Analyse multidimensionnelle appliquée

Denis Larocque, Léo Belzile

Version du 2020-02-14

Table des matières

Chapitre 1

Introduction

1.1 Survol du cours

1.1.1 Analyse factorielle exploratoire

On dispose de p variables $X_1, \ldots, X_p$. 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 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 :

— Détecter les faillites des clients (ou des entreprises)
— Cibler les clients qui seront intéressés par une offre promotionnelle
— Détecter 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.

1.1.3 Régression logistique

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

Exemple 1.2. 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 $\boldsymbol{X}$ 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 $\boldsymbol{X}$.

1.1.4 Analyse de regroupements

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

Exemple 1.3. 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 marketing [1]

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.

1.1.5 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 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 $X_1, \ldots, X_p$.

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

Chapitre 2

Analyse factorielle exploratoire

2.1 Introduction

On dispose de p variables $X_1, \ldots, X_p$.

— 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 : Habileté 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 d'au moins 10 fois le nombre de variables.

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 ie 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

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

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 $i = 1, \ldots, n$.

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 commandes **SAS** (ainsi que plusieurs commentaires) pour faire les analyses se trouvent dans le fichier `MATH60602_cours3.sas`. Les statistiques descriptives ainsi que la matrice des corrélations sont obtenues en exécutant les lignes suivantes :

```
proc corr data=multi.factor2;
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 $X_1, \ldots, X_p$. Le modèle d'analyse factorielle fait l'hypothèse que ces variables dépendent linéairement d'un plus petit nombre m de variables aléatoires, $F_1, \ldots, F_m$, appelées facteurs communs et de p termes d'erreurs (ou facteurs spécifiques) $\varepsilon_1, \ldots, \varepsilon_p$, de moyenne $\mathsf{E}(\varepsilon_i) = 0$ et de

variance $\mathsf{Var}(\varepsilon_i) = \psi_i$ pour $i = 1, \ldots, p$. Spécifiquement, le modèle est

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

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

Les espérances (μ_i), les chargements (l_{ij}) et les variances (ψ_i) sont des quantités fixes, mais inconnues, tandis que les facteurs communs (F_i) et spécifiques (ε_i) sont des variables aléatoires non observables que l'on assume non corrélée aux facteurs F et entre elles.

Des hypothèses supplémentaires sont nécessaires afin de pouvoir utiliser ce modèle (contraintes d'identifiabilité des paramètres). Sans entrer dans les détails, mentionnons que l'une de ces hypothèses est que les facteurs sont non corrélés.

De plus, si les variables ont été préalablement standardisées de telle sorte que $\mathsf{E}(X_i) = 0$ et $\mathsf{Var}(X_i) = 1$ (note : ceci revient à utiliser la matrice de corrélation des observations dans l'analyse ce qui est fait par défaut dans **SAS**), alors $\mathsf{Cor}\left(X_i, F_j\right) = l_{ij}$, 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 $X_1, \ldots, X_p$ 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 $p(m+1)$ paramètres (p variances spécifiques et pm chargements). 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 l_{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 revanche, l'estimation des paramètres requiert une optimisation numérique qui peut être délicate selon les cas de figure.

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 prêt (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.factor2
 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	Khi-2	Pr > 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			

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.

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

m	AIC	BIC	valeur-p
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 2.1.

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. 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

		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 2.1 – Estimés des chargements pour trois facteurs avec rotation varimax

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.

— le diagramme d'éboulis : un graphique des valeurs propres ordonnées de la plus grande à la plus petite en fonction de $1,\ldots,p$. 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.factor2 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 2.2. Pour le diagramme d'éboulis de la figure 2.3, 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.

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,

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

	Valeur propre	Différence	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 2.2 – Valeurs propres et proportion de variance

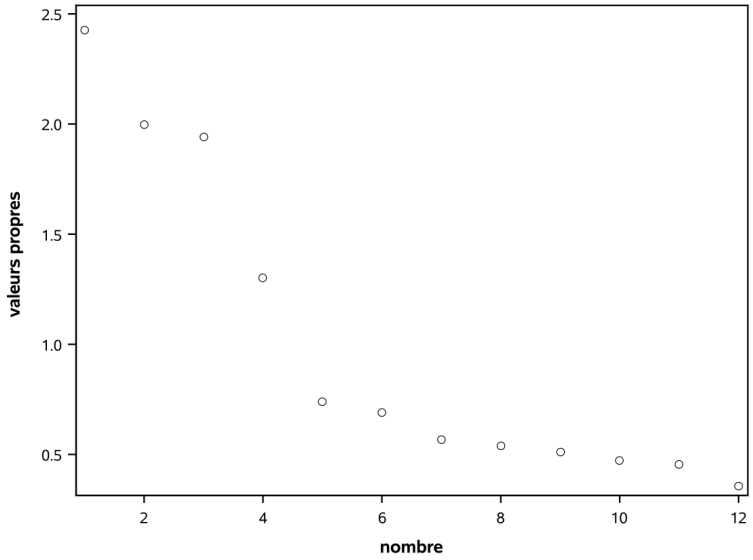


FIGURE 2.3 – Diagramme d'éboulis

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

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

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 ie 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.factor2 alpha;
var x4 x8 x11;
run;
/* pour le facteur produits */
proc corr data=multi.factor2 alpha;
var x3 x6 x9 x12;
run;
/* pour le facteur paiement */
```

```
proc corr data=multi.factor2 alpha;
var x2 x7 x10;
run;
/* pour le facteur prix */
proc corr data=multi.factor2 alpha;
var x1 x5;
run;
```

Coefficient Alpha de Cronbach

Variables	Alpha
Brut	0.780524
Normalisé	0.780611

Coefficient Alpha de Cronbach avec variable supprimée

	Variables brutes		Variables standardisées		
Variable supprimée	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 2.4 – 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 retirait une variable à la fois de l'échelle. Tout est donc cohérent. Les échelles provenant des facteurs *service*, *produits* et *paiement*, sont satisfaisantes. Ces facteurs sont identifiés à la fois dans la solution à quatre, mais aussi dans la solution à troisfacteurs. 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 ordinales

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 ordinales (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

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 2.5 – Alpha de Cronbach pour le facteur *produits*.

Coefficient Alpha de Cronbach	
Variables	**Alpha**
Brut	0.727492
Normalisé	0.734783

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 2.6 – Alpha de Cronbach pour le facteur *paiement*.

Coefficient Alpha de Cronbach	
Variables	**Alpha**
Brut	0.545634
Normalisé	0.545805

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

de corrélation linéaire. Par exemple, on peut utiliser la corrélation polychorique, qui est une mesure de corrélation entre deux variables ordinales. La corrélation tétrachorique correspond au cas spécial de deux variables binaires.

Ma suggestion est d'utiliser la corrélation linéaire ordinaire avec des variables ordinales (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.factor2 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 2.8. Les facteurs obtenus sont les mêmes qu'en utilisant les corrélations linéaires.

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

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 2.8 – Chargements estimés pour la corrélation polychorique

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.factor2
maxiter=500 flag=.3 hey;
var x1-x12;
run;
```

La matrice des corrélations entres facteurs est donnée dans la figure 2.9 et les chargements sont présentés dans la figure 2.10. 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 2.9 – Corrélation interfacteurs pour rotation varimax oblique

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 2.10 – Chargements avec rotation oblique varimax

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

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

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

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 ($\widehat{\gamma}$) plus élevés. Par contre, les scores factoriels ne sont pas uniques car ils dépendent des chargements utilisés (et 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.

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 2.11. 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 2.12. On remarque la forte corrélation entre le score factoriel et les échelles construites avec les

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 2.11 – Coefficients du score normalisés

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 2.12 – Corrélation entre scores et échelles

Chapitre 3

Sélection de variables et de modèles

3.1 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 $X_1, \ldots, X_p$. Le terme variable ici inclut autant des variables distinctes que des transformations d'une ou plusieurs variables.

Par exemple, supposons que les variables `age`, `sexe` 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(age), etc. Nous pourrions aussi considérer des termes d'interactions entre les variables, comme age·revenu ou age·revenu·sexe. 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.

3.2 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 $X_1, \ldots, X_p$. 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 guiderons 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.

3.3 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. Le modèle ajusté est 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 a été 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. Encore plus généralement, il est préférable d'avoir un modèle un peu trop complexe que d'avoir un modèle trop simple.

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.

3.4 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.

3.5 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 $(Y, X_1, \ldots, Xp)$ et que l'on a ajusté un modèle $\widehat{f}(X_1, \ldots, X_p)$, avec ces données, pour prédire une variable continue Y.

Ce modèle peut être un modèle de régression linéaire,

$$\widehat{f}(X_1, \ldots, X_p) = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p$$

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*),

$$\mathsf{EMQ} = \mathsf{E}\left[\left\{(Y - \widehat{f}(X_1, \ldots, X_p)\right\}^2\right]$$

lorsque $(Y, X_1, \ldots, X_p)$ est choisi au hasard dans la population. Cette quantité mesure l'erreur (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 ne connaît pas 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 EMQ avec l'erreur quadratique moyenne de l'échantillon d'apprentissage (*training mean squared error*),

$$\widehat{\mathsf{EMQ}}_a = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widehat{f}(X_{i1}, \ldots, X_{ip})\right\}^2.$$

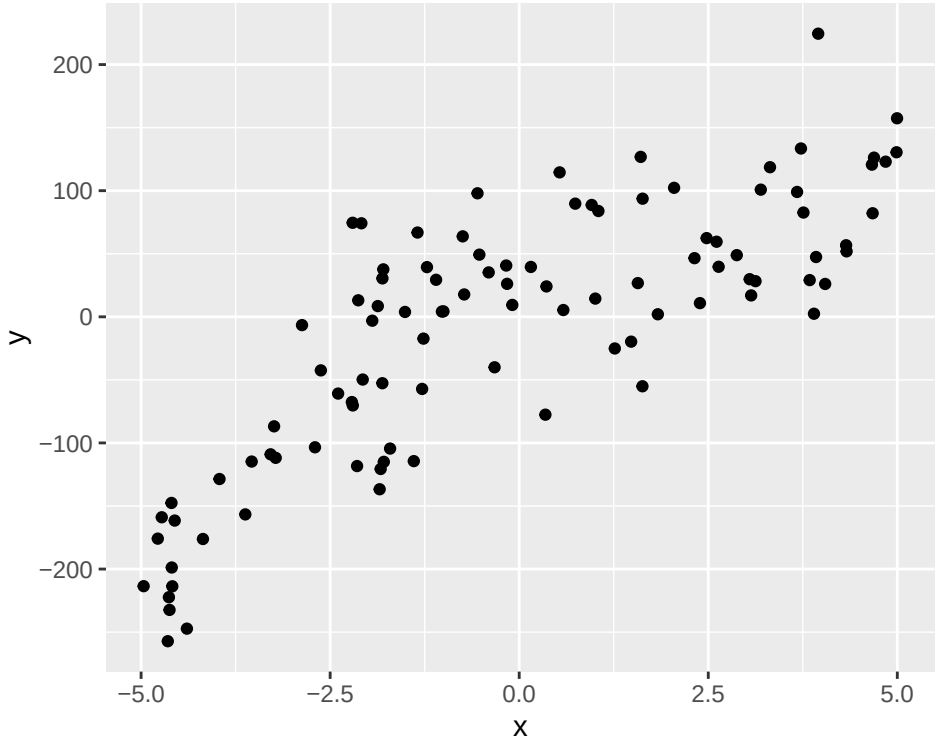
Cette quantité est tout simplement l'équivalent du EMQ, mais est calculée en utilisant seulement notre échantillon.

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

3.5.1 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.xls` 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 $Y = \beta_0 + \beta_1 X + \cdots + \beta_k X^k + \varepsilon$. 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.

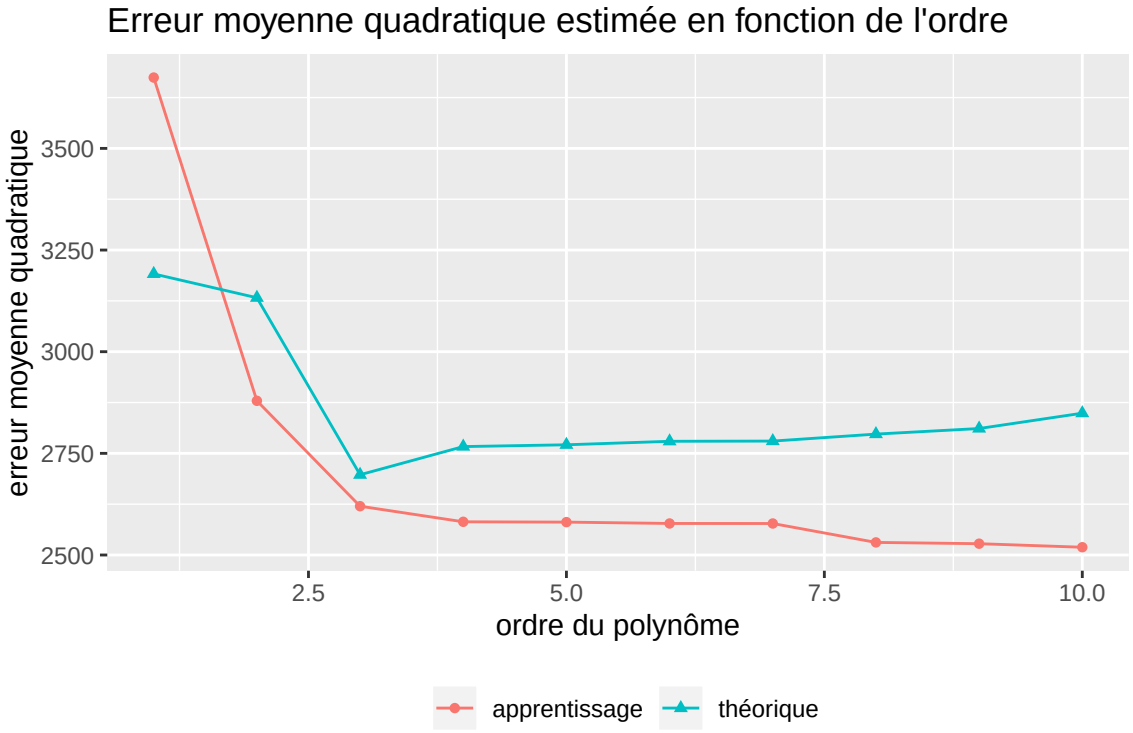


Il s'agit de l'échantillon d'apprentissage. Ces données ont été obtenues par simulation et le vrai modèle sous-jacent (celui qui a généré les données) est le modèle cubique, c'est-à-dire le modèle d'ordre $k = 3$.

Afin de simuler une population, j'ai généré selon le même modèle 100 000 observations supplémentaires. Ces observations ne vont pas servir à estimer les modèles mais seulement à évaluer leur performance afin d'avoir une estimation sans biais. Ces données se trouvent dans `selection1_test.xls`

J'ai ajusté tour à 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 moyenne quadratique d'apprentissage pour chacun de ces modèles. J'ai ensuite utilisé ces modèles afin de prédire les 100 000 autres observations (la population) et calculé les 100 000 observations de l'échantillon test pour obtenir une très bonne approximation de l'erreur quadratique moyenne de généralisation.

Voici le graphe de l'$\widehat{\text{EMQ}}_a$ et de l'EMQ de généralisation en fonction de l'ordre (k) du modèle utilisé.



On voit clairement que l'$\widehat{\text{EMQ}}_a$ diminue en fonction de l'ordre sur l'échantillon d'apprentissage. C'est-à-dire, plus le modèle est complexe, plus l'erreur observée sur l'échantillon d'apprentissage est petite. Mais cela est trompeur. La courbe EMQ donne l'heure juste. 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 car il s'agit du modèle que j'ai utilisé pour générer les données. On peut aussi remarquer d'autres éléments intéressants. Premièrement, on obtient un bon gain en performance (EMQ) 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'EMQ 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

$$\widehat{f}(X_1,\ldots,X_p) = \widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p.$$

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

$$\mathsf{SCE} = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p)^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

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 $\widehat{\beta}_{p+1} = 0$ et on retombera sur le modèle sans cette variable. C'est pourquoi, la quantité $\widehat{\mathsf{EMQ}}_a = \mathsf{SCE}/n$ 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 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

$$R^2 = \{\mathsf{cor}(\boldsymbol{y}, \widehat{\boldsymbol{y}})\}^2 = 1 - \frac{\mathsf{SCE}}{\mathsf{SCT}},$$

où $\mathsf{SCT} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ 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 nécessairement 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 $\widehat{\mathsf{EMQ}}_a$ pour tenir compte de la complexité du modèle (AIC, BIC).
— tenter d'estimer l'EMQ directement sur d'autres données (hors échantillon) en utilisant des méthodes de rééchantillonnage, notamment la validation croisée et la division de l'échantillon.

3.6 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 $\mathscr{L}$ et la log-vraisemblance ℓ mesurent l'adéquation du modèle.

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

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

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

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

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 méchaniquement diminuer, et donc la quantité $n\ln(\mathsf{SCE}/n)$ 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 choisissent le bon model 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 parsimonie 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

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

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 :

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

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

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

Mesures d'adéquation du modèle linéaire et estimés de l'erreur

EMQ

$\widehat{\mathsf{EMQ}}_a$

R^2

R_a^2

AIC

BIC

1

3191.29

3674.20

0.65

0.65

1110.70

1118.51

2

3132.67

2879.24

0.73

0.72

1088.32

1098.74

3

2697.40

2620.05

0.75

0.74

1080.88

1093.91

4

2766.68

2581.70

0.75

0.74

1081.41

1097.04

5

2771.05

2580.86

0.75

0.74

1083.38

1101.61

6

2779.66

2577.60

0.75

0.74

1085.25

1106.09

7

2780.21

2577.49

0.75

0.74

1087.24

1110.69

8

2797.35

2531.00

0.76

0.74

1087.42

1113.48

9

2811.07

2527.85

0.76

0.73

1089.30

1117.96

10

2848.81

2519.14

0.76

0.73

1090.95

1122.22

Les colonnes EMQ et $\widehat{\text{EMQ}}_a$ ont déjà été expliquées à la section précédente et ont été représentées graphiquement. On voit que le augmente toujours au fur et à mesure qu'on ajoute une variable (augmente l'ordre du polynôme). Les critères AIC et BIC choisissent le modèle cubique ($k = 3$), c'est-à-dire le bon modèle. Le R^2 ajusté quant à lui choisit le modèle d'ordre 4 (qui est le deuxième meilleur selon le EMQ). 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 $\widehat{\text{EMQ}}_a$.

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.

3.7 Division de l'échantillon et validation croisée

La deuxième grande approche après celle consistant à pénaliser le $\widehat{\text{EMQ}}_a$ consiste à tenter d'estimer le EMQ directement. Nous allons voir deux telles méthodes ici, la 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.

3.7.1 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 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'EMQ) 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'EMQ 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 implanter. 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. 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.

3.7.2 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 $P_1, P_2, \ldots, P_K$ de taille 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 repondérer 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 prédite 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 varibles catégorielles sont présentes. Un échantillonage stratifié permet de pallier à cette lacune, mais ce problème se présente en pratique quand certaines classes ont peu d'observations.

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 et on recommence pour chaque observation. Pour les modèles linéaires, il existe des formules explicites qui nous permettent d'éviter d'ajuster n régressions par moindre carrés.

Le fichier `selection3_cv.sas` contient une macro SAS permettant de faire une validation croisée pour un modèle de régression linéaire. Revenons à notre exemple où une seule variable explicative est disponible et où l'on cherche à déterminer un bon modèle polynomial. Le Tableau 3.7.2 est le même que le Tableau 3.6 mais avec une colonne en plus, la dernière, VC($K = 10$). Il s'agit des estimations du EMQ 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.

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

EMQ

$\widehat{\text{EMQ}}_a$

R^2

R^2_a

AIC

BIC

VC($K = 10$)

1

3191.29

3674.20

0.65

0.65

1110.70

1118.51

3675.37

2

3132.67

2879.24

0.73

0.72

1088.32

1098.74

2897.94

3

2697.40

2620.05

0.75

0.74

1080.88

1093.91

2675.51

4

2766.68

2581.70

0.75

0.74

1081.41

1097.04

2666.16

5

2771.05
2580.86
0.75
0.74
1083.38
1101.61
2711.11
6
2779.66
2577.60
0.75
0.74
1085.25
1106.09
2757.13
7
2780.21
2577.49
0.75
0.74
1087.24
1110.69
2787.95
8
2797.35
2531.00
0.76
0.74
1087.42
1113.48
2845.78

9

2811.07

2527.85

0.76

0.73

1089.30

1117.96

2895.61

10

2848.81

2519.14

0.76

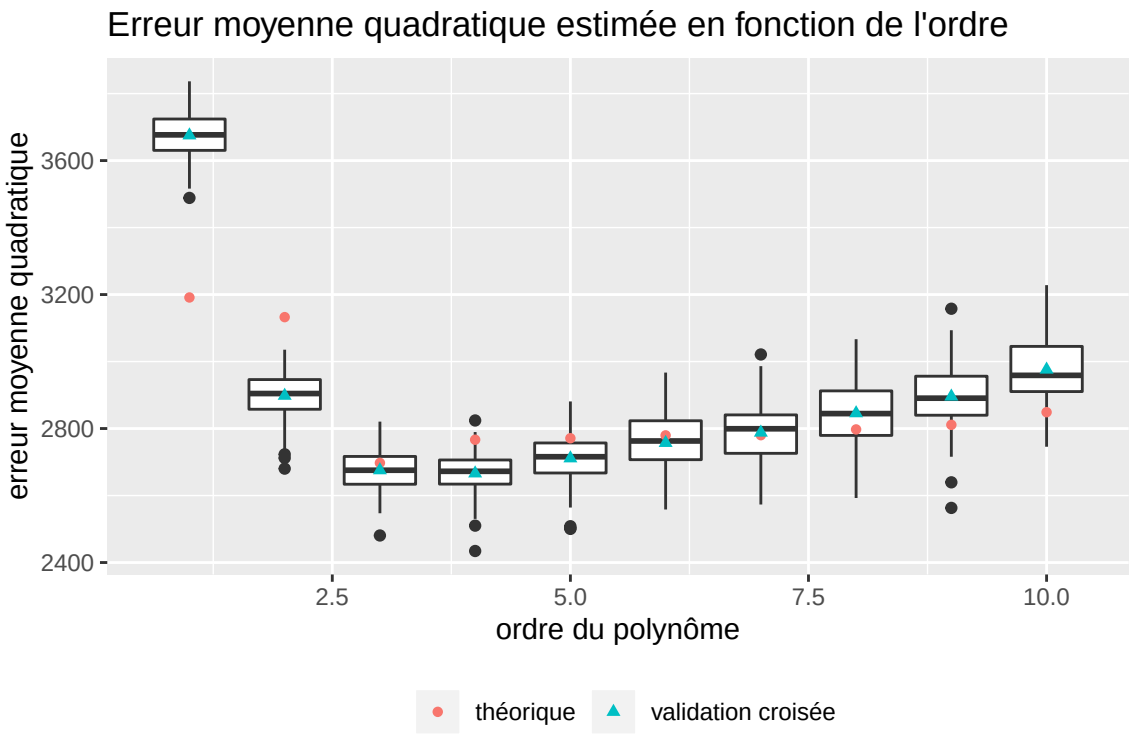
0.73

1090.95

1122.22

2976.04

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 EMQ. On voit que l'estimation par validation croisée suit assez bien la forme du EMQ (qu'il est supposé estimer). Les boîtes à moustache permettent d'apprécier la variabilité des estimés de l'erreur moyenne quadratique telles qu'estimée par validation croisée avec 10 groupes.



3.8 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 fait au moins un achat au cours de la dernière année. Il y a 101 000 clients dans la base de données. La première étape de la stratégie consiste à envoyer le catalogue à un échantillon de 1000 clients. Par la suite, un modèle sera construit avec ces 1000 clients afin de cibler lesquels des 100 000 clients restants seront choisis pour recevoir le catalogue. Les 1000 clients forment l'échantillon d'apprentissage. 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`. Lors d'une vraie application, nous aurions seulement les valeurs des variables cibles `yachat` et `ymontant` pour l'échantillon d'apprentissage (car eux seuls ont reçu le catalogue). Dans notre exemple, elles sont fournies pour tous les clients afin de pouvoir évaluer la performance des différentes stratégies testées. Les modèles seront déterminés (sélectionnés et ajustés) en utilisant seulement l'échantillon d'apprentissage (1000 clients). Les 100 000 autres clients serviront d'échantillon test 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 test nous donnera donc l'heure juste quant aux mérites des différentes approches que nous allons comparer.

Voici d'abord des statistiques descriptives pour l'échantillon d'apprentissage.

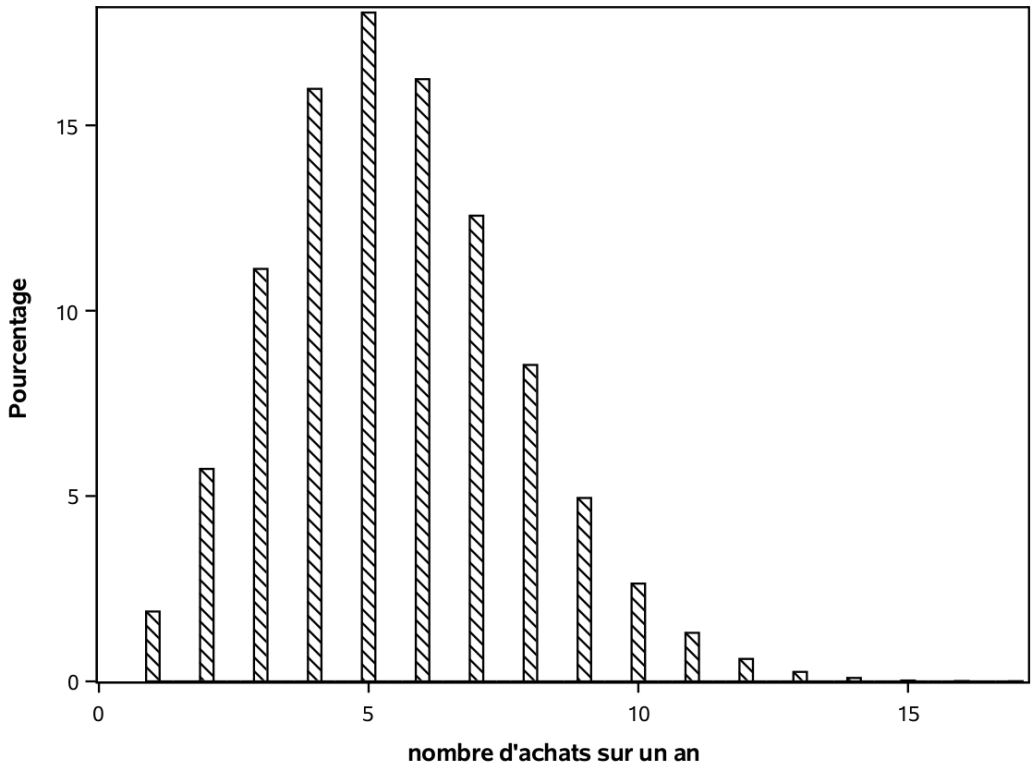
sexe	
x1	Fréquence
0	534
1	466

revenu	
x3	Fréquence
1	397
2	337
3	266

region	
x4	Fréquence
1	216
2	185
3	216
4	191
5	192

conjoint	
x5	Fréquence
0	575
1	425

Il y a donc 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.

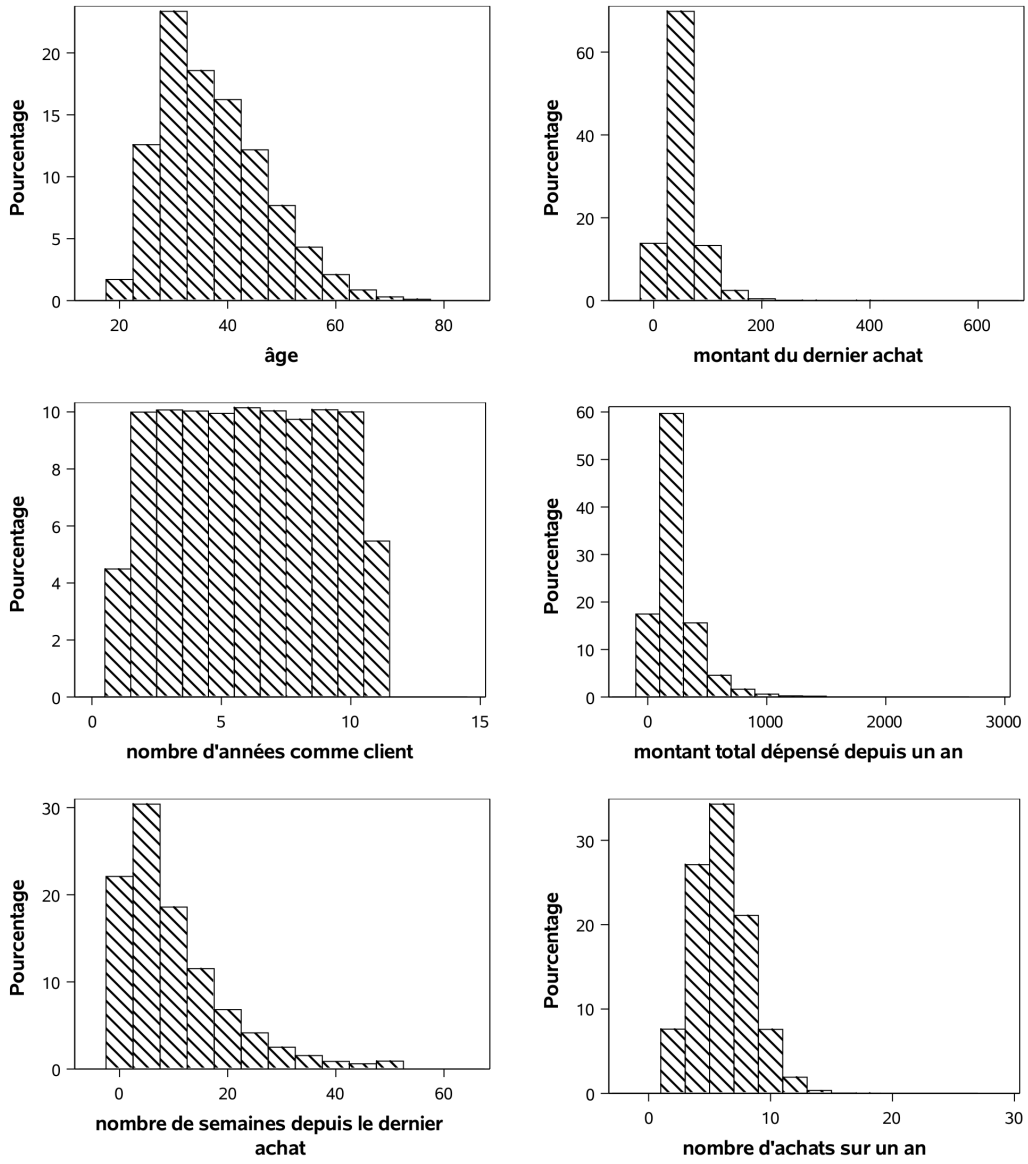


Le nombre d'achats différents depuis un an par ces clients varie entre 1 et 14. Un peu plus de la moitié (51,4%) ont fait 5 achats ou moins. Parmi les 1000 clients de l'échantillon d'apprentissage, 210 ont acheté quelque chose dans le catalogue. La variable yachat sera l'une des variables que nous allons 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 229,3$ 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.

a) 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 Pr(`yachat` = 1). Comme la variable `yachat` 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 `ymontant`. 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 :

$$\begin{aligned}\mathsf{E}(\texttt{ymontant}) &= \mathsf{E}(\texttt{ymontant} \mid \texttt{yachat} = 1)\,\mathsf{P}(\texttt{yachat} = 1) \\ &\quad + \mathsf{E}(\texttt{ymontant} \mid \texttt{yachat} = 0)\,\mathsf{P}(\texttt{yachat} = 0) \\ &= \mathsf{E}(\texttt{ymontant} \mid \texttt{yachat} = 1)\,\mathsf{P}(\texttt{yachat} = 1)\end{aligned}$$

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 $\mathsf{E}(\texttt{ymontant} \mid \texttt{yachat} = 1)$ et $\mathsf{P}(\texttt{yachat} = 1)$, pour ensuite les combiner et avoir une estimation de $\mathsf{E}(\texttt{ymontant})$. Le développement du modèle pour $\mathsf{E}(\texttt{ymontant} \mid \texttt{yachat} = 1)$ peut se faire avec la régression linéaire, en utilisant seulement les clients qui ont acheté dans l'échantillon d'apprentissage, car ymontant est une variable continue dans ce cas. Le développement du modèle pour $\mathsf{P}(\texttt{yachat} = 1)$ 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 $\mathsf{E}(\texttt{ymontant} \mid \texttt{yachat} = 1)$, le montant d'argent dépensé par les clients qui ont acheté quelque chose.

Le fichier `prepare_DBM.sas` contient des commandes afin de préparer les données aux analyses qui seront présentées dans les sections qui suivent. En particulier, nous avons 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. 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 qui suit.

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

En pratique, il suffit d'incorporer les indicatrices (x31 et x32) dans le modèle comme variables explicatives et de ne plus utiliser la variable originale x3. On peut aussi procéder ainsi pour la variable x4, en créant quatre indicatrices.

3.9 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 $2^p = 2 \times 2 \times \cdots \times 2$ (p fois) modèles en tout à considérer. Ce nombre augmente très rapidement comme en témoigne le tableau 3.9.

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

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). II 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.

3.10 Recherche automatique de tous les sous-ensembles

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 ($k = 1, \ldots, 14$) qui a le plus grand R^2 ; voir `selection2_all_subset.sas` pour plus de détails.

```
proc reg data=trainymontant;
model ymontant=x1 x2 x31 x32 x41 x42 x43 x44 x5 x6
  x7 x8 x9 x10 / selection=rsquare best=1 aic bic;
run;
```

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`.

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 $= 1 - \mathsf{SCE}/\mathsf{SCT}$ 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 $\mathsf{AIC} = n(\ln(\mathsf{SCE}/n)) + 2p$, 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.

Nombre dans le modèle	R carré	AIC	SBC	Variables du modèle
1	0.3982	981.2487	987.94291	x5
2	0.6364	877.4520	887.49336	x5 x6
3	0.7947	759.4210	772.80942	x31 x5 x6
4	0.8345	716.1137	732.84928	x31 x5 x6 x10
5	0.8521	694.4859	714.56857	x1 x31 x5 x6 x10
6	0.8675	673.3848	696.81454	x1 x31 x5 x6 x7 x10
7	0.8740	664.9471	691.72400	x1 x31 x5 x6 x7 x8 x10
8	0.8763	663.1003	693.22422	x1 x31 x44 x5 x6 x7 x8 x10
9	0.8790	660.3748	693.84585	x1 x2 x31 x44 x5 x6 x7 x8 x10
10	0.8803	660.1462	696.96442	x1 x2 x31 x44 x5 x6 x7 x8 x9 x10
11	0.8810	660.8715	701.03679	x1 x2 x31 x43 x44 x5 x6 x7 x8 x9 x10
12	0.8815	661.9932	705.50555	x1 x2 x31 x41 x42 x43 x44 x5 x6 x7 x8 x10
13	0.8825	662.2835	709.14305	x1 x2 x31 x41 x42 x43 x44 x5 x6 x7 x8 x9 x10
14	0.8827	663.8411	714.04768	x1 x2 x31 x32 x41 x42 x43 x44 x5 x6 x7 x8 x9 x10

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 allons utiliser les 100 000 autres clients pour évaluer la performance réelle des modèles qui nous sont suggérés par nos différents critères. En pratique, nous ne pourrions pas faire cela car la valeur de la variable cible ne serait pas connue pour ces clients. En fait, dans une vraie application, nous utiliserions plutôt les modèles pour obtenir des prédictions pour les clients à « scorer ». Les valeurs des variables cibles pour les 100 000 clients nous permettront de voir à quel point différentes stratégies auraient été profitables si elles avaient été mises en place. 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 moyenne quadratique (théorique) des modèles retenus par nos critères (voir le fichier `selection2_all_subset.sas` pour les manipulations).

Voici l'estimation de l'erreur moyenne quadratique (moyenne des carrés des erreurs) pour les deux modèles retenus par le AIC et le BIC. Le tableau @ref contient aussi l'estimation de l'erreur moyenne quadratique si on utilise toutes les variables (14 en incluant les indicatrices) sans faire de sélection.

TABLE 3.2: Estimation de l'erreur moyenne quadratique 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	EMQ	méthode
14	25,69	toutes les variables
10	24,72	exhaustive - AIC
7	23,83	exhaustive - BIC

On voit que le modèle choisi par le BIC est le meilleur des trois, car l'erreur moyenne quadratique 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.

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 $\texttt{age}^2$.
— toutes les interactions d'ordre deux entre les variables de base, comme $\texttt{sexe} \cdot \texttt{age}$.

Toutes ces variables sont créées dans `prepare_DBM.sas`. 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. 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.

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 restraignant à ces 40 variables.

3.11 Méthodes classiques de sélection ascendante, descendante et séquentielle

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 $X_1, \ldots, X_p$. Les procédures `glmselect` et `reg` permettent une sélection de modèle avec une approche séquentielle, ascendante ou descendante.

3.11.1 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 (`proc reg` utilise par défaut 0.5).
- *Boucle* soit $X_{(1)}, \ldots, X_{(k)}$, les variables explicatives à l'étape $k < p$.
  - pour chaque j ($j = \{1, \ldots, p\} \setminus \{(1), \ldots (k)\}$), 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 $\mathscr{H}_0 : \beta_{k+1} = 0$ contre l'alternative bilatérale $\mathscr{H}_1 : \beta_{k+1} \neq 0$.
  - 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.

3.11.2 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 (`proc reg` utilise par défaut 0.1).
- *Boucle* soit $X_{(1)}, \ldots, X_{(p-k)}$, les variables explicatives présentes dans le modèle à l'étape $k < p$.
  - pour chaque j ($j = 1, \ldots, p - k$), on calcule la valeur-p du test-t $\mathscr{H}_0 : \beta_j = 0$ contre l'alternative bilatérale $\mathscr{H}_1 : \beta_j \neq 0$.
  - 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.

3.11.3 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 (0.15 dans `proc reg`) et on début 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.

Utilisons la méthode de sélection séquentielle classique avec des critères d'entrée et de sortie de 0.15 et les 104 variables. Le code suivant, extrait de `selection2_all_subset.sas`, donne la syntaxe **SAS**.

```
proc reg data=trainymontant;
model ymontant=
x1 x2 x31 x32 x41 x42 x43 x44 x5 x6 x7 x8 x9 x10
cx2 cx6 cx7 cx8 cx9 cx10
i_x2_x1 i_x2_x5 i_x2_x31 i_x2_x32 i_x2_x41 i_x2_x42 i_x2_x43 i_x2_x44
i_x2_x7 i_x2_x6 i_x2_x8 i_x2_x9 i_x2_x10
i_x1_x5 i_x1_x31 i_x1_x32 i_x1_x41 i_x1_x42 i_x1_x43 i_x1_x44
i_x1_x7 i_x1_x6 i_x1_x8 i_x1_x9 i_x1_x10
i_x5_x31 i_x5_x32 i_x5_x41 i_x5_x42 i_x5_x43 i_x5_x44
i_x5_x7 i_x5_x6 i_x5_x8 i_x5_x9 i_x5_x10
i_x31_x41 i_x31_x42 i_x31_x43 i_x31_x44
i_x31_x7 i_x31_x6 i_x31_x8 i_x31_x9 i_x31_x10
i_x32_x41 i_x32_x42 i_x32_x43 i_x32_x44
i_x32_x7 i_x32_x6 i_x32_x8 i_x32_x9 i_x32_x10
i_x41_x7 i_x41_x6 i_x41_x8 i_x41_x9 i_x41_x10
i_x42_x7 i_x42_x6 i_x42_x8 i_x42_x9 i_x42_x10
i_x43_x7 i_x43_x6 i_x43_x8 i_x43_x9 i_x43_x10
i_x44_x7 i_x44_x6 i_x44_x8 i_x44_x9 i_x44_x10
i_x7_x6 i_x7_x8 i_x7_x9 i_x7_x10
i_x6_x8 i_x6_x9 i_x6_x10
i_x8_x9 i_x8_x10
i_x9_x10 / selection=stepwise sle=.15 sls=.15 ;
run;
```

Notez que les variables `cx2` à `cx10` sont les carrés des variables `x2` à `x10` que nous avons créées au préalable. De plus, les variables débutant par *i* sont les interactions entre les variables binaires et les variables continues. Par exemple, `i_x2_x1` est l'interaction entre `x1` et `x2`, c'est-à-dire le produit des deux.

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 `i_x5_x6` a été ajoutée, suivie de `i_x31_x10` à l'étape 2. Un peu plus loin, à l'étape 6, `i_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 de Sélection Stepwise

Etape	Variable entrée	Variable supprimée	Libellé	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	i_x5_x6			1	0.4327	0.4327	2307.71	158.64	<.0001
2	i_x31_x10			2	0.2434	0.6761	1231.24	155.55	<.0001
3	x6		anneeclient	3	0.1218	0.7979	693.409	124.19	<.0001
4	x5		conjoint	4	0.0419	0.8398	509.795	53.61	<.0001
5		i_x5_x6		3	0.0000	0.8398	507.925	0.04	0.8462
6	i_x31_x8			4	0.0395	0.8793	334.999	67.03	<.0001
7	i_x7_x8			5	0.0243	0.9036	229.269	51.44	<.0001
8	i_x1_x6			6	0.0245	0.9281	122.564	69.27	<.0001
9	i_x1_x10			7	0.0053	0.9334	100.995	16.14	<.0001
10	i_x5_x10			8	0.0058	0.9392	77.3939	19.10	<.0001
11	cx6			9	0.0054	0.9446	55.3352	19.61	<.0001
12		x6	anneeclient	8	0.0000	0.9446	53.3497	0.01	0.9136
13	cx10			9	0.0022	0.9468	45.6183	8.26	0.0045

Synthèse de Sélection Stepwise

Etape	Variable entrée	Variable supprimée	Libellé	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
14	i_x2_x31			10	0.0050	0.9518	25.5173	20.60	<.0001
15		i_x31_x10		9	0.0001	0.9517	23.8101	0.27	0.6020
16	x7		semainedernier	10	0.0027	0.9544	13.8722	11.77	0.0007
17	x8		montantdernier	11	0.0021	0.9566	6.3607	9.79	0.0020
18		i_x7_x8		10	0.0001	0.9565	4.6118	0.26	0.6118
19	i_x31_x44			11	0.0011	0.9576	1.9340	4.93	0.0276
20	i_x1_x42			12	0.0008	0.9584	0.1719	4.02	0.0462
21	i_x31_x41			13	0.0009	0.9594	-1.9858	4.53	0.0346
22	i_x2_x43			14	0.0013	0.9607	-5.6463	6.33	0.0127
23	x10		achatunan	15	0.0007	0.9614	-6.8007	3.57	0.0602
24	i_x5_x8			16	0.0007	0.9621	-7.9495	3.62	0.0587
25	i_x1_x32			17	0.0006	0.9627	-8.7093	3.21	0.0750
26	x44			18	0.0005	0.9632	-9.0496	2.74	0.0993
27		i_x31_x44		17	0.0004	0.9628	-9.2790	2.08	0.1513
28	i_x42_x6			18	0.0006	0.9634	-9.8486	3.03	0.0835
29	i_x41_x8			19	0.0007	0.9641	-11.056	3.83	0.0517
30	i_x1_x43			20	0.0005	0.9647	-11.380	2.80	0.0956
31	i_x8_x10			21	0.0004	0.9651	-11.316	2.35	0.1268
32	i_x32_x8			22	0.0004	0.9655	-11.026	2.09	0.1498
33	i_x44_x9			23	0.0004	0.9659	-10.790	2.17	0.1425
34	x32			24	0.0004	0.9663	-10.562	2.19	0.1402
35		x8	montantdernier	23	0.0003	0.9660	-11.426	1.41	0.2372
36		i_x1_x32		22	0.0003	0.9657	-12.232	1.47	0.2261
37		i_x1_x42		21	0.0004	0.9654	-12.666	1.93	0.1664
38	i_x42_x8			22	0.0006	0.9660	-13.130	3.05	0.0822
39		i_x42_x6		21	0.0003	0.9657	-14.018	1.38	0.2420
40	i_x44_x6			22	0.0004	0.9661	-13.751	2.16	0.1437

Voici les estimés des paramètres du modèle final retenu. On voit bien que toutes les valeurs-p (qui ne sont pas valides à cause de la sélection de modèles) sont toutes inférieures à $0,15$.

Variable	Valeur estimée des paramètres	Erreur type	SC Type II	Valeur F	Pr > F
Intercept	58.34184	1.11772	18095	2724.55	<.0001
x32	-2.77808	1.04895	46.58588	7.01	0.0088
x44	5.04444	1.25148	107.90771	16.25	<.0001
x5	6.29714	1.25363	167.57951	25.23	<.0001
x7	0.20731	0.01728	956.23194	143.98	<.0001
x10	0.84352	0.30010	52.47291	7.90	0.0055
cx6	0.14177	0.00801	2078.53494	312.96	<.0001
cx10	-0.13133	0.02714	155.47349	23.41	<.0001
i_x2_x31	-0.19460	0.02762	329.69938	49.64	<.0001
i_x2_x43	0.09759	0.02263	123.48077	18.59	<.0001

Variable	Valeur estimée des paramètres	Erreur type	SC Type II	Valeur F	Pr > F
i_x1_x43	-2.00266	0.89936	32.93241	4.96	0.0272
i_x1_x6	1.09890	0.10850	681.26470	102.57	<.0001
i_x1_x10	-0.34245	0.15726	31.49507	4.74	0.0307
i_x5_x8	0.03803	0.01075	83.07897	12.51	0.0005
i_x5_x10	1.43423	0.21761	288.50814	43.44	<.0001
i_x31_x41	1.54601	0.90967	19.18377	2.89	0.0909
i_x31_x8	-0.08801	0.01157	384.28942	57.86	<.0001
i_x32_x8	0.04008	0.01321	61.10879	9.20	0.0028
i_x41_x8	0.02367	0.00850	51.48950	7.75	0.0059
i_x42_x8	0.02983	0.00736	109.05701	16.42	<.0001
i_x44_x6	-0.22397	0.15253	14.31950	2.16	0.1437
i_x44_x9	-0.00416	0.00250	18.40660	2.77	0.0976
i_x8_x10	-0.00356	0.00177	26.79622	4.03	0.0460

La performance de ce modèle a, comme pour les modèles précédents, été évaluée avec l'échantillon test de 23 179 observations. Le tableau 3.3 présente la performance de la méthode de sélection séquentielle classique et celle du modèle dans lequel les 104 variables sont incluses sans faire de sélection.

TABLE 3.3: Comparaison des méthodes selon l'erreur moyenne quadratique pour la méthode de sélection séquentielle classique et pour le modèle incluant toutes les variables, les termes quadratiques et les interactions.

nombre de variables	EMQ	méthode
104	19,63	toutes les variables

nombre de variables	EMQ	méthode
22	12,25	séquentielle classique

On voit donc qu'utiliser toutes les 104 variables sans faire de sélection fait mieux (EMQ = 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 (EMQ = 12,25). On voit que dans cet exemple, utiliser les 104 variables fait du surajustement (*over-fitting*).

Le choix de 0,15 comme critère d'entrée et de sortie est assez 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. Le BIC est quant à lui beaucoup plus parcimonieux et choisit 15 de ces variables pour le modèle final. Encore une fois, ces deux modèles sont testés sur les 23 179 clients restants. Les résultats sont présentés dans le tableau 3.4.

TABLE 3.4: Comparaison des méthodes selon l'erreur moyenne quadratique pour la méthode de sélection séquentielle suivie d'une recherche exhaustive.

nombre de variables	EMQ	méthode
38	14,83	séquentielle classique, recherche exhaustive avec 56 variables (AIC)
15	11,96	séquentielle classique, recherche exhaustive avec 56 variables (BIC)

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 (EMQ = 11,96). Le AIC fait moins bien dans ce cas, avec une erreur moyenne quadratique estimée de 14,83. 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.

3.12 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 séquenctielle, 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.

```
proc glmselect data=trainymontant;
model ymontant= /*(mettre les 104 variables ici) */
/  selection=stepwise(select=aic choose=bic)  ;
score data=testymontant out=predglmselectaicbic p=predymontant;
run;
```

Voici l'historique de la procédure séquentielle avec cette combinaison.

Synthèse des sélections Stepwise

Etape	Effet saisi	Effet supprimé	Nombre d'effets dans	AIC	SBC
0	Intercept		1	1297.9002	1089.2473
1	i_x5_x6		2	1180.8614	975.5556
2	i_x31_x10		3	1065.1708	863.2121
3	x6		4	968.0920	769.4804
4	x5		5	921.3081	726.0436
5		i_x5_x6	4	919.3467	720.7351
6	i_x31_x8		5	861.9388	666.6743
7	i_x7_x8		6	816.7201	624.8028
8	i_x1_x6		7	757.0659	568.4957
9	i_x1_x10		8	742.9243	557.7011
10	i_x5_x10		9	725.8598	543.9838
11	cx6		10	708.2144	529.6855
12		x6	9	706.2268	524.3508
13	cx10		10	699.7279	521.1990
14	i_x2_x31		11	681.0439	505.8621
15		i_x31_x10	10	679.3317	500.8028
16	x7		11	669.2665	494.0847
17	x8		12	661.1313	489.2966
18		i_x7_x8	11	659.4052	484.2234*
19	i_x31_x44		12	656.2422	484.4075
20	i_x1_x42		13	653.9957	485.5081
21	i_x31_x41		14	651.2006	486.0601
22	i_x2_x43		15	646.4912	484.6978
23	x10		16	644.6570	486.2108
24	i_x5_x8		17	642.7586	487.6595
25	i_x1_x32		18	641.2814	489.5293
26	x44		19	640.2868	491.8818
27	i_x42_x6		20	639.2953	494.2375
28		i_x31_x44	19	639.2545	490.8495
29	i_x41_x8		20	637.0588	492.0009
30	i_x1_x43		21	635.9653	494.2545
31	i_x8_x10		22	635.3544	496.9907
32	i_x32_x8		23	635.0193	500.0027
33	i_x44_x9		24	634.5839	502.9145
34	x32		25	634.1076	505.7853

*** Valeur optimale du critère**

Synthèse des sélections Stepwise					
Etape	Effet saisi	Effet supprimé	Nombre d'effets dans	AIC	SBC
35		**x8**	24	633.6981	502.0286
36		**i_x1_x32**	23	633.3568	498.3403
37	**i_x6_x8**		24	632.9432	501.2738
38	**i_x41_x6**		25	632.5156	504.1933
39		**i_x41_x8**	24	632.1082	500.4388
40		**i_x1_x42**	23	631.6708*	496.6543
*** Valeur optimale du critère**					

À l'étape 1, la variable `i_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 `i_x31_x10` est ajoutée, À l'étape 6, la variable `i_x5_x6` est retirée car cela fait baisser le AIC. Notez que le AIC décroit 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 dimiuer 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	2.82857
Moyenne dépendante	67.28571
R carré	0.9565
R car. ajust.	0.9543
AIC	659.40521
AICC	660.98897
SBC	484.22340

Résultats estimés des paramètres

Paramètre	DDL	Estimation	Erreur type	Valeur du test t
Intercept	1	59.992385	0.730231	82.16
x5	1	9.448014	1.087431	8.69
x7	1	0.194506	0.017939	10.84
x8	1	0.036506	0.007510	4.86
cx6	1	0.136709	0.007354	18.59
cx10	1	-0.080544	0.015295	-5.27
i_x2_x31	1	-0.110846	0.024212	-4.58
i_x1_x6	1	1.110129	0.114049	9.73
i_x1_x10	1	-0.470860	0.165068	-2.85
i_x5_x10	1	1.310432	0.211687	6.19
i_x31_x8	1	-0.129280	0.009987	-12.94

Il s'avère que ce modèle performe très bien avec une erreur moyenne quadratique estimé à 10,08 sur les 23 179 clients de l'échantillon test. Il s'agit du meilleur jusqu'à maintenant.

3.13 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 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 expérimentale, `modelaverage`, qui permet de faire une moyenne de modèles. Comme elle est expérimentale, les particularités (options et sorties) de cette commande risquent de changer au cours des versions à venir. Le code suivant permet de faire une moyenne de modèles.

```
proc glmselect data=trainymontant seed=57484765;
model ymontant=
  ... /* mettre les 104 variables ici */
/ selection=stepwise(select=bic choose=bic) ;
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=bic choose=bic`).

Moyenne des résultats estimés des paramètres

Paramètre	Nombre différent de zéro	Pourcentage différent de zéro	Estimation de la moyenne	Ecart-type	Estimation quantiles 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 an 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, `i_x1_x6` et `i_x31_x8`. Le tableau rapporte aussi la moyenne des estimations pour ces paramètres.

Il s'avère que cette approche performe très bien sur l'échantillon test de 23 179 clients avec une erreur moyenne quadratique estimé de 10,57. Le tableau 3.5 résume la performance des différentes méthodes que nous avons utilisé sur notre échantillon test.

TABLE 3.5: Comparaison des méthodes selon l'erreur moyenne quadratique.

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

variables	nombre de variables	EMQ	méthode
interactions et termes quadratiques	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)
		10,57	moyenne de modèles

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 y aurait plusieurs autres approches/combinaisons 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 LASSO et 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 performant 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 de corriger les valeurs-p pour faire de l'inférence post-sélection.

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.

Chapitre 4

Régression logistique

4.1 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 $X_1, \ldots, X_p$.

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 $\boldsymbol{X}$ 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 $\boldsymbol{X}$.

4.2 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 $\mathsf{E}(\varepsilon \mid \boldsymbol{X}) = 0$ et $\mathsf{Var}(\varepsilon \mid \boldsymbol{X}) = \sigma^2$, peut être écrit de manière équivalente comme $\mathsf{E}(Y \mid \boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ et $\mathsf{Var}(Y \mid \boldsymbol{X}) = \sigma^2$.

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

$$\mathsf{E}(Y \mid \boldsymbol{X}) = \mathsf{P}(Y = 1 \mid \boldsymbol{X}),$$

soit la probabilité que Y égale 1 étant donné les valeurs des variables explicatives. Pour simplifier la notation, posons $p = \mathsf{P}(Y = 1 \mid \boldsymbol{X})$ 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

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

Le problème 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 $\mathbb{R} = (-\infty, \infty)$. 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,

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

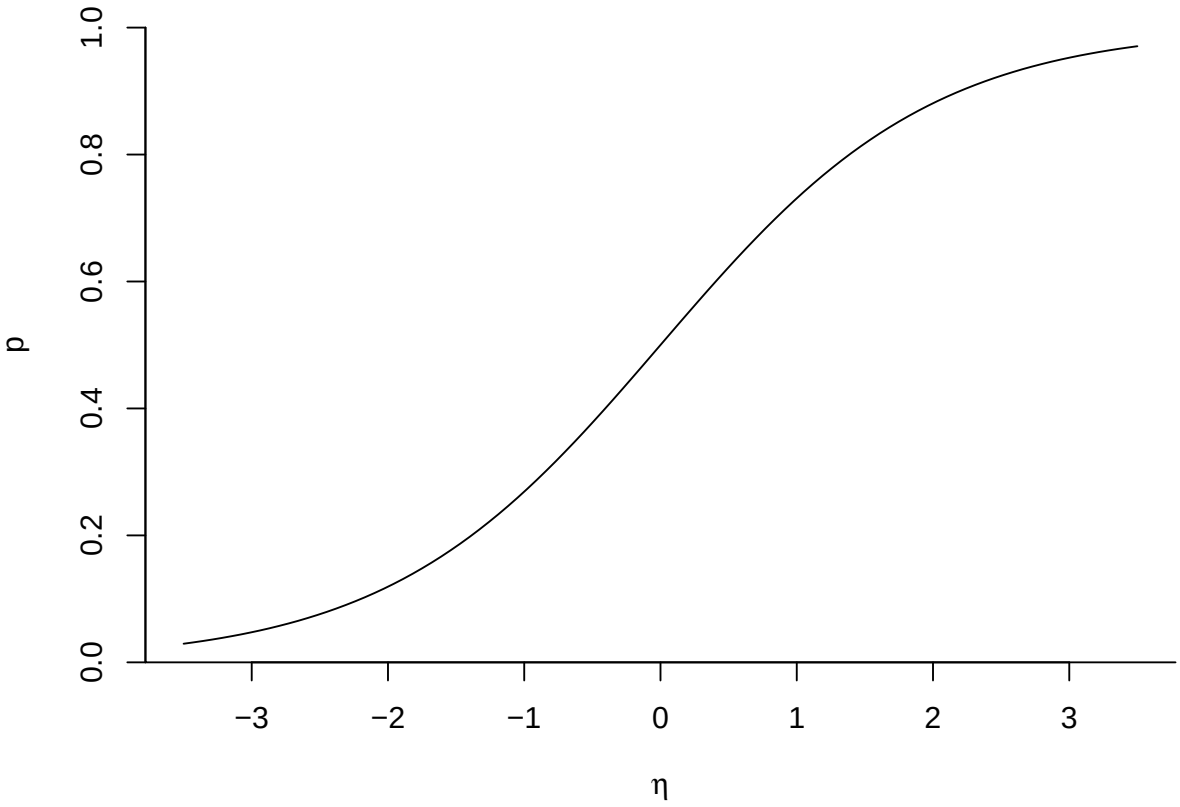
où ln est le logarithme naturel.

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 ($Y=0$),

$$\mathsf{cote}(p) = \frac{p}{1-p} = \frac{\mathsf{P}(Y=1 \mid \boldsymbol{X})}{\mathsf{P}(Y=0 \mid \boldsymbol{X})}.$$

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 4.1 donne un aperçu de cotes pour quelques probabilités p.

TABLE 4.1: Cote et probabilité de succès

$\mathsf{P}(Y=1)$	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	2,33	4	9
cote (frac.)	$\frac{1}{9}$	$\frac{1}{4}$	$\frac{3}{7}$	$\frac{2}{3}$	1	$\frac{3}{2}$	$\frac{7}{3}$	4	9

4.3 Estimation des paramètres

4.3.1 Principes de base

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

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

À l'aide de ces observations, on peut estimer les paramètres $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ 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 $\hat{\boldsymbol{\beta}}$, desquels découle une estimation de $\mathsf{P}(Y = 1)$ pour les valeurs $X_1 = x_1, \ldots, X_p = x_p$ 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 classifie 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 ℓ, le AIC et le BIC. Les deux critères d'information, que nous avons couvert dans les chapitres précédent, servent à la sélection de modèles tandis que la log-vraisemblance servira à construire un test d'hypothèse.

4.3.2 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 (AIC, BIC et -2LL).

La méthode du maximum de vraisemblance (*maximum likelihood*) est possiblement la méthode d'estimation la plus utilisée en statistique. En général, pour un échantillon donné et un modèle avec des paramètres inconnus $\boldsymbol{\theta}$, 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, $\boldsymbol{\theta}$, 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 » ou LL).

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, alors la log-vraisemblance du modèle pour un échantillon de taille n est

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

Puisque le premier terme ne dépend pas des paramètres $\boldsymbol{\beta}$, il est clair que maximiser cette fonction de $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ 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. De plus, 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

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

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) du $\ell(\hat{\boldsymbol{\beta}}$. La quantité $-2\ell(\hat{\boldsymbol{\beta}})$ (`-2 LL`) 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 fonctions de a log-vraisemblance et sont

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

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 analyse factorielle, ces deux critères pourront être utilisés pour faire de la sélection de modèles si on calcule les estimateurs du maximum de vraisemblance.

4.4 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`.

4.4.1 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;
```

Le fichier `logit1_intro.sas` contient ce programme et décrit plus en détail les différentes options. La syntaxe `y(ref='0')` sert à spécifier la catégorie de référence, zéro, de la variable réponse Y : le modèle

décrit donc $\mathrm{P}\left(y = 1 \mid X_5\right)$.

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 uniquement	Constante et Covariables
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

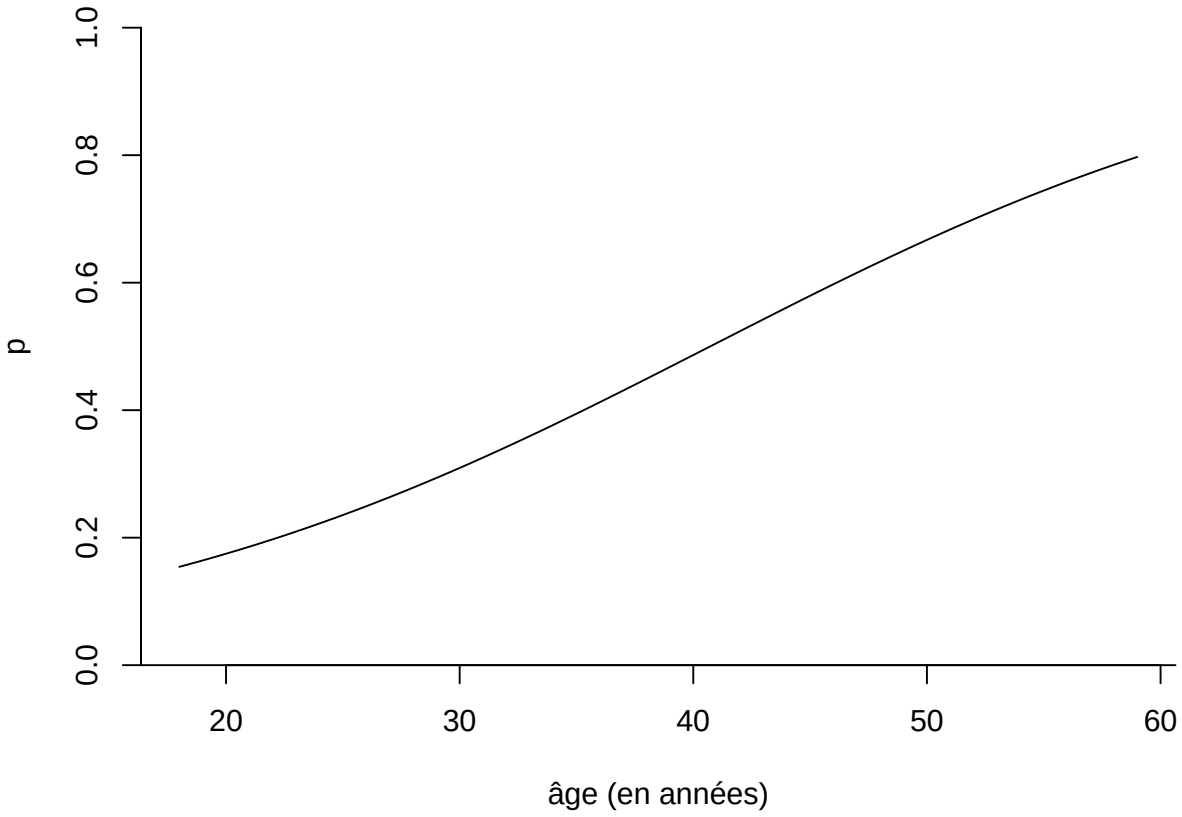
Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil

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 $\hat{\beta}_{\texttt{age}} = 0,0749$.
— Un intervalle de confiance de niveau 95% pour l'effet de l'âge est [0,0465 ; 0,1043].
— Le modèle ajusté est $\text{logit}\{\mathsf{P}(Y = 1 \mid X_5 = x_5)\} = -3,05 + 0,0749x_5$. On peut également exprimer ce modèle directement en terme de la probabilité de succès,

$$\mathsf{P}(Y = 1 \mid X_5 = x_5) = \text{expit}(-3,05 + 0,0749x_5) = \frac{1}{1 + \exp(3,05 - 0,0749x_5)}$$

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 sigmoide de la fonction logit aux deux extrémités.



— La valeur-p pour $\hat{\beta}_{\texttt{age}}$ (`Pr > khi-2`), correspondant aux test des hypothèses $\mathscr{H}_0 : \beta_{\texttt{age}} = 0$ versus $\mathscr{H}_1 : \beta_{\texttt{age}} \neq 0$, 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.

4.4.2 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 $\mathsf{P}(Y = 1)$ augmente. Inversement, Si le coefficient β_j est négatif, plus la variable augmente, plus $\mathsf{P}(Y = 1)$ 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 $\mathsf{P}(Y=1)$ (courbe sigmoide), l'augmentation ou la dimininution de $\mathsf{P}(Y=1 \mid \boldsymbol{X})$ 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,

$$\frac{\mathsf{P}(Y=1 \mid X_5 = x_5)}{\mathsf{P}(Y=0 \mid X_5 = x_5)} = \exp(-3{,}05)\exp(0{,}0749 x_5).$$

Ainsi, lorsque X_5 augmente d'une année, la cote est multipliée par $\exp(0{,}0749) = 1{,}078$ peut 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 reparametrisations. Ainsi, l'intervalle de confiance à niveau 95% pour $\exp(\beta_{\texttt{age}})$ est obtenu en prenant l'exponentielle des bornes de l'intervalle pour $\beta_{\texttt{age}}$, $[\exp(0{,}0465); \exp(0{,}1043)]$, 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 $\hat{\beta} \pm 1.96\mathrm{se}(\hat{\beta})$. 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 $\mathscr{H}_0 : \beta_j = 0$ ou de façon équivalente $\mathscr{H}_0 : \exp(\beta_j) = 1$ à niveau 95%.

4.4.3 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 ordinale 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.

Statistique d'ajustement du modèle		
Critère	**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			
Effet	DDL	Khi-2 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**	1	0.3580	0.4826	0.5503	0.4582	1.430
x1	**2**	1	-0.4677	0.3714	1.5865	0.2078	0.626
x1	**3**	1	-0.3113	0.3503	0.7901	0.3741	0.732
x1	**4**	1	-0.3170	0.4025	0.6201	0.4310	0.728
x2	**1**	1	1.3312	0.5967	4.9772	0.0257	3.786
x2	**2**	1	1.1484	0.5014	5.2469	0.0220	3.153
x2	**3**	1	0.7733	0.4830	2.5628	0.1094	2.167
x2	**4**	1	-1.1088	0.5418	4.1884	0.0407	0.330
x3		1	1.3490	0.2599	26.9385	<.0001	3.853
x4		1	1.8303	0.2979	37.7528	<.0001	6.235
x5		1	0.1095	0.0193	32.3048	<.0001	1.116
x6	**1**	1	2.4122	0.3756	41.2339	<.0001	11.158
x6	**2**	1	1.0446	0.2493	17.5557	<.0001	2.842

Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil				
Effet	**Unité**	**Estimation**	**Intervalle de confiance à95%**	
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

$$\begin{aligned}\text{logit}\{\mathsf{P}(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})\} = &-6,89 + 0,36\mathbf{1}_{X_1=1} - 0,47\mathbf{1}_{X_1=2} - 0,31\mathbf{1}_{X_1=3} - 0,32\mathbf{1}_{X_1=4} \\ &+ 1,33\mathbf{1}_{X_2=1} + 1,15\mathbf{1}_{X_2=2} + 0,77\mathbf{1}_{X_2=3} - 1,11\mathbf{1}_{X_2=4} \\ &+ 1,35X_3 + 1,83X_4 + 0,11X_5 + 2,41\mathbf{1}_{X_6=1} + 1,04\mathbf{1}_{X_6=2}\end{aligned}$$

Notez que les variables $\mathbf{1}_{X_1=1}$ (`x11`), $\mathbf{1}_{X_1=21}$ (`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 exmple. Le paramètre estimé est $\hat{\beta}_{\texttt{age}} = 0,1095$ et il est significativement différent de zéro. Ainsi, plus l'âge augmente, plus $\mathsf{P}(Y=1 \mid \boldsymbol{X})$ 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 5 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 $\mathscr{H}_0 : \beta_{6_1} = \beta_{6_2} = 0$ 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 5 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.

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

Le test est basé sur la statistique

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

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 $\mathscr{H}_0$ 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 $\mathscr{H}_0$). 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 $\mathscr{H}_0 : \beta_{6_1} = \beta_{6_2} = 0$ 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 modeèl 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 $1,22 \times 10^{-11}$.

```
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 $\mathscr{H}_0$, 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).

4.4.5 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 $\boldsymbol{X}$, 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.

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

Une des utilisations les plus fréquentes de la régression logistique est de se servir d'un modèle ajusté pour obtenir des 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 $X_1, \ldots, X_p$.

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

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

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 $\hat{p} < c$, alors $\hat{Y} = 0$ (c'est-à-dire, on assigne cette observation à la catégorie zéro).
— Si $\hat{p} \geq c$, alors $\hat{Y} = 1$ (c'est-à-dire, on assigne cette observation à la catégorie 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 à 0 lorsqu'elle vaut 1 en réalité et que les conséquences de faussement classifier une observation à 1 lorsqu'elle vaut 0 en réalité, ne sont pas les mêmes, il peut être avantageux d'utiliser un autre point de coupure.

Dans un cadre de prédiction, 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, $\mathsf{EMQ} = \mathsf{E}\left((Y - \hat{Y})^2\right)$, 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 prédiction. 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 $\mathsf{P}\left(Y \neq \hat{Y}\right)$, soit la probabilité de mal classifier une observation choisie au hasard dans la population. Ce critère est le **taux de mauvaise classification**. Plus $\mathsf{P}\left(Y \neq \hat{Y}\right)$ est petite, meilleure est la capacité prédictive du modèle.

Tout comme l'erreur moyenne quadratique, on ne peut pas calculer exactement le taux de mauvaise classification ; tout au plus peut-on l'estimer. 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 $\mathsf{P}\left(Y \neq \hat{Y}\right)$.

Cette utilisation d'un modèle de régression logistique sera illustrée avec l'exemple que nous avons traité au chapitre précédent. Rappelez-vous le contexte. Un catalogue a été envoyé à un échantillon de 1000 clients. Le but final est de construire un modèle, avec ces 1000 clients, afin de cibler lesquels des 100 000 clients restants seront choisis pour recevoir le catalogue. Les 1000 clients forment l'échantillon d'apprentissage.

Des données sont disponibles pour les 1000 clients. 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 (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.

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

Afin d'introduire différentes notions, nous allons, dans un premier temps, simplement 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 vet prédire l'intention d'achat.