performance d'un modèle que celles qui ont servi à l'ajuster (à estimer le modèle et ses paramètres), on va surestimer sa performance. Autrement dit, notre estimation de l'erreur que fera le modèle pour prédire des observations futures sera biaisée à la baisse. Ainsi, il aura l'air meilleur que ce qu'il est en réalité. C'est comme si on demandait à un cinéaste d'évaluer son dernier film. Comme c'est son film, il n'aura généralement pas un regard objectif. C'est pourquoi on aura tendance à se fier à l'opinion d'un critique.

On cherchera donc à utiliser des outils et méthodes qui nous donneront l'heure juste (une évaluation objective) quant à la performance prédictive d'un modèle.

4.5 Principes généraux

Les idées présentées ici seront illustrées à l'aide de la régression linéaire. Par contre, elles sont valides dans à peu près n'importe quel contexte de modélisation.

Plaçons-nous d'abord dans un contexte plus général que celui de la régression linéaire. Supposons que l'on dispose de n observations indépendantes sur (Y, X_1, \dots, X_p) et que l'on a ajusté un modèle $\hat{f}(X_1, \dots, X_p)$, avec ces données, pour prédire une variable continue Y .

Ce modèle peut être un modèle de régression linéaire,

$$\hat{f}(X_1, \dots, X_p) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

mais il pourrait aussi avoir été construit selon d'autres méthodes (réseau de neurones, arbre de régression, forêt aléatoire, etc.) Une manière de quantifier la performance prédictive du modèle est l'erreur quadratique moyenne de généralisation (*generalization mean squared error*),

$$\text{EQMG} = \mathbb{E} \left[\{(Y - \hat{f}(X_1, \dots, X_p)\}^2 \right]$$

lorsque (Y, X_1, \dots, X_p) est choisi au hasard dans la population. Cette quantité mesure l'erreur théorique (la différence au carré entre la vraie valeur de Y et la valeur prédite par le modèle) que fait le modèle en moyenne pour l'ensemble de la population. Plus cette quantité est petite, meilleur est le modèle. Le problème est que l'on ne peut pas la calculer car on n'a pas accès à toute la population. Tout au plus peut-on essayer de l'estimer ou bien d'estimer une fonction qui, sans l'estimer directement, classifiera les modèles dans le même ordre qu'elle.

Une première idée est d'estimer l'erreur quadratique moyenne de l'échantillon d'apprentissage (*training mean squared error*),

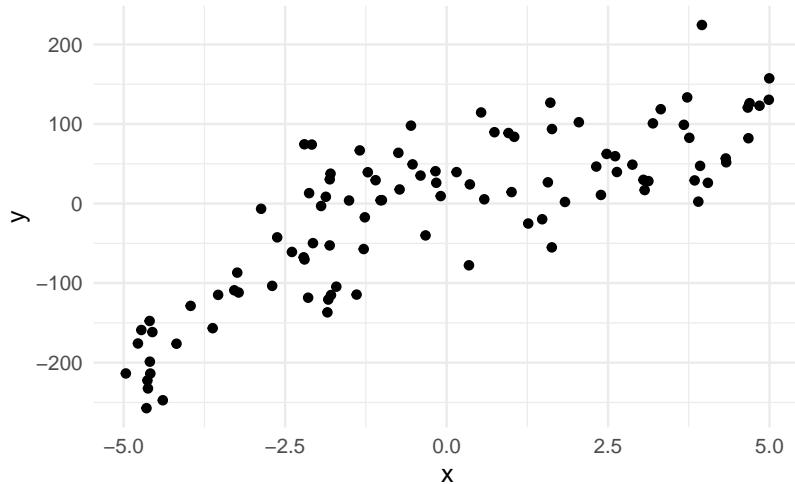
$$\widehat{\text{EQM}}_a = \frac{1}{n} \sum_{i=1}^n \{Y_i - \hat{f}(X_{i1}, \dots, X_{ip})\}^2.$$

Malheureusement, selon le principe fondamental de la section précédente, cette quantité n'est pas un bon estimateur de l'EQMG. En effet, comme on utilise les mêmes observations que celles qui ont estimé le modèle, l' $\widehat{\text{EQM}}_a$ aura tendance à toujours diminuer lorsqu'on augmente la complexité du modèle (par exemple, lorsqu'on augmente le nombre de paramètres). L' $\widehat{\text{EQM}}_a$ tend à surestimer la qualité du modèle en sous-estimant l'EQMG et le modèle a l'air meilleur qu'il ne l'est en réalité.

4.5.1 Choix d'un modèle polynomial en régression linéaire

Cet exemple simple servira à illustrer le fait qu'on ne peut utiliser directement les mêmes données qui ont servi à ajuster un modèle pour évaluer sa performance.

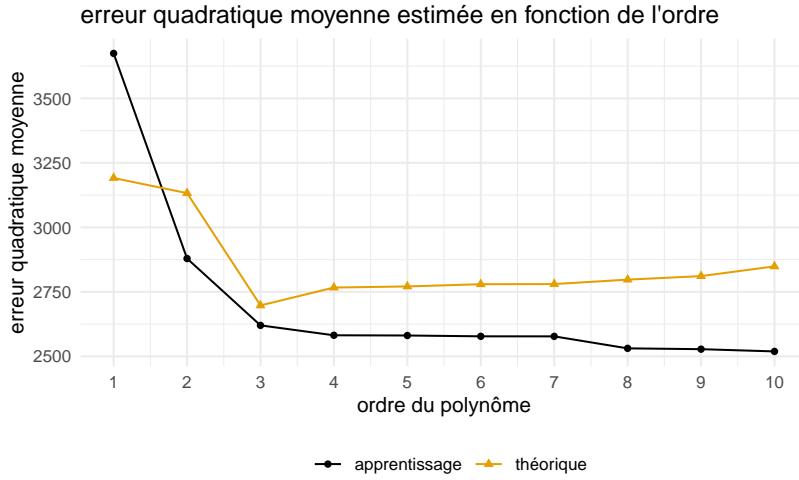
Nous disposons de 100 observations sur une variable cible Y et d'une seule variable explicative X . Le fichier `selection1_train` contient les données. Nous voulons considérer des modèles polynomiaux (en X) afin d'en trouver un bon pour prédire Y . Un modèle polynomial est un modèle de la forme $Y = \beta_0 + \beta_1 X + \dots + \beta_k X^k + \varepsilon$. Le cas $k = 1$ correspond à un modèle linéaire simple, $k = 2$ à un modèle cubique, $k = 3$ à un modèle cubique, etc. Notre but est de déterminer l'ordre (k) du polynôme qui nous donnera un bon modèle. Voici d'abord le graphe de ces 100 observations de l'échantillon d'apprentissage.



Ces données ont été obtenues par simulation et le vrai modèle sous-jacent (celui qui a généré les données) est le modèle cubique, c'est-à-dire le modèle d'ordre $k = 3$.

J'ai ajusté tour à tour les modèles polynomiaux jusqu'à l'ordre 10, avec l'échantillon d'apprentissage de taille 100. C'est-à-dire, le modèle linéaire avec un polynôme d'ordre $k = 1$ (linéaire), $k = 2$ (quadratique), etc., jusqu'à $k = 10$. J'ai ensuite obtenu la valeur de l'erreur quadratique moyenne d'apprentissage pour chacun de ces modèles. En pratique, on ne pourrait pas calculer l'erreur quadratique moyenne de généralisation, mais j'ai approximé cette dernière en simulant 100 000 observations du vrai modèle (`selection1_test`), en obtenant la prédiction pour chacune de ces 100 000 observations en utilisant le modèle d'ordre k ajusté sur les données d'apprentissage et en calculant l'erreur moyenne quadratique par la suite. En pratique, on ne peut réaliser cette opération car on ne connaît pas le vrai modèle.

Voici le graphe de l'erreur moyenne quadratique d'apprentissage ($\widehat{\text{EQM}}_a$) et de l'erreur moyenne quadratique théorique (EQMG) en fonction de l'ordre (k) du modèle utilisé.



On voit clairement que l' $\widehat{\text{EQM}}_a$ diminue en fonction de l'ordre sur l'échantillon d'apprentissage : plus le modèle est complexe, plus l'erreur observée sur l'échantillon d'apprentissage est petite. La courbe EQM donne l'heure juste, car il s'agit d'une estimation de la performance réelle des modèles sur de nouvelles données. On voit que le meilleur modèle est donc le modèle cubique ($k = 3$), ce qui n'est pas surprenant puisqu'il s'agit du modèle que utilisé pour générer les données. On peut aussi remarquer d'autres éléments intéressants. Premièrement, on obtient un bon gain en performance (EQM) en passant de l'ordre 2 à l'ordre 3. Ensuite, la perte de performance en passant de l'ordre 3 à 4, et ensuite à des ordres supérieurs n'est pas si sévère, même si elle est présente. Cela illustre empiriquement qu'il est préférable d'avoir un modèle un peu trop complexe que d'avoir un modèle trop simple. Il serait beaucoup plus grave pour la performance de choisir le modèle avec $k = 2$ que celui avec $k = 4$.

En pratique par contre, on n'a pas accès à la population : les 100 000 observations qui ont servi à estimer l'EQM théorique ne seront pas disponible. Si on a seulement l'échantillon d'apprentissage, soit 100 observations dans notre exemple, comment faire alors pour choisir le bon modèle ? C'est ce que nous verrons à partir de la section suivante.

Mais avant cela, nous allons discuter un peu plus en détail au sujet de la régression linéaire et d'une mesure très connue, le coefficient de détermination (R^2). Supposons que l'on a ajusté un modèle de régression linéaire

$$\hat{f}(X_1, \dots, X_p) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p.$$

La somme du carré des erreurs (SCE) pour notre échantillon est

$$\text{SCE} = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \dots - \hat{\beta}_p X_p)^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

On peut démontrer que si on ajoute une variable quelconque au modèle, la valeur de la somme du carré des erreurs va nécessairement baisser. Il est facile de se convaincre de cela. En régression linéaire, les estimations sont obtenues par la méthode des moindres carrés qui consiste justement à minimiser la SCE. Ainsi, en ajoutant une variable X_{p+1} au modèle, la SCE ne peut que baisser car, dans le pire des cas, le paramètre de la nouvelle variable sera $\hat{\beta}_{p+1} = 0$ et on retombera sur le modèle sans cette variable. C'est pourquoi, la quantité $\bar{EQM}_a = SCE/n$ ne peut être utilisée comme outil de sélection de modèles en régression linéaire.

Nous venons d'ailleurs d'illustrer cela avec notre exemple sur les modèles polynomiaux. En effet, augmenter l'ordre du polynôme de 1 revient à ajouter une variable. Le coefficient de détermination (R^2) est souvent utilisé, à tort, comme mesure de qualité du modèle. Il peut s'interpréter comme étant la proportion de la variance de Y qui est expliquée par le modèle.

Le coefficient de détermination est

$$R^2 = \{\text{cor}(\mathbf{y}, \hat{\mathbf{y}})\}^2 = 1 - \frac{\text{SCE}}{\text{SCT}},$$

où $\text{SCT} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ est la somme des carrés totale calculée en centrant les observations. La somme des carrés totale, SCT, ne varie pas en fonction du modèle. Ainsi, on voit que le R^2 va mécaniquement augmenter lorsqu'on ajoute une variable au modèle (car la SCE diminue). C'est pourquoi on ne peut pas l'utiliser comme outil de sélection de variables.

Le problème principal que nous avons identifié jusqu'à présent afin d'être en mesure de bien estimer la performance d'un modèle est le suivant : si on utilise les mêmes observations pour évaluer la performance d'un modèle que celles qui ont servi à l'ajuster, on va surestimer sa performance.

Il existe deux grandes approches pour contourner ce problème lorsque le but est de faire de la sélection de variables ou de modèle :

- utiliser les données de l'échantillon d'apprentissage (en échantillon) et pénaliser la mesure d'ajustement (ici \bar{EQM}_a) pour tenir compte de la complexité du modèle (par exemple, à l'aide de critères d'informations).
- tenter d'estimer l'EQM directement sur d'autres données (hors échantillon) en utilisant des méthodes de rééchantillonnage, notamment la validation croisée ou la validation externe (division de l'échantillon).

4.6 Pénalisation et critères d'information

Plaçons-nous dans le contexte de la régression linéaire pour l'instant. Nous avons déjà utilisé les critères AIC et BIC en analyse factorielle. Il s'agit de mesures qui découlent d'une méthode d'estimation des paramètres, la méthode du maximum de vraisemblance (*maximum likelihood*).

Il s'avère que les estimateurs des paramètres obtenus par la méthode des moindres carrés en régression linéaire sont équivalents à ceux provenant de la méthode du maximum de vraisemblance

si on suppose la normalité des termes d'erreurs du modèle. Ainsi, dans ce cas, nous avons accès aux AIC et BIC, deux critères d'information définis pour les modèles dont la fonction objective est la vraisemblance (qui mesure la probabilité des observations sous le modèle postulé suivant une loi choisie par l'utilisateur). La fonction de vraisemblance \mathcal{L} et la log-vraisemblance ℓ mesurent l'adéquation du modèle.

Supposons que nous avons ajusté un modèle avec p paramètres en tout (**inclus** l'ordonnée à l'origine). En régression linéaire, le critère d'information d'Akaike, AIC, est

$$\text{AIC} = -2\ell(\hat{\beta}, \hat{\sigma}^2) + 2p = n \ln(\text{SCE}) - n \ln(n) + 2p,$$

tandis que le critère d'information bayésien de Schwartz, BIC, est défini par

$$\text{BIC} = -2\ell(\hat{\beta}, \hat{\sigma}^2) + p \ln(n) = n \ln(\text{SCE}) - n \ln(n) + p \ln(n)$$

Plus la valeur du AIC (ou du BIC) est petite, meilleur est l'adéquation. Que se passe-t-il lorsqu'on ajoute un paramètre à un modèle? D'une part, la somme du carré des erreurs va mécaniquement diminuer, et donc la quantité $n \ln(\text{SCE}/n)$ va diminuer. D'autre part, la valeur de p augmente de 1. Ainsi, le AIC peut soit augmenter, soit diminuer, lorsqu'on ajoute un paramètre; idem pour le BIC. Par exemple, le AIC va diminuer seulement si la baisse de la somme du carré des erreurs est suffisante pour compenser le fait que le terme $2p$ augmente à $2(p+1)$.

Ces critères pénalisent l'ajout de variables afin de se prémunir contre le surajustement. De plus, le BIC pénalise plus que le AIC. Par conséquent, le critère BIC va choisir des modèles contenant soit le même nombre, soit moins de paramètres que le AIC.

Les critères AIC et BIC peuvent être utilisés comme outils de sélection de variables en régression linéaire mais aussi beaucoup plus généralement avec d'autres méthodes basées sur la vraisemblance (analyse factorielle, régression logistique, etc.) En fait, n'importe quel modèle dont les estimateurs proviennent de la méthode du maximum de vraisemblance produira ces quantités. Nous donnerons des formules générales pour le AIC et le BIC dans le chapitre sur la régression logistique.

Le critère BIC est le seul de ces critères qui est convergent. Cela veut dire que si l'ensemble des modèles que l'on considère contient le vrai modèle, alors la probabilité que le critère BIC choisisse le bon modèle tend vers 1 lorsque n tend vers l'infini. Il faut mettre cela en perspective : il est peu vraisemblable que Y ait été généré exactement selon un modèle de régression linéaire, car le modèle de régression n'est qu'une approximation de la réalité. Certains auteurs trouvent que le BIC est quelquefois trop sévère (il choisit des modèles trop simples) pour les tailles d'échantillons finies. Dans certaines applications, cette parcimonie est utile, mais il n'est pas possible de savoir d'avance lequel de ces deux critères (AIC et BIC) sera préférable pour un problème donné.

Avant de revenir à l'exemple, voici la description d'une modification du coefficient de détermination, le R^2 ajusté, qui permet (contrairement au R^2) de faire de la sélection de variables. En

régression linéaire, le R^2 ajusté est

$$R_a^2 = 1 - \frac{SCE/(n-p)}{SCT/(n-1)}.$$

Lorsqu'on ajoute une variable, la somme du carré des erreurs (SCE) diminue mais c'est aussi le cas de la quantité $(n-p)$. Ainsi, le R^2 ajusté peut soit augmenter, soit diminuer lorsqu'on ajoute une variable. On peut donc l'utiliser pour choisir le modèle. Plus R_a^2 est élevé, mieux c'est. Ce critère est moins sévère que le AIC. Ainsi, en général, il va choisir un modèle avec le même nombre ou bien avec plus de paramètres que le AIC. Pour résumer, on aura la situation suivante :

$$\#(BIC) \leq \#(AIC) \leq \#(R_a^2),$$

où $\#$ représente le nombre de paramètres du modèle linéaire.

Il est facile d'obtenir les quantités R_a^2 , AIC et BIC avec la procédure `glmselect` dans **SAS**. Le fichier `selection1_intro.sas` contient les programmes. La sortie qui suit provient des commandes :

```
proc glmselect data=multi.selection1_train;
model y=x x*x x*x*x /selection=none;
run;
```

Il s'agit du modèle cubique (d'ordre 3) en x .

Racine MSE	52.24190
Moyenne dépendante	-9.77202
R carré	0.7499
R car. ajust.	0.7421
AIC	897.09479
AICC	897.73309
SBC	805.51547

Résultats estimés des paramètres

Paramètre	DDL	Estimation	Erreur type	Valeur du test t	Pr > t
Intercept	1	20.973673	7.757440	2.70	0.0081
x	1	16.029603	4.477988	3.58	0.0005
x*x	1	-3.295585	0.668133	-4.93	<.0001
x*x*x	1	0.795758	0.258221	3.08	0.0027

TABLE 4.1 – Mesures d’adéquation du modèle linéaire et estimés de l’erreur

	EQM	$\widehat{\text{EQM}}_a$	R^2	R_a^2	AIC	BIC
1	3191	3674	0.65	0.65	1111	1119
2	3133	2879	0.73	0.72	1088	1099
3	2697	2620	0.75	0.74	1081	1094
4	2767	2582	0.75	0.74	1081	1097
5	2771	2581	0.75	0.74	1083	1102
6	2780	2578	0.75	0.74	1085	1106
7	2780	2577	0.75	0.74	1087	1111
8	2797	2531	0.76	0.74	1087	1113
9	2811	2528	0.76	0.73	1089	1118
10	2849	2519	0.76	0.73	1091	1122

Le tableau qui suit résume ces quantités pour tous les modèles de l’ordre 1 à l’ordre 10.

Les colonnes EQM et $\widehat{\text{EQM}}_a$ ont déjà été expliquées à la section précédente et ont été représentées graphiquement. On voit que l’erreur moyenne quadratique des données d’apprentissage, $\widehat{\text{EQM}}_a$, diminue toujours à mesure qu’on ajoute des variables (c’est-à-dire, qu’on augmente l’ordre du polynôme). Les critères AIC et BIC suggèrent le modèle cubique ($k = 3$), c’est-à-dire le vrai modèle. Le R^2 ajusté quant à lui choisit le modèle d’ordre 4 (qui est le deuxième meilleur selon le EQM). N’oubliez pas que ces trois critères sont calculés avec l’échantillon d’apprentissage ($n = 100$), mais en pénalisant l’ajout de variables. On est ainsi en mesure de contrecarrer le problème provenant du fait qu’on ne peut pas utiliser directement le $\widehat{\text{EQM}}_a$.

Le AIC et le BIC sont des critères très utilisés et très généraux. Ils sont disponibles dès qu’on utilise la méthode du maximum de vraisemblance est utilisée comme méthode d’estimation. Le R^2 ajusté a une portée plus limitée car il est spécialisé à la régression linéaire.

4.7 Division de l’échantillon et validation croisée

La deuxième grande approche après celle consistant à pénaliser le $\widehat{\text{EQM}}_a$ consiste à tenter d’estimer le EQM directement sans utiliser deux fois les mêmes données. Nous allons voir deux telles méthodes ici, la validation externe (division de l’échantillon) et la validation croisée (*cross-validation*).

Ces deux méthodes s’attaquent directement au problème qu’on ne peut utiliser (sans ajustement) les mêmes données qui ont servi à estimer les paramètres d’un modèle pour estimer sa performance. Pour ce faire, l’échantillon de départ est divisé en deux, ou plusieurs parties, qui vont jouer

des rôles différents.

4.7.1 Validation externe et division de l'échantillon

Cette idée est très simple. Nous avons un échantillon de taille n . Nous pouvons le diviser au hasard en deux parties de tailles respectives n_1 et n_2 ($n_1 + n_2 = n$),

- un échantillon d'apprentissage (*training*) de taille n_1 et
- un échantillon de validation (*test*) de taille n_2 .

L'échantillon d'apprentissage servira à estimer les paramètres du modèle. L'échantillon de validation servira à estimer la performance prédictive (par exemple estimer l'EQM) du modèle. Comme cet échantillon n'a pas servi à estimer le modèle lui-même, il est formé de « nouvelles » observations qui permettent d'évaluer d'une manière réaliste la performance du modèle. Comme il s'agit de nouvelles observations, on n'a pas à pénaliser la complexité du modèle et on peut directement utiliser le critère de performance choisi, par exemple, l'erreur quadratique moyenne, c'est-à-dire, la moyenne des erreurs au carré pour l'échantillon de validation. Cette quantité est une estimation valable de l'EQM de ce modèle. On peut faire la même chose pour tous les modèles en compétition et choisir celui qui a la meilleure performance sur l'échantillon de validation.

Cette approche possède plusieurs avantages. Elle est facile à planter. Elle est encore plus générale que les critères AIC et BIC. En effet, ces critères découlent de la méthode d'estimation du maximum de vraisemblance. Plusieurs autres types de modèles ne sont pas estimés par la méthode du maximum de vraisemblance (par exemple, les arbres, les forêts aléatoires, les réseaux de neurones, etc.) La performance de ces modèles peut toujours être estimée en divisant l'échantillon. Cette méthode peut donc servir à comparer des modèles de familles différentes. Par exemple, choisit-on un modèle de régression linéaire, une forêt aléatoire ou bien un réseau de neurones?

Cette approche possède tout de même un désavantage. Elle nécessite une grande taille d'échantillon au départ. En effet, comme on divise l'échantillon, on doit en avoir assez pour bien estimer les paramètres du modèle (l'échantillon d'apprentissage) et assez pour bien estimer sa performance (l'échantillon de validation).

La méthode consistant à diviser l'échantillon en deux (apprentissage et validation) afin de sélectionner un modèle est valide. Par contre, si on veut une estimation sans biais de la performance du modèle choisi (celui qui est le meilleur sur l'échantillon de validation), on ne peut pas utiliser directement la valeur observée de l'erreur de ce modèle sur l'échantillon de validation car elle risque de sous-évaluer l'erreur. En effet, supposons qu'on a 10 échantillons et qu'on ajuste 10 fois le même modèle séparément sur les 10 échantillons. Nous aurons alors 10 estimations différentes de l'erreur du modèle. Il est alors évident que de choisir la plus petite d'entre elles sous-estimerait la vraie erreur du modèle. C'est un peu ce qui se passe lorsqu'on choisit le modèle qui minimise

l'erreur sur l'échantillon de validation. Le modèle lui-même est un bon choix, mais l'estimation de son erreur risque d'être sous-évaluée.

Une manière d'avoir une estimation de l'erreur du modèle retenu consiste à diviser l'échantillon de départ en trois (plutôt que deux). Aux échantillons d'apprentissage et de validation, s'ajoute un échantillon « test ». Cet échantillon est laissé de côté durant tout le processus de sélection du modèle qui est effectué avec les deux premiers échantillons tel qu'expliqué plus haut. Une fois un modèle retenu (par exemple celui qui minimise l'erreur sur l'échantillon de validation), on peut alors évaluer sa performance sur l'échantillon test qui n'a pas encore été utilisé jusque là. L'estimation de l'erreur du modèle retenu sera ainsi valide. Il est évident que pour procéder ainsi, on doit avoir une très grande taille d'échantillon au départ.

4.7.2 Validation croisée

Si la taille d'échantillon n'est pas suffisante pour diviser l'échantillon en deux et procéder comme nous venons de l'expliquer, la validation croisée est une bonne alternative. Cette méthode permet d'imiter le processus de division de l'échantillon.

Voici les étapes à suivre pour faire une validation croisée à K groupes (*K -fold cross-validation*) :

1. Diviser l'échantillon au hasard en K parties P_1, P_2, \dots, P_K contenant toutes à peu près le même nombre d'observations.
2. Pour $j = 1$ à K ,
 - i. Enlever la partie j .
 - ii. Estimer les paramètres du modèle en utilisant les observations des $K - 1$ autres parties combinées.
 - iii. Calculer la mesure de performance (par exemple la somme du carré des erreurs) de ce modèle pour le groupe P_j .
3. Faire la somme des K estimations de performance pour obtenir une mesure de performance finale et répondre au besoin.

On recommande habituellement de prendre entre $K = 5$ et 10 groupes (le choix de 10 groupes est celui qui revient le plus souvent en pratique). Si on prend $K = 10$ groupes, alors chaque modèle est estimé avec 90% des données et on prédit ensuite le 10% restant. Comme on passe en boucle les 10 parties, chaque observation est prédictée une et une seule fois à la fin. Il est important de souligner que les groupes sont formés de façon aléatoire et donc que l'estimé que l'on obtient peut être très variable, surtout si la taille de l'échantillon d'apprentissage est petite. Il arrive également que le modèle ajusté sur un groupe ne puisse pas être utilisé pour prédire les observations mises de côté, notamment si des variables catégorielles sont présentes mais qu'une modalité n'est présente que dans un des groupes; ce problème se présente en pratique si certaines classes ont peu

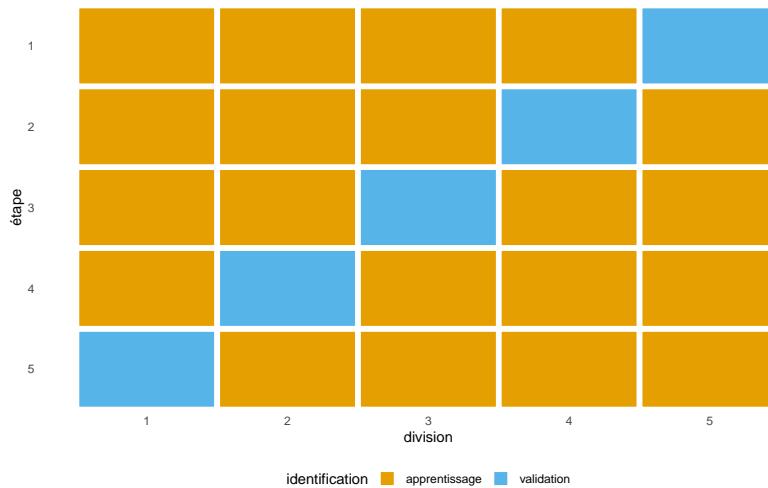


FIGURE 4.1 – Illustration de la validation croisée : on scinde l'échantillon d'apprentissage en cinq groupes (abscisse) et à chaque étape, une portion différente des données est mise de côté et ne sert que pour la validation.

d'observations. Un échantillonnage stratifié permet de pallier à cette lacune et de s'assurer d'une répartition plus homogène des variables catégorielles.

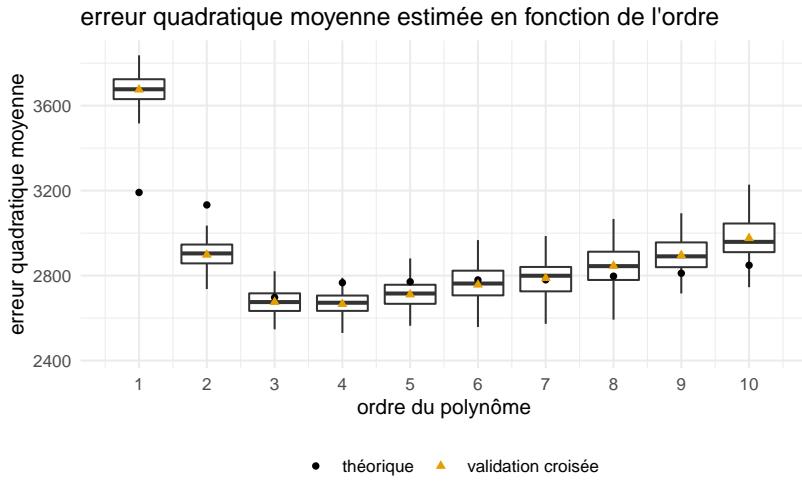
Le cas particulier $K = n$ (en anglais *leave-one-out cross validation*, ou LOOCV) consiste à enlever une seule observation, à estimer le modèle avec les $n - 1$ autres et à valider à l'aide de l'observation laissée de côté : on répète cette procédure pour chaque observation. Pour les modèles linéaires, il existe des formules explicites qui nous permettent d'éviter d'ajuster n régressions par moindre carrés. Cette forme de validation croisée tend à être trop optimiste.

Revenons à notre exemple où une seule variable explicative est disponible et où l'on cherche à déterminer un bon modèle polynomial. Le Tableau 4.2 est le même que le Tableau 4.1 mais avec une colonne en plus, la dernière, $VC(K = 10)$. Il s'agit des estimations de l'EQM obtenues avec la validation croisée à 10 groupes. Notez que si vous exécutez le programme, vous n'obtiendrez pas les mêmes valeurs car il y a un élément aléatoire dans ce processus. La colonne représente la moyenne de 100 réplications.

Le modèle cubique (ordre 3) est aussi choisi par la validation croisée, en moyenne (comme il l'était par le AIC et le BIC). Le graphe qui suit trace les valeurs de l'estimation par validation croisée (courbe de validation croisée) et aussi le EQM. On voit que l'estimation par validation croisée suit assez bien la forme du EQM (qu'il est supposé estimer). Les boîtes à moustache permettent d'apprécier la variabilité des estimés de l'erreur quadratique moyenne telles qu'estimée par validation croisée avec 10 groupes.

TABLE 4.2 – Mesures d'adéquation du modèle linéaire et estimés de l'erreur, incluant la validation croisée.

	EQM	$\widehat{\text{EQM}}_a$	R^2	R_a^2	AIC	BIC	$\text{VC}(K = 10)$
1	3191	3674	0.65	0.65	1111	1119	3675
2	3133	2879	0.73	0.72	1088	1099	2898
3	2697	2620	0.75	0.74	1081	1094	2676
4	2767	2582	0.75	0.74	1081	1097	2666
5	2771	2581	0.75	0.74	1083	1102	2711
6	2780	2578	0.75	0.74	1085	1106	2757
7	2780	2577	0.75	0.74	1087	1111	2788
8	2797	2531	0.76	0.74	1087	1113	2846
9	2811	2528	0.76	0.73	1089	1118	2896
10	2849	2519	0.76	0.73	1091	1122	2976



4.8 Cibler les clients pour l'envoi d'un catalogue

Nous allons présenter un exemple classique de commercialisation de bases de données qui nous servira à illustrer la sélection de modèles, la régression logistique et la gestion de données manquantes.

Le contexte est le suivant : une entreprise possède une grande base de données client. Elle désire envoyer un catalogue à ses clients mais souhaite maximiser les revenus d'une telle initiative. Il

est évidemment possible d'envoyer le catalogue à tous les clients mais ce n'est possiblement pas optimal. La stratégie envisagée est la suivante :

1. Envoyer le catalogue à un échantillon de clients et attendre les réponses. Le coût de l'envoi d'un catalogue est de 10\$.
2. Construire un modèle avec cet échantillon afin de décider à quels clients (parmi les autres) le catalogue devrait être envoyé, afin de maximiser les revenus.

Plus précisément, on s'intéresse aux clients de 18 ans et plus qui ont au moins un an d'historique avec l'entreprise et qui ont effectué au moins un achat au cours de la dernière année. Dans un premier lieu, on a envoyé le catalogue à un échantillon de 1000 clients. Un modèle sera construit avec ces 1000 clients afin de cibler lesquels des clients restants seront choisis pour recevoir le catalogue.

Pour les 1000 clients de l'échantillon d'apprentissage, les deux variables cibles suivantes sont disponibles :

- *yachat*, une variable binaire qui indique si le client a acheté quelque chose dans le catalogue égale à 1 si oui et 0 sinon.
- *ymontant*, le montant de l'achat si le client a acheté quelque chose.

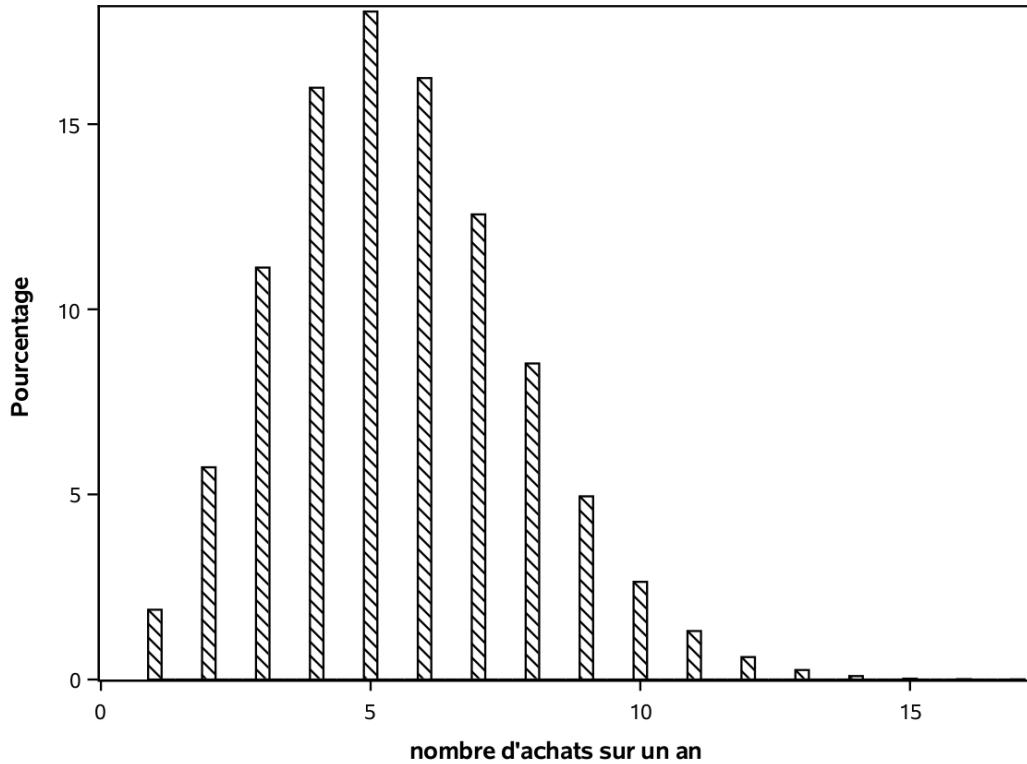
Les 10 variables suivantes sont disponibles pour tous les clients et serviront de variables explicatives pour les deux variables cibles. Il s'agit de :

- *x1* : sexe de l'individu, soit homme (0) ou femme (1);
- *x2* : l'âge (en année);
- *x3* : variable catégorielle indiquant le revenu, soit moins de 35 000\$ (1), entre 35 000\$ et 75 000\$ (2) ou plus de 75 000\$ (3);
- *x4* : variable catégorielle indiquant la région où habite le client (de 1 à 5);
- *x5* : conjoint : le client a-t-il un conjoint (0=non, 1=oui);
- *x6* : nombre d'année depuis que le client est avec la compagnie;
- *x7* : nombre de semaines depuis le dernier achat;
- *x8* : montant (en dollars) du dernier achat;
- *x9* : montant total (en dollars) dépensé depuis un an;
- *x10* : nombre d'achats différents depuis un an.

Les données se trouvent dans le fichier DBM.sas7bdat. Voici d'abord des statistiques descriptives pour l'échantillon d'apprentissage.

sexe		revenu		region	
x1	Fréquence	x3	Fréquence	x4	Fréquence
0	534	1	397	1	216
1	466	2	337	2	185
conjoint		x5 Fréquence		x4 Fréquence	
0	575	3	266	3	216
1	425	4		4	191
		5		5	192

Il y a 46,6% de femmes parmi les 1000 clients de l'échantillon. De plus, 39,7% ont un revenu de moins de 35 000\$, 33,7% sont entre 35 000\$ et 75 000\$ et 26,6% ont plus de 75 000\$. 42,5% de ces clients qui ont un conjoint.

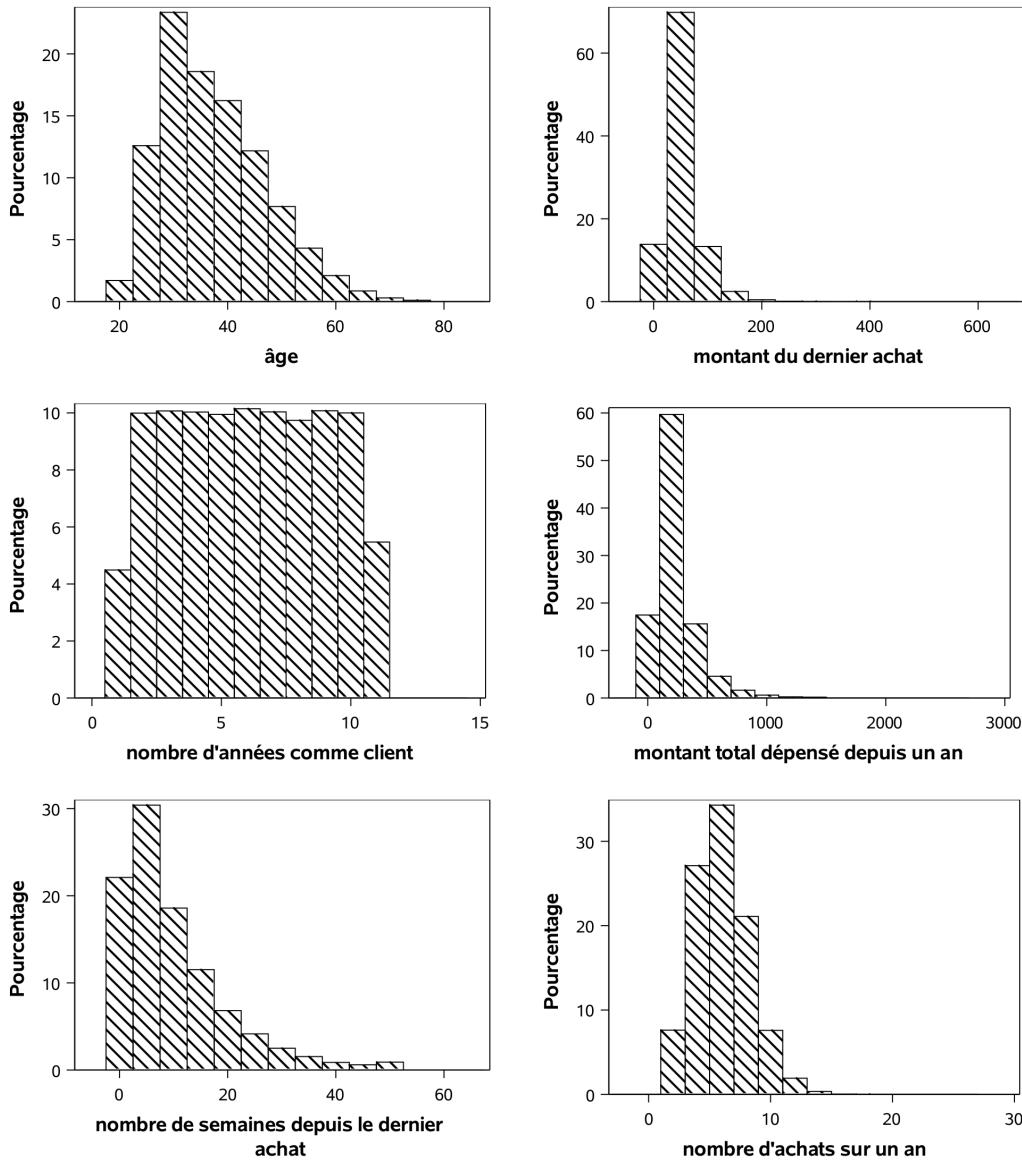


Le nombre d'achats différents depuis un an par ces clients varie entre 1 et 14. Un peu plus de la moitié (51,4%) ont fait cinq achats ou moins. Parmi les 1000 clients de l'échantillon d'apprentis-

sage, 210 ont acheté quelque chose dans le catalogue. La variable yachat sera l'une des variables que nous allons chercher à modéliser en vue d'obtenir des prédictions.

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
x2	âge	1000	37.06	9.27	20.00	70.00
x6	nombre d'années comme client	1000	6.01	2.92	1.00	11.00
x7	nombre de semaines depuis le dernier achat	1000	9.97	9.34	1.00	52.00
x8	montant du dernier achat	1000	48.41	28.27	20.00	252.00
x9	montant total dépensé depuis un an	1000	229.27	173.97	22.00	1407.00
x10	nombre d'achats sur un an	1000	5.64	2.31	1.00	14.00
ymontant	montant de l'achat (catalogue)	210	67.29	13.24	25.00	109.00

L'âge des 1000 clients de l'échantillon d'apprentissage varie entre 20 et 70 avec une moyenne de 37,1 ans. En moyenne, ces clients ont acheté pour 229,30\$ depuis un an. Le dernier achat de ces clients remonte, en moyenne, à 10 semaines. Nous chercherons également à modéliser la variable ymontant. Seuls 210 clients ont acheté quelque chose dans le catalogue et les statistiques rapportées correspondent seulement à ces derniers, car la variable ymontant est manquante si le client n'a rien acheté dans le catalogue. On pourrait également remplacer ces valeurs par des zéros et les modéliser, mais nous aborderons cet aspect ultérieurement. Les clients qui ont acheté quelque chose ont dépensé en moyenne 67,3\$, et au minimum 25\$. Les histogrammes de quelques unes de ces variables permet de mieux visualiser la répartition des observations.



Il y a plusieurs façons d'utiliser l'échantillon d'apprentissage afin de mieux cibler les clients à qui envoyer le catalogue et maximiser les revenus. En voici quelques unes.

- On pourrait développer un modèle afin d'estimer la probabilité qu'un client achète quelque chose si on lui envoie un catalogue. Plus précisément, on peut développer un modèle pour $\text{Pr}(y_{achat} = 1)$. Comme la variable y_{achat} est binaire, un modèle possible est la régression logistique, que nous décrirons au chapitre suivant. Ainsi, en appliquant le modèle aux

100 000 clients restant, on pourra cibler les clients susceptibles d'acheter (ceux avec une probabilité élevée).

- b) Une autre façon serait de tenter de prévoir le montant d'argent dépensé. Nous venons de voir la distribution de la variable $y_{montant}$. Il y a deux situations, ceux qui ont acheté et ceux qui n'ont pas achetés. En conditionnant sur le fait d'avoir acheté quelque chose, il est possible de décomposer le problème de la manière suivante :

$$\begin{aligned} E(y_{montant}) &= E(y_{montant} | y_{achat} = 1)P(y_{achat} = 1) \\ &\quad + E(y_{montant} | y_{achat} = 0)P(y_{achat} = 0) \\ &= E(y_{montant} | y_{achat} = 1)P(y_{achat} = 1) \end{aligned}$$

En mots, la moyenne du montant dépensé est égale à la moyenne du montant dépensé étant donné qu'il y a eu achat, fois la probabilité qu'il ait eu achat.

On peut donc estimer $E(y_{montant} | y_{achat} = 1)$ et $P(y_{achat} = 1)$, pour ensuite les combiner et avoir une estimation de $E(y_{montant})$. Le développement du modèle pour $E(y_{montant} | y_{achat} = 1)$ peut se faire avec la régression linéaire, en utilisant seulement les clients qui ont acheté dans l'échantillon d'apprentissage, car $y_{montant}$ est une variable continue dans ce cas. Le développement du modèle pour $P(y_{achat} = 1)$ peut se faire avec la régression logistique, tel que mentionné plus haut, en utilisant tous les 1000 clients de l'échantillon d'apprentissage. En fait, nous verrons plus loin qu'il est possible d'estimer conjointement les deux modèles avec un modèle Tobit. En appliquant le modèle aux 100 000 clients restants, on pourra cibler les clients qui risquent de dépenser un assez grand montant.

Comme nous n'avons pas encore vu la régression logistique, nous allons nous limiter à illustrer les méthodes qui restent à voir dans ce chapitre avec la régression linéaire en cherchant à développer un modèle pour $E(y_{montant} | y_{achat} = 1)$, le montant d'argent dépensé par les clients qui ont acheté quelque chose.

La base de donnée contient deux variables explicatives catégorielles. Il s'agit de `revenu` (`x3`) et `région` (`x4`). Il faut coder d'une manière appropriée afin de pouvoir les incorporer dans les modèles. La manière habituelle est de créer des variables indicatrices (binaires) qui indiquent si la variable prend ou non une valeur particulière, dans **SAS** avec l'option `class`. En général, si une variable catégorielle possède K valeurs possibles, il est suffisant de créer $K - 1$ indicatrices, en laissant une modalité comme référence. Par exemple, pour `x3`, nous allons créer deux variables,

- `x31` : variable binaire égale à 1 si `x3` égale 1 et 0 sinon,
- `x32` : variable binaire égale à 1 si `x3` égale 2 et 0 sinon.

Ainsi, la valeur 3 est celle de référence. Ces deux indicatrices sont suffisantes pour récupérer toute l'information comme le démontre le tableau 4.3.

TABLE 4.4 – Nombres de modèles en fonction du nombre de paramètres p .

p	nombre de paramètres
5	32
10	1024
15	32768
20	1048576
25	33554432
30	1073741824

TABLE 4.3: Valeur des indicateurs en fonction du niveau de la variable catégorielle

x3	x31	x32
1	1	0
2	0	1
3	0	0

4.9 Recherche automatique du meilleur modèle

Lorsque nous voulons comparer un petit nombre de modèles, il est relativement aisé d'obtenir les critères (AIC, BIC ou autre) pour tous les modèles et de choisir le meilleur. C'était le cas dans l'exemple du choix de l'ordre du polynôme où il y avait seulement 10 modèles en compétitions. Mais lorsqu'il y a plusieurs variables en jeu, le nombre de modèles potentiel augmente très rapidement.

En fait, supposons qu'on a p variables distinctes disponibles. Avant même de considérer les transformations des variables et les interactions entre elles, il y a déjà modèles possibles. En effet, chaque variable est soit incluse ou pas (deux possibilités) et donc il y a $2^p = 2 \times 2 \times \dots \times 2$ (p fois) modèles en tout à considérer. Ce nombre augmente très rapidement comme en témoigne le tableau 4.4.

Ainsi, si le nombre de variables est restreint, il est possible de comparer tous les modèles potentiels et de choisir le meilleur (selon un critère). Il existe même des algorithmes très efficaces qui permettent de trouver le meilleur modèle sans devoir examiner tous les modèles possibles. Le nombre de variables qu'il est possible d'avoir dépend de la puissance de calcul et augmente d'année en année. Par contre, dans plusieurs applications, il ne sera pas possible de comparer tous les modèles et il faudra effectuer une recherche limitée. Faire une recherche exhaustive parmi tous

les modèles possibles s'appelle sélection de tous les sous-ensembles (*best subsets*). La procédure `reg` de **SAS** permet de faire cela pour la régression linéaire.

On veut trouver un bon modèle pour prévoir la valeur de *ymontant* des clients qui ont acheté quelque chose. On a vu qu'il y a 210 clients qui ont acheté dans l'échantillon d'apprentissage. Nous allons chercher à développer un « bon » modèle avec ces 210 clients. Dans ce premier exemple, nous allons seulement utiliser les 10 variables explicatives de base (14 variables avec les indicatrices). Le code suivant montre comment faire une sélection de variables selon le critère du R^2 et demande à **SAS** de présenter le modèle à k variables ($k = 1, \dots, 14$) qui a le plus grand R^2 ; voir `selection2_methodes.sas` pour plus de détails.

Pour un nombre de variables fixé, le meilleur modèle selon le R^2 est aussi le meilleur selon les critères d'information AIC et BIC, pour ce nombre fixé de variables. Pour vous convaincre de cette affirmation, fixons le nombre de variables et restreignons-nous seulement aux modèles avec ce nombre de variables. Comme $= 1 - \text{SCE}/\text{SCT}$ et que SCT est une constante indépendante du modèle, le modèle avec le plus grand coefficient de détermination, R^2 , est aussi celui avec la plus petite somme du carré des erreurs (SCE). Comme $\text{AIC} = n(\ln(\text{SCE}/n)) + 2p$, ce sera aussi celui avec le plus petit AIC car la pénalité $2p$ est la même si on fixe le nombre de variables; la même remarque est valide pour le BIC.

Ainsi, pour trouver le meilleur modèle globalement (sans fixer le nombre de variables), il suffit de trouver le modèle à k variables explicatives ayant le coefficient de détermination le plus élevé pour tous les nombres de variables fixés et d'ensuite de trouver celui qui minimise le AIC (ou le BIC) parmi ces modèles. Cette astuce est utile dans la mesure où **SAS** ne permet pas de faire cette même recherche avec les critères d'information. Ainsi, le modèle linéaire simple qui a le plus grand R^2 est celui qui inclut le conjoint (x5). Le meilleur modèle (selon le R^2) parmi tous les modèles avec deux variables est celui avec x5 et x6.

Etape	Effet saisi	Nombre d'effets dans	AIC	SBC	ASE	Validation - ASE
0	Intercept	1	1297.9002	1089.2473	174.4136	184.3078
1	x5	2	1193.2487	987.9429	104.9585	97.2350
2	x6	3	1089.4520	887.4934	63.4195	56.0533
3	x3_1	4	971.4210	772.8094	35.8089	35.1709
4	x10	5	928.1137	732.8493	28.8597	33.2971
5	x1	6	906.4859	714.5686	25.7886	29.5169
6	x7	7	885.3848	696.8145	23.1022	25.1729
7	x8	8	876.9471	691.7240*	21.9820	23.8341*
8	x4_4	9	875.1003	693.2242	21.5830	24.2716
9	x2	10	872.3748	693.8459	21.1028	24.4139
10	x9	11	872.1462*	696.9644	20.8800	24.7243
11	x4_3	12	872.8715	701.0368	20.7536	24.9117
12	x4_1	13	874.2299	705.7423	20.6903	24.9744
13	x4_2	14	874.2835	709.1431	20.4994	25.4998
14	x3_2	15	875.8411	714.0477	20.4563	25.6859

* Valeur optimale du critère

Dans **SAS**, seule la procédure `reg` permet de faire cette recherche exhaustive et nous garantit de recouvrir le modèle avec le plus grand R^2 . Un algorithme par séparation et évaluation permet d'effectuer cette recherche de manière efficace sans essayer tous les candidats pour ces sous-ensembles. Dans l'exemple, on voit que le modèle avec les variables `x1 x2 x31 x44 x5 x6 x7 x8 x9` et `x10` est celui qui minimise le AIC globalement ($AIC = 660.15$). Le modèle choisi par le BIC contient seulement sept variables explicatives (plutôt que 10), soit `x1 x31 x5 x6 x7 x8 x10`.

Nous avons seulement inclus les variables de base pour ce premier essai. Il est possible qu'ajouter des variables supplémentaires améliore la performance du modèle. Pour cet exemple, nous allons considérer les variables suivantes :

- les variables continues au carré, comme `age2`.
- toutes les interactions d'ordre deux entre les variables de base, comme `sexe · age`.

Si on suppose que la vrai moyenne de la variable réponse `ymontant` est lisse, le modèle ajusté précédemment capture l'approximation de degré 1 (série de Taylor) du vrai modèle et le modèle avec termes quadratiques (incluant les interactions) l'approximation de degré 2. Aux variables de base

(10 variables explicatives, mais 14 avec les indicatrices pour les variables catégorielles), s'ajoutent ainsi 90 autres variables. Il y a donc 104 variables explicatives potentielles si on inclut les interactions et les termes quadratiques. Notez qu'il y a des interactions entre chacune des variables indicatrices et chacune des autres variables, mais il ne sert à rien de calculer une interaction entre deux indicatrices d'une même variable (car une telle variable est zéro pour tous les individus). De même, il ne sert à rien de calculer le carré d'une variable binaire.

Dans la mesure où on aura presque un ratio d'un paramètre pour deux observations, ajuster un modèle avec toutes les variables explicatives est profondément stupide ici. Pensez à la taille de votre échantillon comme à un budget et aux paramètres comme à un nombre d'items : plus vous achetez d'items, moins votre budget est élevé pour chacun et leur qualité en pâtit. Le modèle à 104 variables servira uniquement à illustrer le surajustement. Réalistement, un modèle avec plus d'une vingtaine de variables ici serait difficilement estimable de manière fiable et l'inclusion d'interactions et de termes quadratiques sert surtout à augmenter la flexibilité et les possibilités lors de la sélection de variables.

Lancer une sélection exhaustive de tous les sous-modèles avec 104 variables risque de prendre un temps énorme. Que faire alors ? Il y a plusieurs possibilités. Nous pourrions faire une recherche limitée avec les méthodes que nous allons voir à partir de la section suivante. Nous pourrions aussi combiner les deux approches. Supposons que notre ordinateur permet de faire une recherche exhaustive de tous les sous-modèles avec 40 variables. Nous pourrions alors commencer avec une recherche limitée pour trouver un sous-ensemble de 40 « bonnes » variables et faire une recherche exhaustive, mais en se restringant à ces 40 variables.

4.10 Méthodes classiques de sélection

Les méthodes de sélection ascendante, descendante et séquentielle sont des algorithmes gloutons qui permettent de choisir des variables. Elles ont été développées à une époque où la puissance de calcul était bien moindre, et où il était impossible de faire une recherche exhaustive des sous-modèles. Avec l'approche classique, ces méthodes font une recherche séquentielle guidée parmi un nombre limité de modèles, à l'aide des valeurs- p du test- t pour la significativité des paramètres individuels du modèle avec p prédicteurs potentiels X_1, \dots, X_p . Les procédures `glmselect` et `reg` permettent une sélection de modèle avec une approche séquentielle, ascendante ou descendante.

4.10.1 Sélection ascendante

L'idée de la sélection ascendante est de tester l'ajout de chaque variable individuellement et d'ajouter celle qui est la plus significative selon le test- t si elle a une valeur- p assez petite.

- *Initialisation* : le modèle linéaire de départ est celui qui n'inclut que l'ordonnée à l'origine, $Y = \beta_0 + \varepsilon$, où ε est une erreur centrée.

- *Critère d'entrée*: c , valeur- p minimale à partir de laquelle une variable peut être incluse dans le modèle
- *Boucle* soit $X_{(1)}, \dots, X_{(k)}$, les variables explicatives à l'étape $k < p$.
 - pour chaque j ($j = \{1, \dots, p\} \setminus \{(1), \dots, (k)\}$), on ajuste tour à tour le modèle $Y = \beta_0 + \sum_{i=1}^k \beta_i X_{(i)} + \beta_{k+1} X_j$ et on calcule la valeur- p du test- t pour les hypothèses $\mathcal{H}_0 : \beta_{k+1} = 0$ contre l'alternative bilatérale $\mathcal{H}_1 : \beta_{k+1} \neq 0$.
 - Soit p_{\min} la plus petite des $p - k$ valeurs- p qui correspond à $X_{(k+1)}$, disons.
 - si $p_{\min} < c$, continuer la procédure.
 - si $p_{\min} \geq c$, retourner le modèle $Y = \beta_0 + \sum_{i=1}^k \beta_i X_{(i)} + \varepsilon$.

On continue ainsi à ajouter des variables jusqu'à ce que le critère d'entrée ne soit pas satisfait. Si on se rend jusqu'au bout, on va terminer avec le modèle complet qui contient toutes les variables.

4.10.2 Sélection descendante

- *Initialisation* : le modèle linéaire de départ est celui qui inclut toutes les variables explicatives, $Y = \beta_0 + \sum_{j=1}^p \beta_j X_{(j)} + \varepsilon$, où ε est une erreur centrée.
- *Critère de sortie* : c , valeur- p maximale à partir de laquelle une variable peut être exclue du modèle
- *Boucle* soit $X_{(1)}, \dots, X_{(p-k)}$, les variables explicatives présentes dans le modèle à l'étape $k < p$.
 - pour chaque j ($j = 1, \dots, p - k$), on calcule la valeur- p du test- t $\mathcal{H}_0 : \beta_j = 0$ contre l'alternative bilatérale $\mathcal{H}_1 : \beta_j \neq 0$.
 - si toutes ces valeurs sont inférieures à c , on retourne le modèle $Y = \beta_0 + \sum_{j=1}^{p-k} \beta_j X_{(j)}$.
 - sinon, on enlève la variable qui a la plus grande valeur- p (disons $X_{(p-k)}$), on réajuste le modèle sans cette variable et on recommence la procédure.

L'idée est l'inverse de la méthode ascendante. On va tester le retrait de chaque variable individuellement et retirer celle qui est la moins significative, si sa valeur- p est assez grande. Si la procédure se termine après p itérations, aucune variable explicative n'est retenue.

4.10.3 Méthode séquentielle

Il s'agit d'une méthode hybride entre ascendante et descendante. On sélectionne un critère d'entrée et de sortie pour chacune des deux et on débute la recherche à partir du modèle ne contenant que l'ordonnée à l'origine. À chaque étape, on fait une étape ascendante suivie de une (ou plusieurs) étapes descendantes. On continue ainsi tant que le modèle retourné par l'algorithme n'est pas identique à celui de l'étape précédente. Le dernier modèle est celui retenu.

Avec la méthode séquentielle, une fois qu'on entre une variable (étape ascendante), on fait autant d'étapes descendante afin de retirer toutes les variables qui satisfont le critère de sortie (il peut ne

pas y en avoir). Une fois cela effectué, on refait une étape ascendante pour voir si on peut ajouter une nouvelle variable.

Remarques sur ces méthodes : avec la méthode ascendante, une fois qu'une variable est dans le modèle, elle y reste. Avec la méthode descendante, une fois qu'une variable est sortie du modèle, elle ne peut plus y entrer. Avec la méthode séquentielle, une variable peut entrer dans le modèle et sortir plus tard dans le processus. Par conséquent, parmi les trois, la méthode séquentielle est généralement préférable aux méthodes ascendante et descendante, car elle inspecte potentiellement un plus grand nombre de modèles.

On peut soi-même spécifier les critères d'entrée et de sortie. Plus le critère d'entrée est élevé, plus il y aura de variables dans le modèle final. De même, plus le critère de sortie est élevé, plus il y aura de variables dans le modèle. Il y également moyen avec `glmselect` de spécifier le nombre d'étapes de chaque procédure.

Utilisons la méthode de sélection séquentielle classique avec des critères d'entrée et de sortie de 0.15 et les 104 variables. La sortie **SAS** est assez volumineuse car elle retrace toutes étapes de la sélection séquentielle. L'historique montre qu'à l'étape 1, la variable d'interaction `x5_x6` a été ajoutée, suivie de `x31_x10` à l'étape 2. Un peu plus loin, à l'étape 6, `x5_x6` est retirée et ainsi de suite. Il y a eu 40 étapes en tout et, à la fin, il reste 22 variables (parmi les 104) dans le modèle final. Le R^2 du modèle final est 0,966.

Synthèse des sélections Stepwise						
Etape	Effet saisi	Effet supprimé	Nombre d'effets dans	ASE	Validation - ASE	Valeur F Pr > F
0	Intercept		1	174.4136	184.3078	0.00 1.0000
1	x5*x6		2	98.9464	89.4494	158.64 <.0001
2	x10*x3_1		3	56.4947	64.2056	155.55 <.0001
3	x6		4	35.2457	43.8005	124.19 <.0001
4	x5		5	27.9394	33.2699	53.61 <.0001
5		x5*x6	4	27.9446	33.2492	0.04 0.8462
6	x8*x3_1		5	21.0590	24.7610	67.03 <.0001
7	x7*x8		6	16.8185	20.0424	51.44 <.0001
8	x1*x6		7	12.5395	14.5946	69.27 <.0001
9	x1*x10		8	11.6118	13.7267	16.14 <.0001
10	x5*x10		9	10.6041	13.0942	19.10 <.0001
11	x6*x6		10	9.6570	12.4121	19.61 <.0001
12		x6	9	9.6576	12.4214	0.01 0.9136
13	x10*x10		10	9.2746	11.6676	8.26 0.0045
14	x2*x3_1		11	8.4046	11.3956	20.60 <.0001

Synthèse des sélections Stepwise						
Etape	Effet saisi	Effet supprimé	Nombre d'effets dans	ASE	Validation - ASE	Valeur F Pr > F
33	x9*x4_4		24	5.9520	11.9236	2.17 0.1425
34	x3_2		25	5.8823	12.0426	2.19 0.1402
35		x8	24	5.9270	12.3414	1.41 0.2372
36		x1*x3_2	23	5.9740	12.2359	1.47 0.2261
37		x1*x4_2	22	6.0356	12.1191	1.93 0.1664
38	x8*x4_2		23	5.9386	12.2249	3.05 0.0822
39		x6*x4_2	22	5.9824	12.1808	1.38 0.2420
40	x6*x4_4		23	5.9142	12.2471	2.16 0.1437

On voit bien que toutes les valeurs-*p* (qui ne sont pas valides à cause de la sélection de modèles) sont toutes inférieures à 0.15. Le choix de 0.15 comme critère d'entrée et de sortie est complète-

ment arbitraire. Il est fort possible que d'autres valeurs donnent de meilleurs résultats, mais il n'est pas évident de les choisir.

Une façon de contourner le problème de devoir spécifier les critères d'entrée et de sortie est de procéder en deux étapes. Supposons que notre ordinateur permet de faire une recherche exhaustive de tous les sous-modèles avec près de 60 variables. L'idée est alors de passer de 104 à un sous-ensemble d'environ 60 variables, avec une sélection séquentielle gloutonne, et d'ensuite utiliser une recherche exhaustive avec ce sous-ensemble de variables. Plus précisément :

- 1) On fait une sélection séquentielle classique avec des valeurs élevées pour les critères d'entrée et de sortie afin que le modèle retenu contienne le nombre voulu de variables (par exemple, 60).
- 2) En utilisant seulement ce sous-ensemble de variables, on choisit le meilleur modèle selon le AIC ou le BIC en faisant une recherche exhaustive de tous les sous-modèles.

En fixant, les critères d'entrée et de sortie à 0,6 pour la recherche séquentielle, le modèle retenu aura 56 variables. Il est possible de faire une recherche exhaustive avec 56 variables sur un ordinateur portable avec **SAS**. Le AIC est mène à un modèle avec 38 de ces 56 variables, ce qui est probablement trop. Le BIC est quant à lui beaucoup plus parcimonieux et choisit 15 de ces variables pour le modèle final. Nous verrons à la section suivante qu'il est possible de faire une recherche séquentielle en utilisant d'autres critères que la valeur-*p* du test-*t* pour faire ajouter ou enlever des variables.

4.11 Recherche séquentielle automatique limitée

L'idée de la procédure séquentielle classique est d'inclure ou d'exclure une variable à la fois sur la base des valeurs-*p*. La procédure `glmselect` permet de faire une sélection séquentielle en utilisant d'autres critères, comme le AIC ou le BIC. Cette procédure permet de contrôler très finement le processus de sélection de variables. Le code qui suit fait une recherche séquentielle avec les particularités suivantes. À chaque étape ascendante de la procédure séquentielle, c'est la variable qui améliore le plus le AIC (`select=aic`) qui est entrée. De plus, à chaque étape descendante de la procédure séquentielle, c'est la (ou les) variable(s) qui détériore(nt) le plus le AIC qui est (sont) retirée(s). À la toute fin du processus, c'est le modèle qui a le meilleur BIC (`choose=BIC`) qui est retenu.

La procédure `glmselect` permet la déclarations de variables catégorielles (`class`) ; on évite ainsi de créer les variables binaires une par une. Voici les commandes pour faire une procédure séquentielle avec des critères plus généreux (`entrée=sortie=0.6`). Si le nombre de variables change en sortie, nous aurons ici 56 variables qui seront ensuite utilisées avec une recherche exhaustive : pour ce faire, on reprend la sortie, mais cette fois on choisit le modèle par la suite qui a le plus petit BIC (SBC) ou AIC.

```

proc glmselect data=ymontant outdesign=glmselectoutput;
partition role=train(train="1" validate="0");
class x3(param=ref split) x4(param=ref split);
model ymontant=x1|x2|x3|x4|x5|x6|x7|x8|x9|x10 @2
x2*x2 x6*x6 x7*x7 x8*x8 x9*x9 x10*x10 /
selection=stepwise(slentry=0.6 slstay=0.6 select=60) hier=none;
run;

proc glmselect data=glmselectoutput;
model ymontant= &_GLSMOD /
selection=backward(stop=1 choose=sbc) hier=none;
ods output GLMSelect.Summary.SelectionSummary=historiqueSelection;
run;

```

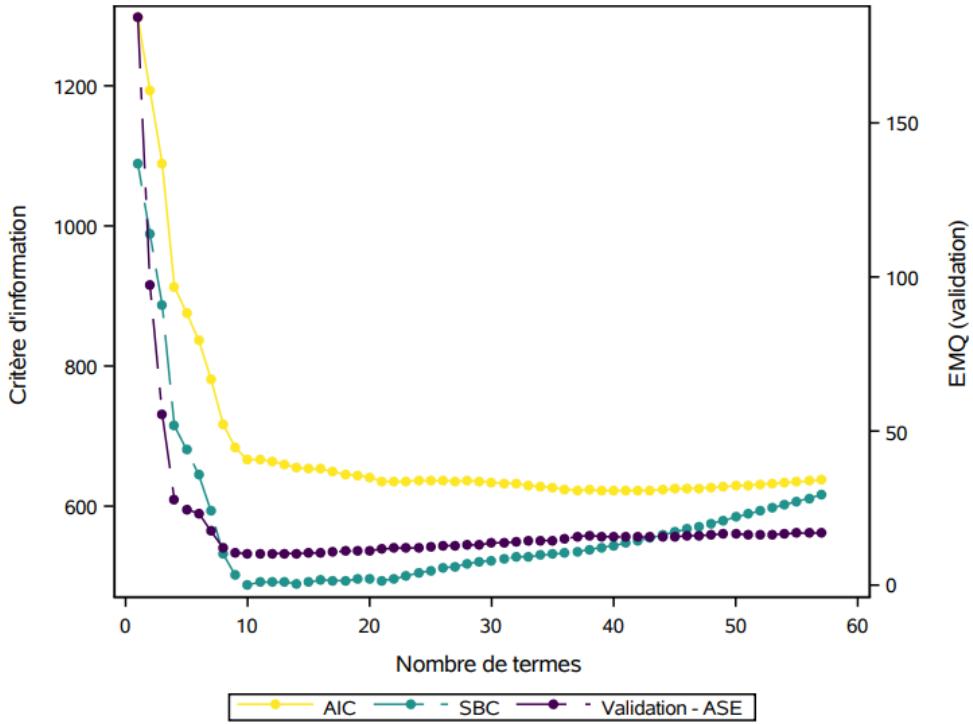
À l'étape 1, la variable x5_x6 est ajoutée au modèle de base car c'est celle qui fait diminuer le plus le AIC. À l'étape 2, la variable x31_x10 est ajoutée, À l'étape 6, la variable x5_x6 est retirée car cela fait baisser le AIC. Notez que le AIC décroît toujours d'une étape à l'autre. SAS garde aussi la trace du BIC car le modèle final sera choisi selon ce critère. Finalement le processus séquentiel se termine à l'étape 40, car il n'y a plus moyen de faire diminuer le AIC. Le modèle final retenu est celui de l'étape 18, car c'est celui qui a le BIC le plus petit parmi tous ces modèles (BIC = 484.22).

Voici différentes statistiques ainsi que les estimations des paramètres de ce modèle qui contient 10 variables.

Racine MSE	2.87859
Moyenne dépendante	67.28571
R carré	0.9548
R car. ajust.	0.9527
AIC	665.82101
AICC	667.15435
SBC	487.29209
ASE (Apprentissage)	7.89172
ASE (Validation)	10.18009

Paramètre	DDL	Paramètres estimés			Erreur type	Valeur du test t
		Estimation	 	 		
Intercept	1	59.669036	0.734137			81.28
x2*x3_1	1	-0.107931	0.024618			-4.38
x5	1	8.986221	1.094331			8.21
x1*x6	1	0.850410	0.069898			12.17
x7	1	0.197533	0.018224			10.84
x8	1	0.041456	0.007436			5.58
x8*x3_1	1	-0.133053	0.010074			-13.21
x5*x10	1	1.405927	0.212720			6.61
x6*x6	1	0.146950	0.006532			22.50
x10*x10	1	-0.115500	0.009314			-12.40

On peut voir sur le graphique @ref(fig :fig2p2_fig17) l'historique des valeurs de AIC et BIC à mesure qu'on diminue le nombre de variables dans le modèle : les mêmes variables sont enlevées à chaque étape, mais la valeur optimale du critère est différente pour la sélection finale. Sur l'axe des abscisses, j'ai ajouté l'erreur quadratique moyenne pour l'échantillon de validation contenant les 23 389 clients (sur 100 000) qui achèteraient si leur envoyait une copie du catalogue. Cet exemple n'est pas réaliste puisqu'on regarde la solution, mais il permet de nous comparer et de voir à quel point ici le critère d'information bayésien suit la même tendance que l'erreur moyenne quadratique de validation.



4.12 Méthodes de régression avec régularisation

Une façon d'éviter le surajustement est d'ajouter une pénalité sur les coefficients : ce faisant, on introduit un biais dans nos estimés, mais dans l'espoir de réduire leur variabilité et ainsi d'obtenir une meilleure erreur quadratique moyenne.

Les estimateurs des moindres carrés ordinaires pour la régression linéaire représentent la combinaison qui minimise la somme du carré des erreurs,

$$SCE = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2.$$

On peut ajouter à cette fonction objective SCE un terme additionnel de pénalité qui va contraindre les paramètres à ne pas être trop grand. On considère une pénalité additionnelle pour la valeur absolue des coefficients,

$$q_1(\lambda) = \lambda \sum_{j=1}^p |\beta_j|.$$

Pour chaque valeur de λ donnée, on obtiendra une solution différente pour les estimés car on minimisera désormais $SCE + q_1(\lambda)$. On ne pénalise pas l'ordonnée à l'origine β_0 , parce que ce

coefficient sert à recentrer les observations et a une signification particulière. Si on standardise les données, de manière à ce que leur moyenne empirique soit zéro et leur écart-type un, alors $\hat{\beta}_0 = \bar{y}$.

L'avantage des moindres carrés est que les valeurs ajustées et les prédictions ne changent pas si on fait une transformation affine (de type $Z = aX + b$). Peu importe le choix d'unités (par exemple, exprimer une distance en centimètres plutôt qu'en mètres, ou la température en Fahrenheit plutôt qu'en Celcius), on obtient le même modèle. En revanche, une fois qu'on introduit un terme de pénalité, notre solution dépendra de l'unité de mesure, d'où l'importance d'utiliser les données centrées et réduites pour que la solution reste la même.

La pénalité $q_1(\lambda)$ a un rôle particulier parce qu'elle a deux effets : elle réduit la taille des paramètres, mais elle force également certains paramètres très proches de zéro à être exactement égaux à zéro, ce qui fait que la régression pénalité agit également comme outil de sélection de variables. Des algorithmes efficaces permettent de trouver la solution du problème d'optimisation

$$\min_{\beta} \{SCE + q_1(\lambda)\} = \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

laquelle est appelée LASSO. La Figure 4.2 montre la fonction objective dans le cas où on a deux paramètres, β_1 et β_2 . La solution des moindres carrés ordinaires, qui minimisent l'erreur quadratique moyenne, est au centre des ellipses de contour et correspond à la solution du modèle avec $\lambda = 0$. À mesure que l'on augmente la pénalité λ , les coefficients rétrécissent vers $(0, 0)$. On peut interpréter la pénalité l_1 comme une contrepartie budgétaire : les coefficients estimés pour une valeur de λ donnée sont ceux qui minimisent la somme du carré des erreurs, mais doivent être à l'intérieur d'un budget alloué (losange). La forme de la région fait en sorte que la solution, qui se trouve sur la bordure du losange, intervient dans un coin avec certaines coordonnées nulles.

Plusieurs variantes existent dans la littérature qui généralisent le modèle à des contextes plus compliqués. Le choix des variables à inclure dans la sélection dépend du choix de la pénalité λ , qui est règle générale estimée par validation croisée à 5 ou 10 groupes.

4.13 Moyenne de modèles

Une idée importante et moderne en statistique est qu'il est souvent préférable de combiner plusieurs modèles plutôt que d'en choisir un seul. La technique des forêts aléatoires (*random forests*) est une des meilleures techniques de prédiction disponibles de nos jours. Elle est basée sur cette idée, en combinant plusieurs arbres de classification (ou de régression) individuels. C'est une des techniques de base en exploitation de données.

Ici, nous allons voir comment cette idée peut être appliquée à notre contexte. Toutes les méthodes que nous avons vues jusqu'à maintenant font une sélection « rigide » de variables, dans le sens

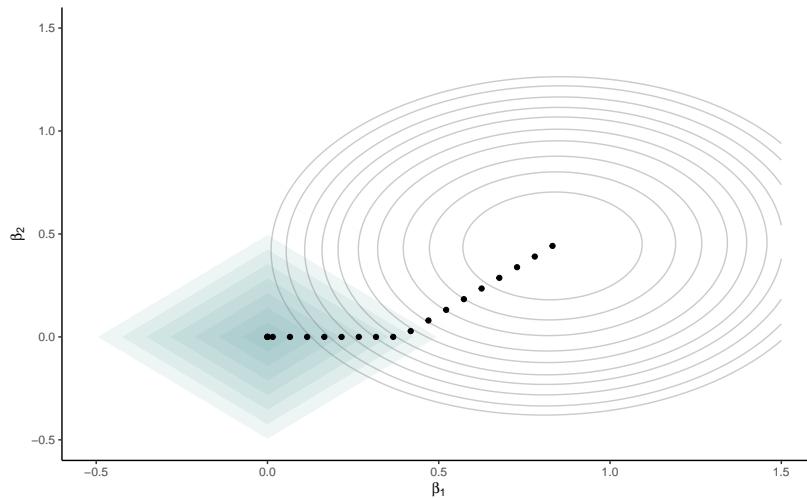


FIGURE 4.2 – Courbes de contour du critère de l'erreur quadratique moyenne (ellipses) et fonction de pénalité (losanges) pour différentes valeurs de λ . Les points dénotent des solutions différentes et intersectent les contours du losange.

que chaque variable est soit sélectionnée pour faire partie du modèle, soit elle ne l'est pas. C'est donc tout ou rien pour chaque variable. Il y a beaucoup de variabilité associée à une telle forme de sélection. Une variable peut avoir été très près d'être choisie, mais elle ne l'a pas été et est éliminée complètement. Construire plusieurs modèles et en faire la moyenne permet d'adoucir le processus de sélection car une variable peut alors être partiellement sélectionnée.

Supposons qu'on dispose de deux échantillons et qu'on fasse une sélection de variables séparément pour les deux échantillons, avec l'une des approches que nous avons vues jusqu'à maintenant. Il est alors très probable qu'on ne va pas avoir exactement les mêmes variables sélectionnées pour les deux échantillons. Supposons ensuite qu'on fasse la moyenne des coefficients pour les deux modèles. Si une variable, disons X_1 , a été choisie les deux fois, alors la moyenne des deux coefficients devrait estimer en quelque sorte un effet global pour cette variable. Si une autre variable, disons X_2 , n'a pas été choisie du tout pour les deux échantillons, alors la moyenne de ses deux coefficients est nulle. Mais si une variable, disons, X_3 , a été choisie pour seulement l'un des deux échantillons, alors la moyenne de ses deux coefficients est la moitié du coefficient pour le modèle dans lequel elle a été choisie (car l'autre est zéro). Ainsi, cette variable est donc représentée par une « moitié » d'effet dans la moyenne des modèles. Donc au lieu d'être totalement là ou totalement absente, elle est présente en fonction de sa probabilité d'être sélectionnée. Ceci diminue de beaucoup la variabilité engendrée par une sélection « rigide » de variables et permet souvent de produire un modèle fort raisonnable.

Le problème est que l'on n'a pas plusieurs échantillons mais un seul. Une solution possible est

de générer nous-mêmes des échantillons différents à partir de l'échantillon original. Cela peut être fait avec l'autoamorçage (*bootstrap*). Un échantillon d'autoamorçage est tout simplement un échantillon choisi au hasard et **avec remise** dans l'échantillon original. Ainsi, une même observation peut être sélectionnée plus d'une fois tandis qu'une autre peut ne pas être sélectionnée du tout.

L'idée est alors la suivante :

- 1) Générer plusieurs échantillons par autoamorçage nonparamétrique à partir de l'échantillon original.
- 2) Faire une sélection de variables pour chaque échantillon.
- 3) Faire la moyenne des paramètres de ces modèles.

La procédure `glmselect` a une commande, `modelaverage`, qui permet de faire une moyenne de modèles. Le code suivant permet de faire une moyenne de modèles.

```
proc glmselect data=ymontant seed=57484765;
partition role=train(train="1" validate="0");
class x3(param=ref split) x4(param=ref split);
model ymontant=x1|x2|x3|x4|x5|x6|x7|x8|x9|x10 @2
x2*x2 x6*x6 x7*x7 x8*x8 x9*x9 x10*x10 /
selection=stepwise(select=sbc choose=sbc) hier=none;
score data=testymontant out=predaverage p=predymontant;
modelaverage nsamples=500 sampling=urs subset(best=500);
run;
```

Chaque modèle est construit à l'aide d'un échantillon aléatoire avec remise (`sampling=urs`). Il y aura 500 échantillons, et donc modèles, en tout (`nsamples=500`). L'option `subset(best=500)` indique à **SAS** de faire la moyenne des paramètres des 500 modèles. Notez l'option `seed` qui permet de reproduire les résultats, car elle fixe une valeur pour le générateur de nombre aléatoire (qui sera utilisé pour générer les échantillons d'autoamorçage). Cette fois-ci la sélection se fait avec le critère BIC à tous les niveaux (`select=aic choose=sbc`).

Paramètre	Moyenne des résultats estimés des paramètres				Estimation quantiles			
	Nombre différent de zéro	Pourcentage différent de zéro	Estimation de la moyenne		Ecart-type	25%	Médiane	75%
Intercept	500	100.00	60.429767	3.352850	58.670538	60.300081	61.949300	
x44	109	21.80	1.067166	2.584582	0	0	0	
x5	412	82.40	7.451372	4.038011	5.993272	8.439527	10.012441	
x7	262	52.40	0.114565	0.121009	0	0.106973	0.215401	
x8	180	36.00	0.017365	0.026464	0	0	0.035063	
cx6	470	94.00	0.119403	0.039397	0.107385	0.128792	0.143434	
cx10	414	82.80	-0.097125	0.059167	-0.128162	-0.105223	-0.065131	
i_x2_x31	371	74.20	-0.114838	0.081225	-0.163845	-0.128573	0	
i_x1_x42	172	34.40	1.067095	1.598967	0	0	2.300445	
i_x1_x44	124	24.80	0.693210	1.336098	0	0	0	
i_x1_x6	500	100.00	1.124126	0.253744	0.946274	1.100470	1.250659	
i_x1_x10	221	44.20	-0.316154	0.394074	-0.627624	0	0	
i_x5_x10	452	90.40	1.259713	0.534894	1.047783	1.329564	1.598489	
i_x31_x41	137	27.40	0.751732	1.337842	0	0	1.908085	
i_x31_x44	111	22.20	0.576940	1.211018	0	0	0	
i_x31_x8	500	100.00	-0.120901	0.016059	-0.131599	-0.120639	-0.111370	
i_x43_x7	166	33.20	0.035325	0.053847	0	0	0.081852	
i_x7_x6	102	20.40	0.003307	0.007050	0	0	0	
i_x7_x8	175	35.00	0.000490	0.000809	0	0	0.001183	

Les paramètres sélectionnés dans moins de 20% échantillons ne sont pas affichés

Ce tableau présente les variables qui ont été choisies dans au moins 20% des modèles, c'est-à-dire, dans au moins 100 des 500 modèles ici. Il y a deux variables qui ont été retenues dans tous les modèles, x1_x6 et x31_x8. Le tableau rapporte aussi la moyenne des estimations pour ces paramètres.

Toutes les méthodes employées jusqu'à maintenant utilisent une méthode de pénalisation pour déterminer le meilleur modèle. Une alternative avec `glmselect` serait de répéter la sélection en utilisant directement l'erreur moyenne quadratique estimée à l'aide de la validation croisée comme critère de sélection (en remplaçant l'option `choose=cv` avec par exemple `cvmethod=random(10)` pour de la sélection avec $K = 10$ groupes formées aléatoirement).

4.14 Évaluation de la performance

La direction de la compagnie a décidé de passer outre vos recommandations et d'envoyer le catalogue aux 100 000 clients restants; nous pouvons donc faire un post-mortem afin de voir ce que chaque modèle aurait donné comme profit, comparativement à la stratégie de référence. Les 100 000 autres clients serviront d'échantillon de validation pour évaluer la performance des modèles et, plus précisément, afin d'évaluer les revenus (ou d'autres mesures de performance) si ces modèles avaient été utilisés. L'échantillon de validation nous donnera donc l'heure juste quant aux mérites des différentes approches que nous allons comparer. En pratique, nous ne pourrions pas faire cela car la valeur de la variable cible ne serait pas connue pour ces clients et nous utiliserions plutôt les modèles pour obtenir des prédictions pour déterminer quels clients cibler avec l'envoi. Parmi, les 100 000 clients restants, il y en a 23 179 qui auraient acheté quelque chose si on leur avait envoyé le catalogue. Ces 23 179 observations vont nous servir pour estimer l'erreur quadratique moyenne (théorique) des modèles retenus par nos critères (voir le fichier `selection2_methodes.sas` pour les manipulations). Comme notre base de données contient les 101 000 observations avec une étiquette `train`, on peut spécifier la partition avec `partition role=train(train="1" validate="0")` pour obtenir directement l'erreur moyenne quadratique pour l'échantillon de validation; cette stratégie fonctionne avec toutes les méthodes, sauf la moyenne de modèles pour laquelle on devra faire le calcul manuellement en obtenant les prédictions.

Commençons par l'estimation de l'erreur quadratique moyenne (moyenne des carrés des erreurs) pour les deux modèles retenus par le AIC et le BIC avec les variables de base. Le tableau 4.5 contient aussi l'estimation de l'erreur quadratique moyenne si on utilise toutes les variables (14 en incluant les indicatrices) sans faire de sélection. On voit que le modèle choisi par le BIC est le meilleur des trois, car l'erreur quadratique moyenne sur l'échantillon test est de 3,6% inférieure à celle du modèle choisi par le AIC. Ces deux méthodes font mieux que le modèle qui inclut toutes les variables sans faire de sélection, mais nous verrons que leur performance est exécrable : les variables de base ne permettent pas de capturer les effets présents dans les données et ce manque de flexibilité coûte cher.

TABLE 4.5: Estimation de l'erreur quadratique moyenne sur l'échantillon test avec les variables de base. Les meilleurs modèles selon les critères d'informations découlent d'une recherche exhaustive de tous les sous-ensembles.

nombre de variables	EQM	méthode
14	25,69	toutes les variables
10	24,72	exhaustive - AIC
7	23,83	exhaustive - BIC

TABLE 4.6: Comparaison des méthodes selon l'erreur quadratique moyenne avec les variables de base, les interactions et les termes quadratiques.

nombre de variables	EQM	méthode
104	19,63	toutes les variables
22	12,25	séquentielle classique
38	14,83	séquentielle classique, recherche exhaustive avec 56 variables (AIC)
15	11,96	séquentielle classique, recherche exhaustive avec 56 variables (BIC)
10	10,08	séquentielle avec critère AIC (choix selon le BIC)
26	11,61	LASSO, validation croisée avec 10 groupes
	10,57	moyenne de modèles

Le tableau 4.6 présente la performance de toutes les méthodes avec les autres variables. On voit d'abord qu'utiliser toutes les 104 variables sans faire de sélection fait mieux ($\text{EQM} = 19.63$) que les modèles précédents basés sur les 10 variables originales. Mais faire une sélection séquentielle classique permet une amélioration très importante de la performance ($\text{EQM} = 12.25$). Utiliser les 104 variables mène à du surajustement (*over-fitting*).

La stratégie consistant à sélectionner un sous-ensemble de 56 variables avec la méthode séquentielle classique pour ensuite faire une recherche exhaustive de tous les sous-modèles possibles avec ces 56 variables, selon le BIC, donne le meilleur résultat jusqu'à présent ($\text{EQM} = 11.96$). Le AIC fait moins bien dans ce cas, avec une erreur quadratique moyenne estimée de 14.83. Les méthodes séquentielles avec un critère d'information (qui pénalisent davantage que les tests d'hypothèse classique) mènent à des modèles plus parcimonieux et qui réduisent encore l'erreur moyenne quadratique. La moyenne de modèles est compétitive, tandis que le LASSO performe moins bien (probablement parce que les coefficients sont tous rétrécis vers zéro, ce qui engendre du biais).

Dans cet exemple, la méthode séquentielle de `glmselect` avec les options `select=aic` et `choose=bic` aurait donné le meilleur résultat pour prévoir le montant acheté des clients restants (de ceux qui auraient acheté quelque chose). Le deuxième meilleur aurait été la moyenne des modèles. Il faut bien comprendre qu'il ne s'agit que d'un seul exemple : il ne faut surtout pas conclure que la méthode séquentielle de `glmselect` avec les options `select=aic` et `choose=bic` sera toujours la meilleure. En fait, il est impossible de prévoir quelle méthode donnera les

meilleurs résultats.

Il y aurait plusieurs autres approches/combinasions qui pourraient être testées. Le but de ce chapitre était simplement de présenter les principes de base en sélection de modèles et de variables ainsi que certaines approches pratiques. Il y a d'autres approches intéressantes, tels le filet élastique ou la régression LARS (*least-angle regression*) qui sont disponibles dans `glmselect`. Ces méthodes sont dans la même mouvance moderne que celle qui consiste à faire la moyenne de plusieurs modèles, en effectuant à la fois une sélection de variables et en permettant d'avoir des parties d'effet par le rétrécissement (*shrinkage*). De récents développements théoriques permettent aussi de corriger les valeurs- p pour faire de l'inférence post-sélection avec le LASSO.

Chapitre 5

Régression logistique

5.1 Introduction

En régression linéaire, on cherche à expliquer le comportement d'une variable quantitative Y que l'on peut traiter comme étant continue (elle peut prendre suffisamment de valeurs différentes).

Supposons à présent que l'on veut expliquer le comportement d'une variable Y prenant seulement deux valeurs que l'on va noter 0 et 1.

Exemples :

- Est-ce qu'un client potentiel va répondre favorablement à une offre promotionnelle ?
- Est-ce qu'un client est satisfait du service après-vente ?
- Est-ce qu'un client va faire faillite ou non au cours des trois prochaines années.

En général, on cherchera à expliquer le comportement d'une variable binaire Y en utilisant un modèle basé sur p variables quelconques X_1, \dots, X_p .

Notre but sera de faire de l'inférence, de la prédiction, ou les deux à la fois, soit

- 1) Comprendre comment et dans quelles mesures les variables X influencent Y (ou bien la probabilité que $Y = 1$).
- 2) Prédiction : développer un modèle pour prévoir des valeurs de Y futures à partir des variables X .

5.2 Modèle de régression logistique

Avec une variable réponse continue, le modèle de régression linéaire,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

avec $E(\varepsilon | \mathbf{X}) = 0$ et $\text{Var}(\varepsilon | \mathbf{X}) = \sigma^2$, peut être écrit de manière équivalente comme $E(Y | \mathbf{X}) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ et $\text{Var}(Y | \mathbf{X}) = \sigma^2$.

Si Y est binaire (0/1), on peut facilement vérifier que

$$E(Y | \mathbf{X}) = P(Y = 1 | \mathbf{X}),$$

soit la probabilité que Y égale 1 étant donné les valeurs des variables explicatives. Pour simplifier la notation, posons $p = P(Y = 1 | \mathbf{X})$ en se rappelant que p est une fonction des variables explicatives.

À première vue, on peut se demander pourquoi ne pas utiliser le même modèle que la régression linéaire, c'est-à-dire

$$\eta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p.$$

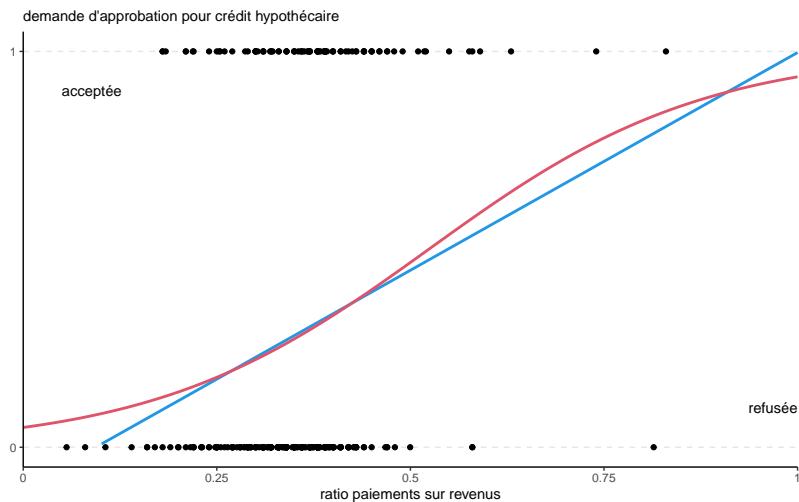


FIGURE 5.1 – Données de la réserve de Boston sur l'approbation de prêts hypothécaires (1990); données tirées de Stock et Watson (2007).

La Figure 5.1 montre le modèle de régression linéaire (bleu) et le modèle logistique. La pente pour la ligne bleue correspond à l'augmentation (réputée constante) de la probabilité d'approbation de crédit, de l'ordre de 11% par augmentation de 0.1 du rapport paiements hypothécaires sur revenu.

Il y a quelques problèmes avec le modèle linéaire. D'abord, les données binaires ne respectent pas le postulat d'égalité des variances, ce qui rend les tests d'hypothèses caducs. Le problème principal est que p est une probabilité. Par conséquent p prend seulement des valeurs entre 0 et 1 alors que rien n'empêche η de prendre des valeurs dans $\mathbb{R} = (-\infty, \infty)$: par exemple, on voit que la droite de la figure 5.1 retourne des prédictions négatives dès que le ratio paiements/revenus est

en dessous de 0.094 : on peut évidemment tronquer ces prédictions à zéro, mais cela sous-tend que la probabilité d'acceptation est nulle, alors même que certaines personnes ont reçu un prêt.

Une façon de résoudre ce problème consiste à appliquer une transformation à p de telle sorte que la quantité transformée puisse prendre toutes les valeurs entre $-\infty$ et ∞ . Le modèle de régression logistique est défini à l'aide de la transformation logit,

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

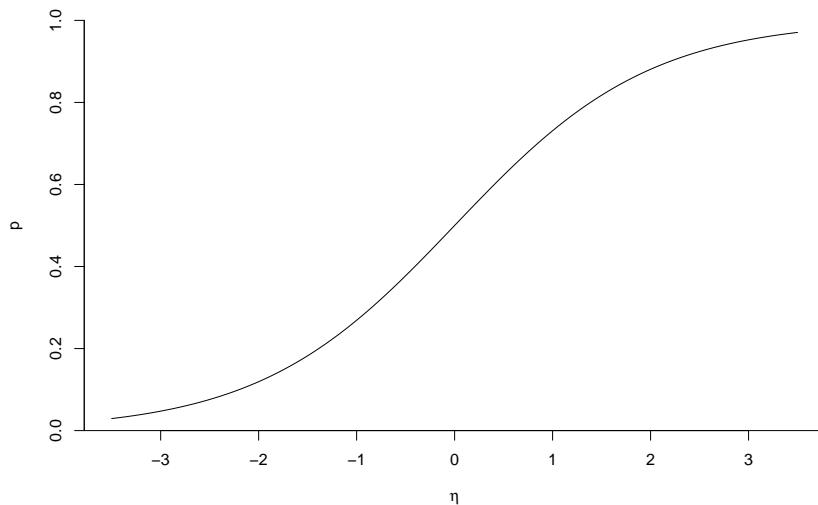
où ln est le logarithme naturel.

En régression linéaire, on suppose que l'espérance de Y étant donné les valeurs des variables explicatives est une combinaison linéaire de ces dernières. En régression logistique, on suppose que le logit de la probabilité que $Y = 1$ étant donné les valeurs des variables explicatives est une combinaison linéaire de ces dernières.

Une simple manipulation algébrique permet d'exprimer ce modèle en terme de la probabilité p ,

$$p = \text{expit}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{1}{1 + \exp(-\eta)}.$$

On peut voir qu'à mesure que le prédicteur linéaire $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ augmente, la probabilité augmente. Si le coefficient β_j est négatif, p diminuera à mesure que X_j augmente.



Pour une variable binaire Y , le quotient $p/(1 - p)$ est appelé **cote** et représente le ratio de la probabilité de succès ($Y = 1$) sur la probabilité d'échec ($Y = 0$),

$$\text{cote}(p) = \frac{p}{1 - p} = \frac{\mathbb{P}(Y = 1 | \mathbf{X})}{\mathbb{P}(Y = 0 | \mathbf{X})}.$$

TABLE 5.1 – Cote et probabilité de succès

$P(Y = 1)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
cote	0.11	0.25	0.43	0.67	1	1.5	2.3	4	9
	$\frac{1}{9}$	$\frac{1}{4}$	$\frac{3}{7}$	$\frac{2}{3}$	1	$\frac{3}{2}$	$\frac{7}{3}$	4	9

Par exemple, une cote de 4 veut dire qu'il y a 4 fois plus de chance que Y soit égale à 1 par rapport à 0. Une cote de 0.25 veut dire le contraire, il y a 4 fois moins de chance que $Y = 1$ par rapport à 0 ou bien, de manière équivalente, il y a 4 fois plus de chance que $Y = 0$ par rapport à 1. Le Tableau 5.1 donne un aperçu de cotes pour quelques probabilités p .

5.3 Estimation des paramètres

5.3.1 Principes de base

On dispose d'un échantillon de taille n sur les variables (Y, X_1, \dots, X_p) , dans le tableau

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & \ddots & \cdots & x_{2p} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

À l'aide de ces observations, on peut estimer les paramètres $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ du modèle de régression logistique

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

On obtient ainsi les estimés des paramètres $\hat{\beta}$, desquels découle une estimation de $P(Y = 1)$ pour les valeurs $X_1 = x_1, \dots, X_p = x_p$ d'un individu donné,

$$\hat{p} = \text{expit}(\hat{\beta}_0 + \cdots + \hat{\beta}_p X_p).$$

Un modèle ajusté peut ensuite être utilisé pour faire de la classification (prédiction) pour de nouveaux individus pour lesquels la variable réponse Y n'est pas observée. Pour ce faire, on choisit un point de coupure c (souvent $c = 0.5$ mais pas toujours) et on classe les observations en deux groupes :

- Si $\hat{p} < c$, alors $\hat{Y} = 0$ (c'est-à-dire, on assigne cette observation à la catégorie 0).

- Si $\hat{p} \geq c$, alors $\hat{Y} = 1$ (c'est-à-dire, on assigne cette observation à la catégorie 1).

On reviendra en détail sur cet aspect dans une section suivante.

La méthode d'estimation des paramètres habituellement utilisée est la méthode du maximum de vraisemblance. Pour les applications, il est suffisant de savoir manipuler trois quantités importantes : la log-vraisemblance, le AIC et le BIC. Les deux critères d'information, que nous avons couvert dans les chapitres précédents, servent à la sélection de modèles tandis que la log-vraisemblance ℓ servira à construire un test d'hypothèse.

5.3.2 Méthode du maximum de vraisemblance

Cette sous-section est facultative. Elle donne plus de détails sur la méthode du maximum de vraisemblance et les quantités en découlant (soit AIC, BIC et $\ell(\hat{\beta})$).

La méthode du maximum de vraisemblance (*maximum likelihood*) est probablement la méthode d'estimation la plus utilisée en statistique. En général, pour un échantillon donné et un modèle avec des paramètres inconnus θ , on peut calculer la « probabilité » d'avoir obtenu les observations de notre échantillon selon les paramètre. Si on traite cette « probabilité » comme étant une fonction des paramètres du modèle, θ , on l'appelle alors la vraisemblance (*likelihood*). La méthode du maximum de vraisemblance consiste à trouver les valeurs des paramètres qui maximisent la vraisemblance. On cherche donc les estimations qui sont les plus vraisemblables étant donné nos observations.

En pratique, il est habituellement plus simple de chercher à maximiser le log de la vraisemblance (ce qui revient au même car le log est une fonction croissante) et on nomme cette fonction la log-vraisemblance (*log-likelihood*).

Vous connaissez déjà des exemples d'estimateurs du maximum de vraisemblance. La moyenne d'un échantillon est l'estimateur du maximum de vraisemblance pour la moyenne de la population μ si les observations représentent un échantillon aléatoire simple tiré d'une loi normale.

Dans le cas d'un modèle de régression linéaire multiple $Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \varepsilon$ avec les erreurs $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ des termes indépendants et identiquement distributions, la log-vraisemblance du modèle pour un échantillon de taille n est

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_p X_{ip})^2.$$

Puisque le premier terme ne dépend pas des paramètres β , il est clair que maximiser cette fonction de β revient à minimiser $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_p X_{ip})^2$, et ce critère est exactement le même que celui des moindres carrés. Par conséquent, les estimations des paramètres β provenant de la méthode des moindres carrés peuvent être vues comme étant des estimateurs du maximum de vraisemblance sous l'hypothèse de normalité des observations ; il est même possible d'écrire une formule explicite pour ces estimations.

Dans le cas de la régression logistique, la fonction de log-vraisemblance s'écrit

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n Y_i (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}) - \sum_{i=1}^n \ln \{1 + \exp(\beta_0 + \cdots + \beta_p X_{ip})\}$$

Contrairement au cas de la régression linéaire, on ne peut trouver une fonction explicite pour les valeurs des paramètres qui maximisent cette fonction. Des méthodes numériques doivent alors être utilisées pour l'optimisation. Une fois la maximisation accomplie, on obtient les estimés du maximum de vraisemblance, $\hat{\boldsymbol{\beta}}$. On peut alors calculer la valeur maximale (numérique) de la log-vraisemblance, $\ell(\hat{\boldsymbol{\beta}})$. La quantité $-2\ell(\hat{\boldsymbol{\beta}})$ ($-2 \log L$) est rapportée dans les sorties **SAS**. Par analogie avec la régression linéaire la valeur de la log-vraisemblance évaluée à $\hat{\boldsymbol{\beta}}$, $\ell(\hat{\boldsymbol{\beta}})$, augmente toujours lorsqu'on ajoute des régresseurs et c'est pourquoi on ne pourra pas l'utiliser comme outil de sélection de variables.

Les critères d'information sont des fonctions de la log-vraisemblance, mais incluent une pénalité pour le nombre de coefficients $\boldsymbol{\beta}$,

$$\begin{aligned} AIC &= -2\ell(\hat{\boldsymbol{\beta}}) + 2(p+1) \\ BIC &= -2\ell(\hat{\boldsymbol{\beta}}) + \ln(n)(p+1) \end{aligned}$$

Ces définitions sont utilisables dans plusieurs situations lorsque le modèle est ajusté par la méthode du maximum de vraisemblance. En particulier, elles sont utilisées par **SAS** en régression logistique. Tout comme en régression linéaire et en analyse factorielle, ces deux critères pourront être utilisés pour faire de la sélection de modèles si on calcule les estimateurs du maximum de vraisemblance.

5.4 Exemple du *Professional Rodeo Cowboys Association*

L'exemple suivant est inspiré de l'article

Daneshvary, R. et Schwer, R. K. (2000) The Association Endorsement and Consumers' Intention to Purchase. *Journal of Consumer Marketing* 17, 203-213.

Dans cet article, les auteurs cherchent à voir si le fait qu'un produit soit recommandé par le *Professional Rodeo Cowboys Association* (PRCA) a un effet sur les intentions d'achats. On dispose de 500 observations sur les variables suivantes :

- Y : seriez-vous intéressé à acheter un produit recommandé par le PRCA
- 0 : non

- 1 : oui
- X_1 : quel genre d'emploi occupez-vous?
 - 1 : à la maison
 - 2 : employé
 - 3 : ventes/services
 - 4 : professionnel
 - 5 : agriculture/ferme
- X_2 : revenu familial annuel
 - 1 : moins de 25 000
 - 2 : 25 000 à 39 999
 - 3 : 40 000 à 59 999
 - 4 : 60 000 à 79 999
 - 5 : 80 000 et plus
- X_3 : sexe
 - 0 : homme
 - 1 : femme
- X_4 : avez-vous déjà fréquenté une université?
 - 0 : non
 - 1 : oui
- X_5 : âge (en années)
- X_6 : combien de fois avez-vous assisté à un rodéo au cours de la dernière année?
 - 1 : 10 fois ou plus
 - 2 : entre six et neuf fois
 - 3 : cinq fois ou moins

Le but est d'examiner les effets de ces variables sur l'intentions d'achat (Y). Les données se trouvent dans le fichier `logit1.sas7bdat`.

5.4.1 Modèle avec une seule variable explicative

Faisons tout d'abord une analyse en utilisant seulement X_5 (âge) comme variable explicative. L'ajustement du modèle de régression incluant uniquement X_5 sera effectuée en exécutant le programme

```
proc logistic data=multi.logit1 ;
model y(ref='0') = x5 / clparm=pl clodds=pl expb;
run;
```

La syntaxe `y(ref='0')` sert à spécifier la catégorie de référence, zéro, de la variable réponse Y : le modèle décrit donc $P(y = 1 | X_5)$.

Voici une partie de la sortie

Profil de réponse		
Valeur ordonnée y		Fréquence totale
1	0	272
2	1	228

La probabilité modélisée est y=1.

**Statistique d'ajustement du
modèle**

Critère	Constante uniquelement	Constante et Covariables	Constante
AIC	691.270	665.397	
SC	695.485	673.827	
-2 Log L	689.270	661.397	

**Test de l'hypothèse nulle
globale : BETA=0**

Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	27.8728	1	<.0001
Score	27.1776	1	<.0001
Wald	25.8122	1	<.0001

Analyse des valeurs estimées du maximum de vraisemblance

Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	-3.0499	0.5745	28.1835	<.0001	0.047
x5	1	0.0749	0.0147	25.8122	<.0001	1.078

Valeurs estimées du paramètre et intervalle de confiance de vraisemblance de profil

Paramètre	Estimation	Intervalle de confiance à95%	
Intercept	-3.0499	-4.1990	-1.9436
x5	0.0749	0.0465	0.1043

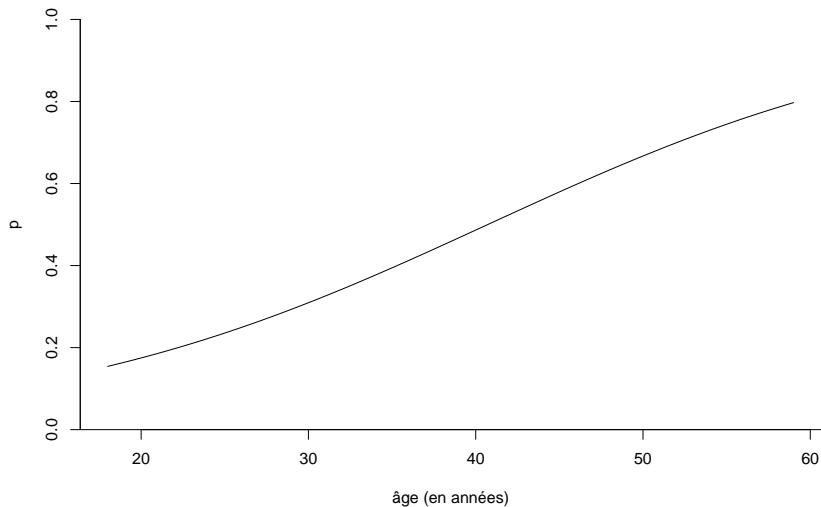
Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil

Effet	Unité	Estimation	Intervalle de confiance à95%	
x5	1.0000	1.078	1.048	1.110

- On voit qu'il y a 272 personnes (0) qui ne sont pas intéressées à acheter un produit recommandé par le PRCA et 228 personnes (1) qui le sont.
- Les estimés des paramètres sont $\hat{\beta}_0 = -3.05$ et $\hat{\beta}_{\text{age}} = 0.0749$.
- Un intervalle de confiance de niveau 95% pour l'effet de l'âge est [0.0465; 0.1043].
- Le modèle ajusté est $\text{logit}\{\mathbb{P}(Y = 1 | X_5 = x_5)\} = -3.05 + 0.0749x_5$. On peut également exprimer ce modèle directement en terme de la probabilité de succès,

$$\begin{aligned}\mathbb{P}(Y = 1 | X_5 = x_5) &= \text{expit}(-3.05 + 0.0749x_5) \\ &= \frac{1}{1 + \exp(-3.05 - 0.0749x_5)}\end{aligned}$$

Le graphe de cette fonction pour X_5 allant de 18 à 59 ans, respectivement les valeurs minimales et maximales observées dans l'échantillon, montre que le lien entre l'âge et p est presque linéaire entre 20 et 60 ans. On décèle tout de même la forme sigmoïde de la fonction logit aux deux extrémités.



- La valeur-*p* pour $\hat{\beta}_{\text{age}}$ ($\Pr > \text{khi-2}$), correspondant aux test des hypothèses $\mathcal{H}_0 : \beta_{\text{age}} = 0$ versus $\mathcal{H}_1 : \beta_{\text{age}} \neq 0$, est plus petite que 10^{-4} et donc l'effet de la variable âge est statistiquement différent de zéro. Plus l'âge augmente, plus la probabilité d'être intéressé à acheter un produit recommandé par le PRCA augmente.
- Le tableau Test de l'hypothèse nulle globale : BETA=0 contient les résultats de trois tests pour l'hypothèse nulle que tous les paramètres sont nuls, contre l'alternative qu'au moins un des paramètres est différent de zéro. Comme il y a un seul paramètre ici, ces tests reviennent à tester l'effet de la variable âge. Le test de Wald est le même que celui que nous venons de voir dans le tableau des coefficients.

5.4.2 Interprétation du paramètre

Si une variable est modélisée à l'aide d'un seul paramètre (pas de terme quadratique et pas d'interaction avec d'autre covariables), une valeur positive du paramètre indique une association positive avec *p* alors qu'une valeur négative indique le contraire.

Ainsi, le signe du paramètre donne le sens de l'association. Si le coefficient β_j de la variable X_j est positif, alors plus la variable augmente, plus $P(Y = 1)$ augmente. Inversement, Si le coefficient β_j est négatif, plus la variable augmente, plus $P(Y = 1)$ diminue.

En régression linéaire, l'interprétation de coefficient β_j est simple : lorsque la variable X_j augmente de un, la variable Y augmente en moyenne de β_j , toute chose étant égale par ailleurs. Cette interprétation ne dépend pas de la valeur de X_j . En régression logistique, comme le modèle est nonlinéaire en fonction de $P(Y = 1)$ (courbe sigmoïde), l'augmentation ou la diminution de $P(Y = 1 | \mathbf{X})$ pour un changement d'une unité de X_j dépend de la valeur de cette dernière. C'est pourquoi il est parfois plus utile d'utiliser la cote pour interpréter globalement l'effet d'une variable.

Dans notre exemple, on peut exprimer le modèle ajusté en termes de cote,

$$\frac{P(Y = 1 | X_5 = x_5)}{P(Y = 0 | X_5 = x_5)} = \exp(-3.05) \exp(0.0749x_5).$$

Ainsi, lorsque X_5 augmente d'une année, la cote est multipliée par $\exp(0.0749) = 1.078$ peu importe la valeur de x_5 . Pour deux personnes dont la différence d'âge est un an, la cote de la personne plus âgée est 7.8% plus élevée. On peut aussi quantifier l'effet d'une augmentation d'un nombre d'unités quelconque. Par exemple, pour chaque augmentation de 10 ans de X_5 , la cote est multiplié par $1.078^{10} = 2.12$, soit une augmentation de 112%.

La cote est rapportée à la dernière colonne du tableau des coefficients. En général, si on veut une interprétation globale de l'effet d'une variable, il faudra baser l'interprétation sur l'exponentielle du coefficient, $\exp(\hat{\beta})$. **SAS** dénomme cette quantité rapport de cote (*odds ratio*).

Un des avantages d'utiliser la vraisemblance comme fonction objective est que les intervalles de confiance et les estimateurs basés sur la vraisemblance (profilée) sont invariant aux reparamétrisations. Ainsi, l'intervalle de confiance à niveau 95% pour $\exp(\beta_{\text{age}})$ est obtenu en prenant l'exponentielle des bornes de l'intervalle pour β_{age} , $[\exp(0.0465); \exp(0.1043)]$, soit $[1.048; 1.110]$ tel que rapporté dans la sortie. Ce n'est **pas** le cas des intervalles de Wald qui ont la forme $\hat{\beta} \pm 1.96\text{se}(\hat{\beta})$. Comme l'exponentielle est une transformation monotone croissante, on a $\beta > 0$ si et seulement si $\exp(\beta) > 1$, etc. On peut ainsi utiliser les intervalles de confiance pour tester l'hypothèse $\mathcal{H}_0 : \beta_j = 0$ ou de façon équivalente $\mathcal{H}_0 : \exp(\beta_j) = 1$ à niveau 95%.

5.4.3 Modèle avec toutes les variables explicatives

Ajustons à présent le modèle avec toutes les variables explicatives. Rappelez-vous que la variable X_1 (quel genre d'emploi occupez-vous) a cinq catégories, X_2 (revenu familial annuel) a cinq catégories, et X_6 (combien de fois avez-vous assisté à un rodéo au cours de la dernière année) a trois catégories. Il faut donc spécifier à **SAS** de les traiter comme des variables catégorielles dans le modèle. Notez qu'on pourrait aussi traiter X_2 comme continue car elle est ordinaire et possède tout de même cinq modalités, mais on la traitera comme variable nominale.

```
proc logistic data=multi.logit1 ;
class x1(ref=last) x2(ref=last) x6 / param=ref;
model y(ref='0') =x1-x6 / clparm=pl clodds=pl expb;
run;
```

Dans **SAS**, les variables incluses dans la commande `class` sont modélisées à l'aide d'un ensemble de variables indicatrices. Cette commande nous évite de créer nous-même les indicatrices; cette option est disponible dans la plupart des procédures **SAS**, bien que la procédure `reg` est une exception notable.

On peut changer la catégorie de référence (ref=) qui est par défaut la dernière modalité (en ordre alphanumérique). L'option param=ref pour class permet d'imprimer un tableau indiquant le code pour les variables indicatrices. Les variables incluses dans la commande class sont modélisées à l'aide d'un ensemble de variables indicatrices. Prenons l'exemple de la variable X_1 : la modalité de référence est (5), soit agriculture est spécifiée dans le tableau Informations sur les niveaux de classe.

Informations sur les niveaux de classe					
	Classe	Valeur	Variables d'expérience		
x1	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	0	0	0	0
x2	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	0	0	0	0
x6	1	1	0		
	2	0	1		
	3	0	0		

Le fichier logit1_intro.sas contient le code pour ajuster le même modèle sans la commande class, c'est-à-dire en créant nous-mêmes les variables indicatrices pour inclure les variables explicatives catégorielles. Vous pouvez l'exécuter afin de vous convaincre qu'il s'agit du même modèle. Les estimés seront les mêmes.

Critère	Statistique d'ajustement du modèle	
	Constante uniquement	Constante et Covariables
AIC	691.270	544.196
SC	695.485	603.201
-2 Log L	689.270	516.196

Analyse des effets Type 3				
			Khi-2	
Effet	DDL	de Wald	Pr > khi-2	
x1	4	4.2455	0.3738	
x2	4	28.5174	<.0001	
x3	1	26.9385	<.0001	
x4	1	37.7528	<.0001	
x5	1	32.3048	<.0001	
x6	2	42.9364	<.0001	

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	-6.8883	1.0224	45.3961	<.0001	0.001
x1	1	0.3580	0.4826	0.5503	0.4582	1.430
x1	2	1	-0.4677	0.3714	1.5865	0.2078
x1	3	1	-0.3113	0.3503	0.7901	0.3741
x1	4	1	-0.3170	0.4025	0.6201	0.4310
x2	1	1	1.3312	0.5967	4.9772	0.0257
x2	2	1	1.1484	0.5014	5.2469	0.0220
x2	3	1	0.7733	0.4830	2.5628	0.1094
x2	4	1	-1.1088	0.5418	4.1884	0.0407
x3		1	1.3490	0.2599	26.9385	<.0001
x4		1	1.8303	0.2979	37.7528	<.0001
x5		1	0.1095	0.0193	32.3048	<.0001
x6	1	1	2.4122	0.3756	41.2339	<.0001
x6	2	1	1.0446	0.2493	17.5557	<.0001

Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil

Effet	Unité	Estimation	Intervalle de confiance à 95%
x1 1 vs 5	1.0000	1.430	0.557 3.715
x1 2 vs 5	1.0000	0.626	0.301 1.294
x1 3 vs 5	1.0000	0.732	0.367 1.453
x1 4 vs 5	1.0000	0.728	0.330 1.603
x2 1 vs 5	1.0000	3.786	1.195 12.520
x2 2 vs 5	1.0000	3.153	1.202 8.685
x2 3 vs 5	1.0000	2.167	0.855 5.752
x2 4 vs 5	1.0000	0.330	0.114 0.964
x3	1.0000	3.853	2.341 6.497
x4	1.0000	6.235	3.526 11.359
x5	1.0000	1.116	1.075 1.160
x6 1 vs 3	1.0000	11.158	5.456 23.882
x6 2 vs 3	1.0000	2.842	1.756 4.675

Le modèle ajusté est

$$\begin{aligned} \text{logit}\{\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x})\} &= -6.89 + 0.36\mathbf{1}_{X_1=1} - 0.47\mathbf{1}_{X_1=2} - 0.31\mathbf{1}_{X_1=3} - 0.32\mathbf{1}_{X_1=4} \\ &\quad + 1.33\mathbf{1}_{X_2=1} + 1.15\mathbf{1}_{X_2=2} + 0.77\mathbf{1}_{X_2=3} - 1.11\mathbf{1}_{X_2=4} \\ &\quad + 1.35X_3 + 1.83X_4 + 0.11X_5 + 2.41\mathbf{1}_{X_6=1} + 1.04\mathbf{1}_{X_6=2} \end{aligned}$$

Notez que les variables $\mathbf{1}_{X_1=1}$ (x11), $\mathbf{1}_{X_1=2}$ (x12), $\mathbf{1}_{X_1=3}$ (x13) et $\mathbf{1}_{X_1=4}$ (x14) représentent les quatre indicatrices pour la variable X_1 (et de même pour X_2 et X_6). L'interprétation se fait comme en régression linéaire multiple. Ici, il n'y a pas de terme quadratique, ni d'interaction. Les paramètres estimés représentent donc l'effet de la variable correspondante sur le logit une fois que les autres variables sont dans le modèle, et demeurent fixes.

Prenons le coefficient associé à l'âge (X_5) comme exemple. Le paramètre estimé est $\hat{\beta}_{\text{age}} = 0.1095$ et il est significativement différent de zéro. Ainsi, plus l'âge augmente, plus $\mathbb{P}(Y = 1 | \mathbf{X})$ augmente, toutes autres choses étant égales par ailleurs. Pour chaque augmentation d'un an de X_5 , la cote est multipliée par $\exp(0.1095) = 1.116$, lorsque les autres variables demeurent fixes.

N'oubliez pas la nuance suivante concernant l'interprétation d'un test lorsque plusieurs variables explicatives font partie du modèle. Si un paramètre n'est pas significativement différent de zéro, cela ne veut pas dire qu'il n'y a pas de lien entre la variable correspondante et Y . Cela veut seulement dire qu'il n'y a pas de lien significatif une fois que les autres variables sont dans le modèle.

Prenons l'exemple de la variable X_6 , qui représente le nombre de fois où l'individu a assisté à un rodéo au cours de la dernière année. Cette variable est modélisée à l'aide de deux variables

indicatrices, $\mathbf{1}_{X_6=1}$ égale à un si $X_6 = 1$ et zéro autrement, et $\mathbf{1}_{X_6=2}$ égale à un si $X_6 = 2$ et zéro sinon. La catégorie de référence est $X_6 = 3$, c'est-à-dire les personnes ayant assisté cinq fois ou moins à un rodéo au cours de la dernière année. Pour tester la significativité globale d'une variable catégorielle qui est modélisée avec plusieurs indicatrices, il faut aller dans le tableau Analyse des effets Type 3. On voit que la statistique de test est 42.9364 et que la valeur- p associée est négligeable : la variable X_6 est donc globalement significative. En fait, il s'agit du test conjoint sur toutes les indicatrices associées à cette variable. Plus précisément, il s'agit du test de l'hypothèse nulle $\mathcal{H}_0 : \beta_{6_1} = \beta_{6_2} = 0$ versus la contre-hypothèse qu'au moins un de ces deux paramètres est différent de zéro.

L'interprétation des variables catégorielles est analogue à celle faite en régression linéaire. On peut aussi interpréter individuellement les paramètres des indicatrices : pour $\mathbf{1}_{X_6=1}$, lorsque les autres variables demeurent fixes, les personnes ayant assisté 10 fois ou plus à un rodéo au cours de la dernière année voient leur cote multipliée par $\exp(2.4122) = 11.158$ par rapport aux personnes ayant assisté cinq fois ou moins. Ce paramètre est significativement différent de zéro car sa valeur- p est négligeable (tableau Analyse des valeurs estimées du maximum de vraisemblance) ; l'intervalle de confiance à 95% pour le rapport de cotes, basé sur la vraisemblance profilée, est [5.456; 23.882] et un n'est pas dans l'intervalle. Ainsi, il y a une différence significative entre les gens qui ont assisté à 10 rodéos ou plus et les gens qui ont assisté à 5 rodéos ou moins, pour ce qui est de l'intérêt à acheter un produit recommandé par le PRCA.

On procède de la même façon pour $\mathbf{1}_{X_6=2}$: lorsque les autres variables demeurent fixes, les personnes ayant assisté entre six et neuf fois à un rodéo au cours de la dernière année voient leur cote multipliée par 2,842 par rapport aux personnes ayant assisté cinq fois ou moins. Ce paramètre est aussi significativement différent de zéro. Il y a donc une progression. Plus une personne a assisté à un grand nombre de rodéo au cours de la dernière année, plus elle est intéressée à acheter un produit recommandé par la PRCA.

Si on désire comparer les deux modalités $X_6 = 1$ et $X_6 = 2$, il suffit de changer la modalité de référence dans la commande `class` et d'exécuter le modèle à nouveau. Une alternative est de calculer le rapport (de rapport) de cotes pour ces deux modalités.

5.4.4 Test du rapport de vraisemblance

Les tests correspondants aux valeurs- p dans le tableau des paramètres sont des tests de Wald. Ces tests feront l'affaire dans la plupart des applications. Par contre, il existe un autre test qui est généralement plus puissant, c'est-à-dire qu'il sera meilleur pour détecter que \mathcal{H}_0 n'est pas vraie lorsque c'est effectivement le cas. Ce test est le test du rapport de vraisemblance (*likelihood ratio test*). Il découle de la méthode d'estimation du maximum de vraisemblance et est donc généralement applicable lorsqu'on estime les paramètres avec cette méthode. Il est basé sur la quantité ℓ que nous avons vue plus tôt.

La procédure consiste à ajuster deux modèles emboîtés :

- Le premier modèle, le modèle complet, contient tous les paramètres et l'estimateur du maximum de vraisemblance $\hat{\beta}$.
- Le deuxième modèle correspondant à l'hypothèse nulle \mathcal{H}_0 , le modèle réduit, contient tous les paramètres avec les restrictions imposées sous \mathcal{H}_0 ; on dénote l'estimateur du maximum de vraisemblance $\hat{\beta}_0$

Le test est basé sur la statistique

$$D = -2\{\ell(\hat{\beta}_0) - \ell(\hat{\beta})\}$$

ou la différence entre $-2 \log L$ pour le modèle réduit et $-2 \log L$ pour le modèle complet. Cette différence D , lorsque l'hypothèse \mathcal{H}_0 est vraie suit approximativement une loi khi-deux avec un nombre de degrés de liberté égal au nombre de paramètre testé (le nombre de restrictions sous \mathcal{H}_0). On peut donc calculer la valeur- p en utilisant la distribution du khi-deux.

Prenons comme exemple le test de la significativité de X_6 , qui est modélisée à l'aide deux variables binaires $\mathbf{1}_{X_6=1}$ et $\mathbf{1}_{X_6=2}$ et dont les paramètres correspondants sont β_{6_1} et β_{6_2} . Nous avons déjà étudié la sortie pour le test de Wald de significativité globale de X_6 , soit le test de l'hypothèse $\mathcal{H}_0 : \beta_{6_1} = \beta_{6_2} = 0$ versus l'alternative qu'au moins un de ces deux paramètres est différent de zéro. La statistique de test (de Wald) est 42.93 et la valeur- p est moins de 10^{-4} . Pour effectuer le test du rapport de vraisemblance, il suffit de retirer la variable X_6 et de réajuster le modèle à nouveau avec toutes les autres variables; cette manipulation est effectuée dans `logit1_intro.sas`. On obtient donc $-2 \log L$ de 516,196 pour le modèle complet sans contrainte et 566.447 pour le modèle excluant la variable X_6 .

La différence $D = 566.447 - 516.196 = 50.25$. Il s'agit de la statistique du test de rapport de vraisemblance. La valeur- p peut-être obtenue de la loi du khi-deux avec 2 degrés de liberté via le code suivant permet d'imprimer la valeur- p , qui est 1.22×10^{-11} .

```
data pval;
pval=1-CDF('CHISQ', 566.447 - 516.196, 2);
run;
proc print data=pval;
run;
```

Comme la statistique du test de rapport de vraisemblance $D = 50.25$ est encore plus grande est encore plus grande que la statistique de Wald (42.9364), qui suit la même loi de probabilité sous \mathcal{H}_0 , cela indique que le test du rapport de vraisemblance est encore plus significatif que le test de Wald. Cela ne fait pas de différence ici mais, dans certains cas, il est possible que le test de Wald ne soit pas significatif (valeur- p plus grande que 0.05) tandis que le test du rapport de vraisemblance le soit (valeur- p inférieure à 0.05).

5.4.5 Multicolinéarité

Rappelez-vous que le terme multicolinéarité fait référence à la situation où les variables explicatives sont très corrélées entre elles ou bien, plus généralement, à la situation où une (ou plusieurs) variable(s) explicative(s) est (sont) très corrélée(s) à une combinaison linéaire des autres variables explicatives.

L'effet potentiellement néfaste de la multicolinéarité est le même qu'en régression linéaire, c'est-à-dire, elle peut réduire la précision des estimations des paramètres (augmenter leurs écarts-types estimés).

En pratique, le problème est qu'il devient difficile de départager l'effet individuel d'une variable explicative lorsqu'elle est fortement corrélée avec d'autres variables explicatives.

Comme la multicolinéarité est une propriété des variables explicatives (le Y n'intervient pas) on peut utiliser les mêmes outils qu'en régression linéaire pour tenter de la détecter, par exemple, le facteur d'inflation de la variance (*variance inflation factor*). Cette quantité ne dépend que des variables explicatives X , pas du modèle ou de la variable réponse.

La multicolinéarité est surtout un problème lorsque vient le temps d'interpréter et tester l'effet des paramètres individuels. Si le but est seulement de faire de la classification (prédiction) et que l'interprétation des paramètres individuels n'est pas cruciale alors il n'y a pas lieu de se soucier de la multicolinéarité. Il faut alors plutôt comparer correctement la performance de classification des modèles en utilisant des méthodes permettant d'obtenir un bon modèle tout en se protégeant contre le surajustement. Certaines de ces méthodes (division de l'échantillon, validation croisée) ont déjà été présentées.

5.5 Classification et prédiction à l'aide de la régression logistique

La finalité du modèle de régression logistique est fréquemment l'obtention de prédictions. Une fois qu'on a ajusté un modèle, on peut l'utiliser pour prévoir la valeur de Y pour de nouvelles observations. Ceci consiste à assigner une classe (0 ou 1) à ces observations (pour lesquels Y est inconnue) à partir des valeurs prises par X_1, \dots, X_p .

Le modèle ajusté nous fournit une estimation de $P(Y = 1 | \mathbf{X} = \mathbf{x})$ pour des valeurs $X_1 = x_1, \dots, X_p = x_p$ données. Cet estimé est

$$\hat{p} = \frac{1}{1 + \exp\{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p)\}}.$$

Classification de base : pour classifier des observations, il suffit de choisir un point de coupure c , souvent $c = 0.5$, et de classifier une observation de la manière suivante :

- Si $\hat{p} < c$, on assigne cette observation à la catégorie zéro et $\hat{Y} = 0$.
- Si $\hat{p} \geq c$, on assigne cette observation à la catégorie un et $\hat{Y} = 1$.

Si on prend $c = 0.5$ comme point de coupure, cela revient à assigner l'observation à la classe (catégorie) la plus probable, un choix fort raisonnable. Nous verrons dans une section suivante que, lorsque les conséquences de faussement classifier une observation (succès, mais échec prédit et vice-versa) ne sont pas les mêmes, il peut être avantageux d'utiliser un autre point de coupure.

Dans un cadre de prédition, il nous faudra un critère pour juger de la qualité de l'ajustement du modèle. Rappelez-vous que pour une réponse continue, nous avons utilisé l'erreur moyenne quadratique, $EMQ = E\{(Y - \hat{Y})^2\}$, où $\hat{Y} = E(Y | \mathbf{X})$, pour juger de la performance d'un modèle. Comme la réponse Y est binaire ici, nous allons utiliser des critères différents.

Voyons d'abord un premier critère pour juger de la qualité d'un modèle de prédition. Soit Y la vraie valeur de la réponse binaire et \hat{Y} (soit 0 ou 1) la valeur de Y prédite par un modèle pour une observation choisie au hasard dans la population. Un premier critère pour juger de la performance d'un modèle est le **taux de mauvaise classification**, un estimé de la probabilité de mal classifier une observation choisie au hasard dans la population, $P(Y \neq \hat{Y})$. Plus $P(Y \neq \hat{Y})$ est petite, meilleure est la capacité prédictive du modèle.

Tout comme l'erreur moyenne quadratique, on ne peut qu'estimer $P(Y \neq \hat{Y})$. Pour les raisons vues au chapitre précédent, l'estimer en calculant le taux de mauvaise classification des observations ayant servi à l'ajustement du modèle sans aucune correction n'est pas une bonne approche. Les approches couvertes dans le dernier chapitre pour l'estimation de l'erreur moyenne quadratique, telles la validation-croisée et la division de l'échantillon, peuvent être utilisées pour estimer le taux de mauvaise classification $P(Y \neq \hat{Y})$.

Cette utilisation d'un modèle de régression logistique sera illustrée avec l'exemple que nous avons traité au chapitre précédent : notre objectif final est de construire un modèle avec les 1000 clients de l'échantillon d'apprentissage et cibler ensuite lesquels des 100 000 clients restants seront choisis pour recevoir le catalogue. Les variables cibles sont :

- *yachat* : variable binaire égale à un si le client a acheté quelque chose dans le catalogue et zéro sinon.
- *ymontant* : le montant de l'achat si le client a acheté quelque chose

Les 10 variables suivantes sont disponibles pour tous les clients et serviront de variables explicatives,

- *x1* : sexe de l'individu, soit homme (0) ou femme (1);
- *x2* : l'âge (en année);
- *x3* : variable catégorielle indiquant le revenu, soit moins de 35 000\$ (1), entre 35 000\$ et 75 000\$ (2) ou plus de 75 000\$ (3);
- *x4* : variable catégorielle indiquant la région où habite le client (de 1 à 5);
- *x5* : conjoint : le client a-t-il un conjoint, soit oui (1) ou non (0);

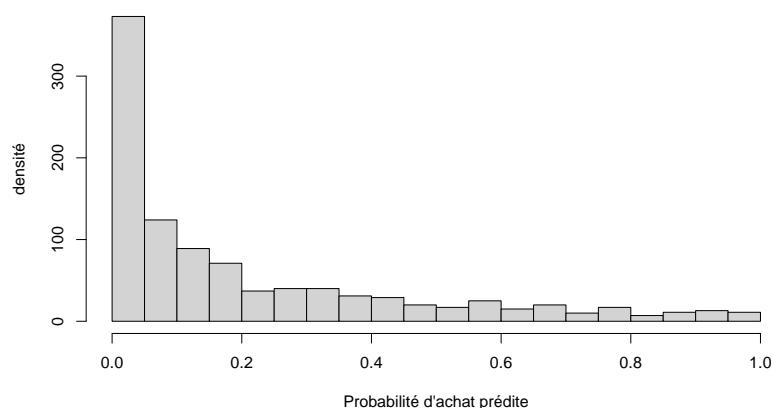
- x6 : nombre d'année depuis que le client est avec la compagnie;
- x7 : nombre de semaines depuis le dernier achat;
- x8 : montant (en dollars) du dernier achat;
- x9 : montant total (en dollars) dépensé depuis un an;
- x10 : nombre d'achats différents depuis un an.

Dans le chapitre précédent, nous avons cherché à développer un modèle pour prévoir $y_{montant}$, le montant dépensé, étant donné que le client achète quelque chose. Cette fois-ci, nous allons travailler avec la variable y_{achat} , qui est binaire, à l'aide de la régression logistique.

Afin d'introduire différentes notions, nous allons, dans un premier temps, utiliser les 10 variables de base. À partir de la section suivante, nous chercherons à optimiser le modèle en considérant les interactions d'ordre deux. Pour ce faire, nous utiliserons des méthodes de sélections de variables. Les commandes se trouvent dans le fichier `logit2_classification_base.sas`. Dans le code qui suit, le fichier `train` contient les 1000 clients de l'échantillon d'apprentissage et le fichier `test` contient les 100 000 clients pour lesquels on veut prédire l'intention d'achat.

```
proc logistic data=train;
model yachat(ref='0') = x1x2 x3 x32 x41-x44 x5-x10;
output out=pred predprobs=crossvalidate;
run;
```

Le modèle utilise seulement les 10 variables de base (en fait 14 avec les indicatrices pour les variables catégorielles). Des prévisions pour les clients restants seront exportées dans le fichier `pred`, grâce à la commande `score`. L'option `ctable` permet d'obtenir la Table de classification (sic). Tel que nous l'avons vu au chapitre précédent, il y a 210 clients qui ont acheté quelque chose parmi les 1000.



coupe	VP	VN	FP	FN	correct (%)	sensibilité (%)	spécificité (%)	FP (%)	FN (%)
0.02	210	209	581	0	41.9	100.0	26.5	73.5	0.0
0.04	207	320	470	3	52.7	98.6	40.5	69.4	0.9
0.06	201	398	392	9	59.9	95.7	50.4	66.1	2.2
0.08	199	451	339	11	65.0	94.8	57.1	63.0	2.4
0.10	193	480	310	17	67.3	91.9	60.8	61.6	3.4
0.12	191	512	278	19	70.3	91.0	64.8	59.3	3.6
0.14	184	547	243	26	73.1	87.6	69.2	56.9	4.5
0.16	176	572	218	34	74.8	83.8	72.4	55.3	5.6
0.18	172	598	192	38	77.0	81.9	75.7	52.7	6.0
0.20	164	611	179	46	77.5	78.1	77.3	52.2	7.0
0.22	162	626	164	48	78.8	77.1	79.2	50.3	7.1
0.24	158	639	151	52	79.7	75.2	80.9	48.9	7.5
0.26	153	645	145	57	79.8	72.9	81.6	48.7	8.1
0.28	150	657	133	60	80.7	71.4	83.2	47.0	8.4
0.30	143	667	123	67	81.0	68.1	84.4	46.2	9.1
0.32	138	679	111	72	81.7	65.7	85.9	44.6	9.6
0.34	134	695	95	76	82.9	63.8	88.0	41.5	9.9
0.36	130	699	91	80	82.9	61.9	88.5	41.2	10.3
0.38	126	708	82	84	83.4	60.0	89.6	39.4	10.6
0.40	120	715	75	90	83.5	57.1	90.5	38.5	11.2
0.42	115	723	67	95	83.8	54.8	91.5	36.8	11.6
0.44	112	731	59	98	84.3	53.3	92.5	34.5	11.8
0.46	109	736	54	101	84.5	51.9	93.2	33.1	12.1
0.48	106	739	51	104	84.5	50.5	93.5	32.5	12.3
0.50	100	744	46	110	84.4	47.6	94.2	31.5	12.9
0.52	98	748	42	112	84.6	46.7	94.7	30.0	13.0
0.54	92	750	40	118	84.2	43.8	94.9	30.3	13.6
0.56	87	753	37	123	84.0	41.4	95.3	29.8	14.0
0.58	83	761	29	127	84.4	39.5	96.3	25.9	14.3
0.60	80	766	24	130	84.6	38.1	97.0	23.1	14.5
0.62	77	769	21	133	84.6	36.7	97.3	21.4	14.7
0.64	74	771	19	136	84.5	35.2	97.6	20.4	15.0
0.66	68	772	18	142	84.0	32.4	97.7	20.9	15.5
0.68	62	774	16	148	83.6	29.5	98.0	20.5	16.1
0.70	54	775	15	156	82.9	25.7	98.1	21.7	16.8

0.72	51	777	13	159	82.8	24.3	98.4	20.3	17.0
0.74	49	778	12	161	82.7	23.3	98.5	19.7	17.1
0.76	46	778	12	164	82.4	21.9	98.5	20.7	17.4
0.78	41	781	9	169	82.2	19.5	98.9	18.0	17.8
0.80	35	783	7	175	81.8	16.7	99.1	16.7	18.3
0.82	33	783	7	177	81.6	15.7	99.1	17.5	18.4
0.84	32	783	7	178	81.5	15.2	99.1	17.9	18.5
0.86	28	784	6	182	81.2	13.3	99.2	17.6	18.8
0.88	25	786	4	185	81.1	11.9	99.5	13.8	19.1
0.90	21	787	3	189	80.8	10.0	99.6	12.5	19.4
0.92	18	787	3	192	80.5	8.6	99.6	14.3	19.6
0.94	14	788	2	196	80.2	6.7	99.7	12.5	19.9
0.96	6	788	2	204	79.4	2.9	99.7	25.0	20.6
0.98	2	790	0	208	79.2	1.0	100.0	0.0	20.8

Le tableau de classification contient des estimations de plusieurs quantités intéressantes, en faisant varier le point de coupure (Niveau de proba dans le tableau SAS). Pour chaque point de coupure, ces estimations ont été obtenues à l'aide d'une approximation de la méthode de validation croisée à n groupes (en anglais, *leave-one-out cross-validation*, ou LOOCV). Ainsi, ces estimations sont meilleures que les estimés sans ajustement aucun car elles ne sont pas obtenues en utilisant les mêmes observations que celles qui ont servi à estimer le modèle.

La colonne **correct** donne une estimation du taux de bonne classification, $P(Y = \hat{Y}) = 1 - P(Y \neq \hat{Y})$, ou de manière équivalente un moins le taux de mauvaise classification.

Avec un point de coupure de 0, on classifie toutes les observations à la classe achat (1), car \hat{p} est forcément plus grande que zéro. Le taux de bonne classification dans ce cas de figure sera de 21%, puisque 210 individus ont acheté un produit dans le catalogue dans l'échantillon d'apprentissage. L'autre extrême, avec un point de coupure $c = 1$, donne un taux de bonne classification de 79%.

On peut chercher dans le tableau les points de coupure qui donnent le meilleur taux de bonne classification. Ce dernier, à savoir 84.6%, est atteint par trois points de coupure, soit 0.52, soit 0.6, soit 0.62. Une recherche plus fine donne 0.465 comme point de coupure optimal, avec un taux de mauvaise classification de 15.3%.

La **matrice de confusion**, qui compare les vraies valeurs avec les prédictions, peut être construite à partir des colonnes **Correct** – Événement, **Correct** – Non-événement, **Incorrect** – Événement, **Incorrect** – Non-événement. Il y a deux classifications possibles et le tableau contient, en partant du coin supérieur gauche et dans le sens des aiguilles d'une montre, le nombre de vrai positif ($Y = 1, \hat{Y} = 1$), de faux positif ($Y = 0, \hat{Y} = 1$), de vrai négatif ($Y = 0, \hat{Y} = 0$) et finalement de faux négatif ($Y = 1, \hat{Y} = 0$). Ces nombres proviennent de la validation croisée à

TABLE 5.3 – Matrice de confusion avec point de coupure 0.465

	\$Y=1\$	\$Y=0\$
\$\widehat{Y}=1\$	109	52
\$\widehat{Y}=0\$	101	738

n groupes et ne sont pas ceux qu'on obtiendrait si on appliquait directement le modèle ajusté à notre échantillon. Le taux de mauvaise classification est $(FP + FN)/n$.

Quatre autres quantités, dérivées à partir de la matrice de confusion, sont parfois utilisées :

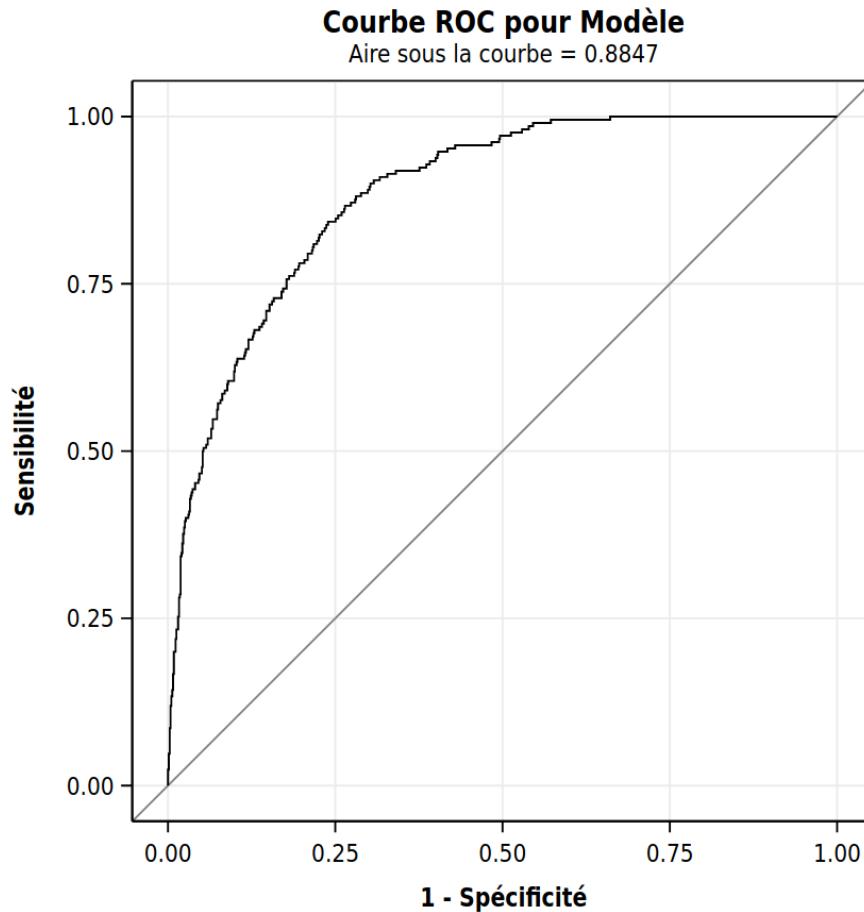
- la **sensibilité** (*sensitivity*), $P(\widehat{Y} = 1 | Y = 1)$, ou $VP/(VP + FN)$;
- la **spécificité** (*specificity*), $P(\widehat{Y} = 0 | Y = 0)$, ou $VN/(VN + FP)$;
- le **taux de vrais positifs**, $P(Y = 1 | \widehat{Y} = 1)$, ou $VP/(VP + FP)$;
- le **taux de vrais négatifs**, $P(Y = 0 | \widehat{Y} = 0)$, ou $VN/(VN + FN)$.

Les estimés empiriques sont simplement obtenus en calculant les rapports du nombre d'observations dans chaque classe. **SAS** rapporte ces quantités, mais notez que les vieilles versions du logiciel retournent le taux de faux positifs et de faux négatifs dans les deux dernières colonnes, tandis que les sorties des nouvelles version du logiciel donnent les taux de vrais positifs et de vrais négatifs.

La sensibilité mesure à quel point notre modèle est performant pour détecter un vrai positif (classe 1). La spécificité mesure à quel point notre modèle est performant pour détecter un résultat négatif (classe 0). Plus le point de coupure augmente, plus la sensibilité et le taux de faux positifs diminuent mais plus la spécificité et le taux de faux négatifs augmentent.

La **fonction d'efficacité du récepteur**, parfois appelée courbe ROC (*receiver operating characteristic*) est parfois utilisée pour représenter globalement la performance du modèle. Elle est obtenue avec l'option `plots(only)=(roc)` dans **SAS**. Il s'agit du graphe de la sensibilité en fonction de un moins la spécificité, en faisant varier le point de coupure. Un modèle parfait aurait une sensibilité et une spécificité égales à 1 (correspondant au coin supérieur gauche de la fonction d'efficacité du récepteur). Ainsi, plus le couple (1 - spécificité, sensibilité) est près de (0, 1), meilleur est le modèle. Par conséquent, plus la courbe ROC tend vers (0, 1) meilleur est le pouvoir prévisionnel des variables.

L'**aire sous la courbe** (*area under the curve*) est souvent utilisée en parallèle est simplement l'aire sous la courbe de la fonction d'efficacité du récepteur. Pour le modèle logistique ajusté, on a une aire sous la courbe de 0.8847. Plus cette valeur est élevée (au plus 1), mieux c'est.



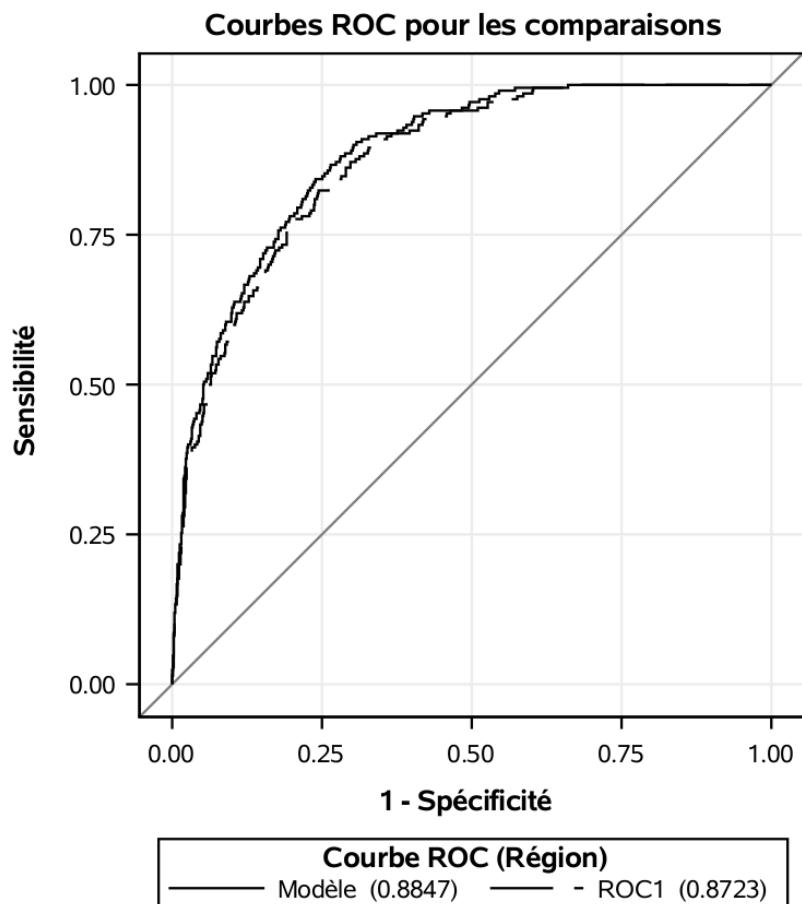
La courbe ROC et la valeur de l'aire sous la courbe (avec l'option `plots(only)=(roc)`), sont calculées avec les données d'apprentissage et ne sont pas corrigées. Si on veut les utiliser pour comparer des modèles, il faut plutôt utiliser l'option `crossvalidate` qui permet d'obtenir des estimations des probabilités par validation-croisée avec n groupes tout comme celle utilisée dans le tableau de classification.

```
proc logistic data=train;
class x3 x4 / ref=glm;
model yachat(ref='0') = x1-x10;
output out=pred predprobs=crossvalidate;
run;

proc logistic data=pred;
```

```
class x3 x4 / ref=glm;
model yachat(ref='0') = x1-x10;
roc pred=xp_1;
run;
```

On sauvegarde d'abord les probabilités estimées par validation-croisée dans le fichier `pred` avec la commande `output out=pred predprobs=crossvalidate`. La variable `xp_1` désigne cette probabilité dans le fichier `pred`. Ensuite, on exécute de nouveau la procédure `logistic` avec ce fichier et la commande `roc`. L'aire sous la courbe pour les prédictions avec la validation-croisée à n groupes est 0.8723 : cet estimé est légèrement inférieur à celui obtenu sans la correction (trop optimiste) qui est 0.8847.



Un autre type de graphe qui est souvent utilisé dans des contextes de gestion est la courbe lift (sic) (en anglais, *lift chart*). Cette courbe est obtenue en ordonnant les probabilités de succès estimées

par le modèle, \hat{p} , en ordre croissant et en regardant quelle pourcentage de ces derniers seraient bien classifiés (le nombre de vrais positifs sur le nombre de succès).

SAS ne permet pas de la tracer directement, mais le fichier `logit3_lift_chart.sas` contient une macro **SAS** qui permet de le faire.

```
proc logistic data=train;
model yachat(ref='0') = x1 x2 x31 x32 x41-x44 x5-x10;
output out=pred predprobs=crossvalidate;
run;
%liftchart1(pred,yachat,xp_1,10);
```

Ici, le tableau présente les 10 déciles. Si on classifiait comme acheteurs les 10% qui ont la plus forte probabilité estimée d'achat, on détecterait 79 des 210 clients (37,6%). En comparaison, on s'attend que 21 clients soient sélectionnés en moyenne si on prend un échantillon aléatoire de 100 personnes. Le ratio 79/21 (dernière colonne) est le *lift* du modèle : il permet de détecter 3,76 fois plus de succès que le hasard.

Obs.	Nombre de 1 qui seraient détectés en choisissant les obs au hasard								lift
	% des obs classées 1 par le modèle	Nombre d'obs classées 1 par le modèle	Nombre détectés en choisissant les obs au hasard	% de 1 qui sont détectés par le modèle	Nombre détectés par le modèle	Nombre de 1 supplémentaire que le modèle détecte comparativement au hasard			
1	10	100	21	37.619	79	58	3.76190		
2	20	200	42	57.619	121	79	2.88095		
3	30	300	63	73.333	154	91	2.44444		
4	40	400	84	84.286	177	93	2.10714		
5	50	500	105	91.905	193	88	1.83810		
6	60	600	126	95.714	201	75	1.59524		
7	70	700	147	99.524	209	62	1.42177		
8	80	800	168	100.000	210	42	1.25000		
9	90	900	189	100.000	210	21	1.11111		
10	100	1000	210	100.000	210	0	1.00000		

Le graphique 5.2 présente le pourcentage d'observations bien classées parmi les variables (pourcentage des probabilités prédites qui correspondent à un succès parmi les k plus susceptibles selon le modèle). La référence est la ligne diagonale, qui correspond à une détection aléatoire.

Il peut être intéressant de vérifier la **calibration** de notre modèle, et une statistique simple proposée par Spiegelhalter (1986) peut être utile à cette fin. Pour une variable binaire $Y \in \{0,1\}$, l'erreur

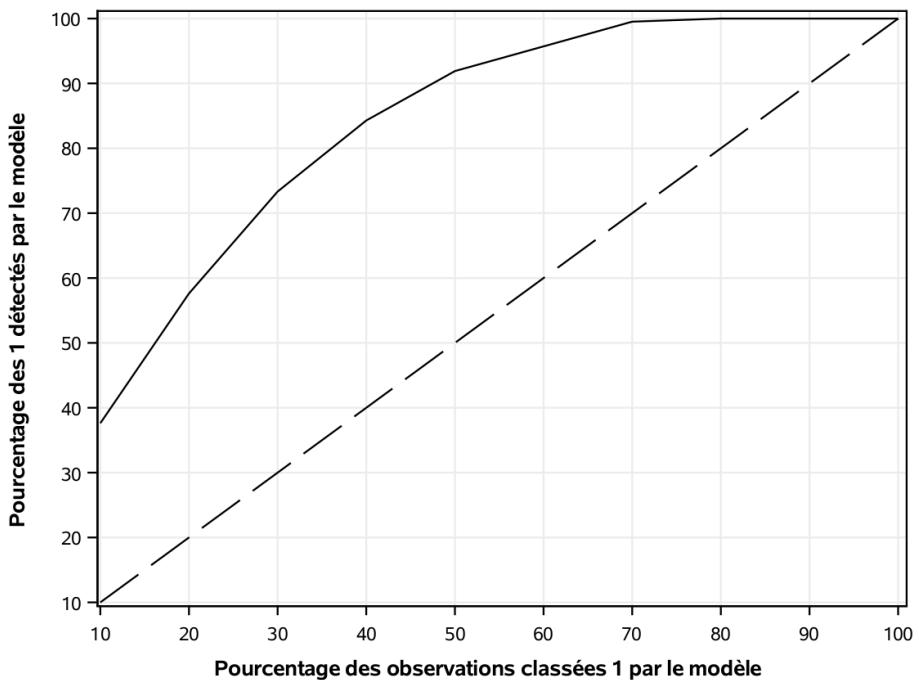


FIGURE 5.2 – Taux de classement en fonction du lift

moyenne quadratique s'écrit

$$\bar{B} = \frac{1}{n} \sum_{i=1}^n (Y_i - p_i)^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - p_i)(1 - 2p_i) + \frac{1}{n} \sum_{i=1}^n p_i(1 - p_i).$$

Le premier terme représente le manque de calibration du modèle, tandis que le deuxième correspond à la séparation entre variables. Si notre modèle était parfaitement calibré, alors $E_0(Y_i) = p_i$ et $\text{Var}_0(Y_i) = p_i(1 - p_i)$. On peut utiliser ce fait pour construire une statistique de Wald, $Z = \{\bar{B} - E_0(\bar{B})\}/\sqrt{\text{Var}_0(\bar{B})}$, où

$$E_0(\bar{B}) = \frac{1}{n} \sum_{i=1}^n p_i(1 - p_i)$$

$$\text{Var}_0(\bar{B}) = \frac{1}{n^2} \sum_{i=1}^n p_i(1 - p_i)(1 - 2p_i)^2$$

Sous l'hypothèse nulle de calibration parfaite, $Z \sim \text{No}(0,1)$ en grand échantillon. Pour le modèle simple avec toutes les covariables, la valeur- p approximative calculée avec les probabilités de succès obtenues par validation-croisée et les données de l'échantillon d'apprentissage est 0.22 et il n'y a pas de preuve que le modèle est mal calibré. Cette technique est utile pour vérifier s'il n'y a pas de surajustement (auquel cas le modèle tend à retourner des probabilités très près de 0/1, mais qui ne correspondent pas à la réalité).

5.6 Classification avec une matrice de gain

Utiliser le taux de mauvaise classification $P(Y \neq \hat{Y})$, comme critère de performance, revient au même que d'utiliser le taux de bonne classification $P(Y = \hat{Y})$, car $P(Y \neq \hat{Y}) = 1 - P(Y = \hat{Y})$. On veut un modèle avec un haut taux de bonne classification (ou un faible taux de mauvaise classification).

Lorsqu'on utilise $P(Y \neq \hat{Y})$ comme critère pour juger de la qualité d'un modèle prévisionnel, on fait l'hypothèse que le gain associé à bien classifier une observation dans la catégorie 0 lorsqu'elle est réellement dans la catégorie 0 est le même que celui associé à classifier une observation dans la catégorie 1 lorsqu'elle est réellement dans la catégorie 1 : cela correspond à la matrice de gain.

TABLE 5.4: Matrice de gain correspondant au taux de bonne classification

		observation	
		gain	$Y = 1$
gain	$Y = 0$		
gain	$Y = 0$		

		observation	
prédiction	$\hat{Y} = 1$	1	0
	$\hat{Y} = 0$	0	1

C'est-à-dire, le gain vaut 1 lorsque la prévision est bonne (les deux cas sur la diagonale) et 0 lorsque le modèle se trompe (les deux autres cas). L'unité de mesure du gain n'est pas importante pour l'instant. Le gain total est

$$\begin{aligned} \text{gain} &= 1P(\hat{Y} = 1, Y = 1) + 1P(\hat{Y} = 0, Y = 0) \\ &\quad + 0P(\hat{Y} = 1, Y = 0) + 0P(\hat{Y} = 0, Y = 1) \\ &= P(Y = \hat{Y}). \end{aligned}$$

Maximiser le gain total revient donc à maximiser le taux de bonne classification.

Dans certaines situations, les gains (ou la perte si le gain est négatif) associés aux bonnes décisions et aux erreurs ne sont pas équivalents. Par exemple, un des types d'erreurs peut être plus grave que l'autre. Il peut alors être souhaitable d'en tenir compte dans le choix du modèle de classification.

Supposons que le gain de classer une observation à i ($i \in \{0,1\}$) lorsqu'elle vaut j ($j \in \{0,1\}$) en réalité est de c_{ij} . La matrice de gain est alors

TABLE 5.5: Matrice de gain pondérée en fonction d'un coût

		observation	
prédiction	gain	$Y = 1$	$Y = 0$
	$\hat{Y} = 1$	c_{11}	c_{10}
	$\hat{Y} = 0$	c_{01}	c_{00}

En pratique, l'une de ces quatre quantités peut être fixée à un car seulement les poids relatifs (les ratios) des gains sont importants. Dans ce cas, le gain moyen est

$$\begin{aligned} \text{gain} &= c_{11}P(\hat{Y} = 1, Y = 1) + c_{00}P(\hat{Y} = 0, Y = 0) \\ &\quad + c_{10}P(\hat{Y} = 1, Y = 0) + c_{01}P(\hat{Y} = 0, Y = 1) \end{aligned}$$

Le meilleur modèle est alors celui qui maximise le gain moyen. Le fichier `logit4_macro_gain.sas` contient des macros **SAS** qui permettent d'estimer le gain moyen à l'aide de la validation croisée.

Nous allons encore une fois seulement utiliser les 10 variables de base. Mais nous allons intégrer des revenus et coûts afin de trouver le meilleur point de coupure. Rappelez-vous que le coût de

l'envoi d'un catalogue est de 10\$. Le tableau des variables descriptives qui suit montre que, pour les 210 clients qui ont acheté quelque chose, le revenu moyen est de 67,29\$ (moyenne de la variable *ymontant*).

Variable d'analyse : <i>ymontant</i> montant de l'achat (catalogue)				
N	Moyenne	Ec-type	Minimum	Maximum
210	67.29	13.24	25.00	109.00

Nous allons travailler en termes de revenu net. Nous pouvons donc spécifier la matrice de gain du tableau 5.6 pour notre problème. Si on n'envoie pas de catalogue, notre gain est nul. Si on envoie le catalogue à un client qui n'achète pas, on perd 10\$ (le coût de l'envoi). En revanche, notre revenu net est de 57\$ (revenu moyen moins coût de l'envoi).

TABLE 5.6: Matrice de gain pour l'envoi de catalogue

		observation	
		$Y = 1$	$Y = 0$
prédiction	gain	57	-10
	$\hat{Y} = 1$	0	0
	$\hat{Y} = 0$		

L'appel de la macro `manycut_cvlogistic`, dont les paramètres sont expliqués dans le script, se fait de la manière suivante :

```
%manycut_cvlogisticclass(
  yvar=yachat, xvar=x1-x10, xvarclass=x3-x4,
  n=1000, k=10, ncv=10, dataset=train,
  c00=0, c01=0, c10=-10, c11=57,
  manycut=.05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .5);
```

Cette macro produit le tableau suivant. Il donne l'estimation du gain moyen (*gain*) pour différents points de coupures (*cutpoint*). Cette estimation provient d'une validation-croisée avec 10 groupes (*k*=10 dans la macro). En fait, on a répété 10 fois (*ncv*=10 dans la macro) la validation croisée avec 10 groupes et fait la moyenne des 10 répétitions afin d'avoir plus de précisions. Il faut essayer plusieurs points de coupure afin de trouver le meilleur.

On voit que le meilleur point de coupure, celui qui maximise le gain est 0.12. Avec ce point de coupure, on estime que le taux de bonne classification est de 0.707 et que la sensibilité est de 0.899. Ainsi, on estime qu'on va détecter 90% des clients qui achètent.

Obs.	cutpoint	gain	good	sensitivity	specificity
1	0.05	7.3729	0.5699	0.97011	0.46379
2	0.06	7.6042	0.6029	0.96024	0.50824
3	0.07	7.7643	0.6269	0.95185	0.54068
4	0.08	7.9075	0.6478	0.94501	0.56884
5	0.09	7.9166	0.6614	0.93201	0.58934
6	0.10	7.9280	0.6757	0.91728	0.61098
7	0.11	7.9891	0.6898	0.90946	0.63110
8	0.12	8.0624	0.7070	0.89920	0.65544
9	0.13	8.0461	0.7190	0.88640	0.67420
10	0.14	8.0064	0.7296	0.87127	0.69162
11	0.15	7.9614	0.7392	0.85719	0.70748
12	0.16	7.9037	0.7480	0.84219	0.72255
13	0.17	7.8419	0.7578	0.82532	0.73934
14	0.18	7.7582	0.7640	0.80995	0.75127
15	0.50	5.3862	0.8464	0.48769	0.94152

On est loin du point de coupure usuel de 0.5 (présenté à la dernière ligne). La raison est simple. Comme il est très coûteux de rater un client qui aurait acheté quelque chose, il est préférable d'envoyer le catalogue à plus de clients, quitte à ce que plusieurs d'entre eux n'achètent rien. En fait, le point de coupure de 0.5 donne un meilleur taux de bonne classification mais un gain moyen plus faible car on rate trop de clients qui achètent (la sensibilité est seulement de 48,8%). Travailler avec la matrice de gain permet de trouver le point de coupure optimal en incorporant des notions de coûts et profits.

Ici, nous avons ajusté un seul modèle, celui contenant uniquement les 10 variables de base et nous nous sommes attardés au choix du point de coupure pour l'assignation aux classes. Il est possible qu'un autre modèle, contenant par exemple des termes d'interactions, des termes quadratiques ou d'autres transformations des variables, soit supérieur à celui-ci. Le choix du modèle de prévision se fait donc souvent en deux étapes

1. trouver les bonnes variables
2. trouver le bon point de coupure.

Nous avons déjà vu des méthodes de sélections de variables au chapitre précédent. La section suivante reviendra sur ces méthodes dans le contexte de la régression logistique.

5.7 Sélection de variables en régression logistique

Les principes généraux, concernant la sélection de variables et de modèles, que nous avons vus au chapitre précédent sont toujours valides. Les critères AIC et BIC sont toujours disponibles puisqu'on estime le modèle par maximum de vraisemblance et les techniques générales de division de l'échantillon et de validation-croisée sont toujours valides. Nous allons voir comment appliquer spécifiquement ces techniques au cas de la régression logistique.

5.7.1 Recherche séquentielle

Rappelez-vous qu'avec une variable cible continue, nous avons utilisé avec la procédure `reg` pour faire une recherche du meilleur sous-ensemble de variables parmi tous les ensembles. Pour ce faire, on sélectionnait le meilleur modèle selon le R^2 pour un nombre de variables fixe et il suffisait ensuite de trouver parmi ces variables le meilleur selon le critère d'information choisi.

Parce qu'il n'y a pas de solution explicite pour les estimateurs du maximum de vraisemblance du modèle logistique, ajuster chacun de ces modèles est coûteux. Les options pour la sélection de modèle avec le modèle de régression logistique est très limité dans **SAS** : toutes les procédures supportent la sélection à des degrés variés (pas de validation externe basée sur la log-vraisemblance, pas de validation croisée). Comble de malheur, le support des options n'est pas cohérent d'une procédure à l'autre. On peut se rabattre sur une recherche séquentielle ou le LASSO : pour cette première, il est possible d'utiliser une stratégie d'élimination rapide avec la statistique du score (ou test des multiplicateurs de Lagrange) pour tester si l'ajout d'une variable est utile ou pas : c'est une approximation de la recherche exhaustive des meilleurs sous-ensembles.

Ce paragraphe est plus technique et peut être omis. La statistique de score, qui est basée sur la vraisemblance, ne nécessite que d'obtenir le maximum de vraisemblance sous l'hypothèse nulle ; cela permet d'éviter des ajustements coûteux lors de comparaisons. L'algorithme employé par **SAS** utilise une méthode de recherche arborescente dite méthode de séparation et d'évaluation, qui ne nécessite pas de tester tous les modèles ; à noter que la solution à k variables n'est pas nécessairement imbriquée dans celle à $k + 1$ variables. Lorsque la taille d'échantillon tend vers l'infini, la statistique du rapport de vraisemblance et la statistique de score sont équivalentes. Choisir le modèle selon la statistique du score équivaut alors à choisir le modèle qui maximise la vraisemblance (ou qui minimise la quantité -2ℓ). Ainsi, pour ce nombre fixé de variables, cela va donner le modèle avec le meilleur AIC (et BIC). Par conséquent, on peut trouver le meilleur modèle, globalement, en minimisant le AIC (ou le BIC) en faisant varier le nombre de variables. Par contre, cela n'est pas nécessairement vrai pour une taille d'échantillon finie. Le meilleur modèle selon le critère score n'est pas nécessairement celui qui maximise la vraisemblance. Mais en pratique, cette approximation est plus que suffisante et on va procéder comme on a fait avec la procédure `reg`.

À la section précédente, nous avons inclus les 10 variables de base (14 avec les indicatrices pour les variables catégorielles) dans notre exemple d'envoi ciblé. Nous allons ici faire une recherche de type all-subset parmi ces 14 variables. Le code est dans le fichier logit5_selection_variables.sas.

```
proc logistic data=train;
model yachat(ref='0') = x1 x2 x3_1 x3_2 x4_1-x4_4 x5-x10 /
  selection=score best=1;
run;
```

Modèles de régression sélectionnés par le critère du score

Nombre de variables du score	Khi-2 Variables incluses dans le modèle
1	108.1349 x8
2	194.3293 x2 x8
3	236.1176 x2 x8 x10
4	260.7184 x2 x31 x7 x8
5	281.1804 x2 x31 x7 x8 x10
6	292.4871 x1 x2 x5 x7 x8 x10
7	297.7814 x1 x2 x32 x5 x7 x8 x10
8	300.4638 x1 x2 x32 x41 x5 x7 x8 x10
9	302.5042 x1 x2 x32 x43 x44 x5 x7 x8 x10
10	303.3246 x1 x2 x31 x32 x43 x44 x5 x7 x8 x10
11	304.0243 x1 x2 x31 x32 x43 x44 x5 x7 x8 x9 x10
12	304.5067 x1 x2 x31 x32 x41 x43 x44 x5 x7 x8 x9 x10
13	304.5551 x1 x2 x31 x32 x41 x43 x44 x5 x6 x7 x8 x9 x10
14	304.5902 x1 x2 x31 x32 x41 x42 x43 x44 x5 x6 x7 x8 x9 x10

Le meilleur modèle avec une seule variable, selon la statistique du score, est celui avec x8, le meilleur avec deux variables est celui avec x2 et x8, et ainsi de suite. Nous voulons ensuite choisir parmi ces 14 modèles, celui qui minimise le AIC ou le BIC. Le problème est que ces critères ne sont pas fournis (contrairement aux sorties de la procédure reg). La solution longue

consiste à ajuster chacun de ces modèles, à extraire le AIC et le BIC et à ainsi trouver le meilleur modèle. Mais le faire manuellement en spécifiant plusieurs modèles est trop long. La macro `logistic_aic_BIC_score`, qui se trouve dans le fichier `logit6_macro_all_subset.sas` ajuste tous ces modèles automatiquement.

```
%logistic_aic_BIC_score(yvariable=yachat,
                          xvariables=x1 x2 x3_1 x3_2 x4_1-x4_4 x5-x10,
                          dataset=train, minvar=1, maxvar=14);
```

Obs.	AIC	SBC	VariablesInModel
1	936.223	946.039	x8
2	826.567	841.290	x2 x8
3	766.187	785.818	x2 x8 x10
4	745.428	769.967	x2 x31 x7 x8
5	712.865	742.311	x2 x31 x7 x8 x10
6	696.755	731.109	x1 x2 x5 x7 x8 x10
7	689.616	728.878	x1 x2 x32 x5 x7 x8 x10
8	689.156	733.326	x1 x2 x32 x41 x5 x7 x8 x10
9	684.889	733.967	x1 x2 x32 x43 x44 x5 x7 x8 x10
10	686.260	740.245	x1 x2 x31 x32 x43 x44 x5 x7 x8 x10
11	687.445	746.338	x1 x2 x31 x32 x43 x44 x5 x7 x8 x9 x10
12	689.388	753.189	x1 x2 x31 x32 x41 x43 x44 x5 x7 x8 x9 x10
13	691.281	759.989	x1 x2 x31 x32 x41 x43 x44 x5 x6 x7 x8 x9 x10
14	693.215	766.832	x1 x2 x31 x32 x41 x42 x43 x44 x5 x6 x7 x8 x9 x10

On voit que le meilleur modèle selon le AIC a neuf variables, contre sept pour le BIC. Nous verrons plus loin, dans un tableau synthèse, comment auraient performé ces modèles s'ils avaient été utilisés pour cibler les clients restants.

5.7.2 Recherche séquentielle

Faire une recherche de tous les sous-modèles possibles devient impraticable lorsqu'il y a trop de variables en jeu. La procédure `logistic` permet aussi une recherche de type séquentielle clas-

sique. Ceci permet aussi d'utiliser la même approche en deux temps présentée au chapitre précédent. Dans un premier temps, on fait une recherche séquentielle pour sélectionner un nombre de variables qui sera assez petit afin qu'une recherche exhaustive de tous les sous-modèles soit possible. Dans un second temps, on fait cette recherche avec ces variables uniquement. Idéalement, on débute la sélection avec le modèle qui contient toutes les variables, soit `start=n` où $n = 104$ dans notre cas. Si on inclut tous les termes quadratiques et les termes d'interactions d'ordre deux, nous avons 104 variables potentielles : c'est trop pour une recherche exhaustive.

On peut faire une recherche descendante avec le test du score pour réduire le nombre de variables à 50, puis passer les variables sélectionnées à la procédure `logistic` et faire une recherche exhaustive approximative. Le modèle qui a le plus petit AIC, soit 585.194, est un modèle avec 27 variables. Le BIC mène à un modèle beaucoup plus parsimonieux qui inclut sept variables, pour une valeur de critère de 667.704.

5.7.3 Algorithme glouton et critères alternatifs avec `hpgenselect`

Nous avons vu au chapitre précédent que la procédure `glmselect` permet de faire une recherche de type séquentielle avec un critère autre que la valeur- p du test de Wald pour rajouter ou retrancher des variables explicatives du modèle final. Cette procédure est limitée à la régression linéaire, mais la procédure `hpgenselect` permet de faire une sélection de variables pour d'autres types de modèles, incluant la régression logistique.

Le code suivant fait une recherche séquentielle en ajoutant ou retranchant les variables selon leur valeur- p (`select=s1`), la seule méthode disponible pour l'instant. En revanche, le modèle final peut-être choisi selon d'autres critères.

```
proc hpgenselect data=train;
class x3(ref='3' split) x4(ref='5' split);
model yachat(ref='0')=x1|x2|x3|x4|x5|x6|x7|x8|x9|x10 @2
  x2*x2 x6*x6 x7*x7 x8*x8 x9*x9 x10*x10 /
link=logit distribution=binary;
selection method=stepwise(select=s1 choose=sbc);
run;
```

Avec le critère BIC, on obtient 12 variables tandis que `choose=aic` donne 13 variables (seule la variable `x41` est ajoutée). Il s'agit des mêmes variables que celles sélectionnées par une sélection séquentielle classique en prenant 0.05 comme critère d'entrée et de sortie.

5.8 Performance des différents modèles pour l'exemple des clients cibles

Nous allons conclure, pour l'instant, notre exemple dans cette section, en évaluant la performance de différentes stratégies. Le critère de performance sera le suivant : revenu net de la stratégie si elle était appliquée aux 100 000 clients restants. Pour chacun des 100 000 clients à catégoriser, nous allons calculer la quantité suivante :

- Si le client n'est pas ciblé pour l'envoi d'un catalogue par le modèle, alors le revenu est nul.
- Si le client est ciblé pour l'envoi d'un catalogue par le modèle et qu'il n'achète rien, le revenu est de -10\$ (le coût de l'envoi).
- Si le client est ciblé pour l'envoi d'un catalogue par le modèle et qu'il achète quelque chose, le revenu est de $(ymontant - 10)$ \$, c'est-à-dire, le montant qu'il dépense moins le 10\$ du coût de l'envoi.

Pour une stratégie donnée, chaque individu n'appartient qu'à une seule des catégories. Le revenu net de la stratégie est la somme des revenus pour les 100 000 clients. Parmi ces derniers, 23 179 auraient acheté si on leur avait envoyé le catalogue et ces clients auraient générés des revenus de 1 601 212\$. Si on enlève le coût des envois ($100\ 000 \times 10\$ = 1\ 000\ 000\$$), on obtient que la stratégie de référence permet un revenu net de 601 212\$.

Nous allons investiguer deux types de stratégies :

- 1) une basée sur la régression logistique seulement en utilisant le modèle pour prévoir l'achat et
- 2) une basée sur la combinaison de la régression logistique et la régression linéaire en utilisant un modèle pour prévoir l'achat et un autre pour prévoir le montant.

5.8.1 Stratégies en utilisant seulement la régression logistique

Dans ce cas, nous allons estimer la probabilité d'achat avec un modèle de régression logistique. Nous allons ensuite trouver le meilleur point de coupure, avec une matrice de gain adéquatement choisie, afin d'avoir une règle d'assignation optimale. Nous avons déterminé des modèles potentiels à la section précédente. De plus, nous avons déjà vu comment trouver le meilleur point de coupure en spécifiant une matrice de gain, afin de maximiser le gain moyen à partir de la matrice de gain du tableau 5.6. Nous allons donc trouver le meilleur point de coupure pour quelques-uns des modèles choisis à la section précédente, pour ensuite évaluer le revenu net de ces modèles.

Il faut encore une fois bien comprendre qu'en pratique, on ne pourrait pas faire cette comparaison, car on ne sait pas d'avance si les clients futurs vont acheter ou non. Mais dans cet exemple, les variables *yachat* et *ymontant* sont fournies pour ces 100 000 clients afin qu'on puisse voir ce qui se serait passé avec les différentes stratégies.

TABLE 5.7 – Résumé des caractéristiques des modèles logistiques avec (a) référence, soit l'envoi sans sélection à tous les clients; (b) 10 variables de base sans sélection; (c) toutes les variables, incluant les termes quadratiques et les interactions d'ordre 2; (d) sélection séquentielle classique avec critère d'entrée et de sortie à 0.05; (e) idem que (d), mais avec meilleur modèle selon le AIC; (f) idem que (d), mais avec meilleur modèle selon le BIC; (g) recherche exhaustive avec 50 variables sélectionnées par recherche séquentielle et modèle final sélectionné selon le BIC; (h), idem mais sélection avec AIC. Les points de coupure optimaux ont été déterminés par validation croisée sur l'échantillon d'apprentissage, tandis que la performance du modèle (sensibilité, taux de faux positifs et taux de bonne classification) ont été calculés à partir de l'échantillon test de 100 000 individus.

modèle	# variables	point de coupure	sensibilité	taux de faux positifs	taux de bonne classification	revenu net
(a)	—	—	100	76.8	23.2	601212
(b)	14	0.12	89	56.2	70.9	940569
(c)	104	0.08	85.8	52.6	74.6	937150
(d), (e)	13	0.14	85.7	49.1	77.5	969350
(f)	12	0.19	81	44.7	80.4	943935
(g)	8	0.16	86	48.1	78.3	985069
(h)	28	0.15	83.5	47.4	78.8	952672

La stratégie de référence est celle qui consiste à envoyer le catalogue aux 100 000 clients sans les sélectionner. Le tableau qui suit montre des statistiques pour les variables *ymontant* et *yachat* pour les 100 000 clients à scorer. Le tableau 5.7 résume la performance des différentes stratégies basées exclusivement sur le modèle logistique.

Table :

Nous avons vu plus tôt, qu'avec les 10 variables de base, le meilleur point de coupure est de 0.12. En utilisant cette stratégie sur les 100 000 clients, le revenu net aurait été de 940 569\$. C'est une énorme amélioration, de plus de 56%, par rapport à la stratégie de référence qui consiste à envoyer le catalogue à tout le monde (revenu net de 601 212\$). Si on inclut tous les termes quadratiques et les termes les interactions d'ordre deux (104 variables en tout), le revenu net est inférieur avec une valeur de 937 150\$. Ici, le modèle est trop complexe et surajusté. Si on fait une sélection de variables (quasi méthodes sont présentées), suivie de la détermination du meilleur point de coupure, on fait alors toujours mieux qu'avec le modèle incluant les 10 variables de base seulement. L'approche qui aurait fait le mieux est la recherche séquentielle pour réduire le nombre de variables considérées à 50, suivi d'une recherche exhaustive pour trouver le modèle avec le plus petit BIC : cette approche aurait généré 985 069\$ de revenus nets. Il s'agit d'un gain de 4.7% par rapport au

modèle avec les 10 variables de base.

5.8.2 Stratégies alternatives

Nous venons tout juste d'étudier des stratégies qui consistent essentiellement, à estimer $P(yachat = 1)$ et un point de coupure afin de décider à qui envoyer le catalogue en partant du postulat que tous les clients dépensent le même montant; le tout est basé uniquement sur la régression logistique. Le revenu moyen peut être estimé à partir de l'équation

$$E(ymontant) = E(ymontant | yachat = 1)P(yachat = 1),$$

c'est-à-dire, la moyenne du montant dépensé est égale à la moyenne du montant dépensé étant donné qu'il y a eu achat, fois la probabilité qu'il ait eu achat. Une autre stratégie possible consiste donc à développer deux modèles : un pour $E(ymontant | yachat = 1)$ et un autre pour $P(yachat = 1)$ et à les combiner afin d'obtenir des prévisions du montant dépensé.

Nous avons déjà développé des modèles de régression linéaire pour $E(ymontant | yachat = 1)$ au chapitre précédent et nous venons de développer des modèles de régression logistique pour $P(yachat = 1)$ dans ce chapitre. Nous avons donc tous les ingrédients pour implanter cette stratégie.

Nous allons cibler les clients dont la prévision du montant dépensé est plus grande que 10\$ (le coût de l'envoi du catalogue).

Les possibilités de modèles sont nombreuses. Par exemple, si on a cinq modèles potentiels pour $E(ymontant | yachat = 1)$ et cinq pour $P(yachat = 1)$, il y a 25 combinaisons possibles. Ici, nous allons seulement présenter les résultats pour deux combinaisons :

- 1) pour $ymontant$, nous allons utiliser les variables choisies par `glmselect` avec les options `select=aic`, `choose=bic`, tandis que pour $yachat$, nous allons utiliser les variables choisies par la procédure séquentielle suivie d'une recherche exhaustive avec le critère BIC
- 2) à la fois pour $ymontant$ et $yachat$, nous allons utiliser les variables choisies en faisant une sélection séquentielle classique ($tests-t$) avec critères d'entrée et de sortie fixés à 0.05.

Pour obtenir les prévisions, nous allons estimer conjointement les modèles pour $E(ymontant | yachat = 1)$ et pour $P(yachat = 1)$ avec un modèle Tobit de type 2 (aussi appelé modèle Heckit), dont une brève description est donnée dans la section 5.8.3. L'avantage de l'estimation simultanée est que l'on a pas à sélectionner le point de coupure, puisque l'on enverra le catalogue uniquement si le montant prédit pour $E(ymontant)$ (non-conditionnel) est supérieur à 10\$. Les résultats du modèle Tobit sur l'échantillon de validation sont rapportés dans le tableau 5.8.

Il s'avère qu'on aurait eu des performances semblables aux stratégies basées uniquement sur la régression logistique vues à la sous-section précédente. La première combinaison aurait tout de

TABLE 5.8 – Matrice de gain pour l'envoi de catalogue avec des modèles Tobit de type II : sensibilité, taux de faux positifs et de bonne classification et gain net de la stratégie.

modèle	sensibilité	FP (%)	bonne classification	revenu net (%)
(1)	88.3	50.9	76.1	997 238
(2)	86.3	49.9	76,9	977 422

même produit un revenu net de 997 238\$, supérieur au revenu net de 985 069\$, qui était le meilleur trouvé à la sous-section précédente.

Pour conclure cet exemple, il s'avère donc que la régression logistique permet d'effectuer un bon ciblage des clients potentiels afin de maximiser les revenus. L'approche générale consistant à obtenir des prévisions pour $P(yachat = 1)$ et ensuite trouver le meilleur point de coupure est très générale. D'autres types de modèles (arbre de classification, forêt aléatoire, réseau de neurones) pourraient être utilisés à la place de la régression logistique.

Nous reviendrons une dernière fois sur cet exemple dans le chapitre traitant des données manquantes. Nous verrons alors comment procéder si des valeurs manquantes sont présentes dans les variables explicatives.

5.8.3 Modèle Tobit de type 2

Cette partie est plus technique et peut être omise.

Il ne serait pas justifié d'ajuster séparément les deux modèles pour $E(ymontant | yachat = 1)$ et $P(yachat = 1)$ et de calculer les prévisions en prenant le produit : $E(ymontant | yachat = 1)P(yachat = 1)$. Cela provient du fait que le modèle pour $E(ymontant | yachat = 1)$ aurait été estimé seulement avec les clients qui ont acheté quelque chose et qu'ensuite on l'appliquerait (au moment de calculer les prévisions) à la fois aux clients qui vont acheter et à ceux qui ne vont pas acheter. Il y a donc un biais de sélection dans l'échantillon qui a servi à ajuster le modèle au départ. Une manière de contourner ce problème est d'ajuster conjointement les deux modèles. Le modèle de Tobit de type 2 permet de faire cela. Ce modèle est basé sur l'hypothèse que les deux variables observées (Y_1 et Y_2) proviennent de deux variables latentes non observées (Y_1^* et Y_2^*), où

$$Y_1 = \begin{cases} 1 & \text{si } Y_1^* \geq 0, \\ 0 & \text{si } Y_1^* < 0, \end{cases} \quad Y_2 = \begin{cases} Y_2^* & \text{si } Y_1^* \geq 0, \\ 0 & \text{si } Y_1^* < 0. \end{cases}$$

Dans notre exemple, Y_1 correspond à $yachat$ et Y_2 à $ymontant$. Ce qui lie les deux équations est le fait qu'on suppose que les variables sont binormales : les deux termes d'erreur sont de loi normale

et sont corrélés, $\boldsymbol{\varepsilon} \sim \mathcal{N}_2(\mathbf{0}_2, \boldsymbol{\Sigma})$. Les variables dépendantes observées sont :

$$\begin{aligned} Y_1^* &= \beta_{01} + \beta_{11}X_{11} + \cdots + \beta_{1p}X_{p1} + \varepsilon_1 \\ Y_2^* &= \beta_{02} + \beta_{12}X_{12} + \cdots + \beta_{1p}X_{q2} + \varepsilon_2 \end{aligned}$$

Notez que les variables explicatives ne sont pas nécessairement les mêmes dans les deux équations. En estimant conjointement les deux équations, on élimine le biais de sélection mentionné plus haut. La procédure `q1im` permet d'estimer ce modèle. Cependant, `q1im` ne fait pas de sélection de variables. Le choix des variables doit être fait avant avec les méthodes qu'on a vues. De plus, pour être précis, le modèle Tobit ajuste un modèle probit et non logistique à la variable binaire.

5.9 Extensions du modèle de régression logistique à plus de deux catégories

Supposons que la variable Y que vous cherchez à modéliser est une variable catégorielle pouvant prendre trois valeurs ou plus. Voici quelques exemples :

- Destination de vacances l'année dernière (Québec, États-Unis, ailleurs).
- Si les élections avaient lieu aujourd'hui au Québec, pour quel parti voteriez-vous (PLQ, PQ, CAQ, QS).
- Combien de fois êtes-vous allé au cinéma l'année dernière : moins de cinq fois (1), entre cinq et 10 fois (2), ou plus de 10 fois (3).
- Quelle importance accordez-vous au service après-vente ? Un parmi « pas important » (1), « peu important » (2), « moyennement important » (3), « assez important » (4), « très important » (5).

Dans les deux premiers exemples, la variable réponse Y est nominale (elle n'a pas d'ordre) alors qu'elle est ordinaire dans les deux derniers. Pour une variable ordinaire, le modèle logit multinomial peut être utilisé mais il existe d'autres possibilités comme le modèle logit cumulé. Nous couvrirons ces deux modèles.

5.9.1 Régression logistique multinomiale

Afin de simplifier la notation, on suppose qu'il y a une seule variable explicative X à disposition et que la variable Y représente trois catégories, une parmi 0, 1 et 2.

En régression logistique, Y est une variable binaire qui vaut soit 0, soit 1 et la probabilité de succès est

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i, \quad p_i = P(Y_i = 1 | X_i) = \text{expit}(\beta_0 + \beta_1 X_i).$$

Dans ce modèle logistique, $\ln(p_i) - \ln(1 - p_i) = \ln\{\mathbb{P}(Y_i = 1 | X_i)\} - \ln\{\mathbb{P}(Y_i = 0 | X_i)\}$ peut être vu comme étant le logit de la catégorie 1 en utilisant 0 comme catégorie de référence. Le modèle logistique multinomial procède de même en fixant une catégorie de référence et en modélisant le logit de chacune des autres catégories par rapport à la catégorie de référence. Avec $K+1$ catégories et en choisissant la catégorie 0 comme référence, le modèle devient

$$\ln\left(\frac{p_{i1}}{p_{i0}}\right) = \beta_{01} + \beta_{11}X_i, \dots, \ln\left(\frac{p_{iK}}{p_{i0}}\right) = \beta_{0K} + \beta_{1K}X_i,$$

où $p_{ik} = \mathbb{P}(Y_i = k | X_i)$ ($k = 0, \dots, K$). Comme en régression logistique, on peut facilement exprimer ce modèle en termes des différentes probabilités,

$$p_{i0} = \mathbb{P}(Y_i = 0 | X_i) = \frac{1}{1 + \sum_{j=1}^K \exp(\beta_{0j} + \beta_{1j}X_i)}$$

$$p_{ik} = \mathbb{P}(Y_i = k | X_i) = \frac{\exp(\beta_{0k} + \beta_{1k}X_i)}{1 + \sum_{j=1}^K \exp(\beta_{0j} + \beta_{1j}X_i)}, \quad k = 1, \dots, K.$$

On voit facilement que la somme des probabilités égale 1, c'est-à-dire $p_{i0} + \dots + p_{iK} = 1$. En fait, le modèle logit multinomial ne fait que combiner plusieurs logit dans un seul modèle. L'interprétation des paramètres se fait comme en régression logistique sauf qu'il faut y aller équation par équation.

Destination vacances. Le fichier `logit6.sas7bdat` contient 100 observations obtenues par voie de sondage auprès d'adultes âgés de 18 à 45 ans. Le fichier contient les réponses aux questions suivantes :

- `y1` : quelle a été votre destination vacances l'année dernière : Québec (0), États-Unis (1) ou ailleurs (2) ?
- `y2` : combien de fois êtes-vous allé au cinéma l'année dernière : moins de 5 fois (1), entre 5 et 10 fois (2), ou plus de 10 fois (3).
- `x` : âge (en année) du répondant.

Nous allons modéliser la destination vacance Y_1 à l'aide d'une régression logistique multinomiale avec l'âge comme variable explicative.

Variable d'analyse : x x					
y1	N	Moyenne	Ec-type	Minimum	Maximum
0	47	26.47	5.84	18.00	40.00
1	32	33.00	8.23	19.00	44.00
2	21	37.29	6.20	23.00	44.00

On voit que les gens qui ont passé leurs vacances au Québec ($Y_1 = 0$) ont 26.5 ans en moyenne. Ils sont plus jeunes que ceux qui ont passé leurs vacances aux États-Unis (âge moyen de 33 ans).

Finalement, ceux dont la destination vacances était ailleurs sont les plus vieux avec une moyenne de 37.3 ans.

Pour le modèle logit multinomial, nous allons prendre $Y_1 = 0$ comme catégorie de référence. Les commandes sont

```
proc logistic data=multi.logit6 ;
model y1(ref='0') = x / clparm=pl clodds=pl expb link=glogit;
run;
```

L'option link=glogit spécifie le type de fonction de lien, ici celle du modèle logistique multinomial.

Profil de réponse		
Valeur ordonnée	y1	Fréquence totale
1	0	47
2	1	32
3	2	21

Les logits modélisés utilisent y1=0 comme catégorie de référence.

Etat de convergence du modèle
Critère de convergence (GCONV=1E-8) respecté.

Statistique d'ajustement du modèle			
Critère	Constante uniquement	Constante et Covariables	
AIC	213.443	183.448	
SC	218.653	193.868	
-2 Log L	209.443	175.448	

Test de l'hypothèse nulle globale : BETA=0				
Test	khi-2	DDL	Pr > khi-2	
Rapport de vrais	33.9955	2	<.0001	
Score	30.2091	2	<.0001	
Wald	23.4463	2	<.0001	

Analyse des effets Type 3				
		Khi-2		
Effet	DDL	de Wald	Pr > khi-2	
x	2	23.4463	<.0001	

Analyse des valeurs estimées du maximum de vraisemblance							
Paramètre	y1	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	1	1	-4.0970	1.0825	14.3231	0.0002	0.017
Intercept	2	1	-7.9793	1.7231	21.4448	<.0001	0.000
x	1	1	0.1253	0.0351	12.7488	0.0004	1.133
x	2	1	0.2233	0.0496	20.2942	<.0001	1.250

Valeurs estimées du paramètre et intervalle de confiance de vraisemblance de profil					
Paramètre	y1	Estimation	Intervalle de confiance à 95%		
Intercept	1	-4.0970	-6.3564		-2.0802
Intercept	2	-7.9793	-11.7817		-4.9473
x	1	0.1253	0.0596		0.1982
x	2	0.2233	0.1343		0.3306

Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil				
Effet	y1	Unité	Estimation	Intervalle de confiance à 95%
x	1	1.0000	1.133	1.061 1.219
x	2	1.0000	1.250	1.144 1.392

Comme il y a trois catégories pour la variable dépendante, il y a deux équations pour le modèle ajusté,

$$\ln \left\{ \frac{P(Y_{1i} = 1 | X_i)}{P(Y_{1i} = 0 | X_i)} \right\} = -4.10 + 0.13X_i, \quad \ln \left\{ \frac{P(Y_{1i} = 2 | X_i)}{P(Y_{1i} = 0 | X_i)} \right\} = -7.98 + 0.22X_i.$$

Plus l'âge du répondant augmente, plus la probabilité qu'il ait passé ses vacances aux États-Unis par rapport au Québec augmente. En fait, pour chaque augmentation de 1 de l'âge, le rapport des cote pour $Y_1 = 1$ par rapport à $Y_1 = 0$ est multipliée par $1.133 = \exp(0.1253)$. Cette valeur est donnée dans la dernière colonne du tableau. De plus, cet effet est significatif car la valeur- p est inférieure à 10^{-4} .

De même, plus l'âge du répondant augmente, plus la probabilité qu'il ait passé ses vacances ailleurs qu'aux États-Unis ou au Québec par rapport au Québec augmente. En fait, pour chaque

augmentation de l'âge d'un an, le rapport des cote pour $Y_1 = 1$ par rapport à $Y_1 = 0$ est multiplié par 1.25. Cet effet est également statistiquement significatif.

Nous venons donc de comparer chacune des catégories $Y_1 = 1$ et $Y_1 = 2$ à la catégorie de référence $Y_1 = 0$. Pour une comparaison directe entre $Y_1 = 1$ et $Y_1 = 2$, il suffit de changer la catégorie de référence et de resoumettre le programme **SAS**.

5.9.2 Régression logistique cumulative à cotes proportionnelles

Si les modalités de la réponse sont ordinaires, la régression logistique multinomiale est toujours appropriée. Il peut néanmoins être préférable d'utiliser un modèle qui utilise l'ordre des modalités pour obtenir un modèle plus facile à interpréter et plus parcimonieux. Le modèle de régression logistique cumulative à cotes proportionnelles est une simplification sous l'hypothèse que les cotes sont proportionnelles.

Supposons que la variable Y est ordinaire et peut prendre les $K + 1$ valeurs ordonnées de la plus petite à la plus grande selon les catégories de Y ($0, 1, 2, \dots, K$). Supposons que l'on dispose de p variables explicatives X_1, \dots, X_p .

Soit $p_{ik} = P(Y_i = k | X_{i1}, \dots, X_{ip})$ ($k = 0, 1, \dots, K$) la probabilité que Y_{ik} prenne la valeur k . On dénote

$$S_{ij} = \sum_{k=j}^K p_{ik} = P(Y_i > j - 1 | X_{i1}, \dots, X_{ip}), \quad j = 1, \dots, K.$$

La quantité S_{ij} est la probabilité que Y_i soit plus grand ou égal à j ; S_{i0} est égal à 1 et $S_{iK} = P(Y_i = K | X_{i1}, \dots, X_{ip}) = p_{iK}$.

Le modèle logistique cumulé spécifie que

$$\ln\left(\frac{S_{ij}}{1 - S_{ij}}\right) = \beta_{0j} + \beta_1 X_{i1} + \dots + \beta_p X_{ip}, \quad j = 1, \dots, K.$$

Il y a donc K équations logistiques. Les paramètres quantifiant les effets des variables explicatives, β_1, \dots, β_p sont les mêmes pour chacune des log-cotes, mais il y a une ordonnée à l'origine différente par log de rapport de cote. Par conséquent, pour modéliser une variable ordinaire Y ayant $K + 1$ valeurs possibles avec p variables explicatives, le modèle cumulatif logistique utilise $p + K$ paramètres. Le modèle logit multinomial, qui peut également être utilisé pour les données ordinaires, utilise $K \cdot (p + 1)$ paramètres. Le modèle multinomial ordonné est donc plus parcimonieux et, pour autant qu'il soit approprié, mènera à des estimations des paramètres plus précises qu'avec le modèle de régression logistique multinomiale. Les deux modèles sont identiques au modèle de régression logistique si la variable ordinaire a deux modalités.

La cote $S_{ij}/(1 - S_{ij})$ mesure à quel point il est plus probable que Y_i prenne une valeur plus grande ou égale à j par rapport à une valeur plus petite que j , viz.

$$\frac{S_{ij}}{1 - S_{ij}} = \exp(\beta_0 j + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}).$$

Dans cet exemple, aucune fonction autre qu'additive en X , ni aucune interaction n'est présente. Si le paramètre β_j est positif, cela indique que plus X_j prend une valeur élevée, plus la variable Y a tendance à prendre aussi une valeur élevée. Inversement, si le paramètre β_j est négatif, cela indique que plus X_j prend une valeur élevée, plus la variable Y a tendance à prendre une valeur basse. Plus précisément, pour chaque augmentation d'une unité de X_j , la cote $S_k/(1 - S_k)$ est multipliée par $\exp(\beta_j)$, peu importe la valeur de Y . Ainsi, la cote d'être dans une catégorie plus élevée, par rapport à une catégorie moins élevée, est multipliée par $\exp(\beta_j)$. En terme de probabilités cumulées d'excéder k ,

$$S_{ik} = P(Y_i \geq k | X_{i1}, \dots, X_{ip}) = \text{expit}(\beta_0 k + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}), \quad j = 1, \dots, K.$$

En utilisant ces expressions, on peut obtenir la probabilité de chaque catégorie,

$$P(Y_i = k | X_{i1}, \dots, X_{ip}) = P(Y_i \geq k | X_{i1}, \dots, X_{ip}) - P(Y_i \geq k+1 | X_{i1}, \dots, X_{ip}) = S_k - S_{k+1}.$$

Nous considérons maintenant la variable Y_2 du fichier `logit6.sas7bdat`, qui donne le nombre de visites au cinéma. Pour cet exemple, nous allons chercher à modéliser Y_2 à l'aide de X (âge) en utilisant le modèle multinomial cumulé à cotes proportionnelles.

y2	n	moyenne	écart-type	minimum	maximum
1	44	33.5	7.23	18	44
2	38	30.4	8.16	18	44
3	18	25.1	6.58	18	39

Ainsi, les répondants qui sont allés moins de cinq fois au cinéma ont en moyenne 33.5 ans, ceux qui sont allés entre cinq et 10 fois ont 30.4 ans en moyenne, et ceux qui sont allés plus de 10 fois ont 25.1 ans en moyenne. Il y a une progression et on voit que les répondants plus jeunes vont plus souvent au cinéma.

Nous utilisons exactement le même programme que pour une régression logistique habituelle. Si la variable Y prend plus de deux valeurs, **SAS** utilisera automatiquement le modèle de régression multinomiale cumulé.

```
proc logistic data=multi.logit6 descending;
model y2 = x / clparm=pl clodds=pl expb;
run;
```

L'option descending impose la paramétrisation discutée précédemment. Sans cette option, ce serait plutôt les probabilités de prendre une valeur plus petite, par rapport à une plus grande qui serait modélisée. Le modèle est le même, mais les signes des paramètres des variables explicatives seraient inversés.

Informations sur le modèle		
Table	MULTI.LOGIT6	
Variable de réponse	y2	y2
Nombre de niveaux de réponse	3	
Modèle	logit cumulé	
Technique d'optimisation	Score de Fisher	

Profil de réponse		
Valeur ordonnée	Fréquence y2	totale
1	3	18
2	2	38
3	1	44

Les probabilités modélisées sont cumulées sur les valeurs ordonnées inférieures.

Etat de convergence du modèle		
Critère de convergence (GCONV=1E-8) respecté.		

Test de score pour l'hypothèse des cotes proportionnelles		
khi-2	DDL	Pr > khi-2
1.2813	1	0.2577

Statistique d'ajustement du modèle			
Critère	Constante uniquement	Constante et Covariables	
AIC	211.515	199.796	
SC	216.726	207.612	
-2 Log L	207.515	193.796	

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	13.7191	1	0.0002
Score	13.1022	1	0.0003
Wald	12.6987	1	0.0004

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept	3	1	1.1597	0.7713	2.2607	0.1327
Intercept	2	1	3.1191	0.8290	14.1559	0.0002
x		1	-0.0916	0.0257	12.6987	0.0004
						0.912

Valeurs estimées du paramètre et intervalle de confiance de vraisemblance de profil				
Paramètre		Estimation	Intervalle de confiance à95%	
Intercept	3	1.1597	-0.3508	2.7023
Intercept	2	3.1191	1.5215	4.8117
x		-0.0916	-0.1438	-0.0423

Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil				
Effet	Unité	Estimation	Intervalle de confiance à95%	
x	1.0000	0.912	0.866	0.959

Avant toute chose, il faut s'assurer que le modèle est approprié. Rappelez-vous que l'une des hypothèses de ce modèle est que les effets des variables explicatives sont les mêmes pour chaque équation. Le tableau « Test de score pour l'hypothèse des cotes proportionnelles » est un test de l'hypothèse nulle

- \mathcal{H}_0 : l'effet de chaque variable est le même pour les K logit du modèle multinomial logistique, soit $\beta_{11} = \dots = \beta_{1K}, \dots, \beta_{p1} = \dots = \beta_{pK}$.

Une très petite valeur- p (rejet de \mathcal{H}_0) pour ce test serait une indication que le modèle de régression multinomiale ordinaire n'est pas approprié. Comme la valeur- p est 0.2577 ici, on ne rejette pas \mathcal{H}_0 et il n'y a pas lieu de douter de cette hypothèse. On peut donc aller de l'avant et interpréter le modèle.

Ici, l'effet estimé de l'âge (X) est -0.0916 et ce paramètre est significativement différent de zéro (valeur- p de 0.0004). Rappelez-vous que Y_2 représente le nombre d'entrées au cinéma dans la dernière année.

Ainsi, plus l'âge augmente, plus Y_2 a tendance à prendre une petite valeur, c'est-à-dire, plus la personne a tendance à aller moins souvent au cinéma. Plus précisément, lorsque l'âge augmente de 1, la cote d'être dans une catégorie plus élevée de Y_2 , par rapport à une catégorie plus basse, est multipliée par 0.912 (la cote diminue donc et aussi la probabilité d'être dans une catégorie plus élevée).

Chapitre 6

Analyse de survie

6.1 Introduction

Le but de cette section est d'étudier l'effet de variables explicatives sur le temps de survie. Plusieurs mécanismes de survie peuvent impacter la survie, de façon aléatoire ou pas. Considérons l'exemple d'une étude sur le chômage dû à la crise du coronavirus. On s'intéresse à tous ceux qui étaient en recherche active d'emploi entre mars et juin ; seuls ceux qui étaient au chômage durant cette période seront considérés. Certaines personnes seront déjà au chômage en avril et donc leur durée de chômage sera plus longue (troncature à gauche). Lors de notre suivi, d'autre mentionneront avoir trouvé un emploi lors de notre appel, mais ne pourront nous renseigner sur la date exacte de leur embauche (censure à gauche) ; cette dernière précèdera notre prise de contact, mais nous est inconnue. D'autres personnes seront toujours au chômage en juin à la fin de l'étude et on ignorera le nombre réel de mois passés au chômage (censure à droite). Enfin, certaines personnes cesseront de chercher activement un emploi et donc quitteront l'étude. Tous ces mécanismes (complexes) peuvent être dictés par certaines covariables (employabilité, découragement) et être aléatoires ou pas. Pour estimer le taux de chômage, il faudra prendre en compte les mécanismes de survie dans notre modèle. On se concentrera sur le cas simple des données censurées à droite de façon aléatoire.

6.1.1 Exemple du temps d'abonnement

Une entreprise oeuvrant dans le secteur des télécommunications s'intéresse aux facteurs influençant le temps qu'un client reste abonné à son service de téléphone cellulaire. Des données provenant d'un échantillon de clients se trouvent dans le fichier `survival1.sas7bdat`, qui contient les variables suivantes :

- t : temps (en semaines) que le client est resté abonné au service de téléphone cellulaire. Il s'agit du vrai temps si le client n'est plus abonné et d'un temps censuré à droite si le client est toujours abonné.
- **censure** : variable binaire qui indique si la variable t est censurée (1 si le client est toujours abonné) ou non (0, la variable t est la durée finale de l'abonnement).
- **age** : âge du client au début de l'abonnement.
- **sexe** : sexe du client, soit femme (1), soit homme (0).
- **service** : nombre de services en plus du cellulaire auquel le client est abonné parmi internet, téléphone fixe, télévision (câble ou antenne parabolique).
- **region** : région où habite le client en ce moment (valeurs entre 1 et 5).

6.1.2 Contexte

On s'intéresse au temps avant qu'un événement survienne. On observe chaque sujet jusqu'à ce que l'une des deux choses suivantes se produise : l'événement survient avant la fin de la période d'observation ou bien l'étude se termine et l'événement n'est toujours pas survenu. Dans l'exemple, l'événement correspond au fait d'interrompre son abonnement. On dispose donc d'une variable « temps », que l'on nomme T , pour chaque individu qui est soit censurée soit non censurée. Si l'individu a expérimenté l'événement avant la fin de la période d'observation, la valeur de T n'est pas censurée. Si l'événement n'est toujours pas survenu à la fin de la période d'observation, la valeur de T est censurée. Pour chaque individu, on dispose également d'un ensemble de variables explicatives X_1, \dots, X_p . Pour l'instant, supposons que les valeurs de ces variables sont fixes dans le temps mais on reviendra plus loin au cas où leurs valeurs peuvent varier dans le temps. Bien que le terme analyse de survie semble implicitement référer à la santé, de nombreux autres exemples sont envisageables

- temps qu'un client demeure abonné à un service offert par notre compagnie.
- temps de survie d'un individu après avoir été diagnostiqué avec un certain type de cancer.
- ancienneté d'un travailleur au service d'une compagnie.
- durée de vie d'une franchise.
- temps avant la faillite d'une entreprise (ou d'un particulier).
- temps avant le prochain achat d'un client.
- temps durant lequel un(e) employé(e) est au chômage.

Si aucune observation n'est censurée, c'est-à-dire, si on a observé le « vrai » temps pour chaque sujet, on pourrait alors simplement modéliser T en incluant des covariables dans les paramètres de vraisemblance d'une loi positive (par exemple, une régression log-linéaire). En revanche, si des observations sont censurées dans l'échantillon, leur omission biaiserait l'analyse.

Ce chapitre se veut une introduction à l'analyse de données de survie. Comme le développement de la théorie de l'analyse de survie est assez complexe (plus encore que celle de la régression linéaire ou logistique), on s'intéressera ici uniquement aux principes de base afin d'être en mesure

d'appliquer les méthodes et de bien interpréter les résultats. Plusieurs extensions sont également possibles. Un survol de ces dernières sera effectué dans des sections plus loin.

Il existe deux grandes approches pour analyser des données de survies :

- i) nonparamétrique ou semi-paramétrique : estimateur de Kaplan-Meier, modèle de Cox (à risques proportionnels).
- ii) paramétrique : modèle paramétrique avec loi continue (Weibull, log-normal, log-logistique, gamma).

Nous discuterons seulement de la première approche dans ce chapitre. Le tableau suivant fait une analogie entre ce que nous ferons dans ce chapitre et des méthodes que vous connaissez.

réponse Y	résumé descriptif	comparaison de deux groupes	modèle général
continue	moyenne	test- t pour deux échantillons	régression linéaire
binaire	proportion	test d'indépendance du khi-deux	régression logistique
temps de survie (censure à droite)	fonction de survie temps de survie médian	test log-rang test de Wilcoxon généralisé (Gehan)	modèle de Cox

La structure de données de base que l'on doit avoir pour travailler en **SAS** (et avec la plupart des autres logiciels) est la suivante :

- 1) une variable temps, T .
- 2) une variable C (censure), qui vaut un si l'observation est censurée et zéro sinon.
- 3) d'autres variables explicatives X_1, \dots, X_p , une par colonne.

6.2 Fonctions de survie et de risque

Un des éléments de base d'une analyse de survie (*survival analysis*) est la **fonction (ou courbe) de survie**. Soit $F(t) = P(T \leq t)$ la fonction de répartition du temps de survie t et $f(t) = d/dtF(t)$. La fonction de survie est

$$S(t) = P(T > t) = 1 - F(t)$$

et donne la probabilité que le temps de survie soit supérieur à t . On verra plus loin comment estimer cette fonction avec un échantillon et comment tester l'égalité de deux (ou plusieurs) fonctions de survie.

La **fonction de risque** (en anglais, *hazard*) est

$$h(t) = \frac{f(t)}{S(t)}$$

où $f(t)$ est la fonction de densité (pour T continu) ou de masse pour T discret. Dans le cas discret où le temps peut seulement prendre les valeurs $0, 1, 2, \dots$, la fonction de risque est donc simplement la probabilité que l'événement survienne au temps t , étant donné qu'il n'était pas survenu avant, $P(T = t | T > t) = P(T = t)/P(T > t) = f(t)/S(t)$; c'est une probabilité conditionnelle. Dans le cas général, la fonction de risque est nécessairement positive mais peut prendre des valeurs supérieures à un. On ne peut donc pas, à strictement parler, la voir comme une probabilité et c'est pourquoi on parle plutôt de risque. En fait, cette fonction mesure le risque instantané que l'événement survienne au temps t , étant donné qu'il n'était pas survenu avant.

Cette fonction est importante car il s'agit de celle que nous allons modéliser avec le modèle de régression de Cox. Si, en régression logistique, on modélise le logarithme des cotes, on modélise plutôt la fonction de risque en analyse de survie. Les fonctions de survie et de risque sont intimement reliées et

$$h(t) = -\frac{d \ln\{S(t)\}}{dt}, \quad S(t) = \exp\left\{-\int_0^t h(u)du\right\}.$$

Ainsi, si on connaît la fonction de survie, on peut retrouver la fonction de risque et vice-versa. Par conséquent, un modèle pour la fonction de survie spécifie une fonction de risque (et vice-versa).

6.3 Estimation d'une courbe de survie et de risque

L'estimateur nonparamétrique le plus couramment utilisé pour l'estimation de la fonction de survie en présence de censure à droite est l'estimateur de Kaplan-Meier. De plus, cette méthode est nonparamétrique en ce sens qu'on ne suppose aucun modèle et qu'on suppose uniquement que la censure est non-informative. Dans **SAS**, la procédure `lifetest` fournit l'estimation de Kaplan-Meier de la courbe de survie et permet aussi d'estimer la fonction de risque correspondante.

Si l'échantillon ne contient aucune observation censurée (on a des temps exacts pour tous les sujets), l'estimateur de Kaplan-Meier de la fonction de survie à un temps t donné est alors simplement la proportion des observations dans l'échantillon qui possède un temps de survie supérieur à t . Par convention, on considère qu'une observation censurée à droite faisait partie de l'ensemble d'observations à risque au temps de censure observé.

On considère l'exemple des temps d'abonnement pour illustrer le concept; les commandes **SAS** sont dans `survival1_fonction_survie.sas`. L'estimation de la fonction de survie selon la méthode Kaplan-Meier est obtenue grâce aux commandes suivantes :

```
proc lifetest data=multi.survival1 method=km
    plots=survival(cl nocensor);
time t*censure(1);
run;
```

Le premier tableau renvoyé par **SAS** contient une ligne par observation et permet de lire l'estimé de la fonction de survie. Par exemple, on voit que l'estimation de la probabilité que le temps d'abonnement soit supérieur à 30 semaines est $\hat{S}(30) = 0,986$.

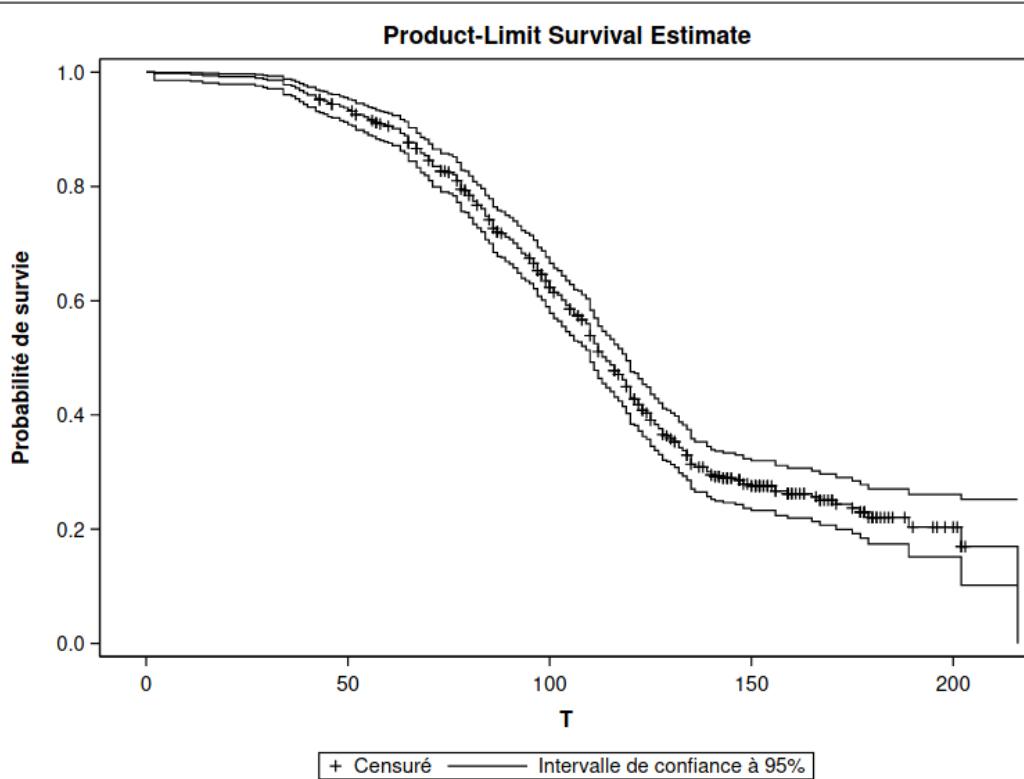
Valeurs estimées de survie de Kaplan-Meier					
T	Survie	Echec	Erreur type de survie	Nombre d'échecs	Nombre restant
0.000	1.0000	0		0	500
2.000	0.9980	0.00200	0.00200	1	499
11.000	0.9960	0.00400	0.00282	2	498
14.000	0.9940	0.00600	0.00345	3	497
18.000	0.9920	0.00800	0.00398	4	496
27.000	0.9900	0.0100	0.00445	5	495
29.000	0.9880	0.0120	0.00487	6	494
30.000	0.9860	0.0140	0.00525	7	493
34.000	.	.		8	492
⋮					
201.000	*	.	.	331	6
202.000	0.1697	0.8303	0.0388	332	5
202.000	*	.	.	332	4
202.000	*	.	.	332	3
203.000	*	.	.	332	2
216.000	.	.	.	333	1
216.000	0	1.0000	.	334	0

La sortie contient également un tableau contenant les quartiles. On utilise généralement le temps de survie médian (au lieu de la moyenne) dans ce type d'étude. Ici, l'estimé du temps de survie médian est de 114 semaines : on estime que la moitié des clients vont avoir une durée d'abonnement

supérieure à 114 semaines. De même, la moitié des clients vont avoir une durée d'abonnement inférieure à 114 semaines. Un intervalle de confiance de niveau 95% pour ce temps médian est [110;119].

Un estimé de la moyenne et de l'écart-type est donné, mais ce dernier est biaisé (trop bas) puisque les données censurées ne donnent qu'une borne inférieure pour la vraie valeur. Avec un modèle paramétrique pour la survie (par ex., une loi exponentielle), les paramètres estimés du modèle dicteraient ces deux valeurs. Le modèle de Kaplan–Meier estime la survie, mais si la plus grande observation est censurée, la courbe n'atteindra pas zéro.

Le graphique de la fonction de survie permet de lire le temps de survie pour une probabilité donnée. Les bandes donnent un intervalle de confiance ponctuel de niveau 95% pour chaque temps donné.



Le tableau 6.1 donne le nombre de données censurées : parmi les 500 observations, il y a 334 clients qui ont terminé leur abonnement et 166 qui sont censurées (le client est toujours abonné et le temps est donc une borne inférieure de la durée d'abonnement).

Récapitulatif du nombre de valeurs censurées et non censurées			
			Pourcentage censuré
Total	A échoué	Censuré	
500	334	166	33.20

FIGURE 6.1 – Fraction de valeurs censurées

6.3.1 Calcul de l'estimateur de Kaplan–Meier

Cette partie peut être omise; elle est incluse seulement par souci de complétude. Soit T_1, \dots, T_n les n réalisations aléatoires de la variable temps (certaines censurées, d'autres pas). Supposons qu'il y a m temps distincts où au moins un individu expérimente l'événement. Soient $t_{(1)} < \dots < t_{(m)}$, ces temps ordonnés en ordre croissant et r_i le nombre d'individus à risque au temps $t_{(i)}$ (les individus qui n'ont pas encore expérimenté l'événement, et qui ne sont pas encore censurés avant $t_{(i)}$). On note d_i le nombre d'individus qui expérimentent l'événement au temps $t_{(i)}$. L'estimateur de Kaplan–Meier de la fonction de survie à un temps t est

$$\hat{S}(t) = \left(1 - \frac{d_1}{r_1}\right) \times \dots \times \left(1 - \frac{d_{i(t)}}{r_{i(t)}}\right), \quad t_{(1)} \leq t \leq t_{(m)},$$

où $i(t) = \max(j \in \{1, \dots, m\} : t \geq t_j)$, soit le plus grand indice parmi $1, \dots, m$ tel que $t \geq t_{i(t)}$. Par convention, si $t < t_{(1)}$, on fixe $\hat{S}(t) = 1$.

6.4 Comparaison de deux courbes de survie

Supposons que les individus ont été divisés en deux groupes et que $S_1(t)$ et $S_2(t)$ dénotent respectivement la fonction de survie du premier groupe et du deuxième groupe. On est souvent intéressé à tester l'égalité des fonctions de survie, c'est-à-dire, les hypothèses $\mathcal{H}_0 : S_1(t) = S_2(t)$ pour tout t et $\mathcal{H}_1 : S_1(t) \neq S_2(t)$ pour au moins une valeur de t .

Par exemple, dans une étude sur le temps de survie après avoir été diagnostiqué avec un certain type de cancer, on pourrait vouloir comparer le temps de survie des individus ayant reçu le traitement standard (groupe 1) au temps de survie des individus ayant reçu un nouveau traitement (groupe 2).

Les deux tests utilisés habituellement sont le test du log-rang (*log-rank test*) et le test de Wilcoxon généralisé (ou test de Gehan).

Testons l'hypothèse que la courbe de survie des clients masculins est la même que celle des clients féminins dans l'exemple des données d'abonnement. Ce test est effectué grâce à l'option `strata` de la procédure `lifetest` à l'aide des commandes suivantes (voir le fichier `survival1_fonction_survie.sas`):

```
proc lifetest data=multi.survival1 method=km plots=(s) censoredsymbol=none;
time t*censure(1);
strata sexe;
run;
```

On retrouve dans la sortie les estimés de la fonction de survie, de même que les quartiles par strate : la première strate correspond aux hommes (`sexe=0`) et la deuxième aux femmes (`sexe=1`).

Estimations du quartile					
Pourcentage	Valeur estimée du point	Transformation	Intervalle de confiance à 95%		
			[Inférieur	Supérieur]	
75	135.000	LOGLOG	128.000	177.000	
50	110.000	LOGLOG	101.000	115.000	
25	78.000	LOGLOG	70.000	84.000	

Estimations du quartile					
Pourcentage	Valeur estimée du point	Transformation	Intervalle de confiance à 95%		
			[Inférieur	Supérieur]	
75	216.000	LOGLOG	171.000	216.000	
50	123.000	LOGLOG	114.000	135.000	
25	98.000	LOGLOG	87.000	104.000	

Récapitulatif du nombre de valeurs censurées et non censurées					
Niveau de discréttisation	sex	Total	A échoué	Censuré	Pourcentage
					censuré
1	0	309	217	92	29.77
2	1	191	117	74	38.74
Total		500	334	166	33.20

On voit qu'il y a 309 hommes et 191 femmes. L'estimation du temps de survie médian est de 110 semaines pour les hommes et de 123 semaines pour les femmes.

Test d'égalité sur les niveaux de discréétisation			
Test	khi-2	DDL	Pr >
Log-rang	16.4347	1	<.0001
Wilcoxon	17.5966	1	<.0001
-2Log(LR)	7.1752	1	0.0074

Un des tableaux contient les statistiques et valeurs-*p* pour trois tests de l'hypothèse d'égalité des fonctions de survie, le test du log-rang et le test de Gehan. Le troisième test dans le tableau ($-2\text{Log}(\text{LR})$) est un test du rapport de vraisemblance sous l'hypothèse que les temps de survie des deux groupes suivent une loi exponentielle. Il est préférable d'utiliser les deux premiers qui ne font pas d'hypothèses quant à la distribution du temps de survie. Les valeurs-*p* des tests log-rang et de Wilcoxon généralisé sont toutes les deux inférieures à 10^{-4} . On rejette donc \mathcal{H}_0 pour conclure qu'il y a donc une différence significative entre les deux courbes de survie.

Le test de Wilcoxon généralisé accorde plus de poids au temps près du début de la distribution qu'au temps plus loin. Il est donc, en général, plus puissant lorsque la différence entre les deux fonctions de survie survient tôt dans la distribution. Le test log-rang quant à lui suppose que le ratio des fonctions de risques des deux groupes est constant pour toute la période d'intérêt.

Ces courbes sont représentées dans la Figure 6.2. On voit que la courbe des femmes est systématiquement au-dessus de celle des hommes. Les femmes ont donc tendance à rester abonnées plus longtemps que les hommes, et cette différence est significative.

Il est également possible de tester l'égalité des courbes de survie avec plus de deux groupes. Par exemple, s'il y a trois groupes, l'hypothèse nulle est alors $\mathcal{H}_0 : S_1(t) = S_2(t) = S_3(t)$ pour tout t , versus l'alternative qu'au moins deux des fonctions ont une valeur différente pour au moins une valeur de t . Les commandes **SAS** pour exécuter le test sont les mêmes : il suffit de mettre la variable identifiant les groupes à la ligne **strata**.

L'estimateur de Kaplan-Meier ne permet pas l'inclusion de variables explicatives à proprement parler : si on peut veux les différences au niveau de la survie selon les modalités d'une variable explicative catégorielle, on divise pour ce faire l'échantillon en sous-groupes et on utilise l'estimateur de Kaplan-Meier pour chacune des modalités en gardant en tête que cela réduit la taille de l'échantillon disponible et que l'estimation résultante est possiblement trop variable pour être utile.

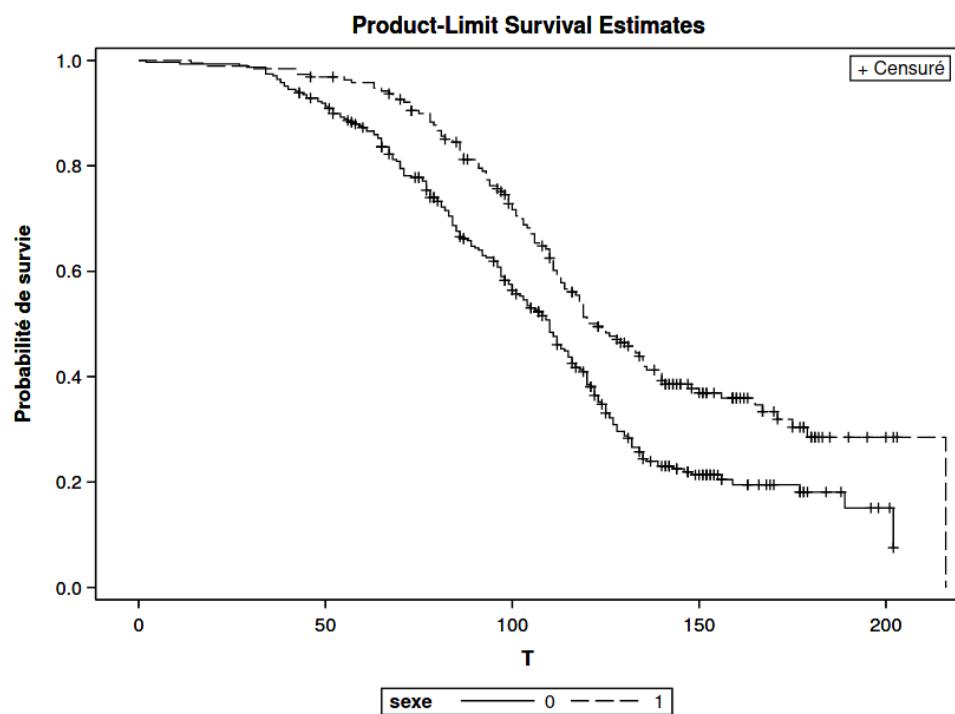


FIGURE 6.2 – Courbes de survie estimées par sexe (Kaplan–Meier)

6.5 Modèle à risques proportionnels de Cox

Le modèle à risques proportionnels de Cox (*proportional hazard model*) est l'un des modèles les plus utilisés pour analyser des données de survie.

6.5.1 Description du modèle de Cox

Soit $h(t; \mathbf{x})$ la valeur de la fonction de risque au temps pour un individu dont les valeurs des variables explicatives sont $X_1 = x_1, \dots, X_p = x_p$. Le modèle à risques proportionnels est

$$h(t; \mathbf{x}) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p)$$

où $h_0(t)$ est la fonction de risque de base ; il n'est pas nécessaire de spécifier cette dernière, d'où la nature semiparamétrique du modèle de Cox. Le postulat de risques proportionnels implique que le terme de droite $\exp(\mathbf{X}\boldsymbol{\beta})$ ne dépend pas du temps, et plus particulièrement β_1, \dots, β_p ne dépend pas du temps. Nous verrons subséquemment une extension qui permet de prendre en compte les variables explicatives dont la valeur change dans le temps en scindant ces observations.

Lorsque toutes les variables explicatives prennent la valeur zéro, $\mathbf{X} = \mathbf{0}$, on recouvre $h(t; \mathbf{0}) = h_0(t)$. Par conséquent, la fonction $h_0(t)$ peut être interprétée comme la fonction de risque lorsque toutes les variables explicatives valent zéro. Toutefois, tout comme la valeur de l'ordonnée à l'origine dans un modèle de régression linéaire, cette interprétation n'est pas nécessairement valide si la situation où toutes les variables explicatives valent zéro n'est pas possible ou si elle ne survient pas dans notre échantillon.

La deuxième partie du modèle, $\exp(\beta_1 x_1 + \dots + \beta_p x_p)$, vient modéliser l'effet d'un changement des valeurs des variables explicatives sur la fonction de risque de base. Tout comme dans le cas de la régression logistique (l'effet des variables sur la cote), c'est un effet multiplicatif, d'où le terme **risques proportionnels**.

Pour l'interprétation des paramètres, il sera plus simple de penser en termes de rapport de risque (*hazard ratio*), qui est défini comme étant le rapport des fonctions de risque pour deux ensembles de valeurs des variables explicatives. Pour simplifier l'illustration, supposons que nous avons seulement une variable explicative X et que $h(t; x) = h_0(t) \exp(\beta x)$. Le rapport de risque lorsque $X = x_1$ par rapport à $X = x_0$ est

$$\frac{h(t; x_1)}{h(t; x_0)} = \exp\{\beta(x_1 - x_0)\}.$$

Par conséquent, l'impact d'une augmentation de X d'une unité (quand $x_1 - x_0 = 1$) est $\exp(\beta)$. Ainsi, pour chaque augmentation d'une unité pour X , le risque que l'événement survienne est multiplié par $\exp(\beta)$.

Le terme **risques proportionnels** fait référence à la situation où le rapport de risque dépend seulement de la différence $x_1 - x_0$ et non pas du temps lui-même. Le rapport de risque est constant par rapport au temps t . Cela implique que l'effet d'une variable est stable dans le temps. Nous verrons plus loin comment faire en sorte que l'effet d'une variable puisse varier dans le temps.

Débutons avec un exemple simple en utilisant les données d'abonnement : on ajuste un modèle de Cox en utilisant seulement la variable binaire sexe (survival12_cox.sas). Ceci peut être fait avec la procédure phreg, comme suit :

```
proc phreg data=multi.survival1;
model t*censure(1)=sexe / ties=exact;
run;
```

La sortie inclut notamment des tests de significativité globale basés sur la vraisemblance pour les variables explicatives (rapport de vraisemblance, score et Wald) ainsi qu'un tableau des coefficients et des statistiques sur la qualité de l'ajustement pour le modèle sans variable explicative et celui qui inclut sexe.

Statistique d'ajustement du modèle			
Critère	Sans covariables	Avec covariables	
-2 LOG L	3219.227	3202.443	
AIC	3219.227	3204.443	
SBC	3219.227	3208.254	

Test de l'hypothèse nulle globale : BETA=0				
Test	khi-2	DDL	Pr > khi-2	
Rapport de vrais	16.7843	1	<.0001	
Score	16.4260	1	<.0001	
Wald	16.1381	1	<.0001	

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Libellé
sex	1	-0.46563	0.11591	16.1381	<.0001	0.628 sex

Il y a une seule variable explicative, le sexe de l'individu. L'estimation du paramètre de l'effet de sexe est $\hat{\beta} = -0,466$. Ce paramètre est significativement différent de 0 (valeur- p inférieure à 10^{-4}).

Pour l'interprétation, on utilise la colonne « Rapport de risque » qui contient la valeur $\exp(\hat{\beta}_{\text{sex}}) = \exp(-0,466) = 0,628$. Ainsi, le rapport du risque d'une femme par rapport à un homme est

$$\frac{\hat{h}(t; \text{sex} = 1)}{\hat{h}(t; \text{sex} = 0)} = 0,628.$$

Par conséquent, le risque qu'une femme interrompe son abonnement est 0,628 fois celui d'un homme. Une femme est donc moins à risque de quitter qu'un homme. Nous avions déjà vu cela à la section précédente lorsque nous avions comparé les courbes de survie des hommes et des femmes. Il est important de se rappeler qu'avec ce modèle, l'effet d'une variable est le même dans le temps (peu importe la valeur de t). Donc, une femme est moins à risque de quitter qu'un homme à tout moment, d'après ce modèle. Inversement, le ratio du risque d'un homme par rapport à une femme est $1/0,628 = 1,59$. Ainsi, à tout moment, un homme a un risque d'interrompre son abonnement qui est 59% plus élevé que celui d'une femme.

Comme il y a un seul paramètre ici, les tests basés sur la vraisemblance pour $\mathcal{H}_0: \boldsymbol{\beta} = \mathbf{0}$ reviennent à tester l'effet de la variable sexe. Le test de Wald est le même que celui du tableau des coefficients. Dans le cas particulier où il y a une seule variable explicative binaire (comme ici), le test du score est équivalent au test du log-rang que nous avons vu à la section précédente (à une petite différence près lorsqu'il y a des ex aequo dans les temps de survie).

On pourrait également utiliser une variable explicative continue plutôt qu'une variable binaire ; le principe est le même.

Statistique d'ajustement du modèle		
Critère	Sans covariables	Avec covariables
-2 LOG L	3219.227	3174.278
AIC	3219.227	3176.278
SBC	3219.227	3180.089

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre	DDL	des paramètres	Valeur estimée	Erreur	Rapport	
			type	khi-2	Pr > khi-2	de risque
age	1		-0.04191	0.00651	41.3880	<.0001
					0.959	age

Le rapport de risque pour âge est 0,959 et donc le risque diminue de 4,1% chaque fois que l'âge augmente d'un an — le risque d'interrompre l'abonnement diminue lorsque l'âge augmente et cet effet est significatif (valeur- p des tests inférieures à 10^{-4}).

Généralement, on considérera le modèle de Cox avec toutes les variables explicatives simultanément. La variable `region` est nominale tandis que la variable `service` est ordinaire (avec quatre modalités). Nous allons les incorporer, comme d'habitude, en utilisant des variables indicatrices avec `region=5` et `service=0` (le client n'est abonné à aucun autre service) comme catégories de référence.

```
proc phreg data=multi.survival1;
class region(ref='5') service(ref='0') / param=ref;
model t*censure(1)=age sexe region service / ties=exact;
run;
```

Statistique d'ajustement du modèle

Critère	Sans covariables	Avec covariables
-2 LOG L	3219.227	2992.289
AIC	3219.227	3010.289
SBC	3219.227	3044.590

Test de l'hypothèse nulle globale : BETA=0

Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	226.9378	9	<.0001
Score	229.5370	9	<.0001
Wald	209.6444	9	<.0001

Tests Type 3

Effet	DDL	Khi-2 de Wald	Pr > khi-2
age	1	61.1027	<.0001
sexe	1	31.2185	<.0001
region	4	7.5638	0.1089
service	3	149.5195	<.0001

Analyse des valeurs estimées du maximum de vraisemblance								
Paramètre	DDL	Valeur estimée	Erreur	Rapport				
		des paramètres	type	khi-2	Pr > khi-2	de risque	Libellé	
age	1	-0.05121	0.00655	61.1027	<.0001	0.950	age	
sexé	1	-0.66470	0.11897	31.2185	<.0001	0.514	sexé	
region	1	0.03372	0.17498	0.0371	0.8472	1.034	region 1	
region	2	1	-0.35994	0.18347	3.8488	0.0498	0.698	region 2
region	3	1	0.06470	0.16976	0.1453	0.7031	1.067	region 3
region	4	1	-0.18771	0.16887	1.2356	0.2663	0.829	region 4
service	1	1	-1.02707	0.12623	66.2024	<.0001	0.358	service 1
service	2	1	-1.73132	0.18543	87.1745	<.0001	0.177	service 2
service	3	1	-2.13557	0.25272	71.4058	<.0001	0.118	service 3

Les effets des variables sont maintenant des effets marginaux. Ainsi, lorsque les autres variables demeurent fixes, le risque de quitter d'une femme est 0,514 fois plus petit que celui d'un homme. L'effet marginal (une fois que les autres variables sont incluses) de la variable sexe est significatif (valeur-*p* inférieure à 10^{-4}).

Toutes autres choses étant égales, chaque augmentation de l'âge d'un an fait diminuer le risque d'interrompre l'abonnement. Plus précisément, le risque est multiplié par 0,95 lorsque l'âge augmente d'un an et cet effet est significatif.

Pour la variable service, l'interprétation se fait par rapport à la catégorie de référence, qui est la catégorie 0 (abonné à aucun autre service). Ainsi, si le client est abonné à un autre service, son risque de quitter est 0,358 fois celui d'un client qui n'est pas abonné à un autre service (toutes autres choses étant égales). Si le client est abonné à deux autres services, son risque de quitter est encore plus petit comparativement à un client qui n'est pas abonné à un autre service (rapport de risque de 0,177) Finalement, si le client est abonné à trois autres services, son risque de quitter est encore plus petit (rapport de risque de 0,118). Les paramètres de ces trois variables sont tous significatifs. Ainsi, les clients qui sont abonnés à un, deux ou trois services ont un risque de quitter qui est significativement plus faible que celui d'un client qui n'est pas abonné à un autre service. Le tableau Tests Type 3 permet de tester globalement la significativité d'une variable explicative modélisée avec plusieurs indicatrices. Pour la variable service, le test présenté teste l'hypothèse nulle $\mathcal{H}_0 : \beta_{\text{service}_1} = \beta_{\text{service}_2} = \beta_{\text{service}_3} = 0$ contre l'alternative qu'un moins un de ces trois paramètres est non-nul. Le test est largement significatif (statistique de Wald du χ^2_4 valant 149,52 avec une valeur-*p* inférieure à 10^{-4}). L'effet de la variable service est donc globalement significatif. Afin de comparer les autres modalités entre elles, par exemple afin de voir si le risque de quitter est différent entre un client qui a deux services et un autre qui a trois services, il suffit de changer la catégorie de référence à la commande class et de réajuster le modèle.

Finalement, la variable `region` n'est pas globalement significative (statistique de Wald de 7,56 avec une valeur- p de 0,11).

6.5.2 Estimation de la fonction de survie pour des valeurs particulières des variables explicatives

Il est possible d'obtenir l'estimation de la fonction de survie pour des valeurs particulières des variables explicatives avec la commande `baseline` (voir le script `survival2_cox.sas` pour plus de détails). Pour ce faire, il faut avoir un autre fichier de données qui contient les valeurs des variables explicatives pour lesquelles on veut une estimation de la fonction de survie.

Si on ajuste le modèle avec aucune variable explicative, on retrouvera alors l'estimation de Kaplan-Meier de la fonction de survie comme avec la procédure `lifetest`.

Supposons qu'on ajuste le modèle avec les variables sexe et âge seulement dans l'exemple du temps d'abonnement, et que l'on désire la fonction de survie pour les hommes de 25 et 60 ans et pour les femmes de 25 et 60 ans. Le fichier `survival2.sas7bdat` contient les données qui seront utilisées à cette fin. Il contient seulement les quatre lignes suivantes.

Obs.	sex	age
1	0	25
2	1	25
3	0	60
4	1	60

Les quatre fonctions de la Figure 6.3 correspondent aux profils pour lesquels nous désirons une estimation de la courbe de survie. La courbe 1 est pour les hommes de 25 ans, la courbe 2 pour les femmes de 25 ans, la courbe 3 pour les hommes de 60 ans et la courbe 4 pour les femmes de 60 ans. On voit donc que, parmi ces quatre profils, les hommes de 25 ans sont le plus à risque de quitter tandis que les femmes de 60 ans sont le moins à risque de quitter.

6.6 Extensions du modèle de Cox

Dans cette section, nous allons voir deux extensions du modèle à risques proportionnels de base.

- i) Inclusion de variables explicatives dont la valeur change dans le temps.
- ii) Modèle à risques compétitifs (*competing risks*) pour étudier la situation où il y a plusieurs manières de quitter l'état.

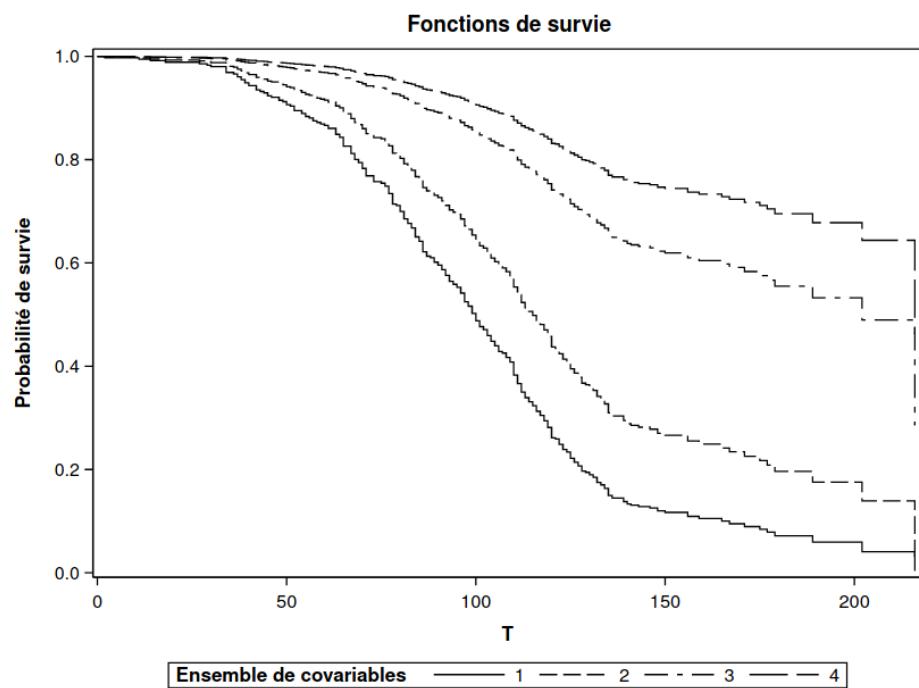


FIGURE 6.3 – Courbes de survie pour hommes et femmes de 25 et 60 ans.

6.6.1 Variables explicatives dont la valeur change dans le temps

Il est clair que certaines caractéristiques d'un individu évoluent dans le temps (*time-varying covariates*). Si le sexe d'un individu est stable dans le temps, son revenu, son statut matrimonial, l'endroit où il habite, sont par contre des caractéristiques qui peuvent changer dans le temps. Il peut alors être intéressant d'en tenir compte dans l'analyse. Rappelez-vous que le modèle à risques proportionnels est

$$h(t; \mathbf{x}) = h_0(t) \exp(\beta_1 x_1 + \cdots + \beta_p x_p).$$

Supposons que la variable X_1 change au fil du temps et que les autres demeurent fixes. On peut alors réécrire le modèle

$$h(t; \mathbf{x}) = h_0(t) \exp\{\beta_1 x_1(t) + \cdots + \beta_p x_p\},$$

où $x_1(t)$ indique que la valeur de X_1 dépend du temps t .

Supposons que la variable `service`, qui représente le nombre d'autres services souscrits, est la seule que nous voulons modéliser comme une variable qui varie dans le temps. Pour l'âge, nous prenons simplement l'âge au début de l'abonnement, idem pour la région.

Le plus difficile est de créer correctement le fichier de données pour effectuer ce genre d'analyse. Supposons pour cet exemple qu'il y a eu au plus un changement dans la variable `service`, comme présenté dans le fichier `survival3.sas7bdat`. Les variables `t`, `censure`, `age`, `sexe` et `region` sont comme précédemment. Trois nouvelles variables remplacent l'ancienne variable `service`.

- `service_avant` : nombre d'autres services auxquels le client est abonné au début de son abonnement.
- `temps_ch` : temps au moment où un changement est survenu quant au nombre d'autres services. En l'absence de changement, l'observation est remplacée par une valeur manquante (.).
- `service_apres` : nombre d'autres services auxquels le client est abonné à partir du temps `temps_ch`.

Obs.	T	censure	age	sexe	region	service_avant	temps_ch	service_apres
1	178	1	48	1	3	2	130	1
2	159	1	31	1	3	2	.	2
3	110	0	36	1	4	0	.	0
4	109	0	30	0	2	0	.	0
5	108	0	22	0	5	1	78	0

On regarde plus en détail le profil des cinq premiers clients, dont seuls deux ont changé le nombre d'abonnements ; les individus 3–5 se sont désabonnés du service cellulaire à un moment donné. Le premier client était abonné à deux autres services au début de son abonnement au téléphone cellulaire mais, après 130 semaines d'abonnement, a effectué un changement à son forfait pour ne conserver qu'un autre service en plus du cellulaire. Pour le deuxième client, comme `temps_ch` est manquante, il est toujours abonné à deux autres services et ce, jusqu'à la fin de l'étude.

Les commandes **SAS** permettant d'ajuster le modèle avec point de rupture se trouvent dans le fichier `survival3_varie_temps.sas`. Notez que lorsqu'on a une variable catégorielle qui varie dans le temps (comme ici avec la variable `service`), on ne peut pas utiliser `class` pour la déclarer catégorielle ; il faut plutôt créer nous-mêmes les variables indicatrices nécessaires à l'intérieur même de l'appel à la fonction `phreg` (voir le script). Ici, nous utiliserons la catégorie 0 (aucun autre service) comme catégorie de référence.

Analyse des valeurs estimées du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur	khi-2	Pr > khi-2	Rapport	
		des paramètres	type			de risque	Libellé
age	1	-0.04928	0.00657	56.2147	<.0001	0.952	age
sex	1	-0.57079	0.11768	23.5261	<.0001	0.565	sex
region	1	0.06705	0.17452	0.1476	0.7008	1.069	region 1
region	2	0.36114	0.18183	3.9449	0.0470	0.697	region 2
region	3	-0.01305	0.16912	0.0060	0.9385	0.987	region 3
region	4	-0.12632	0.16832	0.5632	0.4530	0.881	region 4
service1	1	-0.50342	0.12166	17.1236	<.0001	0.604	
service2	1	-1.32035	0.18636	50.1949	<.0001	0.267	
service3	1	-1.55893	0.24120	41.7746	<.0001	0.210	

L'interprétation se fait comme précédemment. Ici, c'est la valeur d'une variable qui varie dans le temps et non pas son effet. Ainsi, le ratio de risque de quitter pour un client qui a un autre service est 0,604 fois celui d'un client qui n'a aucun autre service (référence). Le fait d'avoir deux ou trois services diminue encore plus le risque de quitter (ratios de risque de 0,267 et 0,21, respectivement).

6.6.2 Modèle à risques compétitifs

Parfois, la raison pour laquelle un individu quitte l'état étudié peut avoir un intérêt en soi. Par exemple si on s'intéresse au temps qu'un employé demeure au service de la compagnie, la distinction entre le fait qu'il ait démissionné ou bien qu'il ait été renvoyé peut avoir un impact sur

l'effet des variables explicatives. Comme autre exemple, si on s'intéresse au temps de survie d'un individu après qu'il ait été diagnostiqué avec un certain type de cancer, il pourrait être important de distinguer selon la cause exacte de la mort.

De manière générale, supposons qu'il y a K manières possibles que l'événement survienne. On peut alors spécifier K fonctions de risques (une pour chaque manière) et obtenir le modèle de Cox à risques compétitifs (*competing risks*),

$$\begin{aligned} h_1(t; \mathbf{x}) &= h_{01}(t) \exp(\beta_{11}x_1 + \cdots + \beta_{p1}x_p) \\ &\vdots \\ h_K(t; \mathbf{x}) &= h_{0K}(t) \exp(\beta_{1K}x_1 + \cdots + \beta_{pK}x_p) \end{aligned}$$

Notez que les coefficients K sont différents d'une équation à l'autre. En estimant ce modèle, on obtient donc une estimation de l'effet des variables selon la raison du départ de l'état. De plus, on peut aussi inclure des variables dont la valeur change dans le temps, comme vu précédemment. Ce qui simplifie énormément la situation est qu'il est prouvé qu'on peut estimer les paramètres de chaque équation séparément sans perte de précision. Par conséquent, en pratique, il suffira d'ajuster K modèles séparément.

Dans notre exemple d'abonnement cellulaire, supposons que nous avons trois causes possibles pour la perte d'un client : soit il a interrompu son abonnement pour aller chez le compétiteur A, soit pour aller chez le compétiteur B, soit il n'a plus de cellulaire du tout.

Les données pour cet exemple se trouvent dans le fichier `survival4.sas7bdat` et le programme dans le fichier `survival4_risques_competitifs.sas`. La seule nouveauté par rapport au fichier original est la variable `censure` qui est maintenant codée ainsi

- 1, si le temps est censuré (l'individu est toujours abonné à notre service)
- 2, si l'individu a quitté pour aller chez le compétiteur A
- 3, si l'individu a quitté pour aller chez le compétiteur B
- 4, si l'individu a quitté parce qu'il n'a plus besoin de cellulaire.

On peut calculer la fréquence de chaque modalité.

		censure			
		Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
censure		Fréquence	Pourcentage		
	1	166	33.20	166	33.20
	2	170	34.00	336	67.20
	3	121	24.20	457	91.40
	4	43	8.60	500	100.00

Ainsi, il y a donc 166 clients toujours abonnés, 170 qui nous ont quitté pour aller chez A, 121 pour aller chez B, et 43 qui n'ont plus de cellulaires.

Pour ajuster le modèle lorsque la cause du départ est le compétiteur A, le code est

```
proc phreg data=multi.survival4;
class region(ref='5') service(ref='0') / param=ref;
model t*censure(1,3,4)=age sexe region service / ties=exact;
run;
```

Notez qu'on précise que les valeurs 1, 3 et 4 sont des observations censurées. Ici, l'événement d'intérêt est que le client est parti chez le compétiteur A. S'il est toujours abonné (censure=1), s'il est parti chez le compétiteur B (censure=3) ou s'il nous a quitté car il n'a plus de cellulaire (censure=4), alors l'événement « quitter pour aller chez A » n'est pas survenu. C'est pourquoi on doit traiter ces situations comme des censures.

Récapitulatif du nombre d'événements et de valeurs censurées

	Total	Evénement	Censuré	Pourcentage censuré
	500	170	330	66.00

Analyse des valeurs estimées du maximum de vraisemblance

Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Libellé
age	1	-0.04646	0.00910	26.0722	<.0001	0.955	age
sex	1	-0.81137	0.17057	22.6262	<.0001	0.444	sex
region	1	-0.04751	0.23272	0.0417	0.8382	0.954	region 1
region	2	-0.65685	0.25942	6.4112	0.0113	0.518	region 2
region	3	-0.24149	0.23984	1.0138	0.3140	0.785	region 3
region	4	-0.37356	0.22969	2.6450	0.1039	0.688	region 4
service	1	-0.96060	0.17641	29.6525	<.0001	0.383	service 1
service	2	-1.71341	0.26263	42.5628	<.0001	0.180	service 2
service	3	-2.23135	0.37031	36.3089	<.0001	0.107	service 3

Ainsi, on voit que l'événement est survenu 170 fois et qu'il y a 330 censures. L'interprétation des paramètres se fait comme précédemment. Sauf qu'il faut préciser qu'il s'agit du risque de quitter pour aller chez le compétiteur A. Par exemple, le risque de quitter pour aller chez le compétiteur

A d'une femme est 0,444 fois le risque de quitter pour aller chez le compétiteur A d'un homme. Ainsi, les femmes sont moins à risque de quitter pour aller chez le compétiteur A que les hommes.

Pour ajuster le modèle lorsque la cause du départ est le compétiteur B, procède de la même manière. Notez que cette fois-ci, ce sont les valeurs 1, 2 et 4 de la variable censure qui correspondent au fait que l'événement n'est pas survenu; on spécifie donc `censure(1,2,4)` dans l'appel à la fonction `phreg`. Il y a 121 clients qui ont quitté pour aller chez B et 379 cas autre (censure). L'interprétation des paramètres se fait en termes de risque de quitter pour aller chez le compétiteur B.

Récapitulatif du nombre d'événements et de valeurs censurées			
Total	Evénement	Censuré	Pourcentage censuré
500	121	379	75.80

Analyse des valeurs estimées du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée	Erreur	Rapport			
		des paramètres	type	khi-2	Pr > khi-2	de risque	Libellé
age	1	-0.04993	0.01078	21.4428	<.0001	0.951	age
sex	1	-0.48798	0.19292	6.3981	0.0114	0.614	sex
region	1	0.08719	0.29647	0.0865	0.7687	1.091	region 1
region	2	-0.12432	0.29637	0.1760	0.6749	0.883	region 2
region	3	0.19191	0.28071	0.4674	0.4942	1.212	region 3
region	4	-0.22848	0.29409	0.6036	0.4372	0.796	region 4
service	1	-0.86407	0.20809	17.2418	<.0001	0.421	service 1
service	2	-1.57303	0.29915	27.6499	<.0001	0.207	service 2
service	3	-1.98386	0.41235	23.1472	<.0001	0.138	service 3

Si on ajuste le modèle pour le cas de figure où la cause de départ est que le client n'a plus besoin de cellulaire, on obtient une sortie similaire. On voit, que contrairement aux deux premiers modèles, l'effet de la variable sexe n'est pas significatif ici.

Récapitulatif du nombre d'événements et de valeurs censurées			
Total	Evénement	Censuré	Pourcentage censuré
500	43	457	91.40

Analyse des valeurs estimées du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée	Erreur	Rapport			
		des paramètres	type	khi-2	Pr > khi-2	de risque	Libellé
age	1	-0.07415	0.01975	14.0929	0.0002	0.929	age
sex	1	-0.58995	0.32996	3.1967	0.0738	0.554	sex
region	1	0.39191	0.60931	0.4137	0.5201	1.480	region 1
region	2	0.35437	0.59549	0.3541	0.5518	1.425	region 2
region	3	1.02514	0.53628	3.6540	0.0559	2.787	region 3
region	4	0.83667	0.53069	2.4856	0.1149	2.309	region 4
service	1	-1.78026	0.39257	20.5653	<.0001	0.169	service 1
service	2	-2.15812	0.55416	15.1664	<.0001	0.116	service 2
service	3	-2.03798	0.62506	10.6306	0.0011	0.130	service 3

En fait, comme il y a seulement 43 événements (quitter car on n'a plus besoin de cellulaire), les estimations des paramètres sont moins précises, ce qu'on peut voir avec les erreurs-types qui sont plus élevés. Les observations censurées contiennent moins d'information que les événements, d'où cette perte de précision.

6.7 Risques non proportionnels

Comme son nom l'indique, le modèle à risques proportionnels suppose que les risques sont proportionnels. Cela implique que l'effet d'une variable est stable dans le temps. Nous verrons dans cette section deux façons de modéliser le cas de risques non proportionnels.

6.7.1 Non-proportionnalité avec un terme d'interaction avec le temps

Pour simplifier l'exposition, supposons que nous avons une seule variable explicative X . L'équation du modèle à risques proportionnels est $h(t; x) = h_0(t) \exp(\beta x)$ et suppose que la fonction de risque de base $h_0(t)$ est indépendante de la variable explicative X . Une manière de modéliser la non-proportionnalité est d'inclure un terme d'interaction entre la variable et le temps. Il existe plusieurs façons de le faire, l'une d'entre elles consiste à inclure une nouvelle variable qui est le produit entre le temps et la variable X . Le modèle est alors

$$h(t; x) = h_0(t) \exp(\beta_1 x + \beta_2 t).$$

Pour ce modèle, le rapport de risque, pour une augmentation d'une unité de X est $\exp(\beta_1 + \beta_2 t)$ et dépend du temps t : c'est un modèle avec risques non proportionnels. On retombe sur le modèle à risques proportionnels lorsque $\beta_2 = 0$.

Ajustons le modèle pour l'abonnement cellulaire en ajoutant une interaction entre l'âge et le temps. Les commandes se trouvent dans `survival5_non_proportionnel.sas`, dont l'extrait montre comment créer la variable produit.

```
proc phreg data=temp;
class region(ref='5') service(ref='0') / param=ref;
model t*censure(1)=age iaget sexe region service / ties=exact;
iaget=age*t;
run;
```

L'interaction entre l'âge et le temps est spécifiée en incluant une nouvelle variable, `iaget`, créée à l'intérieur de l'appel à la procédure `phreg`, et qui est égale au produit entre l'âge et `t`. On remarque que le terme d'interaction est tout juste non significatif (valeur-*p* de 0,061).

Analyse des valeurs estimées du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Libellé
<code>age</code>	1	-0.09005	0.02196	16.8120	<.0001	0.914	age
<code>iaget</code>	1	0.0003921	0.0002093	3.5096	0.0610	1.000	
<code>sexe</code>	1	-0.65412	0.11881	30.3099	<.0001	0.520	sexe
<code>region</code>	1	0.02842	0.17505	0.0264	0.8710	1.029	region 1
<code>region</code>	2	-0.32803	0.18390	3.1819	0.0745	0.720	region 2
<code>region</code>	3	0.06360	0.16970	0.1405	0.7078	1.066	region 3
<code>region</code>	4	-0.19777	0.16898	1.3697	0.2419	0.821	region 4
<code>service</code>	1	-1.01671	0.12615	64.9541	<.0001	0.362	service 1
<code>service</code>	2	-1.70574	0.18535	84.6957	<.0001	0.182	service 2
<code>service</code>	3	-2.10448	0.25233	69.5602	<.0001	0.122	service 3

6.7.2 Stratification

Une autre manière de modéliser la non-proportionnalité est par la stratification. Il est important de comprendre qu'on ne pourra pas estimer l'effet de la variable de stratification. On devrait donc seulement utiliser une variable qui ne nous intéresse pas en soi, mais qui peut avoir un effet sur le temps de survie.

Encore une fois, pour simplifier, supposons que nous avons deux variables *X* et *Z*; cette dernière est binaire et prend les valeurs 0 et 1. On s'intéresse à l'effet de la variable *X* mais pas à celui de la variable *Z*; néanmoins, on croit que *Z* impacte le temps de survie et que, de ce fait, le postulat

de proportionnalité de la fonction de risque n'est pas validé. Le modèle de Cox avec stratification (pour la variable Z) est

$$h(t; x, z) = h_0(t) \exp(\beta x),$$

où $h_0(t)$ est la fonction de risque de base quand $Z = 0$ et $h_1(t)$ est la fonction de risque de base lorsque $Z = 1$. L'effet de la variable X est supposé être le même pour toute valeur de Z , mais la fonction de risque de base peut différer. Le rapport de risque pour $Z = 1$ versus $Z = 0$ est $h_1(t)/h_0(t)$; cette quantité dépend du temps t . L'effet de la variable est donc variable dans le temps (et non pas constant). Ce modèle permet donc de modéliser la non-proportionnalité pour la variable Z . Si on stratifie par rapport à une variable, il ne faut pas l'inclure dans le modèle en plus car elle est déjà modélisée via la stratification. Notez que les paramètres β seront estimés à l'aide des données de toutes les strates.

L'avantage de la stratification est que cette méthode permet de modéliser n'importe quel changement dans l'effet d'une variable dans le temps sans devoir spécifier un type de changement particulier, comme lorsqu'on doit choisir la forme de l'interaction. Par contre, on perd la possibilité de tester l'effet de la variable de stratification on réduit la taille de l'échantillon pour l'estimation de la fonction de risque de base. On devrait donc utiliser la stratification seulement avec des variables pour lesquelles nous n'avons pas besoin d'estimer l'effet (variables secondaires ou de contrôles).

On considère la stratification par rapport à la région pour notre modèle pour le temps d'abonnement à un forfait cellulaire; un test permet de voir que l'hypothèse de non-proportionnalité des fonctions de risque n'est pas valide pour région. Le modèle contient toutes les variables explicatives, hormis région qui est utilisée pour la stratification.

Les commandes pour ajuster le modèle sont

```
proc phreg data=multi.survival1;
class service(ref='0') / param=ref;
model t*censure(1)=age sexe service / ties=exact;
strata region;
run;
```

Récapitulatif du nombre d'événements et de valeurs censurées					
Niveau de discrétilisation	region	Total	Événement	Censuré	Pourcentage censuré
1	1	91	62	29	31.87
2	2	86	56	30	34.88
3	3	107	71	36	33.64
4	4	108	73	35	32.41
5	5	108	72	36	33.33
Total		500	334	166	33.20

Analyse des valeurs estimées du maximum de vraisemblance								
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Libellé	
age	1	-0.05149	0.00672	58.7764	<.0001	0.950	age	
sex	1	-0.66552	0.12106	30.2221	<.0001	0.514	sex	
service	1	-1.00637	0.12765	62.1585	<.0001	0.366	service 1	
service	2	1	-1.71799	0.18730	84.1360	<.0001	0.179	service 2
service	3	1	-2.12496	0.25671	68.5193	<.0001	0.119	service 3

On voit à la lecture de la sortie SAS qu'il n'y a pas de paramètres pour la variable région. Les paramètres des autres variables s'interprètent comme d'habitude.

Le commentaire suivant est technique et peut être omis. Le postulat de risques proportionnels est rarement validé en pratique (et difficile à tester). Cela aura pour conséquences que les erreurs-types des variables explicatives autres que binaires sont faussées, problème qu'on peut régler en utilisant des procédures d'autoamorçage. La valeur du rapport de risque dépend de la distribution des pertes de suivi, même quand ces dernières surviennent de manière aléatoire ; une recommandation récente est d'utiliser des rapports de risques pondérés par la probabilité inverse d'appartenance, le tout complémenté par des mesures d'effets comme la différence de survie, la moyenne de survie (estimation restreinte) à des temps préspécifiés.

Chapitre 7

Données manquantes

Il arrive fréquemment d'avoir des valeurs manquantes dans notre échantillon. Ces valeurs peuvent être manquantes pour diverses raisons. Si on prélève nous-mêmes nos données, un répondant peut refuser de répondre à certaines questions. Si on acquiert nos données d'une source externe, les valeurs de certaines variables peuvent être manquantes directement dans le fichier obtenu. Si on ne prend pas en compte le mécanisme générant les valeurs manquantes, ces dernières peuvent également biaiser nos analyses. Le but de ce chapitre est de faire un bref survol de ce sujet.

7.1 Terminologie

Soit X une variable pour laquelle des données sont manquantes. Voici la définition de trois processus de génération de données manquantes.

- 1) Les données manquantes de X sont dites **manquantes de façon complètement aléatoire** (MCAR, de l'anglais *missing completely at random*) si la probabilité que la valeur de X soit manquante ne dépend ni de la valeur de X (qui n'est pas observée), ni des valeurs des autres variables.

Le fait qu'une variable est manquante peut être relié au fait qu'une autre soit manquante. Des gens peuvent refuser systématiquement de répondre à deux questions dans un sondage. Dans ce cas, si la probabilité qu'une personne ne réponde pas ne dépend pas des valeurs de ces variables (et de toutes les autres), nous sommes encore dans le cas MCAR. Si par contre, la probabilité que les gens ne répondent pas à une question sur leur revenu augmente avec la valeur de ce revenu, alors nous ne sommes plus dans le cas MCAR.

Le cas MCAR peut se présenter par exemple si des questionnaires, ou des pages ont été égarés ou détruits par inadvertance (effacées du disque rigide, etc.) Si les questionnaires manquants constituent un sous-ensemble choisi au hasard de tous les questionnaires, alors le processus est MCAR. L'hypothèse que les données manquantes sont complètement aléatoires est en général considérée comme trop restrictive.

- 2) Les données manquantes de X sont dites **données manquantes de façon aléatoire** (MAR, de l'anglais *missing at random*) si la probabilité que la valeur de X soit manquante ne dépend pas de la valeur de X (qui n'est pas observée) une fois qu'on a contrôlé pour les autres variables.

Il est possible par exemple que les femmes refusent plus souvent que les hommes de répondre à une question (et donc, le processus n'est pas MCAR). Si pour les femmes, la probabilité que X est manquante ne dépend pas de la valeur de X et que pour les hommes, la probabilité que X est manquante ne dépend pas de la valeur de X , alors le processus est MAR. Les probabilités d'avoir une valeur manquante sont différentes pour les hommes et les femmes mais cette probabilité ne dépend pas de la valeur de X elle-même. L'hypothèse MAR est donc plus faible que l'hypothèse MCAR.

- 3) Les données manquantes de X sont dites **manquantes de façon non-aléatoire** (MNAR, de l'anglais *missing not at random*) si la probabilité que la valeur de X soit manquante dépend de la valeur de X elle-même.

Par exemple, les gens qui ont un revenu élevé pourraient avoir plus de réticences à répondre à une question sur leur revenu. La méthode de traitement que nous allons voir dans ce chapitre, l'imputation multiple, est très générale et est valide dans le cas MAR (et donc aussi dans le cas MCAR). Le cas MNAR est beaucoup plus difficile à traiter et ne sera pas considéré ici.

7.2 Quelques méthodes

7.2.1 Cas complets

La première idée naïve pour une analyse est de retirer les observations avec données manquantes pour conserver les cas complets (*listwise deletion*, ou *complete case analysis*).

Cette méthode consiste à garder seulement les observations qui n'ont aucune valeur manquante pour les variables utilisées dans l'analyse demandée. Dès qu'une variable est manquante, on enlève le sujet au complet. C'est la méthode utilisée par défaut dans la plupart des logiciels, dont **SAS**.

- Si le processus est MCAR, cette méthode est valide car l'échantillon utilisé est vraiment un sous-échantillon aléatoire de l'échantillon original. Par contre, ce n'est pas nécessairement la meilleure solution car on perd de la précision en utilisant moins d'observations.
- Si le processus est seulement MAR ou MNAR, cette méthode produit généralement des estimations biaisées des paramètres.

En général, l'approche des cas complets est la première étape d'une analyse afin d'obtenir des estimés initiaux que nous corrigerais pas d'autre méthode. Elle n'est vraiment utile que si la proportion d'observations manquantes est très faible et le processus est MCAR. Évidemment, la présence de valeurs manquantes mène à une diminution de la précision des estimateurs (caractérisée par une augmentation des erreurs-types) et à une plus faible puissance pour les tests d'hypothèse et donc ignorer l'information partielle (si seulement certaines valeurs des variables explicatives sont manquantes) est sous-optimal.

7.2.2 Imputation simple

La deuxième approche est l'**imputation simple**. L'idée ici est de ne pas enlever les observations avec des valeurs manquantes mais de remplacer ces valeurs par des valeurs raisonnables. Par exemple, on peut remplacer les valeurs manquantes d'une variable par la moyenne de cette variable dans notre échantillon. On peut aussi ajuster un modèle de régression avec cette variable comme variable dépendante et d'autres variables explicatives comme variables indépendantes et utiliser les valeurs prédites comme remplacement. Une fois que les valeurs manquantes ont été remplacées, on fait l'analyse avec toutes les observations.

Il existe d'autres façons d'imputer les valeurs manquantes mais le problème de toutes ces approches est que l'on ne tient pas compte du fait que des valeurs ont été remplacées et on fait comme si c'était de vraies observations. Cela va en général sous-évaluer la variabilité dans les données. Par conséquent, les écarts-type des paramètres estimés seront en général sous-estimés et l'inférence (tests et intervalles de confiance) ne sera pas valide. Cette approche n'est donc **pas recommandée**.

Une manière de tenter de reproduire correctement la variabilité dans les données consiste à ajouter un terme aléatoire dans l'imputation. C'est ce que fait la méthode suivante, qui possédera l'avantage de corriger automatiquement les écarts-type des paramètres estimés.

7.2.3 Imputation multiple

Cette méthode peut être appliquée dans à peu près n'importe quelle situation et permet d'ajuster les écarts-type des paramètres estimés. Elle peut être appliquée lorsque le processus est MAR (et donc aussi MCAR).

L'idée consiste à procéder à une imputation aléatoire, selon une certaine technique, pour obtenir un échantillon complet et à ajuster le modèle d'intérêt avec cet échantillon. On répète ce processus plusieurs fois et on combine les résultats obtenus.

L'estimation finale des paramètres du modèle est alors simplement la moyenne des estimations pour les différentes répétitions et on peut également obtenir une estimation des écarts-type des paramètres qui tient compte du processus d'imputation.

Plus précisément, supposons qu'on s'intéresse à un seul paramètre θ dans un modèle donné. Ce modèle pourrait être un modèle de régression linéaire, de régression logistique, etc. Le paramètre θ serait alors un des β du modèle.

Supposons qu'on procède à K imputations, c'est-à-dire, qu'on construit K ensemble de données complets à partir de l'ensemble de données initial contenant des valeurs manquantes. On estime alors les paramètres du modèle séparément pour chacun des ensembles de données imputés. Soit $\hat{\theta}_k$, l'estimé du paramètre θ pour l'échantillon $k \in \{1, \dots, K\}$ et $\hat{\sigma}_k^2 = \text{Var}(\hat{\theta}_k)$ l'estimé de la variance de $\hat{\theta}_k$ produite par le modèle estimé.

L'estimation finale de θ , dénotée $\hat{\theta}$, est obtenue tout simplement en faisant la moyenne des estimations de tous les modèles, c'est-à-dire,

$$\hat{\theta} = \frac{\hat{\theta}_1 + \dots + \hat{\theta}_K}{K}.$$

Une estimation ajustée de la variance de $\hat{\theta}$ est

$$\begin{aligned}\text{Var}(\hat{\theta}) &= W + \frac{K+1}{K}B, \\ W &= \frac{1}{K} \sum_{k=1}^K \hat{\sigma}_k^2 = \frac{\hat{\sigma}_1^2 + \dots + \hat{\sigma}_K^2}{K}, \\ B &= \frac{1}{K-1} \sum_{k=1}^K (\hat{\theta}_k - \hat{\theta})^2.\end{aligned}$$

Ainsi, le terme W est la moyenne des variances et B est la variance entre les imputations. Le terme $(1+1/K)B$ est celui qui vient corriger le fait qu'on travaille avec des données imputées et non pas des vraies données en augmentant la variance estimée du paramètre.

C'est ici qu'on voit l'intérêt à procéder à de l'imputation multiple. Si on procédait à une seule imputation (même en ajoutant une part d'aléatoire pour essayer de reproduire la variabilité des données), on ne serait pas en mesure d'estimer la variance inter-groupe de l'estimateur. Notez que la formule présentée n'est valide que pour le cas unidimensionnel; l'estimation de la variance dans le cas multidimensionnel est différente (voir Little et Rubin, 2002).

Les procédures `mi` et `mianalyze` de **SAS** permettent d'effectuer de l'imputation multiple et d'obtenir des estimations valides des paramètres et de leurs écarts-type pour à peu près n'importe

quel modèle. Plus spécifiquement, la procédure `mi` permet d'examiner la configuration des valeurs manquantes et de procéder à l'imputation. De son côté, la procédure `mianalyze` permet de combiner les résultats des estimations provenant des échantillons combinés afin d'obtenir les estimations finales. La structure habituelle du code **SAS** est la suivante :

1. procédure `mi` pour obtenir des échantillons imputés.
2. procédure quelconque pour ajuster le modèle voulu (par exemple, `proc reg` ou `proc logistic`). Il faut ajuster les modèles séparément pour chaque échantillon imputé avec la commande `by`.
3. procédure `mianalyze` pour combiner les résultats.

La méthode d'imputation multiple possède l'avantage d'être applicable avec n'importe quel modèle sous-jacent. Une fois qu'on a des échantillons complets (imputés), on ajuste le modèle comme d'habitude. Mais une observation imputée ne remplacera jamais une vraie observation. Il faut donc faire tout ce qu'on peut pour limiter le plus possible les données manquantes.

Il faut utiliser son jugement. Par exemple, si la proportion d'observations perdues est petite (moins de 5%), ça ne vaut peut-être pas la peine de prendre des mesures particulières et on peut faire une analyse avec les données complètes seulement. S'il y a un doute, on peut faire une analyse avec les données complètes seulement et une autre avec imputations multiples afin de valider la première.

Si, à l'inverse, une variable secondaire cause à elle seule une grande proportion de valeurs manquantes, on peut alors considérer l'éliminer afin de récupérer des observations. Par exemple, si vous avez une proportion de 30% de valeurs manquantes en utilisant toutes vos variables et que cette proportion baisse à 3% lorsque vous éliminez quelques variables peu importantes pour votre étude (ou qui peuvent être remplacées par d'autres jouant à peu près le même rôle qui elles sont disponibles), alors vous pourriez considérer la possibilité de les éliminer. Il est donc nécessaire d'examiner la configuration des valeurs manquantes avant de faire quoi que ce soit. La procédure `mi` permet de faire cela, comme nous le verrons dans l'exemple qui suit.

7.3 Example d'application de l'imputation

On examine l'exemple de recommandations de l'association professionnelle des vachers de la section 5.4.

Le but est d'examiner les effets des variables X_1 à X_6 sur les intentions d'achat (Y) ; la base de données `missing1.sas7bdat` contient les observations et les commandes **SAS** se trouvent dans le fichier `missing1.sas`. Il s'agit des mêmes données que celles du fichier `logit1.sas7bdat` mais avec des valeurs manquantes.

Les points (.) indiquent des valeurs manquantes. Le premier sujet n'a pas de valeur manquante. Le deuxième a une valeur manquante pour X_1 (emploi) et X_4 (éducation), etc.

TABLE 7.1 – Tableau des configurations des données manquantes.

X_1	X_2	X_3	X_4	X_5	X_6	y
1	4	0	1	35	2	0
.	1	0	.	33	3	0
2	3	1	.	46	3	0
5	2	1	.	32	1	1
3	2	1	.	38	3	1
.	4	0	0	36	3	0
.	3	0	.	35	3	0
.	5	1	0	26	2	0
.	3	1	1	39	2	1
5	2	1	.	38	3	0

Une première façon de voir combien il y a de valeurs manquantes consiste à faire sortir les statistiques descriptives avec la procédure `means`. Ainsi, il y 192 valeurs manquantes pour X_1 , 48 pour X_2 et 184 pour X_4 . Les autres variables n'ont pas de valeurs manquantes, incluant la variable dépendante Y . La procédure unidimensionnelle nous permet seulement de voir combien il y a de valeurs manquantes variable par variable, séparément.

Variable	N	Nbre manquant
x1	308	192
x2	451	49
x3	500	0
x4	316	184
x5	500	0
x6	500	0
y	500	0

Afin d'avoir plus d'information au sujet de la configuration des valeurs manquantes, on peut utiliser la procédure `mi`. Comme on veut seulement voir comment sont distribuées les valeurs manquantes et non pas imputer les valeurs manquantes, on fixe `nimpute=0`.

```
proc mi data=multi.missing1 nimpute=0;
var y x1-x6;
run;
```

Groupe	y	x1	x2	x3	x4	x5	x6	Fréq	Pourcentage	Moyennes de groupe						
										y	x1	x2	x3	x4	x5	
1	X	X	X	X	X	X	X	180	36.00	0.427778	3.155556	2.711111	0.550000	0.255556	37.911111	2.216667
2	X	X	X	X	.	X	X	99	19.80	0.474747	3.171717	2.666667	0.474747	.	38.010101	2.262626
3	X	X	.	X	X	X	X	23	4.60	0.434783	3.130435	.	0.565217	0.130435	37.608696	2.217391
4	X	X	.	X	.	X	X	6	1.20	0.166667	3.666667	.	0.500000	.	37.666667	2.166667
5	X	.	X	X	X	X	X	99	19.80	0.474747	.	2.808081	0.565657	0.282828	38.373737	2.222222
6	X	.	X	X	.	X	X	73	14.60	0.479452	.	2.753425	0.452055	.	38.917808	2.068493
7	X	.	.	X	X	X	X	14	2.80	0.571429	.	0.357143	0.357143	41.214286	2.142857	
8	X	.	.	X	.	X	X	6	1.20	0.500000	.	0.333333	.	39.333333	2.166667	

Dans la partie de gauche, les X indiquent que la variable est présente dans cette configuration et le point indique qu'elle est manquante. Ainsi, il y a 180 sujets (36% de l'échantillon) avec aucune observation manquante. Il y en a 99 avec seulement X_4 manquante et ainsi de suite. On voit donc, par exemple, que pour 14 sujets, à la fois X_1 et X_2 sont manquantes. La partie de droite donne la moyenne des variables dans chaque configuration.

La recommandation d'usage est d'imputer au moins le pourcentage de cas incomplet, ici 72% donc 72 imputations. Si la procédure est trop gloutonne en calcul, on peut diminuer le nombre d'imputations, mais au moins cinq ou 10 réplications sont de mise.

On peut comparer l'inférence avec toutes les variables explicatives pour les données sans valeurs manquantes ($n = 500$ observations), avec les cas complets uniquement ($n = 180$ observations). Le Tableau 7.2 présente les estimations des paramètres du modèle de régression logistique s'il n'y avait pas eu de valeurs manquantes, avec les cas complets et les résultats de l'imputation multiple.

Revenons à présent à notre fichier avec des valeurs manquantes. Si on demande à **SAS** d'ajuster le modèle de régression logistique, il va par défaut retirer les observations qui ont au moins une valeur manquante pour une des variables nécessaires à l'analyse. Ainsi, le modèle va être ajusté avec seulement 180 observations. Cette analyse est présentée seulement pour montrer l'effet des valeurs manquantes.

Ceci est clairement indiqué dans la partie de la sortie qui est reproduite ci-après.

Nb d'observations lues	500
Nb d'observations utilisées	180
<hr/>	
Profil de réponse	
Valeur ordonnée	Fréquence totale
1	0 103
2	1 77

La probabilité modélisée est $y=1$.

Note: 320 observations were deleted due to missing values for the response or explanatory variables.

Il ne serait pas raisonnable de faire l'analyse avec seulement 180 observations et de laisser tomber les 320 autres. De plus, comme nous l'avons vu plus haut, ce n'est pas valide à moins que le processus ne soit MCAR. La partie du milieu du Tableau 7.2 présente les estimations obtenues. On voit par exemple que la variable X_3 (sexe), qui était clairement significative lorsqu'il n'y avait pas de valeurs manquantes, est maintenant non-significative (valeur- p de 0,068). La même chose est vraie pour quelques autres variables incluant $\mathbf{1}_{X_6=2}$. Il y a même pire, non seulement la variable $\mathbf{1}_{X_2=1}$ est passée de significative à non significative, mais en plus l'estimé de son paramètre a changé de signe.

Nous allons donc faire l'analyse avec l'imputation multiple, en prenant la méthode d'imputation par défaut dans la procédure **mi**. Le code **SAS** est

```
proc mi data=multi.missing1 out=outmi
    n impute=pctmissing seed=746382643;
var x11 x12 x13 x14 x21 x22 x23 x24 x3 x4 x5 x61 x62;
run;

proc logistic data=outmi outest=outlogistic
    covout descending noprint;
model y(ref='0') = x11 x12 x13 x14 x21 x22 x23 x24 x3 x4 x5 x61 x62;
by _imputation_;
run;

proc mianalyze data=outlogistic;
var intercept x11 x12 x13 x14 x21 x22 x23 x24 x3 x4 x5 x61 x62;
run;
```

Avec la procédure `mi` on demande un nombre d'imputations égal au nombre de données manquantes, mais au moins cinq et au plus 50 (`nimpute=pctmissing`). Le fichier de sortie (`outmi`) va contenir ces m ensembles de données pour un total de $m \times 500$. Dans ce fichier, l'ensemble est identifié par la variable `_imputation_`, qui prend des valeurs de 1 à m .

On estime pour chaque jeu de données un modèle de régression logistique grâce à la commande `by _imputation_`. Les paramètres estimés sont placés dans le fichier `outlogistic`. On sauvegarde aussi l'estimé de la matrice de variance des paramètres avec l'option `covout`.

Finalement, la procédure `mianalyze` est utilisée pour combiner les résultats de ces m modèles et pour fournir les estimations finales des paramètres. Une partie de la sortie est présentée

Informations sur la variance (50 Imputations)								
Variable	Variance			Augmentation relative		Informations manquantes		Efficacité relative
	Inter	Dans	Total	DDL	dans variance	fraction		
x11	0.000076460	0.000153	0.000231	186.41	0.509794	0.340719	0.993232	
x12	0.000293	0.000367	0.000666	128.76	0.814936	0.453495	0.991012	
x13	0.000147	0.000371	0.000521	221.59	0.403781	0.290032	0.994233	
x14	0.000214	0.000434	0.000652	188.66	0.501845	0.337166	0.993302	
x21	0.000018962	0.000155	0.000174	397.62	0.124908	0.111486	0.997775	
x22	0.000037558	0.000474	0.000512	436.93	0.080814	0.074982	0.998503	
x23	0.000048376	0.000432	0.000482	407.16	0.114176	0.102860	0.997947	
x24	0.000029246	0.000276	0.000306	412.68	0.107998	0.097821	0.998047	
x4	0.000132	0.000385	0.000519	244.85	0.349518	0.261015	0.994807	

Paramètres estimés (50 Imputations)										
Variable	Moyenne	Erreur type	Intervalle de confiance à 95%	DDL	Minimum	Maximum	Mu0	Moyenne=Mu0	t pour H0:	
x11	0.078481	0.015198	0.048499 - 0.108463	186.41	0.062555	0.095177	0	5.16	<.0001	
x12	0.236297	0.025800	0.185251 - 0.287344	128.76	0.187639	0.271770	0	9.16	<.0001	
x13	0.236459	0.022834	0.191459 - 0.281460	221.59	0.214142	0.262100	0	10.36	<.0001	
x14	0.324745	0.025529	0.274387 - 0.375104	188.66	0.299024	0.353519	0	12.72	<.0001	
x21	0.082097	0.013198	0.056151 - 0.108044	397.62	0.074961	0.091278	0	6.22	<.0001	
x22	0.378534	0.022635	0.334046 - 0.423021	436.93	0.361310	0.396048	0	16.72	<.0001	
x23	0.315150	0.021943	0.272013 - 0.358287	407.16	0.294572	0.328372	0	14.36	<.0001	
x24	0.166203	0.017494	0.131814 - 0.200591	412.68	0.155876	0.179026	0	9.50	<.0001	
x4	0.267935	0.022789	0.223047 - 0.312824	244.85	0.248519	0.296367	0	11.76	<.0001	

On peut remarquer que la précision est systématiquement meilleure avec l'imputation multiple; les erreurs-type pour l'imputation multiple sont plus petits que celle du modèle qui retire les données incomplètes.

TABLE 7.2 – Estimés, erreurs-type et valeurs-*p* des paramètres, avec les 500 données complètes (gauche), avec les 180 cas complets (milieu) et imputation multiple (droite).

	$\hat{\beta}$	se($\hat{\beta}$)	valeur- <i>p</i>	$\hat{\beta}$	se($\hat{\beta}$)	valeur- <i>p</i>	$\hat{\beta}$	se($\hat{\beta}$)	valeur- <i>p</i>
Intercept	-6.888	1.022	0.000	-5.251	1.700	0.002	-6.575	1.037	0.000
$x_1 = 1$	0.358	0.483	0.458	-0.086	0.850	0.919	0.546	0.540	0.312
$x_1 = 2$	-0.468	0.371	0.208	-0.568	0.656	0.387	-0.125	0.446	0.779
$x_1 = 3$	-0.311	0.350	0.374	-0.471	0.658	0.474	0.072	0.436	0.869
$x_1 = 4$	-0.317	0.402	0.431	-0.926	0.738	0.210	-0.043	0.485	0.930
$x_2 = 1$	1.331	0.597	0.026	-0.735	1.144	0.521	1.098	0.647	0.090
$x_2 = 2$	1.148	0.501	0.022	0.460	0.908	0.613	1.032	0.545	0.059
$x_2 = 3$	0.773	0.483	0.109	-0.409	0.886	0.644	0.524	0.520	0.314
$x_2 = 4$	-1.109	0.542	0.041	-2.741	1.020	0.007	-1.043	0.568	0.066
x_3	1.349	0.260	0.000	0.802	0.440	0.068	1.190	0.268	0.000
x_4	1.830	0.298	0.000	2.254	0.578	0.000	1.516	0.372	0.000
x_5	0.109	0.019	0.000	0.107	0.033	0.001	0.104	0.019	0.000
$x_6 = 1$	2.412	0.376	0.000	2.233	0.665	0.001	2.261	0.380	0.000
$x_6 = 2$	1.045	0.249	0.000	0.831	0.437	0.057	0.995	0.250	0.000

On voit que la variable X_3 (sexe) est significative avec l'imputation multiple. Son paramètre estimé est 1,202, comparativement à 1,349 s'il n'y avait pas eu de valeurs manquantes. La précision dans l'estimation avec l'imputation multiple est seulement un peu moins bonne (erreur-type de 0,282) que celle s'il n'y avait pas eu de manquantes (erreur type de 0,26). Le paramètre de $1_{X_6=2}$ redevient aussi significatif, alors qu'il ne l'était plus si on retirait les manquantes. Il est peu probable que les données soit MCAR et donc les résultats de l'analyse des cas complets seraient biaisés.

7.4 Valeurs manquantes dans un contexte de prédiction

Nous avons vu que l'imputation multiple permet de corriger les écarts-type des paramètres estimés afin d'obtenir une inférence valide. Mais cette fois-ci, le but n'est pas d'estimer un modèle afin de tester formellement certaines hypothèses, mais plutôt de développer un modèle pour obtenir des prédictions. Dans ce cas, l'imputation multiple peut aussi être utile.

Nous allons revenir une dernière fois sur l'exemple de ciblage de clients pour l'envoi d'un catalogue. Rappelez-vous qu'on a un échantillon d'apprentissage de 1000 clients pour lesquels la variable *yachat* est disponible (est-ce que le client a acheté quelque chose lorsqu'on lui a envoyé un catalogue). Nous avons développé des modèles avec ces 1000 clients afin de décider à qui, parmi les 100 000 clients restants, envoyer le catalogue. Nous avions alors utilisé des données sans valeurs

manquantes. Cette fois-ci, nous allons faire comme s'il y avait des valeurs manquantes pour les variables explicatives à la fois dans l'échantillon d'apprentissage mais aussi dans l'échantillon à prédire. Nous allons chercher à développer un modèle de régression logistique pour $P(yachat = 1)$. Plusieurs approches sont possibles et il n'est pas clair à priori laquelle est la meilleure. Voici la description de deux approches.

Approche 1 :

- 1) Obtenir K ensembles de données complets par imputations multiples (simultanément pour les échantillons tests et d'apprentissage).
- 2) Pour chaque ensemble de données complet,
 - a. Faire une sélection de variables
 - b. Obtenir les estimations de $P(yachat = 1)$
- 3) Pour chaque observation dans les deux échantillons faire la moyenne des K estimations de $P(yachat = 1)$ de manière à avoir une seule prédiction par observation de la probabilité d'achat.
- 4) Trouver le meilleur point de coupure avec les probabilités calculées en 3) pour l'échantillon d'apprentissage.
- 5) Assigner les observations de l'échantillon à prédire avec ce point de coupure en utilisant les probabilités calculées en 3) pour les données de l'échantillon test.

Approche 2 :

- 1) Obtenir K ensembles de données complets par imputations multiples (simultanément pour les échantillons tests et d'apprentissage).
- 2) Pour chaque ensemble de données complet,
 - a. Faire une sélection de variables
 - b. Trouver le meilleur point de coupure par validation-croisée
 - c. Obtenir les prédictions (0 ou 1) pour l'échantillon à scorer, avec ce point de coupure.
- 3) Assigner l'observation à la classe majoritaire (celle qui a le plus de votes parmi zéro ou un pour les K prédictions) pour chaque observation à prédire.

Nous allons seulement utiliser une approximation de la première approche ici et ignorer les valeurs de $yachat$ et $ymontant$ lors de l'imputation car ces dernières sont manquantes dans l'échantillon test et on se trouverait imputer à partir de modèles différentes dans les deux échantillons (test et apprentissage). Le code se trouve dans le fichier `manquantes2_prevision.sas` et les données dans `dbmmissing.sas7bdat`. Des informations sur les valeurs manquantes dans l'échantillon d'apprentissage peuvent être obtenues à l'aide des procédures `means` et `mi`. Il y a des valeurs manquantes dans chaque variable : par exemple, 99 valeurs manquantes pour X_1 . Globalement, seulement 201 des 1000 clients n'ont aucune valeur manquante sur les 10 variables. Il y a 164 configurations différentes de valeurs manquantes.

En utilisant l'approche 1 présentée plus haut, nous allons imputer simultanément les valeurs manquantes pour l'échantillon d'apprentissage et l'échantillon test à prédire avec $K = 5$ échantillons imputés. La méthode de sélection de variables utilisée est la procédure séquentielle classique avec 0,05 comme critère d'entrée et de sortie. De plus, afin de simplifier le tout, le point de coupure a été fixé à 0,14 (celui que l'on avait obtenu dans le cas où il n'y a pas de valeurs manquantes) et non pas estimé par validation-croisée.

Il s'avère que le revenu net, sur les 100 000 clients restants, aurait été de 926 917\$. S'il n'y avait pas eu de valeurs manquantes, la sélection basée sur une procédure séquentielle classique aurait généré un revenu net de 969 350\$. Les données manquantes rendent plus difficile le développement du modèle. Mais on fait quand même encore beaucoup mieux que la stratégie de référence qui consiste à envoyer le catalogue aux 100 000 clients, qui aurait généré un revenu net de 601 112\$.