Ricco Rakotomalala

Pratique de la Régression Logistique

Régression Logistique Binaire et Polytomique

Version 2.0

Université Lumière Lyon 2

Avant-propos

Ce fascicule est dédié à la Régression Logistique. Il s'agit d'une technique de modélisation qui, dans sa version la plus répandue, vise à prédire et expliquer les valeurs d'une variable catégorielle binaire Y (variable à prédire, variable expliquée, variable dépendante, attribut classe, variable endogène) à partir d'une collection de variables X continues ou binaires (variables prédictives, variables explicatives, variables indépendantes, descripteurs, variables exogènes). Elle fait partie des méthodes d'apprentissage supervisé [13] ; elle peut s'inscrire dans le cadre de la régression linéaire généralisée [7] (Chapitre 5, pages 83-97) ; elle peut être vue comme une variante de la régression linéaire multiple, bien connue en économétrie [6] (Chapitre IV, pages 67-77).

Pendant longtemps, trouver de la documentation en français sur la Pratique de la Régression Logistique a été un problème. Les seuls ouvrages disponibles étudiaient le sujet sous l'angle de *l'économétrie des variables qualitatives*, excellents par ailleurs, mais avec un prisme plutôt théorique, assez éloigné des préoccupations du praticien qui souhaite mettre en oeuvre l'outil dans le cadre du scoring ou du data mining sans entrer dans les arcanes des propriétés des estimateurs, biais, convergence, etc. Les questions que tout un chacun se pose face à ce type de méthode sont assez simples et demandent des réponses tout aussi simples : De quoi s'agit-il ? A quel type de problème répond la technique ? Comment peut-on la mettre en oeuvre ? Quelles en sont les conditions d'utilisation ? Comment lire et interpréter les résultats ? Comment les valider ?

Fort heureusement, dans la période récente, la situation a radicalement changé. Des chapitres entiers sont consacrés aux aspects pratiques de la régression logistique dans de nombreux ouvrages en français que nous citons en bibliographie. Certains le font de manière approfondie en détaillant les formules. D'autres se concentrent sur la mise en oeuvre et les interprétations. En tous les cas, le lecteur exclusivement francophone a de quoi lire.

La situation est en revanche moins reluisante concernant la documentation accessible librement sur internet. Certes, nous pouvons glaner ici ou là quelques "slides" sur des serveurs. Mais, d'une part, il ne s'agit que de supports très peu formalisés et, d'autre part, leur durée de vie est souvent très faible. Je fais certes systématiquement des copies locales en ce qui me concerne, mais il est hors de question bien entendu de les diffuser moi même. Leurs auteurs ne les ont pas retirés par hasard.

Ce fascicule est une version formalisée et complétée de mes "slides" accessibles sur mon site de cours (`http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`, "Régression Logis-

tique" [14] et "Régression Logistique Polytomique" [15]). Nous faisons la part belle à la régression logistique binaire dans les 2 premières parties. Nous élargirons notre propos à la régression logistique polytomique (Y peut prendre plus 2 modalités, elles sont éventuellement ordonnées) dans les autres.

Enfin, nous nous focalisons avant tout sur la mise en oeuvre de la régression logistique. Les formules sont détaillées uniquement lorsqu'elles permettent de mieux comprendre les mécanismes sous-jacents, de mieux appréhender la teneur des résultats et, par là, de mieux les interpréter. Comme à notre habitude, une des particularités de cet ouvrage est que nous reproduisons autant que possible les calculs dans un tableur. Nous mettons en relation directe les formules, qui sont parfois assez abstraites, et les étapes numériques qui permettent d'aboutir aux résultats[1]. Au besoin nous croiserons les résultats avec les sorties des logiciels spécialisés. Nous utiliserons prioritairement les outils libres, TANAGRA (`http://eric.univ-lyon2.fr/~ricco/tanagra`) et R (`http://www.r-project.org`), pour que le lecteur puisse reproduire les exemples illustratifs. Tous les fichiers de données et de calculs utilisés pour l'élaboration de cet ouvrage sont accessibles en ligne (voir Annexe B, page 247).

Un document ne vient jamais du néant. Comme il n'y a pas 10.000 manières de présenter la régression logistique, toute ressemblance avec des références existantes n'est pas fortuite. Elle est complètement assumée. Le plus important dans ce cas est de veiller à les citer[2]. Rendons donc à César ce qui lui appartient, les sources suivantes m'ont beaucoup inspiré :

1. L'ouvrage de Hosmer et Lemeshow est certainement **LA** référence anglo-saxonne [9]. Quiconque souhaite mettre en pratique la régression logistique dans une application réelle se doit d'avoir lu cet ouvrage. Le discours est clair. Il va directement à l'essentiel, sans néanmoins faire l'impasse sur les aspects théoriques importants. Tout est disséqué, discuté, les références sont systématiquement croisées, recoupées. J'ai rarement lu un livre d'une telle qualité. C'est simple. J'ouvre une page au hasard, je trouve intéressant ce qui y est écrit. Les ouvrages qui m'ont autant impressionné se comptent sur les doigts de la main. Je remarque d'ailleurs que je ne suis pas le seul à l'avoir apprécié. De nombreux auteurs s'en inspirent grandement dans leur présentation. On retrouve, entres autres, quasiment partout le fameux exemple de la prédiction de la CHD (coronary heart disease) en fonction de l'âge, avec les figures 1.1 et 1.2 ([9], pages 4 et 5). C'est plutôt bon signe je trouve. J'avoue moi même avoir fait comme tout le monde. Autant prendre les informations là où elles sont de bonne qualité.
2. L'autre référence anglo-saxonne qui m'a beaucoup plu est l'ouvrage de Scott Menard [10] de la série *Quantitative Applications in the Social Sciences* (Sage University Paper). Il s'agit d'une petite brochure qui ne paie pas de mine, écrit un peu à la manière des "Que sais-je". Mais à l'usage, on se rend compte très rapidement de la richesse du propos (comme les "Que sais-je" d'ailleurs). En plus, la lecture est très fluide. C'est toujours agréable. L'auteur prend beaucoup de recul par rapport aux techniques. Il faut prendre cet ouvrage comme un guide de lecture des résultats de la régression

1. C'est devenu un peu une marque de fabrique de mes écrits. Je pense que savoir reproduire les formules sur un tableur est le signe qu'on les a parfaitement comprises. Je montre les calculs sous Excel parce que je l'utilise pour mes enseignements, mais la transposition à Open Office Calc ne présente aucune difficulté.

2. Reprendre le travail des autres sans les citer, c'est du plagiat ; reprendre le travail des autres en les citant, c'est une manière d'honorer leurs efforts. Ca ne coûte rien de le faire et ça fait plaisir. Pourquoi s'en priver ?

logistique. Il nous aide à comprendre ce qui est important dans les sorties de logiciels. Il fait beaucoup référence aux principaux outils du marché (SAS, SPSS).

3. En français, après une longue période de disette, la documentation existe maintenant. Il n'y a certes pas de livres exclusivement consacrés au sujet. Mais bien souvent les chapitres que l'on retrouve dans les différents ouvrages sont d'excellente facture. Nous les détaillons volontiers dans la bibliographie en indiquant les numéros de chapitre et le nombre de pages dédiées au sujet pour que le lecteur puisse faire sa recherche bibliographique en connaissance de cause.
4. En ligne, en français, de la documentation à la fois pérenne et suffisamment approfondie est très rare. Il y a bien la page Wikipédia [25], mais elle est plutôt concise. Comme je le disais plus haut, en cherchant bien on trouve ici ou là des "slides". Mais d'une part, ils sont très laconiques (c'est plutôt normal pour des slides) ; d'autre part, ils restent peu de temps en ligne. C'est un peu (beaucoup) dommage. Ceci est vrai aujourd'hui (Août 2009). Peut être qu'entre temps d'autres supports de qualité en français sont maintenant disponibles. Ca ne peut être que positif pour tout le monde.
5. En anglais, la situation est tout autre. Les excellentes références abondent, avec une pérennité qui semble assurée. Je citerai le cours complet avec des exemples commentés sous SAS et R du département de Statistique de l'Université de Pennsylvania [22], ou encore la page de David Garson qui, fidèle à sa démarche, trace les contours de la méthode puis explique de manière approfondie les sorties du logiciel SPSS [5].

Il ne m'a pas été possible de rédiger d'une traite la totalité de ce fascicule. Plutôt que d'attendre indéfiniment sa finalisation, j'ai préféré sortir une première version, consacrée exclusivement à la régression logistique binaire. Le reste, les chapitres relatifs à la régression logistique polytomique, viendra au fil du temps. J'ai mis en annexes les indications qui permettent de suivre les différentes versions et les dates de modifications (Annexe A).

Enfin, selon l'expression consacrée, ce support n'engage que son auteur. Toutes suggestions ou commentaires qui peuvent en améliorer le contenu sont les bienvenus.

Table des matières

Partie III La régression logistique multinomiale

Notations

L'objectif est de prédire les valeurs prises par la variable aléatoire Y définie dans $\{y_1, y_2, \ldots, y_K\}$. Pour la régression logistique binaire, Y prend uniquement deux modalités $\{+, -\}$ (ou $\{1, 0\}$ pour simplifier). Nous disposons d'un échantillon Ω de taille n. La valeur prise par Y pour un individu ω est notée $Y(\omega)$.

Le fichier comporte J descripteurs $\{X_1, X_2, \ldots, X_J\}$. Le vecteur de valeurs pour un individu ω s'écrit $(X_1(\omega), X_2(\omega), \ldots, X_J(\omega))$.

Dans le cadre binaire, pour un individu donné, sa probabilité a priori d'être positif s'écrit $P[Y(\omega) = +] = p(\omega)$. Lorsqu'il ne peut y avoir d'ambiguïtés, nous la noterons simplement p.

Lorsque l'échantillon est issu d'un tirage aléatoire dans la population, sans distinction des classes d'appartenance, si n_+ est le nombre d'observations positives dans Ω, p peut être estimée par $\frac{n_+}{n}$. On parle de "schéma de mélange" ([3], page 5).

La **probabilité a posteriori** d'un individu ω d'être positif c.-à-d. sachant les valeurs prises par les descripteurs est notée $P[Y(\omega) = +/X(\omega)] = \pi(\omega)$. Ici également, lorsqu'il ne peut y avoir de confusions, nous écrirons π. Ce dernier terme est très important. En effet, c'est la probabilité **que l'on cherche à modéliser en apprentissage supervisé**.

Le LOGIT d'un individu ω s'écrit

$$\ln\left[\frac{\pi(\omega)}{1-\pi(\omega)}\right] = a_0 + a_1 X_1(\omega) + \cdots + a_J X_J(\omega)$$

$a_0, a_1, \ldots, a_J$ sont les paramètres que l'on souhaite estimer à partir des données.

Lorsque nous adoptons une écriture matricielle, nous écrirons

$$\ln\left[\frac{\pi(\omega)}{1-\pi(\omega)}\right] = X(\omega) \times a$$

avec $X(\omega) = (1, X_1(\omega), X_2(\omega), \ldots, X_J(\omega))$, la première composante ($X_0(\omega) = 1,\ \forall\omega$) symbolise la constante ; $a' = (a_0, a_1, \ldots, a_J)$ est le vecteur des paramètres.

Enfin, toujours pour alléger l'écriture, nous omettrons le terme ω lorsque cela est possible.

Données

Autant que faire se peut, nous utiliserons le même jeu de données fictif comportant 20 observations et 3 variables prédictives pour illustrer la régression logistique binaire. L'objectif est de prédire la présence ou l'absence d'un problème cardiaque (COEUR - Y ; avec "présence" = "+" et "absence" = "-") à partir de son AGE (quantitative - X_1), du TAUX MAX (pression sanguine, quantitative - X_2) et l'occurence d'une ANGINE de poitrine (binaire - X_3) (Figure 0.1).

Nous obtenons une série d'indicateurs lorsque nous le traitons avec le Tanagra (Figure 0.2) ou lorsque nous le traitons avec le logiciel R (Figure 0.3). Certaines permettent d'évaluer la qualité globale de la

age	taux_max	angine	coeur
50	126	1	presence
49	126	0	presence
46	144	0	presence
49	139	0	presence
62	154	1	presence
35	156	1	presence
67	160	0	absence
65	140	0	absence
47	143	0	absence
58	165	0	absence
57	115	1	absence
59	145	0	absence
44	175	0	absence
41	153	0	absence
54	152	0	absence
52	169	0	absence
57	168	1	absence
50	158	0	absence
44	170	0	absence
49	171	0	absence

Fig. 0.1. Fichier COEUR

régression, d'autres permettent de juger la contribution individuelle de chaque variable. **Expliciter les principes qui régissent la méthode et décrire les formules associées pour que nous sachions lire en connaissance de cause les résultats constituent les objectifs de ce support**.

Le fichier est suffisamment petit pour que l'on puisse détailler tous les calculs. Le faible effectif en revanche induit une certaine instabilité des résultats. Dans certains cas ils ne concordent pas avec nos connaissances usuelles. Il ne faudra pas s'en formaliser. L'intérêt d'avoir recours à un expert du domaine justement est qu'il a la possibilité de valider ou d'invalider le fruit de calculs purement mécaniques.

Bien entendu, lorsque la situation ne s'y prête pas, nous utiliserons ponctuellement d'autres fichiers de données. Nous l'indiquerons au fur et à mesure.

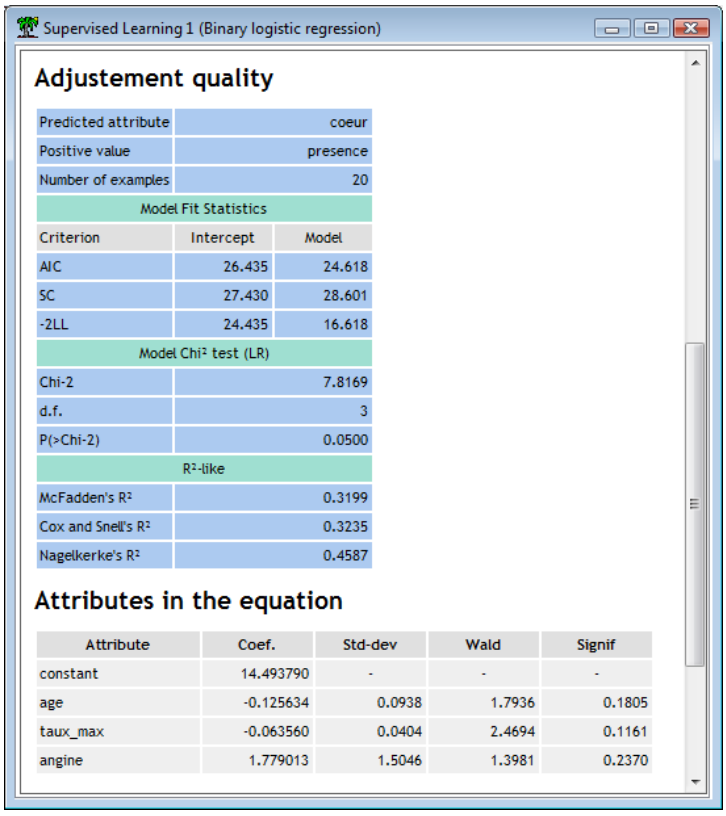


Fig. 0.2. Traitement du fichier COEUR avec le logiciel Tanagra

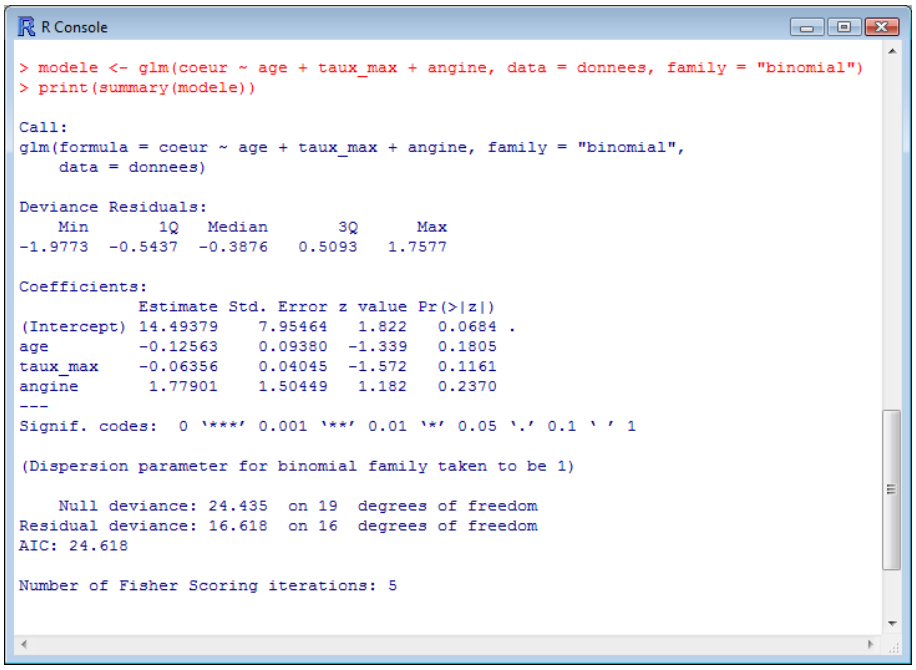


Fig. 0.3. Traitement du fichier COEUR avec le logiciel R

Partie I

Régression Logistique Binaire

1

Régression Logistique Binaire - Principe et estimation

1.1 Un cadre bayesien pour l'apprentissage supervisé

1.1.1 Apprentissage supervisé - Problématique

En apprentissage supervisé, l'objectif est de prédire et/ou expliquer une variable catégorielle Y à partir d'une collection de descripteurs $X = (X_1, X_2, \ldots, X_J)$. Il s'agit en quelque sorte de mettre en évidence l'existence d'une liaison fonctionnelle sous-jacente (en anglais, *underlying concept*) de la forme

$$Y = f(X, \alpha)$$

entre ces variables.

La fonction $f(.)$ est le modèle de prédiction, on parle aussi de classifieur ; α est le vecteur des paramètres de la fonction, on doit en estimer les valeurs à partir des données disponibles.

Dans le cadre de la discrimination binaire, nous considérons que la variable dépendante Y ne prend que 2 modalités : positif "+" ou négatif "-". Nous cherchons à prédire correctement les valeurs de Y, mais nous pouvons également vouloir quantifier la propension (la probabilité) d'un individu à être positif (ou négatif).

Les applications sont nombreuses, certains touchent directement à notre vie quotidienne :

1. Déterminer la viabilité d'un client sollicitant un crédit à partir de ses caractéristiques (age, type d'emploi, niveau de revenu, autres crédits en cours, etc.) ;
2. Quantifier le risque de survenue d'un sinistre pour une personne sollicitant un contrat d'assurance (ex. un jeune homme venant d'obtenir son permis de conduire et demandant une assurance tous risques pour une 205 Turbo-kittée avec un aileron de requin sur le toit aura très peu de chances de trouver une compagnie conciliante) ;
3. Discerner les facteurs de risque de survenue d'une maladie cardio-vasculaire chez des patients (ex. l'âge, le sexe, le tabac, l'alcool, regarder les matches de l'équipe de France de foot, etc.) ;
4. Pour une enseigne de grande distribution, cibler les clients qui peuvent être intéressés par tel ou tel type de produit.

Comme dans toute démarche de modélisation, plusieurs questions se posent immédiatement [23] (pages 104-105) :

1. Choisir la forme de la fonction.
2. Estimer les paramètres du modèle à partir d'un échantillon Ω.
3. Évaluer la précision des estimations.
4. Mesurer le pouvoir explicatif du modèle.
5. Vérifier s'il existe une liaison significative entre l'ensemble des descripteurs et la variable dépendante.
6. Identifier les descripteurs pertinents dans la prédiction de Y, évacuer celles qui ne sont pas significatives et/ou celles qui sont redondantes.
7. Mesurer l'influence de chaque observation, au besoin détecter celles qui peuvent avoir une influence exagérée au point de fausser les résultats.
8. Pour un nouvel individu à classer, déterminer la valeur de π à partir des valeurs prises par les X.
9. Construire un intervalle de variation (fourchette) de π.

La régression logistique permet de répondre précisément à chacune de ces questions. Elle le fait surtout de manière complètement cohérente avec sa démarche d'apprentissage, la maximisation de la vraisemblance. Ce n'est pas un de ses moindres mérites par rapport à d'autres méthodes supervisées.

1.1.2 Apprentissage supervisé - Évaluation

Avant de proposer une démarche pour résoudre le problème ci-dessus, penchons-nous sur un aspect fondamental de l'apprentissage supervisé : comment évaluer la qualité de la modélisation ? Comment comparer 2 approches concurrentes et dégager celle qui serait la meilleure ?

Une attitude simple consiste à mesurer la qualité de la prédiction c.-à-d. l'aptitude du modèle à prédire correctement dans la population Ω^{pop}. Nous la quantifions avec **l'erreur théorique** que l'on interprète comme la **probabilité de mal classer** un individu pris au hasard dans la population :

$$\epsilon = \frac{1}{card(\Omega^{pop})} \sum_{\omega} \Delta \left[Y(\omega), f(X(\omega)) \right]$$

où Δ est une fonction indicatrice qui, pour un individu ω donné, prend la valeur 1 lorsque la prédiction ne concorde pas avec la vraie valeur ; elle prend la valeur 0 lorsque le modèle prédit à bon escient.

On confronte ainsi les vraies valeurs prises par la variable dépendante dans la population et les prédictions du modèle. Dans le cas idéal, toutes les prédictions sont correctes, l'erreur théorique est égal 0. L'autre extrême serait que le modèle se trompe systématiquement, dans ce cas le taux serait égal à 1. Mais en réalité, il est plus judicieux de prendre comme borne haute le classement au hasard [1]. Lorsque les classes sont équi-distribuées c.-à-d. les proportions de positifs et de négatifs sont identiques dans la population, nous obtiendrions un taux d'erreur théorique égal à 0.5. Le classifieur doit faire mieux.

1. Nous verrons plus loin (sections 1.6 et 2.1.4) qu'il y a une approche plus rigoureuse pour définir le classifieur de référence (le modèle trivial), celui que l'on doit absolument surpasser.

Notre indicateur est théorique dans la mesure où nous ne disposerons jamais de la population complète pour la calculer. Il faudra que l'on produise une estimation à partir d'un échantillon. La solution la plus simple consiste à mesurer la proportion de mauvais classement sur le fichier qui a servi à construire le modèle, on parle de *taux d'erreur en resubstitution*. Pour simple quelle soit, cette estimation n'est cependant pas très fiable, on sait que le taux d'erreur ainsi calculé est souvent [2] trop optimiste, il faut le produire autrement. Nous reviendrons sur ce sujet plus loin (section 2.1).

1.1.3 Un cadre bayesien pour l'apprentissage supervisé

Le classifieur bayesien est celui qui répond de manière optimale aux spécifications ci-dessus. Pour un individu ω, il s'agit de calculer les probabilités conditionnelles (probabilité a posteriori)

$$P[Y(\omega) = y_k / X(\omega)]$$

pour chaque modalité y_k de Y.

On affecte à l'individu la modalité la plus probable y_{k^*} c.-à-d.

$$y_{k^*} = \arg\max_k P[Y(\omega) = y_k / X(\omega)]$$

On associe donc l'individu à la classe la plus probable compte tenu de ses caractéristiques $X(\omega)$.

Cette approche est optimale au sens de l'erreur théorique [3]. Mais un problème apparaît tout de suite : comment estimer correctement ces probabilités conditionnelles ?

1.1.4 Un exemple : prédire COEUR en fonction de ANGINE

Apprentissage

Pour notre fichier ci-dessus (Figure 0.1), nous souhaitons prédire les valeurs de COEUR en fonction de ANGINE. La variable prédictive ANGINE étant binaire (et de manière plus générale catégorielle), nous pouvons utiliser les fréquences pour estimer les probabilités conditionnelles. Nous utilisons pour cela l'outil "Tableaux croisés dynamiques d'Excel" (Figure 1.1).

Pour ANGINE = 0, nous avons les fréquences conditionnelles

– $P(COEUR = + / ANGINE = 0) = 0.2$

2. Souvent, pas toujours. L'importance du biais d'optimisme dépend aussi des caractéristiques du classifieur et des données. Dans certains cas, lorsque la méthode a tendance à fortement coller aux données (ex. la méthode des plus proches voisins dans un espace sur-dimensionné), le taux d'erreur en resubstitution n'est d'aucune utilité ; dans d'autres, méthodes linéaires dans un espace où le ratio entre le nombre d'observations et le nombre de descripteurs est favorable, il donne des indications tout à fait crédibles.

3. Il est possible de généraliser l'approche à une configuration où l'on associerait des coûts de mauvais classement aux affectations (cf. [3], page 4)

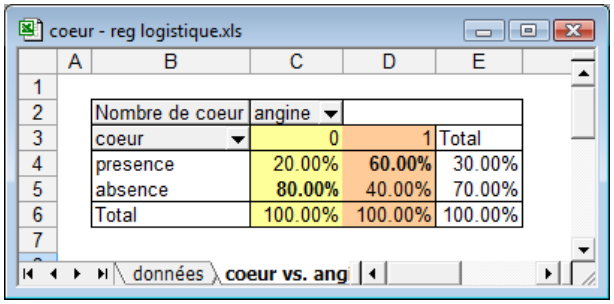


Fig. 1.1. COEUR vs. ANGINE - Probabilités conditionnelles

– $P(COEUR = -/ANGINE = 0) = 0.8$

En vertu du principe bayesien, nous adoptons la règle suivante :

$$Si\ ANGINE = 0\ Alors\ COEUR = -\ (absence)$$

De la même manière, pour ANGINE $= 1$, nous calculons

– $P(COEUR = +/ANGINE = 1) = 0.6$
– $P(COEUR = -/ANGINE = 1) = 0.4$

Nous en déduisons

$$Si\ ANGINE = 1\ Alors\ COEUR = +\ (presence)$$

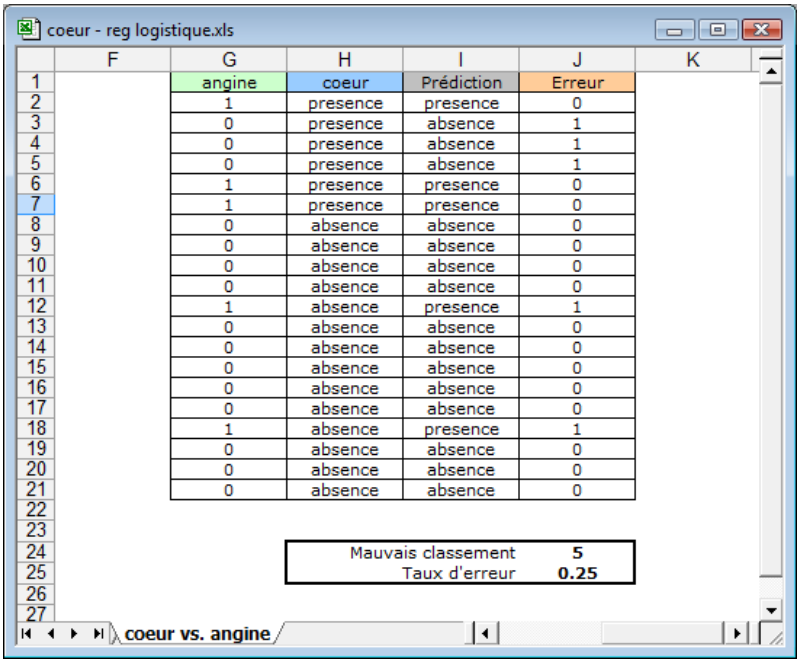
Évaluation

Maintenant que nous avons un modèle de prédiction $COEUR = f(ANGINE)$, il faut en évaluer les performances. Pour cela, nous confrontons les vraies valeurs de la variable dépendante avec celles prédites par le modèle.

Dans notre feuille Excel (Figure 1.2), la colonne "Prédiction" correspond aux valeurs prédites par le modèle, nous utilisons simplement une fonction "SI(...)" s'appuyant sur la colonne ANGINE. "Erreur" correspond à la fonction Δ. Elle prend la valeur 1 lorsque la prédiction erronée, 0 autrement.

Dans la partie basse de la feuille, nous comptons le nombre d'erreurs de prédiction : 5 individus ont été mal classés. Nous en déduisons le taux d'erreur $\epsilon_{resub} = \frac{5}{20} = 0.25$ c.-à-d. si nous classons un individu pris au hasard dans la population, nous avons 25% de chances de faire une prédiction erronée. A l'inverse, nous avons 75% de chances de faire une prédiction correcte.

Attention, il s'agit bien d'une erreur en resubstitution puisque le modèle a été élaboré (dans notre cas, les probabilités conditionnelles ont été calculées) à partir des mêmes données. Les performances annoncées sont donc sujettes à caution, surtout pour un modèle construit sur un effectif aussi faible.



	angine	coeur	Prédiction	Erreur
1	presence	presence	0
0	presence	absence	1
0	presence	absence	1
0	presence	absence	1
1	presence	presence	0
1	presence	presence	0
0	absence	absence	0
0	absence	absence	0
0	absence	absence	0
0	absence	absence	0
1	absence	presence	1
0	absence	absence	0
0	absence	absence	0
0	absence	absence	0
0	absence	absence	0
0	absence	absence	0
1	absence	presence	1
0	absence	absence	0
0	absence	absence	0
0	absence	absence	0

Mauvais classement	**5**
Taux d'erreur	**0.25**

Fig. 1.2. COEUR vs. ANGINE - Évaluation des performances

1.1.5 Insuffisances de l'approche basée sur les fréquences

La démarche basée sur les fréquences est extrêmement séduisante par sa simplicité. Un simple comptage permet de produire les probabilités conditionnelles et déduire les règles d'affectation. Toutefois, elle n'est pas viable en situation réelle, lorsque nous avons plus d'une variable prédictive, pour différentes raisons :

1. Dans le cas où toutes les variables sont binaires, le nombre de probabilités à calculer devient rapidement prohibitif, impossible à gérer même sur des ordinateurs. Par exemple, si nous avons 20 variables, il faudrait procéder à $2 \times 2^{20} = 2.097.152$ comptages.
2. Et même si cela était possible, nous aurions la valeur 0 dans la plupart des cases de notre tableau croisé, ou tout du moins de très faibles effectifs, rendant inutilisables les estimations.
3. L'affaire se corse lorsque nous avons des descripteurs continus (ex. l'AGE dans notre fichier COEUR). Procéder par comptage global n'a plus de sens. Il faut passer par d'autres stratégies : soit en discrétisant ces variables (les découper en intervalles) ; soit en estimant par comptage les probabilités, mais localement, en se limitant au voisinage de l'observation à classer (cf. par exemple la méthode des plus proches voisins, les noyaux de Parzen [3], pages 28-35).
4. Et on ne parle même pas de la situation où l'on a un mélange de variables prédictives continues et catégorielles. La solution pourrait passer par un découpage en classes des variables continues, mais il faudrait proposer des découpages pertinents, au moins en relation avec la variable à prédire, et peut être aussi en relation avec les autres variables prédictives pour tenir compte des possibles interactions.
5. Enfin, en admettant que tous les problèmes ci-dessus aient été résolus, il reste un écueil : il n'y a pas de processus de sélection de variables inhérent à la méthode. Elle ne nous indique pas quelles sont les

variables pertinentes qu'il faut conserver, quelles sont les variables qui ne servent rien et que l'on peut évacuer. Pourtant, cet aspect est incontournable dès que l'on est confronté à un problème un tant soit peu réaliste. L'expert du domaine a certes une idée plus ou moins vague des "bonnes" variables, mais bien souvent il compte sur les techniques numériques pour préciser ses idées.

1.2 Hypothèse fondamentale de la régression logistique

Pour rendre calculable la quantité $P(Y = y_k/X)$, il nous faudra donc introduire une ou plusieurs hypothèses sur les distributions. Nous sommes dans le cadre des méthodes dites "paramétriques" (ou "semi-paramétriques", nous préciserons la distinction plus loin). Elles semblent plus contraignantes par rapport aux méthodes dites non-paramétriques qui, elles, procèdent à l'estimation des probabilités sans jamais introduire des hypothèses sur les distributions (ex. les arbres de décision, la méthode des plus proches voisins, etc.). En effet, lors du traitement d'un problème réel, il faudrait en toute rigueur s'assurer de la crédibilité des hypothèses avant de pouvoir mettre en oeuvre la technique.

En pratique, ce n'est pas nécessaire. On se rend compte que les méthodes paramétriques sont souvent robustes. Elles restent opérationnelles même lorsque l'on s'écarte assez fortement des hypothèses qui les sous-tendent. L'idée la plus importante à retenir finalement est que les hypothèses pèsent sur la forme de la frontière induite pour distinguer les classes dans l'espace de représentation. La régression logistique par exemple produit un séparateur linéaire[4], c'est la principale information qu'il faut retenir.

Avant de décrire les hypothèses introduites dans la régression logistique, reconsidérons la probabilité conditionnelle $P(Y = y_k/X)$:

$$\begin{aligned} P(Y = y_k/X) &= \frac{P(Y = y_k) \times P(X/Y = y_k)}{P(X)} \\ &= \frac{P(Y = y_k) \times P(X/Y = y_k)}{\sum_k P(Y = y_k) \times P(X/Y = y_k)} \end{aligned}$$

Dans le cas à deux classes, nous devons comparer simplement $P(Y = +/X)$ et $P(Y = -/X)$. Formons-en le rapport,

$$\frac{P(Y = +/X)}{P(Y = -/X)} = \frac{P(Y = +)}{P(Y = -)} \times \frac{P(X/Y = +)}{P(X/Y = -)} \tag{1.1}$$

La règle de décision devient

$$\text{Si } \frac{P(Y = +/X)}{P(Y = -/X)} > 1 \text{ Alors } Y = +$$

Revenons à l'expression ci-dessus (Équation 1.1),

4. Je me rappelle d'une discussion animée avec un ami qui soutenait que la régression logistique est une régression non-linéaire. Oui, effectivement il a raison, c'est une régression non-linéaire parce que la fonction de transfert est non linéaire, la fonction logistique en l'occurrence. C'est un point de vue que l'on retrouve souvent en statistique ou en économétrie. En revanche, pour séparer les positifs et les négatifs, elle construit une frontière linéaire, basée sur une combinaison linéaire des variables. C'est en ce sens qu'on parle d'un classifieur linéaire. On retrouve volontiers ce point de vue en reconnaissance des formes.

- Le rapport $\frac{P(Y=+)}{P(Y=-)}$ est facile à estimer dès lors que l'échantillon est issu d'un tirage aléatoire dans la population, indépendamment des classes d'appartenance des individus. Il suffit de prendre le rapport entre le nombre d'observations positives et négatives $\frac{n_+}{n_-}$.
- Et quand bien même l'échantillon serait issu d'un tirage à deux niveaux – on parle de tirage rétrospectif (ou "données cas-témoin" lorsque l'on fixe à l'avance le nombre d'observations positives et négatives que l'on souhaite obtenir, on procède alors par tirage aléatoire dans chaque groupe (voir [3], page 5 ; [9], pages 205 à 210 ; [23], pages 431 à 434) – il est possible de procéder à des redressements si l'on connaît par ailleurs la vraie valeur de la prévalence $p = P(Y = +)$ (voir [2], pages 67 et 68, ou [3], pages 79 et 80, pour une présentation rapide ; [9], chapitre 6, pour une présentation plus détaillée et l'étude d'autres schémas d'échantillonnage).

Le véritable enjeu réside donc dans l'estimation du rapport de probabilité $\frac{P(X/Y=+)}{P(X/Y=-)}$. La régression logistique introduit l'hypothèse fondamentale suivante :

$$\ln\left[\frac{P(X/Y=+)}{P(X/Y=-)}\right] = b_0 + b_1X_1 + \cdots + b_JX_J \tag{1.2}$$

Cette hypothèse couvre une large palette de lois de distribution des données ([2], page 64) :

- La loi normale (comme pour l'analyse discriminante) ;
- Les lois exponentielles ;
- Les lois discrètes ;
- Les lois Beta, les lois Gamma et les lois de Poisson ;
- Un mélange de variables explicatives binaires (0/1) et continues, cette propriété est très importante car elle rend opérationnelle la régression logistique dans de très nombreuses configurations.

Contrairement à l'Analyse Discriminante Linéaire, que l'on qualifie de méthode paramétrique car on émet une hypothèse sur les distributions respectives de $P(X/Y = +)$ et $P(X/Y = -)$ (loi normale), **la régression logistique est une méthode semi-paramétrique** car l'hypothèse porte uniquement sur le rapport de ces probabilités. Elle est moins restrictive. Son champ d'action est donc théoriquement plus large [5].

1.3 Le modèle LOGIT

La régression logistique peut être décrite d'une autre manière. Pour un individu ω, on appelle transformation LOGIT de $\pi(\omega)$ l'expression ([9], page 6 pour la régression simple, page 31 pour la régression multiple)

5. En théorie seulement. En pratique, ces deux méthodes présentent souvent des performances similaires (voir [7], chapitre 7, en particulier la section 7.1.5, page 145 ; [21], page 480 ; [8], pages 103 à 105). Entre autres parce qu'elles induisent un séparateur linéaire dans l'espace de représentation ([8], chapitre 4, pages 79 à 113). La régression logistique ne se démarque vraiment que lorsque l'une des hypothèses de l'Analyse Discriminante Linéaire, l'homoscédasticité, est très fortement remise en cause. Toujours selon ce même point de vue, lorsque les classes ne sont pas linéairement séparables dans l'espace de représentation, la régression logistique, tout comme l'analyse discriminante linéaire, ne nous est d'aucun secours.

$$\ln\left[\frac{\pi(\omega)}{1-\pi(\omega)}\right] = a_0 + a_1X_1 + \cdots + a_JX_J \tag{1.3}$$

La quantité $\frac{\pi}{1-\pi} = \frac{P(Y=+/X)}{P(Y=-/X)}$ exprime un **odds** c.-à-d. un rapport de chances. Par exemple, si un individu présente un odds de 2, cela veut dire qu'il a 2 fois plus de chances d'être positif que d'être négatif.

Posons $C(X) = a_0 + a_1X_1 + \cdots + a_JX_J$, nous pouvons revenir sur π avec la fonction logistique

$$\pi = \frac{e^{C(X)}}{1+e^{C(X)}} \tag{1.4}$$

$$= \frac{1}{1+e^{-C(X)}} \tag{1.5}$$

Quelques commentaires et remarques

A propos de la fonction de transformation,

- Le LOGIT $= C(X)$ est théoriquement défini entre $-\infty$ et $+\infty$.
- En revanche, $0 \leq \pi \leq 1$ issue de la transformation de $C(X)$ (Figure 1.3) représente une probabilité.

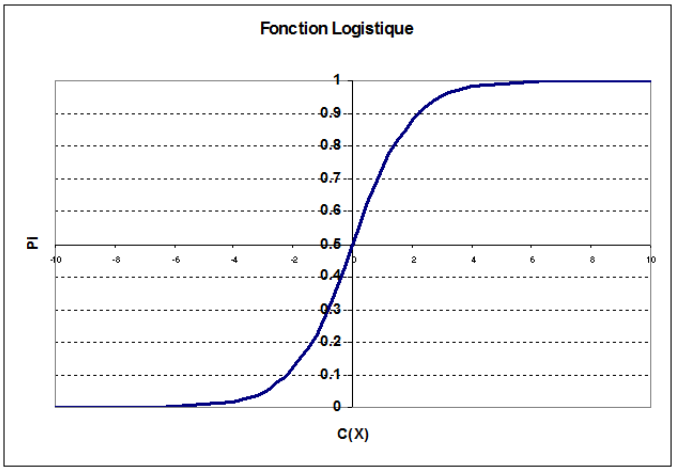


Fig. 1.3. Fonction Logistique

A propos de la règle d'affectation,

- La règle d'affectation peut être basée sur π de différentes manières :
  - Si $\frac{\pi}{1-\pi} > 1$ Alors $Y = +$
  - Si $\pi > 0.5$ Alors $Y = +$
- Elle peut être aussi basée simplement sur $C(X)$ avec :
  - Si $C(X) > 0$ Alors $Y = +$

Autres remarques,

- $C(X)$ et π permettent tous deux de "scorer" les individus, et par là de les classer selon leur propension à être "positif". Cette fonctionnalité est très utilisée dans le ciblage marketing par exemple. On parle de "scoring".
- Sauf que π représente une probabilité, avec les propriétés inhérentes à une probabilité, entres autres $P(Y = +/X) + P(Y = -/X) = 1$.
- D'autres fonctions de transformation existent. Si on utilise la fonction de répartition normale par exemple, on parle de modèle PROBIT (voir [23], page 395).
- Comme nous avions pu le dire déjà plus haut, la fonction de transfert logistique est non linéaire (Figure 1.3), c'est en ce sens que l'on qualifie la régression logistique de régression non-linéaire dans la littérature.

Équivalence entre les approches

Les deux approches ci-dessus correspondent à deux facettes d'un même problème. En effet :

$$\begin{aligned}
\ln\left[\frac{\pi}{1-\pi}\right] &= a_0 + a_1 X_1 + \cdots + a_J X_J \\
&= \ln\left[\frac{P(Y=+)}{P(Y=-)} \times \frac{P(X/Y=+)}{P(X/Y=-)}\right] \\
&= \ln\left[\frac{P(Y=+)}{P(Y=-)}\right] + \ln\left[\frac{P(X/Y=+)}{P(X/Y=-)}\right] \\
&= \ln\left[\frac{p}{1-p}\right] + (b_0 + b_1 X_1 + \cdots + a_J X_J)
\end{aligned}$$

Les deux formulations (Équations 1.2 et 1.3) sont identiques à une constante près

$$a_0 = \ln\left[\frac{p}{1-p}\right] + b_0$$

Il faudra s'en souvenir lorsque les données sont issues d'un mode d'échantillonnage autre que le tirage aléatoire simple (schéma de mélange) dans la population.

1.4 Estimation des paramètres par la maximisation de la vraisemblance

Pour estimer les paramètres de la régression logistique par la méthode du maximum de vraisemblance, nous devons tout d'abord déterminer la loi de distribution de $P(Y/X)$.

Y est une variable binaire définie dans $\{+, -\}$, (ou $\{1, 0\}$ pour simplifier les écritures). Pour un individu ω, on modélise la probabilité à l'aide de la loi binomiale $\mathcal{B}(1, \pi)$, avec

$$P[Y(\omega)/X(\omega)] = \pi(\omega)^{y(\omega)} \times (1 - \pi(\omega))^{(1-y(\omega))} \tag{1.6}$$

Cette modélisation est cohérente avec ce qui a été dit précédemment, en effet :

- Si $y(\omega) = 1$, alors $P[Y(\omega) = 1/X(\omega)] = \pi$;
- Si $y(\omega) = 0$, alors $P[Y(\omega) = 0/X(\omega)] = 1 - \pi$;

Vraisemblance

La vraisemblance (en anglais *likelihood*) d'un échantillon Ω s'écrit

$$L = \prod_{\omega} \pi(\omega)^{y(\omega)} \times (1 - \pi(\omega))^{(1-y(\omega))} \tag{1.7}$$

Pour alléger l'écriture, nous utiliserons pour la suite

$$L = \prod_{\omega} \pi^{y} \times (1 - \pi)^{(1-y)}$$

N'oublions pas que la vraisemblance correspond à la probabilité d'obtenir l'échantillon Ω à partir d'un tirage dans la population. Elle varie donc entre 0 et 1. La méthode du maximum de vraisemblance consiste à produire les paramètres $a = (a_0, a_1, \ldots, a_J)$ de la régression logistique qui rendent maximum la probabilité d'observer cet échantillon [11] (page 81).

Log-vraisemblance

Pour faciliter les manipulations, on préfère souvent travailler sur la log-vraisemblance (*log-likelihood*)

$$LL = \sum_{\omega} y \times \ln \pi + (1 - y) \times \ln(1 - \pi) \tag{1.8}$$

Le logarithme étant une fonction monotone, le vecteur **a** qui maximise la vraisemblance est le même que celui qui maximise la log-vraisemblance. Cette dernière en revanche varie entre $-\infty$ et 0.

Puisque $\hat{a}$ est un estimateur du maximum de vraisemblance, il en possède toutes les propriétés :

1. Il est asymptotiquement sans biais ;
2. Il est de variance minimale ;
3. Il est asymptotiquement gaussien.

Ces éléments, notamment le dernier, seront très importants pour l'inférence statistique (intervalle de confiance, test de significativité, etc.).

Déviance

Bien souvent, on utilise la quantité

$$D_M = -2LL = -2 \times LL \tag{1.9}$$

appelée déviance [9] (page 13) (ou déviance résiduelle, en anglais *residual deviance*, dans certains logiciels tels que R) (D_M). Contrairement à la log-vraisemblance, elle est positive. L'objectif de l'algorithme d'optimisation est de minimiser cette déviance. On peut faire le parallèle avec la somme des carrés des résidus de la régression linéaire multiple. La *null deviance* (D_0) calculée sur le modèle uniquement composée de la constate correspondrait alors à la somme des carrés totaux [10] (pages 20 à 27).

Dans certains ouvrages, on définit la déviance D de manière plus générique (cf. [9], page 13 ; [23], page 405 ; [7], page 115) :

$$\begin{aligned} D &= 2 \times \ln\left[\frac{L(\text{Modèle saturé})}{L(\text{Modèle étudié})}\right] \\ &= -2 \times LL(\text{Modèle étudié}) - [-2 \times LL(\text{Modèle saturé})] \\ &= D_M - [-2 \times LL(\text{Modèle saturé})] \\ &= -2\sum_{\omega}\left[y \ln\left(\frac{\hat{\pi}}{y}\right) + (1-y)\ln\left(\frac{1-\hat{\pi}}{1-y}\right)\right] \end{aligned}$$

Un modèle saturé pour des données individuelles [6] est un modèle reconstituant parfaitement les valeurs de la variable dépendante c.-à-d. $\hat{\pi}(\omega) = y(\omega)$. Sa vraisemblance est égale à 1 (Équation 1.4), et sa log-vraisemblance 0 (Équation 1.8). Dans ce contexte, $D = D_M$.

Optimisation

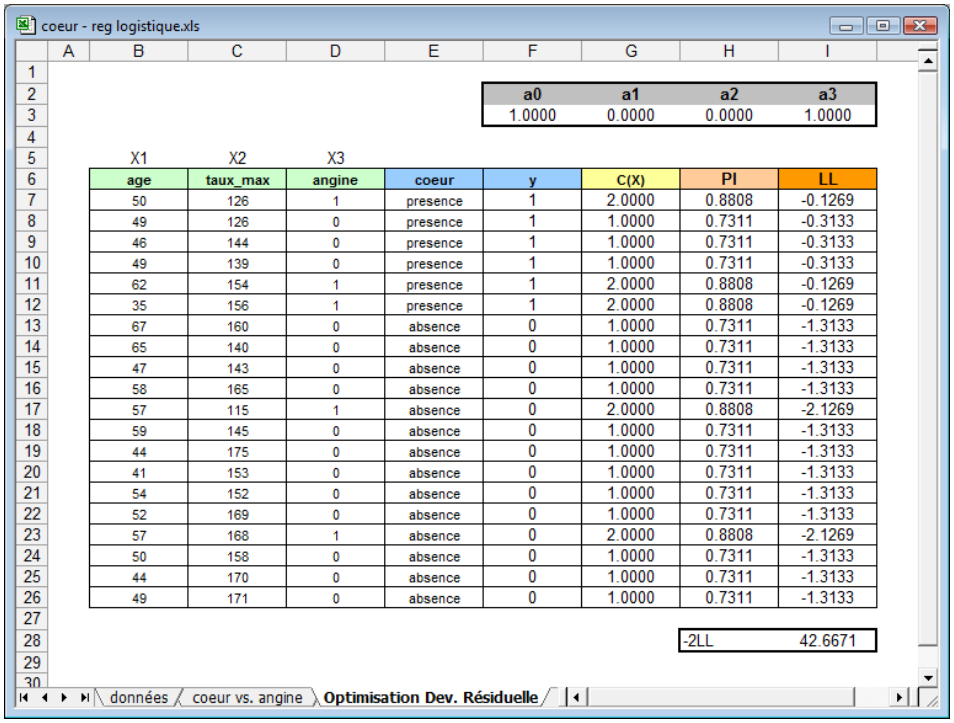
Bonne nouvelle, la log-vraisemblance est une fonction convexe. Il existe donc une solution unique $\hat{a}$. Mauvaise nouvelle, il n'existe pas de solution analytique. Il faut passer par des heuristiques. Ce qui explique que l'on obtienne parfois des résultats différents d'un logiciel à l'autre : le résultat obtenu dépend de l'algorithme utilisé, du paramétrage adopté, et parfois même des choix d'implémentation de l'informaticien. Ces différences déroutent le néophyte. En réalité, il n'y aucune raison de s'en inquiéter si on connaît un peu la technique. Les divergences entre les logiciels ne doivent nous alerter que si elles sont trop importantes.

Plusieurs techniques d'optimisation existent, les logiciels s'appuient souvent sur l'algorithme de Newton-Raphson [23] (pages 398 à 400) ou de ses variantes (ex. Fisher Scoring). Nous en reparlerons en détail plus loin (section 1.5). Cet aspect est très important. En effet, il peut influencer les résultats, il explique également les éventuels plantages des logiciels (ah ! le fameux "ça marche pas !").

6. A distinguer de la situation de "covariate pattern" où plusieurs observations, dont certaines sont positives, d'autres négatives, partagent la même description [9] (page 144). C'est le cas lorsque les données sont issues d'expérimentations ou lorsque les variables explicatives sont toutes catégorielles [1] (pages 91 à 97). On parle aussi de *situation de données groupées* [23] (pages 434 à 438). Le modèle saturé correspond alors au modèle où l'on aura tenu compte de toutes les interactions possibles entre les variables explicatives. Pour une étude plus approfondie, voir le chapitre 9.

COEUR = f (AGE, TAUX MAX, ANGINE)

Pour illustrer notre propos, nous allons estimer les paramètres de la régression logistique pour notre problème de prédiction de maladie cardiaque. Nous organisons les calculs dans le tableur Excel, puis nous utiliserons le solveur [7] pour minimiser la déviance.



	A	B	C	D	E	F	G	H	I
1									
2						a0	a1	a2	a3
3						1.0000	0.0000	0.0000	1.0000
4									
5		X1	X2	X3					
6		age	taux_max	angine	coeur	y	C(X)	PI	LL
7		50	126	1	presence	1	2.0000	0.8808	-0.1269
8		49	126	0	presence	1	1.0000	0.7311	-0.3133
9		46	144	0	presence	1	1.0000	0.7311	-0.3133
10		49	139	0	presence	1	1.0000	0.7311	-0.3133
11		62	154	1	presence	1	2.0000	0.8808	-0.1269
12		35	156	1	presence	1	2.0000	0.8808	-0.1269
13		67	160	0	absence	0	1.0000	0.7311	-1.3133
14		65	140	0	absence	0	1.0000	0.7311	-1.3133
15		47	143	0	absence	0	1.0000	0.7311	-1.3133
16		58	165	0	absence	0	1.0000	0.7311	-1.3133
17		57	115	1	absence	0	2.0000	0.8808	-2.1269
18		59	145	0	absence	0	1.0000	0.7311	-1.3133
19		44	175	0	absence	0	1.0000	0.7311	-1.3133
20		41	153	0	absence	0	1.0000	0.7311	-1.3133
21		54	152	0	absence	0	1.0000	0.7311	-1.3133
22		52	169	0	absence	0	1.0000	0.7311	-1.3133
23		57	168	1	absence	0	2.0000	0.8808	-2.1269
24		50	158	0	absence	0	1.0000	0.7311	-1.3133
25		44	170	0	absence	0	1.0000	0.7311	-1.3133
26		49	171	0	absence	0	1.0000	0.7311	-1.3133
27									
28								-2LL	42.6671
29									
30									

Fig. 1.4. Préparation de la feuille de calcul - Minimisation de la déviance

Dans un premier temps, nous devons préparer la feuille Excel (Figure 1.4) :

- En F3..I3, nous introduisons les valeurs de départ des coefficients, le solveur a besoin de cette initialisation, elle nous permet également de vérifier l'intégrité de la feuille de calcul. Nous mettons, au hasard [8], $a = (1.0, 0.0, 0.0, 1.0)$.
- En colonnes B, C, D et E, nous avons le jeu de données.
- En F, nous plaçons la variable Y recodée en 0/1.
- Nous calculons alors $C(X)$. Pour le première observation, nous avons $C(X) = 1.0 + 0.0 \times 50 + 0.0 \times 126 + 1.0 \times 1 = 2.0$.
- Nous en déduisons alors π. Toujours pour la première observation, nous obtenons $\pi = \frac{1}{1+e^{-2.0}} = 0.8808$.

7. A propos de l'utilisation du solveur, des sites de cours en ligne sont référencés sur ma page consacrée à Excel : `http://eric.univ-lyon2.fr/~ricco/cours/cours_excel.html`

8. L'initialisation est faite au hasard. En théorie, n'importe quelle valeur conviendrait. En pratique, on a intérêt à mettre des valeurs proches de la solution définitive. A défaut, on conseille généralement de tenter plusieurs valeurs de départ.

– La fraction de la log-vraisemblance correspondante est égale à $LL = y \times \ln(\pi) + (1-y) \times \ln(1-\pi) = 1 \times \ln(0.8808) + (1-1) \times \ln(1-0.8808) = -0.1269$.
– Il ne nous reste plus qu'à calculer la déviance $D_M = -2LL = -2 \times (-0.1269 - 0.3133 \dots) = 42.6671$.

Nous pouvons actionner le solveur à ce stade. Nous souhaitons minimiser la cellule cible I28 contenant l'expression de la déviance. Les cellules variables sont celles contenant les paramètres de la régression logistique, à savoir les cellules F3 à I3. Il n'y a pas de contraintes dans cette optimisation.

coeur - reg logistique.xls

	a0	a1	a2	a3
	14.4937	-0.1256	-0.0636	1.7790

	X1	X2	X3					
	age	taux_max	angine	coeur	y	C(X)	PI	LL
1	50	126	1	presence	1	1.9825	0.8789	-0.1290
2	49	126	0	presence	1	0.3291	0.5815	-0.5421
3	46	144	0	presence	1	-0.4381	0.3922	-0.9360
4	49	139	0	presence	1	-0.4972	0.3782	-0.9723
5	62	154	1	presence	1	-1.3048	0.2134	-1.5448
6	35	156	1	presence	1	1.9601	0.8765	-0.1318
7	67	160	0	absence	0	-4.0933	0.0164	-0.0165
8	65	140	0	absence	0	-2.5708	0.0710	-0.0737
9	47	143	0	absence	0	-0.5001	0.3775	-0.4740
10	58	165	0	absence	0	-3.2804	0.0362	-0.0369
11	57	115	1	absence	0	1.8022	0.8584	-1.9549
12	59	145	0	absence	0	-2.1348	0.1058	-0.1118
13	44	175	0	absence	0	-2.1572	0.1037	-0.1094
14	41	153	0	absence	0	-0.3820	0.4057	-0.5203
15	54	152	0	absence	0	-1.9516	0.1244	-0.1328
16	52	169	0	absence	0	-2.7809	0.0584	-0.0601
17	57	168	1	absence	0	-1.5665	0.1727	-0.1896
18	50	158	0	absence	0	-1.8304	0.1382	-0.1487
19	44	170	0	absence	0	-1.8394	0.1371	-0.1475
20	49	171	0	absence	0	-2.5311	0.0737	-0.0766

-2LL	16.6177

données / coeur vs. angine / Optimisation Dev. Résiduelle

Fig. 1.5. Feuille de calcul après minimisation de la déviance

Nous obtenons une nouvelle version de la feuille de calcul à la sortie (Figure 1.5). La déviance est passée à $D_M = 16.6117$. Les valeurs des paramètres qui ont permis de l'obtenir sont

$$\hat{a} = (14.4937, -0.1256, -0.0636, 1.7790)$$

En d'autres termes, le LOGIT estimé permettant de prédire l'occurrence d'une maladie cardiaque à partir de l'âge, le taux max et l'angine, s'écrit :

$$C(X) = 14.4937 - 0.1256 \times X_1 - 0.0636 \times X_2 + 1.7790 \times X_3$$

1.5 L'algorithme de Newton-Raphson

Bien entendu, notre exemple implémenté sous Excel est à visée pédagogique. Dans les études réelles, nous avons intérêt à utiliser les logiciels spécialisés qui produisent directement les résultats. L'idée était de détailler les calculs de manière à ce que le lecteur puisse retracer les formules décrites en amont.

Mais justement, qu'en est-il des logiciels ? Quel est l'algorithme utilisé ? Ce choix peut-il avoir des répercussions sur les résultats ? Peut-on en obtenir d'autres informations qui pourraient être utiles pour l'inférence statistique ?

L'algorithme de Newton-Raphson est une des méthodes numériques les plus utilisées pour optimiser la log-vraisemblance ([23], page 398 ; [9], page 33 ; [11], page 162). Il démarre avec une initialisation quelconque du vecteur de paramètre $\mathbf{a}$; pour passer de l'étape (i) à l'étape $(i+1)$, il se rapproche de la solution finale $\hat{a}$ en utilisant la formule suivante

$$a^{i+1} = a^i - \left(\frac{\partial LL}{\partial a.\partial a'}\right)^{-1} \times \frac{\partial LL}{\partial a} \tag{1.10}$$

1.5.1 Quelques remarques

Plusieurs règles d'arrêt sont possibles pour stopper le processus de recherche :

- On fixe à l'avance le nombre maximum d'itérations pour limiter le temps de calcul. C'est un peu fruste mais souvent bien utile pour éviter les boucles infinies faute de convergence.
- On stoppe les itérations lorsque l'évolution de la log-vraisemblance d'une étape à l'autre n'est pas significative. Pour cela, on fixe souvent une valeur seuil ϵ, on arrête le processus si l'écart d'une étape à l'autre est plus petit que le seuil.
- On stoppe les itérations lorsque l'écart entre les vecteurs solutions $\hat{a}$ est faible d'une étape à l'autre. Ici également, souvent il s'agit de fixer un seuil à l'avance auquel on compare la somme des écarts aux carrés ou la somme des écarts absolus entre les composantes des vecteurs solutions.

Dans ce contexte, il ne faut pas s'étonner qu'il y ait des disparités entre les logiciels. Quand bien même ils utiliseraient le même algorithme d'optimisation, avec les mêmes valeurs de départ, rien que le paramétrage de la règle d'arrêt peut produire des solutions différentes. Certains logiciels donnent à l'utilisateur la possibilité d'affiner les seuils. D'autres utilisent des seuils prédéfinis connus d'eux seuls [9]. Et on ne parle même pas des astuces destinées à accélérer les calculs ou à les sécuriser. Ce dernier point est important. En effet on remarque que le processus comporte une étape d'inversion de matrice (la matrice hessienne). Voilà un danger qu'il convient de circonscrire. Les stratégies adoptées pèsent sur le résultat obtenu à l'issue du processus d'optimisation.

9. A ce propos, le logiciel libre, *open source*, est une garantie de transparence qui nous donne l'opportunité d'inspecter le code source et de comprendre les divergences entre les logiciels, voire entre les versions du même logiciel ! A défaut, nous sommes condamnés à subir le bon vouloir des éditeurs.

1.5.2 Vecteur des dérivées partielles premières de la log-vraisemblance

Dans l'équation 1.10, le vecteur de dimension $(J+1\times 1)$ correspondant aux dérivées partielles premières de la log-vraisemblance $\nabla(a) = \frac{\partial LL}{\partial a}$ retient notre attention. On parle de **vecteur score** ou de **vecteur gradient**. Voyons-en le détail pour la variable X_j :

$$\nabla(a_j) = \sum_{\omega}[y(\omega) - \pi(\omega)] \times x_j(\omega) \tag{1.11}$$

Lorsque la solution a été trouvée c.-à-d. le vecteur $\hat{a}$ permettant d'optimiser LL est obtenu, toutes les composantes du vecteur gradient sont égales à 0. C'est tout à fait normal. On cherche un optimum dans un espace convexe. La solution annule la dérivée première par rapport aux paramètres [10].

1.5.3 Matrice des dérivées partielles secondes de la log-vraisemblance

Autre expression qui retient notre attention toujours dans l'équation 1.10, il s'agit de la matrice des dérivées partielles seconde $H(a) = \frac{\partial LL}{\partial a.\partial a'}$, dite **matrice hessienne**. Elle est très importante car son inverse correspond à la matrice des variances covariances des coefficients, précieuse lors de l'inférence statistique (tests et intervalle de confiance).

$H(a)$ est de dimension $(J + 1 \times J + 1)$ d'expression générale :

$$H(j_1, j_2) = \sum_{\omega} x_{j_1}(\omega) \times x_{j_2}(\omega) \times \pi(\omega) \times [1 - \pi(\omega)] \tag{1.12}$$

Il est parfois plus commode de passer par une notation matricielle, nous pouvons écrire

$$H(a) = -X'VX \tag{1.13}$$

où V est une matrice diagonale de taille $(n \times n)$ composée de $\pi(\omega) \times (1 - \pi(\omega))$.

1.6 Première évaluation de la régression : les pseudo-R^2

Une question cruciale est de pouvoir déterminer si le modèle obtenu est "intéressant" ou non. Le premier à pouvoir trancher est l'expert. En se basant sur les contraintes du domaine, il peut nous dire si le modèle est suffisamment concluant. En son absence, il ne faut surtout pas se lancer dans des considérations plus ou moins vaseuses, basées essentiellement sur le taux d'erreur en resubstitution. La seule attitude viable est de poser la question "à quel classifieur de référence peut-on se comparer ?".

10. Rappelons-nous les exercices d'optimisation d'équations du second degré (une parabole). La démarche consiste à calculer la dérivée première, qui est une équation du premier degré, puis de l'annuler. L'idée est *grosso modo* la même ici.

Dans le cadre de l'apprentissage supervisé, le *classifieur de référence* est le modèle qui n'utilise pas les informations en provenance des variables indépendantes X_j. On parle également de *classifieur par défaut* (en anglais *default classifier*). En régression logistique, il correspond au modèle M_0 (on parle également de "modèle initial", de "modèle trivial" ; en anglais *null model*) n'incluant que la constante a_0.

Dans ce qui suit, nous montrons (1) comment estimer directement le paramètre a_0 du modèle réduit à la simple constante, (2) comment obtenir la déviance sans avoir à la calculer explicitement, (3) nous présenterons alors plusieurs indicateurs, de type $Pseudo - R^2$, basés sur la comparaison des déviances respectives du modèle étudié (D_M) et du modèle par défaut (D_0).

Remarque : L'analogie avec le coefficient de détermination R^2 de la régression linéaire multiple est tout à fait intéressante. En effet, il est usuellement interprété comme la part de variance expliquée par le modèle. Mais il peut être également compris comme une confrontation entre les performances du modèle analysé (traduite par la somme des carrés des résidus $SCR = \sum_\omega (y - \hat{y})^2$) et celles du modèle par défaut réduite à la simple constante (dans ce cas, la constante est estimée par la moyenne de l'endogène $\bar{y}$, la somme des carrés totaux correspond donc à la somme des carrés des résidus du modèle réduit à la simple constante $SCT = \sum_\omega (y - \bar{y})^2$). N'oublions pas que $R^2 = \frac{SCT - SCR}{SCT} = 1 - \frac{SCR}{SCT}$. Sa définition répond exactement à la notion d'efficacité prédictive ([10] (page 28), [1], pages 110 à 112).

1.6.1 Estimation du paramètre a_0 et de la déviance du modèle trivial

Le modèle trivial est réduit à la seule constante c.-à-d.

$$LOGIT(M_0) = \ln\left[\frac{\pi}{1-\pi}\right] = a_0$$

Nous ne tenons pas compte des variables explicatives X_j. De fait :

$$\begin{aligned}\frac{\pi}{1-\pi} &= \frac{p}{1-p} \times \frac{P(X/Y=+)}{P(X/Y=-)} \\ &= \frac{p}{1-p}\end{aligned}$$

On devine aisément **l'estimation** $\hat{a}_0$ de la régression

$$\begin{aligned}\hat{a}_0 &= \ln\left[\frac{\hat{p}}{1-\hat{p}}\right] \\ &= \ln\left[\frac{n_+}{n_-}\right]\end{aligned}$$

Le nombre de positifs n_+ et négatifs n_- dans l'échantillon suffit pour estimer le paramètre du modèle trivial. Pour prédire la probabilité a posteriori pour un individu d'être positif $\hat{\pi}(\omega)$, nous utilisons simplement la proportion des positifs $\hat{p} = \frac{n_+}{n}$ dans la base, soit

$$\hat{\pi}(\omega) = \hat{p},\ \forall\ \omega$$

Voyons maintenant ce qu'il en est de la **log-vraisemblance**. A partir de l'équation 1.8 et des considérations ci-dessus, nous obtenons

$$\begin{aligned} LL_0 &= \sum_{\omega} y \times \ln(\hat{p}) + (1-y) \times \ln(1-\hat{p}) \\ &= \sum_{\omega} y \times \ln(\hat{p}) + \sum_{\omega} (1-y) \times \ln(1-\hat{p}) \\ &= n_+ \times \ln(\hat{p}) + n_- \times \ln(1-\hat{p}) \\ &= n \times \ln(1-\hat{p}) + n_+ \times \ln(\frac{\hat{p}}{1-\hat{p}}) \end{aligned}$$

Nous pouvons en déduire la déviance du modèle trivial

$$D_0 = -2 \times LL_0$$

COEUR = f(∅) - Estimation directe

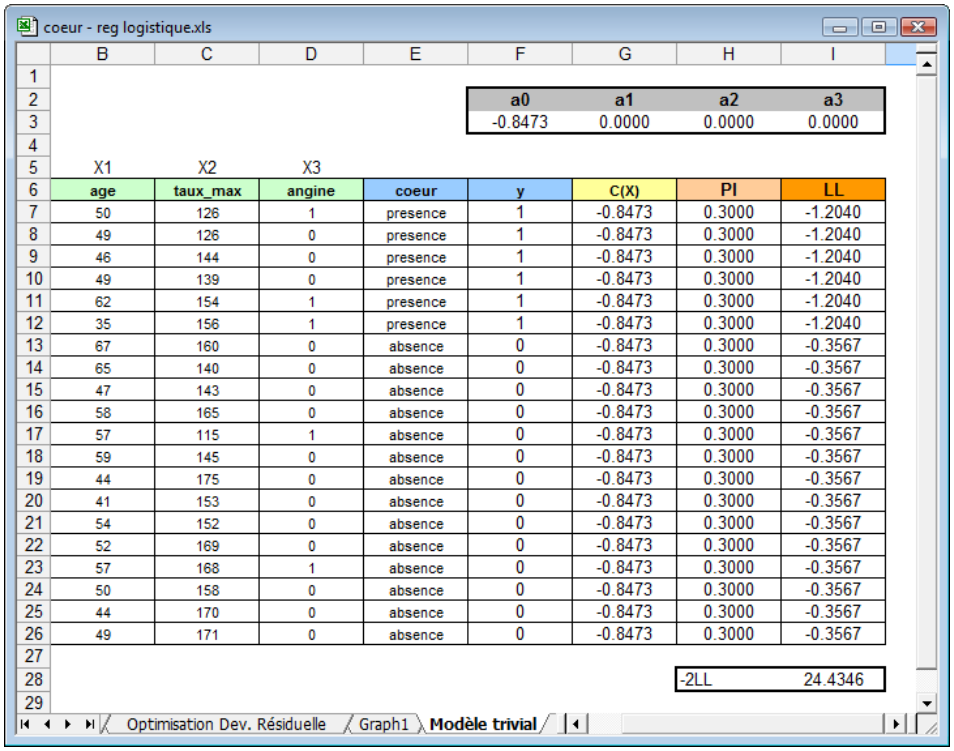
Reprenons l'exemple du fichier COEUR (Figure 0.1). Nous y observons $n_+ = 6$ observations positives parmi $n = 20$. Nous obtenons directement :

- Le nombre de négatifs $n_- = 20 - 6 = 14$
- La proportion de positifs $\hat{p} = \frac{6}{20} = 0.3$
- L'estimation de la constante $\hat{a}_0 = \ln\left[\frac{n_+}{n_-}\right] = \ln\left[\frac{6}{14}\right] = -0.8473$
- La log-vraisemblance $LL(0) = 20 \times \ln(1-0.3) + 6 \times \ln\left[\frac{0.3}{1-0.3}\right] = -12.217$
- La déviance $D_0 = -2 \times LL(0) = -2 \times (-12.217) = 24.4346$

COEUR = f(∅) - Estimation usuelle

Par curiosité, nous souhaitons vérifier si les résultats de l'estimation directe concordent avec ceux de la procédure usuelle. Nous reprenons notre feuille Excel (Figure 1.4). Nous la modifions en 2 temps : (1) nous annulons les coefficients associés aux variables explicatives c.-à-d. $a_1 = a_2 = a_3 = 0$; (2) nous lançons le solveur en spécifiant uniquement a_0 (cellule F3) en cellule variable.

Les résultats (Figure 1.6) sont totalement cohérents avec l'approche directe : l'estimation $\hat{a}_0 = -0.8473$ et la déviance $D_0 = 24.4346$. Ce qui est plutôt encourageant. Le calcul direct nous épargne une optimisation compliquée. Nous remarquerons également que $\hat{\pi}(\omega) = \hat{p} = 0.3,\ \forall\omega$.



	a0	a1	a2	a3
	-0.8473	0.0000	0.0000	0.0000

	X1	X2	X3					
	age	taux_max	angine	coeur	y	C(X)	PI	LL
1	50	126	1	presence	1	-0.8473	0.3000	-1.2040
2	49	126	0	presence	1	-0.8473	0.3000	-1.2040
3	46	144	0	presence	1	-0.8473	0.3000	-1.2040
4	49	139	0	presence	1	-0.8473	0.3000	-1.2040
5	62	154	1	presence	1	-0.8473	0.3000	-1.2040
6	35	156	1	presence	1	-0.8473	0.3000	-1.2040
7	67	160	0	absence	0	-0.8473	0.3000	-0.3567
8	65	140	0	absence	0	-0.8473	0.3000	-0.3567
9	47	143	0	absence	0	-0.8473	0.3000	-0.3567
10	58	165	0	absence	0	-0.8473	0.3000	-0.3567
11	57	115	1	absence	0	-0.8473	0.3000	-0.3567
12	59	145	0	absence	0	-0.8473	0.3000	-0.3567
13	44	175	0	absence	0	-0.8473	0.3000	-0.3567
14	41	153	0	absence	0	-0.8473	0.3000	-0.3567
15	54	152	0	absence	0	-0.8473	0.3000	-0.3567
16	52	169	0	absence	0	-0.8473	0.3000	-0.3567
17	57	168	1	absence	0	-0.8473	0.3000	-0.3567
18	50	158	0	absence	0	-0.8473	0.3000	-0.3567
19	44	170	0	absence	0	-0.8473	0.3000	-0.3567
20	49	171	0	absence	0	-0.8473	0.3000	-0.3567

-2LL	24.4346

Fig. 1.6. Fichier COEUR - Modèle trivial

Indicateur	Formule	Valeur Min/Max et Commentaires	Fichier COEUR
R^2 de McFadden	$R^2_{MF} = 1 - \frac{LL_M}{LL_0}$	Min = 0 si $LL_M = LL_0$, on ne fait pas mieux que le modèle trivial. Max = 1 si $LL_M = 0$, notre modèle est parfait. L'analogie avec le R^2 de la régression linéaire multiple est totale	$R^2_{MF} = 1 - \frac{-8.3088}{-12.2173} = 0.3199$
R^2 de Cox and Snell	$R^2_{CS} = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}$	Min = 0. Max si $L_M = 1$, avec $max\left[R^2_{CS}\right] = 1 - L_0^{(\frac{2}{n})}$. L'indicateur n'est pas normalisé, c'est un peu gênant.	$R^2_{CS} = 1 - \left(\frac{4.94\times10^{-6}}{2.46\times10^{-4}}\right)^{\frac{2}{20}} = 0.3235$
R^2 de Nagelkerke	$R^2_N = \frac{R^2_{CS}}{max\left[R^2_{CS}\right]}$	Min = 0. Max = 1. C'est une simple normalisation du R^2 de Cox and Snell.	$R^2_N = \frac{0.3235}{0.7053} = 0.4587$

Tableau 1.1. Quelques pseudo-$R2$ - Application au fichier COEUR

1.6.2 Quelques pseudo-R^2

Les pseudo-R^2 résultent de l'opposition, sous différentes formes, de la vraisemblance du modèle étudié L_M avec celle du modèle trivial L_0. Ils quantifient la contribution des descripteurs dans l'explication de

la variable dépendante. *Grosso modo*, il s'agit de vérifier si notre modèle fait mieux que le modèle trivial c.-à-d. s'il présente une vraisemblance ou une log-vraisemblance plus favorable.

Plusieurs formes de pseudo-R^2 sont proposés dans la littérature, nous en distinguons quelques uns (Tableau 1.1) (voir [23], page 407 ; [9], page 166).

Les R^2 de Mac Fadden et de Nagelkerke sont les plus simples à appréhender : lorsque la régression ne sert à rien, les variables explicatives n'expliquent rien, l'indicateur vaut 0 ; lorsque la régression est parfaite, l'indicateur vaut 1. Menard ([10], page 27) suggère que le R^2_{MF} de McFadden est le plus adapté à la régression logistique : il est le plus proche conceptuellement du coefficient de détermination de la régression linéaire multiple ; il n'est pas sensible à des modifications de la proportion de positifs dans le fichier d'apprentissage.

Dans notre exemple, avec $R^2_{MF} = 0.3199$, il semble que notre modèle se démarque du modèle trivial. On ne saurait pas dire en revanche si l'apport est significatif ou non, nous en saurons d'avantage lorsque nous aborderons l'évaluation statistique (Chapitre 3).

2

Évaluation de la régression

Maintenant que nous avons construit un modèle de prédiction, il faut en évaluer l'efficacité. Nous pouvons le faire de différentes manières :

- Confronter les valeurs observées de la variable dépendante $Y(\omega)$ avec les prédictions $\hat{Y}(\omega)$.
- Comparer les vraies valeurs π avec celles prédites par le modèle $\hat{\pi}$. En effet, n'oublions pas que la régression logistique sait fournir une bonne approximation de cette quantité [16]. Elle peut se révéler très utile lorsque nous souhaitons classer les individus selon leurs degrés de positivité ou introduire d'autres calculs ultérieurement (ex. intégrer les coûts de mauvais classement).

Dans ce chapitre, nous nous consacrons à ce que l'on appellerait des méthodes d'évaluation externes [1], basées sur les prédictions $\hat{y}(\omega)$ et/ou les probabilités a posteriori $\hat{\pi}(\omega)$ fournies par le classifieur. A aucun moment nous n'exploitons des informations spécifiques (internes) à la régression logistique (log-vraisemblance). De fait, **les techniques et ratios présentés dans ce chapitre peuvent s'appliquer à tout classifieur issu d'un processus d'apprentissage supervisé**, pourvu qu'il sache fournir $\hat{y}(\omega)$ et $\hat{\pi}(\omega)$ (ex. analyse discriminante, arbres de décision, réseaux de neurones, etc.). On s'étonne d'ailleurs que certaines procédures, très populaires dans le cadre de la régression logistique (la construction de la courbe ROC par exemple), ne soient pas plus utilisées par ailleurs.

2.1 La matrice de confusion

2.1.1 Construction et indicateurs associés

Pour évaluer la capacité à bien classer du modèle, nous pourrions reproduire la démarche utilisée précédemment (Figure 1.2) : construire la colonne prédiction, puis la colonne erreur $\Delta(Y, \hat{Y})$, comptabiliser le nombre de mauvais classement, en déduire le taux d'erreur.

Il est plus judicieux de construire ce que l'on appelle une **matrice de confusion** (en anglais *classification table*). Elle confronte toujours les valeurs observées de la variable dépendante avec celles qui sont prédite, puis comptabilise les bonnes et les mauvaises prédictions. Son intérêt est qu'elle permet à la fois

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	Total
$+$	a	b	$a+b$
$-$	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Tableau 2.1. Matrice de confusion - Forme générique

d'appréhender la quantité de l'erreur (le taux d'erreur) et de rendre compte de la structure de l'erreur (la manière de se tromper du modèle).

Dans un problème à 2 classes (+ vs. -), à partir de la forme générique de la matrice de confusion (Tableau 2.1), plusieurs indicateurs peuvent être déduits pour rendre compte de la concordance entre les valeurs observées et les valeurs prédites (voir [10], pages 27 à 36). Nous nous concentrons sur les ratios suivants :

- a sont les *vrais positifs* c.-à-d. les observations qui ont été classées positives et qui le sont réellement.
- c sont les *faux positifs* c.-à-d. les individus classés positifs et qui sont réalité des négatifs.
- de la même manière, b sont les faux négatifs et d sont les vrais négatifs. Mais ces termes sont peu utilisés en pratique car les positifs et les négatifs n'ont pas le même statut dans la majorité des études (ex. les positifs sont les fraudeurs que l'on cherche à isoler ; les positifs sont les personnes atteintes d'une maladie que l'on chercher à détecter ; etc.).
- Le *taux d'erreur* est égal au nombre de mauvais classement rapporté à l'effectif total c.-à-d.

$$\epsilon = \frac{b+c}{n} = 1 - \frac{a+d}{n}$$

  Il estime la probabilité de mauvais classement du modèle.
- Le taux de succès correspond à la probabilité de bon classement du modèle, c'est le complémentaire à 1 du taux d'erreur

$$\theta = \frac{a+d}{n} = 1 - \epsilon$$

- La *sensibilité* (ou le *rappel*, ou encore le *taux de vrais positifs [TVP]*) indique la capacité du modèle à retrouver les positifs

$$S_e = \text{Sensibilité} = TVP = rappel = \frac{a}{a+b}$$

- La *précision* indique la proportion de vrais positifs parmi les individus qui ont été classés positifs

$$precision = \frac{a}{a+c}$$

  Elle estime la probabilité d'un individu d'être réellement positif lorsque le modèle le classe comme tel. Dans certains domaines, on parle de *valeur prédictive positive* (VPP).
- La *spécificité*, à l'inverse de la sensibilité, indique la proportion de négatifs détectés

$$S_p = \text{Spécificité} = \frac{d}{c+d}$$

1. Il faut être précis sur les terminologies. Chez certains auteurs, validation externe correspond à une évaluation du modèle sur un échantillon à part, dit échantillon test, n'ayant pas participé à la construction du modèle ([9], pages 186 à 188).

– Parfois, on utilise le *taux de faux positifs (TFP)*, il correspond à la proportion de négatifs qui ont été classés positifs c.-à-d.

$$TFP = \frac{c}{c+d} = 1 - \text{Spécificité}$$

– La F-Mesure est très utilisée en recherche d'information. Elle synthétise (moyenne harmonique) le rappel et la précision, l'importance accordée à l'une ou à l'autre est paramétrable avec β

$$F_\beta = \frac{(1+\beta^2) \times rappel \times precision}{\beta^2 \times precision + rappel}$$

Lorsque

* $\beta = 1$ est la valeur usuelle, on accorde la même importance au rappel et à la précision, la F-Mesure devient

$$F_{\beta=1} = \frac{2 \times rappel \times precision}{precision + rappel}$$

* $\beta < 1$, on accorde plus d'importance à la précision par rapport au rappel. Une valeur fréquemment utilisée est $\beta = 0.5$, on accorde deux fois plus d'importance à la précision.
* $\beta > 1$, on accorde plus d'importance au rappel par rapport à la précision. Une valeur fréquemment rencontrée est $\beta = 2$.

La F-Mesure est une moyenne harmonique entre le rappel et la précision, en effet nous pouvons l'écrire de la manière suivante

$$F = \frac{1}{\alpha \frac{1}{precision} + (1-\alpha)\frac{1}{rappel}}$$

où $\beta^2 = \frac{1-\alpha}{\alpha}$.

Quelques remarques sur le comportement de ces indicateurs

– Un "bon" modèle doit présenter des valeurs faibles de taux d'erreur et de taux de faux positifs (proche de 0) ; des valeurs élevées de sensibilité, précision et spécificité (proche de 1).
– Le taux d'erreur est un indicateur symétrique, il donne la même importance aux faux positifs (c) et aux faux négatifs (b).
– La sensibilité et la précision sont asymétriques, ils accordent un rôle particulier aux positifs.
– Enfin, en règle générale, lorsqu'on oriente l'apprentissage de manière à améliorer la sensibilité, on dégrade souvent la précision et la spécificité. Un modèle qui serait meilleur que les autres sur ces deux groupes de critères antinomiques est celui qu'il faut absolument retenir.

2.1.2 Autres indicateurs

La sensibilité et la spécificité jouent un rôle particulier dans l'évaluation des classifieurs. En effet :

– Un "bon" modèle doit présenter des valeurs élevées sur ces deux critères d'évaluation.
– Comme nous le disions plus haut, lorsqu'on oriente l'apprentissage pour améliorer la sensibilité, on dégrade (souvent) la spécificité. Raison de plus pour les surveiller simultanément.

- Tous deux partagent une propriété importante : ils ne dépendent pas du schéma d'échantillonnage. Même si l'échantillon n'est pas représentatif c.-à-d. la proportion des positifs (resp. des négatifs) ne reflète pas la probabilité d'être positif (resp. négatif), la sensibilité et la spécificité n'en sont pas affecté. Tout simplement parce que nous utilisons le "profil-ligne" de la matrice de confusion. Lorsque nous travaillons sur des données où la proportion des positifs a été fixée arbitrairement (schéma d'échantillonnage rétrospectif), cette propriété est précieuse car elle nous évite d'avoir à procéder à des redressements périlleux.
- Enfin, pour couronner le tout, la grande majorité des indicateurs d'évaluation des classifieurs peuvent s'écrire en fonction de la sensibilité et la spécificité.

Dans ce qui suit, nous ré-écrivons quelques indicateurs décrits précédemment de manière à faire ressortir la synthèse entre sensibilité et spécificité. Nous proposerons aussi d'autres indicateurs moins connus en apprentissage automatique.

Taux d'erreur

La probabilité de mal classer peut être décomposée de la manière suivante :

$$\begin{aligned} P(erreur) &= P[(Y = + \; et \; \hat{Y} = -) \; ou \; (Y = - \; et \; \hat{Y} = +)] \\ &= P(Y = + \; et \; \hat{Y} = -) + P(Y = - \; et \; \hat{Y} = +) \\ &= P(Y = +) \times P(\hat{Y} = -/Y = +) + P(Y = -) \times P(\hat{Y} = +/Y = -) \\ &= p \times (1 - S_e) + (1 - p) \times (1 - S_p) \end{aligned}$$

On le devine aisément dans cette expression, le taux d'erreur sera d'autant plus faible que la sensibilité et la spécificité sont élevés (proches de 1).

Selon le schéma d'échantillonnage, deux situations sont envisageables :

1. Nous avons un échantillon représentatif c.-à-d. nous pouvons estimer p à l'aide de $\hat{p} = \frac{a+b}{n}$, nous avons

$$\begin{aligned} \epsilon &= \frac{a+b}{n}(1 - \frac{a}{a+b}) + \frac{c+d}{n}(1 - \frac{d}{c+d}) \\ &= \frac{a+b}{n}(\frac{c}{a+b}) + \frac{c+d}{n}(\frac{c}{c+d}) \\ &= \frac{c+d}{n} \end{aligned}$$

   Nous retrouvons l'expression du taux d'erreur issu de la matrice de confusion ci-dessus.

2. L'échantillon n'est pas représentatif mais nous disposons par ailleurs de la vraie valeur de p (connaissances du domaines, études précédentes, etc.). Nous formons

$$\epsilon = p \times (1 - \frac{a}{a+b}) + (1 - p) \times (1 - \frac{d}{c+d})$$

   Les estimations de la sensibilité et de la spécificité à partir de la matrice de confusion restent valables parce que ce sont des "profils lignes" du tableau, ils ne dépendent pas de la proportion des positifs et négatifs dans le fichier.

Taux de succès

Le taux de succès θ est le complémentaire à 1 du taux d'erreur, nous pouvons naturellement l'écrire en fonction de S_e et S_p

$$\begin{aligned}\theta &= 1 - \epsilon \\ &= 1 - [p \times (1 - S_e) + (1 - p) \times (1 - S_p)] \\ &= p \times S_e + (1 - p) \times S_p\end{aligned}$$

Précision

Toujours en partant de la définition probabiliste, la précision (valeur prédictive positive) peut s'écrire

$$VPP = \frac{p \times S_e}{p \times S_e + (1 - p) \times (1 - S_p)}$$

Indice de Youden

L'indice de Youden est bien connue en biostatistique, moins en apprentissage supervisé. Il s'écrit

$$IY = S_e + S_p - 1 \tag{2.1}$$

Son mérite est de caractériser le classifieur selon la sensibilité et la spécificité. Il prend la valeur maximum 1 lorsque le modèle est parfait. En effet, dans ce cas $S_e = 1$ et $S_p = 1$. Il peut être utilisé pour comparer les performances de plusieurs modèles.

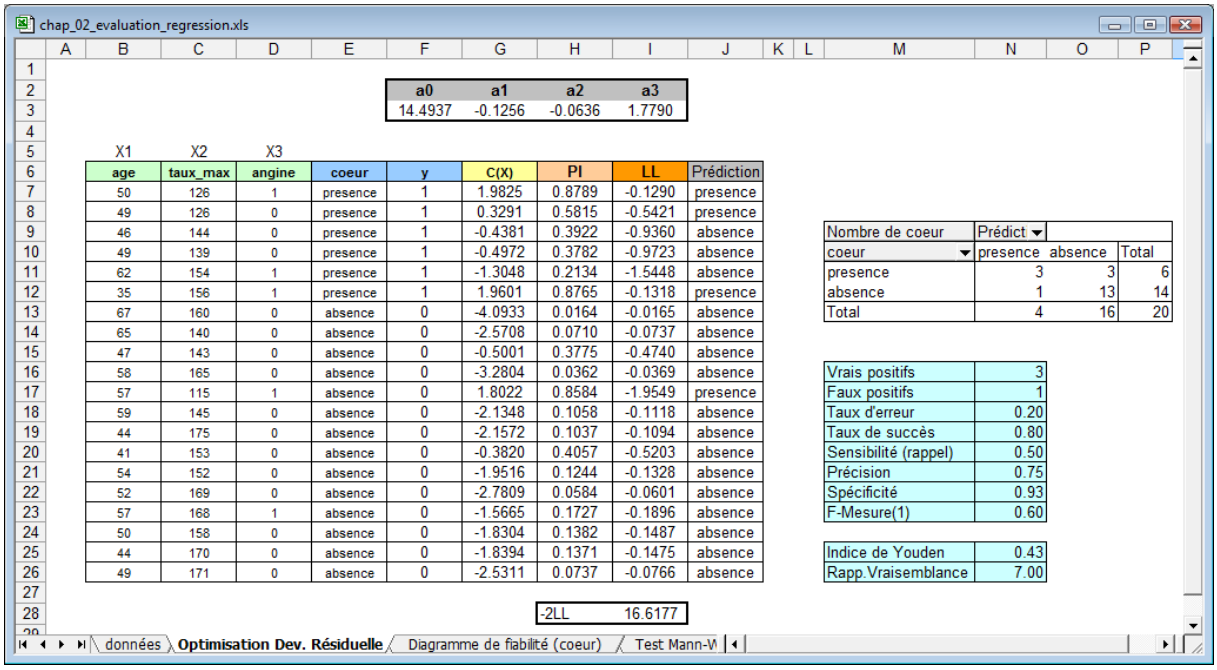
Son interprétation n'est pas très évidente en revanche. C'est le principal frein à son utilisation.

Rapport de vraisemblance

Le rapport de vraisemblance décrit le surcroît de chances des positifs (par rapport aux négatifs) d'être classés positifs. Sa définition est la suivante :

$$\begin{aligned}L &= \frac{P(\hat{Y} = +/Y = +)}{P(\hat{Y} = +/Y = -)} \\ &= \frac{P(\hat{Y} = +/Y = +)}{1 - P(\hat{Y} = -/Y = -)} \\ &= \frac{S_e}{1 - S_p}\end{aligned}$$

Le rapport de vraisemblance ne dépend pas de la proportion des positifs. Il donne donc des indications valables même si l'échantillon n'est pas représentatif. Plus grande est sa valeur, meilleur sera le modèle.


chap_02_evaluation_regression.xls

a0	a1	a2	a3
14.4937	-0.1256	-0.0636	1.7790

	X1	X2	X3						
	age	taux_max	angine	coeur	y	C(X)	PI	LL	Prédiction
1	50	126	1	presence	1	1.9825	0.8789	-0.1290	presence
2	49	126	0	presence	1	0.3291	0.5815	-0.5421	presence
3	46	144	0	presence	1	-0.4381	0.3922	-0.9360	absence
4	49	139	0	presence	1	-0.4972	0.3782	-0.9723	absence
5	62	154	1	presence	1	-1.3048	0.2134	-1.5448	absence
6	35	156	1	presence	1	1.9601	0.8765	-0.1318	presence
7	67	160	0	absence	0	-4.0933	0.0164	-0.0165	absence
8	65	140	0	absence	0	-2.5708	0.0710	-0.0737	absence
9	47	143	0	absence	0	-0.5001	0.3775	-0.4740	absence
10	58	165	0	absence	0	-3.2804	0.0362	-0.0369	absence
11	57	115	1	absence	0	1.8022	0.8584	-1.9549	presence
12	59	145	0	absence	0	-2.1348	0.1058	-0.1118	absence
13	44	175	0	absence	0	-2.1572	0.1037	-0.1094	absence
14	41	153	0	absence	0	-0.3820	0.4057	-0.5203	absence
15	54	152	0	absence	0	-1.9516	0.1244	-0.1328	absence
16	52	169	0	absence	0	-2.7809	0.0584	-0.0601	absence
17	57	168	1	absence	0	-1.5665	0.1727	-0.1896	absence
18	50	158	0	absence	0	-1.8304	0.1382	-0.1487	absence
19	44	170	0	absence	0	-1.8394	0.1371	-0.1475	absence
20	49	171	0	absence	0	-2.5311	0.0737	-0.0766	absence

-2LL	16.6177

Nombre de coeur	Prédict		
coeur	presence	absence	Total
presence	3	3	6
absence	1	13	14
Total	4	16	20

Vrais positifs	3
Faux positifs	1
Taux d'erreur	0.20
Taux de succès	0.80
Sensibilité (rappel)	0.50
Précision	0.75
Spécificité	0.93
F-Mesure(1)	0.60

Indice de Youden	0.43
Rapp.Vraisemblance	7.00

données | Optimisation Dev. Résiduelle | Diagramme de fiabilité (coeur) | Test Mann-W

Fig. 2.1. COEUR - Matrice de Confusion

2.1.3 Exemple : $coeur = f(age, taux\ max, angine)$

Voyons ce qu'il en est sur notre fichier COEUR. Nous avons complété la feuille de calcul en lui adjoignant la colonne "Prédiction" (Figure 2.1). La règle de classement utilisée est la suivante

$$\text{Si } \hat{\pi}(\omega) > 0.5 \text{ alors } \hat{Y}(\omega) = +$$

De manière complètement équivalente, nous pouvions nous baser sur le LOGIT avec

$$\text{Si } C(\omega) > 0 \text{ alors } \hat{Y}(\omega) = +$$

Nous pouvons former la matrice de confusion en confrontant les colonnes "Coeur" et "Prédiction". Nous en déduisons les principaux indicateurs d'évaluation des classifieurs :

- Taux d'erreur $= \frac{1+3}{20} = 0.20$
- Taux de succès $= \frac{3+13}{20} = 0.80$
- Sensibilité = Rappel $= \frac{3}{6} = 0.50$
- Précision $= \frac{3}{4} = 0.75$
- Spécificité $= \frac{13}{14} = 0.93$ (0.92857143)
- F-Mesure $= F_{\beta=1} = \frac{(1+1^2)\times 0.5\times 0.75}{1^2\times 0.75+0.5} = 0.60$
- Indice de Youden $= 0.5 + 0.93 - 1 = 0.43$
- Rapport de vraisemblance $= \frac{0.5}{1-0.92857143} = 7$

En termes de performances, nous constatons que le modèle issu de la régression logistique semble (pourquoi cette prudence ? nous verrons pourquoi plus loin, section 2.1.5) meilleur que le précédent basé sur les probabilité conditionnelles $P(COEUR/ANGINE)$ qui présentait un taux d'erreur égal à 0.25 (Figure 1.2).

2.1.4 Le modèle est-il "intéressant" ?

Une de mes questions favorites en cours, après avoir présenté ces concepts, est la suivante : "j'ai un taux d'erreur de 0.20, c'est bien ou c'est pas bien ?". Généralement, un silence gêné s'installe avant que ne vienne des réponses plus ou moins loufoques. Puis arrive la bonne réponse qui est en réalité une question "à quoi peut-on comparer cette valeur ?".

Nous avions déjà abordé cette idée lors de la présentation des pseudo-R^2 (section 1.6). Nous avions obtenu une réponse claire, l'élément de comparaison est la déviance du modèle ne comportant que la constante a_0. Il faut généraliser l'approche en sortant du seul cadre de la régression logistique, et proposer un indicateur qui confronte des taux d'erreur.

Le modèle par défaut est défini comme un modèle qui n'utilise pas les informations en provenance des variables explicatives. Si l'on s'en tient au cadre bayesien (section 1.1.3), la règle d'affectation devient

$$y_{k^*} = \arg\max_k P[Y(\omega) = y_k]$$

La règle de décision du classifieur par défaut est donc très simple : on affecte, pour tout individu à classer, la modalité majoritaire dans l'échantillon d'apprentissage.

Pour le fichier COEUR, sachant que la proportion des "présence (+)" est $\frac{6}{20} = 0.3$, celle des "absence (-)" $\frac{14}{20}$. En l'absence de toute information en provenance de variables explicatives, nous avons intérêt à affecter systématiquement la conclusion "absence" à tous les individus que l'on souhaite classer.

La matrice de confusion du classifieur par défaut est facile à construire

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	Total
$+$	0	6	6
$-$	0	14	14
Total	0	20	20

Et le taux d'erreur associé est

$$\epsilon_{(def)} = \frac{6}{20} = 0.3$$

Pseudo-R^2 basé sur le taux d'erreur

Maintenant, nous avons un élément de référence. Le taux d'erreur de la régression logistique est $\epsilon_{(M)} = \frac{4}{20} = 0.2$; celui du modèle par défaut est $\epsilon_{(def)} = \frac{6}{20} = 0.3$.

Nous pouvons en dégager une sorte de pseudo-R^2 qui s'écrit

$$R^2_\epsilon = 1 - \frac{\epsilon_{(M)}}{\epsilon_{(def)}}$$

Si notre modèle est parfait, avec un taux d'erreur nul, nous obtenons $R^2_\epsilon = 1$; si notre modèle ne sait pas faire mieux que le classifieur par défaut, nous avons $R^2_\epsilon = 0$.

Elle est notée λ_p à cause de sa similitude avec le λ de Goodman et Kruskal (1954) - une mesure d'association pour les tableaux de contingence - dans certains ouvrages [10] (page 32). Son inconvénient est qu'elle peut prendre des valeurs négatives lorsque le modèle étudié est moins bon que le modèle par défaut. Cette configuration arrive principalement lorsque les classes sont très déséquilibrées dans le fichier de données. Le taux d'erreur du classifieur par défaut est d'office très faible, il est difficile de faire mieux. C'est une des critiques que l'on adresse à la matrice de confusion en tant qu'outil d'évaluation d'ailleurs. Pour nous, ce n'est pas rédhibitoire. Il faut en être conscient simplement et ne pas pousser des hauts cris parce qu'on obtient quelque chose que l'on désigne par R^2 et qui s'avère être négatif.

Pour le fichier COEUR, le pseudo-R^2 est

$$R^2_\epsilon = 1 - \frac{0.2}{0.3} = 1 - 0.67 = 0.33$$

La régression logistique fait mieux que le classifieur par défaut.

Un test de comparaison des taux d'erreur

Le modèle M produit par la régression logistique semble faire mieux si l'on compare son taux d'erreur avec celui du classifieur par défaut. Mais est-ce réellement significatif ? Est-ce que l'écart va au-delà des simples fluctuations d'échantillonnage ?

Nous avons deux proportions à comparer. L'hypothèse nulle est "notre modèle ne fait pas mieux que le classifieur par défaut" en termes de probabilité de mauvais classement ; l'hypothèse alternative est "notre modèle est meilleur" (probabilité de mal classer plus faible).

Il est hors de question d'utiliser le test usuel car les échantillons ne sont pas indépendants. La piste serait plutôt du côté de la comparaison à un standard (taux d'erreur du classifieur par défaut), sachant que ce dernier a lui aussi été mesuré sur le fichier. Bulmer (1979) [10] (page 34) propose la statistique suivante pour répondre à notre question

$$d = \frac{\epsilon_{(def)} - \epsilon_{(M)}}{\sqrt{\frac{1}{n}\epsilon_{(def)}(1 - \epsilon_{(def)})}} \tag{2.2}$$

Elle suit une loi binomiale, mais elle se rapproche très rapidement de loi normale centrée réduite dès que n augmente (dès que $n \times \epsilon_{(def)} \times (1 - \epsilon_{(def)}) > 9$).

La région critique du test, rejet de l'hypothèse nulle, au risque α pour le test unilatéral s'écrit

$$R.C. : d > u_{1-\alpha}$$

Où $u_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi normale centrée est réduite.

Voyons ce qu'il en est pour notre exemple COEUR, la formation de la statistique ne pose aucun problème, nous avons

$$d = \frac{0.3 - 0.2}{\sqrt{\frac{1}{20}0.3(1 - 0.3)}} = \frac{0.1}{\sqrt{0.0105}} = \frac{0.1}{0.1025} = 0.9759$$

La régression logistique ne semble pas si bonne que cela finalement ?

Méfiance. Ce test a le mérite d'exister, mais c'est bien son seul mérite. En effet, on se rend compte à l'usage que la moindre différence entre les taux d'erreur est quasi-systématiquement entérinée pour peu que la taille du fichier dépasse la centaine d'observations (dans le *data mining*, on traite souvent des fichiers avec plusieurs milliers voire centaines de milliers d'observations !). Elle est systématiquement réfutée sur des petits échantillons (notre fichier COEUR). Ses indications sont finalement très peu utilisables. Mieux vaut s'en tenir à des indicateurs simples tel que le pseudo-R^2 qui donne avant tout un ordre d'idées sur la pertinence du modèle.

2.1.5 Subdivision "apprentissage - test" des données pour une évaluation plus fiable

Le modèle issu de la régression logistique avec les 3 variables ($\epsilon_{resub} = 0.2$) semble meilleur que celui basé uniquement sur "angine" ($\epsilon_{resub} = 0.25$) si l'on se réfère au taux d'erreur en resubstitution. Faut-il s'en tenir à cela ?

Non, car nous utilisons les mêmes données pour construire le modèle et pour l'évaluer. Or, dans ce contexte, les classifieurs plus complexes ayant tendance à "coller" aux données laissent à penser, à tort, qu'ils présentent de meilleures performances. En règle générale, plus une observation pèse sur son propre classement en généralisation, plus optimiste sera le taux d'erreur en resubstitution. Bref, **le taux d'erreur en resubstitution est totalement inutilisable dès lors que l'on souhaite comparer les performances de modèles de complexité différente (ou reposant sur des représentations différentes ex. arbre de décision vs. régression logistique).**

Parmi les solutions envisageables, la plus simple consiste à évaluer le classifieur sur des données à part qui n'ont pas participé au processus d'apprentissage. Nous procédons de la manière suivante lorsque l'on dispose d'un échantillon Ω de taille n :

1. Nous tirons au hasard n_a individus parmi n, il s'agit de *l'échantillon d'apprentissage*, nous les utilisons pour construire le modèle de prédiction M_a. On dédie généralement 70% des données à l'apprentissage. Mais ce n'est pas aussi simple, nous en rediscuterons plus loin.

2. Sur les n_t observations restantes, *l'échantillon test*, nous appliquons le modèle M_a, et nous élaborons la matrice de confusion en confrontant les valeurs observées et les valeurs prédites. Habituellement, $\frac{n_t}{n} = 1 - \frac{n_a}{n} = 30\%$.

Principal atout de cette approche, les indicateurs ainsi obtenus sont non-biaisés. Ils permettent de comparer les mérites respectifs de plusieurs modèles, même s'ils sont de complexité différente, même s'ils ne reposent pas sur des systèmes de représentation identiques (ex. un classifieur linéaire vs. un classifieur non linéaire). C'est la démarche à privilégier si l'on dispose de suffisamment d'observations.

Et c'est bien là le principal défaut de cette démarche. Lorsque nous travaillons sur un petit échantillon, en réserver une partie pour l'évaluation pénalise la construction du modèle, sans pour autant que l'on ait une évaluation fiable des performances puisque l'effectif est trop faible. Nous sommes face à 2 exigences contradictoires :

- Réserver une grande partie des données à l'apprentissage favorise la construction d'un modèle de bonne qualité. En revanche, l'échantillon test sera trop réduit pour espérer obtenir une estimation viable des performances en prédiction.
- Réserver une fraction plus forte au test permet certes d'obtenir une évaluation fiable. Mais dans ce cas nous nous tirons une balle dans le pied (aïe!) car le modèle élaboré peut être dégradé faute d'informations (d'observations) suffisantes.

Bref, les proportions habituellement mises en avant (70% vs. 30%) ne doivent pas être prises au pied de la lettre. Tout est affaire de compromis : il en faut suffisamment pour l'apprentissage afin de produire un modèle consistant ; il en faut suffisamment pour le test afin d'obtenir une évaluation fiable des performances. Les "bonnes" proportions dépendent souvent des caractéristiques du classifieur et des données analysées (rapport entre le nombre d'observations et le nombre de variables, degré de difficulté du concept à apprendre, etc.).

Remarque : A propos des méthodes de ré-échantillonnage. Lorsque les effectifs sont très faibles, nous avons intérêt à construire le modèle M sur la totalité des données, puis à utiliser des techniques de ré-échantillonnage pour en mesurer les performances (ex. la validation croisée, le bootstrap). L'intérêt est double. Nous utilisons la totalité des données (la totalité de l'information disponible) pour construire le classifieur. Et nous pouvons obtenir une évaluation (plus ou moins) faiblement biaisée de son erreur de prédiction [17].

2.1.6 Inconvénients de la matrice de confusion

Pour intéressante qu'elle soit, elle est très utilisée en apprentissage supervisé, la matrice de confusion présente une faiblesse importante : elle repose essentiellement sur les prédictions $\hat{y}(\omega)$, sans tenir compte des probabilités estimées $\hat{\pi}(\omega)$. Se baser uniquement sur les prédictions est un peu réducteur. En effet, un individu avec $\hat{\pi}(\omega) = 0.495$ sera désigné "négatif", un autre avec $\hat{\pi}(\omega') = 0.505$ sera désigné "positif". Pourtant, si l'on se réfère aux probabilités, ils sont finalement assez proches. La matrice de confusion ne nous rapporte pas ce type d'information [9] (pages 156 à 160).

Autre écueil auquel sont confrontés la matrice de confusion et le taux d'erreur qui en est dérivé, ils sont sensibles à l'importance relative des groupes c.-à-d. la proportion des "positifs" et "négatifs" dans

le fichier. Le classement dans le groupe le plus important est toujours favorisé. Par exemple, si nous avons 99% de positifs, nous avons intérêt à classer systématiquement les observations dans cette classe, nous avons la garantie que le taux d'erreur sera égal à 1%. On pourrait penser alors que construire un classifieur dans ce contexte ne sert à rien.

2.2 Diagramme de fiabilité

2.2.1 Calcul et interprétation du diagramme de fiabilité

Contrairement à certaines méthodes supervisées (ex. support vector machine, classifieur bayesien naïf), la régression logistique produit une bonne approximation de la quantité $\pi(\omega)$. La première idée qui vient à l'esprit est de confronter les probabilités estimées par le modèle et celles observées dans le fichier de données. On construit pour cela le diagramme de fiabilité (en anglais *reliability diagram*) [16].

Ici également, si nous en avons la possibilité, nous avons tout intérêt à construire le diagramme à partir des données tests n'ayant pas participé à l'élaboration du classifieur. Les indications obtenues n'en seront que plus crédibles.

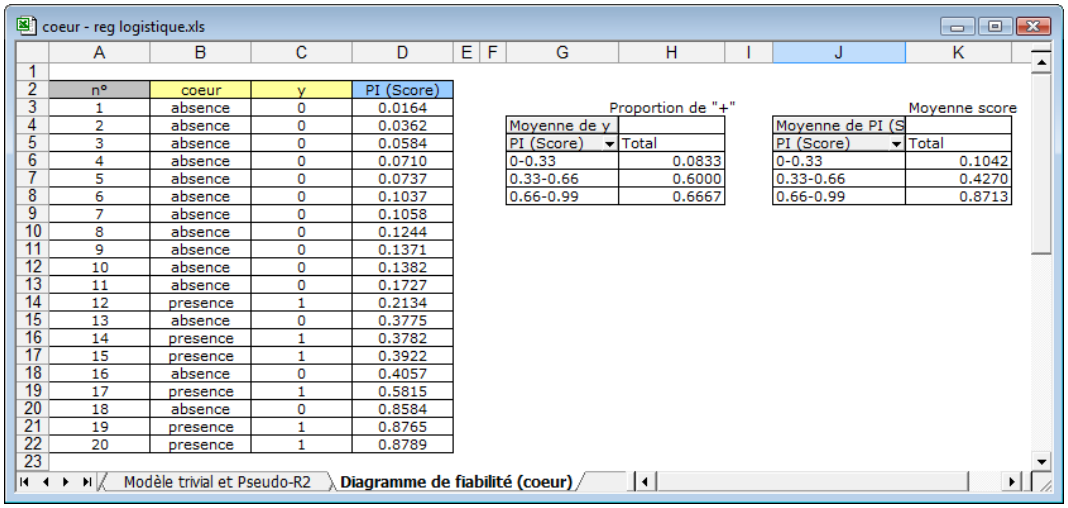
Voici les principales étapes de la construction du diagramme de fiabilité :

1. Appliquer le classifieur sur les données pour obtenir le score $\hat{\pi}(\omega)$.
2. Trier le fichier selon le score croissant.
3. Sur la base du score, subdiviser les données en intervalles (ex. 0.0-0.2, 0.2-0.4, etc.).
4. Dans chaque intervalle, calculer la proportion de positifs.
5. Dans le même temps, toujours dans chaque intervalle, calculer la moyenne des scores.
6. Si les chiffres concordent dans chaque intervalle, les scores sont bien calibrés, le classifieur est de bonne qualité.
7. Nous pouvons résumer l'information dans un graphique *nuage de points* appelé **diagramme de fiabilité**, avec en abcisse la moyenne des scores, en ordonnée la proportion de "positifs".
8. **Si les scores sont bien calibrés, les points devraient être alignés sur une droite, la première bissectrice**.
9. Les points s'écartant sensiblement de la première bissectrice doivent attirer notre attention.

2.2.2 Exemple : COEUR = f(age, taux max, angine)

Nous reprenons notre exemple de détection des problèmes cardiaques (Figure 0.1). L'effectif étant très faible, $n = 20$, nous réaliserons un découpage en 3 groupes selon le score, avec les intervalles 0.00-0.33, 0.34-0.66, 0.67-1.00.

Nous appliquons à la lettre la démarche ci-dessus, nous obtenons une nouvelle feuille de calcul sous Excel (Figure 2.2) :



coeur - reg logistique.xls

n°	coeur	y	PI (Score)
1	absence	0	0.0164
2	absence	0	0.0362
3	absence	0	0.0584
4	absence	0	0.0710
5	absence	0	0.0737
6	absence	0	0.1037
7	absence	0	0.1058
8	absence	0	0.1244
9	absence	0	0.1371
10	absence	0	0.1382
11	absence	0	0.1727
12	presence	1	0.2134
13	absence	0	0.3775
14	presence	1	0.3782
15	presence	1	0.3922
16	absence	0	0.4057
17	presence	1	0.5815
18	absence	0	0.8584
19	presence	1	0.8765
20	presence	1	0.8789

Proportion de "+"

Moyenne de y	
PI (Score)	Total
0-0.33	0.0833
0.33-0.66	0.6000
0.66-0.99	0.6667

Moyenne score

Moyenne de PI (S	
PI (Score)	Total
0-0.33	0.1042
0.33-0.66	0.4270
0.66-0.99	0.8713

Modèle trivial et Pseudo-R2 | Diagramme de fiabilité (coeur)

Fig. 2.2. COEUR - Calcul du Diagramme de fiabilité

- Le tableau a été trié selon un le score croissant.
- Dans le 1^{er} groupe, avec un score variant entre 0.00 et 0.33 c.-à-d. de l'observation n^o1 au n^o12, la proportion de "+" est égale à $\frac{1}{12} = 0.0833\%$. Dans le même temps, la moyenne des scores est égale à $\frac{0.0164+0.0362+\cdots+0.2134}{12} = 0.1042$. Nous obtenons le premier point du graphique.
- Nous faisons de même pour les autres groupes, nous obtenons le diagramme de fiabilité (Figure 2.3).

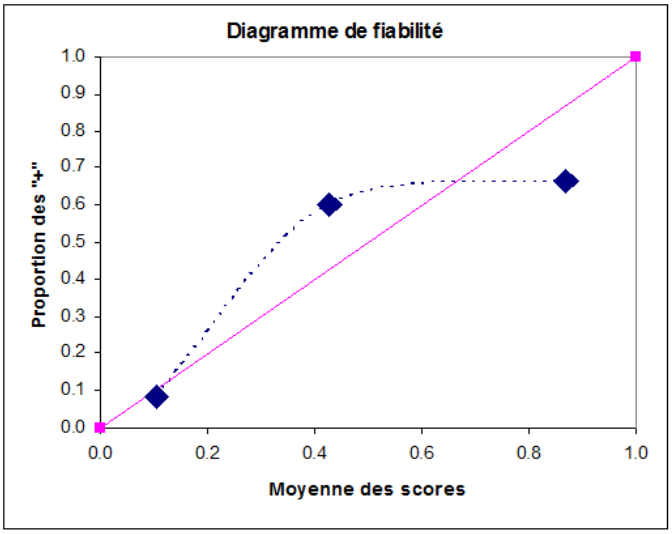


Fig. 2.3. COEUR - Diagramme de fiabilité

Manifestement, il y a un problème dans notre régression. Les points ne sont pas alignés du tout. Mais on ne devrait pas trop s'en étonner. Les effectifs sont tellement faibles ($n = 20$) qu'il pouvait difficilement

en être autrement. Les résultats sont de mauvaise qualité. Le classifieur est certainement très instable de surcroît, ajouter ou retirer une observation peut le modifier fortement.

2.2.3 Exemple : Acceptation de crédit

Penchons-nous sur des données un peu plus réalistes pour montrer l'intérêt de cette procédure. Dans le problème qui suit, nous souhaitons expliquer l'accord d'un prêt par un organisme de crédit à partir l'âge du référant, le revenu par tête dans le ménage, le fait d'être propriétaire de son habitation ou non, occuper une profession indépendante ou non, le nombre de problèmes rencontrés avec sa banque. Nous disposons de $n = 100$ observations, avec $n_+ = 73$ positifs.

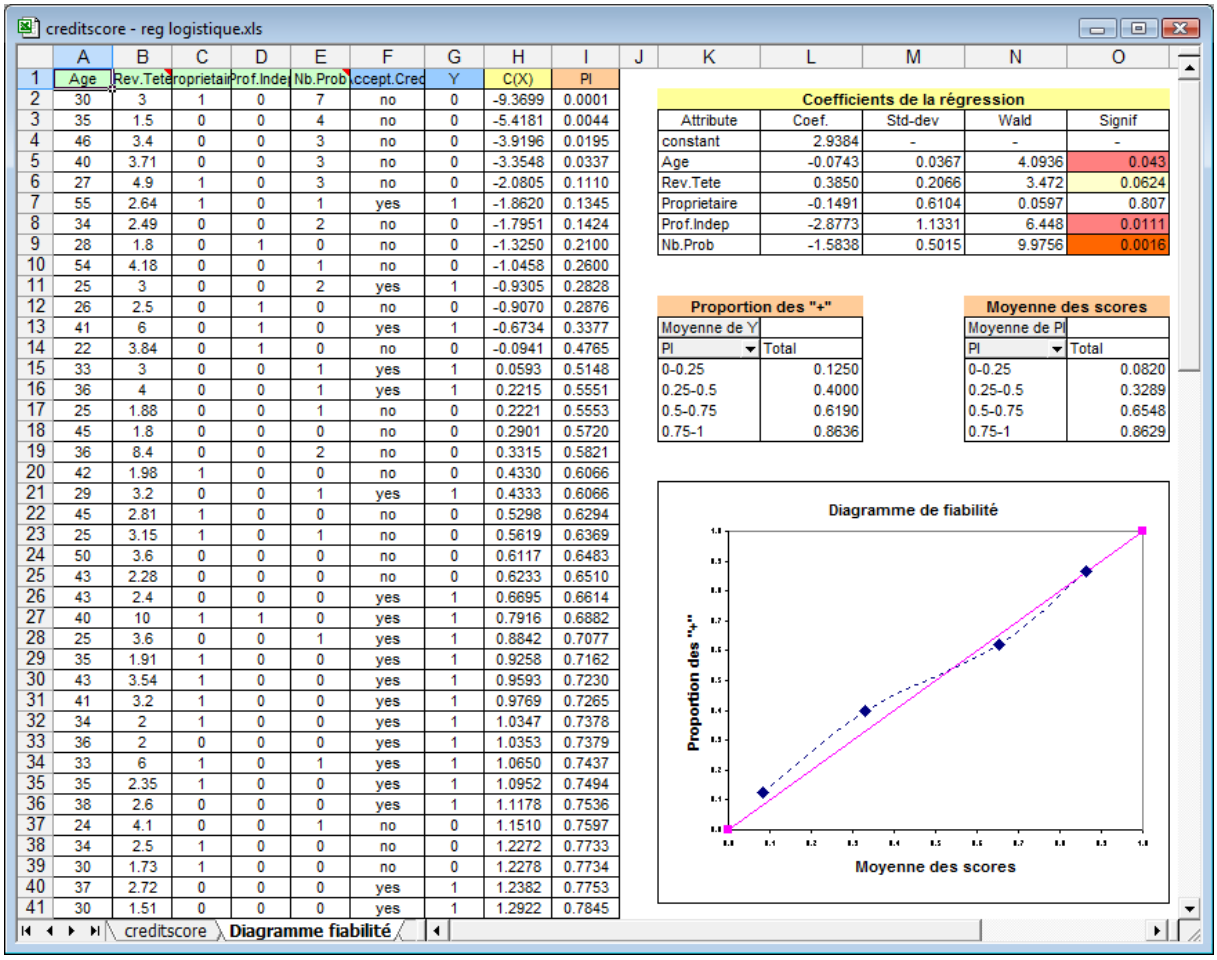


Fig. 2.4. CREDIT - Diagramme de fiabilité

De nouveau, nous reproduisons les étapes permettant d'obtenir le diagramme de fiabilité (Figure 2.4, nous ne visualisons que les 40 premières observations ici) :

– Nous avons estimé les paramètres du modèle à l'aide de Tanagra, nous obtenons les valeurs suivantes (Remarque : nous ignorons pour l'instant les autres informations)

Variable	Coefficient
Constante	2.9384
Age	−0.0734
Rev.Tete	0.3850
Propriétaire	−0.1491
Prof.Indep	−2.8773
Nb.Prob	−1.5838

– Nous calculons alors le LOGIT pour chaque individu. Pour la première observation, nous avons

$$C(1) = 2.9384 - 0.0743 \times 30 + 0.3850 \times 3 - 0.1491 \times 1 - 2.8773 \times 0 - 1.5838 \times 7 = -9.3699$$

– Nous en déduisons le score

$$\hat{\pi}(1) = \frac{1}{1 + e^{-(-9.3699)}} = 0.0001$$

– Une fois calculé tous les scores, et le tableau trié, nous décidons de procéder à un découpage en 4 intervalles, définies par $0.00 - 0.25$, $0.26 - 0.50$, etc [2].

– Dans chaque intervalle nous comptabilisons la proportion de positifs et, dans le même temps, nous calculons la moyenne des scores (nous avons utilisé les tableaux croisés dynamiques pour cela).

– Il ne reste plus qu'à produire le diagramme de fiabilité. Concernant le fichier CREDIT, nous constatons que le modèle produit une bonne estimation des quantités $\pi(\omega)$, les points sont quasiment alignés sur une droite.

2.3 Test de Hosmer-Lemeshow

2.3.1 Construction du test de Hosmer-Lemeshow

Le test de Hosmer-Lemeshow [9] (pages 147 à 156; des variantes sont proposées) relève à peu près de la même logique que le diagramme de fiabilité. A la différence qu'au lieu de se baser simplement sur une impression visuelle, on extrait du tableau de calcul un indicateur statistique qui permet de quantifier la qualité des estimations $\hat{\pi}(\omega)$.

Concrètement, nous procédons de la manière suivante :

1. Appliquer le classifieur sur les données pour obtenir les estimations $\hat{\pi}(\omega)$ (score).
2. Trier les données selon le score croisant.
3. Subdivisez les données en G groupes en se basant sur les quantiles (ex. les quantiles d'ordre 4 correspondent aux quartiles, les quantiles d'ordre 10 aux déciles, etc.). Les auteurs proposent prioritairement les déciles ($G = 10$). Il semble par ailleurs plus judicieux d'utiliser les quantiles plutôt que les seuils sur les scores comme cela a été fait pour le diagramme de fiabilité. L'approximation de la loi de distribution de la statistique du test sous H_0 est de meilleure qualité [9] (page 149).

2. Attention, le nombre d'intervalles est déterminant dans cette procédure. Nous avons toujours intérêt à fixer un nombre assez faible de manière à obtenir un bon "lissage" de la courbe. S'il est trop élevé, la courbe devient chaotique, très peu utilisable et laissant à penser que les classifieurs sont toujours de mauvaise qualité

4. Dans chaque groupe g, d'effectif m_g, nous devons calculer plusieurs quantités :
   - m_{g1}, le nombre de positifs observés ;
   - m_{g0}, le nombre de négatifs observés ;
   - $\hat{m}_{g1} = \sum_{\omega \in g} \hat{\pi}(\omega)$, la somme des scores des observations situées dans le groupe g. On la désigne comme la fréquence théorique des positifs dans le groupe ;
   - $\bar{\pi}_{g1} = \frac{\hat{m}_{g1}}{m_g}$, la moyenne des scores observés dans le groupe g ;
   - $\hat{m}_{g0} = m_g - \hat{m}_{g1}$, la fréquence théorique des négatifs.
5. Nous calculons alors la statistique de Hosmer et Lemeshow en utilisant une des formules suivantes ([23], page 407 ; [9], page 148)

$$\hat{C} = \sum_g \left[\frac{(m_{g1} - \hat{m}_{g1})^2}{\hat{m}_{g1}} + \frac{(m_{g0} - \hat{m}_{g0})^2}{\hat{m}_{g0}} \right] \tag{2.3}$$

$$= \sum_g \left[\frac{m_g (m_{g1} - \hat{m}_{g1})^2}{\hat{m}_{g1}(m_g - \hat{m}_{g1})} \right] \tag{2.4}$$

$$= \sum_g \frac{(m_{g1} - \hat{m}_{g1})^2}{\hat{m}_{g1}(1 - \bar{\pi}_{g1})} \tag{2.5}$$

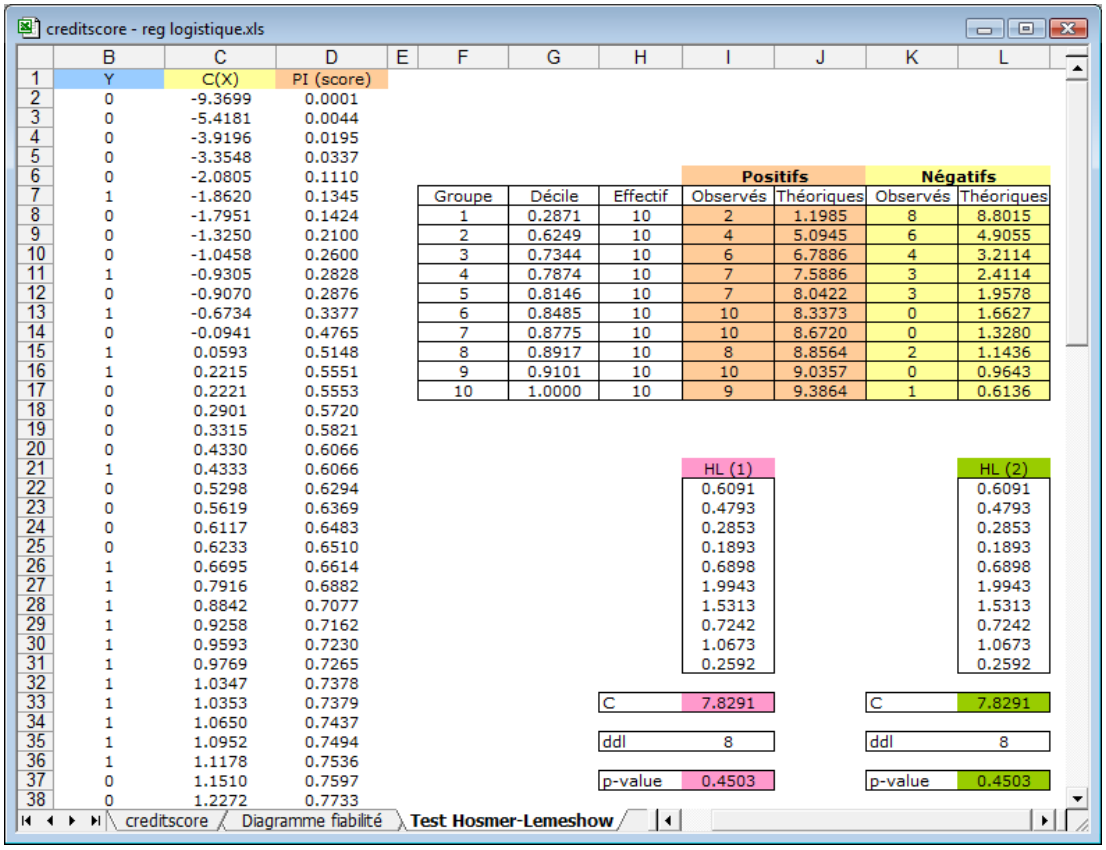
6. Lorsque le modèle est correct (H_0), la statistique $\hat{C}$ suit approximativement une loi du χ^2 à $(G-2)$ degrés de liberté.
7. Lorsque la probabilité critique du test (p-value) est plus grand que le risque choisi, le modèle issu de la régression logistique est accepté.
8. Les réserves usuelles concernant ce type de test restent de mise ici. Il faudrait entres autres que tous les effectifs théoriques soient supérieurs à 5 dans toutes les cases du tableau. Si ce n'est pas le cas, on devrait procéder à des regroupements et corriger en conséquence les degrés de liberté. Mais il ne faut pas non plus s'arc-bouter à cette idée. Il s'agit d'un outil d'évaluation du classifieur, il donne avant tout une indication sur la qualité des $\hat{\pi}(\omega)$ [9] (page 150).
9. Enfin, au delà de la statistique elle-même, l'étude du tableau de calcul, en particulier la détection des situations où les effectifs observés et théoriques sont fortement dissemblables, donnent des indications précieuses sur le comportement du classifieur [9] (page 151). Nous nous rapprochons en cela à une étude qualitative déjà mise en avant lors de la présentation du diagramme de fiabilité.

Remarque : Hosmer et Lemeshow sur un échantillon test. Tout comme pour la matrice de confusion, nous pouvons subdiviser les données en 2 parties : la première pour construire le modèle, la seconde pour l'évaluer. La procédure de Hosmer et Lemeshow peut être élaborée sur ce second échantillon. La statistique de test reste identique, les degrés de liberté en revanche sont modifiés puisqu'aucun paramètre n'a été estimé sur ces données (voir [9], pages 186 à 188 ; d'autres statistiques sont proposées, toujours dans le contexte d'une évaluation sur un échantillon test).

2.3.2 Acceptation de crédit - Test de Hosmer-Lemeshow

Nous travaillons sur le fichier "Acceptation de crédit" (section 2.2.3). Il comporte suffisamment d'observations $n = 100$ pour que la subdivision en $G = 10$ groupes ne pose pas trop de problèmes.

Nous avons déjà obtenu précédemment, lors de l'étude du diagramme de fiabilité, la colonne de score et trié le fichier (Figure 2.4). Il ne nous reste plus qu'à constituer les groupes en nous basant sur les déciles ($G = 10$). Nous devrions obtenir les mêmes effectifs $m_g = \frac{100}{10} = 10$ dans chaque groupe (à peu près, tout dépend s'il y a des ex-aequo ou non).



	B	C	D
1	Y	C(X)	PI (score)
2	0	-9.3699	0.0001
3	0	-5.4181	0.0044
4	0	-3.9196	0.0195
5	0	-3.3548	0.0337
6	0	-2.0805	0.1110
7	1	-1.8620	0.1345
8	0	-1.7951	0.1424
9	0	-1.3250	0.2100
10	0	-1.0458	0.2600
11	1	-0.9305	0.2828
12	0	-0.9070	0.2876
13	1	-0.6734	0.3377
14	0	-0.0941	0.4765
15	1	0.0593	0.5148
16	1	0.2215	0.5551
17	0	0.2221	0.5553
18	0	0.2901	0.5720
19	0	0.3315	0.5821
20	0	0.4330	0.6066
21	1	0.4333	0.6066
22	0	0.5298	0.6294
23	0	0.5619	0.6369
24	0	0.6117	0.6483
25	0	0.6233	0.6510
26	1	0.6695	0.6614
27	1	0.7916	0.6882
28	1	0.8842	0.7077
29	1	0.9258	0.7162
30	1	0.9593	0.7230
31	1	0.9769	0.7265
32	1	1.0347	0.7378
33	1	1.0353	0.7379
34	1	1.0650	0.7437
35	1	1.0952	0.7494
36	1	1.1178	0.7536
37	0	1.1510	0.7597
38	0	1.2272	0.7733

			Positifs		Négatifs	
Groupe	Décile	Effectif	Observés	Théoriques	Observés	Théoriques
1	0.2871	10	2	1.1985	8	8.8015
2	0.6249	10	4	5.0945	6	4.9055
3	0.7344	10	6	6.7886	4	3.2114
4	0.7874	10	7	7.5886	3	2.4114
5	0.8146	10	7	8.0422	3	1.9578
6	0.8485	10	10	8.3373	0	1.6627
7	0.8775	10	10	8.6720	0	1.3280
8	0.8917	10	8	8.8564	2	1.1436
9	0.9101	10	10	9.0357	0	0.9643
10	1.0000	10	9	9.3864	1	0.6136

HL (1)	HL (2)
0.6091	0.6091
0.4793	0.4793
0.2853	0.2853
0.1893	0.1893
0.6898	0.6898
1.9943	1.9943
1.5313	1.5313
0.7242	0.7242
1.0673	1.0673
0.2592	0.2592

C	7.8291	C	7.8291
ddl	8	ddl	8
p-value	0.4503	p-value	0.4503

Fig. 2.5. CREDIT - Test de Hosmer et Lemeshow

La feuille de calcul est construite comme suit (Figure 2.5, l'affichage est limité aux 37 premières observations) :

- Tout d'abord, nous calculons les déciles. Le 1^{er} décile est égal à 0.271, le 2^{nd} à 0.6249.
- Nous vérifions le nombre d'observations dans chaque groupe, nous avons bien $m_g = 10$, $\forall g$ puisque $n = 100$.
- Dans chaque groupe, nous comptons le nombre de positifs et de négatifs. Pour le 1^{er} groupe par exemple, nous avons $m_{11} = 2$ et $m_{10} = 10 - 2 = 8$.
- Puis nous calculons les effectifs espérés en faisant la somme des scores dans le groupe. Pour le 1^{er} groupe, nous avons $\hat{m}_{11} = 0.0001 + 0.0044 + 0.0195 + \cdots + 0.2828 = 1.1985$. Nous en déduisons $\hat{m}_{10} = 10 - 1.1985 = 8.8015$.
- Il ne reste plus qu'à calculer la statistique de Hosmer et Lemeshow en utilisant une des formules ci-dessus. Pour la première, nous avons

$$\hat{C} = \left[\frac{(2-1.1985)^2}{1.1985} + \frac{(8-8.8015)^2}{8.8015}\right] + \cdots + \left[\frac{(9-9.3864)^2}{9.3864} + \frac{(1-0.6136)^2}{0.6136}\right] = 7.8291$$

Pour la seconde,

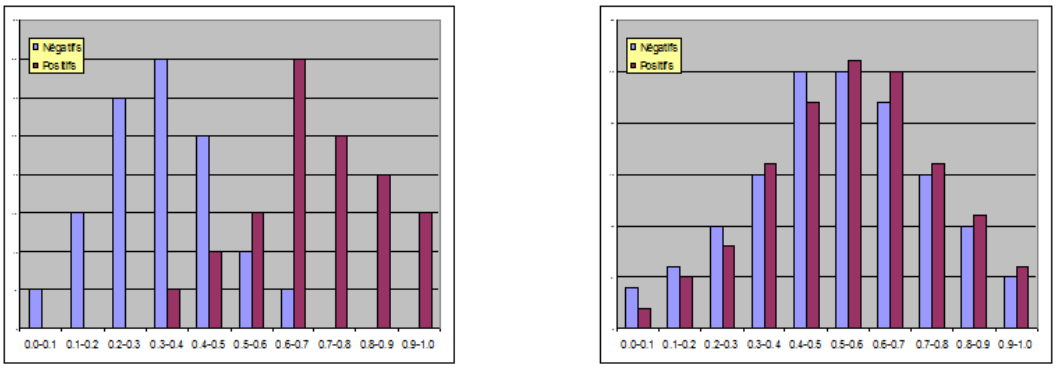
$$\hat{C} = \left[\frac{10(2-1.1985)^2}{1.1985(10-1.1985)}\right] + \cdots + \left[\frac{10(9-9.3864)^2}{9.3864(10-9.3864)}\right] = 7.8291$$

- Les degrés de liberté étant égales à $G - 2 = 10 - 2 = 8$, nous obtenons une p-value de 0.4503 avec la loi du χ^2.
- La p-value est supérieure au risque usuel de 5%. Le modèle est validé, il est compatible avec les données.

2.4 Le test de Mann-Whitney

2.4.1 Pourquoi un test de comparaison de populations ?

La discrimination sera d'autant meilleure que les positifs ont un score $\hat{\pi}(\omega)$ élevé et les négatifs un score faible. Dans les tableaux où l'on trie les observations selon un score croissant, les négatifs seraient agglutinés en haut, les positifs en bas. On peut illustrer ce point de vue en comparant les distributions des scores conditionnellement aux classes d'appartenance. Lorsque le modèle est de bonne qualité, les distributions conditionnelles des scores sont bien différenciées (Figure 2.6, A) ; dans le cas contraire, elles sont confondues (Figure 2.6, B).



[A] Bonne discrimination. Les « négatifs » ont un score faible, les « positifs » un score élevé. Les distributions conditionnelles sont décalées.

[B] Mauvaise discrimination. Les « négatifs » et les « positifs » ont le même profil de score. Les distributions conditionnelles sont confondues.

Fig. 2.6. Distributions types des scores conditionnellement aux classes

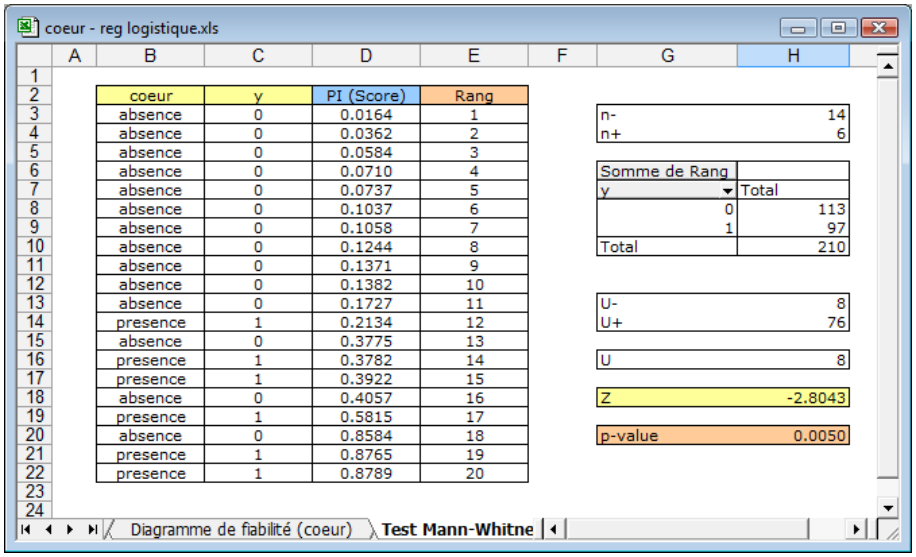
Il faut pouvoir quantifier cette impression visuelle. Pour ce faire, un test de comparaison de populations semble approprié. L'objectif est de répondre à la question : "est-ce que les positifs ont des scores (significativement) plus élevés que les négatifs ?". **Le test non paramétrique de Mann-Whitney** est celui que l'on retient le plus souvent dans la littérature. Pour différentes raisons [18] (page 34). Dans le cadre de l'apprentissage supervisé, il **convient surtout parce qu'il est en relation avec le critère**

AUC (Area Under Curve) associé à la courbe ROC que nous présenterons plus loin (section 2.5) [23] (pages 410-411). A défaut, nous aurions pu utiliser tout autre test permettant de caractériser un décalage entre les paramètres de localisation des distributions.

Rappelons brièvement les formules associées à ce test :

1. A partir des scores $\hat{\pi}(\omega)$, nous calculons le rang des $r(\omega)$ des individus dans l'ensemble de l'échantillon, sans distinction de classes.
2. Nous calculons alors les sommes conditionnelles de rangs, pour les positifs
$$r_+ = \sum_{\omega:y(\omega)=1} r(\omega)$$
et pour les négatifs
$$r_- = \sum_{\omega:y(\omega)=0} r(\omega)$$
3. Nous en déduisons les statistiques
$$U_+ = r_+ - \frac{n_+(n_+ + 1)}{2}$$
et
$$U_- = r_- - \frac{n_-(n_- + 1)}{2}$$
4. La statistique de Mann-Whitney correspond au minimum de ces deux quantités, soit
$$U = \min(U_+, U_-)$$
5. Sous H_0, les distributions sont confondues, la statistique centrée et réduite Z suit une loi normale $\mathcal{N}(0,1)$
$$Z = \frac{U - \frac{n_+ n_-}{2}}{\sqrt{\frac{1}{12}(n_+ + n_- + 1)n_+ n_-}}$$
6. Il s'agit usuellement d'un test bilatéral. Mais en vérité on imagine mal que les positifs puissent présenter des scores significativement plus faibles que les négatifs. Ou alors, il faudrait prendre le complémentaire à 1 des valeurs produites par le classifieur.

Deux types de corrections peuvent être introduites pour préciser les résultats dans certaines circonstances : une correction de continuité lorsque les effectifs sont faibles ; une correction du dénominateur de la statistique centrée et réduite lorsqu'il y a des ex-aequo, on utilise habituellement les rangs moyens [18] (pages 40 et 41-44).



coeur	y	PI (Score)	Rang
absence	0	0.0164	1
absence	0	0.0362	2
absence	0	0.0584	3
absence	0	0.0710	4
absence	0	0.0737	5
absence	0	0.1037	6
absence	0	0.1058	7
absence	0	0.1244	8
absence	0	0.1371	9
absence	0	0.1382	10
absence	0	0.1727	11
presence	1	0.2134	12
absence	0	0.3775	13
presence	1	0.3782	14
presence	1	0.3922	15
absence	0	0.4057	16
presence	1	0.5815	17
absence	0	0.8584	18
presence	1	0.8765	19
presence	1	0.8789	20

n-	14
n+	6

Somme de Rang	
y	Total
0	113
1	97
Total	210

U-	8
U+	76
U	8
Z	-2.8043
p-value	0.0050

Fig. 2.7. Fichier COEUR - Test de Mann-Whitney

2.4.2 Fichier COEUR - Test de Mann-Whitney

Nous souhaitons implémenter le test de Mann-Whitney sur le fichier COEUR. Nous formons la feuille de calcul (Feuille 2.7) :

- Il y a $n_- = 14$ négatifs et $n_+ = 6$ positifs dans le fichier.
- Nous construisons la colonne "Rang". Puisque le fichier a été trié selon un score croissant, elle prend mécaniquement les valeurs $1, 2, \ldots, n$.
- Nous calculons la somme des rangs pour les individus négatifs $r_- = 113$, et pour les positifs $r_+ = 97$.
- Nous en déduisons $U_- = 113 - \frac{14(14+1)}{2} = 8$, $U_+ = 97 - \frac{6(6+1)}{2} = 76$, et $U = \min(U_+, U_-) = 8$.
- La statistique centrée et réduite est égale à

$$Z = \frac{8 - \frac{6\times 14}{2}}{\sqrt{\frac{1}{12} \times (6 + 14 + 1) \times 6 \times 14}} = -2.8043$$

- Nous obtenons la probabilité critique du test avec la loi de répartition normale centrée et réduite p-value $= 0.0050$. Au risque usuel de 5%, nous concluons que les distributions conditionnelles des scores sont décalées.

2.4.3 Acceptation de crédit - Test de Mann-Whitney

Étudions maintenant notre fichier d'acceptation de crédit (section 2.2.3). Nous introduisons 2 nouveautés, (1) le graphique des distributions conditionnelles ; (2) le calcul de la statistique de Mann-Whitney (Figure 2.8, limité aux 40 premières observations) :

- Dans un premier temps, nous calculons le nombre de positifs et de négatifs dans les blocs d'observations définis par les scores $0.0 - 0.2, 0.2 - 0.4, \ldots, 0.8 - 1.0$. Nous noterons au passage qu'il y a $n_- = 27$ observations négatives et $n_+ = 73$ positives.

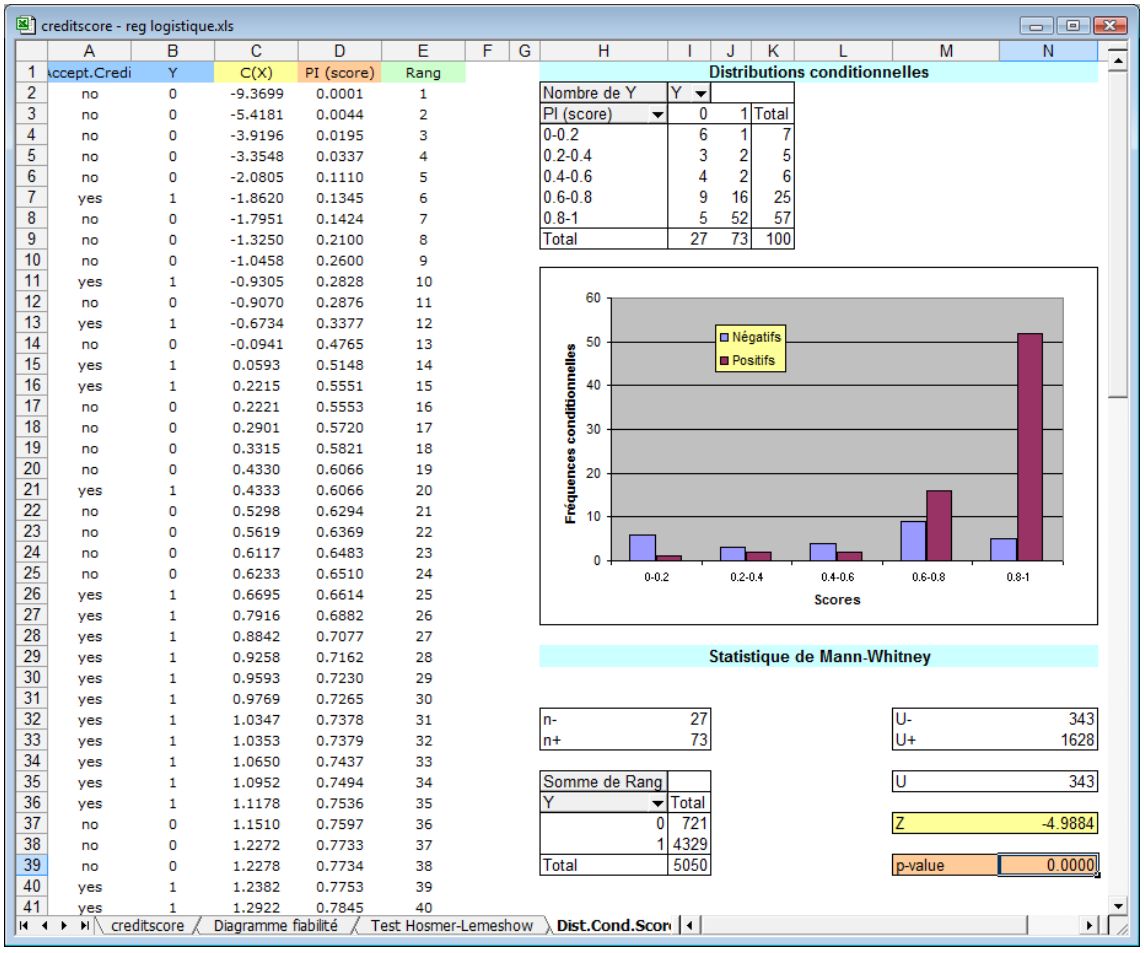


Fig. 2.8. Acceptation de crédit - Test de Mann-Whitney

- Nous en dérivons le graphique des distributions conditionnelles. Il y a manifestement un décalage, notamment pour les valeurs élevées du score où les positifs sont sur-représentés.
- Passons maintenant au calcul de la statistique de Mann-Whitney. Les données ont été triées selon un score croissant. La colonne "Rang" prend les valeurs $1, 2, 3, \ldots, n$.
- Nous réalisons la somme des rangs pour chaque catégorie, toujours à l'aide de l'outil "Tableaux croisés dynamiques" d'Excel. Nous obtenons $r_- = 721$ et $r_+ = 4329$.
- Nous en dérivons $U_- = 721 - \frac{27(27+1}{2} = 343$ et $U_+ = 4329 - \frac{73(73+1}{2} = 1628$
- La statistique de Mann-Whitney s'écrit $U = \min(343, 1628) = 343$
- Et la statistique centrée et réduite

$$Z = \frac{343 - \frac{73.27}{2}}{\sqrt{\frac{1}{12}(73.27 - +1)73.27}} = -4.9884$$

- Nous obtenons ainsi une p-value < 0.0001. Les distributions sont effectivement décalées. Les scores permettent de distinguer les positifs des négatifs.

2.5 La courbe ROC

2.5.1 Justification et construction de la courbe ROC

La courbe ROC est un outil très riche. Son champ d'application dépasse largement le cadre de l'apprentissage supervisé. Elle est par exemple très utilisée en épidémiologie [3]. Pour nous, elle présente surtout des caractéristiques très intéressantes pour l'évaluation et la comparaison des performances des classifieurs [19] :

1. Elle propose un outil graphique qui permet d'évaluer et de comparer globalement le comportement des classifieurs.
2. Elle est indépendante des coûts de mauvaise affectation. Elle permet par exemple de déterminer si un classifieur surpasse un autre, quelle que soit la combinaison de coûts utilisée.
3. Elle est opérationnelle même dans le cas des distributions très déséquilibrées. Mieux, même si les proportions des classes ne sont pas représentatives des probabilités a priori dans le fichier - c'est le cas lorsque l'on procède à un tirage rétrospectif c.-à-d. on fixe le nombre de positifs et négatifs à obtenir, et on tire au hasard dans chaque sous-population - la courbe ROC reste valable.
4. Enfin, on peut lui associer un indicateur synthétique, le critère AUC (aire sous la courbe, en anglais *area under curve*), que l'on sait interpréter.

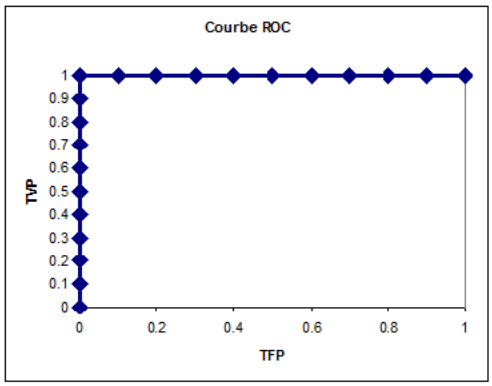
La courbe ROC met en relation le taux de vrais positifs TVP (la sensibilité, le rappel) et le taux de faux positifs TFP (TFP = 1 - Spécificité) dans un graphique nuage de points. Habituellement, nous comparons $\hat{\pi}(\omega)$ à un seuil $s = 0.5$ pour effectuer une prédiction $\hat{y}(\omega)$. Nous pouvons ainsi construire la matrice de confusion et en extraire les 2 indicateurs précités. La courbe ROC généralise cette idée en faisant varier s sur tout le continuum des valeurs possibles entre 0 et 1. Pour chaque configuration, nous construisons la matrice de confusion et nous calculons TVP et TFP.

C'est l'idée directrice. Elle est un peu lourde à mettre en place. Dans la pratique, il n'est pas nécessaire de construire explicitement la matrice de confusion, nous procédons de la manière suivante :

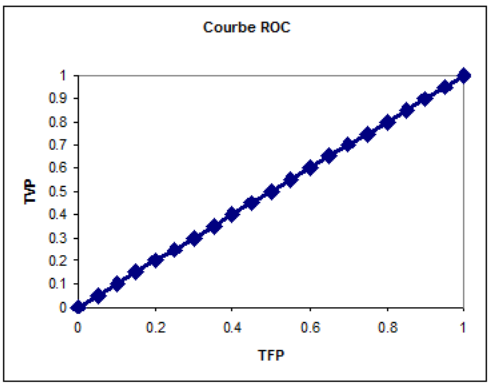
1. Calculer le score $\hat{\pi}(\omega)$ de chaque individu à l'aide du modèle de prédiction.
2. Trier le fichier selon un *score décroissant*.
3. Considérons qu'il n'y a pas d'ex-aequo. Chaque valeur du score peut être potentiellement un seuil s. Pour toutes les observations dont le score est supérieur ou égal à s, les individus dans la partie haute du tableau, nous pouvons comptabiliser le nombre de positifs $n_+(s)$ et le nombre de négatifs $n_-(s)$. Nous en déduisons $TVP = \frac{n_+(s)}{n_+}$ et $TFP = \frac{n_-(s)}{n_-}$.
4. La courbe ROC correspond au graphique nuage de points qui relie les couples (TVP, TFP). Le premier point est forcément $(0, 0)$, le dernier est $(1, 1)$.

Deux situations extrêmes peuvent survenir. La discrimination est parfaite. Tous les positifs sont situés devant les négatifs, la courbe ROC est collée aux extrémités Ouest et Nord du repère (Figure 2.9, A). Les scores sont totalement inopérants, le classifieur attribuent des valeurs au hasard, dans ce cas les positifs et les négatifs sont mélangés. La courbe ROC se confond avec la première bissectrice (Figure 2.9, B).

3. Voir A. Renaud, *Statistique Epidémiologique*, Collection "Que Sais-Je", PUF, 1986 ; pages 103 à 109.



[A] Discrimination parfaite. Tous les positifs sont situés devant les négatifs lorsque l'on trie le tableau selon un score décroisant.



[B] Pas de discrimination. Les positifs et les négatifs sont mélangés c.-à-d. présentent des scores en moyenne identiques.

Fig. 2.9. Courbe ROC - Deux situations extrêmes

2.5.2 Le critère AUC

Il est possible de caractériser numériquement la courbe ROC en calculant la surface située sous la courbe. C'est **le critère AUC**. Elle **exprime la probabilité de placer un individu positif devant un négatif**. Ainsi, dans le cas d'une discrimination parfaite, les positifs sont sûrs d'être placés devant les négatifs, nous avons $AUC = 1$. A contrario, si le classifieur attribue des scores au hasard, il y a autant de chances de placer un positif devant un négatif que l'inverse, la courbe ROC se confond avec la première bissectrice, nous avons $AUC = 0.5$. C'est la situation de référence, notre classifieur doit faire mieux. On propose généralement différents paliers pour donner un ordre d'idées sur la qualité de la discrimination [9] (page 162) (Tableau 2.2).

Valeur de l'AUC	Commentaire
$AUC = 0.5$	Pas de discrimination.
$0.7 \leq AUC < 0.8$	Discrimination acceptable
$0.8 \leq AUC < 0.9$	Discrimination excellente
$AUC \geq 0.9$	Discrimination exceptionnelle

Tableau 2.2. Interprétation des valeurs du critère AUC

Pour calculer l'AUC, nous pouvons utiliser une bête intégration numérique, la méthode des trapèzes par exemple. Nous verrons plus loin que sa valeur peut être obtenue autrement, en faisant le parallèle avec le test de Mann-Whitney.

Au final, il apparaît que le critère AUC est un résumé très commode. Il permet, entre autres, les comparaisons rapides entre les classifieurs. Mais il est évident que si l'on souhaite analyser finement leur comportement, rien ne vaut la courbe ROC.

2.5.3 Fichier COEUR - Courbe ROC

Pour illustrer la construction de la courbe ROC, nous revenons sur le fichier COEUR (Figure 0.1). Voyons le détail des calculs (Figure 2.10) :

n°	coeur	y	PI (Score)	TVP	TFP
				0.0000	0.0000
1	presence	1	0.8789	0.1667	0.0000
2	presence	1	0.8765	0.3333	0.0000
3	absence	0	0.8584	0.3333	0.0714
4	presence	1	0.5815	0.5000	0.0714
5	absence	0	0.4057	0.5000	0.1429
6	presence	1	0.3922	0.6667	0.1429
7	presence	1	0.3782	0.8333	0.1429
8	absence	0	0.3775	0.8333	0.2143
9	presence	1	0.2134	1.0000	0.2143
10	absence	0	0.1727	1.0000	0.2857
11	absence	0	0.1382	1.0000	0.3571
12	absence	0	0.1371	1.0000	0.4286
13	absence	0	0.1244	1.0000	0.5000
14	absence	0	0.1058	1.0000	0.5714
15	absence	0	0.1037	1.0000	0.6429
16	absence	0	0.0737	1.0000	0.7143
17	absence	0	0.0710	1.0000	0.7857
18	absence	0	0.0584	1.0000	0.8571
19	absence	0	0.0362	1.0000	0.9286
20	absence	0	0.0164	1.0000	1.0000

n+	6
n-	14

Fig. 2.10. COEUR - Tableau de calcul de la courbe ROC

- Nous savons qu'il y a $n_+ = 6$ positifs et $n_- = 14$ négatifs dans le fichier.
- Nous avons calculé la colonne des scores $\hat{y}(\omega)$, puis nous avons trié le tableau selon le score décroissant.
- Nous insérons arbitrairement le couple $(0, 0)$.
- Il y a 1 individu ayant un score supérieur ou égal à 0.8789. Il est positif, soit $n_+(0.8789) = 1$ et $TVP_1 = \frac{1}{6} = 0.1667$; par conséquent $n_-(0.8789) = 0$ et $TFP_1 = \frac{0}{14} = 0.0000$.
- Prenons le cas de l'individu n^o4 avec un score de 0.5815. Il a 4 observations avec un score plus grand que ce seuil, avec $n_+(0.5815) = 3$ et $TVP_4 = \frac{3}{6} = 0.5$; concernant les négatifs, nous avons $n_-(0.5815) = 1$ et $TFP_4 = \frac{1}{14} = 0.0714$.
- En procédant ainsi, nous obtenons l'ensemble des points. Il est d'usage d'ajouter la première bissectrice dans le graphique pour que l'on se rende compte visuellement de l'écartement de la courbe ROC par rapport à la situation de référence (Figure 2.11).

Passons maintenant au calcul de l'AUC. Nous utilisons la méthode des trapèzes, une technique d'intégration numérique. On peut toujours discuter de sa précision par rapport à d'autres approches, elle suffit amplement dans notre contexte. Pour calculer la surface du tuyaux d'orgue entre les individus consécutifs $i-1$ et i, nous utilisons

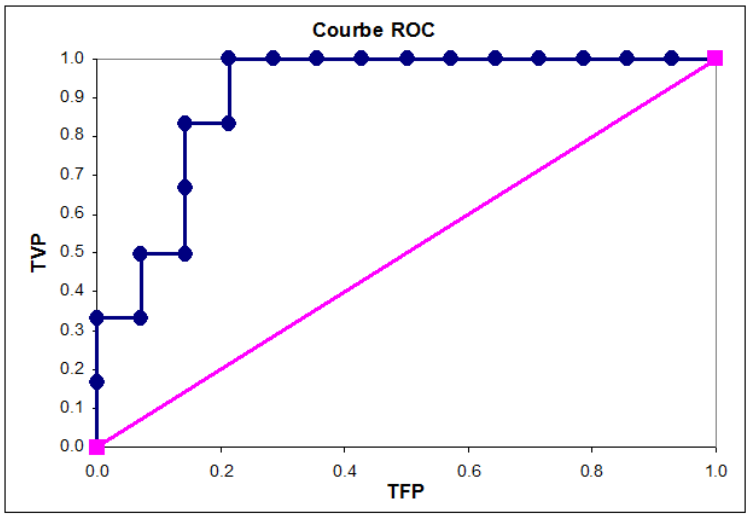


Fig. 2.11. COEUR - Courbe ROC

$$s_i = (TFP_i - TFP_{i-1}) \times \frac{TVP_i + TVP_{i-1}}{2}$$

Nous faisons la somme $AUC = \sum_{i=1}^{n} s_i$ pour obtenir l'aire sous la courbe.

Dans notre exemple (Figure 2.12), nous calculons les s_i successif. Par exemple, pour $i = 3$, nous avons $s_3 = (0.0714 - 0.0000) \times \frac{0.3333+0.3333}{2} = 0.0238$. Au final, nous avons $AUC = 0.0000 + 0.0000 + 0.0238 + 0.0000 + 0.0357 + \cdots + 0.0714 = 0.9048$.

n°	coeur	y	PI (Score)	TVP	TFP	Aire
				0.0000	0.0000	
1	presence	1	0.8789	0.1667	0.0000	0.0000
2	presence	1	0.8765	0.3333	0.0000	0.0000
3	absence	0	0.8584	0.3333	0.0714	0.0238
4	presence	1	0.5815	0.5000	0.0714	0.0000
5	absence	0	0.4057	0.5000	0.1429	0.0357
6	presence	1	0.3922	0.6667	0.1429	0.0000
7	presence	1	0.3782	0.8333	0.1429	0.0000
8	absence	0	0.3775	0.8333	0.2143	0.0595
9	presence	1	0.2134	1.0000	0.2143	0.0000
10	absence	0	0.1727	1.0000	0.2857	0.0714
11	absence	0	0.1382	1.0000	0.3571	0.0714
12	absence	0	0.1371	1.0000	0.4286	0.0714
13	absence	0	0.1244	1.0000	0.5000	0.0714
14	absence	0	0.1058	1.0000	0.5714	0.0714
15	absence	0	0.1037	1.0000	0.6429	0.0714
16	absence	0	0.0737	1.0000	0.7143	0.0714
17	absence	0	0.0710	1.0000	0.7857	0.0714
18	absence	0	0.0584	1.0000	0.8571	0.0714
19	absence	0	0.0362	1.0000	0.9286	0.0714
20	absence	0	0.0164	1.0000	1.0000	0.0714

n+	6
n-	14

AUC	0.9048

Fig. 2.12. COEUR - Calcul de l'AUC à partir des TFP et TVP

Nous avons 90.5% de chances de placer un positif devant un négatif en "scorant" avec notre classifieur, à comparer avec les 50% de la situation de référence. Ce résultat est plutôt encourageant. On pouvait facilement le deviner d'ailleurs en observant le graphique (Figure 2.11). La courbe s'écarte sensiblement de la première bissectrice. Elle semble indiquer - avec les réserves toujours de mise tant que nous évaluons notre modèle sur les données d'apprentissage - que notre modèle est plutôt exceptionnel (cf. Tableau 2.2) avec des estimations $\hat{\pi}(\omega)$ discriminatoires. Ce que ne laissait pas entendre le taux d'erreur en resubstitution de $\epsilon_{resub} = 0.2$ issu de la matrice de confusion, basé uniquement sur les prédictions $\hat{y}(\omega)$.

2.5.4 Critère AUC et Statistique de Mann-Whitney

Il existe une relation entre la statistique U_+ de Mann-Whitney et le critère AUC [9] (page 164). La meilleure justification est certainement du côté de l'interprétation de ces quantités sous l'angle des comparaisons par paires [23] (pages 409-411). La relation est la suivante

$$AUC = \frac{U_+}{n_+ \times n_-} \tag{2.6}$$

Reprenons notre exemple COEUR, le tableau de calcul de la statistique de Mann-Whitney (Figure 2.7) nous fournit $U_+ = 76$. Lorsque nous formons l'expression ci-dessus, nous retrouvons $AUC = \frac{76}{6 \times 14} = 0.9048$. Exactement la valeur de l'aire sous la courbe obtenue par la méthode de trapèze (Figure 2.12). Il est effectivement possible d'obtenir directement l'AUC via la statistique de Mann-Whitney.

2.6 La courbe rappel-précision

2.6.1 Principe de la courbe rappel-précision

La courbe rappel-précision est très utilisée en recherche d'information (en anglais, *information retrieval*). Suite à une requête, nous obtenons un ensemble d'individus que nous appellerons la "cible", nous sommes face à deux exigences contradictoires : nous aimerions retrouver une fraction élevée des positifs potentiels (rappel) ; nous aimerions que la cible ne contienne que des positifs (précision). La courbe traduit l'arbitrage entre ces deux critères lorsque l'on fait varier le seuil d'affectation s.

Elle est conceptuellement proche de la courbe ROC. Pour chaque valeur de s, nous formons (virtuellement) la matrice de confusion et nous calculons les deux indicateurs. Il y a quand même une différence très importante. La précision étant un "profil-colonne" de la matrice de confusion, il faut donc travailler sur un échantillon représentatif (la proportion des positifs $\frac{n_+}{n}$ doit être le reflet de la probabilité d'être positif p) pour pouvoir l'exploiter convenablement. Si cette condition est respectée, elle paraît plus adaptée que la courbe ROC lorsque les classes sont très déséquilibrées (la proportion des positifs est très faible), notamment pour différencier le comportement des algorithmes d'apprentissage supervisé.

Pour élaborer la courbe rappel-précision, nous procédons comme suit :

1. Calculer le score $\hat{\pi}$ de chaque individu.

2. Trier les données selon un score décroissant.
3. Mettons qu'il n'y a pas d'ex-aequo, chaque valeur du score est un seuil potentiel s. Pour les individus situés dans la partie haute du tableau c.-à-d. dont le score est supérieur ou égal à s, il s'agit de la cible, nous comptabilisons le nombre de positifs $n_+(s)$ et le nombre total d'observations $n(s)$.
4. Nous en déduisons le $rappel(s) = \frac{n_+(s)}{n_+}$ et la précision $precision(s) = \frac{n_+(s)}{n(s)}$.

Dans les parties hautes du tableau, lorsque le seuil est élevé, la taille de la cible sera réduite. La précision sera forte, dans la cible ne seront présents que des positifs ; mais le rappel sera faible, une faible fraction de l'ensemble des positifs y sont inclus. A mesure que s diminue, la taille de la cible augmente, elle sera de plus en plus polluée (la précision diminue) mais intégrera une plus grande fraction des positifs (le rappel augmente). La courbe est donc globalement décroissante, mais elle n'est pas forcément monotone.

2.6.2 Fichier COEUR - Courbe rappel-précision

n°	coeur	y	PI (Score)	Rappel	Précision
1	presence	1	0.8789	0.1667	1.0000
2	presence	1	0.8765	0.3333	1.0000
3	absence	0	0.8584	0.3333	0.6667
4	presence	1	0.5815	0.5000	0.7500
5	absence	0	0.4057	0.5000	0.6000
6	presence	1	0.3922	0.6667	0.6667
7	presence	1	0.3782	0.8333	0.7143
8	absence	0	0.3775	0.8333	0.6250
9	presence	1	0.2134	1.0000	0.6667
10	absence	0	0.1727	1.0000	0.6000
11	absence	0	0.1382	1.0000	0.5455
12	absence	0	0.1371	1.0000	0.5000
13	absence	0	0.1244	1.0000	0.4615
14	absence	0	0.1058	1.0000	0.4286
15	absence	0	0.1037	1.0000	0.4000
16	absence	0	0.0737	1.0000	0.3750
17	absence	0	0.0710	1.0000	0.3529
18	absence	0	0.0584	1.0000	0.3333
19	absence	0	0.0362	1.0000	0.3158
20	absence	0	0.0164	1.0000	0.3000

n+	6
n-	14

Fig. 2.13. COEUR - Tableau de calcul de la courbe rappel-précision

Nous reprenons le fichier COEUR. La structure du tableau de calcul est très similaire à celle de la courbe ROC. Les données sont triées selon $\hat{\pi}$ décroissant (Figure 2.13) :

- Il y a $n_+ = 6$ positifs et $n_- = 14$ négatifs.
- Pour le seuil $s = 0.8789$, la cible contient un seul individu $n(s) = 1$ et c'est un positif. Nous avons $rappel = \frac{1}{6} = 0.1667$ et $precision = \frac{1}{1} = 1$.
- En passant au second individu, qui est toujours un positif, nous obtenons $rappel = \frac{2}{6} = 0.3333$ et $precision = \frac{2}{2} = 1$.

– Nous procédons jusqu'au dernier individu, nous obtenons pour celui-ci $rappel = \frac{6}{6} = 1$ et $precision = \frac{6}{20} = 0.3$.
– Nous obtenons ainsi tous les points qui composent la courbe (Figure 2.14).

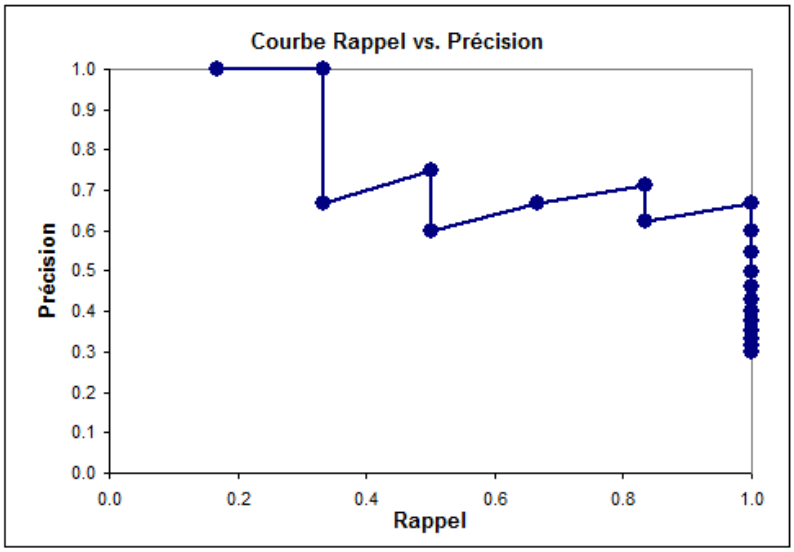


Fig. 2.14. COEUR - Courbe rappel-précision

3

Tests de significativité des coefficients

3.1 Quoi et comment tester ?

3.1.1 Écriture des hypothèses à tester

L'objectif des tests de significativité est d'éprouver le rôle d'une, de plusieurs, de l'ensemble, des variables explicatives. Formellement, les hypothèses nulles peuvent se décliner comme suit :

1. Évaluer la contribution individuelle d'une variable

   $$H_0 : a_j = 0$$

   Ce test de significativité est systématiquement donné par les logiciels. Nous verrons plus loin que seule une de ses formes (test de Wald) est en réalité proposée. L'autre (test du rapport de vraisemblance) est passée sous silence. Or ces approches ne se comportent pas de la même manière. Il faut le savoir pour interpréter les résultats en connaissance de cause.

2. Évaluer la contribution d'un bloc de "q" variables. Sans restreindre la généralité du propos (les coefficients à tester ne sont pas forcément consécutifs dans la régression), nous écrirons H_0 de la manière suivante

   $$H_0 : a_j = a_{j+1} = \cdots = a_{j+q} = 0$$

   On ne peut pas le transformer en une succession de tests individuels. En effet, les coefficients ne sont pas indépendants (en tous les cas, ils ont une covariance non-nulle). Il faut bien tester la nullité simultanée des q coefficients.

3. Évaluer l'apport de l'ensemble des variables explicatives. Nous avons ici une formulation statistique du problème abordé lors de la définition des pseudo-R^2 (section 1.6) .

   $$H_0 : a_1 = a_2 = \cdots = a_J = 0$$

   Il s'agit d'une évaluation globale de la régression. En effet, si l'hypothèse nulle est compatible avec les données, cela signifierait qu'aucun des descripteurs ne contribue à l'explication de la variable dépendante. Le modèle peut être jeté aux orties.

Dans tous les cas, l'hypothèse alternative correspond à : "un des coefficients au moins est non-nul".

Notons que ces tests s'inscrivent dans le cadre d'une formulation générale de la forme

$$H_0 : Ma = 0$$

où M est une matrice de contrastes indépendants à m lignes et $J+1$ colonnes, de rang m. La procédure et les formules sont un peu complexes, mais nous pouvons évaluer tout type de configuration ([23], page 421 ; [7], page 90).

3.1.2 Deux approches pour les tests

Nous disposons de 2 stratégies pour implémenter ces tests :

1. S'appuyer sur le principe du **rapport de vraisemblance**. L'approche est générique, elle est en cohérence avec la démarche d'estimation des paramètres. Elle est puissante c.-à-d. elle détecte mieux l'hypothèse alternative lorsqu'elle est vraie. L'inconvénient est qu'elle est plus gourmande en ressources machines : chaque hypothèse à évaluer donne lieu à une nouvelle estimation des paramètres, donc à un processus d'optimisation. Certes les logiciels et les ordinateurs actuels sont très performants. Il reste que le surcroît de calcul n'est pas négligeable lorsque nous traitons de grandes bases de données.
2. S'appuyer sur la normalité asymptotique des estimateurs (du maximum de vraisemblance). On parle de **test de Wald**. Le principal avantage est que les informations que l'on souhaite exploiter sont toutes disponibles à l'issue de l'estimation du modèle complet, incluant l'ensemble des variables. L'obtention des résultats est donc immédiate. L'inconvénient est que le test de Wald est conservateur. Il a tendance à favoriser l'hypothèse nulle.

Dans ce chapitre, nous présentons tour à tour ces deux démarches pour les configurations énumérées ci-dessus. Nous confronterons les résultats sur le fichier COEUR. Vu la très faible taille du fichier, $n = 20$, nous fixerons le risque de première espèce à 10%.

3.2 Tests fondés sur le rapport de vraisemblance

3.2.1 Principe du rapport de vraisemblance

Le test du rapport de vraisemblance consiste à comparer les vraisemblances de 2 modèles emboîtés M_r et M_s ([9], pages 36 à 40 ; [10], page 22). M_r comporte r variables, avec donc $r + 1$ paramètres à estimer (le nombre de degrés de liberté du modèle est égal à $[n - (r + 1) = n - r - 1]$) ; M_s en comporte s ($s < r$), avec pour contrainte, et c'est pour cela qu'on parle de modèles emboîtés, qu'elles se retrouvent toutes dans M_r.

La statistique de test s'écrit :

$$LR = -2 \times \ln \frac{L(M_s)}{L(M_r)} \tag{3.1}$$

où $L(M)$ représente la vraisemblance du modèle. Sous H_0, les coefficients des variables supplémentaires que l'on retrouve dans M_r sont tous nuls, LR suit une loi du χ^2 à $(r-s)$ degrés de liberté [7] (page 114).

L'expression ci-dessus peut se décliner sous plusieurs formes

$$\begin{aligned} LR &= -2 \times \ln \frac{L(M_s)}{L(M_r)} \\ &= -2 \times LL(M_s) - (-2 \times LL(M_r)) \\ &= D_s - D_r \end{aligned}$$

où $LL(M)$ est la log-vraisemblance, D la déviance.

Quelques remarques :

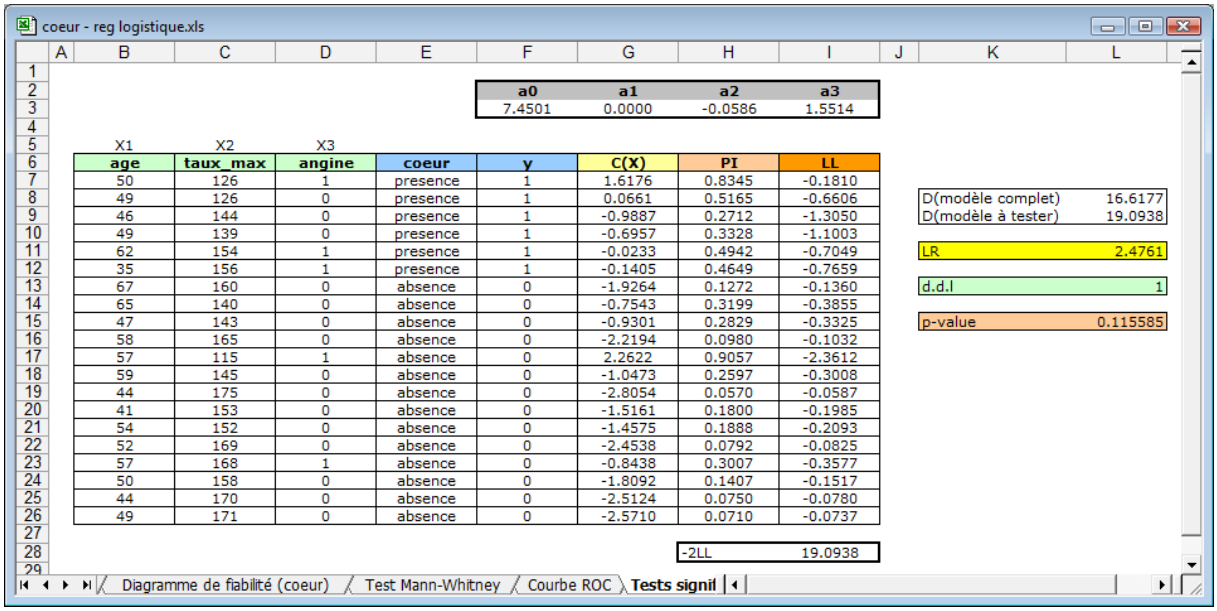
1. $LR \geq 0$, plus on rajoute de variables dans la régression, mêmes non pertinentes, plus faible sera la déviance [1].
2. Dans les tests qui nous intéressent (section 3.1.1), le modèle M_r correspond au modèle complet intégrant les J variables explicatives. Pour le fichier COEUR, la déviance du modèle est égale à $D_M = 16.618$.

3.2.2 Tester la nullité d'un des coefficients

Pour tester la significativité d'un des coefficients, il suffit de comparer la déviance du modèle avec et sans la variable incriminée.

Nous souhaitons tester le coefficient de AGE dans la régression COEUR. Nous devons tout d'abord réaliser une nouvelle estimation, optimiser la vraisemblance, en excluant cette variable c.-à-d. en mettant arbitrairement le coefficient a_1 à 0. Nous réalisons l'opération à l'aide du solveur d'Excel (Figure 3.1). Nous obtenons une déviance $D_{(taux,angine)} = 19.0938$. Nous pouvons former le rapport de vraisemblance $LR = 19.0938 - 16.6177 = 2.4761$. La probabilité critique avec la loi de répartition du $\chi^2(1)$ à $3-2=1$ degré de liberté est p-value $= 0.115585$.

Au risque 10%, les données sont compatibles avec l'hypothèse nulle $a_1 = 0$, la variable AGE ne contribue pas à l'explication des valeurs de COEUR.



coeur - reg logistique.xls

	a0	a1	a2	a3
	7.4501	0.0000	-0.0586	1.5514

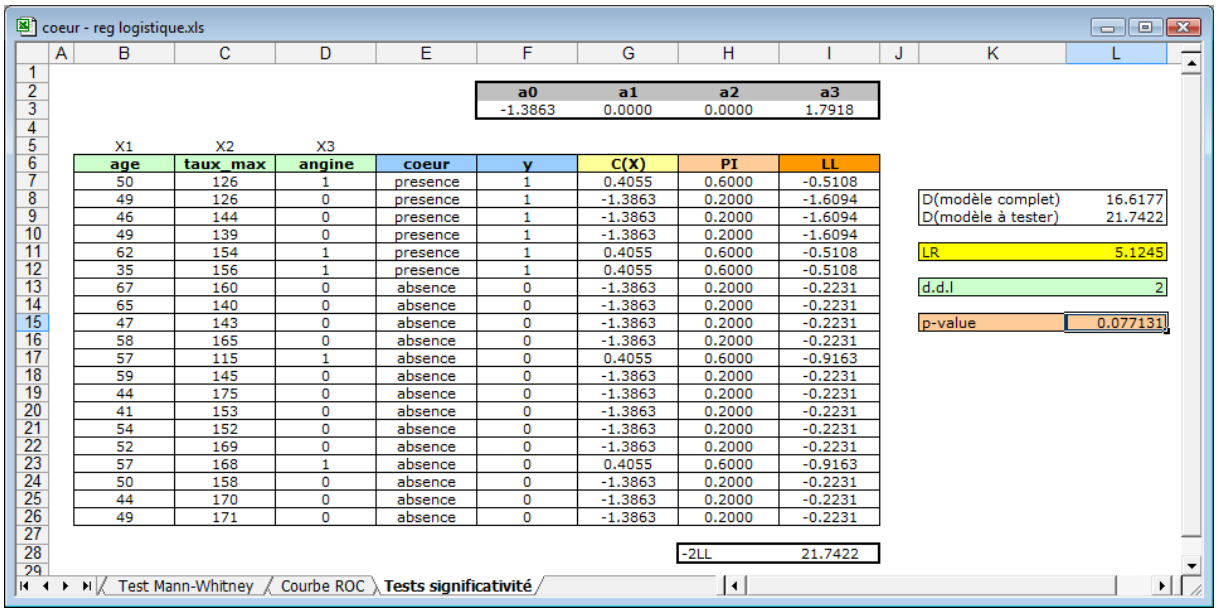
	X1	X2	X3					
	age	taux_max	angine	coeur	y	C(X)	PI	LL
1	50	126	1	presence	1	1.6176	0.8345	-0.1810
2	49	126	0	presence	1	0.0661	0.5165	-0.6606
3	46	144	0	presence	1	-0.9887	0.2712	-1.3050
4	49	139	0	presence	1	-0.6957	0.3328	-1.1003
5	62	154	1	presence	1	-0.0233	0.4942	-0.7049
6	35	156	1	presence	1	-0.1405	0.4649	-0.7659
7	67	160	0	absence	0	-1.9264	0.1272	-0.1360
8	65	140	0	absence	0	-0.7543	0.3199	-0.3855
9	47	143	0	absence	0	-0.9301	0.2829	-0.3325
10	58	165	0	absence	0	-2.2194	0.0980	-0.1032
11	57	115	1	absence	0	2.2622	0.9057	-2.3612
12	59	145	0	absence	0	-1.0473	0.2597	-0.3008
13	44	175	0	absence	0	-2.8054	0.0570	-0.0587
14	41	153	0	absence	0	-1.5161	0.1800	-0.1985
15	54	152	0	absence	0	-1.4575	0.1888	-0.2093
16	52	169	0	absence	0	-2.4538	0.0792	-0.0825
17	57	168	1	absence	0	-0.8438	0.3007	-0.3577
18	50	158	0	absence	0	-1.8092	0.1407	-0.1517
19	44	170	0	absence	0	-2.5124	0.0750	-0.0780
20	49	171	0	absence	0	-2.5710	0.0710	-0.0737

-2LL	19.0938

D(modèle complet)	16.6177
D(modèle à tester)	19.0938
LR	2.4761
d.d.l	1
p-value	0.115585

Diagramme de fiabilité (coeur) / Test Mann-Whitney / Courbe ROC / Tests signif

Fig. 3.1. COEUR - Tester la significativité du coefficient a_1



coeur - reg logistique.xls

	a0	a1	a2	a3
	-1.3863	0.0000	0.0000	1.7918

	X1	X2	X3					
	age	taux_max	angine	coeur	y	C(X)	PI	LL
1	50	126	1	presence	1	0.4055	0.6000	-0.5108
2	49	126	0	presence	1	-1.3863	0.2000	-1.6094
3	46	144	0	presence	1	-1.3863	0.2000	-1.6094
4	49	139	0	presence	1	-1.3863	0.2000	-1.6094
5	62	154	1	presence	1	0.4055	0.6000	-0.5108
6	35	156	1	presence	1	0.4055	0.6000	-0.5108
7	67	160	0	absence	0	-1.3863	0.2000	-0.2231
8	65	140	0	absence	0	-1.3863	0.2000	-0.2231
9	47	143	0	absence	0	-1.3863	0.2000	-0.2231
10	58	165	0	absence	0	-1.3863	0.2000	-0.2231
11	57	115	1	absence	0	0.4055	0.6000	-0.9163
12	59	145	0	absence	0	-1.3863	0.2000	-0.2231
13	44	175	0	absence	0	-1.3863	0.2000	-0.2231
14	41	153	0	absence	0	-1.3863	0.2000	-0.2231
15	54	152	0	absence	0	-1.3863	0.2000	-0.2231
16	52	169	0	absence	0	-1.3863	0.2000	-0.2231
17	57	168	1	absence	0	0.4055	0.6000	-0.9163
18	50	158	0	absence	0	-1.3863	0.2000	-0.2231
19	44	170	0	absence	0	-1.3863	0.2000	-0.2231
20	49	171	0	absence	0	-1.3863	0.2000	-0.2231

-2LL	21.7422

D(modèle complet)	16.6177
D(modèle à tester)	21.7422
LR	5.1245
d.d.l	2
p-value	0.077131

Test Mann-Whitney / Courbe ROC / Tests significativité

Fig. 3.2. COEUR - Tester la nullité simultanée des coefficients a_1 et a_2

3.2.3 Tester la nullité de q $(q < J)$ coefficients

La démarche est toujours la même si nous souhaitons tester la nullité de q coefficients. Le rapport de vraisemblance suit une loi du χ^2 à (q) degrés de liberté.

Testons la nullité simultanée des coefficients a_1 (AGE) et a_2 (TAUX MAX) dans notre régression. Dans la feuille Excel, nous fixons arbitrairement $a_1 = a_2 = 0$, nous réalisons la minimisation de la déviance en introduisant a_0 et a_3 (ANGINE) en cellules variables dans le solveur. Nous obtenons $D_{angine} = 21.7422$.

Le rapport de vraisemblance est égal à $LR = 21.7422 - 16.6177 = 5.1245$. Avec une loi du χ^2 à 2 degrés de liberté, nous aboutissons à une p-value = 0.077131 (Figure 3.2).

Au risque 10%, les données ne sont pas compatibles avec l'hypothèse nulle $a_1 = a_2 = 0$ c.-à-d. on ne peut pas conclure à la nullité simultanée des 2 coefficients.

Remarque : Nous ne montrons pas les calculs mais, dans cet exemple, le rejet de l'hypothèse nulle est avant tout consécutif à la significativité individuelle du coefficient de TAUX MAX ($LR = 3.0840$, p-value pour $\chi^2(1)$ égale à 0.079067). Il arrive parfois que tous les coefficients pris individuellement soient non significatifs. En revanche, lorsque l'on teste leur nullité simultanée, on est amené à rejeter l'hypothèse nulle.

3.2.4 Tester globalement la nullité des J coefficients $(a_1, \ldots, a_J)$

Ce test revient à comparer la vraisemblance du modèle complet avec celle du modèle trivial constitué uniquement de la constante. Nous avions déjà analysé cette configuration auparavant, nous pouvions estimer directement "le" paramètre du modèle $\hat{a}_0$ et en déduire la déviance D_0 (section 1.6).

Dans le cas du fichier COEUR, nous avions $\hat{a}_0 = -0.8473$ et $D_0 = 24.4346$ (Figure 1.6). Nous calculons la statistique

$$LR = D_0 - D_M = 24.4346 - 16.6177 = 7.8169$$

Le nombre de degrés de liberté est égale à $J - 0 = 3 - 0 = 3$. Avec la fonction de répartition de la loi du χ^2, nous obtenons la probabilité critique p-value = 0.049952.

Au risque 10%, nous rejetons l'hypothèse nulle, les données ne sont pas compatibles avec l'hypothèse de nullité de tous les coefficients c.-à-d. le modèle est globalement significatif.

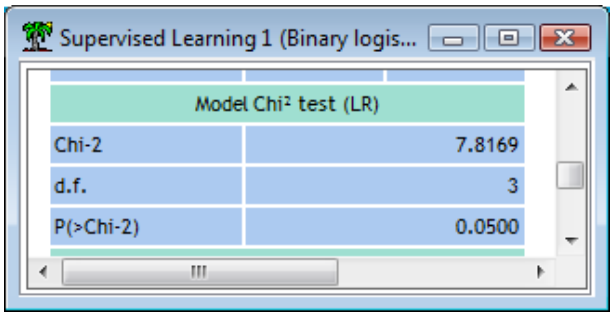


Fig. 3.3. COEUR - Tester la siginifcativité globale du modèle - Tanagra

Notons que les logiciels proposent toujours, d'une manière ou d'une autre, ce test pour évaluer le modèle. Dans Tanagra, le tableau *Model Chi2 test (LR)* fournit la statistique $LR = 7.8169$, le degré de liberté 3, et la p-value 0.0500 (Figure 3.3). Le logiciel R, lui, fournit la *null deviance*, la déviance du modèle trivial, de 24.435 avec $(n - 1 = 19)$ degrés de liberté ; et la déviance du modèle étudié, *residual deviance*, de 16.618 avec $(n - 3 - 1 = 16)$ degrés de liberté. En calculant l'écart entre ces quantités, nous retrouvons le test de significativité globale (Figure 3.4).

1. Nous pouvons faire l'analogie avec la somme des carrés des résidus en régression linéaire multiple.

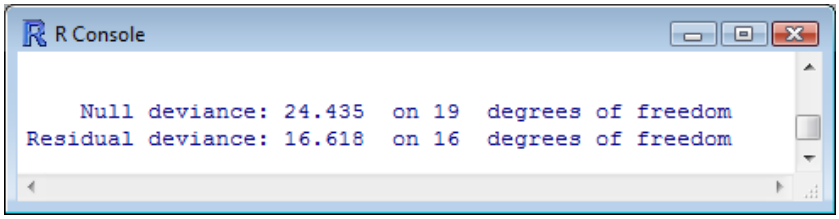


Fig. 3.4. COEUR - Tester la siginifcativité globale du modèle - R

3.3 Tests fondés sur la normalité asymptotique des coefficients - Tests de Wald

Les estimateurs du maximum de vraisemblance sont asymptotiquement normaux. Par conséquent, lorsque les effectifs sont assez élevés, le vecteur $\hat{a}$ suit une loi normale multidimensionnelle. Il importe tout d'abord de déterminer l'expression de sa matrice de variance covariance. Nous pourrons par la suite décliner les différents tests de significativité (section 3.1.1).

3.3.1 Matrice de variance-covariance des coefficients

Matrice Hessienne. Lors de la description de l'algorithme d'optimisation de Newton-Raphson, nous avons défini une matrice des dérivées partielles secondes, dite matrice hessienne (section 1.5). Nous en reprenons l'expression matricielle ici

$$H = X'VX$$

Où X est la matrice des données, la première colonne correspondant à la constante. Elle est de dimension $n \times (J+1)$. Pour les données COEUR (Figure 0.1), les valeurs s'écrivent

$$X = \begin{pmatrix} 1 & 50 & 126 & 1 \\ 1 & 49 & 126 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 49 & 171 & 0 \end{pmatrix}$$

V est une matrice diagonale de taille $n \times n$, composée des valeurs de $\pi(\omega) \times (1 - \pi(\omega))$, les probabilités $\pi(\omega)$ étant obtenues après estimation des paramètres. En reprenant les valeurs issues des calculs (Figure 1.5), nous avons $\pi(1) = 0.8798, \pi(2) = 0.5815, \pi(3) = 0.3922, \ldots, \pi(20) = 0.0737$), et par conséquent

$$V = \begin{pmatrix} 0.8789(1-0.8789) = 0.1064 & 0 & 0 & \ldots & 0 \\ 0 & 0.5815(1-0.5815) = 0.2434 & 0 & \ldots & 0 \\ 0 & 0 & 0.2384 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & \ldots & 0.0683 \end{pmatrix}$$

Ainsi, nous pouvons former la matrice hessienne H de taille $(J+1)\times(J+1)$,

$$H = \begin{pmatrix} 2.61 & 130.24 & 386.30 & 0.65 \\ 130.24 & 6615.41 & 19211.02 & 34.59 \\ 386.30 & 19211.02 & 57709.57 & 94.12 \\ 0.65 & 34.59 & 94.12 & 0.65 \end{pmatrix}$$

Matrice de variance covariance des coefficients. L'affaire devient intéressante lorsque l'on sait que **l'inverse de la matrice hessienne** correspond à la matrice de variance covariance des coefficients estimés. En particulier, nous obtenons les variances des coefficients sur la diagonale principale.

$$\hat{\Sigma} = H^{-1} \tag{3.2}$$

Dans notre exemple COEUR, la matrice qui en résulte est

$$\hat{\Sigma} = \begin{pmatrix} 63.2753 & -0.4882 & -0.2627 & 1.0563 \\ -0.4882 & 0.0088 & 0.0004 & -0.0413 \\ -0.2627 & 0.0004 & 0.0016 & 0.0030 \\ 1.0563 & -0.0413 & 0.0030 & 2.2634 \end{pmatrix}$$

Nous lisons dans ce tableau, entre autres :

- $\hat{\sigma}_1^2 = 0.0088$ est la variance estimée du coefficient $\hat{a}_1$.
- $\widehat{COV}(\hat{a}_1, \hat{a}_2) = 0.0004$ est la covariance estimée entre les coefficients $\hat{a}_1$ et $\hat{a}_2$.
- Etc.

Test de Wald. Nous disposons de $\hat{a}$, vecteur des estimations des paramètres de la régression logistique ; nous savons qu'il suit une loi normale multidimensionnelle ; nous disposons de la matrice de variance covariance associée. Tout est en place pour que nous puissions réaliser les différents tests de significativité. Ils sont regroupés sous l'appellation test de Wald ([7], pages 90 et 113 ; [23], page 421).

3.3.2 Tester la nullité d'un des coefficients

Très facile à mettre en oeuvre puisque l'on dispose directement de la variance des coefficients, le test s'appuie sur la statistique de Wald W_j qui, sous H_0, suit une loi du χ^2 à 1 degré de liberté.

$$W_j = \frac{\hat{a}_j^2}{\hat{\sigma}_{\hat{a}_j}^2} \tag{3.3}$$

Où $\hat{\sigma}_{\hat{a}_j}^2$ est la variance du coefficient $\hat{a}_j$, lue sur la diagonale principale de la matrice de variance covariance de coefficients $\hat{\Sigma}$.

Dans notre exemple du fichier COEUR, puisque nous avons les valeurs des coefficients et la matrice de variance covariance associée, nous pouvons réaliser le test que nous résumons dans le tableau suivant.

Coefficient	Estimation	$\hat{\sigma}^2_{\hat{a}_j}$	W_j	p-value
a_0	14.494	63.2753	3.3200	0.0684
a_1	-0.126	0.0088	1.7936	0.1805
a_2	-0.064	0.0016	2.4694	0.1161
a_3	1.779	2.2634	1.3981	0.2370

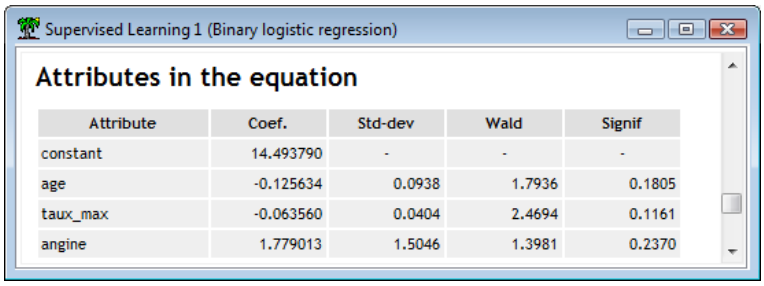


Fig. 3.5. COEUR - Test de Wald - Tanagra

A titre de comparaison, nous reproduisons les sorties du logiciel Tanagra (Figure 3.5). Nous obtenons les mêmes valeurs, à la différence que Tanagra affiche plutôt les écarts-type estimés $\hat{\sigma}_{\hat{a}_j}$. Et il ne réalise pas le test de significativité de la constante.

Le logiciel R, lui, propose la statistique Z_j (Figure 3.6) à la place de W_j, avec

$$Z_j = \frac{\hat{a}_j}{\hat{\sigma}_j} = \text{signe}(\hat{a}_j) \times \sqrt{W_j} \sim \mathcal{N}(0,1)$$

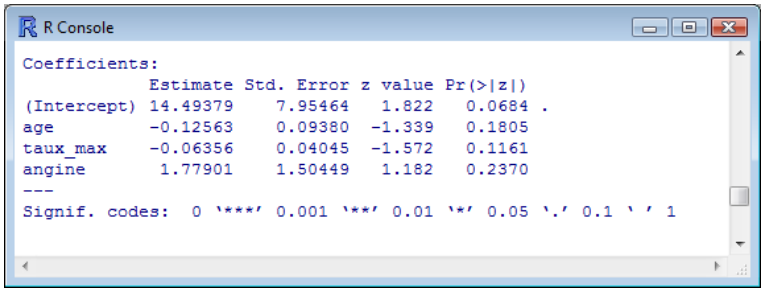


Fig. 3.6. COEUR - Test de Wald - Logiciel R

Z_j peut prendre des valeurs négatives. Le test étant bilatéral, nous retrouvons exactement les mêmes probabilités critiques (p-value) qu'avec la statistique de Wald W_j.

3.3.3 Intervalle de confiance de Wald pour un coefficient

$\hat{a}_j$ suit asymptotiquement une loi normale que l'on soit ou non au voisinage de $a_j = 0$. De fait, nous pouvons construire l'intervalle de confiance au niveau de confiance $1 - \alpha$ pour tout coefficient pris individuellement ([9], pages 18 et 40; [7], page 91). Les bornes sont obtenues de la manière suivante

$$\hat{a}_j \pm u_{1-\alpha/2} \times \hat{\sigma}_{\hat{a}_j} \tag{3.4}$$

$u_{1-\alpha/2}$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

Pour notre fichier COEUR, nous calculons les intervalles des coefficients au niveau de confiance $1-\alpha = 90\%$, avec $u_{0.95} = 1.6449$.

Coefficient	$\hat{a}_j$	$\hat{\sigma}_{\hat{a}_j}$	Borne basse	Borne haute
a_0	14.494	7.9546	1.41	27.58
a_1	-0.126	0.0938	−0.28	0.03
a_2	−0.064	0.0404	−0.13	0.00
a_3	1.779	1.5045	−0.70	4.25

Remarque : Il est possible de construire un intervalle de confiance basé sur le rapport de vraisemblance [7] (page 91). Nous n'en avons pas fait mention dans la section précédente tout simplement parce que la formulation est compliquée, pas vraiment utilisée dans la pratique car peu décisive par rapport à l'intervalle de Wald, et de ce fait non implémentée dans les logiciels (à ma connaissance).

3.3.4 Tester la nullité de q ($q < J$) coefficients

Pour tester la nullité simultanée de q coefficients, nous utilisons la généralisation de la statistique de Wald $W_{(q)}$. Elle suit une loi du χ^2 à q degrés de liberté.

$$W_{(q)} = \hat{a}'_{(q)} \times \hat{\Sigma}^{-1}_{(q)} \times \hat{a}_{(q)} \tag{3.5}$$

où $\hat{a}_{(q)}$ est le sous-vecteur des valeurs observées des coefficients que l'on souhaite tester ; $\hat{\Sigma}_{(q)}$ est la sous-matrice de variance covariance associée à ces coefficients.

Rien ne vaut un petit exemple pour préciser tout cela. Nous souhaitons, pour le fichier COEUR, tester la nullité simultanée des coefficients rattachés à AGE et TAUX MAX. L'hypothèse nulle s'écrit :

$$H_0 : a_1 = a_2 = 0$$

Sous une forme vectorielle

$$H_0 : \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Le vecteur des coefficients estimés est égal à

$$\hat{a}_{(2)} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} -0.126 \\ -0.064 \end{pmatrix}$$

La matrice de variance covariance associée à ces coefficients, extraite de la matrice globale[2] $\hat{\Sigma}$ s'écrit

$$\hat{\Sigma}_{(2)} = \begin{pmatrix} 0.0088 & 0.0004 \\ 0.0004 & 0.0016 \end{pmatrix}$$

Nous inversons cette matrice pour obtenir

$$\hat{\Sigma}_{(2)}^{-1} = \begin{pmatrix} 114.97 & -28.58 \\ -28.58 & 618.40 \end{pmatrix}$$

Il ne reste plus qu'à calculer la forme quadratique définissant $W_{(2)}$, soit

$$\begin{aligned} W_{(2)} &= \begin{pmatrix} -0.126 & -0.064 \end{pmatrix} \times \begin{pmatrix} 114.97 & -28.58 \\ -28.58 & 618.40 \end{pmatrix} \times \begin{pmatrix} -0.126 \\ -0.064 \end{pmatrix} \\ &= 3.8565 \end{aligned}$$

Avec une loi du χ^2 à 2 degrés de liberté, nous obtenons une p-value $= 0.1454$. Au risque 10%, nous ne pouvons pas rejeter l'hypothèse nulle. Nos données sont compatibles avec l'hypothèse de nullité simultanée des coefficients a_1 et a_2. Ce résultat est en contradiction avec celui du test de rapport de vraisemblance. Nous y reviendrons par la suite.

3.3.5 Tester globalement la nullité des J coefficients

Dernier test à mettre en place, évaluer la significativité globale du modèle c.-à-d. tester la nullité simultanée de tous les coefficients relatifs aux variables explicatives dans le modèle. L'hypothèse nule s'écrit

$$H_0 : a_1 = a_2 = \cdots = a_J = 0$$

Attention, la constante a_0 ne doit pas être prise en compte dans cette procédure.

Le test de Wald ici correspond à une simple généralisation du précédent. La statistique $W_{(J)}$ suit une loi du χ^2 à J degrés de liberté sous H_0. Elle s'écrit

$$W_{(J)} = \hat{a}'_{(J)} \times \hat{\Sigma}_{(J)}^{-1} \times \hat{a}_{(J)} \tag{3.6}$$

Pour le fichier COEUR, voici le vecteur estimé des coefficients concernés

2. La situation est facilitée par le fait que les coefficients sont consécutifs dans notre exemple. Mais nous pouvons appliquer ce test en toute généralité, pour toute combinaison de coefficients, qu'ils soient consécutifs ou non.

$$\hat{a}_{(3)} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{pmatrix} = \begin{pmatrix} -0.126 \\ -0.064 \\ 1.779 \end{pmatrix}$$

Et la sous-matrice de variance covariance

$$\hat{\Sigma}_{(3)} = \begin{pmatrix} 0.0088 & 0.0004 & -0.0413 \\ 0.0004 & 0.0016 & 0.0030 \\ -0.0413 & 0.0030 & 2.2635 \end{pmatrix}$$

Nous inversons cette dernière

$$\hat{\Sigma}^1_{(3)} = \begin{pmatrix} 126.37 & -35.73 & 2.36 \\ -35.73 & 622.89 & -1.48 \\ 2.36 & -1.48 & 0.49 \end{pmatrix}$$

Puis nous calculons la forme quadratique qui représente la statistique de test

$$\begin{aligned} W_{(3)} &= \begin{pmatrix} -0.126 & -0.064 & 1.779 \end{pmatrix} \times \begin{pmatrix} 126.37 & -35.73 & 2.36 \\ -35.73 & 622.89 & -1.48 \\ 2.36 & -1.48 & 0.49 \end{pmatrix} \times \begin{pmatrix} -0.126 \\ -0.064 \\ 1.779 \end{pmatrix} \\ &= 4.762 \end{aligned}$$

Avec un loi du χ^2 à 3 degrés de liberté, nous obtenons une p-value de 0.1900. Manifestement, au risque 10%, l'hypothèse nulle ne peut pas être rejetée. Le modèle n'est pas globalement significatif. Comme le précédent (tester simultanément "âge" et "taux max"), ce résultat contredit celui du rapport de vraisemblance.

3.3.6 Écriture générique des tests de significativité

Nous avions évoqué l'idée plus haut, tous les tests de significativité décrits dans ce chapitre peuvent s'écrire sous une forme générique :

$$H_0 : Ma = 0$$

où M est une matrice de dimension $[m \times (J+1)]$ de rang m ; a étant de dimension $(J+1) \times 1$, n'oublions pas la constante.

La statistique de test s'écrit alors ([23], page 421 ; voir [7], page 90 pour une écriture équivalente) :

$$W_{(M)} = \hat{a}' M' \left[M \hat{\Sigma} M' \right]^{-1} M\hat{a} \tag{3.7}$$

Elle suit une loi du χ^2 à m degrés de liberté.

Pour le fichier COEUR, voici l'écriture de la matrice M pour les différentes configurations.

Hypothèse nulle	Matrice M
$H_0 : a_1 = 0$	$\mathrm{M} = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}$
$H_0 : a_1 = a_2 = 0 \Leftrightarrow H_0 : \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\mathrm{M} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
$H_0 : a_1 = a_2 = a_3 = 0 \Leftrightarrow H_0 : \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\mathrm{M} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

Application au test $H_0 : a_1 = a_2 = 0$

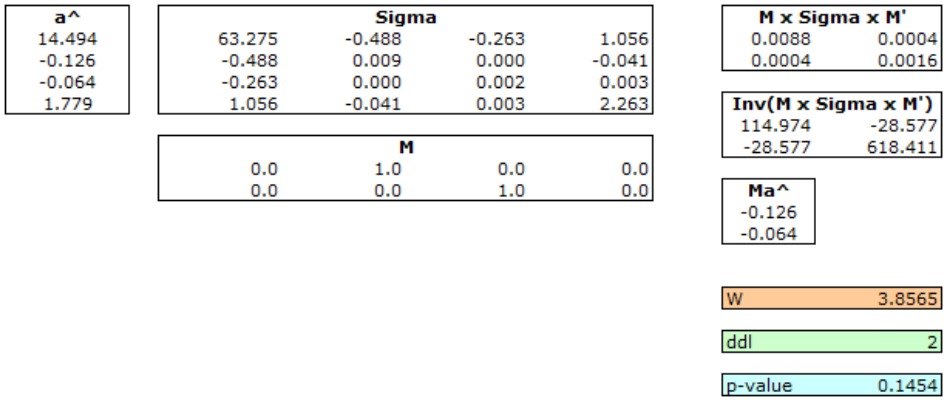


Fig. 3.7. COEUR - Test de Wald avec l'approche générique - $H_0 : a_1 = a_2 = 0$

Curieux comme nous sommes, voyons si les résultats concordent si nous utilisons la forme générique. Nous avons une feuille Excel dont voici la teneur (Figure 3.7) :

- La matrice M pour ce test s'écrit
$$M = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$
- La matrice $\hat{\Sigma}$ de variance covariance des coefficients est connue.
- Nous pouvons former
$$M\hat{\Sigma}M' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 63.2753 & -0.4882 & -0.2627 & 1.0563 \\ -0.4882 & 0.0088 & 0.0004 & -0.0413 \\ -0.2627 & 0.0004 & 0.0016 & 0.0030 \\ 1.0563 & -0.0413 & 0.0030 & 2.2634 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0.0088 & 0.0004 \\ 0.0004 & 0.0016 \end{pmatrix}$$
- Que nous inversons
$$(M\hat{\Sigma}M')^{-}1 = \begin{pmatrix} 114.974 & -28.577 \\ -28.577 & 618.411 \end{pmatrix}$$

- Nous disposons des paramètres estimés $\hat{a} = \begin{pmatrix} 14.494 \\ -0.0126 \\ -0.064 \\ 1.779 \end{pmatrix}$
- Nous calculons
$$M\hat{a} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 14.494 \\ -0.0126 \\ -0.064 \\ 1.779 \end{pmatrix} = \begin{pmatrix} -0.126 \\ -0.064 \end{pmatrix}$$
- Nous disposons de toutes les informations nécessaires à la formation de la statistique de test
$$W_{(M)} = \begin{pmatrix} -0.126 & -0.064 \end{pmatrix} \begin{pmatrix} 114.974 & -28.577 \\ -28.577 & 618.411 \end{pmatrix} \begin{pmatrix} -0.126 \\ -0.064 \end{pmatrix} = 3.8565$$
- Exactement la valeur obtenue avec la méthode directe (section 3.2.3).
- Le nombre de degré de liberté est $m = 2$ (nombre de lignes de la matrice M). Nous obtenons une p-value de 0.1454.

3.3.7 Aller plus loin avec la forme générique des tests

L'intérêt de la forme générique n'est pas que théorique. Certes, il est toujours plaisant de produire une écriture unique qui englobe toutes les autres. Mais elle nous permet surtout d'aller plus loin, de mettre en oeuvre des tests plus complexes.

Mettons que pour les mêmes données COEUR, nous souhaitons tester

$$H_0 : \begin{pmatrix} a_3 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1.5 \\ 2 \times a_2 \end{pmatrix}$$
H_1 : une des deux égalités au moins est fausse

L'enjeu est de savoir écrire correctement la matrice M. Reprenons les hypothèses

$$\begin{pmatrix} a_3 = 1.5 \\ a_1 = 2 \times a_2 \end{pmatrix} \Leftrightarrow \begin{pmatrix} a_3 = 1.5 \\ a_1 - 2 \times a_2 = 0 \end{pmatrix}$$
$$\Leftrightarrow \begin{pmatrix} -1.5 + 0 \times a_1 + 0 \times a_2 + 1 \times a_3 = 0 \\ 0 + 1 \times a_1 + (-2) \times a_2 + 0 \times a_3 = 0 \end{pmatrix}$$

Nous en déduisons facilement la matrice M

$$M = \begin{pmatrix} -1.5 & 0 & 0 & 1 \\ 0 & 1 & -2 & 0 \end{pmatrix}$$

Nous introduisons ces valeurs dans la feuille Excel (Figure 3.8). Le résultat est immédiat, nous obtenons $W_{(M)} = 2.8292$, avec une p-value de 0.2430 pour un $\chi^2(2)$. Au risque 5%, les données sont compatibles avec l'hypothèse nulle.

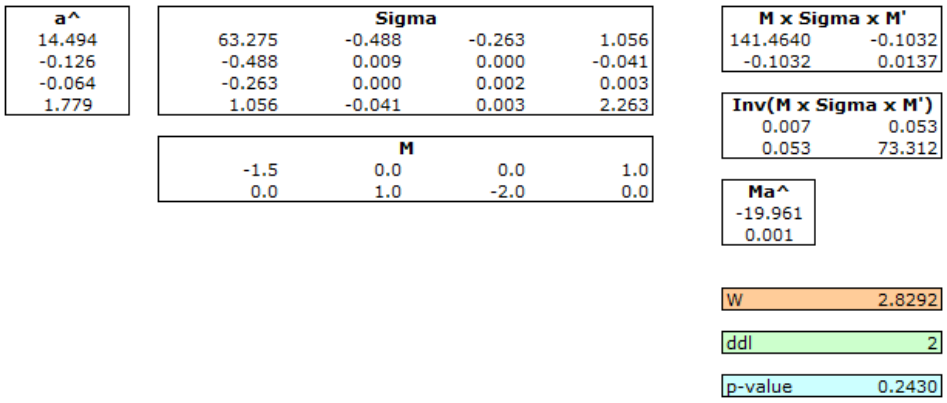


Fig. 3.8. COEUR - Test de Wald - $H_0 : a_3 = 1.5$ et $a_1 = 2 \times a_2$

3.4 Bilan : Rapport de vraisemblance ou Wald ?

Nous avons deux procédures pour les mêmes tests d'hypothèses : celle du rapport de vraisemblance et celle de Wald. Parfois elles se contredisent. C'est fâcheux. Récapitulons pour mémoire les résultats obtenus sur le fichier COEUR.

Tests à 10%	Rapp. de Vraisemblance	Test de Wald
Signif. "âge"	Accep. H_0, p-value = 0.1156	Accep. H_0, p-value = 0.1805
Signif. "âge" et "taux max"	Rejet H_0, p-value = 0.0771	Accep. H_0, p-value = 0.1454
Signif. globale	Rejet H_0, p-value = 0.0499	Accep. H_0, p-value = 0.1900

Nous retrouvons dans ces résultats les comportements que l'on attribue généralement à ces tests dans la littérature, à savoir :

1. Concernant le test du rapport de vraisemblance
   - Il est plus puissant. Il détecte mieux l'hypothèse alternative lorsque cela est justifié.
   - Il est revanche plus gourmand en ressources car il impose de recalculer le modèle sous la contrainte de l'hypothèse nulle. Encore une fois, le problème ne se pose véritablement que lorsque nous avons à traiter une grande base de données.
2. Concernant le test de Wald
   - Il est moins puissant, plus conservateur. Il favorise l'hypothèse nulle H_0. C'est flagrant dans nos résultats sur le fichier COEUR, H_0 n'a jamais été rejetée quel que soit le test mis en place.
   - Lorsque la valeur du coefficient est élevé, l'estimation de l'écart type gonfle exagérément. De nouveau H_0 est favorisé lors des tests individuels, cela nous emmène à supprimer à tort des variables importantes du modèle.
   - Il repose sur des propriétés asymptotiques de l'estimateur. Il est par conséquent peu précis lorsque nous traitons de petits effectifs comme c'est le cas pour le fichier COEUR.

– Accordons lui quand même une qualité, il est peu gourmand en ressources. Nous travaillons à partir des résultats fournis par la régression sur la totalité des variables, sans avoir à produire des calculs supplémentaires compliqués (une inversion de matrice quand même, ce n'est jamais anodin).

Pour mettre tout le monde d'accord, lorsque les effectifs sont importants, les deux procédures fournissent des résultats cohérents [7] (page 91).

Partie II

Pratique de la régression logistique binaire

4

Prédiction et intervalle de prédiction

Un des principaux objectifs de l'apprentissage supervisé est de fournir un système de classement qui, pour un nouvel individu quelconque ω' issu de la population (ex. un nouveau client pour une banque, un malade qui arrive au service des urgences, etc.), fournit une prédiction $\hat{y}(\omega')$. Avec exactitude si possible.

La régression logistique sait faire cela. Mais, à la différence d'autres méthodes, elle peut fournir en plus un indicateur de fiabilité de la prédiction avec une estimation de la probabilité $\hat{\pi}(\omega')$. Ainsi, lorsque $\hat{\pi}$ est proche de 1 ou de 0, la prédiction est plutôt sûre ; lorsqu'elle prend une valeur intermédiaire, proche du seuil d'affectation s ($s = 0.5$ habituellement), la prédiction est moins assurée. Dans les domaines où les conséquences des mauvaises affectations peuvent être dramatiques (dans le domaine de la santé par exemple), on pourrait même imaginer un système qui ne classe qu'à coup (presque) sûr du type :

- Si $\hat{\pi} \leq s_1$ Alors $\hat{y} = -$
- Si $\hat{\pi} \geq s_2$ Alors $\hat{y} = +$, avec $s_2 \gg s_1$ bien entendu.
- Sinon, indétermination. On demande des analyses complémentaires ou on présente le sujet à un expert.

Obtenir une estimation $\hat{\pi}$ et une indication sur sa précision nous est donc fort utile. Dans ce chapitre, nous montrons comment calculer $\hat{\pi}$ pour un nouvel individu à classer, puis nous étudierons la construction d'un intervalle (fourchette) de prédiction. Ce dernier point constitue aussi une avancée considérable par rapport aux d'autres méthodes supervisée. Nous disposons d'une indication sur la plage de valeurs crédibles de $\hat{\pi}$.

4.1 Prédiction ponctuelle

Pour obtenir une prédiction du LOGIT pour un nouvel individu ω' à classer, il nous suffit d'appliquer les coefficients estimés de la régression logistique, soit

$$\hat{c}(x(\omega')) = \hat{a}_0 + \hat{a}_1 \times x_1(\omega') + \cdots + \hat{a}_J \times x_J(\omega') \tag{4.1}$$

Si nous adoptons une écriture matricielle, avec $x(\omega') = (1, x_1(\omega')), \ldots, x_J(\omega'))$ la description de l'individu à classer et $\hat{a}' = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_J)$ le vecteur des paramètres estimés, nous écrivons

$$\hat{c}(x(\omega')) = x(\omega').\hat{a}$$

Pour alléger l'écriture, nous écrirons simplement $\hat{c}$ dans ce qui suit.

A partir du LOGIT, nous pouvons déduire une estimation de la probabilité a posteriori d'être positif de l'individu, soit

$$\hat{\pi}(\omega') = \frac{1}{1+e^{-\hat{c}}} \tag{4.2}$$

Et en appliquant la règle d'affectation standard, nous obtenons $\hat{y}$

$$\text{Si } \hat{\pi} > 0.5 \text{ alors } \hat{y} = + \text{ sinon } \hat{y} = - \tag{4.3}$$

Application aux données COEUR. Rappelons que le vecteur estimé des paramètres de la régression est $\hat{a}' = (14.4937, -0.1526, -0.0636, 1.7790)$ (Figure 1.5). Nous souhaitons classer un nouvel individu avec AGE = 35, TAUX MAX = 156, et ANGINE = 1. Nous réalisons la succession de calculs suivante :

- $\hat{c} = 14.4937 - 0.1526 \times 35 - 0.0636 \times 156 + 1.7790 \times 1 = 1.9601$
- $\hat{\pi} = \frac{1}{1+e^{-1.9601}} = 0.8765$
- $\hat{y} =$ présence

La prédiction est correcte. En effet il s'agit de l'individu $n^o 6$ dans notre tableau de données (Figure 0.1), il est positif ("présence").

4.2 Intervalle de prédiction

L'obtention des prédictions ponctuelles est assez facile finalement. Il n'y a pas à s'attarder dessus. Plus intéressant pour nous est la capacité de la régression logistique à produire un intervalle de variation pour $\pi(\omega')$. Et en matière de prévision, une fourchette est toujours plus utile qu'une valeur ponctuelle.

Pour construire l'intervalle de confiance du LOGIT, nous avons besoin de l'estimation de sa variance [9] (page 41). Elle s'écrit :

$$\hat{V}(\hat{c}) = \sum_{j=0}^{J} x_j^2 \hat{V}(\hat{a}_j) + \sum_{j=0}^{J} \sum_{k=j+1}^{J} 2x_j x_k \widehat{COV}(\hat{a}_j, \hat{a}_k) \tag{4.4}$$

Une écriture matricielle serait peut être plus simple

$$\hat{V}(\hat{c}) = x\hat{\Sigma}x' \tag{4.5}$$

On reconnaît dans l'expression ci-dessus l'estimation de la variance covariance des coefficients estimés. x est le vecteur de description de l'individu à classer.

L'intervalle de confiance du LOGIT au niveau $(1-\alpha)$ est défini par

$$\hat{c} \pm u_{1-\alpha/2} \times \hat{\sigma}_{\hat{c}} \tag{4.6}$$

où $u_{1-\alpha/2}$ est le fractile de la loi normale centrée et réduite ; $\hat{\sigma}_{\hat{c}} = \sqrt{\hat{V}(\hat{c})}$ est l'écart type du LOGIT.

Reprenons notre exemple COEUR ci-dessus. Nous souhaitons calculer l'intervalle de confiance de $\pi(\omega')$ au niveau $(1-\alpha) = 90\%$. D'ores et déjà, nous savons que $u_{0.95} = 1.6449$. Concernant la matrice de variance covariance des paramètres estimés, elle a déjà été calculée par ailleurs (section 3.3.1)

$$\hat{\Sigma} = \begin{pmatrix} 63.2753 & -0.4882 & -0.2627 & 1.0563 \\ -0.4882 & 0.0088 & 0.0004 & -0.0413 \\ -0.2627 & 0.0004 & 0.0016 & 0.0030 \\ 1.0563 & -0.0413 & 0.0030 & 2.2634 \end{pmatrix}$$

Pour calculer la variance du LOGIT, nous appliquons la formule 4.5 :

$$\hat{V}(\hat{c}) = \begin{pmatrix} 1 & 35 & 156 & 1 \end{pmatrix} \begin{pmatrix} 63.2753 & -0.4882 & -0.2627 & 1.0563 \\ -0.4882 & 0.0088 & 0.0004 & -0.0413 \\ -0.2627 & 0.0004 & 0.0016 & 0.0030 \\ 1.0563 & -0.0413 & 0.0030 & 2.2634 \end{pmatrix} \begin{pmatrix} 1 \\ 35 \\ 156 \\ 1 \end{pmatrix} = 4.5689$$

Et l'écart-type

$$\hat{\sigma}_{\hat{c}} = \sqrt{4.5689} = 2.1375$$

Nous pouvons produire les bornes basses (c_1) et hautes (c_2) du LOGIT pour l'individu à classer :

$$c_1 = 1.9601 - 1.6449 \times 2.1375 = -1.5557$$
$$c_2 = 1.9601 + 1.6449 \times 2.1375 = 5.4760$$

Nous en déduisons les bornes de l'intervalle de prédiction des probabilités a posteriori π

$$\pi_1 = \frac{1}{1 + e^{-(-1.5557)}} = 0.1743$$
$$\pi_2 = \frac{1}{1 + e^{-(5.4760)}} = 0.9958$$

Notre intervalle est très peu précis. Ce n'est guère étonnant. L'estimation des paramètres de la régression logistique repose sur un très petit échantillon ($n = 20$). Ce qui engendre une certaine instabilité traduite par des intervalles de confiance larges, que ce soit pour les estimations des coefficients (section 3.3.3) ou pour les prédictions.

Tous ces calculs ont été réalisés à l'aide d'une feuille Excel que nous reproduisons ici (Figure 4.1).

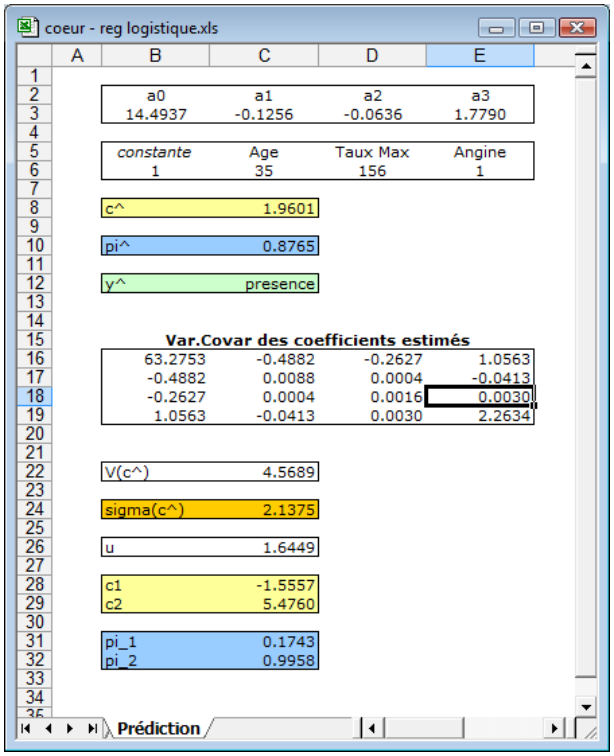


Fig. 4.1. COEUR - Calcul de l'intervalle de prédiction

5

Lecture et interprétation des coefficients

Dans certains domaines, l'explication est bien plus importante que la prédiction [1]. On souhaite comprendre les phénomènes de causalité, mettre à jour les relations de cause à effet. Bien entendu, les techniques statistiques n'ont pas vocation à répondre mécaniquement à des problèmes complexes. En revanche, elles ont pour rôle de donner aux experts les indications adéquates pour qu'ils puissent se concentrer sur les informations importantes. La régression logistique propose des outils qui permettent d'interpréter les résultats sous forme de risques, de chances, de rapports de chances. C'est certainement une des raisons pour laquelle elle a gagné les faveurs d'un large public d'utilisateurs. Un signe qui ne trompe pas, une large documentation est dédiée à l'interprétation des sorties de la régression logistique dans les ouvrages qui font référence ([9], chapitre 3 ; [10], chapitre 3).

5.1 Risque relatif, odds, odds-ratio

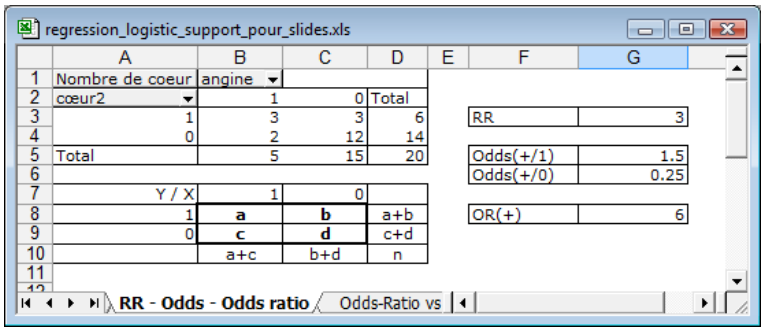


Fig. 5.1. Tableau de contingence - Croisement *coeur* vs. *angine*

Pour illustrer notre propos, nous utiliserons un tableau de contingence issu du fichier COEUR, il croise la variable dépendante *coeur* (avoir une maladie cardiaque ou pas +/-) avec la variable explicative

1. Par exemple, dans le domaine de la santé, on cherche certes à détecter automatiquement les personnes qui développent une maladie particulière, mais il est peut être plus important que l'on comprenne pourquoi ils la développent pour qu'on puisse l'anticiper. On distingue ainsi l'analyse "pronostic" à visée prédictive de l'analyse "étiologique" à visée explicative.

angine (groupe "exposé" vs. groupe "témoin" 1/0). Nous pouvons construire le tableau parce que les deux variables ne sont pas continues. Nous adjoignons à la copie d'écran les notations que nous utiliserons par la suite (Figure 5.1) [2].

Quelques définitions

Risque relatif. On appelle risque relatif le surcroît de chances d'être positif du groupe exposé par rapport au groupe témoin.

$$\begin{aligned} RR &= \frac{P(+/1)}{P(+/0)} \\ &= \frac{a/(a+c)}{b/(b+d)} \\ &= \frac{3/5}{3/15} \\ &= 3 \end{aligned}$$

Nous l'interprétons de la manière suivante : les personnes qui ont une angine de poitrine ont 3 fois plus de chances que les autres (ceux qui n'en ont pas) de développer une maladie cardiaque. Il caractérise un lien entre l'apparition de la maladie et l'occurrence de l'angine de poitrine. Lorsque $RR = 1$, cela veut dire que l'angine n'a pas d'incidence sur la maladie.

Odds. L'odds ou *rapport de chances* est défini comme un rapport de probabilités dans un groupe. Par exemple, dans le groupe exposé, il s'écrit

$$\begin{aligned} odds(1) &= \frac{P(+/1)}{P(-/1)} \\ &= \frac{a/(a+c)}{c/(a+c)} \\ &= \frac{3/5}{2/5} \\ &= 1.5 \end{aligned}$$

Dans le groupe des personnes ayant une angine de poitrine, on a 1.5 fois plus de chances d'avoir une maladie cardiaque que de ne pas en avoir. Nous pouvons de la même manière définir l'odds dans le groupe témoin $odds(0)$.

Odds-ratio. L'odds ratio est égal au rapport entre l'odds du groupe exposé et l'odds du groupe témoin.

2. Pour une étude approfondie des indicateurs présentés dans cette section, notamment les définitions, les estimations, les tests de significativité et les intervalles de confiance, voir [20], chapitre 5, pages 49 à 62.

$$\begin{aligned} OR &= \frac{odds(1)}{odds(0)} \\ &= \frac{\frac{a/(a+c)}{c/(a+c)}}{\frac{b/(b+d)}{d/(b+d)}} \\ &= \frac{a \times d}{b \times c} \\ &= \frac{3 \times 10}{3 \times 2} \\ &= 6 \end{aligned}$$

L'OR indique à peu près la même chose que le risque relatif, à savoir : dans le groupe exposé, on a 6 fois plus de chances d'avoir la maladie que dans le groupe témoin.

Il est toujours un peu gênant d'avoir deux formulations, avec des valeurs différentes, pour le même concept. A priori, le risque relatif est l'indicateur le plus simple à appréhender. Pourtant, on lui préfère souvent l'odds-ratio, principalement pour 2 raisons :

1. La prévalence, la probabilité a priori p d'être positif, est souvent très faible dans les études réelles. Les malades sont rares, les fraudeurs ne sont pas légion, etc. Dans ce cas, l'odds-ratio et le risque relatif prennent des valeurs similaires. En effet, lorsque $a \ll c$ alors $a + c \approx c$; de même, lorsque $b \ll d$ alors $b + d \approx d$. Par conséquent
$$RR = \frac{a/(a+c)}{b/(b+d)} \approx \frac{a/c}{b/d} = \frac{a \times d}{b \times c} = OR$$

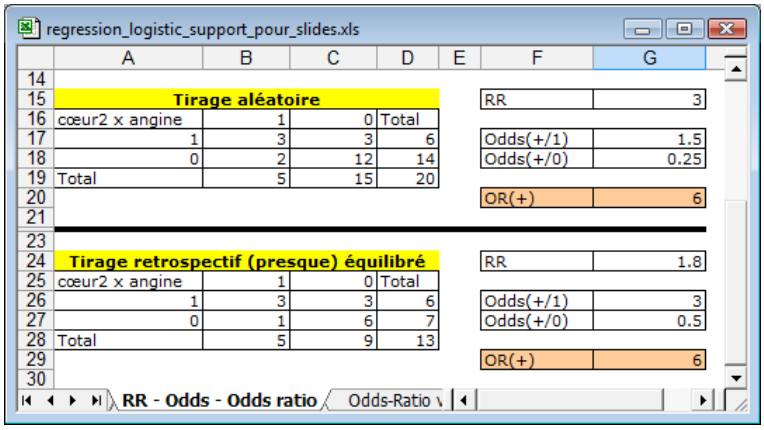


Fig. 5.2. Odds ratio et mode d'échantillonnage

2. L'odds-ratio possède une propriété très précieuse, il est invariant par rapport au mode d'échantillonnage. Que l'on procède à un tirage aléatoire simple des données (schéma de mélange) ou à un tirage rétrospectif, il présentera toujours la même valeur. Voyons un exemple pour nous en persuader (Figure 5.2). Dans le premier cas (celui du haut), l'échantillon a été tiré au hasard, nous obtenons les valeurs $RR = 3$ et $OR = 6$. Dans le second cas (celui du bas), nous avons un tirage (presque) équilibré. Nous avons choisi $n_+ = 6$ individus au hasard parmi les positifs, $n_- = 7$ parmi les négatifs. En calculant de nouveau nos indicateurs, nous avons $RR = 1.8$ et $OR = 6$. L'OR prend la même valeur que

précédemment, le RR a été modifié. Dans les applications réelles, cette propriété est essentielle. L'OR nous évite d'avoir à procéder à des redressements toujours compliqués. Surtout que, souvent, nous avons peu d'informations sur la prévalence réelle p (qui pourrait nous dire le véritable pourcentage des fraudeurs ?).

Log odds-ratio. Il s'agit simplement du logarithme de l'odds-ratio. Développons son expression, nous verrons ainsi le rapport avec la régression logistique.

$$\begin{aligned}
\ln(OR) &= \ln \frac{odds(1)}{odds(0)} \\
&= \ln(odds(1)) - \ln(odds(0)) \\
&= \ln \frac{P(Y=+/1)}{P(Y=-/1)} - \ln \frac{P(Y=+/0)}{P(Y=-/0)} \\
&= \ln \frac{P(Y=+/1)}{1-P(Y=+/1)} - \ln \frac{P(Y=+/0)}{1-P(Y=+/0)} \\
&= LOGIT(1) - LOGIT(0)
\end{aligned}$$

D'ores et déjà, sans rentrer dans les détails, on constate que le log-odds ratio peut s'interpréter comme un écart entre 2 LOGIT. Nous garderons à l'esprit cette idée dans tout ce qui suit.

5.2 Le cas de la régression simple

Dans cette section, nous étudions le cas de la régression simple, avec un LOGIT de la forme

$$LOGIT = a_0 + a_1 \times X \tag{5.1}$$

L'interprétation des coefficients dépend du type de la variable explicative X.

5.2.1 Variable explicative binaire

Le cas de la variable explicative binaire est en relation directe avec le tableau de contingence que nous avons utilisé pour présenter l'odds-ratio (section 5.1). Dans cette configuration, le coefficient a_1 correspond au logarithme de l'odds-ratio calculé à partir du tableau de contingence ([9], pages 49 et 50 ; [11], pages 86 à 88).

L'idée est relativement simple :

$X = 1 \rightarrow LOGIT(1) = a_0 + a_1 \times 1 = a_0 + a_1$
$X = 0 \rightarrow LOGIT(0) = a_0 + a_1 \times 0 = a_0$
$\Rightarrow \ln(OR) = LOGIT(1) - LOGIT(0) = a_1$
$\Rightarrow OR = e^{a_1}$

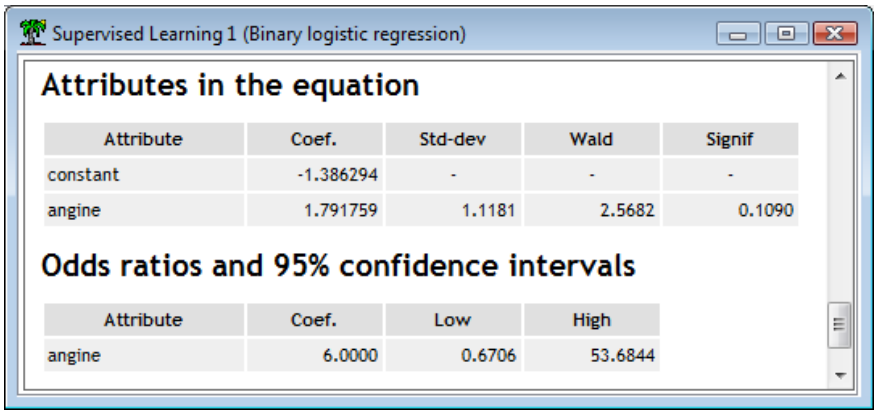


Fig. 5.3. Coefficients de la régression logistique - COEUR = f(ANGINE)

coeur = f (angine)

Reprenons notre exemple croisant *coeur* et *angine*, l'odds-ratio était égal à $OR = 6$ (Figure 5.1). Maintenant, nous réalisons une régression logistique expliquant *coeur* avec *angine* comme seule variable explicative à l'aide du logiciel Tanagra (Figure 5.3).

Nous obtenons $\hat{a}_1 = 1.791759$. En prenant l'exponentielle, nous obtenons $OR(angine) = e^{1.791759} = 6$. Ainsi, la régression logistique nous permet de mesurer directement le surcroît de risque associé à un facteur explicatif binaire. Si $\hat{a}_j < 0 \rightarrow OR < 1$, il y a une diminution du risque ; si $\hat{a}_j > 0 \rightarrow OR > 1$, il y a une augmentation.

Nous pouvons nous appuyer sur le mécanisme de formation des intervalles de confiance des coefficients (section 3.3.3) pour obtenir ceux des odds-ratios. La grande majorité des logiciels fournissent automatiquement ce type de résultat (Figure 5.3, avec un niveau de confiance fixé automatiquement à 95%).

Détaillons les calculs puisque nous disposons de l'estimation du coefficient et de son écart type. Pour un intervalle à 95%, le fractile de la loi normale utilisée est $u_{0.975} = 1.96$. Nous produisons les bornes de la manière suivante :

1. Borne basse

$$bb(a_1) = \hat{a_1} - u_{0.975} \times \hat{\sigma}_{\hat{a_1}}$$
$$= 1.791759 - 1.96 \times 1.1181 = -0.399$$

La borne basse de l'intervalle de variation de l'odds-ratio s'obtient avec $bb(OR) = e^{-0.399} = 0.67$

2. Borne haute

$$bh(a_1) = \hat{a_1} + u_{0.975} \times \hat{\sigma}_{\hat{a_1}}$$
$$= 1.791759 + 1.96 \times 1.1181 = 3.983$$

Et la borne haute de l'intervalle de variation de l'odds-ratio : $bh(OR) = e^{3.983} = 53.68$

Lorsque l'intervalle de variation de l'odds-ratio couvre la valeur 1, ou de manière équivalente lorsque l'intervalle du coefficient couvre la valeur 0, il n'y a pas de lien significatif entre la variable explicative et la variable dépendante.

Calcul direct de l'intervalle de variation du log odds-ratio

Nous pouvons, à partir des données observées dans le tableau de contingence, obtenir l'intervalle de confiance du log odds-ratio, sans passer par une régression logistique. Au niveau de confiance $(1-\alpha)$, il s'écrit (voir [20], page 56)

$$\ln(OR) \pm u_{1-\alpha/2} \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \tag{5.2}$$

Les résultats concordent avec ceux obtenus à l'aide de la régression logistique. En effet, lorsque nous calculons la quantité $\left(\sqrt{\frac{1}{3}+\frac{1}{3}+\frac{1}{2}+\frac{1}{12}} = \sqrt{1.25} = 1.1181\right)$ à partir des données du tableau de contingence (Figure 5.1), nous retrouvons la valeur $\hat{\sigma}_{\hat{a}_1} = 1.1181$ de l'écart-type du coefficient obtenue lors de la régression (Figure 5.3).

La constante a_0

Nous savons lire le coefficient a_1, qu'en est-il de la constante ? $X = 0$ est la catégorie de référence, le groupe témoin. Dans notre exemple, il s'agit des individus qui n'ont pas une angine de poitrine. Le LOGIT associé au cas $X = 0$ s'écrit :

$$LOGIT(0) = a_0 + a_1 \times 0 = a_0$$

Développons l'expression :

$$\begin{aligned} a_0 &= LOGIT(0) \\ &= \ln \frac{P(Y=+/0)}{1-P(Y=+/0)} \\ &= \ln \frac{P(Y=+/0)}{P(Y=-/0)} \\ &= \ln[odds(0)] \end{aligned}$$

Ainsi, la constante a_0 s'interprète comme le logarithme de l'odds (**log-odds**) dans la catégorie de référence. En passant à l'exponentielle, nous avons $odds(0) = e^{a_0}$.

Pour le fichier COEUR, à partir de notre tableau de contingence (Figure 5.1, nous pouvons former

$$odds(0) = \frac{3/15}{12/15} = \frac{3}{12} = 0.25$$

Si nous prenons cette fois-ci les résultats de la régression logistique (Figure 5.3), nous trouvons $\hat{a}_0 = -1.386294$. Et en passant à l'exponentielle :

$$e^{\hat{a}_0} = e^{-1.386294} = 0.25$$

CQFD.

5.2.2 Variable explicative quantitative

Pour comprendre l'interprétation des coefficients dans le cas d'une variable explicative quantitative, voyons l'évolution du LOGIT lorsqu'on fait varier X d'une unité.

$$LOGIT(X+1) = a_0 + a_1 \times (X+1) = a_0 + a_1 \times X + a_1$$
$$LOGIT(X) = a_0 + a_1 \times X$$
$$\Rightarrow LOGIT(X+1) - LOGIT(X) = a_1$$

Dans ce cas, la quantité e^{a_1} s'interprète comme l'odds ratio consécutif à l'augmentation d'une unité de la variable explicative. Nous formulerons quelques remarques :

- Si l'on augmente de b unités la variable explicative, l'odds-ratio devient alors $e^{b \times a_1}$.
- L'intervalle de variation du log odds-ratio lorsque l'on augmente de b unités la variable X s'écrit

$$b \times \hat{a}_1 \pm u_{1-\alpha/2} \times b \times \hat{\sigma}_{\hat{a}_1}$$

- Attention, la valeur de l'odds-ratio dépend de l'unité de mesure utilisée. Prenons l'âge, si on la mesure en mois au lieu d'années, une variation d'une unité n'a pas le même effet sur la variable dépendante. Ce qui paraît assez normal.
- L'outil doit être manipulé avec une grande prudence. *On suppose que le LOGIT est linéaire par rapport à la variable explicative*. C'est une hypothèse un peu forte. Prenons un exemple simple. Si l'on veut étudier le risque d'apparition d'une maladie cardiaque, il est évident que passer de 10 ans à 20 ans n'a pas la même signification que de passer de 40 ans à 50 ans. *Il faut rester raisonnable dans les effets que l'on souhaite tester* [9] (page 63).

coeur = f (taux max)

Nous essayons de prédire COEUR en fonction de TAUX MAX. Nous réalisons la régression logistique, nous obtenons les coefficients estimés (Figure 5.4).

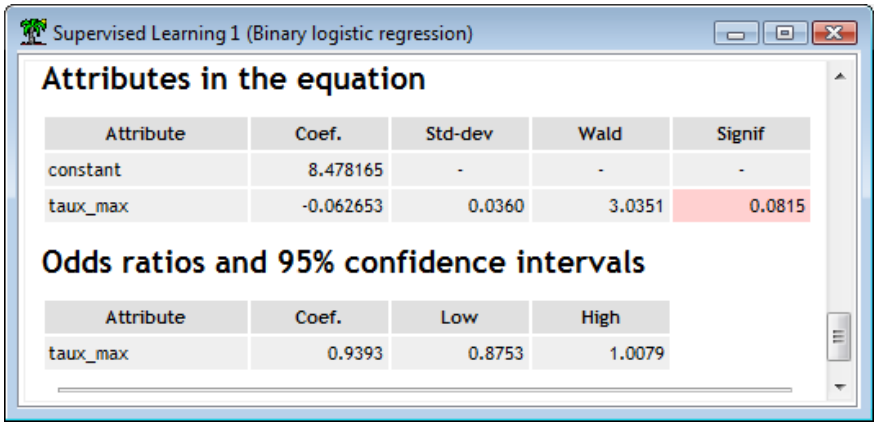


Fig. 5.4. Coefficients de la régression logistique - COEUR = f(TAUX MAX)

Nous obtenons $\hat{a}_1 = -0.062653$, et par conséquent $e^{\hat{a}_1} = 0.9393$. Lorsque le taux max augmente d'une unité, les individus ont $\frac{1}{0.9393} = 1.0646$ fois plus de chances de **ne pas** développer une maladie cardiaque.

La constante a_0

Ici également, la constante peut être comprise comme le log-odds lorsque X prend la valeur de référence $X = 0$. Dans notre exemple $coeur = f(taux\ max)$ c'est un peu gênant. En effet, lorsque $taux\ max = 0$, cela veut simplement dire que la personne est morte, son coeur ne bat plus. Nous avons donc tout intérêt à centrer la variable pour obtenir une interprétation plus séduisante de la constante. C'est ce que nous avons fait, nous avons relancé la régression logistique (Figure 5.5).

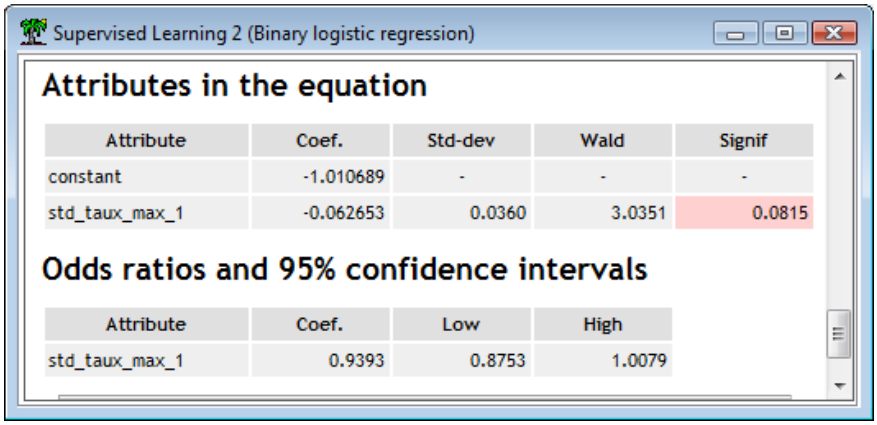


Fig. 5.5. Coefficients de la régression logistique - COEUR = f(TAUX MAX centré)

Premier constat, l'estimation de la pente $\hat{a}_1$ n'a pas été modifiée. C'était attendu. L'odds ratio dépend uniquement des variations de X et non pas de la valeur de X. Que l'origine du repère soit 0 ou la moyenne, cela ne change rien à l'affaire.

Second constat, la constante $\hat{a}_0$ est, elle, tout à fait différente. Voyons comment nous pouvons la lire :

$$e^{-1.010689} = 0.3640$$

Une personne présentant un "taux max" moyen (dont le "taux max" est égal à la moyenne de la population) a $\frac{1}{0.3640} = 2.7475$ fois plus de chances d'être non malade (que d'être malade).

5.2.3 Variable explicative qualitative nominale

Calcul de l'odds-ratio à partir d'un tableau de contingence

Il n'est pas possible d'introduire directement une variable qualitative à $L(L > 2)$ modalités dans la régression logistique. Il faut la recoder. Du choix de codage dépend l'interprétation des coefficients.

Nous traitons un nouveau fichier de $n = 209$ observations dans cette section. La variable dépendante est toujours la présence/absence d'une maladie cardiaque (coeur). La variable explicative est "chest pain" (type douleur dans la poitrine) avec 4 modalités : "typ. angina" (code 1), "atyp. angina" (2), "asympt." (3) et "non anginal" (4).

Former le tableau de contingence ne pose pas de problèmes particuliers. Il en est de même lors du calcul des odds. Pour obtenir **les odds-ratio** en revanche, nous devons définir la catégorie de référence. Ils **seront alors définis en opposition à cette situation de référence**.

Attention, le choix de la modalité de référence est crucial pour l'interprétation. Il ne peut pas être dissocié de l'analyse qualitative des résultats que l'on veut mener par la suite. Dans notre exemple, admettons qu'il s'agisse de la dernière (non anginal - code 4). Nous aurons à calculer $L-1=4-1=3$ odds-ratio. Nous résumons cela dans une feuille de calcul (Figure 5.6).

Calcul direct dans un tableau croisé

Nombre de cœur	chest_pain				
cœur	typ_angina	atyp_angina	asympt	non_anginal	Total
presence	4	6	75	7	92
absence	2	59	27	29	117
Total	6	65	102	36	209

Odds(+/-)	2.000	0.102	2.778	0.241
OR(x/ non_anginal)	8.286	0.421	11.508	

Fig. 5.6. Calcul des odds-ratio dans un tableau de contingence - Variable qualitative nominale

L'odds de la catégorie 1 est obtenue avec $odds(1)=\frac{4}{2}=2.0$: les personnes présentant une douleur de type "typ. angina" ont 2.0 fois plus de chances d'avoir une maladie cardiaque (que de ne pas en avoir). De même pour les autres catégorie, nous pouvons calculer : $odds(2)=\frac{6}{59}=0.102$; $odds(3)=\frac{75}{27}=2.778$; et $odds(4)=\frac{7}{29}=0.241$.

La 4^{eme} catégorie représentant la situation de référence, nous calculons les 3 odds-ratio en l'opposant aux autres c.-à-d. $OR(1/4)=\frac{odds(1)}{odds(4)}=\frac{2.0}{0.241}=8.286$, nous le lisons de la manière suivante "les personnes qui ont une douleur dans la poitrine de type *typ. angina* ont 8.286 fois plus de chances de développer une maladie cardiaque que ceux qui présentent une douleur de type *non anginal*" ; de même, nous pouvons produire $OR(2/4)=\frac{0.102}{0.241}=0.421$ et $OR(3/4)=\frac{2.778}{0.241}=11.508$.

Obtenir les odds-ratio à l'aide de la régression logistique

La question que l'on se pose maintenant est "comment obtenir les mêmes valeurs à partir de la régression logistique ?". Pour y répondre, nous devons poser une autre question "comment coder la variable catégorielle pour que la régression logistique produise les mêmes odds-ratio ?".

La solution repose sur un codage $0/1$ de chacune des modalités de la variable catégorielle, en excluant la modalité de référence. Si X est la variable catégorielle initiale, nous en tirons donc 3 nouvelles variables binaires X_1, X_2, X_3 avec :

- $X_1(\omega)=1$ si $X(\omega)=$"$typ.angina$", 0 sinon
- $X_2(\omega)=1$ si $X(\omega)=$"$atyp.angina$", 0 sinon
- $X_3(\omega)=1$ si $X(\omega)=$"$asympt$", 0 sinon

– par conséquent, si $X(\omega) =$ "*non anginal*", alors $X_1(\omega) = X_2(\omega) = X_3(\omega) = 0$. Nous avons l'information adéquate, il n'est pas nécessaire de créer une variable X_4 pour la 4^{eme} modalité. Elle devient la modalité de référence.

Nous montrons une copie d'écran des 15 premières observations (Figure 5.7).

chest_pain	typ_angina	atyp_angina	asympt
asympt	0	0	1
atyp_angina	0	1	0
_non_anginal	0	0	0
_non_anginal	0	0	0
asympt	0	0	1
asympt	0	0	1
asympt	0	0	1
asympt	0	0	1
asympt	0	0	1
asympt	0	0	1
_non_anginal	0	0	0
asympt	0	0	1
atyp_angina	0	1	0
atyp_angina	0	1	0
asympt	0	0	1

Fig. 5.7. Codage 0/1 - X = CHEST PAIN vs. X_1 = TYP ANGINA, X_2 = ATYP ANGINA, X_3 = ASYMPT

Nous avons réalisé la régression logistique avec ces 3 nouvelles variables c.-à-d. $Y = f(X_1, X_2, X_3)$, Tanagra nous fournit une série de résultats (Figure 5.8) :

– La régression est globalement significative. Le test du rapport de vraisemblance montre que les coefficients relatifs aux variables (a_1, a_2, a_3) ne sont pas tous simultanément nuls ($\chi^2 = 85.7164$, et p-value < 0.0001). Un des coefficients au moins est significativement différent de 0.

– Comme il n'y a que les variables recodées 0/1 de CHEST PAIN dans notre modèle, cela indique (1) que CHEST PAIN a une incidence sur l'apparition de la maladie cardiaque ; (2) qu'il y a un surcroît (ou réduction) de risque significatif associé à au moins une des 3 modalités, par rapport à la modalité de référence NON ANGINAL.

– Voyons le détail des coefficients justement (nous signalons par un astérisque les coefficients significatifs à 5%) :

j	$\hat{a}_j$	Wald	p-value	$OR(j/4) = e^{\hat{a}_j}$
1	2.114534	4.8216	0.0281*	8.286
2	−0.864392	2.0700	0.1502	0.421
3	2.443037	26.2102	0.0000*	11.508

– Nous retrouvons les valeurs des odds-ratio calculées à partir du tableau de contingence (Figure 5.6).

– De plus, nous savons maintenant quelles sont les situations où les surcroîts (réductions) de risques sont significatifs. En effet, si le coefficient est significativement différent de 0, l'odds-ratio qui en est dérivé est significativement différent de 1. Nous n'avions pas cette information auparavant. Dans notre tableau ci-dessus, nous constatons que TYP ANGINA et ASYMPT se démarquent significativement de la situation de référence NON ANGINAL, pas ATYP ANGINA en revanche.

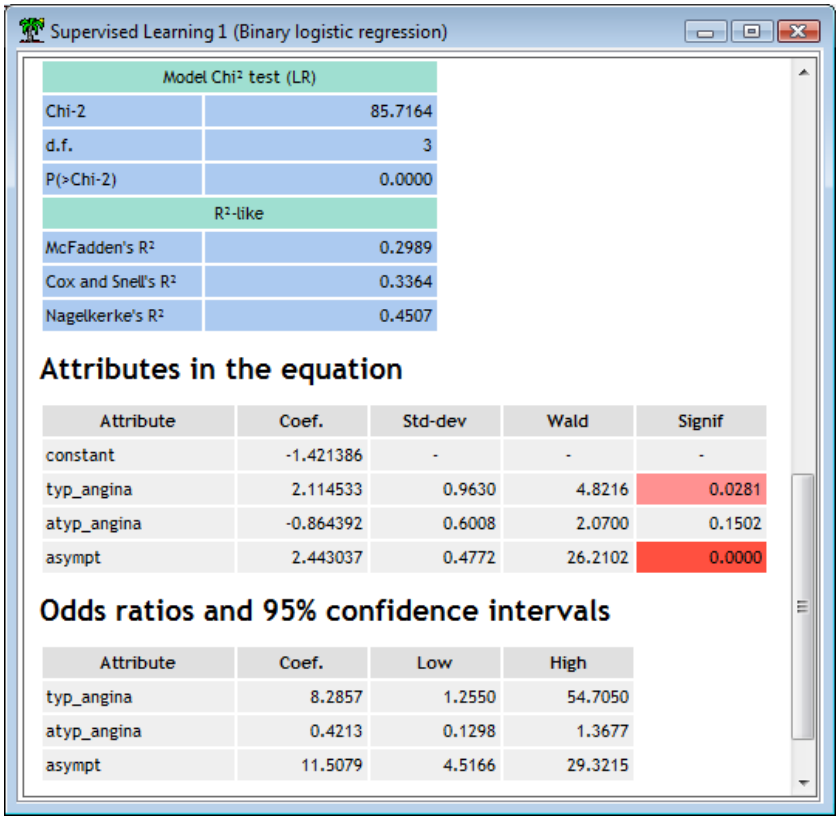


Fig. 5.8. Régression sur les variables explicatives codées 0/1

– Ceci est confirmé par le tableau des intervalles de confiance à 95% fourni par Tanagra dans la partie basse de la fenêtre de résultats (Figure 5.8). L'odds-ratio est considéré significatif si l'intervalle ne contient pas la valeur 1.

En conclusion, lorsque nous procédons à un codage 0/1 simple, **les coefficients de la régression logistique correspondent à des log odds-ratio** de chaque modalité par rapport à la modalité de référence, celle qui a été exclue lors du recodage.

La constante a_0

A l'instar de la variable explicative binaire, la constante s'interprète comme le log-odds de la situation de référence (groupe témoin). Dans notre exemple (Figure 5.8), $\hat{a}_0 = -1.421386$. Lorsque nous passons à l'exponentielle, nous obtenons $e^{-1.421386} = 0.241$, qui est bien la valeur de l'$odds(4)$ obtenu à partir du tableau de contingence (Figure 5.6).

Exclure tout ou partie des indicatrices ?

Face à ce type de résultat (Figure 5.8), le praticien est parfois perplexe. Que faire ? Exclure l'indicatrice ATYP ANGINA parce qu'elle n'est pas significative ? La conserver parce que les deux autres le sont ? La situation sera d'autant plus compliquée que nous travaillons sur une régression multiple.

On conseille généralement de traiter les indicatrices d'une variable nominale comme un groupe, elles ne doivent pas être dissociées. Nous devons travailler en deux temps : (1) tester si les coefficients des indicatrices sont simultanément nuls, nous évaluons l'impact de la variable nominale sur la variable dépendante ; (2) une fois acquise la significativité globale, regarder les modalités qui s'écartent de la situation de référence [10] (page 60).

Un autre point de vue peut être défendu. Nous pouvons traiter individuellement les indicatrices. L'important est de bien en mesurer les conséquences sur l'interprétation des résultats. Si nous retirons uniquement l'indicatrice ATYP ANGINA du modèle, et conservons les deux autres, cela veut dire que la situation de référence est maintenant composée des deux modalités {NON ANGINAL et ATYP ANGINA}. Les coefficients des autres indicatrices s'interprètent comme des log odds-ratio par rapport à cette nouvelle catégorie témoin. Dans notre tableau de contingence (Figure 5.6), cela revient à créer une nouvelle colonne de référence qui serait le fruit de la fusion des colonnes ATYP ANGINA et NON ANGINAL.

Calcul direct dans un tableau croisé

	chest_pain			
cœur	typ_angina	asympt	{atyp angina, non anginal}	Total
presence	4	75	13	92
absence	2	27	88	117
Total	6	102	101	209

Odds	2.000	2.778	0.148
Odds-ratio	13.54	18.80	

Résultat de la régression logistique

OR(Reg.Logistic)	13.54	18.80	0.000

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.912387	-	-	-
typ_angina	2.605535	0.9156	8.0985	0.0044
asympt	2.934039	0.3724	62.0852	0.0000

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
typ_angina	13.5385	2.2503	81.4531
asympt	18.8034	9.0631	39.0119

Fig. 5.9. Régression sur les variables codées 0/1 - Redéfinition de la modalité de référence

A titre de vérification, nous avons calculé les odds-ratio dans le tableau de contingence après fusion des modalités ATYP ANGINA et NON ANGINAL. Nous avons dans le même temps calculé la régression sur les indicatrices TYP ANGINA et ASYMPT (Figure 5.9). Les résultats concordent, fort heureusement. Nous noterons surtout que les odds-ratios obtenus sont plus élevés : la création de la nouvelle situation de référence a permis de mieux caractériser le décalage entre les modalités.

5.2.4 Variable explicative qualitative ordinale

La variable explicative est qualitative mais les $L(L > 2)$ modalités sont ordonnées. Bien évidemment, il faut coder les variables. Il nous faut produire un codage qui sache tenir compte de l'ordre des modalités, sans pour autant introduire une fausse information sur l'amplitude des écarts. Nous reviendrons plus longuement sur cet aspect plus loin. Parmi les différentes stratégies possibles, nous présentons dans cette section le codage 0/1 emboîté ([11], page 92 ; le codage par "polynômes orthogonaux" est l'autre approche proposée par l'auteur).

Quel type d'odds-ratio peut-on produire ?

Dans un premier temps, travaillons toujours à partir d'un tableau de contingence. Notre fichier comporte $n = 209$ observations. La variable à prédire est l'occurrence ou non d'une maladie cardiaque, la variable explicative cette fois-ci est SYSTOLIC avec 3 niveaux : normal (1), élevé (2) et très élevé (3). Nous la croisons avec la variable dépendante, puis nous calculons les odds et les odds-ratio (Figure 5.10).

Calcul sur un tableau de contingence

Nombre de coeur	systolic_level			
coeur	3	2	1	Total
presence	14	31	47	92
absence	10	36	71	117
Total	24	67	118	209

Odds	1.400	0.861	0.662
Odds-Ratio(précédent)	**1.626**	**1.301**	

Fig. 5.10. Calcul des odds-ratio dans un tableau de contingence - Variable qualitative ordinale

Les odds sont simples à calculer : $odds(1) = \frac{47}{71} = 0.662$; $odds(2) = \frac{31}{36} = 0.861$; $odds(3) = \frac{14}{10} = 1.400$. L'interprétation est toujours la même, par exemple, $odds(3) = 1.4$ signifie qu'on a 1.4 fois plus de chances d'avoir une maladie cardiaque (que de ne pas en avoir) lorsqu'on a un SYSTOLIC de niveau TRES ELEVE.

Venons-en à l'odds-ratio maintenant. Dans le cas des variables ordinales, il se calcule par rapport à la modalité précédente. On quantifie le surcroît de risque lors du passage d'un niveau au suivant. Nous n'avons pas à le calculer pour NORMAL puisque c'est la modalité la plus basse. En revanche, pour le passage de NORMAL à ELEVE, nous pouvons produire $OR(2/1) = \frac{odds(2)}{odds(1)} = \frac{0.861}{0.662} = 1.301$. Nous l'interprétons ainsi : en passant du SYSTOLIC NORMAL vers le niveau ELEVE, les individus ont 1.301 fois plus de chances de développer une maladie cardiaque. De la même manière, pour le passage de ELEVE à TRES ELEVE, nous calculons $OR(3/2) = \frac{1.4}{0.662} = 1.626$. Le gap paraît plus important.

Dans le cas des variables ordinales, la modalité de référence est tout simplement la précédente. Nous quantifions le surcroît de risque consécutif à un changement de niveau.

Obtenir les odds-ratio à partir de la régression logistique

Comment coder la variable explicative pour obtenir les mêmes résultats à l'aide de la régression logistique ? La solution la plus simple est d'utiliser le codage 0/1 emboîté. Les coefficients issus de la régression logistique correspondent alors aux log-odds ratio d'un passage d'une modalité à une autre.

Reprenons notre exemple SYSTOLIC pour illustrer le codage emboîté. La variable X possède 3 modalités, nous en dérivons 2 nouvelles variables X_2 et X_3 définies de la manière suivante :

- Si $X(\omega) = 1$ alors $X_2(\omega) = 0$ et $X_3(\omega) = 0$
- Si $X(\omega) = 2$ alors $X_2(\omega) = 1$ et $X_3(\omega) = 0$
- Si $X(\omega) = 3$ alors $X_2(\omega) = 1$ et $X_3(\omega) = 1$ (!!!). L'astuce est ici. Pour avoir le niveau 3, il faut être passé par le niveau 2.

systolic_level	sys2	sys3
2	1	0
1	0	0
3	1	1
3	1	1
2	1	0
2	1	0
2	1	0
3	1	1
1	0	0
3	1	1
2	1	0
1	0	0
3	1	1
2	1	0
1	0	0

Fig. 5.11. Codage de SYSTOLIC - variable qualitative ordinale - en 2 variables 0/1 imbriquées SYS2 et SYS3

Voyons ce qu'il en est sur les 15 premières observations de notre fichier (Figure 5.11). Les colonnes SYS2 et SYS3 sont indissociables, elle permettent de reconstituer parfaitement la variable SYSTOLIC LEVEL.

Nous pouvons lancer la régression logistique. Nous obtenons une série de résultats (Figure 5.12) :

- $\hat{a}_{sys2} = 0.263001$ et $e^{0.263001} = 1.3008 = OR(1/2)$. Nous retrouvons l'odds-ratio du passage du niveau 1 au niveau 2.
- De même, $\hat{a}_{sys3} = 0.486004$ et $e^{0.486004} = 1.6258 = OR(3/2)$.
- Nous constatons avec la régression qu'aucun des deux odds-ratio n'est significativement différent de 1, via le test de Wald pour les coefficients ou via les intervalles de variation des odds-ratio.

La constante a_0

Dans cette configuration, la constante a_0 s'interprète comme le log-odds de la première modalité de la variable explicative ordinale. Voyons cela sur notre exemple : $\hat{a}_0 = -0.412532$ et $e^{-0.412532} = 0.662 = odds(1)$.

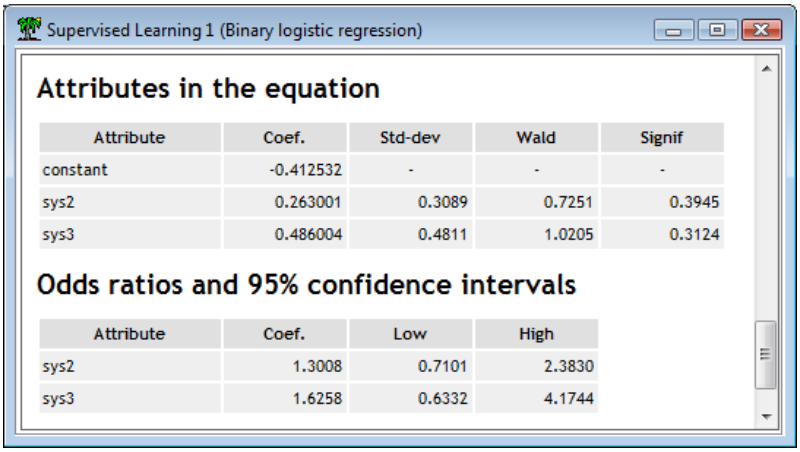


Fig. 5.12. Régression sur 2 variables issues d'un codage 0/1 emboîté

Une erreur (?) fréquente : le codage $\{1, 2, 3, \ldots\}$

Une erreur (?) fréquente est de travailler directement sur la variable explicative ordinale codée $\{1, 2, 3, \ldots, L\}$ c.-à-d. de l'introduire comme une variable quantitative dans la régression. Ce faisant, nous sommes en train d'indiquer à la technique statistique que les écarts entre les modalités sont identiques. En réalité, nous n'en savons rien. Si c'est effectivement le cas, le codage suggéré tient la route. Sinon, nous sommes en train d'induire la régression logistique en erreur, avec des résultats faussés.

En effet, n'oublions pas que dans la régression simple avec une variable explicative quantitative, le coefficient de la pente correspond au log odds-ratio d'une augmentation d'une unité de la variable explicative. On peut l'interpréter comme le changement de niveau dans notre contexte, mais ceci *quel que soit le niveau*. Or nous avons bien vu que ce n'est pas vrai en général. Le surcroît de risque lors du passage d'un niveau au suivant dépend du niveau sur lequel nous nous situons dans notre exemple (Figure 5.10). Notre codage $\{1, 2, \ldots\}$ introduit une contrainte supplémentaire qui pèse sur les résultats : la linéarité du LOGIT par rapport à la variable ordinale. Encore une fois, ce n'est pas forcément faux. Il faut en être conscient tout simplement lors de la lecture et l'interprétation des sorties du logiciel. La pire des choses est de faire sans savoir ou laisser le logiciel choisir à notre place.

A titre de curiosité, nous avons lancé la régression simple sur la variable explicative SYSTOLIC codée $\{1, 2, 3\}$. Tanagra l'a intégrée comme une variable quantitative. Nous obtenons un odds-ratio égal à $e^{\hat{a}_1} = e^{0.3407} = 1.4060$. Nous le lisons de la manière suivante : le changement de niveau entraîne 1.4 fois plus de chances de développer une maladie cardiaque, que ce soit le passage de 1 à 2 ou de 2 à 3 (Figure 5.13). La conclusion n'est pas du tout de la même teneur que celle obtenue avec le codage emboîté où le passage de 2 à 3 ($OR(2/3) = 1.626$) semblait entraîner un risque plus élevé que lors du passage de 1 à 2 ($OR(2/1) = 1.301$) (Figure 5.10).

5.3 Le cas de la régression multiple

Dans la régression multiple, plusieurs variables explicatives doivent cohabiter. Elles ont plus ou moins liées. Certaines sont redondantes. D'autres sont complémentaires. Certaines enfin peuvent masquer ou

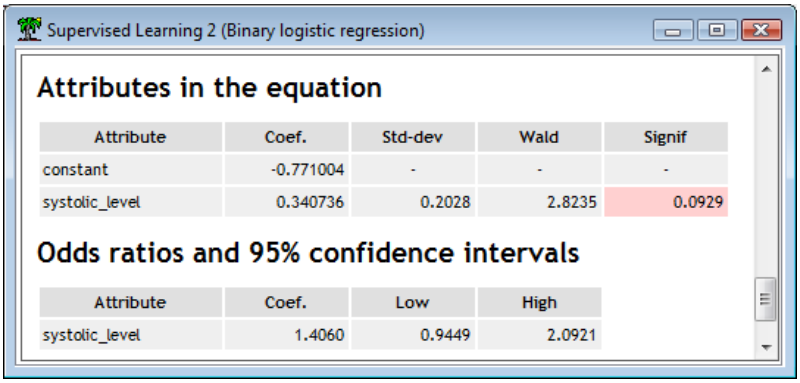


Fig. 5.13. Régression sur la variable ordinale SYSTOLIC codée $\{1, 2, 3\}$

exacerber le rôle d'autres variables. Il nous faut discerner les informations importantes en interprétant correctement les coefficients et les indicateurs fournis par la régression logistique.

5.3.1 Odds-ratio partiel

Les interprétations sous forme de log odds-ratio des coefficients font tout le charme de la régression logistique. Le principe est assez simple dans la régression simple. Lorsque l'on passe à la régression multiple, comment lire les coefficients ? Est-ce que les interprétations vues précédemment restent valables ?

La réponse est oui, mais avec une petite modification : nous avons maintenant des log odds-ratio partiels. Si l'on prend l'exemple d'une variable binaire, le coefficient est bien un log odds-ratio, mais pour lequel nous contrôlons (fixons) le rôle des autres variables. L'analogie avec la corrélation partielle bien connue en économétrie peut aider à la compréhension.

Prenons tout de suite le fichier CREDIT pour expliciter tout cela. Nous nous intéressons à la prédiction de l'acceptation de crédit en fonction, d'une variable indicatrice indiquant si on exerce une profession indépendante (PROFINDEP) dans un premier temps, puis en ajoutant la variable "nombre de problèmes rencontrés avec la banque" (NBPROB) dans un second temps. Cette dernière est traitée comme une variable quantitative pour simplifier.

La première régression ACCEPTATION = f(PROFIDENP) nous indique que la variable explicative n'est pas significative à 5%. Néanmoins, si l'on s'intéresse quand même à la valeur du coefficient, nous avons $\hat{a}_1 = -1.49$ c.-à-d. $OR(PROFINDEP) = e^{-1.49} = 0.2254$. Un individu profession indépendante a $\frac{1}{0.2254} = 4.44$ fois plus de chances de se voir refuser son crédit par rapport à un salarié. On n'aime pas trop les professions indépendantes dans cet organisme de crédit (Figure 5.14).

Nous introduisons la variable NBPROB (Figure 5.15). Surprise ! Non seulement NBPROB est très significative, ça paraît logique, ce n'est pas très indiqué d'avoir des problèmes avec sa banque, mais PROFINDEP devient aussi significative à 5%. L'introduction de NBPROB dans la régression a exacerbé son rôle. En effet, si l'on passe au odds-ratio, nous avons OR(PRODINDEP/NBPROB)$= e^{-2.028} = 0.136$. Ils ont $\frac{1}{0.1316} = 7.60$ plus de chances de se voir refuser leur crédit.

Pour comprendre le mécanisme, nous avons calculé la moyenne des problèmes rencontrés selon le type de profession.

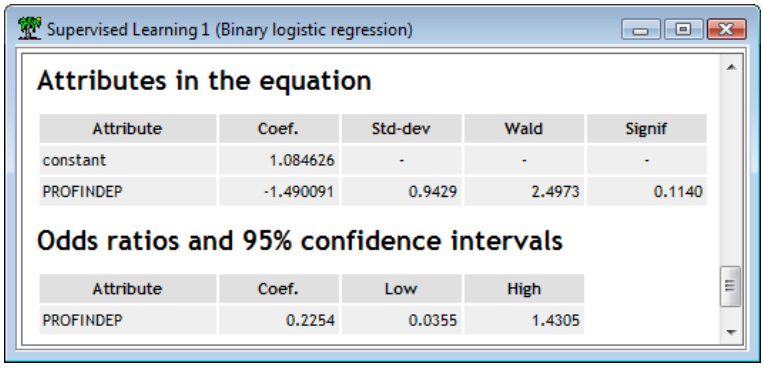


Fig. 5.14. CREDIT = f (PROFINDEP)

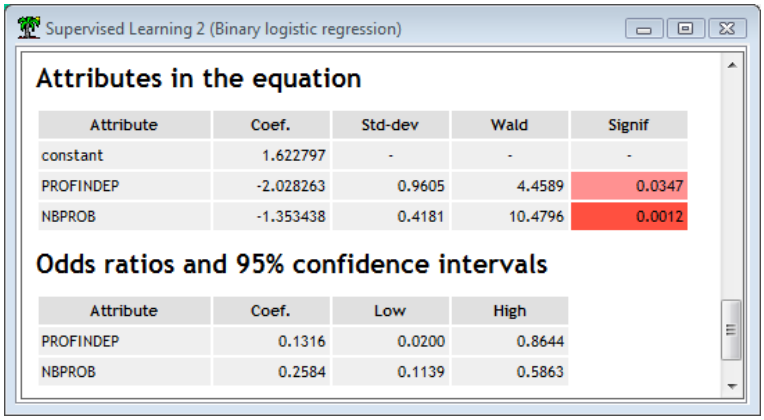


Fig. 5.15. CREDIT = f (PROFINDEP, NBPROB)

PROFINDEP	Moyenne(NBPROB)
0	0.38
1	0.00

Les professions indépendantes sont des personnes qui n'ont jamais de problèmes avec leur banque. Nous pouvons mieux lire les résultats maintenant :

- La prise en compte du nombre de problème augmente l'effet de PROFINDEP.
- En contrôlant le nombre de problèmes, le fameux "toutes choses égales par ailleurs" c.-à-d. si les professions indépendantes et les salariés présentaient le même nombre de problèmes, les premiers auraient 7.60 plus de chances de se voir refuser leur crédit. Les banques sont sans pitié.
- Les banques sont donc enclins à la clémence vis à vis des professions indépendantes (4.44 fois plus de refus quand même) uniquement parce que ce sont des gens sans problèmes [3].
- On retrouve le concept de corrélation partielle. Nous mesurons le lien d'une explicative avec la variable dépendante, à valeurs fixées pour les autres descripteurs.

Pour valider cette interprétation, nous avons filtré la base de manière à ne conserver que les individus sans problèmes (NBPROB = 0). Nous n'avons plus que $n = 82$ observations (sur les 100 initiaux).

3. Les fichiers que l'on récupère sur internet sont parfois cocasses. Comme je ne suis pas banquier, j'éviterai de trop m'étaler sur les interprétations et me concentrer sur les techniques.

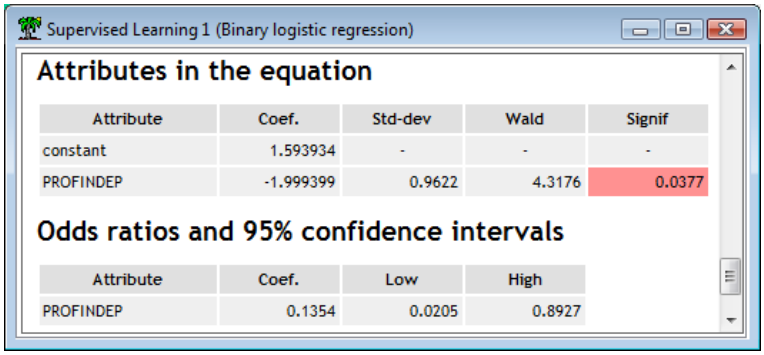


Fig. 5.16. CREDIT = f (PROFINDEP) - Limité aux NBPROB = 0 c.-à-d. $n = 82$ obs.

Nous avons lancé la régression simple CREDIT = f (PROFINDEP). Nous obtenons un résultat qui confirme l'idée ci-dessus : lorsque la population est homogène selon le nombre de problèmes, les professions indépendantes sont manifestement martyrisées ($\hat{a}_1 = -1.9994$) (Figure 5.16).

L'interprétation des coefficients en termes de log odds-ratio reste encore d'actualité dans la régression multiple. A la différence que nous contrôlons l'effet des autres variables. On parle d'odds-ratio partiels.

5.3.2 Coefficients standardisés en régression logistique

Lorsque les explicatives sont exclusivement quantitatives, il peut être intéressant de comparer leur impact sur la variable dépendante. Quelle est celle qui joue le rôle le plus important ? Dans quel sens ?

Comparer les odds-ratio paraît une solution immédiate. Mais comme les explicatives ne sont pas exprimées sur une même échelle, la variation d'une unité n'a absolument pas la même signification d'une variable à l'autre. Les odds-ratio ne sont pas comparables en l'état. La solution la plus simple est de centrer et réduire les explicatives. Ainsi nous pouvons mieux jauger leur influence et, de plus, nous pouvons disposer d'interprétations sous forme de variations d'écarts-type.

Dans cette section, nous souhaitons mettre en place un dispositif qui permet de

1. Comparer les influences respectives des variables explicatives.
2. Mesurer l'impact de la variation d'un écart-type d'une explicative sur le logit, soit en termes absolus c.-à-d. écarts absolus entre logit (l'exponentielle de l'écart entre deux logit est un odds-ratio, ne l'oublions pas), soit en termes relatifs c.-à-d. variation en écarts-type du logit.

Auparavant, faisons un petit détour par la régression linéaire multiple pour décrire la démarche. Nous montrerons qu'il est possible d'obtenir les coefficients standardisés sans avoir à réaliser la régression sur les données centrées et réduites.

Modele	Puissance	Poids	Consommation
Daihatsu Cuore	32	650	5.7
Suzuki Swift 1.0 GLS	39	790	5.8
Fiat Panda Mambo L	29	730	6.1
VW Polo 1.4 60	44	955	6.5
Opel Corsa 1.2i Eco	33	895	6.8
Subaru Vivio 4WD	32	740	6.8
Toyota Corolla	55	1010	7.1
Opel Astra 1.6i 16V	74	1080	7.4
Peugeot 306 XS 108	74	1100	9
Renault Safrane 2.2. V	101	1500	11.7
Seat Ibiza 2.0 GTI	85	1075	9.5
VW Golt 2.0 GTI	85	1155	9.5
Citroen ZX Volcane	89	1140	8.8
Fiat Tempra 1.6 Liberty	65	1080	9.3
Fort Escort 1.4i PT	54	1110	8.6
Honda Civic Joker 1.4	66	1140	7.7
Volvo 850 2.5	106	1370	10.8
Ford Fiesta 1.2 Zetec	55	940	6.6
Hyundai Sonata 3000	107	1400	11.7
Lancia K 3.0 LS	150	1550	11.9
Mazda Hachtback V	122	1330	10.8
Mitsubishi Galant	66	1300	7.6
Opel Omega 2.5i V6	125	1670	11.3
Peugeot 806 2.0	89	1560	10.8
Nissan Primera 2.0	92	1240	9.2
Seat Alhambra 2.0	85	1635	11.6
Toyota Previa salon	97	1800	12.8
Volvo 960 Kombi aut	125	1570	12.7
Moyenne	77.7143	1196.9643	9.0750
Ecart-type	32.2569	308.9928	2.2329

A - Coefficients sur données originelles

	Poids	Puissance	Constante
coef.	0.0044	0.0256	1.7696
ecart-type	0.0009	0.0083	
t	5.1596	3.0968	
p-value	0.00002	0.00478	

B - Coefficients standardisés à partir des données centrées réduites

	Poids	Puissance	Constante
coef.	0.615016	0.369136	pas de constante !

C -Coeff. Standardisés à partir de la formule de correction (cf. Ménard)

	Poids	Puissance	Constante
coef.	0.615016	0.369136	pas de constante !

Fig. 5.17. CONSO = f (POIDS, PUIS) - Régression linéaire - Coefficients standardisés

Comparer l'impact des explicatives dans la régression linéaire

Nous souhaitons expliquer la consommation (CONSO) des véhicules à partir de leur poids (POIDS) et de leur puissance (PUIS). Nous disposons de $n = 28$ observations (Figure 5.17).

Nous réalisons la régression linéaire multiple sur les données non transformées (Figure 5.17; tableau A). Nous obtenons les coefficients $\hat{a}_{poids} = 0.0044$ et $\hat{a}_{puis} = 0.0256$, tous deux sont significatifs. Nous lisons : lorsque la puissance (resp. le poids) augmente d'unité, la consommation augmente de 0.0256 l/100km (resp. 0.0044). Est-ce à dire que la puissance a plus d'influence que le poids ? Ce serait une erreur de le penser. En effet, la puissance est exprimée en chevaux, le poids en kilogramme, nous ne pouvons pas les rapprocher. Si nous avions exprimé le poids en tonne, avec ce principe nous aurions conclu exactement l'inverse.

Une solution immédiate consiste à regarder les t de Student ou les p-value des tests de significativité. Nous avons $t_{poids} = 5.1596$ et $t_{puis} = 3.0968$. Finalement, ce serait plutôt l'inverse. Le poids a plus d'influence sur la consommation. Nous avons une partie de la réponse. En revanche, les chiffres que nous avons ne sont absolument pas interprétables.

Solution 1 : travailler sur les données centrées réduites

Pour obtenir des coefficients que l'on sait lire en termes d'écarts-type, il est d'usage de centrer et réduire les variables (y compris la variable dépendante). C'est ce que nous faisons, puis nous relançons la régression (Figure 5.17; tableau B).

Premier constat : puisque les données sont centrées, la constante est mécaniquement nulle. En ce qui concerne les autres coefficients, nous avons $\hat{a}^{std}_{poids} = 0.615$ et $\hat{a}^{std}_{puis} = 0.369$. Ce que nous avions subodoré précédemment est confirmé : le poids pèse plus sur la consommation que la puissance.

L'énorme avantage avec cette solution est que nous disposons d'une lecture cohérente des coefficients : lorsque le poids (resp. la puissance) augmente de 1 écart-type, la consommation augmente de 0.615 (resp. 0.369) fois son écart type.

Les deux objectifs que nous nous sommes fixés sont atteints : nous pouvons comparer les mérites respectifs des explicatives ; nous savons lire les coefficients en termes de variations d'écarts-type de la variable dépendante.

Solution 2 : correction des coefficients non standardisés

Il est possible de retrouver les coefficients standardisés à partir des résultats de la régression sur les données initiales. Cela nous dispense d'avoir à relancer touts les calculs. Il suffit d'introduire la correction suivante [10] (page 51) :

$$\hat{a}^{std}_j = \hat{a}_j \times \frac{\hat{\sigma}_j}{\hat{\sigma}_y} \tag{5.3}$$

où $\hat{\sigma}_j$ est l'écart-type de la variable X_j, $\hat{\sigma}_y$ celle de l'endogène.

Nous avons introduit ces nouvelles modifications (Figure 5.17 ; tableau C) en utilisant les informations situées sous le tableau de données. Pour la variable *poids* par exemple

$$\hat{a}^{std}_{poids} = 0.0044 \times \frac{308.9928}{2.2329} = 0.615016$$

Nous retrouvons exactement les coefficients standardisés de la régression sur données centrées réduites.

Comparer l'impact des explicatives dans la régression logistique

Les mêmes idées sont transposables à la régression logistique, avec deux objectifs toujours :

1. Comparer les mérites respectifs des explicatives.
2. Obtenir une interprétation des coefficients de la forme : une augmentation de 1 écart-type de la variable X entraîne une variation de θ écarts-type du LOGIT.

Données originelles		
age	taux_max	coeur
50	126	presence
49	126	presence
46	144	presence
49	139	presence
62	154	presence
35	156	presence
67	160	absence
65	140	absence
47	143	absence
58	165	absence
57	115	absence
59	145	absence
44	175	absence
41	153	absence
54	152	absence
52	169	absence
57	168	absence
50	158	absence
44	170	absence
49	171	absence
Statistiques descriptives		
51.75	151.45	moyenne
8.16	16.66	é.t.

Coefficients de la régression sur les données originelles

Attribute	Coef.	Std-dev	Wald	Signif
constant	16.254441	-	-	-
age	-0.1201	0.0843	2.0305	0.1542
taux_max	-0.0744	0.0387	3.6886	0.0548

Fig. 5.18. COEUR = f (AGE, TAUX MAX) - Coefficients non standardisés

Régression sur les données originelles : les coefficients non standardisés

Nous travaillons sur le fichier COEUR. Les écarts-type des explicatives sont $\hat{\sigma}_{age} = 8.16$ et $\hat{\sigma}_{taux\ max} = 16.66$. Nous aurons besoin de ces informations par la suite.

Nous implémentons la régression COEUR = f (age, taux max). Les coefficients estimés sont $\hat{a}_{age} = -0.1201$ et $\hat{a}_{taux\ max} = -0.0744$ (Figure 5.18). Ce sont des log odds-ratio consécutifs à la variation d'une unité des variables. Mais comme ces dernières sont exprimées sur des échelles différentes, nous ne pouvons rien conclure concernant l'importance relative des explicatives. On peut néanmoins le deviner via les p-value, taux max semble plus influent puisque sa p-value est plus petite.

Mais cela ne répond pas à notre seconde question : comment lire les coefficients en termes de variation du logit ? Pour pouvoir y répondre, nous devons calculer l'écart-type du logit $\hat{\sigma}_{logit}$ prédit par le modèle. Nous avons donc construit le logit prédit $\hat{c}$ et la probabilité prédite $\hat{\pi}$ (Figure 5.19)

Essayons d'analyser les implications des variations de la variable âge (Δ_{age}), toutes choses égales par ailleurs c.-à-d. en fixant par exemple la valeur de $taux\ max$ à 150, sur la variation absolue $\Delta_{logit}(\Delta_{age})$ et relative $\delta_{logit}(\Delta_{age}) = \frac{\Delta_{logit}(\Delta_{age})}{\hat{\sigma}_{logit}}$ du logit (Figure 5.20) :

- Lorsque $\Delta_{age} = 1$, nous obtenons $\Delta_{logit}(1) = -0.1201 = \hat{a}_{age}$. C'est l'interprétation usuelle des coefficients de la régression logistique sur les variables explicatives quantitatives.
- Si nous ramenons la variation du logit à son écart-type c.-à-d. $\delta_{logit}(1) = \frac{\Delta_{logit}(1)}{\hat{\sigma}_{logit}} = \frac{-0.1201}{1.4851} = -0.0809$, nous obtenons une valeur dont on ne voit pas très bien la teneur.
- Enfin, pour une variation de 1 écart-type de l'âge, $\Delta(age) = \hat{\sigma}_{age} = 8.16$, nous observons un écart absolu $\Delta_{logit}(\hat{\sigma}_{age}) = -0.9803$ et un écart relatif $\delta_{logit}(\hat{\sigma}_{age}) = -0.6601$ que rien dans les résultats de la régression logistique ne nous permet de deviner. Nous sommes obligés de les calculer explicitement.

age	taux_max	coeur	cœur	prediction(logit)	prediction(PI)
50	126	presence	1	0.876683	0.706134386
49	126	presence	1	0.996793	0.730427577
46	144	presence	1	0.018229	0.504557124
49	139	presence	1	0.029814	0.507452948
62	154	presence	1	-2.647361	0.066151849
35	156	presence	1	0.446843	0.609888367
67	160	absence	0	-3.694209	0.024263746
65	140	absence	0	-1.966329	0.122783735
47	143	absence	0	-0.027498	0.493125933
58	165	absence	0	-2.985134	0.048102006
57	115	absence	0	0.854126	0.701431948
59	145	absence	0	-1.617584	0.165538336
44	175	absence	0	-2.047424	0.114312931
41	153	absence	0	-0.050668	0.487335709
54	152	absence	0	-1.537715	0.176867692
52	169	absence	0	-2.562006	0.071624041
57	168	absence	0	-3.088173	0.043597752
50	158	absence	0	-1.503573	0.181893228
44	170	absence	0	-1.675509	0.157691068
49	171	absence	0	-2.350442	0.087030646

e.t. logit	1.4851

Fig. 5.19. COEUR = f (AGE, TAUX MAX) - Non standardisé - Calcul de l'écart-type du logit

Test age	
age	40
taux-max	150
Logit	0.2926
age	41.00
taux-max	150
Logit	0.1725
Ecart(Logit)	-0.1201
Ecart ramené à l'écart-type	-0.0809

A - Variation de 1 unité de « âge »

Test age	
age	40
taux-max	150
Logit	0.2926
age	48.16
taux-max	150
Logit	-0.6877
Ecart(Logit)	-0.9803
Ecart ramené à l'écart-type	-0.6601

B - Variation de 1 écart-type de « âge »

Fig. 5.20. Variations du logit consécutives aux variations de "âge"

Aucune des questions que nous avons mis en avant n'ont obtenu de réponses avec les coefficient non standardisés : nous ne savons rien sur les influences comparées des explicatives ; nous ne mesurons l'impact sur le logit, en termes relatifs, des variations des explicatives. Dans ce qui suit, nous étudions différents types de standardisation proposés dans la littérature [10] (pages 51 à 56).

Solution 1 *: Standardisation sur les explicatives seulement*

Nous calculons le coefficient standardisé de la manière suivante

$$\hat{a}_j^{std.1} = \hat{a}_j \times \hat{\sigma}_j \tag{5.4}$$

Nous obtenons les nouveaux coefficients

Variable	Coefficient
Constante	non nulle mais non interprétable
age	$-0.1201 \times 8.16 = -0.9803 = \Delta_{logit}(\hat{\sigma}_{age})$
taux max	$-0.0744 \times 16.66 = -1.2389$

Plusieurs informations apparaissent :

- Les coefficients mesurent la variation absolue du logit consécutive à une augmentation de 1 écart-type des variables c.-à-d. $\hat{a}_j^{std.1} = \Delta_{logit}(\hat{\sigma}_j)$.
- Comme nous mesurons l'impact sur le logit des variations en écarts-type des explicatives, nous pouvons comparer leur poids relatif dans la régression. Manifestement "taux max" a un impact plus élevé (en écart absolu du logit) que l'âge.
- Nous ne disposons pas d'informations sur la variation relative δ_{logit}.
- Enfin, dernier commentaire important, cette standardisation nous fournit directement **les coefficients que l'on aurait obtenu si on avait lancé la régression logistique sur les données centrées réduites** [4].

***Solution 2** : Standardisation sur les explicatives et l'écart-type du logit*

Une autre standardisation est proposée dans la littérature

$$\hat{a}_j^{std.2} = \hat{a}_j \times \frac{\hat{\sigma}_j}{\hat{\sigma}_{logit}} \tag{5.5}$$

Sur le fichier COEUR, nous aurons

Variable	Coefficient
Constante	non nulle mais non interprétable
age	$-0.1201 \times \frac{8.16}{1.4851} = -0.6601 = \delta_{logit}(\hat{\sigma}_{age})$
taux max	-0.8342

Quelques commentaires :

- Les nouveaux coefficients mesurent la variation relative du logit lorsqu'on augmente de 1 écart-type l'explicative c.-à-d. $\hat{a}_j^{std.2} = \delta_{logit}(\hat{\sigma}_j)$
- Ils permettent aussi de comparer l'impact des explicatives.

***Solution 3** : Standardisation sur les explicatives et l'écart-type théorique de la loi de répartition logistique*

La dernière standardisation vaut surtout parce qu'elle est proposée dans le logiciel SAS [10] (page 55)

$$\hat{a}_j^{std.3} = \hat{a}_j \times \frac{\hat{\sigma}_j}{\sigma_{theorique}} \tag{5.6}$$

4. Merci à Samuel K.L. de m'avoir indiqué le bon emplacement de ce commentaire !

où $\sigma_{theorique} = \frac{3.14159265}{\sqrt{3}} = 1.8138$ est l'écart-type théorique de la loi de distribution logistique standard [5].

Sur le fichier COEUR, nous aurons

Variable	Coefficient
Constante	non nulle mais non interprétable
age	$-0.1201 \times \frac{8.16}{1.8138} = -0.5405$
taux max	-0.6830

Comme pour toutes les autres standardisations, les coefficients permettent de comparer l'impact des explicatives. Mais elles ne s'interprètent pas en termes de variation du logit.

5. Voir B. Scherrer, *Biostatistique - Volume 1*, Gaëtan Morin Editeur, 2007 ; pages 303 et 304.

6

Analyse des interactions

On parle d'interaction lorsque l'effet d'une explicative sur la variable dépendante dépend du niveau (de la valeur) d'une autre explicative. Boire est mauvais pour la santé (*paraît-il*). Boire et fumer en même temps, c'est pire, on a intérêt à faire son testament tout de suite (*je lègue mes pdf à mes étudiants*). Il faut (1) que l'on puisse décrire l'interaction sous la forme d'une nouvelle variable que la régression logistique saura prendre en compte ; (2) que l'on vérifie si cette conjonction produit un effet significatif sur la variable dépendante ; (3) le mesurer en termes de surcroît de risque, d'odds-ratio ; (4) définir une stratégie d'exploration des différentes interactions que l'on pourrait former à partir des variables disponibles ; (5) interpréter correctement les coefficients fournis par l'estimation.

On parle d'interaction d'ordre 1 lorsque l'on croise 2 variables ; interaction d'ordre 2 lorsque l'on croise 3 variables ; etc. L'analyse des interactions est un sujet très riche en régression logistique. Notre texte doit beaucoup à l'excellente monographique de Jaccard [4]. On trouvera des sections entières consacrées à ce sujet dans plusieurs ouvrages en français ([11], pages 96 à 106, pour deux variables explicatives ; [23], pages 441 à 446).

6.1 Définir les interactions entre variables explicatives

6.1.1 Interaction par le produit de variables

On caractérise généralement l'interaction par le produit de deux (ou plusieurs) variables. La signification n'est pas la même selon leur type. Lorsque les variables sont des indicatrices, soit parce qu'elles binaires par nature, soit parce qu'il s'agit d'une indicatrice de modalité d'une variable qualitative, le produit indique la conjonction des caractéristiques. Par exemple, si $X_1 = fumeur$ et $X_2 = alcoolique$, la variable $Z = X_1 * X_2$ prend la valeur 1 lorsque l'on a affaire à un fumeur alcoolique. Elle prend la valeur 0 lorsqu'il s'agit d'un fumeur qui ne boit pas ; ou d'un soiffard qui ne fume pas ; ou lorsque la personne n'est ni fumeur, ni alcoolique. L'insertion de la variable Z dans la régression permet de vérifier l'interaction. Si l'impact du tabac est constant que l'on soit alcoolique ou pas, le coefficient associé à Z ne devrait pas être significatif ; dans le cas contraire, s'il est significativement différent de 0, cela veut dire que l'impact du tabac n'est pas le même chez les alcooliques et les non-alcooliques. On parle de ***modèle saturé*** lorsque l'on intègre toutes les interactions possibles dans la régression.

On utilise également le produit quand nous traitons des variables quantitatives. Il faut être conscient simplement que l'on caractérise un certain type d'interaction. Admettons que X_1 maintenant représente la consommation de cigarettes par jour, X_2 la consommation d'alcool. Que penser de $Z = X_1 * X_2$ quand elle est introduite dans la régression logistique ?

Le LOGIT s'écrit

$$\begin{aligned} LOGIT &= a_0 + a_1 X_1 + a_2 X_2 + a_3 Z \\ &= a_0 + a_1 X_1 + a_2 X_2 + a_3 X_1 * X_2 \\ &= a_0 + (a_1 + a_3 X_2) X_1 + a_2 X_2 \end{aligned}$$

Voyons ce qu'il en est si l'on fait varier la variable X_1 d'une unité

$$\begin{aligned} LOGIT(\Delta X_1 = 1) &= a_0 + (a_1 + a_3 X_2)(X_1 + 1) + a_2 X_2 \\ &= a_0 + (a_1 + a_3 X_2) X_1 + a_2 X_2 + (a_1 + a_3 X_2) \end{aligned}$$

De fait, la variation du logit consécutive à une variation d'une unité de X_1 est une fonction linéaire de la seconde variable X_2

$$\Delta LOGIT(\Delta X_1 = 1) = LOGIT(\Delta X_1 = 1) - LOGIT = a_1 + a_3 X_2$$

De manière plus générale, la variation du logit lorsque X_1 évolue de d unités s'écrit

$$\Delta LOGIT(\Delta X_1 = d) = (a_1 + a_3 X_2) \times d \tag{6.1}$$

Il faut garder cette idée en tête. Concernant les variables quantitatives, utiliser le produit caractérise un certain type d'interaction : le log odds-ratio consécutif à une variation d'une des explicatives est fonction linéaire des autres explicatives. Ce n'est pas une limitation, il faut en être conscient simplement lorsque nous analysons les résultats.

6.1.2 Étude du ronflement

On cherche à déterminer les facteurs de ronflement à partir d'un fichier comportant $n = 100$ adultes. Les variables explicatives étudiées sont le sexe (homme = 1) et le tabac (fumeur = 1). Nous réalisons la régression sur ces deux indicatrices dans un premier temps. Il semble, au risque 10%, qu'être un homme est propice au ronflement (messieurs, demandez ce qu'il en est à vos épouses). Le tabac joue un rôle également (Figure 6.1). Le critère BIC (SC) est égal à 136.966.

Introduisons la variable $Z = homme \times tabac$. Nous souhaitons savoir si la conjonction "être un homme fumeur" entraîne une augmentation du risque de ronfler.

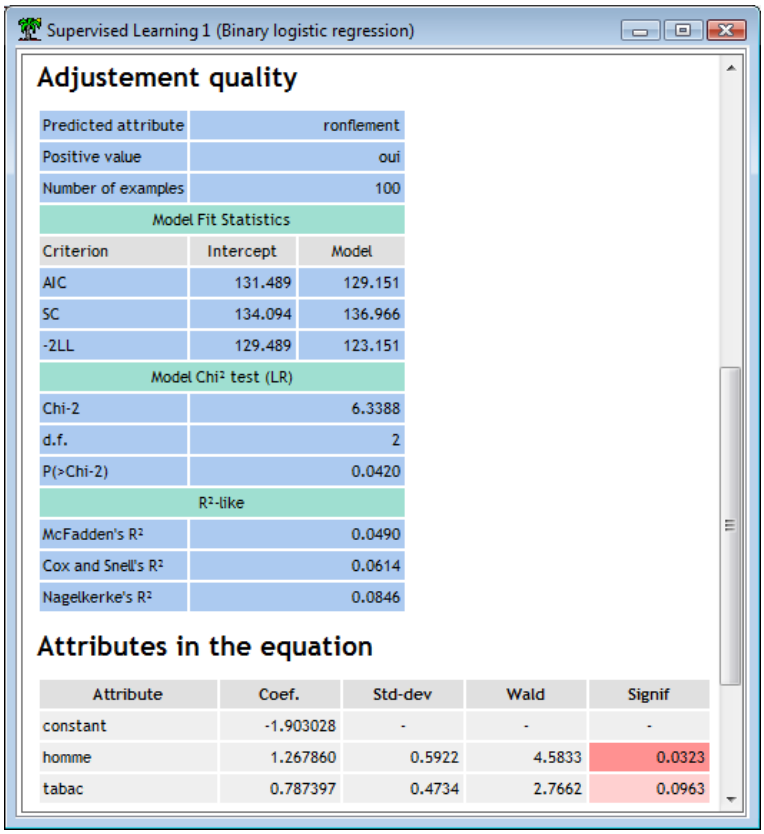


Fig. 6.1. $Ronflement = f(homme, tabac)$

Remarque : La lecture en termes de conjonctions en est une parmi les autres. Bien souvent, dans les études réelles, les variables explicatives ne jouent pas le même rôle. Dans notre exemple, on peut par exemple étudier l'effet du tabac (**facteur de risque**) sur le ronflement. Puis analyser si cet effet est différent selon que l'on est un homme ou une femme. La variable "sexe" (homme) est alors appelée **variable modératrice**.

Nous relançons la régression avec la troisième variable Z (Figure 6.2). Nous constatons plusieurs choses :

- La régression est moins bonne que la précédente si l'on en juge au critère BIC (SC). Il est passé à 141.432 (plus le BIC est élevé, moins bon est le modèle, rappelons-le). C'est le danger qui nous guette à mesure que l'on introduit de nouvelles variables peu ou prou pertinentes dans la régression.
- La variable traduisant l'interaction n'est pas significative : les hommes fumeurs ne ronflent pas plus que les autres (ou, si nous sommes dans le schéma "facteur de risque vs. effet modérateur", le tabac ne joue pas un rôle différencié selon le sexe).

6.1.3 Coefficients des indicatrices seules

Un autre résultat doit attirer notre attention : curieusement, les autres indicatrices ne sont plus significatives à 10%. Cela laisse à penser que les variables ne pèsent pas individuellement sur le risque de

Supervised Learning 1 (Binary logistic regression)

Adjustement quality

Predicted attribute	ronflement
Positive value	oui
Number of examples	100

Model Fit Statistics

Criterion	Intercept	Model
AIC	131.489	131.011
SC	134.094	141.432
-2LL	129.489	123.011

Model Chi² test (LR)

Chi-2	6.4784
d.f.	3
P(>Chi-2)	0.0905

R²-like

McFadden's R²	0.0500
Cox and Snell's R²	0.0627
Nagelkerke's R²	0.0864

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-2.197225	-	-	-
homme	1.586316	1.0920	2.1102	0.1463
tabac	1.185624	1.2051	0.9680	0.3252
homme_tabac	-0.479404	1.3132	0.1333	0.7151

Fig. 6.2. $Ronflement = f(homme, tabac, homme \times tabac)$

ronfler. Or on sait que ce n'est pas vrai au regard du résultat de la régression sans le terme d'interaction. En fait, croire que les coefficients associées aux indicatrices seules correspondent aux effets individuelles des variables est une erreur [4] (page 20). Ils indiquent l'effet de la variable conditionnellement au fait que l'autre indicatrice prend la valeur 0.

Prenons le coefficient de *homme* (sexe = homme) qui est égal à $\hat{a}_{homme} = 1.586316$ (on oublie que la variable est non significative à 10%). En passant à l'exponentielle, nous avons OR(sexe=homme) = $e^{1.586316} = 4.9$ c.-à-d. les hommes ont 4.9 fois plus de chances de ronfler que les femmes *chez les non-fumeurs c.-à-d.* $tabac = 0$!

Pour nous en persuader, nous avons filtré la base en ne retenant que les non-fumeurs. Nous avons $n = 64$ observations. Nous avons réalisé une régression simple $ronflement = f(homme)$ (Figure 6.3). Nous retrouvons le coefficient de *homme* de la régression avec interaction, avec une p-value identique.

6.2 Stratégie pour explorer les interactions

6.2.1 Modèle hiérarchiquement bien formulé

Les considérations de la section précédente nous amènent à un aspect très important de ce chapitre : les stratégies d'exploration des interactions. Il est évident que l'on ne peut par s'appuyer sur des procédures

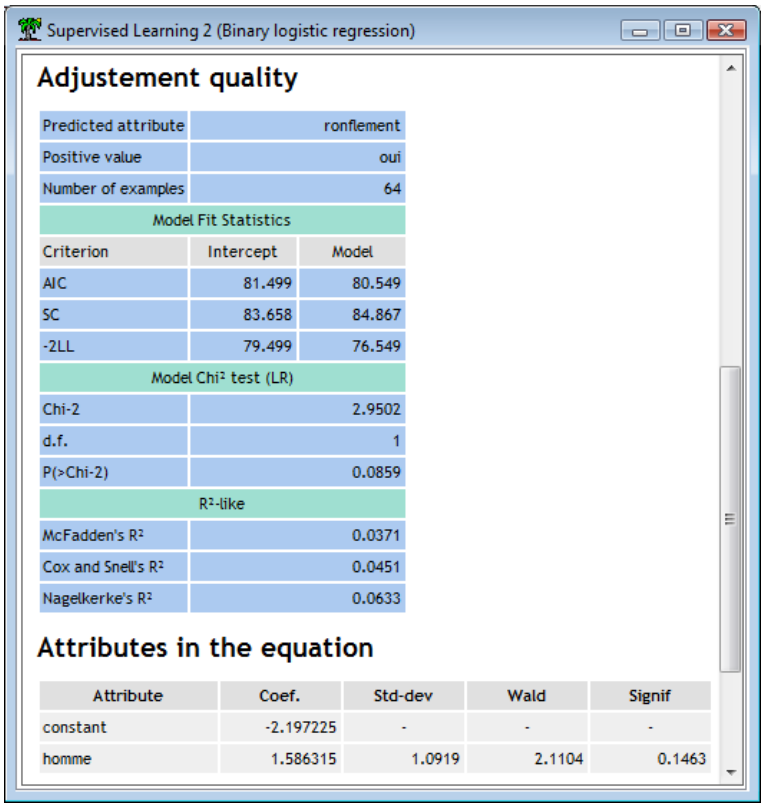


Fig. 6.3. $Ronflement = f(homme)$ chez les non-fumeurs (tabac $= 0$) - $n = 64$ obs.

purement mécaniques comme celles qui sont décrites dans le chapitre consacré à la sélection de variables (chapitre 7). Il faut tenir compte du rôle des variables dans les différents niveaux d'interactions.

Un modèle est dit **"hiérarchiquement bien formulé"** (HBF) (en anglais, *hierarchically well formulated model* ; [4], page 15) si toutes les interactions d'ordre inférieurs de l'interaction d'ordre le plus élevé sont présents.

Vite un exemple pour bien comprendre. Si l'interaction $X_1 * X_2 * X_3$ est présent dans la régression, nous devons y retrouver également les interactions d'ordre 1 c.-à-d. $X_1 * X_2$, $X_1 * X_3$ et $X_2 * X_3$; mais aussi les interactions d'ordre 0 (les variables prises individuellement) c.-à-d. X_1, X_2 et X_3. Cette contrainte doit être respectée lors du processus de sélection de variables.

Deux situations sont envisageables :

1. Si $X_1 * X_2 * X_3$ est significatif, nous arrêtons le processus de sélection, toutes les autres interactions sont conservées.
2. Dans le cas contraire, nous pouvons la supprimer. Reste à définir une stratégie d'élimination parmi les multiples interactions du même ordre (d'ordre 1 concernant notre exemple), toujours en respectant la règle édictée ci-dessus :
   a) Une première approche consiste à confronter le modèle complet incluant toutes les interactions d'ordre supérieur $Y = f(X_1, X_2, X_3, X_1 * X_2, X_1 * X_3, X_2 * X_3)$ avec celle où elles sont absentes c.-à-d. $Y = f(X_1, X_2, X_3)$, en utilisant le test du rapport de vraisemblance ou le test de Wald. Si

on accepte H_0, les coefficients associées aux termes d'interactions sont tous nuls, nous pouvons les supprimer en bloc. Dans le cas contraire, rejet de H_0, la situation se complique. Nous devons comparer le modèle complet avec un modèle n'incluant que certaines interactions [4] (page 64). Admettons que nous souhaitons évaluer le terme $X_2 * X_3$

b) Nous pouvons la confronter avec la régression $Y = f(X_1, X_2, X_3, X_1 * X_2, X_1 * X_3)$. Ce modèle est toujours HBF si l'on se réfère à la définition ci-dessus. Après il faut savoir interpréter correctement les coefficients.

c) Si $X_2 * X_3$ est retirée de la régression, nous pouvons choisir l'autre terme d'interaction ($X_1 * X_2$ ou $X_1 * X_3$) à éliminer en les évaluant tour à tour.

d) Ou bien si une des variables joue un rôle prééminent, nous focaliser sur la suppression de cette variable. Par exemple, si X_3 joue un rôle particulier, après avoir retiré $X_2 * X_3$, nous cherchons à évaluer $X_1 * X_3$, puis le cas échéant X_3.

6.2.2 Étude du ronflement avec 3 variables

Nous revenons à notre étude du ronflement avec 3 variables explicatives $X_1 = homme$, $X_2 = age$ et $X_3 = tabac$. On cherche en priorité à isoler l'effet du tabac sur le ronflement. Dans un premier temps, avec le logiciel R, nous calculons le modèle avec toutes les interactions

```
#régression logistique - complet
modele.full <- glm(ronflement ~ homme+age+tabac+homme_age+homme_tabac+age_tabac
+hom_tab_age, data = donnees, family = "binomial")
print(summary(modele.full))
```

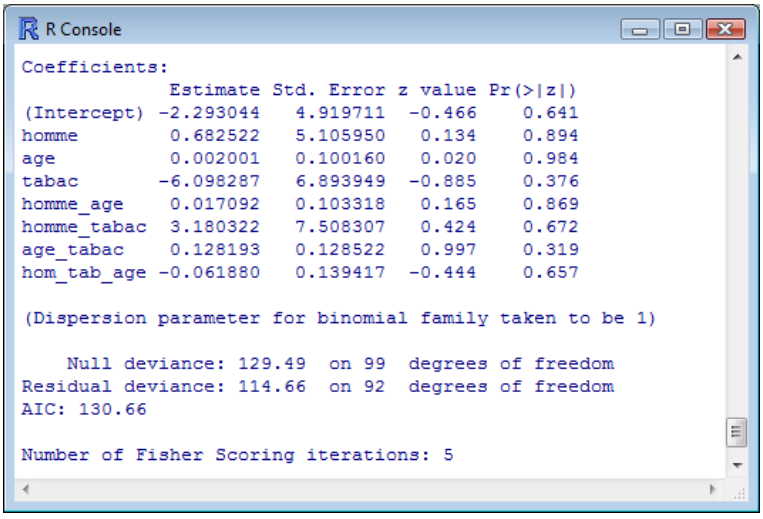


Fig. 6.4. $Ronflement = f(homme, tabac, age, \dots)$ - Modèle complet

La déviance du modèle est $D_{\{0,1,2\}} = 114.66$ (Figure 6.4). Aucun coefficient ne semble significatif. Il ne faut pas trop s'en formaliser, il doit y avoir de fortes corrélations entre les variables.

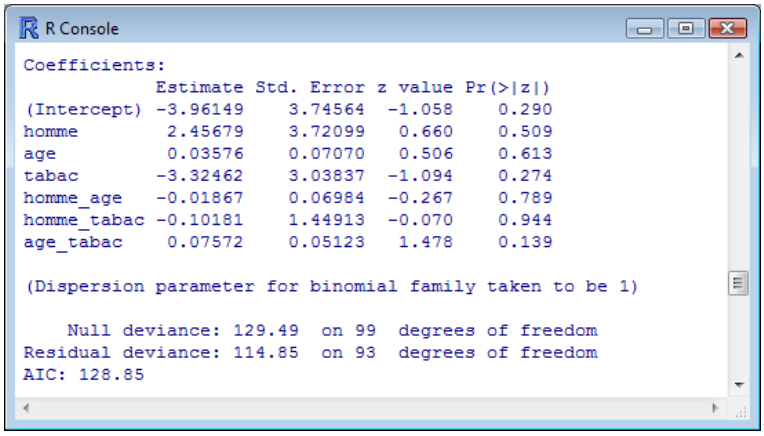


Fig. 6.5. $Ronflement = f(homme, tabac, age, \dots)$ - Interactions d'ordre 1

Nous passons au modèle avec les interactions d'ordre 1 (Figure 6.5).

```
#régression logistique - avec interactions d'ordre 1
modele.1 <- glm(ronflement ~ homme+age+tabac+homme_age+homme_tabac+age_tabac,
data = donnees, family = "binomial")
print(summary(modele.1))
```

La déviance est $D_{\{0,1\}} = 114.85$. La statistique du rapport de vraisemblance est $LR = D_{\{0,1\}} - D_{\{0,1,2\}} = 114.85 - 114.66 = 0.19$. Avec la loi du χ^2 à $(93 - 92) = 1$ degré de liberté, nous avons une p-value de 0.663. Manifestement, au risque 10%, l'interaction d'ordre 2 ne joue aucun rôle dans l'explication du ronflement.

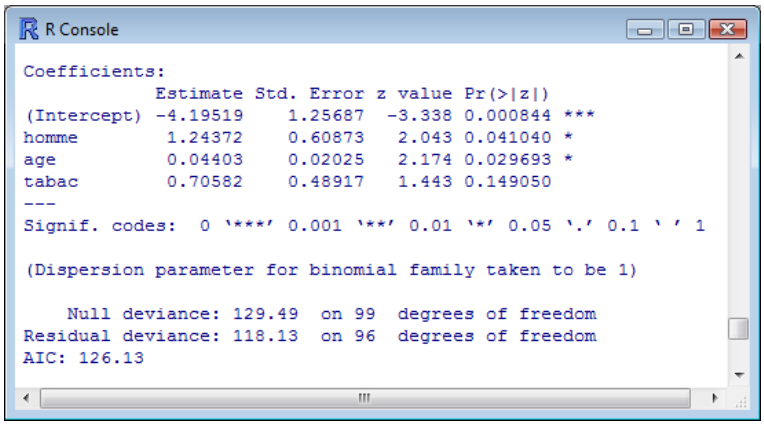


Fig. 6.6. $Ronflement = f(homme, tabac, age)$ - Interactions d'ordre 0

Évaluons maintenant le bloc d'interactions d'ordre 1. Nous réalisons la régression avec uniquement les variables individuelles.

```
#régression logistique - sans interactions d'ordre 1
modele.0 <- glm(ronflement ~ homme+age+tabac, data = donnees, family = "binomial")
print(summary(modele.0))
```

Nous obtenons la déviance $D_{\{0\}} = 118.13$ (Figure 6.6). La statistique du test est $LR = D_{\{0\}} - D_{\{0,1\}} = 118.13 - 114.85 = 3.28$. Avec un χ^2 à $(96 - 93) = 3$ degrés de liberté, nous avons une p-value de 0.350. Nous pouvons éliminer le bloc complet des termes d'interaction d'ordre 1.

Enfin, en considérant cette dernière régression, on se rend compte que *tabac* n'est pas significatif au sens du test de Wald (le test du rapport de vraisemblance aboutit à la même conclusion). Nous pouvons l'éliminer. Le modèle finalement sélectionné inclut *homme* et *age* (Figure 6.7). Moralité : les hommes ronflent plus que les femmes à age égal ; à sexe égal, plus on est âgé, plus on ronfle.

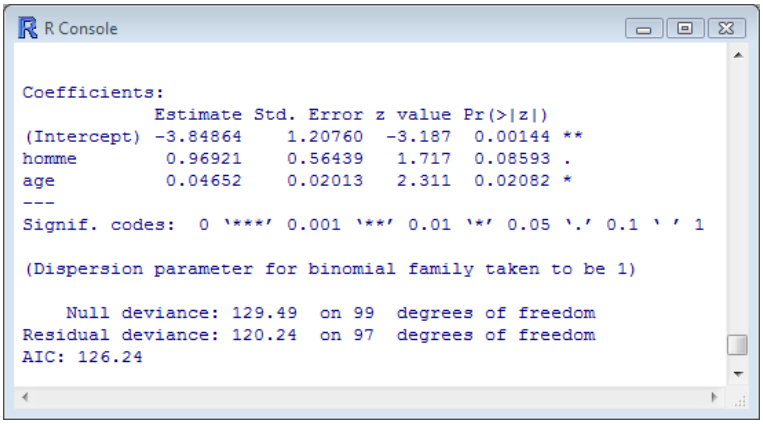


Fig. 6.7. $Ronflement = f(homme, age)$

6.3 Calcul de l'odds-ratio en présence d'interaction

Le calcul de l'odds-ratio d'une variable (présence vs. absence d'un caractère pour une indicatrice ; variation d'une unité pour une variable quantitative) dépend des valeurs des autres variables lorsqu'il y a des termes d'interaction dans la régression [9] (pages 74 à 79). Si l'estimation ponctuelle est assez simple à produire, il en est tout autrement en ce qui concerne l'estimation par intervalle. Nous devons tenir compte des variances **et** covariances des coefficients pour obtenir la variance du log odds-ratio.

6.3.1 Estimation ponctuelle

Prenons un exemple à deux variables $\{X_1, X_2\}$ pour fixer les idées. Le logit s'exprime de la manière suivante

$$logit = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_1 X_2$$

X_2 est binaire, nous souhaitons obtenir son odds-ratio. Le logit pour $X_2 = 0$ s'écrit

$$logit(X_2 = 0) = a_0 + a_1 X_1$$

Pour $X_2 = 1$, il devient

$$logit(X_2 = 1) = a_0 + a_1 X_1 + a_2 + a_3 X_1$$

L'écart entre les logit, le log odds-ratio, est obtenu par différenciation

$$\begin{aligned} \Delta_{logit}(X_2) &= logit(X_2 = 1) - logit(X_2 = 0) \\ &= a_2 + a_3 X_1 \end{aligned}$$

Ainsi, l'odds ratio $OR(X_3) = e^{\Delta_{logit}(X_3)}$ dépend à la fois des coefficients a_2, a_3, mais aussi de la valeur de X_1. Nous ne pouvons plus nous contenter d'analyser uniquement le coefficient a_2 associé à la variable individuelle.

Ronflement en fonction du sexe et du tabac

Nous prenons le modèle complet dans l'explication du ronflement à partir du sexe et du tabac (Figure 6.2). La régression s'écrit

$$logit = -2.1972 + 1.5863 \times homme + 1.1856 \times tabac - 0.4794 \times homme * tabac$$

Nous utilisons tous les termes de la régression même s'ils ne sont pas significatifs. Nous souhaitons connaître le surcroît de risque associé au tabac chez un homme (homme = 1). L'estimation du log odds-ratio est égal à

$$\begin{aligned} \Delta_{logit}(X_2) &= a_2 + a_3 X_1 \\ &= 1.1856 - 0.4794 \times 1 \\ &= 0.7062 \end{aligned}$$

Nous pouvons dire qu'un fumeur masculin $e^{0.7062} = 2.03$ fois plus de chances de ronfler qu'un non fumeur avec les mêmes spécifications (sous réserve de la significativité des coefficients encore une fois).

Cette valeur n'est par contre pas valable pour une femme ($X_1 = 0$). Il faudrait relancer la procédure de calcul. Nous aurions un log odds-ratio de $1.1856 - 0.4794 \times 0 = 1.1856$. Une femme fumeuse a $e^{1.1856} = 3.27$ fois plus de chances de ronfler qu'une non fumeuse.

6.3.2 Estimation par intervalle

Pour construire l'intervalle de variation, nous devons tout d'abord produire une estimation de la variance du log odds-ratio. Toujours pour notre exemple à 2 variables, il s'écrit [1]

$$\hat{V}(\Delta_{logit}) = \hat{V}(\hat{a}_2) + X_1^2 \times \hat{V}(\hat{a}_3) + 2 \times X_1 \times \widehat{COV}(\hat{a}_2, \hat{a}_3)$$

Les bornes de l'intervalle au niveau de confiance $(1-\alpha)$ sont définies par

$$\Delta_{logit} \pm u_{1-\alpha/2} \times \sqrt{\hat{V}(\Delta_{logit})}$$

où $u_{1-\alpha/2}$ est le fractile de la loi normale centrée réduite.

Ronflement en fonction du sexe et du tabac

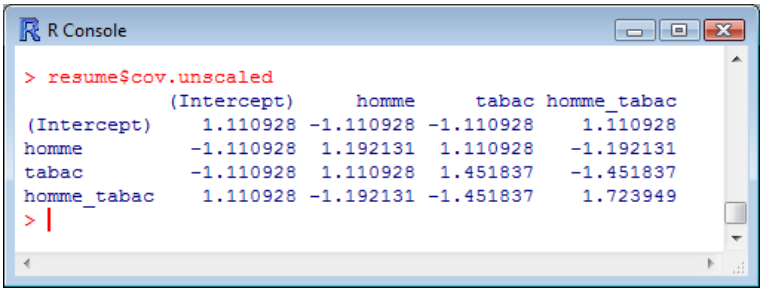


Fig. 6.8. $Ronflement = f(homme, tabac, homme * tabac)$ - Matrice de variance covariance

Poursuivons notre exemple ci-dessus, le logiciel R sait produire la matrice de variance covariance (Figure 6.8). Nous pouvons calculer la variance du log odds-ratio pour un homme $(X_1 = 1)$

$$\begin{aligned}\hat{V}(\Delta_{logit}) &= \hat{V}(\hat{a}_2) + X_1^2 \times \hat{V}(\hat{a}_3) + 2 \times X_1 \times \widehat{COV}(\hat{a}_2, \hat{a}_3) \\ &= 1.451837 + 1^2 \times 1.723949 + 2 \times 1 \times (-1.451837) \\ &= 0.272112\end{aligned}$$

Les bornes de l'intervalle de confiance à 90% du log odds-ratio s'écrivent

$$bb(\Delta_{logit}) = 0.7062 - 1.6449 \times \sqrt{0.272112} = -0.1518$$
$$bh(\Delta_{logit}) = 0.7062 + 1.6449 \times \sqrt{0.272112} = 1.5642$$

1. On devine aisément qu'à mesure que le nombre de variables augmente, avec des interactions d'ordre élevé, la formule devient rapidement assez complexe.

Et par conséquent, ceux de l'odds-ratio

$$bb(OR) = e^{0.1518} = 0.8591$$
$$bh(OR) = e^{1.5642} = 4.7790$$

L'intervalle contient la valeur 1, le tabac ne pèse pas significativement sur le ronflement *chez les hommes*. Cela ne préjuge pas des résultats chez les femmes, il faudrait reproduire la démarche complète pour savoir ce qu'il en est.

6.4 Interpréter les coefficients de la régression en présence d'interactions

L'obtention des odds-ratio est difficile pour les modèles avec interaction. Ils sont plus ou moins liés avec les coefficients de la régression, nous devons tenir compte des valeurs prises par les autres explicatives. Dans le cas de régression à deux variables cependant, nous pouvons déduire les log odds-ratio à partir des coefficients. Tout dépend du type des explicatives (voir [4], chapitres 2, 3 et 4 ; [11], pages 96 à 106).

Pour donner un tour plus concret à notre propos, nous ferons tenir un rôle différent aux variables explicatives : l'une (X) sera le facteur de risque dont on veut étudier l'impact sur la variable dépendante, généralement il s'agit d'une variable sur laquelle nous pouvons raisonnablement influer (ex. fumer ou pas, le poids, etc.) ; l'autre (Z) sera la variable modératrice qui peut masquer ou exacerber cet impact, il s'agit le plus souvent d'une variable sur laquelle nous n'avons pas réellement prise (ex. l'âge, le sexe, etc.).

6.4.1 Deux explicatives binaires

Toujours sur le fichier ronflement, nous posons $X = tabac$ et $Z = sexe$. On souhaite savoir si le tabac a une influence, et si elle est différente selon que l'on est un homme (1) ou une femme (0 = groupe de référence). Rappelons rapidement les coefficients de la régression, nous avions (Figure 6.2)

Coef.	$\hat{a}$	p-value
$\hat{a}_0$	-2.1972	-
$\hat{a}_X$	1.1856	0.3252
$\hat{a}_Z$	1.5863	0.1463
$\hat{a}_{XZ}$	-0.4794	0.7151

En passant par les tableaux croisés, nous pouvons calculer directement les odds-ratio (Figure 6.9). Nous constatons que l'odds ratio est plus élevé chez la femme ($OR(femme) = 3.27$) que chez l'homme $OR(homme) = 2.03$. Reste à déterminer s'il est significatif ou non.

Passons au logarithme de l'odds-ratio, nous pouvons les retrouver directement à partir des coefficients de la régression avec les relations suivantes :

Nombre de ronflement	homme	tabac		
	0		1	
ronflement	0	1	0	1
non	9	11	35	10
oui	1	4	19	11

OR(femme)	3.27	OR(homme)	2.03
log OR	1.1856	log OR	0.7062

Fig. 6.9. Calcul des odds-ratio par un tableau de contingence, deux variables binaires

- $\ln[OR(femme)] = 1.1856 = \hat{a}_X$, le log odds-ratio associé au facteur de risque X dans le groupe de référence correspond au coefficient du facteur de risque $\hat{a}_X$.
- $\ln[OR(homme)] = 0.7032 = \hat{a}_X + \hat{a}_{XZ}$, le log odds-ratio dans le groupe des hommes correspond à la somme des coefficients associés au facteur de risque et au terme d'interaction.
- Nous l'avons constaté précédemment, il y a un écart entre les odds-ratio. Nous savons maintenant qu'il est non significatif à 10% car le coefficient $\hat{a}_{XZ}$ du terme d'interaction ne l'est pas dans la régression (p-value $= 0.7151$).

6.4.2 Un explicative continue et une explicative binaire

Toujours dans notre problème de ronflement, nous souhaitons identifier l'impact de l'indice de masse corporelle ($X = imc$, variable quantitative) sur la variable dépendante, en contrôlant le rôle du sexe ($Z = sexe$) (homme $= 1$, femme $= 0$). Deux questions sont posées : est-ce que l'imc influe sur le ronflement, est-ce qu'il influe différemment selon que l'on est un homme ou une femme. Avec des séries de régressions simples, nous parvenons aux conclusions suivantes :

Impact	$\ln(OR)$	OR
imc chez les hommes ($n = 75$)	-0.083342	0.9200
imc chez les femmes ($n = 25$)	0.876508	2.4025

Est-ce que nous pouvons retrouver ces valeurs à partir de la régression incluant X, Z et le terme d'interaction XZ (Figure 6.10) ?

La réponse est oui, le principe est assez similaire à celui des deux variables binaires :

- $\ln[OR(femme)] = 0.876508 = \hat{a}_X$, le log odds-ratio consécutif à une variation d'une unité d'IMC chez les femmes ($Z = 0$) correspond au coefficient $\hat{a}_X$ de la régression.
- $\ln[OR(homme)] = -0.083342 = \hat{a}_X + \hat{a}_{XZ}$, le log odds-ratio consécutif à une variation d'une unité d'IMC chez les hommes ($Z = 1$) correspond à la somme des coefficients du facteur de risque et du terme d'interaction.
- Nous savons que l'écart entre ces odds-ratio n'est pas significatif à 10% parce que le coefficient du terme d'interaction ne l'est pas dans la régression.

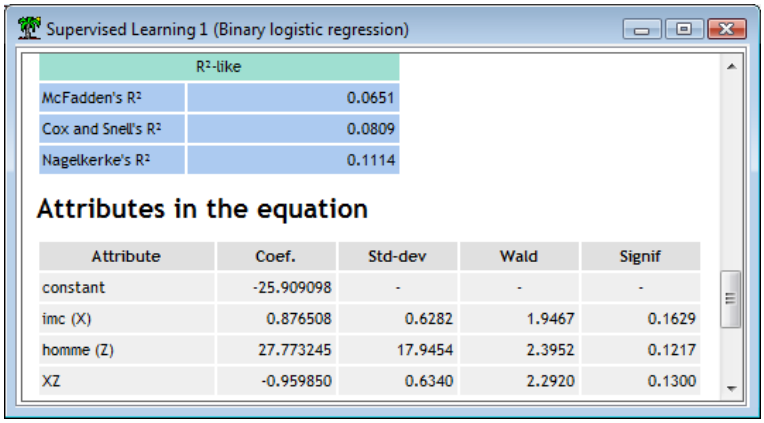


Fig. 6.10. Régression $ronflement = f(imc, homme, imc * homme)$ - $n = 100$ obs.

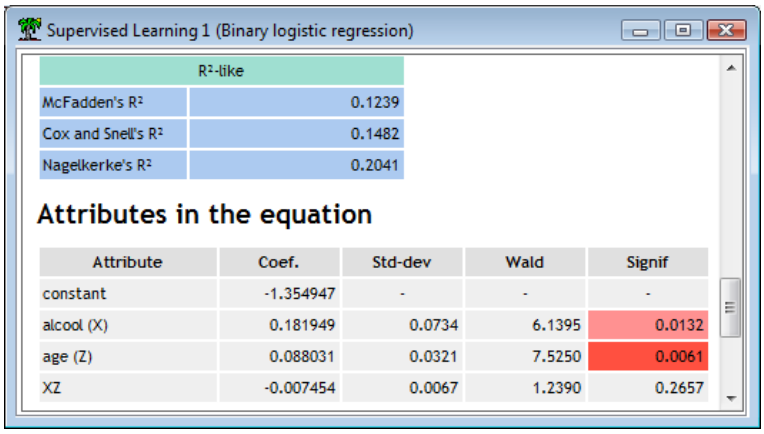


Fig. 6.11. Régression $ronflement = f(alcool, age, alcool * age)$ - $n = 100$ obs.

6.4.3 Deux explicatives continues

Dans cette section, nous essayons d'expliquer le ronflement à partir de la consommation d'alcool (X) en contrôlant l'âge (Z), nous avons centré cette dernière pour faciliter les interprétations.

Les questions que l'on se pose sont les suivantes : est-ce que l'alcool pèse sur le ronflement ? est-ce que son impact varie en fonction de l'âge? Pour répondre à cela nous avons calculé la régression $Y = f(X, Z, XZ)$ (Figure 6.11). Nous lisons les coefficients de la manière suivante :

- $\hat{a}_X = 0.181949$ correspond au log odds-ratio consécutif à une augmentation d'une unité de la consommation d'alcool pour des personnes ayant $Z = 0$ c.-à-d. l'âge moyen de la population (car la variable a été centrée).
- $\hat{a}_{XZ} = -0.007454$ est la variation du log odds-ratio associé à X lorsque Z augmente d'une unité.

Cette dernière idée mérite quelques éclaircissements. Nous pouvons ré-écrire le logit :

$$\begin{aligned} logit &= a_0 + a_X X + a_Z Z + a_{XZ} XZ \\ &= a_0 + (a_X + a_{XZ} Z)X + a_Z Z \end{aligned}$$

Le log odds-ratio relatif à une variation d'une unité de X est égal à $a_X + a_{XZ}Z$, il dépend de la valeur de Z ! Lorsque $Z = 0$, il sera a_X, nous l'avions vu précédemment ; lorsque $Z = 1$, il sera $a_X + a_{XZ}$. La différence entre ces deux quantités correspond bien à a_{XZ}.

Ceci étant, l'interaction $alcool * age$ n'est pas significative à 10%, nous pouvons la retirer de la régression. Le "bon" modèle serait finalement $ronflement = f(alcool, age)$ avec comme principales conclusions : à âge égal, boire fait ronfler ; et à consommation d'alcool égal, plus on vieillit, plus on ronfle (Figure 6.12). Bref, mesdames, si vous voulez passer des nuits en toute quiétude, mieux vaut épouser un jeune sobre qu'un vieux soûlard. Ca tombe un peu sous le sens quand même. Je ne suis pas sûr qu'il était nécessaire de faire des calculs statistiques aussi compliqués pour parvenir à cette conclusion.

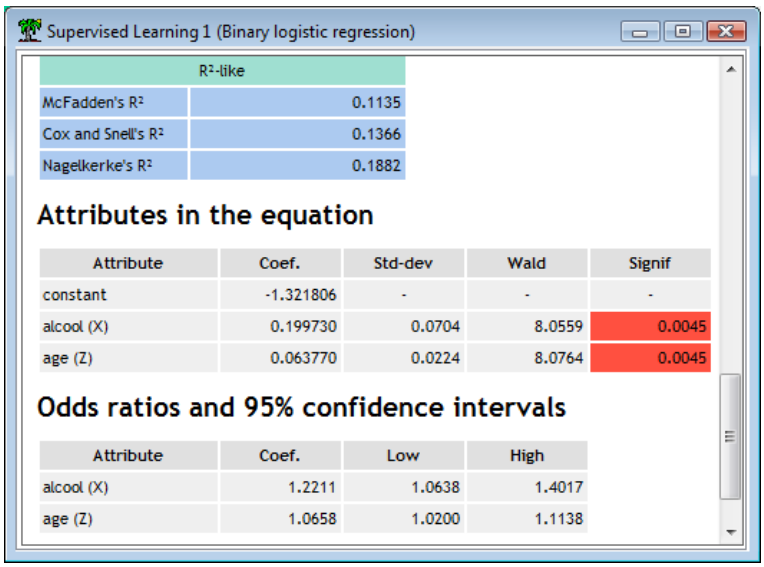


Fig. 6.12. Régression $ronflement = f(alcool, age)$ - $n = 100$ obs.

7

La sélection de variables

7.1 Pourquoi la sélection de variables ?

La sélection de variables est une étape clé de la modélisation. Dans les études réelles, nous sommes confrontés à des bases de données avec un nombre considérable de descripteurs. Ce sont autant de variables explicatives potentielles. Certaines d'entre elles sont redondantes, d'autres n'ont aucun rapport avec la variable dépendante. La méthode statistique doit nous donner des indications sur le sous-ensemble des *bonnes* variables à inclure dans le modèle. Dans l'idéal, elles devraient être orthogonales entre elles et toutes fortement liées avec la variable dépendante. Certains auteurs encensent la sélection automatique de variables parce qu'elle constitue un outil fort utile pour une première approche sur des données que l'on ne connaît pas très bien ; d'autres par contre la critiquent vertement car elle nous rend dépendante des fluctuations aléatoires dans les données, d'un échantillon à l'autre nous sommes susceptibles d'obtenir des solutions différentes [10] (page 63). Il reste qu'elle est précieuse lorsque la qualité de prédiction est l'objectif principal ou lorsque nous sommes dans un contexte exploratoire. Même si l'expert du domaine a une certaine idée des explicatives à retenir, une sélection automatique peut l'aiguiller sur les pistes à étudier.

Plusieurs raisons nous poussent à réduire le nombre de variables explicatives :

- **Moins il y aura de variables, plus facile sera l'interprétation**. En évacuant les descripteurs qui ne sont pas nécessaires à l'explication de la variable dépendante, nous pouvons plus facilement cerner le rôle de celles qui sont retenues. N'oublions pas que dans de nombreux domaines, l'explication est au moins aussi importante que la prédiction. La régression logistique nous propose des outils merveilleux pour lire les coefficients en termes de surcroît de risque. Réduire le nombre de variables permet d'en profiter pleinement.
- **Le déploiement sera facilité**. Lorsque le modèle sera mis en production, on a toujours intérêt à poser peu de questions pour identifier la classe d'appartenance d'un individu. Imaginez vous arriver au service des urgences d'un hôpital, une personne vous pose une trentaine de questions pour identifier votre problème, vous aurez eu le temps de mourir plusieurs fois. Idem, vous sollicitez un crédit auprès d'un banque, elle commence à vous demander la date de naissance de votre arrière grand-père, la question d'après vous êtes déjà dans l'établissement d'à-côté. Au fil du temps, je me suis rendu compte qu'un système aussi efficace soit-il n'est vraiment adopté par les utilisateurs que s'il est peu contraignant, simple d'utilisation.

– Dernier argument en faveur de la sélection, pour un même nombre d'observations, **un modèle avec peu de variables a de meilleures chances d'être plus robuste en généralisation**. C'est le principe du *Rasoir d'Occam*. En effet, lorsque le nombre de paramètres du modèle est trop élevé, le sur-apprentissage nous guette (*overfitting* en anglais). Le classifieur "colle" trop aux données et, au lieu d'intégrer les informations essentielles qui se rapportent à la population, il ingère les particularités de l'échantillon d'apprentissage. Introduire des variables explicatives non-pertinentes augmente artificiellement les variances des coefficients [10] (page 68), les estimations sont numériquement instables [9] (page 92). Bref, les probabilités conditionnelles $P(X/Y)$ sont mal estimées. On pense généralement qu'il faut respecter un certain ratio entre le nombre de paramètres à estimer et la taille de l'échantillon. Il est malheureusement très difficile à quantifier. Il dépend aussi de la difficulté du concept à apprendre. A titre indicatif, nous citerons la règle empirique suivante [9] (page 346) [1]

$$J + 1 \leq \frac{\min(n_+, n_-)}{10} \tag{7.1}$$

Il faut donc réduire le nombre de variables. Reste à savoir comment. La sélection manuelle est une solution possible. En se basant sur le test de Wald ou le test du rapport de vraisemblance, l'expert peut choisir le meilleur sous-ensemble, en accord avec les connaissances du domaine. Idéale dans l'absolu, cette stratégie n'est pas tenable en pratique, surtout lorsque nous avons à traiter de grandes bases de données avec un nombre considérable de variables explicatives potentielles (quelques centaines habituellement dans les bases de données marketing). Il nous faut utiliser des procédures automatisées.

Dans ce chapitre, nous étudierons deux approches : la **sélection par optimisation** implémentée dans R, et la **sélection basée sur des critères statistiques** implémentée dans Tanagra. Tous deux se rejoignent sur le mode d'exploration de l'espace des solutions, il s'agit de procédures pas-à-pas qui évaluent une succession de modèles emboîtés : la sélection FORWARD part du modèle trivial, puis rajoute une à une les variables explicatives jusqu'à ce que l'on déclenche la règle d'arrêt ; la sélection BACKWARD part du modèle complet, incluant la totalité des descripteurs, puis enlève une à une les variables non significatives ; R, de plus, dispose de la méthode STEPWISE (qu'elle appelle BOTH), elle alterne forward et backward, elle consiste à vérifier si chaque ajout de variable ne provoque pas le retrait d'une explicative qui aurait été intégrée précédemment.

Nous le disons encore une fois, **ces techniques numériques nous proposent des scénarios de solutions**. Il ne faut surtout pas prendre pour argent comptant les sous-ensembles de variables explicatives proposées. D'autant qu'ils peuvent varier d'une stratégie à une autre, et même d'un échantillon d'apprentissage à un autre. Il faut plutôt les considérer comme des alternatives que l'on peut soumettre et faire valider par un expert du domaine. La sélection de variables est un maillon de la démarche exploratoire. Nous pouvons nous appuyer sur ses résultats pour essayer des combinaisons de variables, des transformations, réfléchir sur la pertinence de ce que l'on est en train de faire, etc.

1. On lira avec bonheur la section 8.5, pages 339 à 347, consacrée à la détermination d'une taille "suffisante" d'échantillon dans le même ouvrage.

7.2 Sélection par optimisation

7.2.1 Principe de la sélection par optimisation

La sélection par optimisation consiste à trouver le sous-ensemble de variables prédictives qui minimise un critère. Celui-ci ne peut pas être la déviance. En effet elle diminue de manière mécanique lorsque l'on rajoute de nouvelles variables, à l'instar de la somme des carrés des résidus dans la régression linéaire. Il nous faut un critère qui contrebalance la réduction de la déviance, qui comptabilise la qualité de l'ajustement, par un indicateur qui comptabilise la complexité du modèle. Lorsque nous rajoutons des variables pertinentes, le critère doit continuer à décroître ; lorsque nous rajoutons des variables qui ne sont pas en rapport avec la prédiction ou qui sont redondantes par rapport aux variables déjà choisies, il doit augmenter.

Deux critères répondent à ces spécifications. Le critère AIC d'Akaike

$$AIC = -2LL + 2 \times (J + 1) \tag{7.2}$$

et le critère BIC de Schwartz

$$BIC = -2LL + \ln(n) \times (J + 1) \tag{7.3}$$

où $-2LL$ est la déviance ; $(J+1)$ est le nombre de paramètres à estimer, avec J le nombre de variables explicatives.

Quelques remarques avant de passer à un exemple illustratif :

- Ces deux critères sont assez similaires finalement. BIC pénalise plus la complexité du modèle dès que l'effectif n augmente (dès que $\ln(n) > 2$). Ca ne veut pas dire qu'il est meilleur ou moins bon. Il privilégie simplement les solutions avec moins de variables explicatives par rapport à AIC.
- Selon la stratégie de recherche (forward, backward, stepwise), nous pouvons aboutir à des sous-ensembles différents.
- Ce n'est pas parce que la variable a été sélectionnée via cette procédure d'optimisation qu'elle sera significative au sens du test du rapport de vraisemblance ou du test de Wald dans la régression. Cela entraîne souvent le praticien dans un abîme de perplexité. Mais ce n'est pas du tout étonnant à bien y regarder. Les critères utilisés ne sont pas les mêmes. La conduite à tenir dépend des objectifs de notre étude.

7.2.2 Sélection de variables avec R

Nous utilisons un nouvel ensemble de données dans ce chapitre. Il s'agit toujours de prédire l'occurrence d'une maladie cardiaque (HEART) (décidément !) à l'aide de 10 variables explicatives candidates. Nous disposons de $n = 208$ observations, avec $n_+ = 117$ négatifs (absence) et $n_- = 91$ positifs (présence).

Nous utilisons le logiciel R qui, avec la commande stepAIC du package MASS, implémente la sélection de variables par optimisation[2]. Le code source des commandes décrites ci-dessous est livré avec ce document dans une archive à part (Annexe B).

Avant de lancer les calculs, nous devons spécifier les explicatives du modèle trivial (il n'y en a pas) et celles du modèle complet (toutes). Dans R, nous définissons deux variables de type *chaîne de caractères* pour les décrire

```
#modèle trivial réduit à la constante
str_constant <- "~ 1"
#modèle complet incluant toutes les explicatives potentielles
str_full <- "~ age+restbpress + max_hrate + chest_pain_asympt_1 +
chest_pain_atyp_angina_1 + chest_pain_non_anginal_1 + blood_sugar_f_1 +
restecg_normal_1 + restecg_left_vent_hyper_1 + exercice_angina_yes_1"
```

Sélection FORWARD

Pour initier une sélection forward, nous utilisons la commande `stepAIC`. Elle utilise par défaut le critère AIC, mais nous pouvons le paramétrer de manière à ce qu'elle optimise le critère BIC. Le modèle constitué uniquement de la constante (`modele`) sert de point de départ. `stepAIC` lance la procédure de recherche, et `modele.forward` réceptionne la régression finale intégrant les variables sélectionnées.

```
#départ modele avec la seule constante + sélection forward
modele <- glm(heart ~ 1, data = donnees, family = binomial)
modele.forward <- stepAIC(modele,scope = list(lower = str_constant,
upper = str_full), trace = TRUE, data = donnees, direction = "forward")
#affichage du modèle final
summary(modele.forward)
```

Disséquons les sorties de R durant le processus de recherche, on s'en tiendra uniquement aux trois premières étapes (Figure 7.1)

1. Initialement, nous avons $AIC = 287.09$ pour le modèle trivial.
2. R cherche le modèle à 1 variable qui minimise l'AIC. Il affiche toutes les configurations qu'il a testées et les trie selon l'AIC croissant :
   - `heart = f(chest_pain_asympt_1)` $\rightarrow AIC = 211.86$
   - `heart = f(exercice_angina_ yes_1)` $\rightarrow AIC = 214.88$
   - ...

   Notons que les variables qui viennent après `<none>` proposent un modèle pire, c.-à-d. l'AIC est plus élevé, que le modèle courant (le modèle trivial ici).

   Au final, R a intégré la première variable de la liste "chest_pain_asympt_1". Il essaie de voir quelle serait la seconde meilleure variable qu'il pourrait lui adjoindre.

2. Pour ceux qui ne sont pas très familiarisés avec R, vous trouverez très facilement de la documentation sur le web, entre autres, celles que j'ai rassemblées sur mon site de cours `http://eric.univ-lyon2.fr/~ricco/cours` *glm()* est la fonction qui permet de réaliser une régression logistique

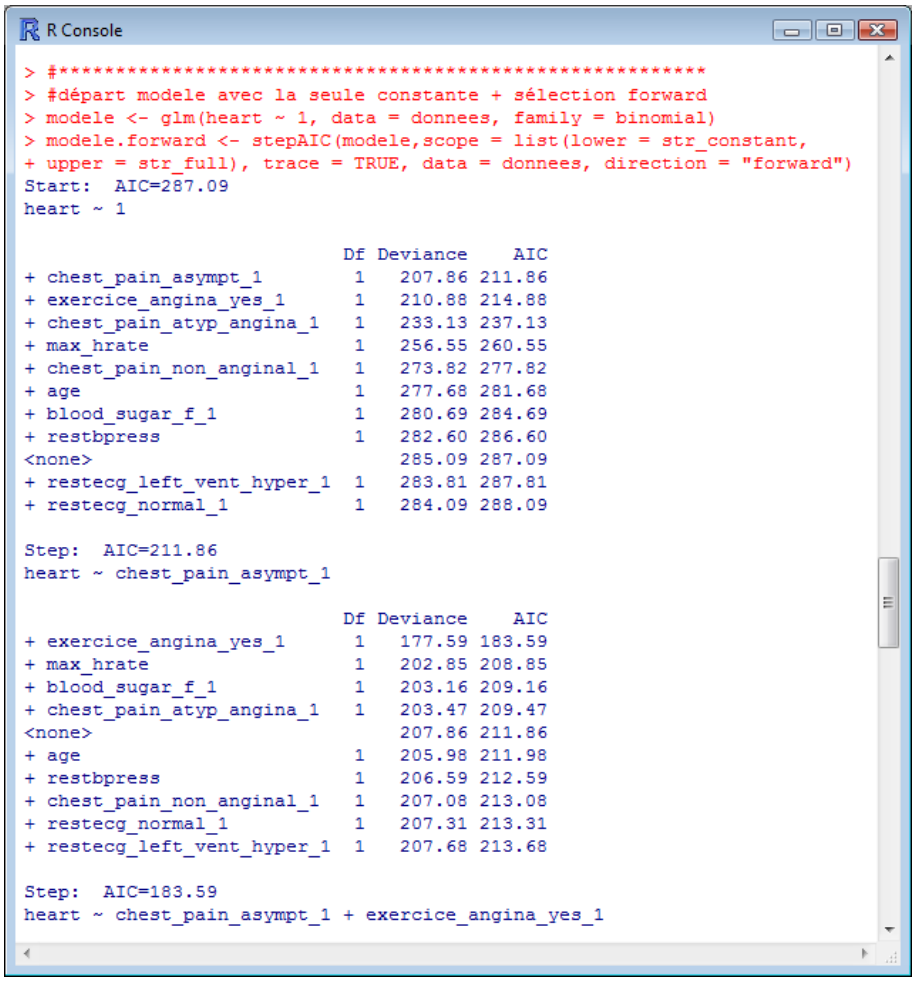


Fig. 7.1. Processus de sélection de variables - `stepAIC` de R - Forward

3. C'est reparti pour un tour. Il teste tous les modèles à deux variables, sachant que "chest_pain_asympt_1" ne peut plus être remis en cause
   - `heart = f(chest_pain_asympt_1,exercice_angina_yes_1)` $\rightarrow AIC = 183.59$
   - `heart = f(chest_pain_asympt_1,max_rate)` $\rightarrow AIC = 208.85$
   - ...

   Le meilleur modèle à 2 variables présentant un AIC (183.59) plus faible que le précédent à 1 variable (211.86), la variable "exercice_angina_yes_1" est acceptée.
4. Le processus se poursuit tant que l'on réduit le critère AIC. Dès que le critère stagne ou repart à la hausse, le processus de recherche est stoppé.

Au final, 5 variables explicatives sont sélectionnées. Dans le modèle qui en découle, nous constatons avec surprise que 2 d'entre elles (`chest_pain_asympt_1` et `blood_sugar_f_1`) ne sont pas significatives au sens du test de Wald à 5% (Figure 7.2). Cela rejoint la remarque que nous avions formulée plus haut : une variable peut être intégrée au sens du critère AIC, sans pour autant être significative au sens du test de Wald ou du rapport de vraisemblance.

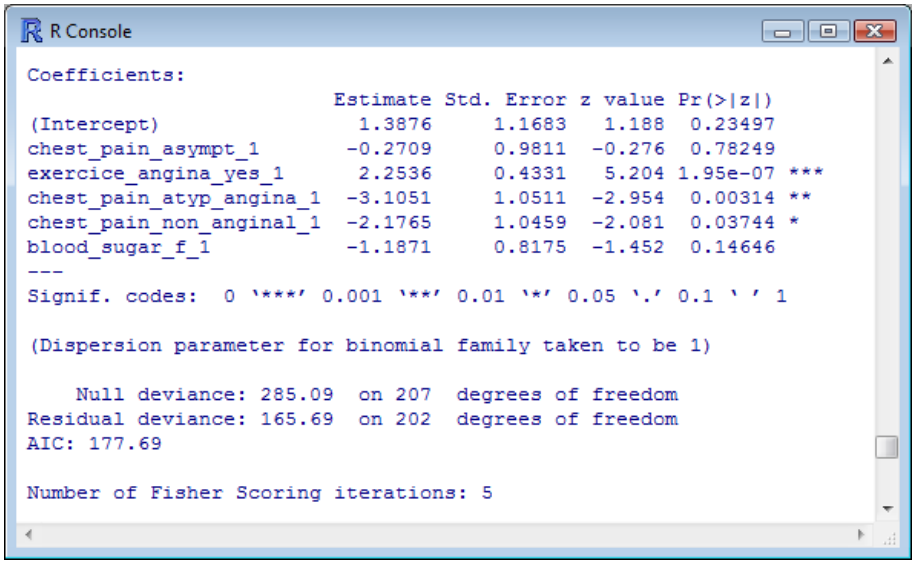


Fig. 7.2. Modèle sélectionné par le module `stepAIC` de R - Option Forward

Sélection BACKWARD

La sélection backward agit exactement à l'inverse du forward : R part du modèle incluant toutes les variables, il les enlève au fur et à mesure tant que le critère AIC décroît. Le processus est stoppé dès que l'AIC stagne ou augmente.

Voici les commandes pour R

```
modele <- glm(paste("heart",str_full), data = donnees, family = binomial)
modele.backward <- stepAIC(modele,scope = list(lower = str_constant,
upper = str_full), trace = TRUE, data = donnees, direction = "backward")
#affichage
summary(modele.backward)
```

Détaillons les premières étapes (Figure 7.3) :

1. Le modèle incluant les 10 variables propose un $AIC = 186.48$.
2. R teste le retrait de chaque variable explicative. Elles sont affichées selon un AIC croissant (le meilleur est celui qui propose l'AIC le plus faible)
   - Si on retire `restecg_normal_1`, le modèle à 9 variables qui en résulte présentera un AIC de 184.49.
   - Si on retire `restbpress`, nous aurons $AIC = 184.53$
   - Etc.

   Celle qui faudrait supprimer est `restecg_normal_1`, l'AIC du modèle à 9 variables est plus faible que le modèle précédent à 10 variables. Le retrait est entériné.
3. A partir de la configuration à 9 variables, R teste tous les modèles à 8 variables en retirant tour à tour chaque explicative. Il apparaît que la suppression de `restbpress` améliore encore le résultat avec $AIC = 182.53$. Elle est supprimée.

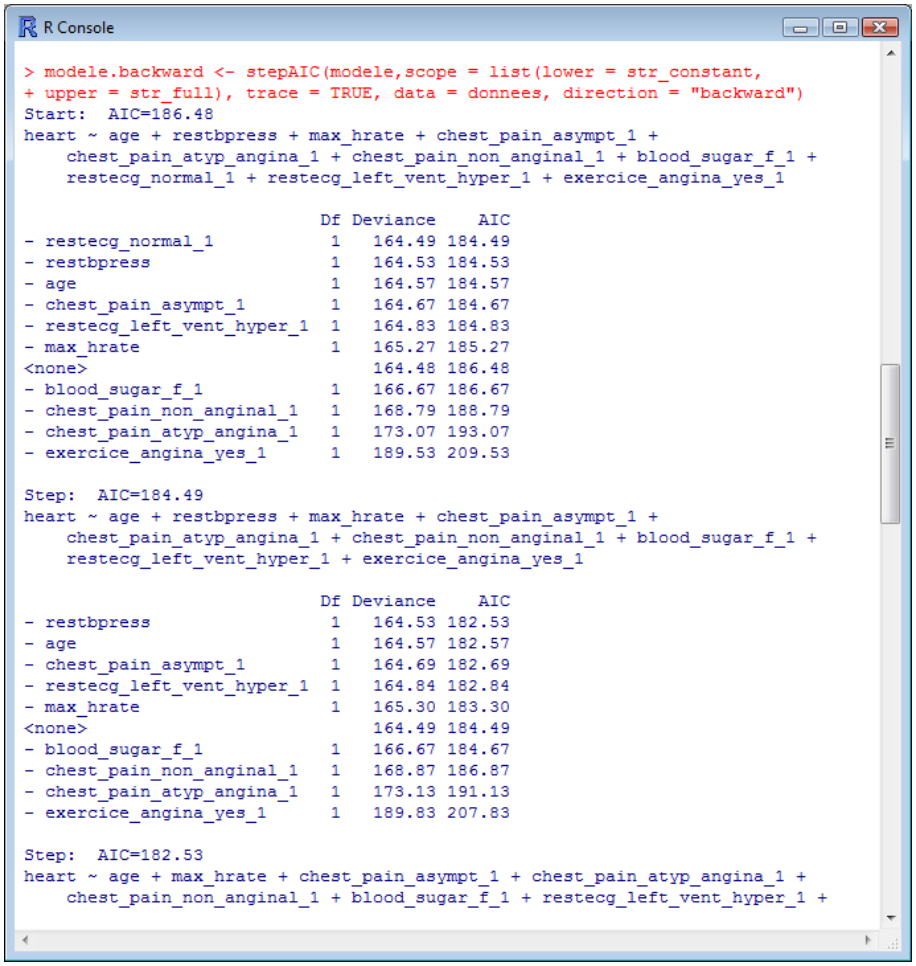


Fig. 7.3. Processus de sélection de variables - `stepAIC` de R - Backward

4. Etc.

Finalement, un modèle à 4 variables explicatives est mis en avant (Figure 7.4). Nous noterons plusieurs choses : nous n'obtenons pas le même sous ensemble de variables, il y en 4 pour l'option backward, il y en avait 5 pour le forward, `chest_pain_asympt_1` a disparu (corps et biens) ; et pourtant l'AIC de backward (175.77) est meilleur que celui de forward (177.69) ; enfin, parmi les variables retenues, certaines s'avèrent non-significatives au sens du test de Wald.

Sélection BOTH

L'option BOTH est a priori plus performante que les deux précédentes parce qu'elle les mixe justement. Voyons ce qu'il en est avec R. Les commandes utilisées sont les suivantes

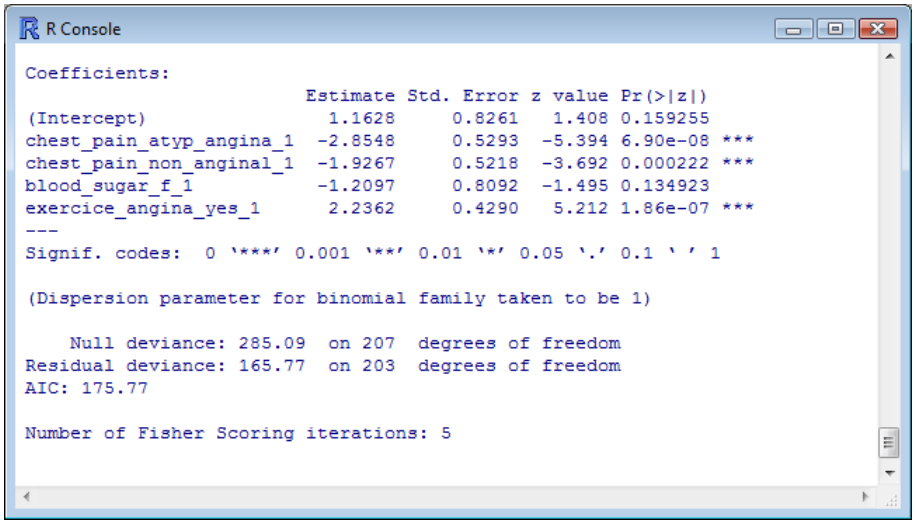


Fig. 7.4. Modèle sélectionné par le module `stepAIC` de R - Option Backward

```
#départ modele avec la seule constante + sélection both
modele <- glm(heart ~ 1, data = donnees, family = binomial)
modele.both <- stepAIC(modele,scope = list(lower = str_constant,
upper = str_full), trace = TRUE, data = donnees, direction = "both")
#affichage
summary(modele.both)
```

La recherche est un peu plus complexe dans ce cas. Nous ne rentrerons pas dans les détails. Allons directement sur le modèle final. Nous constatons que le sous-ensemble de variables retenu par l'option both (Figure 7.5) est le même que celui de l'option backward. Bref, avant de nous exciter inutilement sur les mérites de telle ou telle approche, prendre du recul par rapport aux résultats est toujours salutaire.

Utiliser le critère BIC

Dernier point pour conclure cette section, si nous passons au critère BIC qui pénalise plus la complexité du modèle, nous sélectionnons (presque) mécaniquement moins de variables. Nous l'avons testé avec l'option de recherche "both". On notera dans la fonction `stepAIC` le rôle du paramètre k. Pour obtenir le critère BIC, nous avons fixé $k = \ln(n) = \ln(208) = 5.34$: 2 variables seulement ont été sélectionnées (Figure 7.6). Ca ne veut pas dire que le classifieur est moins bon (ou meilleur). Il s'agit là simplement d'un autre scénario où l'on pénalise plus les modèles complexes. Cette attitude se justifie lorsque nous traitons des bases avec un nombre important de variables candidates dont une partie ne paraissent pas, de prime abord, pertinentes.

7.3 Sélection statistique

La sélection de variables ne peut pas se résumer à une affaire d'optimisation. C'est une démarche possible, mais elle n'est pas la seule. Les incohérences avec le test de Wald ou du rapport de vraisemblance

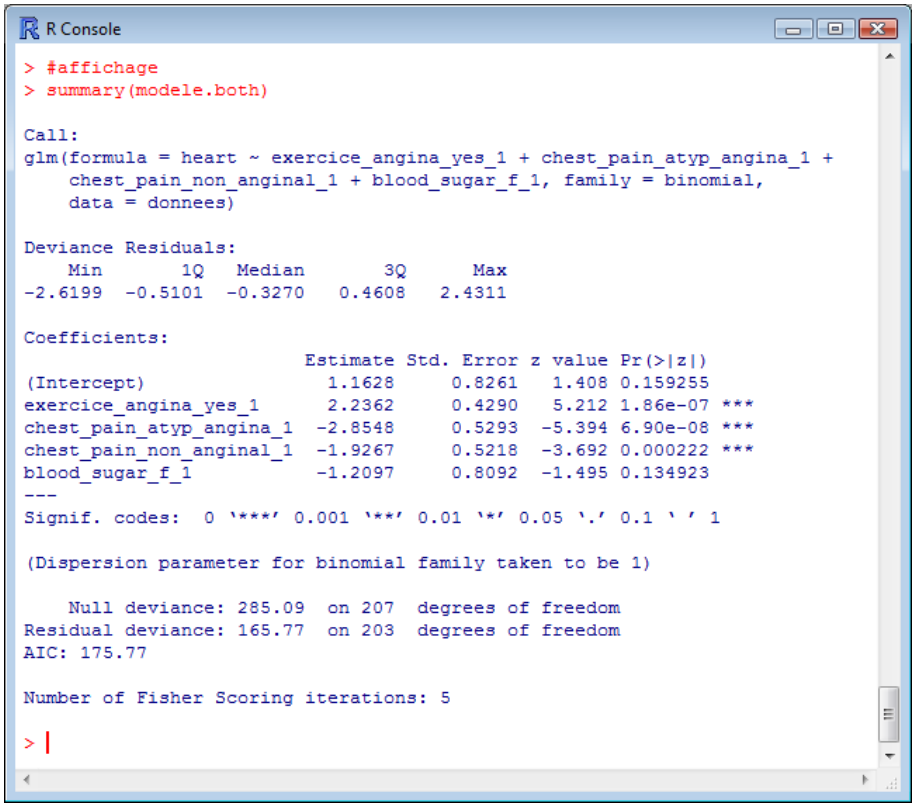


Fig. 7.5. Modèle sélectionné par le module `stepAIC` de R - Option Both

le montrent bien, des pistes alternatives en relation directe avec les tests de significativité des coefficients peuvent être explorées.

Un autre aspect important est le coût, en termes de temps de calcul, nécessaire à la sélection. Si l'on s'en tient à la procédure *forward*, le logiciel effectue J régressions à la première étape ; $(J-1)$ à la seconde ; etc. Dans le pire cas où toutes les variables sont finalement retenues, il aura réalisé $\frac{J(J+1)}{2}$ régressions (et autant d'optimisation de la fonction de log-vraisemblance). Même s'il faut reconnaître que R est particulièrement rapide, ce n'est pas tenable sur de très grandes bases de données.

Dans ce section, nous étudions les techniques de sélection exclusivement fondées sur les tests de significativité. Le stratégies d'exploration sont toujours les mêmes, *forward* et *backward*. L'énorme avantage est que nous construisons J régressions dans le pire des cas : retenir toutes les variables pour *forward*, supprimer toutes les variables pour *backward*. Commençons par l'option la plus facile, la sélection *backward* basée sur le test de Wald.

7.3.1 Sélection BACKWARD basée sur le Test de Wald

La procédure peut être résumée comme suit :

1. Commencer avec la totalité des variables.
2. Estimer les paramètres de la régression logistique.

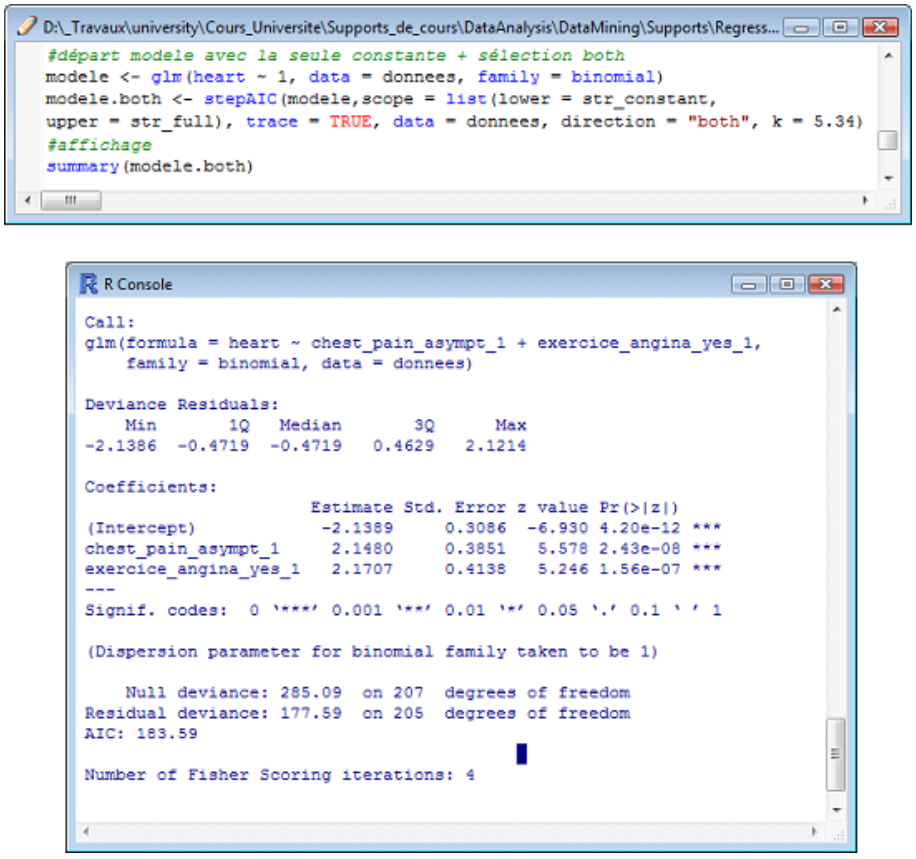


Fig. 7.6. Modèle sélectionné par `stepAIC` - Critère BIC avec $k = \ln(n) = 5.34$ - Option Both

3. Détecter parmi les coefficients celui qui présente la statistique de Wald la plus faible.
4. Vérifier s'il est non significatif en comparant la p-value du test avec le risque de première espèce α que l'on s'est choisi. Si p-value $\leq \alpha$, la variable est conservée. C'est l'arrêt du processus, l'ensemble de variables courant est la solution. Si p-value $> \alpha$, la variable est retirée de l'ensemble courant et, si ce dernier n'est pas vide, retour en [2], sinon c'est l'arrêt du processus, aucune variable n'aura été sélectionnée.

Quelques remarques concernant la démarche et les résultats obtenus :

- Il n'y a rien que l'on ne connaisse déjà dans tous les éléments qui composent ce processus. Nous ne sommes pas dépaysés.
- Toutes les variables retenues sont significatives au sens du test de Wald dans la régression finale. Il n'y a pas d'incohérences comme nous avions pu le constater lors l'optimisation de l'AIC.
- Dans le pire des cas, il n'y a que J régressions à opérer. Le temps de calcul est (à peu près) connu à l'avance.
- Par rapport au *forward*, la stratégie *backward* propose une propriété intéressante : elle prend mieux en compte les combinaisons de variables. En effet, il arrive qu'une variable explicative ne soit vraiment décisive qu'en présence d'une autre. Comme *backward* part de la totalité des variables, elle ne peut pas laisser passer ce type de situation [10] (page 64). A l'usage, on se rend compte qu'il n'y pas de différences réellement flagrantes entre ces deux stratégies sur des bases réelles.

- Lorsque le nombre de variables est très élevé (plusieurs centaines), les premières régressions risquent d'être problématiques. Il y a, entre autres, l'inversion de la matrice hessienne qui est délicate à mener, source de plantage des logiciels.
- Ceci est d'autant plus dommageable que dans la pratique, on ne retient que les modèles assez simples. Ils sont généralement composés au maximum d'une dizaine de variables pour des questions d'interprétation et de déploiement.
- Enfin, un statisticien vous dira tout de suite que le risque associé au test de significativité à l'étape [4] n'est certainement pas α. Chaque test est précédé d'un processus de détection de la variable la moins significative. Il faudrait corriger le véritable risque comme il est d'usage de le faire en comparaisons multiples. Le raisonnement tient la route, c'est indéniable. Mais je pense qu'il ne faut pas se tromper de cible. L'objectif n'est pas de forcer les données à cracher la vérité (si tant est qu'il y ait une vérité à cracher d'ailleurs), mais plutôt de mettre en évidence des scénarios de solutions. Le risque α joue le rôle de tournevis qui traduit nos préférences et que l'on adapte aux caractéristiques de la base traitée. Si l'on souhaite une solution avec peu de variables face un base très bruitée, on peut littéralement serrer la vis (réduire α) pour être plus exigeant avec le sous-ensemble final et obtenir moins de variables. A contrario, sur une petite base, avec des variables qui ont été soigneusement choisies par le praticien, être plus permissif paraît plus judicieux (augmenter α).

Sélection *backward* sur la base HEART

Nous reprenons la même base HEART et nous la traitons à l'aide du logiciel Tanagra. Nous utilisons le composant BACKWARD-LOGIT. Voyons le contenu de la fenêtre de résultats (Figure 7.7) :

- Les tests sont automatiquement réalisés à 1% dans Tanagra. Nous pouvons le paramétrer.
- 3 variables ont été sélectionnées finalement : `chest_pain_atyp_angina_1`, `chest_pain_non_anginal_1` et `exercice_angina_yes_1`. Elles étaient déjà apparues lors de la sélection par optimisation de l'AIC.
- Le tableau dans la partie basse de la fenêtre détaille le processus de calcul, en s'en tenant uniquement aux 5 variables extrêmes.
- La régression avec la totalité des variables est globalement significative, la statistique du test du rapport de vraisemblance est égale à $LR = 120.61$, avec une p-value < 0.0001. Nous observons aussi l'$AIC = 186.48$.
- On constate qu'à l'étape 1, `restecg_normal_1` est la moins bonne variable avec une statistique de Wald de 0.006 ; la suivante est `restecg_normal_1` avec $W = 0.006$; puis `restbpress` avec $W = 0.046$; etc.
- La moins bonne variable, `restecg_normal_1`, n'est pas significative à 1% avec une p-value du test de Wald égale à 0.9398. Elle est retirée.
- La régression avec les 9 variables restantes est globalement significative avec $LR = 120.60$ (p-value < 0.0001) et $AIC = 184.49$. La moins bonne variable est `restbpress`, elle n'est pas significative, elle est donc retirée.

Backward-logit 1

Selected attributes' subset

N°	Selected atts
1	chest_pain_atyp_angina_1
2	chest_pain_non_anginal_1
3	exercice_angina_yes_1

Detailed results

N°	Current Reg.	Moved	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	AIC : 186.48 CHI-2 : 120.61 d.f. : 10 p-value : 0.0000	restecg_normal_1 Chi-2 : 0.006 p : 0.9398	restecg_normal_1 Chi-2 : 0.006 p : 0.9398	restbpress Chi-2 : 0.046 p : 0.8298	age Chi-2 : 0.086 p : 0.7694	chest_pain_asympt_1 Chi-2 : 0.187 p : 0.6652	restecg_left_vent_hyper_1 Chi-2 : 0.312 p : 0.5767
2	AIC : 184.49 CHI-2 : 120.60 d.f. : 9 p-value : 0.0000	restbpress Chi-2 : 0.042 p : 0.8367	restbpress Chi-2 : 0.042 p : 0.8367	age Chi-2 : 0.085 p : 0.7709	chest_pain_asympt_1 Chi-2 : 0.196 p : 0.6580	restecg_left_vent_hyper_1 Chi-2 : 0.307 p : 0.5793	max_hrate Chi-2 : 0.811 p : 0.3679
3	AIC : 182.53 CHI-2 : 120.56 d.f. : 8 p-value : 0.0000	age Chi-2 : 0.102 p : 0.7489	age Chi-2 : 0.102 p : 0.7489	chest_pain_asympt_1 Chi-2 : 0.205 p : 0.6509	restecg_left_vent_hyper_1 Chi-2 : 0.295 p : 0.5872	max_hrate Chi-2 : 0.787 p : 0.3751	blood_sugar_f_1 Chi-2 : 2.074 p : 0.1499
4	AIC : 180.63 CHI-2 : 120.46 d.f. : 7 p-value : 0.0000	chest_pain_asympt_1 Chi-2 : 0.200 p : 0.6550	chest_pain_asympt_1 Chi-2 : 0.200 p : 0.6550	restecg_left_vent_hyper_1 Chi-2 : 0.317 p : 0.5737	max_hrate Chi-2 : 0.692 p : 0.4056	blood_sugar_f_1 Chi-2 : 2.002 p : 0.1571	chest_pain_non_anginal_1 Chi-2 : 4.489 p : 0.0341
5	AIC : 178.84 CHI-2 : 120.25 d.f. : 6 p-value : 0.0000	restecg_left_vent_hyper_1 Chi-2 : 0.323 p : 0.5695	restecg_left_vent_hyper_1 Chi-2 : 0.323 p : 0.5695	max_hrate Chi-2 : 0.561 p : 0.4537	blood_sugar_f_1 Chi-2 : 2.240 p : 0.1344	chest_pain_non_anginal_1 Chi-2 : 11.830 p : 0.0006	exercice_angina_yes_1 Chi-2 : 22.828 p : 0.0000
6	AIC : 177.21 CHI-2 : 119.88 d.f. : 5 p-value : 0.0000	max_hrate Chi-2 : 0.560 p : 0.4543	max_hrate Chi-2 : 0.560 p : 0.4543	blood_sugar_f_1 Chi-2 : 2.096 p : 0.1476	chest_pain_non_anginal_1 Chi-2 : 12.431 p : 0.0004	exercice_angina_yes_1 Chi-2 : 22.852 p : 0.0000	chest_pain_atyp_angina_1 Chi-2 : 26.978 p : 0.0000
7	AIC : 175.77 CHI-2 : 119.32 d.f. : 4 p-value : 0.0000	blood_sugar_f_1 Chi-2 : 2.235 p : 0.1349	blood_sugar_f_1 Chi-2 : 2.235 p : 0.1349	chest_pain_non_anginal_1 Chi-2 : 13.634 p : 0.0002	exercice_angina_yes_1 Chi-2 : 27.169 p : 0.0000	chest_pain_atyp_angina_1 Chi-2 : 29.093 p : 0.0000	-
8	AIC : 176.07 CHI-2 : 117.02 d.f. : 3 p-value : 0.0000	-	chest_pain_non_anginal_1 Chi-2 : 13.359 p : 0.0003	exercice_angina_yes_1 Chi-2 : 28.445 p : 0.0000	chest_pain_atyp_angina_1 Chi-2 : 29.136 p : 0.0000	-	-

Fig. 7.7. Processus de sélection *backward* basée sur le test de Wald - Tanagra

– Ainsi de suite jusqu'à la ligne $n^o 8$, nous constatons que la moins bonne variable au sens de la statistique de Wald, `chest_pain_non_anginal_1`, ne peut pas être retirée parce qu'elle est significative (p-value = 0.0003).

Supervised Learning 2 (Binary logistic regression)

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	0.006950	-	-	-
chest_pain_atyp_angina_1	-2.790023	0.5169	29.1356	0.0000
chest_pain_non_anginal_1	-1.905038	0.5212	13.3591	0.0003
exercice_angina_yes_1	2.270263	0.4257	28.4447	0.0000

Fig. 7.8. Régression sur les variables sélectionnées - Backward basé sur le test de Wald

La régression sur les 3 variables retenues nous donne un résultat (Figure 7.8) que l'on pouvait déjà deviner dans la dernière ligne du tableau de sélection.

7.3.2 Sélection FORWARD basée sur le Test du Score

La recherche *forward* fonctionne de la manière suivante :

1. Construire le modèle initial c.-à-d. réaliser la régression avec exclusivement la constante, sans aucune variable explicative.
2. Parmi les variables candidates, détecter celle qui maximise une statistique lorsque nous la rajoutons au modèle courant.
3. Vérifier si elle est significative c.à-d. p-value $\leq \alpha$. Si oui, intégrer la variable dans le modèle puis estimer les paramètres de la régression. S'il reste des variables candidates, retour en [2]. Si la variable n'est pas significative, elle n'est pas sélectionnée. Fin du processus.

L'étape $n^o 2$ est cruciale dans le processus. Si l'on voulait utiliser le test de Wald pour passer du modèle à p variables à celui comportant $p+1$ explicatives, il faudrait réaliser $J-p$ régressions et choisir celle qui maximise la statistique. Avec un temps de calcul qui peut se révéler prohibitif sur les grandes bases. Pour éviter cet écueil, nous utilisons un autre test de significativité des coefficients : le test du score.

Principe du Test du Score

Le test du score permet de tester la nullité simultanée de q coefficients. Il répond aux mêmes spécifications que les tests que nous avons étudiés dans le chapitre 3. Les hypothèses s'écrivent [3]

$$H_0 : a_{p+1} = \cdots = a_{p+q} = 0$$
$$H_1 : \text{un des coefficients est non nul}$$

L'énorme différence par rapport au test du rapport de vraisemblance et au test de Wald est que nous nous appuyons sur les résultats de la régression sous H_0 portant sur p variables. **Les q variables pour lesquelles nous voulons tester la significativité des coefficients sont traitées comme des variables supplémentaires.**

La statistique de test s'écrit :

$$S = U' H^{-1} U \tag{7.4}$$

Où U est le vecteur gradient de taille $(p+q+1) \times 1$, avec pour la composante j

3. Les variables ne sont pas forcément consécutives dans le modèle. Nous cherchons simplement à simplifier l'écriture ici.

$$U_j = \sum_{\omega} [y(\omega) - \hat{\pi}(\omega)] x_j(\omega) \tag{7.5}$$

H est la matrice hessienne de taille $(p+q+1) \times (p+q+1)$, avec pour la composante (j_1, j_2)

$$H(j_1, j_2) = \sum_{\omega} x_{j_1}(\omega) x_{j_2}(\omega) \hat{\pi}(\omega)(1 - \hat{\pi}(\omega)) \tag{7.6}$$

Sous H_0, la quantité S suit une moi du χ^2 à q degrés de liberté.

Le vecteur gradient U dans l'expression 7.4 peut paraître étrange. En effet, les paramètres de la régression ayant maximisé la log-vraisemblance, toutes les composantes de U devraient être nuls. De fait, S devrait toujours être égal à 0. La réponse est non, U est non nul, parce que **les prédictions $\hat{\pi}(\omega)$ sont produites à l'aide du modèle à p variables**.

Exemple : Reprenons le fichier COEUR (Figure 0.1) pour illustrer la procédure. Nous réalisons la régression COEUR = f (TAUX MAX). Nous souhaitons savoir si l'adjonction de la variable AGE produirait un coefficient significatif.

Dans un premier temps (Figure 7.9), nous optimisons la vraisemblance avec la variable TAUX MAX et la constante (en vert). AGE n'est pas utilisée à ce stade. Nous obtenons l'équation du LOGIT

$$C(X) = 8.7484 - 0.0627 \times taux\ max$$

A partir de ce résultat, nous obtenons la colonne $\hat{C}$ dans la feuille Excel, puis la colonne $\hat{\pi}$.

La formule 7.5 nous permet de compléter le vecteur gradient, nous trouvons les composantes :

$$\begin{aligned} U_{const} &= 0 \\ U_{taux\ max} &= 0 \\ U_{age} &= -22.6863 \end{aligned}$$

Les deux premiers termes sont nuls. En effet, ils ont participé à la maximisation de la vraisemblance. Il est tout à fait normal que les dérivées partielles premières soient nuls. Il en est tout autrement pour AGE. Il n'a pas participé à l'optimisation. Lorsque nous calculons son score, nous obtenons une valeur différente de 0, en l'occurrence $U_{age} = -22.6863$.

A l'aide de la formule 7.6, nous calculons la matrice hessienne [4]

$$H = \begin{pmatrix} 3.41 & 501.80 & 177.41 \\ 501.80 & 74552.70 & 26056.59 \\ 177.41 & 26056.59 & 9440.07 \end{pmatrix}$$

4. Nous avons utilisé la forme matricielle dans la feuille Excel, $H = X'VX$, où V est la matrice diagonale de taille $(n \times n)$ de terme générique $\hat{\pi} \times (1 - \hat{\pi})$.

Coef							
8.4784	-0.0627						
const.	taux_max	age	coeur	cœur	C(X)	π	LL
1	126	50	presence	1	0.584	0.642	-0.443
1	126	49	presence	1	0.584	0.642	-0.443
1	144	46	presence	1	-0.544	0.367	-1.002
1	139	49	presence	1	-0.231	0.443	-0.815
1	154	62	presence	1	-1.170	0.237	-1.441
1	156	35	presence	1	-1.296	0.215	-1.538
1	160	67	absence	0	-1.546	0.176	-0.193
1	140	65	absence	0	-0.293	0.427	-0.557
1	143	47	absence	0	-0.481	0.382	-0.481
1	165	58	absence	0	-1.860	0.135	-0.145
1	115	57	absence	0	1.273	0.781	-1.520
1	145	59	absence	0	-0.607	0.353	-0.435
1	175	44	absence	0	-2.486	0.077	-0.080
1	153	41	absence	0	-1.108	0.248	-0.285
1	152	54	absence	0	-1.045	0.260	-0.301
1	169	52	absence	0	-2.110	0.108	-0.114
1	168	57	absence	0	-2.048	0.114	-0.121
1	158	50	absence	0	-1.421	0.194	-0.216
1	170	44	absence	0	-2.173	0.102	-0.108
1	171	49	absence	0	-2.236	0.097	-0.102
						-2LL	20.682

0.0000	0.0000	-22.6863
U(const)	U(taux_max)	U(age)
0.36	45.11	17.90
0.36	45.11	17.54
0.63	91.11	29.11
0.56	77.48	27.31
0.76	117.54	47.32
0.79	122.48	27.48
-0.18	-28.10	-11.77
-0.43	-59.81	-27.77
-0.38	-54.62	-17.95
-0.13	-22.23	-7.81
-0.78	-89.85	-44.53
-0.35	-51.16	-20.82
-0.08	-13.45	-3.38
-0.25	-37.99	-10.18
-0.26	-39.54	-14.05
-0.11	-18.27	-5.62
-0.11	-19.20	-6.51
-0.19	-30.73	-9.72
-0.10	-17.38	-4.50
-0.10	-16.52	-4.73

H=X'VX

3.41	501.80	177.41
501.80	74552.70	26056.59
177.41	26056.59	9440.07

SIGMA=H^(-1)

43.461	-0.200	-0.265
-0.200	0.001	0.000
-0.265	0.000	0.005

Statistique du score pour age

Valeur	2.3766
d.d.l	1
p-value	0.1232

Fig. 7.9. Construction du test de score - Tester la variable supplémentaire AGE

que nous inversons

$$H' = \begin{pmatrix} 43.461 & -0.200 & -0.265 \\ -0.200 & 0.001 & 0.000 \\ -0.265 & 0.000 & 0.005 \end{pmatrix}$$

Il ne reste plus qu'à calculer S

$$S = \begin{pmatrix} 0.0000 & 0.0000 & -22.6863 \end{pmatrix} \times \begin{pmatrix} 43.461 & -0.200 & -0.265 \\ -0.200 & 0.001 & 0.000 \\ -0.265 & 0.000 & 0.005 \end{pmatrix} \times \begin{pmatrix} 0.0000 \\ 0.0000 \\ -22.6863 \end{pmatrix} = 2.3766$$

Avec la fonction de répartition du χ^2 à 1 degré de liberté, nous obtenons une p-value = 0.1232. Le coefficient de AGE n'est pas significatif à 10% si on l'ajoutait dans la régression.

A titre d'information, si on s'ingénie à introduire quand même la variable AGE dans la régression, la statistique de Wald serait égale à 2.0302 avec un p-value de 0.1542 (Figure 7.10). Le résultat est cohérent avec le test du score. On montre dans la littérature qu'il existe une passerelle entre ces deux tests [7] (page 110).

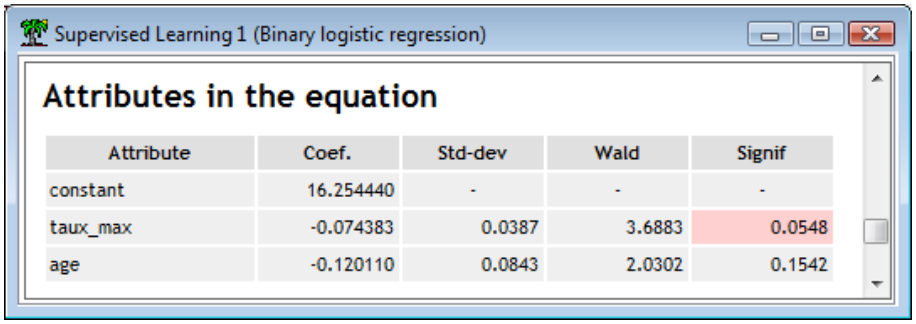


Fig. 7.10. Test de Wald si la variable AGE est introduite dans la régression

Implémentation de la sélection avec le test du score

Revenons à la sélection *forward* de variables. Nous testons l'adjonction d'une variable supplémentaire dans la modèle. Le test d'hypothèses s'écrit

$$H_0 : a_{p+1} = 0$$
$$H_1 : a_{p+1} \neq 0$$

Nous pouvons maintenant détailler le processus complet [5]

1. $p = 0$.
2. Étape courante, nous réalisons la régression avec les p variables déjà sélectionnées (lorsque $p = 0$, il n'y a que la constante dans le modèle).
3. Pour les $J - p$ variables candidates. Calculer, en intégrant la $(p+1)^{ème}$ variable à évaluer comme variable supplémentaire

5. Dans SPSS, cette procédure est désignée par "METHOD - FORWARD : CONDITIONAL" dans les options de gestion des variables.

a) Le vecteur gradient U

b) La matrice hessienne H

c) Inverser la matrice hessienne $\rightarrow H^{-1}$

d) La statistique de test $S = U'H^{-1}U$

4. Choisir la variable qui maximise S. Vérifier que nous rejetons H_0 au risque α que l'on s'est choisi c.-à-d. p-value $< \alpha$. Si oui, l'ajouter dans l'ensemble des explicatives sélectionnées. S'il reste encore des variables candidiates, retour en [2]. Si non, le coefficient associé n'est pas significatif ou il n'y a plus de variables candidates, arrêt de la procédure.

Quelques remarques sur la stratégie *forward* basée sur le test du score :

- Premier avantage très intéressant, nous ne réalisons que J optimisations de la vraisemblance dans le pire des cas (toutes les variables sont finalement retenues).
- Il faut noter quand même que l'évaluation d'une variable induit une série de calculs non négligeables, notamment une inversion de matrice qui peut toujours être problématique.
- Il faut privilégier cette stratégie lorsque nous traitons une grande base de données, avec un grand nombre de variables candidates, alors que nous savons pertinemment que nous n'en retiendrons que quelques unes.
- Lors du test de significativité de la variable que l'on souhaite introduire à chaque étape, le véritable risque du test n'est pas vraiment égal au risque nominal α que l'on a choisi. Il est un peu plus grand. En effet, nous avons d'abord sélectionné la variable portant la statistique S la plus élevée avant de la tester. Mais encore une fois, il faut plutôt voir le paramètre α comme un outil de contrôle qui permet d'orienter l'algorithme vers les solutions qui conviennent compte tenu de nos objectifs et des caractéristiques des données.
- Attention, le test du score et le test de Wald sont similaires mais ne sont pas totalement identiques. Il se peut qu'une explicative validée par le test du score, n'apparaisse pas significative au sens du test de Wald lorsque nous réalisons la régression avec le sous-ensemble de variables sélectionnées.

Sélection *forward* sur la base HEART

Nous revenons sur le fichier HEART ($n = 208$) utilisé tout au long de ce chapitre consacré à la sélection de variables. A l'aide de Tanagra, nous réalisons la sélection *forward* basée sur le test du score avec $\alpha = 0.01$ (Figure 7.11) [6] :

- 2 variables seulement ont été sélectionnées, les mêmes que la stratégie de sélection par optimisation avec le critère BIC (Figure 7.6).
- Nous disposons du détail du processus dans le tableau. Le modèle initial est le modèle trivial composé uniquement de la constante. Bien évidemment, la statistique du test du rapport de vraisemblance évaluant le modèle global est $LR = 0$, le critère $AIC = 287.09$.

6. Le logiciel Tanagra propose une option qui permet de limiter arbitrairement le nombre de variables sélectionnées. Elle s'avère utile lorsque nous traitons des bases avec un très grand nombre de variables et que nous souhaitons obtenir un modèle volontairement simple.

ht]

Forward-logit 1

Selected attributes' subset

N°	Selected atts
1	chest_pain_asympt_1
2	exercice_angina_yes_1

Detailed results

N°	Current Reg.	Moved	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	AIC : 287.09 CHI-2 : 0.00 d.f. : 0 p-value : 0.0000	chest_pain_asympt_1 Chi-2 : 72.126 p : 0.0000	chest_pain_asympt_1 Chi-2 : 72.126 p : 0.0000	exercice_angina_yes_1 Chi-2 : 70.111 p : 0.0000	chest_pain_atyp_angina_1 Chi-2 : 45.779 p : 0.0000	max_hrate Chi-2 : 27.081 p : 0.0000	chest_pain_non_anginal_1 Chi-2 : 10.451 p : 0.0012
2	AIC : 211.86 CHI-2 : 77.23 d.f. : 1 p-value : 0.0000	exercice_angina_yes_1 Chi-2 : 32.078 p : 0.0000	exercice_angina_yes_1 Chi-2 : 32.078 p : 0.0000	max_hrate Chi-2 : 5.050 p : 0.0246	blood_sugar_f_1 Chi-2 : 4.543 p : 0.0331	chest_pain_atyp_angina_1 Chi-2 : 4.509 p : 0.0337	age Chi-2 : 1.860 p : 0.1727
3	AIC : 183.59 CHI-2 : 107.50 d.f. : 2 p-value : 0.0000	-	chest_pain_atyp_angina_1 Chi-2 : 4.761 p : 0.0291	blood_sugar_f_1 Chi-2 : 2.861 p : 0.0908	chest_pain_non_anginal_1 Chi-2 : 0.528 p : 0.4673	max_hrate Chi-2 : 0.521 p : 0.4703	restecg_left_vent_hyper_1 Chi-2 : 0.225 p : 0.6355

Fig. 7.11. Processus de sélection *forward* - Test du Score

- La meilleure variable que l'on pourrait introduire au sens du test du score est `chest_pain_asympt_1` avec $S = 72.126$; la seconde est `exercice_angina_yes_1` avec $S = 70.111$; etc. La première est largement significative avec une p-value < 0.0001. Elle est donc entérinée.
- La régression $heart = f(chest\_pain\_asympt\_1)$ est globalement significative au sens du test du rapport de vraisemblance, avec $LR = 77.23$. Tanagra cherche à introduire une seconde variable.
- La meilleure est `exercice_angina_yes_1`, avec une statistique du score = 32.078 et une p-value < 0.0001. Elle est également sélectionnée.
- La régression avec les deux variables déjà introduites est globalement significative, la statistique du test de rapport de vraisemblance est $LR = 107.50$. Lorsque Tanagra essaie de rajouter une $3^{ème}$ variable, la meilleure est `chest_pain_atyp_angina_1` avec $S = 4.761$. Mais elle n'est pas significative avec $p - value = 0.0291 > \alpha = 0.01$. Le processus est stoppé.

Supervised Learning 1 (Binary logistic regression)

R²-like	
McFadden's R²	0.3771
Cox and Snell's R²	0.4036
Nagelkerke's R²	0.5410

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-2.138928	-	-	-
chest_pain_asympt_1	2.147992	0.3851	31.1131	0.0000
exercice_angina_yes_1	2.170657	0.4138	27.5154	0.0000

Fig. 7.12. Régression sur les variables sélectionnées par le test du score

A titre de curiosité, nous donnons la régression fournie par Tanagra sur ces deux variables explicatives (Figure 7.12). Les coefficients associés sont tous deux fortement significatifs au sens du test de Wald.

8

Diagnostic de la régression logistique

8.1 Analyse des résidus

L'analyse des résidus permet de diagnostiquer la qualité de la régression. Plusieurs questions se posent à l'issue du processus de modélisation, nous devons y apporter des réponses :

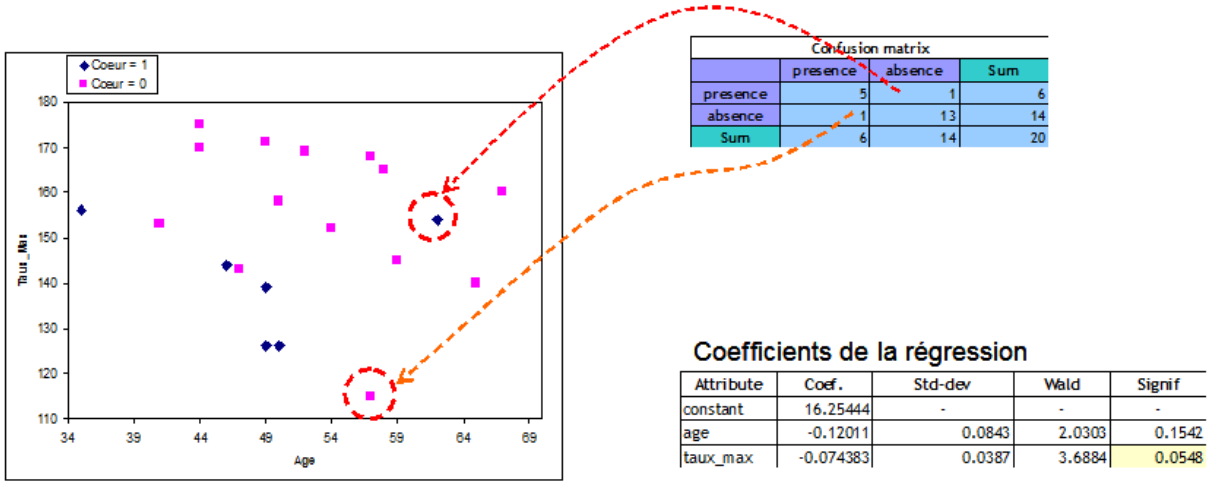
1. Déterminer les points qui "clochent" dans les données, qui s'écartent fortement des autres dans l'espace de représentation. On parle de données "atypiques".
2. Déterminer les points qui sont mal modélisés (mal expliqués) par la régression logistique. On parle de résidus.
3. Déterminer les points qui pèsent fortement dans la régression. On parle de points "leviers".
4. Déterminer les points qui pèsent exagérément sur les résultats. Si on les retirait de l'ensemble d'apprentissage, le modèle obtenu serait très différent. On parle de points "influents".

L'analyse des résidus telle que nous la présentons ici est surtout intéressante lorsque nous traitons des bases de taille modérée. Nous pouvons distinguer les individus, voire leur associer un label. Nous comprendrons mieux leur rôle dans la régression. Par exemple, nous modélisons l'acceptation ou le refus d'une compagnie d'assurance de prendre en charge un véhicule. On comprendra mieux le rôle d'un véhicule mal modélisé si l'on se rend compte qu'il s'agit d'une Aston Martin DB5 (celle de James Bond dans Goldfinger). Ce n'est visiblement pas une voiture comme les autres, son positionnement va largement au delà de ses qualités intrinsèques.

Dernière précision importante, nous présentons les concepts dans le cadre des données individuelles dans cette section. La probabilité que deux individus aient la même description est nulle. C'est le cas lorsqu'ils sont décrits par plusieurs variables continues. Les mêmes concepts (résidus, leviers, etc.) seront présentés dans le canevas des données groupées dans le chapitre 9. Plusieurs individus partagent la même description, appelée "covariate pattern". Dans cette configuration, l'analyse des résidus va au delà de son rôle habituel en régression, elle peut devenir une aide à l'interprétation.

Numéro	const	age	taux_max	coeur
1	1	50	126	1
2	1	49	126	1
3	1	46	144	1
4	1	49	139	1
5	1	62	154	1
6	1	35	156	1
7	1	67	160	0
8	1	65	140	0
9	1	47	143	0
10	1	58	165	0
11	1	57	115	0
12	1	59	145	0
13	1	44	175	0
14	1	41	153	0
15	1	54	152	0
16	1	52	169	0
17	1	57	168	0
18	1	50	158	0
19	1	44	170	0
20	1	49	171	0

Fig. 8.1. Fichier Coeur - Tableau de données, individus numérotés



Confusion matrix

	presence	absence	Sum
presence	5	1	6
absence	1	13	14
Sum	6	14	20

Coefficients de la régression

Attribute	Coef.	Std-dev	Wald	Signif
constant	16.25444	-	-	-
age	-0.12011	0.0843	2.0303	0.1542
taux_max	-0.074383	0.0387	3.6884	0.0548

Fig. 8.2. Coeur = f(age, taux max) - Nuage de points et résultats de la régression

8.1.1 Notre exemple de référence : $coeur = f(age, taux\ max)$

Tout au long de cette section, nous utiliserons les données COEUR (Figure 0.1). Nous modélisons la variable dépendante à l'aide des deux variables explicatives quantitatives AGE et TAUX MAX. Nous reproduisons le tableau de données en numérotant les observations pour que nous puissions les retrouver facilement dans les différents graphiques (Figure 8.1).

Nous avons projeté les observations dans le plan afin de visualiser le positionnement des points. Nous reproduisons les résultats de la régression logistique et la matrice de confusion. Nous constatons que 2 observations sont mal classés (mal modélisés). Un positif noyé au milieu des négatifs (individu n^o5), et un négatif (n^o11) qui est un peu éloigné du nuage global et, qui plus est, situé du mauvais côté de la barrière (nous en reparlerons plus longuement plus loin, voir section 11.3), mal modélisé également. Nous

les mettons en relation avec la matrice de confusion (Figure 8.2). Nous devons bien garder à l'esprit cette disposition des points. La compréhension des indicateurs qui viendront par la suite sera facilitée.

8.1.2 Résidus de Pearson et Résidus déviance

Résidus de Pearson

La modélisation de la variable $Y \in \{1, 0\}$ peut s'écrire sous la forme suivante

$$Y(\omega) = \pi(\omega) + \varepsilon(\omega) \tag{8.1}$$

$\varepsilon(\omega)$ est l'erreur de modélisation, avec $\varepsilon(\omega) = Y(\omega) - \pi(\omega)$, elle peut prendre deux valeurs possibles :

$$\begin{aligned}
&\varepsilon(\omega) = 1 - \pi(\omega) \text{ avec la probabilité } \pi(\omega) \\
&\varepsilon(\omega) = -\pi(\omega) \text{ avec la probabilité } 1 - \pi(\omega)
\end{aligned}$$

Nous calculons aisément :

$$\begin{aligned}
E(\varepsilon) &= \pi(1-\pi) + (1-\pi)(-\pi) = 0 \\
V(\varepsilon) &= \pi(1-\pi)
\end{aligned}$$

La variance de l'erreur n'est pas constante, elle dépend des individus. Il y a hétéroscédasticité.

Pour un individu ω, le résidu de Pearson permet d'identifier les points mal modélisés

$$r(\omega) = \frac{y(\omega) - \hat{\pi}(\omega)}{\sqrt{\hat{\pi}(\omega)(1 - \hat{\pi}(\omega))}} \tag{8.2}$$

Le résidu de Pearson prend une valeur d'autant plus élevée que $\hat{\pi}$ est proche de 0 ou de 1.

Certains auteurs affirment que la distribution de r est approximativement gaussienne $\mathcal{N}(0, 1)$. Ainsi, tout point en dehors de l'intervalle ± 2 (au niveau de confiance 95%) sont suspects [10] (page 82). D'autres pensent que cette approximation n'est licite que dans le cadre des données groupées, lorsque un nombre suffisamment élevé d'observations partagent la même description [9] (page 175). Notre opinion est qu'il ne faut pas trop se focaliser sur des hypothétiques valeurs seuils. Il est plus important de détecter les éventuels décrochements, les observations qui prennent des valeurs inhabituelles par rapport aux autres. Un graphique est très précieux pour cela.

Voyons ce qu'il en est du résidu de Pearson sur nos données COEUR. Construisons les 2 graphiques des résidus : (age, r) et $(taux\ max, r)$.

Pour obtenir les résidus de Pearson, nous avons d'abord estimé les paramètres de la régression, puis calculé les projections $\hat{C}$ et $\hat{\pi}$. Nous avons formé le terme d'erreur $e = y - \hat{\pi}$. Enfin, nous produisons le

	16.254	-0.120	-0.074						
Numéro	const	age	taux_max	coeur	C(X)	PI	PI x (1 - PI)	e	r.pears
1	1	50	126	1	0.88	0.706	0.208	0.294	0.645
2	1	49	126	1	1.00	0.730	0.197	0.270	0.608
3	1	46	144	1	0.02	0.505	0.250	0.495	0.991
4	1	49	139	1	0.03	0.507	0.250	0.493	0.985
5	1	62	154	1	-2.65	0.066	0.062	0.934	3.757
6	1	35	156	1	0.45	0.610	0.238	0.390	0.800
7	1	67	160	0	-3.69	0.024	0.024	-0.024	-0.158
8	1	65	140	0	-1.97	0.123	0.108	-0.123	-0.374
9	1	47	143	0	-0.03	0.493	0.250	-0.493	-0.986
10	1	58	165	0	-2.99	0.048	0.046	-0.048	-0.225
11	1	57	115	0	0.85	0.701	0.209	-0.701	-1.533
12	1	59	145	0	-1.62	0.166	0.138	-0.166	-0.445
13	1	44	175	0	-2.05	0.114	0.101	-0.114	-0.359
14	1	41	153	0	-0.05	0.487	0.250	-0.487	-0.975
15	1	54	152	0	-1.54	0.177	0.146	-0.177	-0.464
16	1	52	169	0	-2.56	0.072	0.066	-0.072	-0.278
17	1	57	168	0	-3.09	0.044	0.042	-0.044	-0.214
18	1	50	158	0	-1.50	0.182	0.149	-0.182	-0.472
19	1	44	170	0	-1.68	0.158	0.133	-0.158	-0.433
20	1	49	171	0	-2.35	0.087	0.079	-0.087	-0.309

Fig. 8.3. Coeur = f(age, taux max) - Tableau de calcul des résidus de Pearson

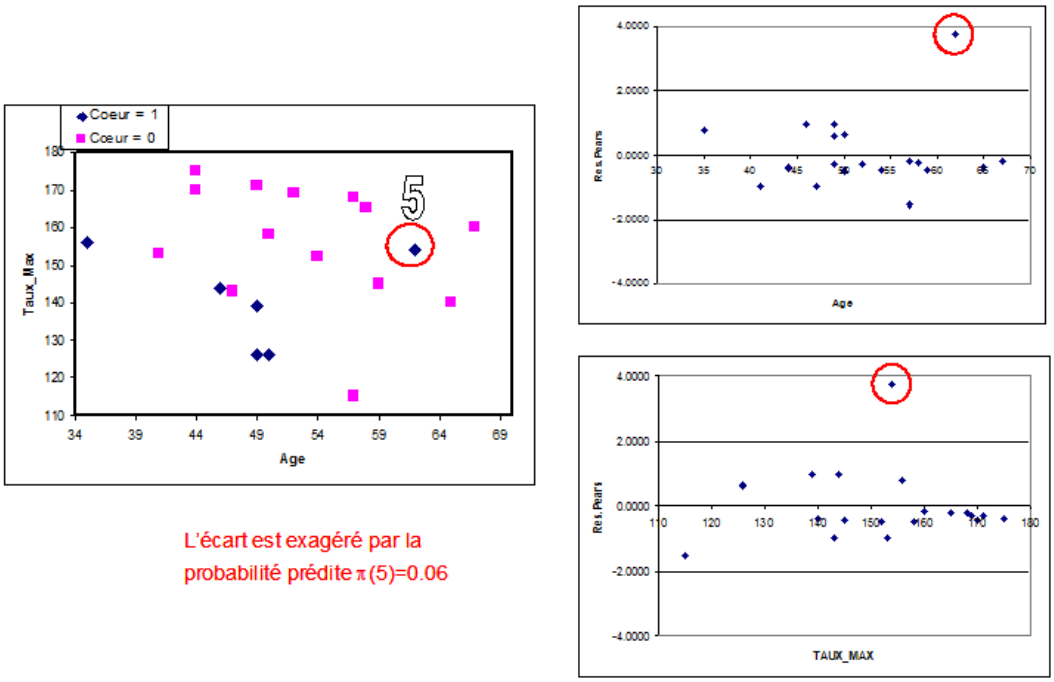


Fig. 8.4. Coeur = f(age, taux max) - Résidus de Pearson - Graphiques

résidu de Pearson (Figure 8.3). Les graphiques des résidus sont édifiants, surtout en les mettant en rapport avec le nuage de points dans l'espace de représentation : manifestement, le point n^o5 pose problème. Il est particulièrement mal modélisé (Figure 8.4). Le résidu $r(5) = 3.757$ prend une valeur d'autant plus extrême que $\hat{\pi}(5) = 0.06$. Le point n^o11, qui lui aussi est mal modélisé, se démarque moins parce que $\hat{\pi}(11) = 0.70$.

A partir du résidu de Pearson, nous pouvons dériver un indicateur, **la statistique χ^2 de Pearson**. Plus faible sera sa valeur, meilleure sera la régression.

$$\chi^2 = \sum_{\omega} r^2(\omega) \tag{8.3}$$

Certains auteurs comparent sa valeur avec un seuil critique issu de la loi du χ^2. Ce n'est pas très conseillé lorsque nous travaillons sur des données individuelles. L'approximation n'est pas très bonne, les p-value sont faussées [9] (page 146). Il en sera autrement lorsque nous traitons des données groupées (chapitre 9).

Résidus déviance

Le résidu déviance pour un individu ω est définie de la manière suivante

$$d(\omega) = \begin{cases} +\sqrt{2 \times |\ln(\hat{\pi}(\omega))|} & \text{si } y(\omega) = 1 \\ -\sqrt{2 \times |\ln(1 - \hat{\pi}(\omega))|} & \text{si } y(\omega) = 0 \end{cases} \tag{8.4}$$

Nous pouvons en déduire la statistique D appelée déviance

$$D = \sum_{\omega} d^2(\omega) \tag{8.5}$$

Sur les données individuelles, la déviance ainsi calculée coïncide avec la déviance du modèle D_M que nous avons présentée plus haut, lorsque nous décrivions les quantités à optimiser lors du processus d'apprentissage (cf. page 17).

Ici également, les distributions approximées, loi normale pour d et loi du χ^2 pour D, ne sont vraiment précises que dans le cadre des données groupées. On s'attachera avant tout à détecter les points qui "décrochent" par rapport aux autres.

Concernant le fichier COEUR, le point n^o5 mal modélisé se démarque encore dans les graphiques des résidus (Figure 8.5), moins fortement néanmoins qu'avec le résidu de Pearson.

8.1.3 Le levier

Levier, détecteur de points atypiques

Le levier d'une observation quantifie son écartement par rapport aux autres. Il permet de détecter les points atypiques dans un espace multivarié.

La *hat-matrix* est de dimension $(n \times n)$. Il s'écrit

$$H = V^{\frac{1}{2}} X (X'VX)^{-1} X' V^{\frac{1}{2}} \tag{8.6}$$

où X $((n \times (J+1))$ est la matrice des descripteurs incluant la constante, et V est la matrice diagonale des $\hat{\pi}(1 - \hat{\pi})$.

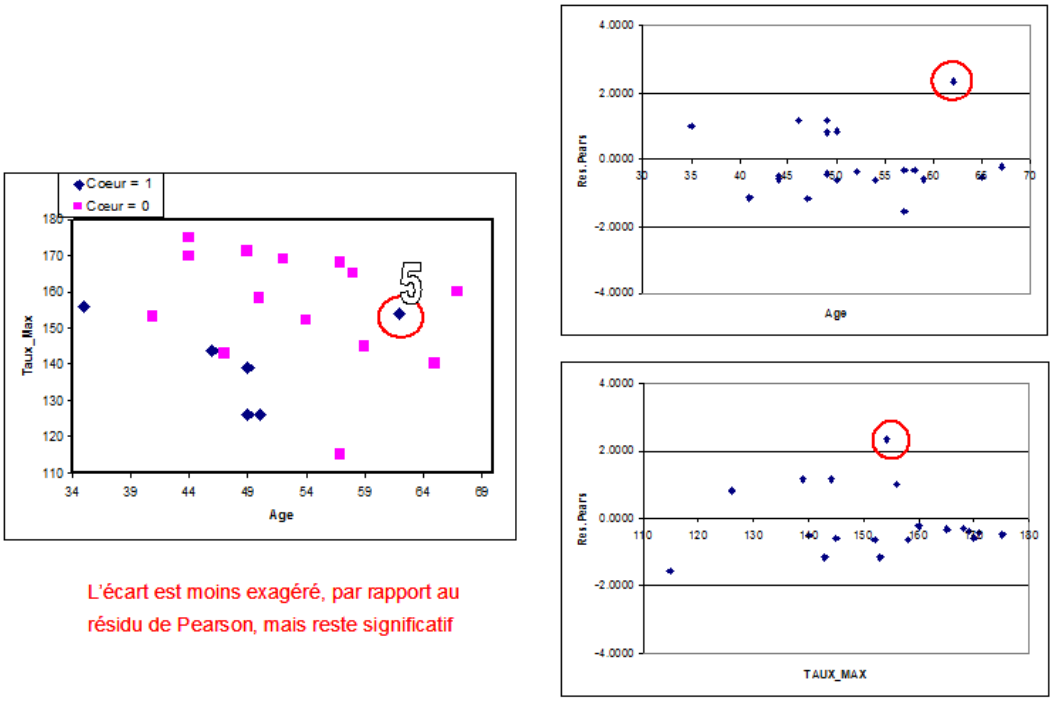


Fig. 8.5. Coeur = f(age, taux max) - Résidus déviance - Graphiques

Pour une observation ω, le levier est lu sur la diagonale principale. Il correspond à la distance du point par rapport au barycentre du nuage, pondéré par $\hat{\pi}(1-\hat{\pi})$, nous avons

$$h(\omega) = \hat{\pi}(\omega)(1-\hat{\pi}(\omega))x(\omega)(X'VX)^{-1}x'(\omega) \tag{8.7}$$

$x(\omega)$ est la description de l'individu ω c.-à-d. $x(\omega) = (1, x_1(\omega), \ldots, x_J(\omega))$.

Attention, de par sa formule, $h(\omega)$ est sur-estimé lors $\hat{\pi}(\omega) \approx 0.5$; il est sous-estimé lorsque $\hat{\pi}(\omega) \approx 0$ ou $\hat{\pi}(\omega) \approx 1$.

On montre facilement que [9] (page 169)

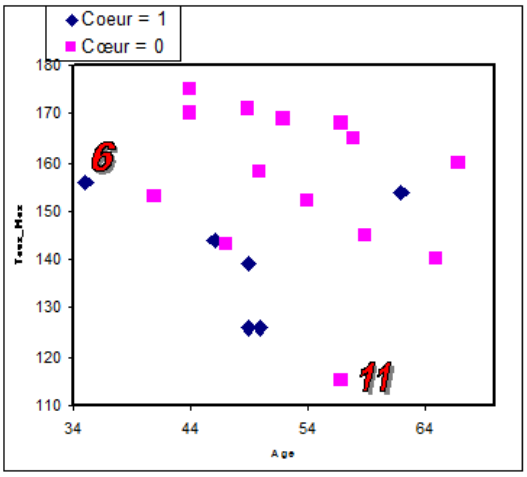
$$\bar{h} = \frac{\sum_\omega h(\omega)}{n} = \frac{J+1}{n}$$

Une règle de détection des points atypiques habituellement utilisée est

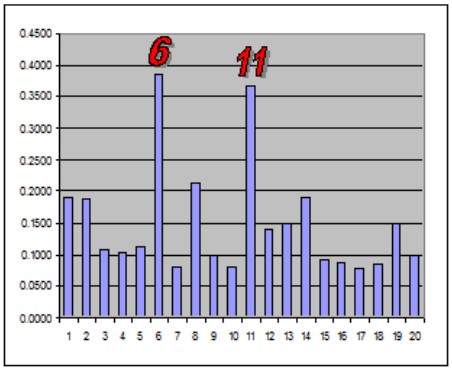
$$h(\omega) \geq 2 \times \bar{h}$$

Mais comme d'habitude, mieux vaut surtout distinguer visuellement dans un graphique les points qui prennent des valeurs inusuelles.

Dans le fichier COEUR, on notera que les points n^o6 et n^o11 sont éloignés des autres dans l'espace de représentation (Figure 8.6). Le levier les met en évidence avec des valeurs supérieures au seuil $2 \times \frac{2+1}{20} = 0.3$. Notons que le point n^o5 qui était si mal modélisé (cf. résidus de Pearson et résidu déviance) ne ressort pas particulièrement avec cet indicateur. C'est tout à fait normal. Il est noyé au milieu des autres points, il n'est en rien atypique au sens des descripteurs *age* et *taux max*.



				Seuil	0.3
Numéro	age	taux_max	coeur	PI	h
1	50	126	1	0.706	0.1914
2	49	126	1	0.730	0.1885
3	46	144	1	0.505	0.1073
4	49	139	1	0.507	0.1048
5	62	154	1	0.066	0.1121
6	35	156	1	0.610	0.3848
7	67	160	0	0.024	0.0804
8	65	140	0	0.123	0.2149
9	47	143	0	0.493	0.1000
10	58	165	0	0.048	0.0809
11	57	115	0	0.701	0.3663
12	59	145	0	0.166	0.1401
13	44	175	0	0.114	0.1489
14	41	153	0	0.487	0.1908
15	54	152	0	0.177	0.0899
16	52	169	0	0.072	0.0871
17	57	168	0	0.044	0.0768
18	50	158	0	0.182	0.0855
19	44	170	0	0.158	0.1498
20	49	171	0	0.087	0.0997



Les points n°6 et n°11 se démarquent des autres dans l'espace de représentation. Le levier traduit cela.

Fig. 8.6. Coeur = f(age, taux max) - Levier - Calcul et nuage de points

Levier, mesure d'influence

Il existe une autre lecture du levier : il mesure l'influence globale d'un point sur la prédiction des valeurs $\hat{\pi}$ des autres observations.

En régression linéaire multiple, nous avons la relation

$$\hat{y}(\omega) = \sum_{\omega'} h(\omega', \omega) \times y(\omega')$$

c.-à-d. dans la colonne $n^o\omega$ de la hat-matrix H, lorsque nous réalisons la somme du produit $h(\omega, \omega') \times y(\omega)$, nous obtenons la prédiction du modèle pour l'individu ω.

Or, on montre, et ce résultat s'applique à la régression logistique,

$$h(\omega) = h(\omega, \omega) = \sum_{\omega'} h^2(\omega', \omega)$$

Ainsi, la valeur lue sur la diagonale principale de la hat-matrix s'avère être en réalité un indicateur de l'influence globale du point ω sur la prédiction des valeurs de tout autre point ω' de l'ensemble de données.

Numéro	PI	h	PI(-17)	PI(-6)
1	0.706	0.1914	0.702	0.675
2	0.730	0.1885	0.726	0.689
3	0.505	0.1073	0.505	0.405
4	0.507	0.1048	0.508	0.450
5	0.066	0.1121	0.071	0.099
6	0.610	0.3848	0.606	
7	0.024	0.0804	0.027	0.048
8	0.123	0.2149	0.130	0.208
9	0.493	0.1000	0.494	0.408
10	0.048	0.0809	0.052	0.058
11	0.701	0.3663	0.698	0.752
12	0.166	0.1401	0.172	0.211
13	0.114	0.1489	0.120	0.068
14	0.487	0.1908	0.487	0.323
15	0.177	0.0899	0.183	0.179
16	0.072	0.0871	0.076	0.063
17	0.044	0.0768		0.050
18	0.182	0.0855	0.188	0.152
19	0.158	0.1498	0.163	0.096
20	0.087	0.0997	0.092	0.066

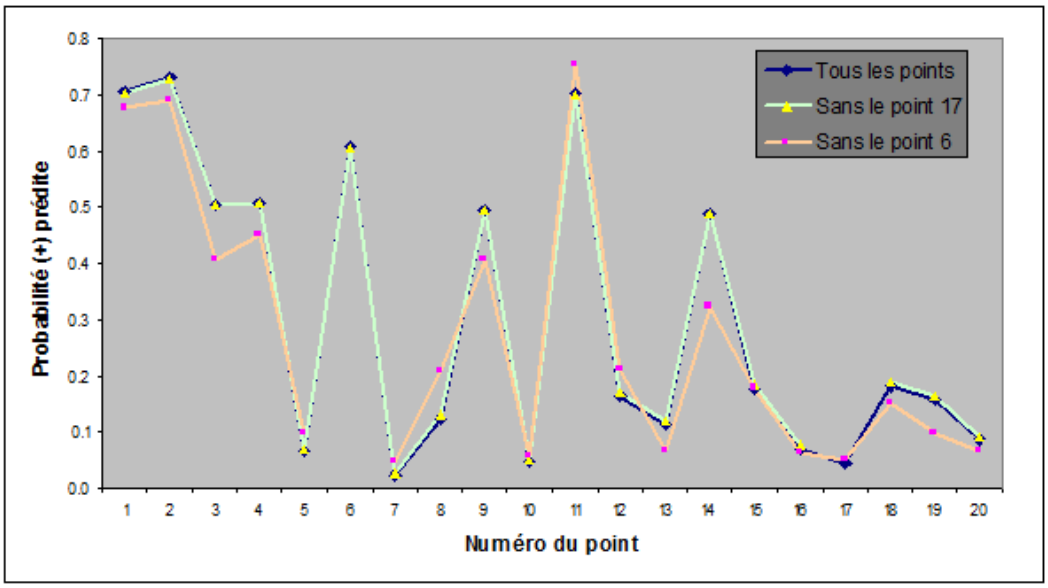


Fig. 8.7. Coeur = f(age, taux max) - Levier - Influence sur les prédictions

Vérifions ce comportement sur le fichier COEUR (Figure 8.7) :

- Dans un premier temps, nous avons réalisé la régression sur les $n = 20$ observations. Nous avons calculé les probabilités prédites $\hat{\pi}$, puis nous avons construit un graphique avec en abcisse le numéro de point, en ordonnée $\hat{\pi}$ (courbe bleue).
- Nous avons ensuite calculé le levier de chaque point. On note par exemple que le point n^o17 avec $h(17) = 0.0768$ ne pèse pas beaucoup sur la prédiction des probabilités des autres.
- Pour le vérifier, nous avons relancé la régression sur $n = 19$ points en excluant l'observation n^o17. Puis, de nouveau, nous avons calculé $\hat{\pi}$, nous avons reporté les valeurs dans notre graphique (courbe jaune). Pour le point n^o17 nous avons pris la valeur initialement fournie par la régression sur tous les points. On constate que les deux courbes (bleue et jaune) se superposent (presque) complètement. Manifestement, l'observation n^o17 n'a aucune incidence sur les prédictions.

– Tournons nous maintenant vers le point n^o6 avec un levier élevé $h(6) = 0.3848$. Nous réitérons les mêmes opérations c.-à-d. retirer le point des données, relancer la régression avec $n = 19$ observations, calculer les prédictions $\hat{\pi}$ (courbe orange). La situation est tout autre. La courbe se démarque des deux précédentes. On notera entre autres les fortes différences pour les points n^o3, n^o4, n^o8, etc. L'observation n^o6 pèse énormément dans la prédiction. Le levier met en évidence ce comportement.

8.1.4 Résidus de Pearson et Résidus déviance standardisés

Les résidus (de Pearson et déviance) indiquent la bonne ou mauvaise modélisation d'un point. A bien y regarder, on se rend compte que ces indicateurs ne sont pas très honnêtes. Considérons une observation ω de l'échantillon d'apprentissage : il a participé à la construction du modèle, par la suite on se pose la question de savoir s'il est bien modélisé ou pas. Pour peu que ω pèse énormément dans la régression, il peut être lui même très bien modélisé, tout en faussant la prédiction des autres. La solution est de corriger le résidu selon l'influence du point. Le levier justement traduit cette idée [1].

On appelle résidu de Pearson standardisé pour l'individu ω

$$r_s(\omega) = \frac{r(\omega)}{\sqrt{1 - h(\omega)}} \tag{8.8}$$

et résidu déviance standardisée

$$d_s(\omega) = \frac{d(\omega)}{\sqrt{1 - h(\omega)}} \tag{8.9}$$

Le résidu des observations à forte influence ($h \approx 1$) est exacerbé ; à l'inverse, celles qui ont une faible influence ($h \approx 0$) voient leur valeur du résidu réduite.

Sur le fichier COEUR, nous constatons que les résidus, du fait de la standardisation, sont un peu modifiés (Figure 8.8) : certes, le point n^o5 particulièrement mal modélisé, même s'il a un levier assez faible, se démarque toujours ; le point n^o11 se distingue très nettement maintenant, il est mal classé et il a un levier fort.

1. Il existe une manière plus "mathématique" de justifier les résidus standardisés. On sait que la variance de l'erreur théorique du modèle est $V(\varepsilon) = \pi(1 - \pi)$, c'est en ce sens que l'on a défini de résidu de Pearson. En revanche, la variance du résidu, l'erreur observée sur les données, s'écrit

$$V(\hat{\varepsilon}) = \pi(1 - \pi)(1 - h)$$

D'où la nouvelle correction aboutissant au résidu de Pearson standardisé. Le mécanisme est identique en régression linéaire multiple. Voir R. Rakotomalala, *Pratique de la Régression Linéaire Multiple - Diagnostic et Sélection de Variables*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf`, pages 33 à 36.

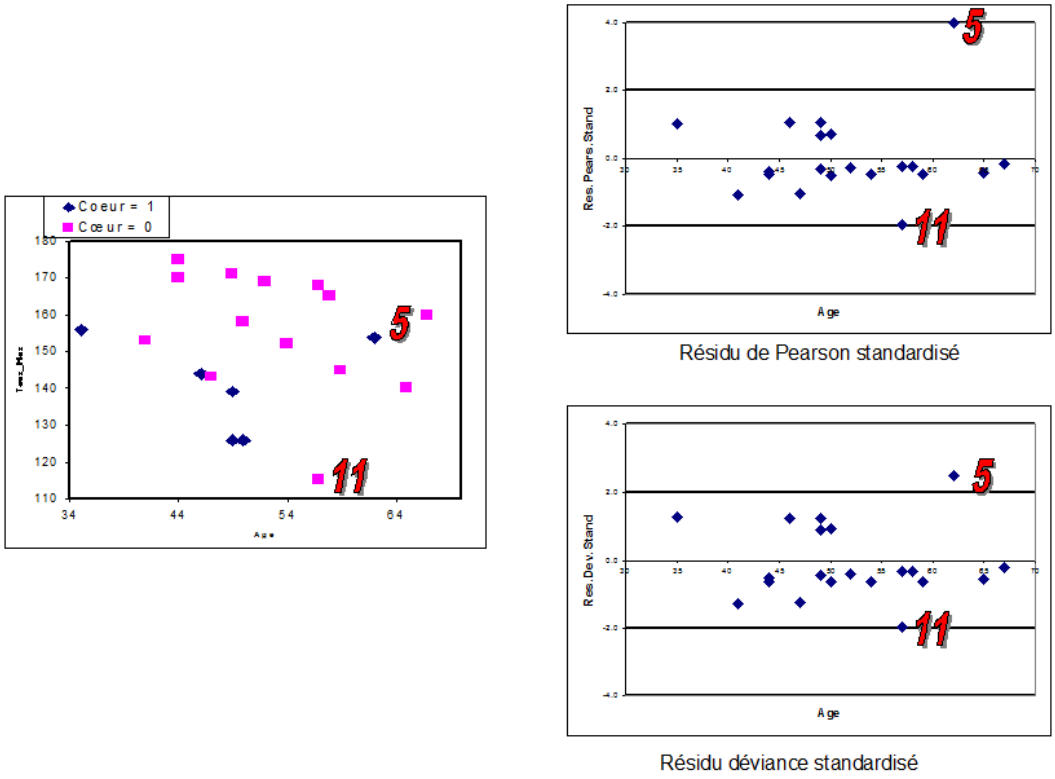


Fig. 8.8. Coeur = f(age, taux max) - Résidus de Pearson et Résidus déviance standardisés

8.1.5 Distance de Cook

La distance de Cook permet de quantifier l'écart entre les vecteurs de paramètres estimées en présence et en l'absence du point ω. On peut le voir sous l'angle d'un test d'hypothèses [2]

$$H_0 : \text{les coefficients sont identiques}$$
$$H_1 : \text{un des coefficients au moins est différent}$$

Bien entendu, il est hors de question de réaliser les n régressions en omettant tour à tour chaque observation. De nouveau le levier nous sera très précieux.

La distance de Cook peut être écrit à partir du résidu déviance standardisé ou du résidu de Pearson standardisé. Si nous prenons la seconde définition, nous aurons

$$C(\omega) = \frac{r_s^2(\omega)}{J+1} \times \frac{h(\omega)}{1-h(\omega)} \tag{8.10}$$

2. Pour une discussion plus approfondie sur les différentes manières de voir la Distance de Cook et sur les règles de détection des points influents associées, voir R. Rakotomalala, *Pratique de la régression linéaire multiple - Diagnostic et sélection de variables*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf` ; pages 41 à 43.

Une lecture rapide de la formule nous indique que la conjonction d'un résidu et un levier élevés produit une distance de Cook élevée.

La règle de détection usuelle est

$$C(\omega) \geq \frac{4}{n - J - 1}$$

Deux remarques essentiellement :

- La distance de Cook peut être définie à partir du résidu déviance.
- Certains logiciels (SPSS) ne normalisent pas par le nombre de paramètres $(J + 1)$. La règle de détection devient $C(\omega) \geq 1$ (voir page 169).

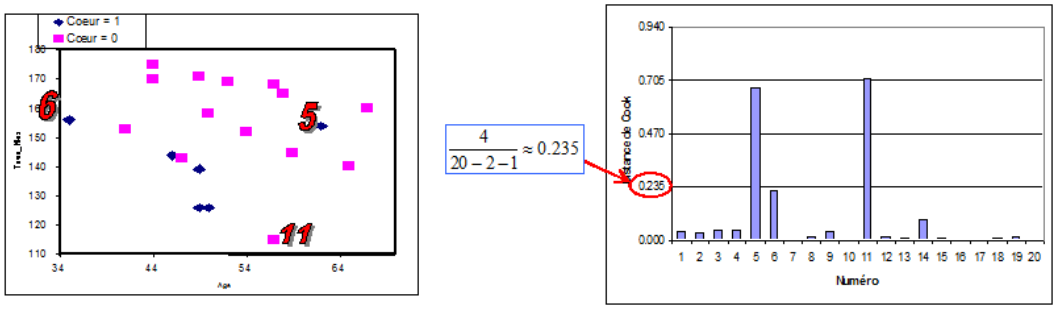


Fig. 8.9. Coeur = f(age, taux max) - Distance de Cook - Valeur seuil : 0.235

Sur le fichier COEUR (Figure 8.9), plusieurs points attirent notre attention :

- Le point n^o11 révèle sa vraie nature. Avec ou sans lui, les paramètres estimés de la régression sont très différents. Il s'agit là d'un point réellement influent. Il est mal modélisé (mal classé) et on notera via le levier qu'il est un peu éloigné des autres.
- Le point n^o5 pèse fortement aussi parce qu'il est mal modélisé, avec un résidu standardisé très élevé.
- Le point n^o6 pèse essentiellement parce qu'il est éloigné des autres (levier élevé). Il est bien modélisé (classé) par ailleurs, le résidu reste raisonnable.

8.1.6 DFBETAS

Les DFBETAS sont complémentaires à la distance de Cook. Ils permettent d'identifier le coefficient sur lequel pèse la présence/absence du point ω. Nous pouvons les voir sous l'angle d'un test de comparaison de coefficients. Ils nous donnent des éléments de réponse à la question : de quelle manière le point ω est atypique ?

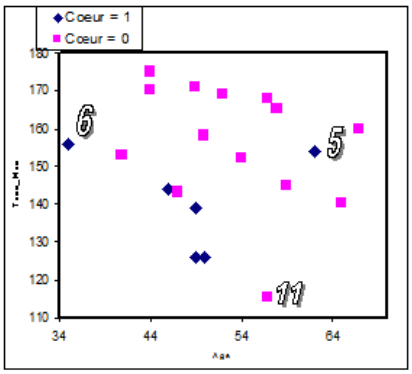
Le DFBETAS du coefficient a_j est calculé comme suit

$$DFBETAS_j(\omega) = \frac{(X'VX)^{-1}x'(\omega)}{\sqrt{(X'VX)^{-1}_j}} \times \frac{y(\omega) - \hat{\pi}(\omega)}{1 - h(\omega)} \tag{8.11}$$

Avec $\sqrt{(X'VX)^{-1}_j} = \hat{\sigma}_{\hat{a}_j}$ est l'écart-type estimé de $\hat{a}_j$.

La règle de détection usuelle est

$$|DFBETAS_j| \geq \frac{2}{\sqrt{n}}$$



Numéro	age	taux_max	dfbetas(const)	dfbetas(age)	dfbetas(taux)
1	50	126	0.242	-0.060	-0.276
2	49	128	0.242	-0.083	-0.261
3	46	144	0.172	-0.160	-0.093
4	49	139	0.160	-0.052	-0.155
5	62	154	-1.072	1.241	0.670
6	35	156	0.373	-0.680	0.007
7	67	160	0.040	-0.044	-0.028
8	65	140	0.115	-0.198	-0.027
9	47	143	-0.154	0.118	0.098
10	58	165	0.060	-0.051	-0.050
11	57	115	-0.666	-0.324	1.111
12	59	145	0.102	-0.158	-0.040
13	44	175	0.082	0.012	-0.132
14	41	153	-0.191	0.362	-0.031
15	54	152	0.089	-0.094	-0.068
16	52	169	0.071	-0.041	-0.075
17	57	168	0.056	-0.044	-0.050
18	50	158	0.081	-0.043	-0.096
19	44	170	0.082	0.027	-0.146
20	49	171	0.076	-0.028	-0.092
		seuil.bas	*-0.447*	*-0.447*	*-0.447*
		seuil.haut	*0.447*	*0.447*	*0.447*

Fig. 8.10. Coeur = f(age, taux max) - DFBETAS - Valeur seuil : ±0.447

Les résultats sur le fichier COEUR nous éclaire sur le rôle des points incriminés jusqu'à présent (Figure 8.10) :

- Le point n^o6 pèse surtout sur le coefficient associé à AGE. Ce n'est guère étonnant vu le positionnement de ce point dans l'espace de représentation. Il a 35 ans lorsque la moyenne d'âge (sans lui) est de 52.63.
- Le point n^o11 lui se distingue par sa faible valeur de TAUX MAX. Là également, l'individu porte une valeur qui semble plutôt faible (115) par rapport à la moyenne du reste de l'échantillon (153.37). Il pèse donc sur le coefficient de TAUX MAX c.-à-d. il modifie la pente de la droite séparatrice des positifs et négatifs, mais aussi sur la constante, il décale la frontière (voir section 11.3 pour apprécier pleinement ce commentaire).
- Enfin, le point n^o5 est un vrai problème Que fait ce positif au milieu de tous ces négatifs ? Si on le retire de l'échantillon d'apprentissage, rien ne serait pareil.

Il existe une version non standardisée de cet indicateur : les DFBETA. Elles se justifient surtout lorsque les variables sont mesurées sur une même échelle, ou lorsqu'elles sont exclusivement composées d'indicatrices (voir les "covariate pattern", section 9.4.3). Lorsque les explicatives sont quantitatives et définies sur des unités différentes, passer à une mesure standardisée (DFBETAS, divisée par l'écart-type du coefficient $\hat{\sigma}_{\hat{a}_j}$) nous autorise à comparer les valeurs d'une variable à l'autre.

8.2 Non-linéarité sur le LOGIT

On dit que le LOGIT est linéaire par rapport à une variable X si la variation de X d'une unité modifie la valeur du LOGIT de la même manière quelle que soit la valeur de X. Nous avions déjà entre-aperçu

cette idée lors de l'interprétation du coefficient associée à une variable explicative continue (section 5.2.2). La variation du LOGIT, dans le cas où la relation est linéaire, est égale au coefficient de X.

Comme nous avions pu le dire déjà, cette contrainte est assez forte. En effet, comment peut-on imaginer qu'une variation de 10 ans ait le même impact sur une éventuelle maladie cardiaque que l'on ait 20 ans ou 40 ans. Il nous faut donc, d'une part, vérifier que la variation du LOGIT ne dépend pas de la valeur de X, et si l'hypothèse de linéarité ne tient pas la route, proposer des méthodes pour prendre en compte la non-linéarité dans le modèle final.

Remarque : Attention, il **ne** faut **pas confondre non-linéarité et non-additivité**. Dans le premier cas, l'impact de la variation d'une variable dépend de sa valeur ; dans le second cas, l'impact de la variation d'un variable dépend de la valeur d'une autre variable explicative. Cela peut arriver lorsque nous manipulons des modèles avec interaction. En vérité, détecter ce type de problème est très difficile en l'absence de connaissances du domaine qui nous aiguillerait sur les configurations à tester. Il n'y a pas vraiment de solutions simples en la matière [10] (page 75).

8.2.1 Identification graphique univariée

Construction du graphique

Une procédure graphique simple permet de vérifier la linéarité du LOGIT par rapport à une variable X (voir [9], page 107) :

1. Découper X en déciles (ou autres) ;
2. Dans chaque intervalle, calculer la proportion de positifs π ;
3. Le graphique nuage de points est constitué de
   - En abcisse, la moyenne de X des intervalles ;
   - En ordonnée, le LOGIT observé c.-à-d.$\ln \frac{\pi}{1-\pi}$
4. Si le LOGIT est linéaire par rapport à X, le nuage de points forme une droite.
5. Le seconde caractéristique à vérifier est l'évolution monotone ou non du LOGIT par rapport à X.

Cette procédure peut poser problème lorsque tous les individus sont positifs (resp. négatifs) dans un intervalle. Il est conseillé dans cas de mettre arbitrairement $\pi = 0.99$ (resp. $\pi = 0.01$) ([10], page 70). L'énorme avantage de cette méthode est qu'elle nous renseigne non seulement sur le caractère linéaire ou non du LOGIT, mais aussi sur la forme de la relation dans le cas où elle ne serait pas linéaire.

Un exemple numérique : prédiction du diabète

Nous utilisons le fichier PIMA dans cette section. Nous souhaitons prédire l'occurrence du diabète (Y = DIABETE) à partir de l'indice de masse corporelle (X = BODYMASS) chez des amérindiens. Le fichier comporte $n = 757$ observations.

Nous avons réalisé la régression entre DIABETE et BODYMASS. Cette dernière s'est avérée significative (p-value < 0.0001), avec un odds-ratio $OR = e^{0.1025} = 1.1079$. Si l'hypothèse de linéarité est licite,

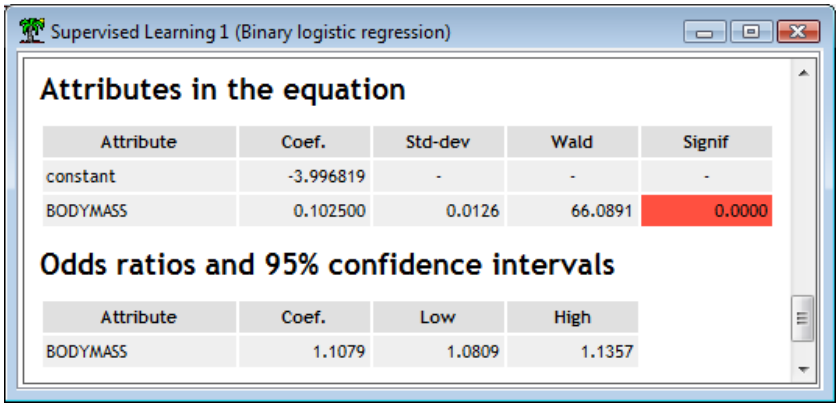


Fig. 8.11. Régresion logistique simple - DIABETE = f (BODYMASS)

N°	Bornes	n(cumul)	n	n+(cumul)	n+	pi	odds	log-odds
1	24.00	78	78	5	5	0.06	0.07	-2.68
2	26.20	155	77	12	7	0.09	0.10	-2.30
3	28.40	229	74	33	21	0.28	0.40	-0.93
4	30.34	303	74	55	22	0.30	0.42	-0.86
5	32.30	380	77	90	35	0.45	0.83	-0.18
6	33.80	457	77	124	34	0.44	0.79	-0.23
7	35.50	533	76	157	33	0.43	0.77	-0.26
8	37.88	605	72	188	31	0.43	0.76	-0.28
9	41.62	681	76	218	30	0.39	0.65	-0.43
10	67.10	757	76	266	48	0.63	1.71	0.54

Fig. 8.12. Évolution du log-odds (LOGIT) en fonction de X - Tableau de calcul

nous lisons le coefficient de la manière suivante : lorsque le BODYMASS augmente d'une unité, l'individu a 1.1079 fois plus de chances d'avoir du diabète, ceci quel que soit son poids (Figure 8.11).

Voyons maintenant si l'hypothèse de linéarité est susceptible d'être remise en cause en construisant notre graphique d'identification. Nous avons élaboré notre tableau de calcul de la manière suivante (Figure 8.12) :

1. La première colonne n^o sert uniquement à numéroter les intervalles.
2. La seconde correspond aux déciles.
3. Nous avons les effectifs cumulés.
4. Par différenciation nous avons les effectifs dans chaque intervalle. Ils ne sont pas égaux parce que n n'est pas divisible par 10, et il y a parfois des ex-aequo.
5. Nous comptabilisons également les effectifs cumulés des positifs.
6. Nous obtenons par différenciation le nombre de positifs dans chaque intervalle.
7. Nous en déduisons la proportion de positifs π.
8. L'odds $\frac{\pi}{1-\pi}$.
9. Et le log-odds ou le LOGIT $\ln \frac{\pi}{1-\pi}$.

Il ne nous reste plus qu'à construire le graphique en prenant en abcisse la moyenne de X dans chaque intervalle, et en ordonnée le LOGIT (Figure 8.13). Plusieurs commentaires nous viennent immédiatement :

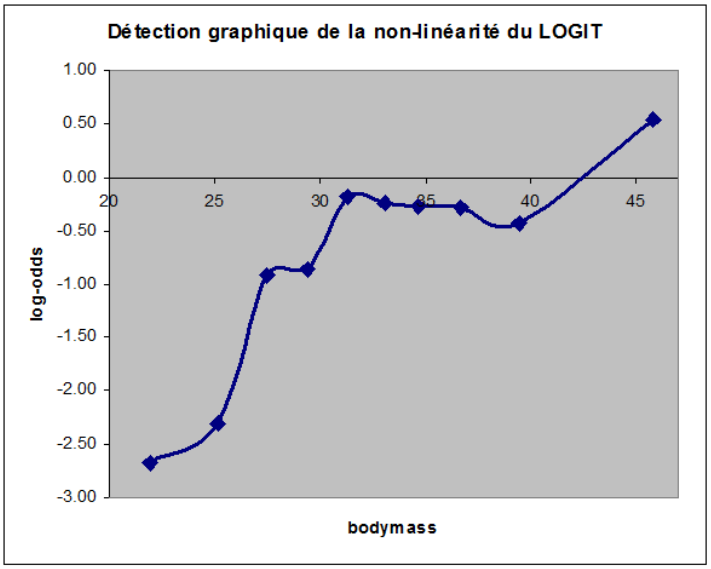


Fig. 8.13. Évolution du log-odds (LOGIT) en fonction de X - Graphique

- Manifestement, la relation n'est pas linéaire.
- Elle est néanmoins monotone et évolue par paliers.
- Nous pouvons visuellement détecter les seuils où l'évolution s'accélère ou ralentit. Nous avons mis en couleur les différentes zones dans le tableau de calcul (Figure 8.12). Cette information est importante car nous pourrons nous en servir pour recoder correctement la variable X dans la régression logistique.

8.2.2 Une solution simple : la discrétisation de la variable X

La transformation de variable est la stratégie privilégiée pour remédier au problème de non linéarité. Nous pouvons essayer différentes fonctions mathématiques usuelles (log, carré, racine carrée, etc.) ou adopter une démarche générique avec les polynômes fractionnaires (en anglais, *fractional polynomials* ; [9], pages 100 et 101). L'efficacité de ces méthodes n'est pas mise en doute, mais elles sont assez fastidieuses à mettre en oeuvre. D'autant plus qu'il faudra par la suite interpréter le coefficient associé à la variable transformée.

Une solution simple est la discrétisation c.-à-d. le découpage en intervalles (ou le regroupement en classes) de la variable explicative. A partir de X, nous dérivons une série d'indicatrices $D_1, D_2, \ldots$ destinées à matérialiser chaque intervalle. Nous devons répondre à une série de questions pour produire un codage efficace :

1. Combien d'intervalles devons-nous produire ? La question est d'importance, il s'agit de ne pas les multiplier inutilement. Il importe surtout que dans chaque groupe, le comportement de la variable dépendante Y, ou plus précisément du LOGIT, soit cohérent. Dans notre exemple (Figure 8.13), nous détectons visuellement 4 paliers. On peut envisager un découpage en 4 classes.

2. Seconde question corollaire à la première, comment définir les bornes de découpage ? La réponse est liée à la précédente. Dans notre exemple, il s'agit de matérialiser chaque palier. Sur la base de notre tableau de calcul (Figure 8.12), nous choisirions $b_1 = 26.2$, $b_2 = 30.34$ et $b_3 = 41.62$.
3. Dernier point important, quel type de codage des indicatrices adopter ? Si la relation est monotone, nous avons tout intérêt à adopter un codage 0/1 emboîté pour relater le caractère monotone de l'évolution du LOGIT. Les coefficients de la régression traduisent alors le surcroît de risque en passant d'un niveau (un intervalle) à celui qui lui succède. Dans le cas contraire, la relation est non monotone, cette contrainte fausse les calculs, nous devons adopter un codage disjonctif simple. La lecture devient moins aisée cependant. Il faut avoir une idée précise sur la modalité (l'intervalle) de référence pour que l'interprétation des coefficients tienne la route.

Discrétisation de la variable BODYMASS

Suite à ces différentes considérations, nous décidons de produire 3 indicatrices emboîtés à partir de la variable BODYMASS, codées de la manière suivante :

1. $D_1 = 1$, si $BODYMASS > 26.20$; 0 sinon
2. $D_2 = 1$, si $BODYMASS > 30.34$; 0 sinon. On remarquera que $D_2 = 1 \Rightarrow D_1 = 1$, ce qui caractérise l'emboîtement.
3. $D_3 = 1$, si $BODYMASS > 41.62$; 0 sinon. De même, nous constatons que $D_3 = 1 \Rightarrow D_2 = D_1 = 1$.

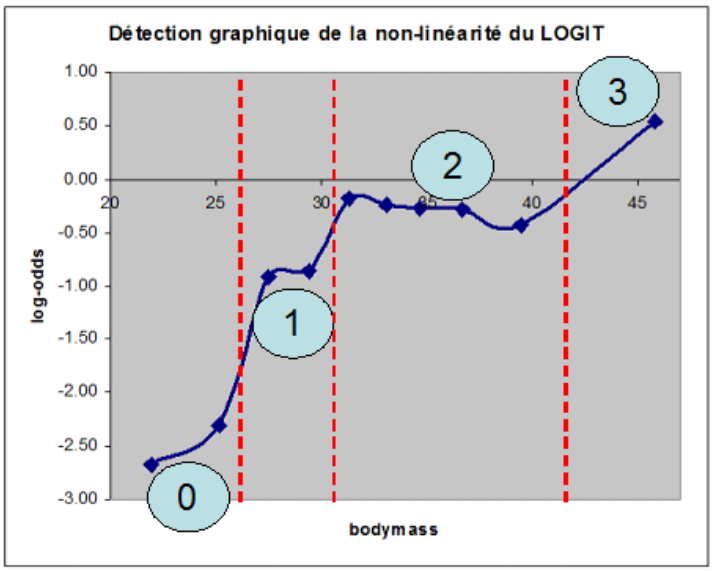


Fig. 8.14. Évolution du log-odds (LOGIT) en fonction de X - Indicatrices et bornes de discrétisation

Nous matérialisons dans le graphique mettant en relation le LOGIT et BODYMASS ces indicatrices (numéro) et les bornes de discrétisation (en rouge pointillés) (Figure 8.14). Il ne nous reste plus qu'à relancer la régression logistique avec ces nouvelles variables (Figure 8.15) :

- Manifestement, un changement de palier de BODYMASS induit un surcroît de risque de diabète significatif.
- Le premier palier est le plus important. On s'en serait douté à la vue du graphique des LOGIT en fonction de BODYMASS. Apparemment, les individus ont $OR(1/0) = e^{1.585178} = 4.8802$ fois plus chances d'avoir le diabète lorsque nous passons du palier n^o0 au n^o1.
- Les autres changements de paliers sont moins spectaculaires. Ils n'en sont pas moins significatifs : $OR(2/1) = 1.8513$ et $OR(3/2) = 2.2612$.

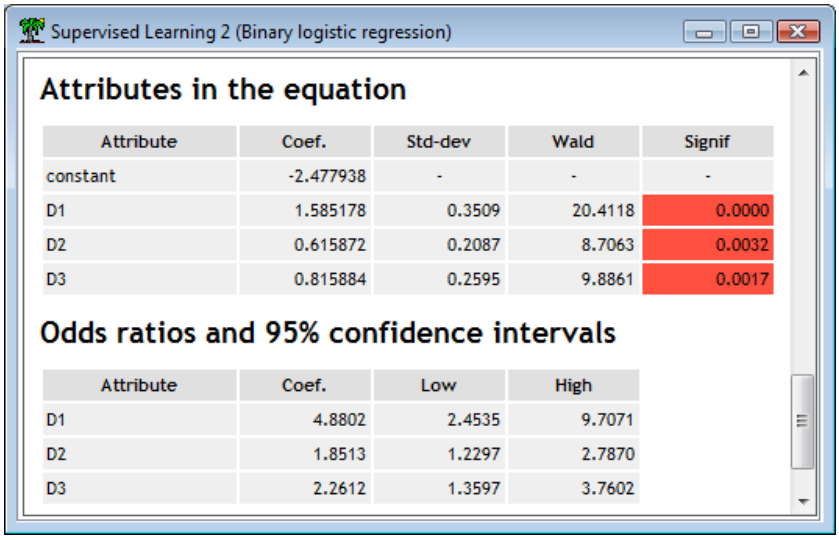


Fig. 8.15. Régresion logistique - $DIABETE = f(D_1, D_2, D_3)$

Remarque : la discrétisation n'est pas la panacée

Certains auteurs préconisent l'usage systématique de la discrétisation dès lors que nous sommes en présence de variables explicatives quantitatives. Ce n'est pas aussi automatique (comme les antibiotiques). Certes, en introduisant des indicatrices, nous diminuons le biais du modèle. Il est plus à même de prendre en compte des relations complexes existantes entre Y et les X_j. Mais dans le même temps, nous en augmentons la variance, la dépendance au fichier de données. On risque le fameux sur-apprentissage (en anglais *overfitting*) avec un degré de liberté qui baisse dangereusement. A force de multiplier les indicatrices, nous aboutirons à un modèle qui marche très bien effectivement sur le fichier de données, mais qui s'effondre totalement dès que nous le déployons dans la population. La discrétisation n'est donc certainement pas la panacée. C'est un outil qu'il faut savoir utiliser avec discernement, comme tous les outils. Dans certaines situations, il est plus judicieux de passer par des transformations de X à l'aide de fonctions mathématiques pour répondre à la non-linéarité.

8.2.3 Détection numérique multivariée : le test de Box-Tidwell

Dans la régression multiple, l'analyse graphique devient plus compliquée (section 8.2.4), et surtout très fastidieuse dès que le nombre de variables explicatives augmente. Il nous faut une procédure numérique

pour détecter automatiquement les situations de non-linéarité, quitte à revenir par la suite sur le graphique pour étudier de manière approfondie la forme de la relation.

Le principe du test de Box-Tidwell est le suivant :

1. Pour une variable X que l'on souhaite évaluer ;
2. Nous créons la variable transformée $Z = X \times \ln X$;
3. Que nous rajoutons parmi les explicatives. Nous conservons toutes les autres variables, y compris X ;
4. Si le coefficient de Z est significatif, cela indique que la variable X intervient de manière non linéaire sur le LOGIT ;
5. Il reste alors à identifier la forme de la relation, l'outil graphique reste le moyen privilégié dans ce cas.

Avec les logiciels proposant un langage de programmation (le logiciel R par exemple), implémenter cette procédure est très facile. Nous pouvons tester un grand nombre de variables. On note néanmoins une faible puissance du test. Il détecte mal les faibles écarts à la linéarité ([10], page 70). De plus, il ne nous donne aucune indication sur la forme de la relation.

Détection de la non-linéarité par rapport à BODYMASS

La régression avec les variables explicatives $BODYMASS$ et $Z = BODYMASS \times \ln(BODYMASS)$ nous indique que cette dernière est très significative (p-value = 0.0009) (Figure 8.16). Cela confirme, si besoin était, la non-linéarité du LOGIT par rapport à BODYMASS.

Attribute	Coef.	Std-dev	Wald	Signif
constant	-13.756516	-	-	-
BODYMASS	1.403581	0.3933	12.7355	0.0004
Z	-0.286109	0.0861	11.0508	0.0009

Fig. 8.16. Test de Box-Tidwell pour la non-linéarité du LOGIT par rapport à $BODYMASS$

8.2.4 Détection graphique multivariée : les résidus partiels

Admettons que le test de Box-Tidwell nous indique que le LOGIT n'est pas linéaire par rapport à une variable explicative en particulier. Il faut que l'on identifie la forme appropriée de la transformation avant de pouvoir ajouter la variable modifiée dans la régression. Pour cela, rien ne vaut les procédures graphiques. Encore faut-il utiliser la bonne.

En effet, la détection univariée décrite précédemment n'est plus valable (section 8.2.1). Il faut que l'on tienne compte du rôle des autres variables. Nous utiliserons les résidus partiels. Dans un premier temps, nous les présentons dans le cadre de la régression linéaire pour en comprendre le principe.

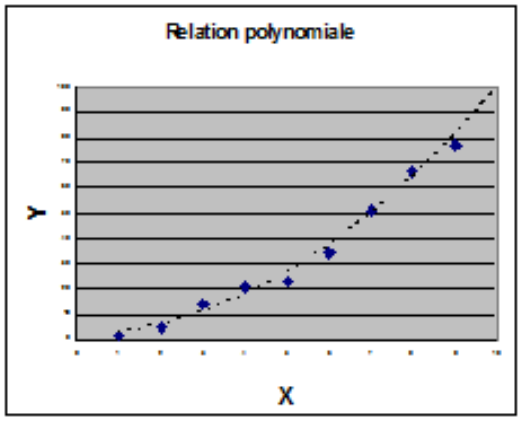


Fig. 8.17. Relation polynomiale entre Y et X

Les résidus partiels en régression linéaire

Dans la régression linéaire simple $Y = a_0 + a_1 X$, le graphique "nuage de points" mettant en relation X et Y permet d'identifier la forme de la relation entre ces variables. Dans la figure 8.17, nous avons généré des données fictives pour réaliser la régression linéaire. Il semble y avoir une relation polynomiale (ben voyons). Nous pouvons créer une variable transformée $Z = X^2$, puis l'adjoindre à la régression c.-à-d. former $Y = b_0 + b_1 X + b_2 Z$. Nous pouvons également la substituer simplement à X.

Dans la régression linéaire multiple à J variables, le graphique individuel (X_j, Y) n'est plus valable parce qu'il ne tient pas compte des autres explicatives, certaines notamment sont plus ou moins liées avec X_j. Dans ce contexte, on utilise les "résidus partiels"

$$\hat{\varepsilon}_j = (y - \hat{y}) + \hat{a}_j x_j \tag{8.12}$$

où $\hat{a}_j$ est le coefficient estimé relatif à la variable X_j dans la régression incluant toutes les variables.

Si la relation est linéaire, le nuage $(X_j, \hat{\varepsilon})$ ne doit pas présenter de forme particulière. Ou si on utilise une forme de lissage des points, la courbe lissée doit former une droite [3].

Concernant notre exemple fictif, on se rend compte dans le graphique des résidus partiels que X entretient bien une relation de type X^2 avec la variable dépendante (Figure 8.18). Nous passons donc à la régression $Y = b_0 + b_1 X + b_2 Z$ et nous souhaitons savoir si cette transformation est suffisante. Nous estimons les paramètres à l'aide des données. Nous formons ensuite les résidus partiels tels que nous les avons définis ci-dessus [4].

3. Nous reviendrons sur cet aspect lorsque nous présenterons les résidus partiels dans le cadre de la régression logistique.

4. Une autre possibilité serait d'utiliser les résidus partiels "augmentés" pour lesquels nous introduisons tous les coefficients et formes de la variable

$$\varepsilon = (y - \hat{y}) + \hat{b}_1 X + \hat{b}_2 Z$$

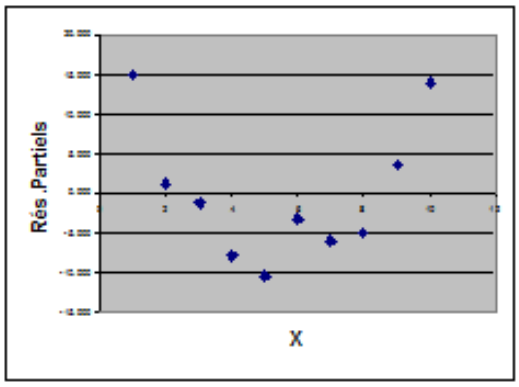


Fig. 8.18. Résidus partiels $\varepsilon = (y - \hat{y}) + \hat{a}_1 X$

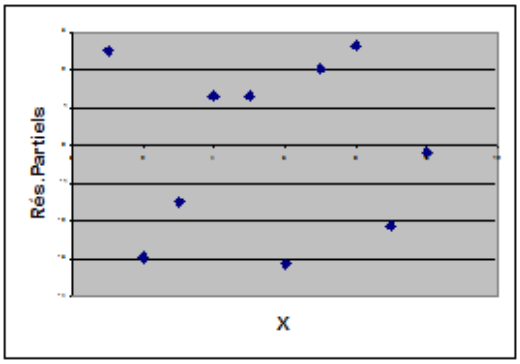


Fig. 8.19. Résidus partiels $\varepsilon = (y - \hat{y}) + \hat{b}_1 X + \hat{b}_2 X^2$

Nous créons le graphique nuage de points $(X, \hat{\varepsilon})$. Si les bonnes transformations ont été introduites, le graphique ne doit plus présenter de "formes" particulières (ou le graphique lissé doit avoir la forme d'une droite). C'est le cas pour notre exemple (Figure 8.19).

Transposition des résidus partiels à la régression logistique

L'idée des résidus partiels a été transposée à la régression logistique. Pour la variable X_j, ils sont calculés de la manière suivante

$$r_j = \frac{y - \hat{\pi}}{\hat{\pi}(1 - \hat{\pi})} + \hat{a}_j x_j \tag{8.13}$$

Nous élaborons la forme lissée du nuage de points (x_j, r_j). Si elle forme une droite, on peut conclure à la linéarité du LOGIT par rapport à la variable X_j. Sinon, en nous inspirant de la forme de la courbe, nous introduisons la variable transformée dans la régression, puis nous calculons de nouveau les résidus partiels.

Deux éléments important doivent attirer notre attention :

1. Nous utilisons la courbe lissée et non pas le nuage de points brut pour évaluer la forme de la relation. En effet, la disposition des observations est trop erratique dans le repère. Nous voulons avant tout dégager une tendance. Dans notre support, nous utilisons une procédure de lissage très fruste qui consiste à découper X_j en L intervalles pour lesquelles nous calculons les moyennes $\bar{x}_{j,l}$; puis les moyennes des résidus $\bar{r}_{j,l}$; pour tracer enfin une suite de segments reliant les L points $(\bar{x}_{j,l}; \bar{r}_{j,l})$.
   Dans les logiciels tels que R (package Design), le graphique est réalisé via un lissage de type loess (*locally weighted regression*). La procédure consiste à définir une série de points équidistants sur l'axe des abcisses ; de calculer une régression pondérée dans le voisinage de ces points ; puis d'utiliser les équations de régression pour calculer la coordonnée en ordonnée. Il ne reste plus qu'à relier les points par des segments [5]. Il faut avouer que le graphique a nettement plus d'allure avec cette procédure.
2. Certains logiciels (R avec le package Design pour ne pas le nommer encore) utilisent un autre formulation des résidus partiels

$$r_j = \frac{y - \hat{\pi}}{\hat{\pi}(1-\hat{\pi})} + \hat{a}_0 + \hat{a}_j x_j$$

   Cela induit un simple décalage sur l'axe des ordonnées. Il n'y a aucune incidence sur les conclusions que l'on pourrait tirer du graphique des résidus partiels.

Un exemple d'application

Nous reprenons le fichier PIMA, nous utilisons 3 variables explicatives maintenant : BODYMASS, PLASMA et AGE.

La régression sour Tanagra nous indique que les 3 explicatives sont toutes très significatives (Figure 8.20). L'AIC (critère Akaike) du modèle est $AIC = 732.958$. On pourrait s'en satisfaire et s'en tenir là. Essayons quand même de voir comment sont disposés les résidus de la régression partiellement à la variable AGE. Nous détaillons la démarche dans une feuille Excel (Figure 8.21) :

- Nous avons reportés les coefficients estimés de la régression dans le feuille Excel. Nous en tirons le LOGIT prédit

$$\hat{c} = -9.03238 + 0.089753 \times BODYMASS + 0.035548 \times PLASMA + 0.028699 \times AGE$$

  et la probabilité prédire

$$\hat{\pi} = \frac{1}{1+e^{-\hat{c}}}$$

- A partir de ces informations, nous formons les résidus partiels (nous utilisons la constante comme dans R pour rendre les résultats comparable)

$$r_{age} = \frac{y - \hat{\pi}}{\hat{\pi}(1-\hat{\pi})} - 9.03238 + 0.028699 \times AGE$$

  Ainsi, pour le 1^{er} individu, nous avons

$$r_{age} = \frac{1 - 0.0612}{0.0612(1-0.0612)} - 9.03238 + 0.028699 \times 26 = 8.0638$$

5. Pour une description approfondie de *loess*, voir W.G. Jacoby, *Statistical Graphics for Univariate and Bivariate Data*, Quantitative Applications in the Social Sciences n^o117, Sage Publications, 1997 ; pages 64 à 83.

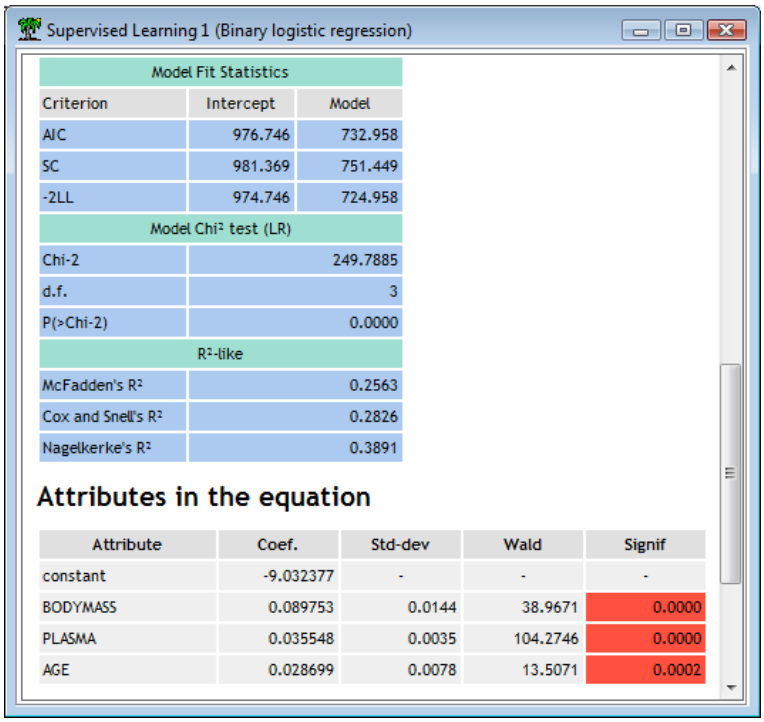


Fig. 8.20. Régression - Fichier PIMA - DIABETE = f(BODYMASS, PLASMA, AGE)

- En utilisant les *tableaux croisés dynamiques* d'Excel, nous découpons l'âge en 6 intervalles ($21-30$, $31-40$, etc.) et, dans chaque bloc, nous calculons la moyenne de l'âge et celle des résidus partiels r_{age}.
- Il ne nous reste plus qu'à former le graphique (Figure 8.21). On notera que la relation n'est absolument pas linéaire mais quadratique en AGE. Il serait tout à fait judicieux de rajouter la variable synthétique $AGE2 = AGE^2$ parmi les explicatives.
- Nous avons calculé la nouvelle régression (Figure 8.22). Nous notons que la variable AGE2 est très significative dans la régression [6] et, surtout, nous constatons que le modèle ainsi élaboré est nettement meilleur que le précédent. Le critère Akaike est passé de $AIC = 732.958$ à $AIC = 701.998$ (idem pour le critère BIC qui baisse fortement en passant de 751.449 à 725.111).
- Il fallait bien cette transformation. Lorsque nous recalculons les résidus partiels par rapport à AGE dans le nouveau modèle. Nous constatons maintenant que les point sont (sagement) alignés sur une droite (Figure 8.23). L'adjonction de AGE^2 nous a permis de mieux prendre en compte la contribution de l'âge dans l'explication de la variable dépendante.

Les résidus partiels dans le logiciel R

Les calculs étant assez complexes et les références rares, nous avons voulu croiser nos résultats avec ceux du logiciel R (package Design). Ce dernier présente un avantage certain, il utilise un lissage LOESS [7]

6. On notera que la contribution de AGE a été modifiée aussi, sa significativité est plus forte.
7. Avec un peu de recul, on se rend compte que la procédure que nous utilisons sous Excel est une version très fruste de LOESS, sauf que : nous ne pondérons pas les points dans le voisinage ; nous utilisons un polynôme

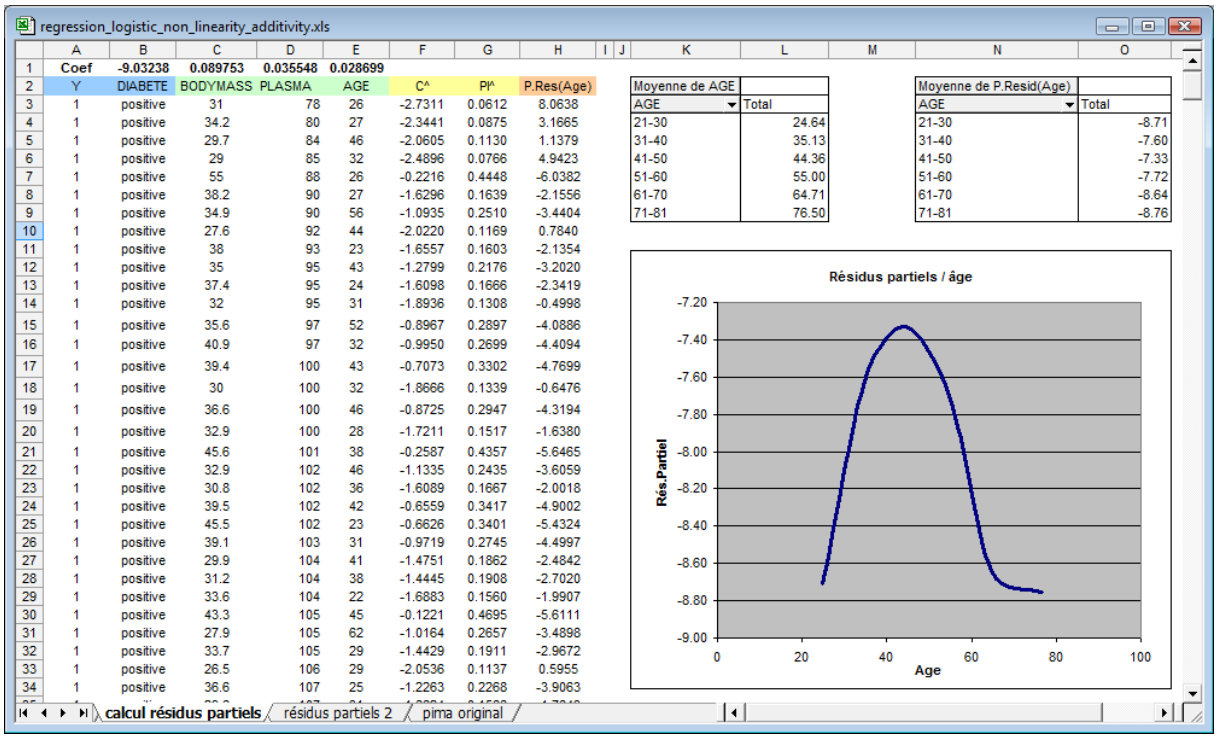


Fig. 8.21. DIABETE = f(BODYMASS, PLASMA, AGE) - Résidus partiels par rapport à AGE

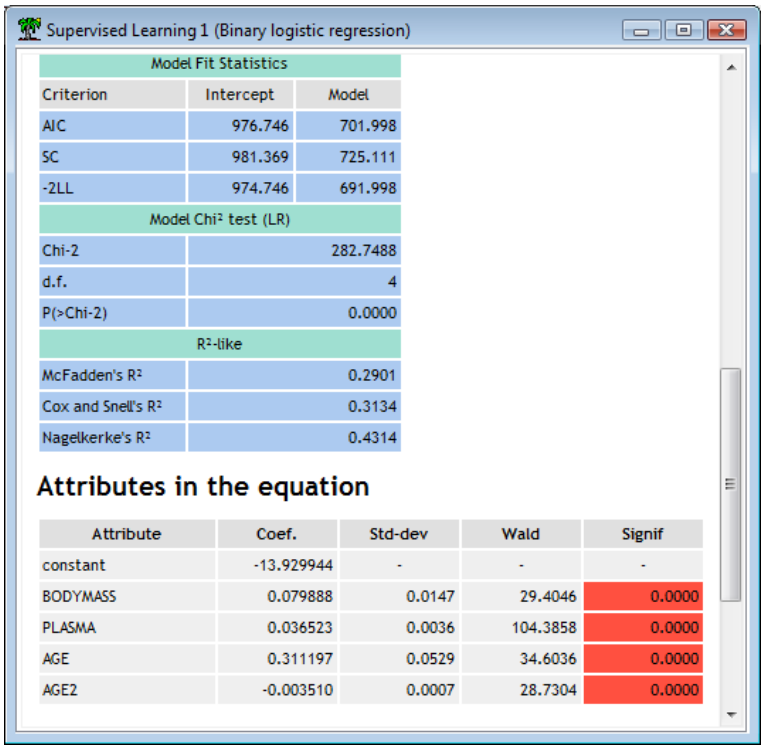
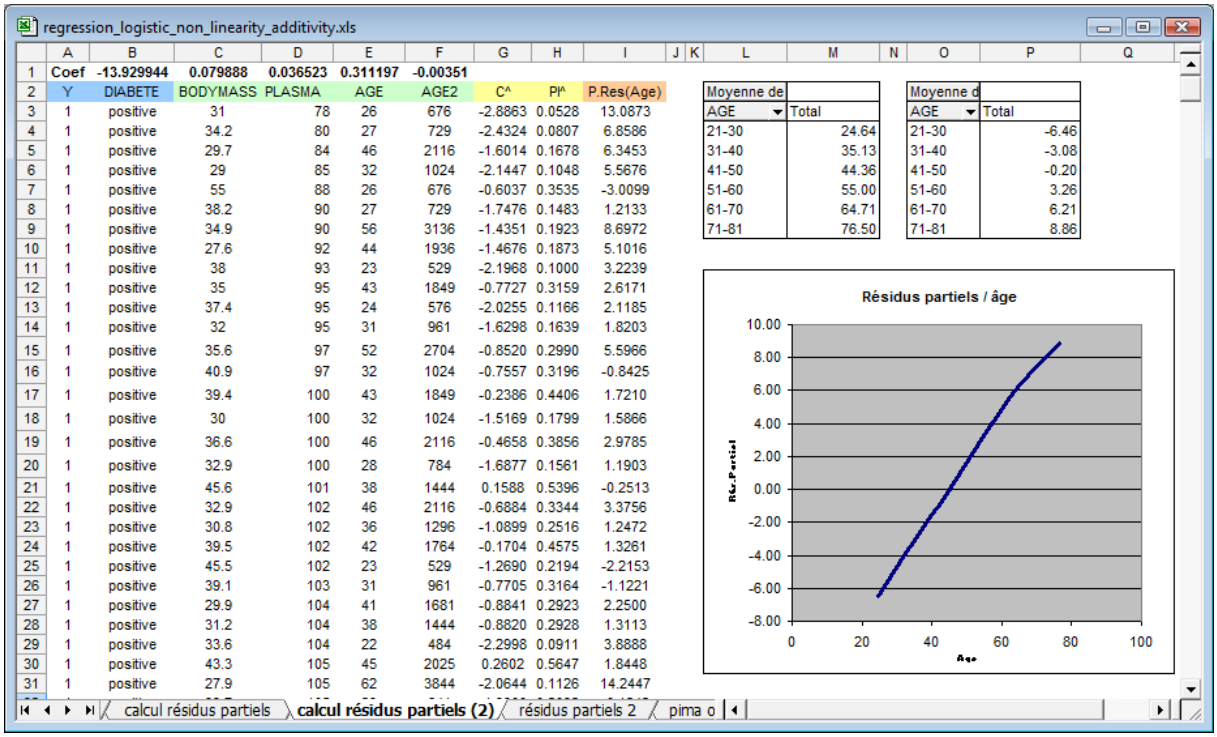


Fig. 8.22. Régression - Fichier PIMA - DIABETE = f(BODYMASS, PLASMA, AGE, AGE^2)

de degré zéro pour estimer la position du point sur l'ordonnée. D'où des graphiques qui sont assez similaires finalement.



	A	B	C	D	E	F	G	H	I
1	Coef	-13.929944	0.079888	0.036523	0.311197	-0.00351			
2	Y	DIABETE	BODYMASS	PLASMA	AGE	AGE2	C^	PI^	P.Res(Age)
3	1	positive	31	78	26	676	-2.8863	0.0528	13.0873
4	1	positive	34.2	80	27	729	-2.4324	0.0807	6.8586
5	1	positive	29.7	84	46	2116	-1.6014	0.1678	6.3453
6	1	positive	29	85	32	1024	-2.1447	0.1048	5.5676
7	1	positive	55	88	26	676	-0.6037	0.3535	-3.0099
8	1	positive	38.2	90	27	729	-1.7476	0.1483	1.2133
9	1	positive	34.9	90	56	3136	-1.4351	0.1923	8.6972
10	1	positive	27.6	92	44	1936	-1.4676	0.1873	5.1016
11	1	positive	38	93	23	529	-2.1968	0.1000	3.2239
12	1	positive	35	95	43	1849	-0.7727	0.3159	2.6171
13	1	positive	37.4	95	24	576	-2.0255	0.1166	2.1185
14	1	positive	32	95	31	961	-1.6298	0.1639	1.8203
15	1	positive	35.6	97	52	2704	-0.8520	0.2990	5.5966
16	1	positive	40.9	97	32	1024	-0.7557	0.3196	-0.8425
17	1	positive	39.4	100	43	1849	-0.2386	0.4406	1.7210
18	1	positive	30	100	32	1024	-1.5169	0.1799	1.5866
19	1	positive	36.6	100	46	2116	-0.4658	0.3856	2.9785
20	1	positive	32.9	100	28	784	-1.6877	0.1561	1.1903
21	1	positive	45.6	101	38	1444	0.1588	0.5396	-0.2513
22	1	positive	32.9	102	46	2116	-0.6884	0.3344	3.3756
23	1	positive	30.8	102	36	1296	-1.0899	0.2516	1.2472
24	1	positive	39.5	102	42	1764	-0.1704	0.4575	1.3261
25	1	positive	45.5	102	23	529	-1.2690	0.2194	-2.2153
26	1	positive	39.1	103	31	961	-0.7705	0.3164	-1.1221
27	1	positive	29.9	104	41	1681	-0.8841	0.2923	2.2500
28	1	positive	31.2	104	38	1444	-0.8820	0.2928	1.3113
29	1	positive	33.6	104	22	484	-2.2998	0.0911	3.8888
30	1	positive	43.3	105	45	2025	0.2602	0.5647	1.8448
31	1	positive	27.9	105	62	3844	-2.0644	0.1126	14.2447

Moyenne de	
AGE	Total
21-30	24.64
31-40	35.13
41-50	44.36
51-60	55.00
61-70	64.71
71-81	76.50

Moyenne d	
AGE	Total
21-30	-6.46
31-40	-3.08
41-50	-0.20
51-60	3.26
61-70	6.21
71-81	8.86

Fig. 8.23. DIABETE = f(BODYMASS, PLASMA, AGE, AGE^2) - Résidus partiels par rapport à AGE

pour construire la courbe des résidus partiels. Comme nous le disions plus haut, elle a quand même plus d'allure, elle est moins heurtée.

Le code utilisé est le suivant

```
#régression avec lrm
modele <- lrm(DIABETE ~ BODYMASS + PLASMA + AGE, x=T, y=T, data=donnees)
print(modele)
#graphique des résidus partiels
par(mfrow=c(2,2))
plot.lrm.partial(modele)
#construire le carré de AGE et le rajouter aux données
age2 <- donnees$AGE^2
donnees <- cbind(donnees,age2)
#régression avec lrm
modele.bis <- lrm(DIABETE ~ BODYMASS+PLASMA+AGE+age2, x=T, y=T, data=donnees)
print(modele.bis)
#nouveau graphique des résidus partiels
par(mfrow=c(2,2))
plot.lrm.partial(modele.bis)
```

Nous retrouvons la trame ci-dessus (Tanagra + Excel). Voyons les principaux résultats :

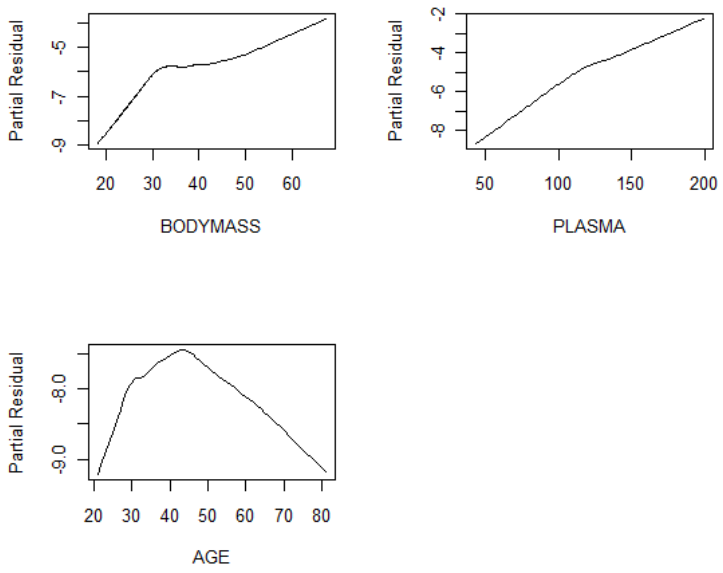


Fig. 8.24. DIABETE = f(BODYMASS, PLASMA, AGE) - Résidus partiels pour chaque explicative

- Les coefficients de la régression sont bien évidemment les mêmes que ceux de Tanagra. La grande nouveauté ici est que nous disposons automatiquement des résidus partiels par rapport à toutes les variables explicatives (Figure 8.24).
- Un seul coup d'oeil suffit à détecter les configurations à problèmes.
- On y constate que PLASMA est pris en compte correctement avec une relation linéaire ; nous savions déjà à quoi nous en tenir par rapport à BODYMASS (voir section 8.2.1) ; la relation par rapport à AGE est manifestement quadratique.
- Nous avons donc créé la variable AGE^2, nous l'avons insérée dans la régression, puis nous avons de nouveau demandé les résidus partiels (Figure 8.25).
- C'est quand même beau la science. Avec cette nouvelle variable, le rôle de l'âge est parfaitement pris en compte dans la détermination du diabète chez les indiens PIMA. Les résidus partiels par rapport à AGE et AGE^2 suivent une droite presque parfaite.

A titre de vérification, nous affichons les 10 premières valeurs des résidus partiels pour le 1^{er} et le 2^{nd} modèle (Figure 8.26). Il faut comparer les valeurs de la colonne AGE avec ceux produits sous Excel (Figure 8.22 et 8.23). La correspondance est exacte. C'est toujours rassurant.

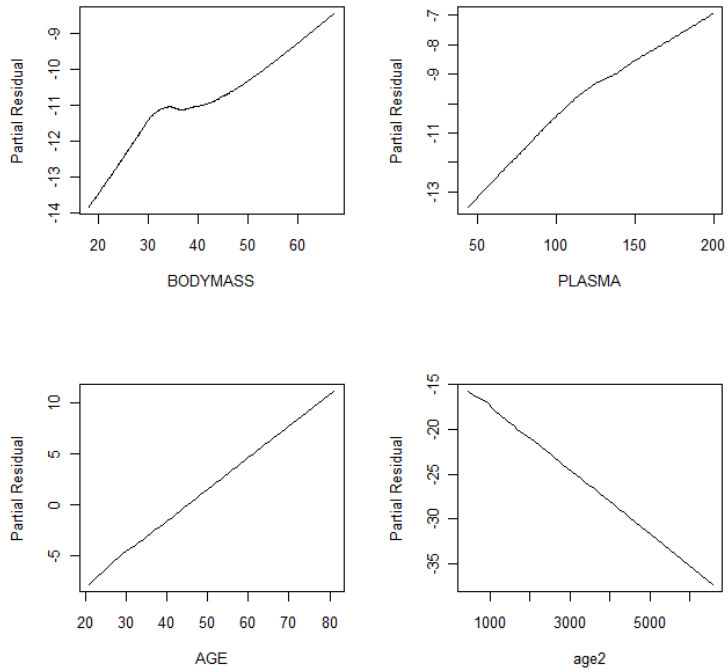


Fig. 8.25. DIABETE = f(BODYMASS, PLASMA, AGE, AGE^2) - Résidus partiels pour chaque explicative

```
R Console
> res1 <- residuals(modele,"partial")
> res1[1:10,]
     BODYMASS     PLASMA        AGE
1  10.0999570 10.0903520  8.0637894
2   5.4611679  5.2354491  3.1664896
3   2.4833660  2.8037278  1.1378601
4   6.6267990  7.0455359  4.9423324
5  -1.8479346 -3.6561337 -6.0381758
6   0.4981159  0.2688643 -2.1555747
7  -1.9152084 -1.8482749 -3.4404390
8   1.9983711  2.7915979  0.7839484
9   0.6150958  0.5104386 -2.1354407
10 -1.2947326 -1.0590346 -3.2020272
> res2 <- residuals(modele.bis,"partial")
> res2[1:10,]
     BODYMASS     PLASMA        AGE        age2
1   7.4700950  7.8423408 13.084695   2.621010
2   1.1866407  1.3762912  6.856797  -4.104100
3  -5.5998529 -4.9046176  6.342538 -15.399075
4  -2.0760174 -1.2883381  5.565539  -7.986709
5  -6.7074612 -7.8872931 -3.010165 -13.473850
6  -4.1381750 -3.9028491  1.212431  -9.748466
7  -5.9452518 -5.4462966  8.693691 -19.739801
8  -6.3883390 -5.2331587  5.099422 -15.388049
9  -0.8987214 -0.5378502  3.223075  -5.791092
10 -7.9692877 -7.2957083  2.616108 -17.254821
>
```

Fig. 8.26. 10 premières valeurs des résidus partiels pour les 2 modèles étudiés

9

"Covariate Pattern" et statistiques associées

9.1 Notion de "Covariate pattern"

Lorsque les données sont constituées de variables qualitatives ou lorsqu'elles sont produites par expérimentation, il arrive que plusieurs observations partagent la même description c.-à-d. elles portent les mêmes valeurs sur les variables explicatives. On parle aussi de "données groupées" [23] (pages 434 à 438). On appelle "covariate pattern" une combinaison de valeurs des variables explicatives [9] (page 144). Elle est partagée par plusieurs individus. Dans ce qui suit, les termes "covariate pattern", "groupe" ou "profil" auront la même signification dans notre esprit.

alcool	surpoids	n	y	pp.obs
1	1	47	16	0.34
1	2	27	11	0.41
1	3	55	39	0.71
2	1	52	20	0.38
2	2	25	15	0.60
2	3	64	41	0.64
3	1	63	38	0.60
3	2	26	15	0.58
3	3	40	33	0.83
Total		399	228	0.57

Fig. 9.1. Tableau de comptage des effectifs pour chaque "covariate pattern"

Prenons un exemple pour illustrer notre propos. Le fichier HYPERTENSION est composé de $n = 399$ observations. La variable dépendante HYPERTENSION prend 2 valeurs possibles $\{high : +, normal : -\}$; les variables explicatives sont SURPOIDS (3 valeurs possibles, $\{1, 2, 3\}$) et ALCOOL (3 valeurs possibles, $\{1, 2, 3\}$).

Dans le fichier, il y a $3 \times 3 = 9$ combinaisons distinctes des variables explicatives. On dit qu'il y a $M = 9$ "covariate pattern" (ou groupes). A chaque combinaison sont associés n_m individus, dont une partie sont positifs. Nous notons y_m le nombre d'observations positives dans le groupe m, f_m est la proportion observée de positifs, et π_m la probabilité a posteriori d'être positif que l'on veut modéliser à l'aide de la régression logistique. Nous avons résumé ces informations dans un tableau (Figure 9.1) :

- Pour la première combinaison, $m = 1$, composée de (ALCOOL = 1, SURPOIDS = 1), nous disposons de $n_1 = 47$ observations, dont $y_1 = 16$ sont positifs. La proportions de positifs est donc égale à $f_1 = \frac{16}{14} = 0.34$.
- Nous pouvons faire de même pour chaque groupe.
- Nous disposons $n = 399$ observations.
- Et le nombre total de positifs dans le fichier est $n_+ = 228$.
- La prévalence des positifs (si le fichier est issu d'un tirage aléatoire simple dans la population) est donc estimé avec $\hat{p} = \frac{228}{399} = 0.57$.

Pourquoi s'intéresser à cette configuration qui n'est qu'un cas particulier finalement ? La première différence est dans la modélisation de la variable aléatoire y_m, elle suit une loi binomiale $\mathcal{B}(n_m, \pi_m)$, la varisemblance et la log-vraisemblance s'écrivent différemment [23] (pages 435 et 436).

En pratique, les cas des données groupées nous emmène à considérer 2 nouveaux éléments :

1. Nous disposons de nouvelles statistiques d'évaluation de la régression basées sur les résidus.
2. Nous pouvons analyser finement le rôle de chaque groupe pour détecter ceux qui présentent des caractéristiques particulières ou qui pèsent de manière exagérée sur les résultats. Lorsque les données sont issues d'expérimentations, cette fonctionnalité nous permet de situer le rôle de chaque groupe expérimental dans la régression.

9.2 Levier associé aux "Covariate pattern"

Avant d'aborder ces sujets, présentons tout d'abord le "levier" associé à chaque "covariate pattern". Il joue un rôle très important dans la régression. Il indique l'écartement d'un groupe par rapport aux autres dans l'espace de représentation. Il caractérise également l'influence d'un groupe dans la prédiction des probabilités des autres groupes. Notons h_m le levier du groupe m. Si $h_m = 0$, le groupe n'a aucune influence.

Le levier du covariate pattern m s'écrit

$$h_m = n_m \hat{\pi}_m (1 - \hat{\pi}_m) x_m (X'VX)^{-1} x'_m \tag{9.1}$$

Remarque : A propos de la matrice de variance covariance des coefficients. $(X'VX)^{-1} = \hat{\Sigma}$ est la matrice de variance covariance des coefficients. Elle peut être obtenue dans la régression sur les n observations individuelles (voir section 3.3.1) (Figure 9.2). Mais nous pouvons également la calculer à partir des données réduites aux "covariate pattern". Dans ce cas, la matrice X comporte M lignes et $J + 1$ colonnes ; V est une matrice diagonale de terme générique $n_m \times \hat{\pi}_m \times (1 - \hat{\pi}_m)$ (Figure 9.3). Sur les données HYPERTENSION, on notera que la matrice X comporte les 9 combinaisons de valeurs que nous pouvons former avec les variables explicatives, la première colonne étant toujours la constante. La matrice V est de taille (9×9). La matrice de variance covariance obtenue concorde avec celle calculée sur les données individuelles produite par le logiciel R.

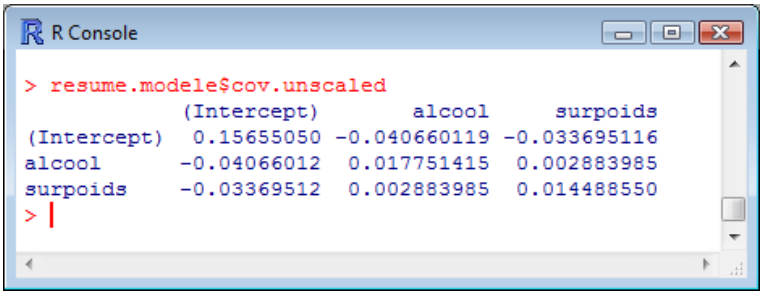


Fig. 9.2. Hypertension - $\hat{\Sigma}$ à partir des données individuelles - Logiciel R

Matrice X

1	1	1
1	1	2
1	1	3
1	2	1
1	2	2
1	2	3
1	3	1
1	3	2
1	3	3

Matrice V

10.493	0	0	0	0	0	0	0	0
0	6.735	0	0	0	0	0	0	0
0	0	12.961	0	0	0	0	0	0
0	0	0	12.769	0	0	0	0	0
0	0	0	0	6.097	0	0	0	0
0	0	0	0	0	13.156	0	0	0
0	0	0	0	0	0	15.671	0	0
0	0	0	0	0	0	0	5.713	0
0	0	0	0	0	0	0	0	6.693

X'VX

90.2868	178.4622	174.4509
178.4622	410.9674	333.2343
174.4509	333.2343	408.3984

(X'VX)^{-1} = Mat. Var.Covar

0.15655	-0.04066	-0.03370
-0.04066	0.01775	0.00288
-0.03370	0.00288	0.01449

Fig. 9.3. Hypertension - $\hat{\Sigma}$ à partir des données groupées

Leviers pour les données HYPERTENSION

Nous devons tout d'abord lancer la régression logistique pour pouvoir produire les éléments permettant de calculer les leviers. Tanagra nous indique que le modèle est globalement significatif, même si par ailleurs les pseudo-R^2 paraissent singulièrement faibles. Les deux variables explicatives sont significatives au risque 5% (Figure 9.4).

A partir d'ici, nous pouvons produire le LOGIT pour chaque "covariate pattern" et en déduire la quantité $\hat{\pi}_m$. Voyons ce qu'il en est pour le premier profil de coordonnées $(alcool = 1, surpoids = 1)$:

- $\hat{C}_1 = -1.673659 + 0.410675 \times 1 + 0.582889 \times 1 = -0.6791$
- $\hat{\pi}_1 = \frac{1}{1+e^{-(-0.6791)}} = 0.3365$
- $\hat{\pi}_1 \times (1 - \hat{\pi}_1) = 0.2233$
-

$$h_1 = 47 \times 0.2233 \times \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 0.15655 & -0.04066 & -0.03370 \\ -0.04066 & 0.01775 & 0.00288 \\ -0.03370 & 0.00288 & 0.01449 \end{pmatrix} \times \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 0.4811$$

Nous avons complété le tableau des leviers (Figure 9.5). Essayons d'en analyser le contenu :

Predicted attribute		hypertension
Positive value		high
Number of examples		399
Model Fit Statistics		
Criterion	Intercept	Model
AIC	546.961	519.871
SC	550.95	531.838
-2LL	544.961	513.871
Model Chi² test (LR)		
Chi-2		31.0897
d.f.		2
P(>Chi-2)		0
R²-like		
McFadden's R²		0.057
Cox and Snell's R²		0.075
Nagelkerke's R²		0.1006

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.673659	-	-	-
alcool	0.410675	0.1332	9.5008	0.0021
surpoids	0.583889	0.1204	23.5307	0

Fig. 9.4. Hypertension - Résultats de la régression logistique

N° covariate	alcool	surpoids	n(m)	y(m)	logit(m)	pi^(m)	pi x (1 - pi)	h(m)
1	1	1	47	16	-0.6791	0.3365	0.2233	**0.4811**
2	1	2	27	11	-0.0952	0.4762	0.2494	0.1865
3	1	3	55	39	0.4887	0.6198	0.2356	**0.4991**
4	2	1	52	20	-0.2684	0.4333	0.2456	0.3007
5	2	2	25	15	0.3155	0.5782	0.2439	0.0681
6	2	3	64	41	0.8994	0.7108	0.2056	0.3651
7	3	1	63	38	0.1423	0.5355	0.2487	**0.5760**
8	3	2	26	15	0.7261	0.6740	0.2197	0.1722
9	3	3	40	33	1.3100	0.7875	0.1673	0.3513
							Somme	3

Fig. 9.5. Hypertension - Calcul des leviers pour chaque "covariate pattern"

- Première vérification, on sait que $\sum_m h_m = J + 1$. Notre tableau a été correctement construit puisque $0.4811 + 0.1865 + 0.4991 + \cdots + 0.3513 = 3$.
- Dans le cadre des "covariate pattern", les données sont souvent binaires ou correspondent à des échelles (notre configuration), essayer de détecter des points atypiques à l'aide du levier n'a pas trop de sens.
- Le levier prend des valeurs élevées essentiellement lorsque la conjonction de 2 évènements survient : l'effectif du groupe n_m est élevé, il est mal modélisé c.-à-d. $\hat{\pi}_m \approx 0.5$. Pour les données HYPERTENSION, nous distinguerons les covariate pattern n^o1, n^o3 et surtout le n^o7. Ils pèsent fortement sur les résultats de la régression.
- Souvent un histogramme des leviers permet de repérer facilement ces groupes (Figure 9.6).

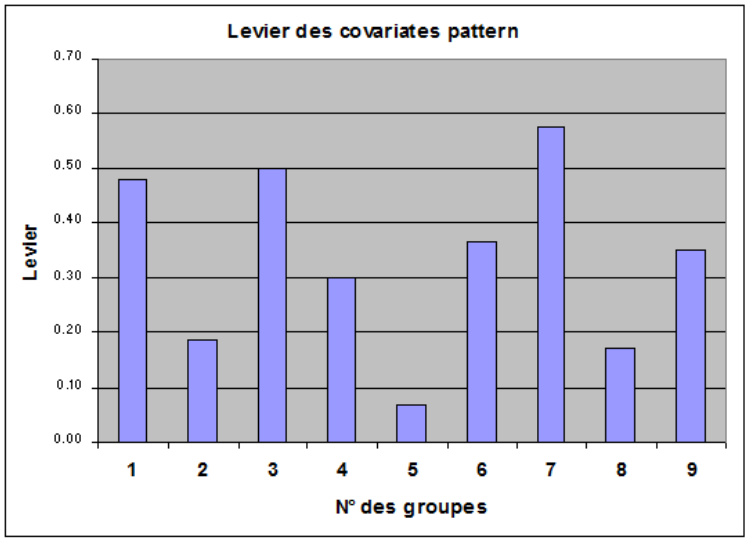


Fig. 9.6. Hypertension - Histogramme des leviers pour chaque "covariate pattern"

9.3 Résidu de Pearson et Résidu déviance

Les résidus que nous présentons dans cette section confrontent, d'une manière ou d'une autre, les probabilités prédites par le modèle et les probabilités observées pour chaque covariate pattern. **Ils mettent en exergue les profils mal modélisés**.

Les statistiques de tests qui en sont dérivées, sous l'hypothèse de l'adéquation du modèle aux données, suivent une loi du χ^2. Nous pouvons ainsi vérifier si le modèle est correct. Notons que l'approximation de la loi statistique n'est plus valable dés que nous nous rapprochons de la configuration des données individuelles, avec $M \approx n$. **Ce qui limite l'utilisation de ces tests aux seuls cas des données groupées avec** $M \ll n$.

9.3.1 Résidu de Pearson

On appelle résidu de Pearson pour le profil m la quantité

$$r_m = \frac{y_m - \hat{y}_m}{\sqrt{n_m \hat{\pi}_m (1 - \hat{\pi}_m)}} \tag{9.2}$$

où $\hat{y}_m = n_m \times \hat{\pi}_m$ est le nombre prédit de positifs dans le groupe m, estimée par la régression logistique.

Le résidu de Pearson sera d'autant plus grand que :

1. La prédiction $\hat{y}_m$ est mauvaise ;
2. Les effectifs n_m sont faibles ;
3. La probabilité estimée $\hat{\pi}_m$ est proche de 0 ou de 1.

La statistique de Pearson est définie de la manière suivante

$$\chi^2 = \sum_{m=1}^{M} r_m^2 \tag{9.3}$$

Si le modèle étudié est exact, et si n_m est assez grand quel que soit m, alors la statistique de Pearson suit une loi du χ^2 à $(M - J - 1)$ degrés de liberté. Nous pouvons utiliser ce test pour vérifier l'adéquation du modèle au donnée. Nous rejetons le modèle si la p-value du test est plus petit que le risque de première espèce que nous nous sommes fixés.

Attention, dans le cas des données individuelles, avec $M \approx n$, ce test n'est plus valable. Il ne faut surtout pas l'utiliser [9] (page 146).

On appelle résidu standardisé de Pearson la quantité [9] (page 173)

$$r_{sm} = \frac{r_m}{\sqrt{1 - h_m}} \tag{9.4}$$

Enfin, on appelle "contribution à la statistique de Pearson" [9] (page 174),

$$\Delta\chi_m^2 = \frac{r_m^2}{1 - h_m} = r_{sm}^2 \tag{9.5}$$

Elle indique (une approximation de) la diminution du χ^2 de Pearson si on supprime le profil m de la régression. Elle est basée sur une approximation linéaire d'une courbe qui ne l'est pas [9] (page 174). Il n'en reste pas moins, nous le verrons dans l'exemple ci-dessous, qu'elle donne une idée assez précise de la variation.

Application aux données HYPERTENSION

N° covariate	alcool	surpoids	n(m)	y(m)	pp.pred(m).pi	y^(m)	pi x (1 - pi)	h(m)	r.pears	chi.pears	delta(chi.pears)
1	1	1	47	16	0.336	15.8	0.223	0.481	0.057	0.003	0.006
2	1	2	27	11	0.476	12.9	0.249	0.186	-0.716	0.513	0.630
3	1	3	55	39	0.620	34.1	0.236	0.499	1.364	1.861	3.716
4	2	1	52	20	0.433	22.5	0.246	0.301	-0.708	0.502	0.718
5	2	2	25	15	0.578	14.5	0.244	0.068	0.221	0.049	0.052
6	2	3	64	41	0.711	45.5	0.206	0.365	-1.239	1.534	2.416
7	3	1	63	38	0.536	33.7	0.249	0.576	1.077	1.160	2.735
8	3	2	26	15	0.674	17.5	0.220	0.172	-1.056	1.114	1.346
9	3	3	40	33	0.788	31.5	0.167	0.351	0.580	0.336	0.518
									Somme	**7.0711**	
									d.f	**6**	
									p-value	**0.3143**	

Fig. 9.7. Hypertension - Tableau de calcul du résidu de Pearson

Nous appliquons les différentes formules ci-dessus pour obtenir les résidus et la statistique de Pearson (Figure 9.7) :

- Pour rappel, pour le groupe n^o1, nous avons obtenu le LOGIT estimé avec $\hat{C}_1 = -1.673659 + 0.410675 \times 1 + 0.582889 \times 1 = -0.6791$
- Puis la probabilité estimée $\hat{\pi}_1 = \frac{1}{1+e^{-(-0.6791)}} = 0.3365$
- L'effectif estimé $\hat{y}_1 = n_1 \times \hat{\pi}_1 = 47 \times 0.3365 = 15.8$.
- Nous pouvons dès lors former les résidus pour chaque groupe. Pour le premier, nous $r_1 = \frac{16-15.8}{\sqrt{47\times0.3365\times(1-0.3365)}} = 0.057$.
- Nous obtenons la statistique de Pearson en faisant la somme des carrés des résidus individuels $\chi^2 = (0.057)^2 + (-0.716)^2 + \cdots + (0.580)^2 = 0.003 + 0.513 + \cdots + 0.336 = 7.0711$. Sous H_0 (le modèle est exact), elle suit une loi du χ^2 à $(M - J - 1 = 9 - 2 - 1 = 6)$ degrés de liberté. Nous aboutissons à un p-value de 0.3143, supérieur au risque usuel de 5% que nous souhaitons utiliser. Les probabilités prédites dans les groupes (représenté par les effectifs prédits $\hat{y}_m$) sont compatibles avec les probabilités observés (représenté par les effectifs observés y_m). Les données sont donc compatibles avec l'hypothèse d'exactitude du modèle.
- Pour repérer les groupes mal modélisés, nous pourrions comparer $|r_m|$ avec la valeur seuil approximative de 2 pour détecter les écarts significatif. Ce seuil mime un peu le fractile de la loi normale pour un test bilatéral à 5% (le véritable seuil est 1.96). C'est une première approche un peu fruste. Elle n'est valable que si n_m est suffisamment grand pour que l'approximation de la loi binomiale par la loi normale soit justifiée [9] (page 175) c.-à-d. $n_m\pi_m(1-\pi_m) > 9$ ([1])
- Pour une meilleure évaluation des profils, nous nous penchons plutôt sur la dernière colonne $\Delta\chi^2_m$ (Équation 9.5). Nous constatons que le profil n^o3 (alcool = 1 et surpoids = 3 ; les gros qui boivent pas) est mal modélisé, c'est celui qui perturbe le plus les résultats avec $\Delta\chi^2_3 = 3.716$.
- Une autre manière de repérer les profils à problème serait de comparer cette valeur avec le seuil critique définie par la loi du $\chi^2(1)$. A 5%, il est de 3.84. Nous sommes à la lisière de la région critique.
- L'interprétation de la quantité $\Delta\chi^2_3 = 3.716$ est assez démoniaque : si nous supprimons le profil n^o3 des données (c.-à-d. tous les individus correspondant au profil n^o3) et que nous relançons la régression, la statistique de Pearson que nous obtiendrons devrait être aux alentours de $(7.0711 - 3.716) = 3.355$. Avec une loi du $\chi^2(5)$, nous obtiendrions une p-value de 0.6454. La compatibilité des données avec le modèle serait renforcée. Bien entendu, empressons-nous de vérifier cela en réalisant les calculs sans le covariate pattern incriminé.

Régression sans le covariate pattern n^o3

Nous avons relancé les calculs sur les mêmes données, sans les observations du covariate pattern n^o3. Le nouveau fichier comporte $n = 344$ individus. L'objectif est de vérifier si la statistique de Pearson obtenue à l'issue des opérations correspond peu ou prou à ce qui est annoncé ci-dessus.

Les coefficients de ALCOOL et SURPOIDS restent significatifs et, après formation du tableau de calcul de la statistique de Pearson, la véritable valeur de la statistique de Pearson est $\chi^2 = 3.314$ (Figure 9.8). L'approximation donnée ci-dessus était effectivement assez bonne, puisqu'en retranchant la contribution au χ^2, nous avions prévu une valeur de $(7.0711 - 3.716) = 3.355$. Les valeurs divergent seulement à partir de la seconde décimale.

1. B. Grais, *Méthodes statistiques*, Dunod, 2003 ; page 103.

Coefficients de la régression logistique sans le covariate pattern n°3

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.8260439	-	-	-
alcool	0.5516103	0.1531	12.9804	0.0003
surpoids	0.4634875	0.1341	11.9373	0.0006

Tableau de calcul du KHI-2 de Pearson

N° covariate	alcool	surpoids	n(m)	y(m)	C^	Pi^	y^(m)	r.pears	chi.pears
1	1	1	47	16	-0.811	0.308	14.5	0.486	0.236
2	1	2	27	11	-0.347	0.414	11.2	-0.070	0.005
4	2	1	52	20	-0.259	0.436	22.6	-0.740	0.548
5	2	2	25	15	0.204	0.551	13.8	0.494	0.244
6	2	3	64	41	0.668	0.661	42.3	-0.344	0.118
7	3	1	63	38	0.292	0.573	36.1	0.491	0.241
8	3	2	26	15	0.756	0.680	17.7	-1.132	1.281
9	3	3	40	33	1.219	0.772	30.9	0.800	0.640
								Somme	3.314
								d.f	5
								p-value	0.6517

Fig. 9.8. Hypertension - Calcul du résidu de Pearson sans le profil $n^o 3$

9.3.2 Résidu déviance

On appelle résidu déviance pour le profil m la quantité

$$d_m = \text{signe}(y_m - \hat{y}_m) \times \sqrt{2\left[y_m \ln \frac{y_m}{\hat{y}_m} + (n_m - y_m) \ln \frac{n_m - y_m}{n_m - \hat{y}_m}\right]} \tag{9.6}$$

Lorsque $y_m = 0$, nous utilisons [9] (page 146)

$$d_m = -\sqrt{2n_m |\ln(1 - \hat{\pi}_m)|}$$

et pour $y_m = n_m$

$$d_m = \sqrt{2n_m |\ln(\hat{\pi}_m)|}$$

On en déduit la déviance

$$D = \sum_{m=1}^{M} d_m^2 \tag{9.7}$$

Cette statistique quantifie l'écart entre les probabilités estimées et les probabilités observées. Dans les mêmes conditions que pour le résidu de Pearson (n_m assez grand, $\forall m$; $M \ll n$), sous l'hypothèse d'exactitude du modèle, D suit une loi du χ^2 à $(M - J - 1)$ degrés de liberté. A l'usage, on se rend compte, non sans raisons [23] (page 437), que la déviance est très proche de la statistique de Pearson.

Comme précédemment, nous pouvons calculer la contribution d'un profil à la déviance

$$\Delta D_m = d_m^2 + r_m^2 \frac{h_m}{1 - h_m} \tag{9.8}$$

Elle indique la réduction de la déviance si on retire le profil m de la régression. Ici également, nous pouvons la comparer avec un seuil critique définie à l'aide d'une loi du $\chi^2(1)$ pour détecter les écarts significatifs (à 5%, le seuil est 3.84).

Application aux données HYPERTENSION

N° covariate	alcool	surpoids	n(m)	y(m)	y^(m)	h(m)	r.pears	r.dev	dev.	delta(dev)
1	1	1	47	16	15.8	0.481	0.057	0.057	0.003	0.006
2	1	2	27	11	12.9	0.186	-0.716	-0.719	0.516	0.634
3	1	3	55	39	34.1	0.499	1.364	1.390	1.932	3.787
4	2	1	52	20	22.5	0.301	-0.708	-0.712	0.507	0.723
5	2	2	25	15	14.5	0.068	0.221	0.221	0.049	0.052
6	2	3	64	41	45.5	0.365	-1.239	-1.213	1.471	2.353
7	3	1	63	38	33.7	0.576	1.077	1.082	1.171	2.746
8	3	2	26	15	17.5	0.172	-1.056	-1.033	1.068	1.300
9	3	3	40	33	31.5	0.351	0.580	0.593	0.352	0.534
								Somme	7.0690	
								d.f.	6	
								p-value	0.3145	

Fig. 9.9. Hypertension - Tableau de calcul du résidu déviance

Comme dans la section précédente (section 9.3.1), le point de départ est l'estimation des paramètres de la régression. Nous énumérons les grandes étapes de calcul (Figure 9.9) :

- Avec les paramètres estimés, nous sommes en mesure de produire le LOGIT, la probabilité d'être positif et les effectifs prédits pour chaque groupe. Ainsi, pour le groupe n^o1, $\hat{y}_1 = 15.8$.
- Nous pouvons calculer le résidu déviance (Équation 9.6), par exemple

$$d_1 = + \times \sqrt{2\left[16 \ln \frac{16}{15.8} + (47-16) \ln \frac{47-16}{47-15.8}\right]} = 0.057$$

- Il reste à faire la somme des carrés des résidus, soit $D = (0.057)^2 + (-0.719)^2 + (1.390)^2 + \cdots + (0.593)^2 = 0.003 + 0.516 + 1.932 + \cdots + 0.352 = 7.0690$.
- Avec un χ^2 à $(M - J - 1 = 9 - 2 - 1 = 6)$ degrés de liberté, nous obtenons une p-value de 0.3145. La p-value est plus grande que le risque usuel de 5% que l'on s'est choisi. Le modèle est correct.
- Avec la dernière colonne, nous pouvons évaluer la contribution de chaque profil à la déviance. Si l'on regarde encore une fois le profil n^o3, nous avons

$$\Delta D_3 = (1.390)^2 + (1.364)^2 \times \frac{0.499}{1 - 0.499} = 3.787$$

En comparant cette valeur avec le seuil de $\chi^2_{1-0.05}(1) = 3.84$, le profil mérite vraiment que l'on s'y penche sérieusement. Si l'on retire ce profil des données et que nous ré-estimons le modèle, nous obtiendrons une déviance de $(7.0690 - 3.787) = 3.28$. Avec maintenant un $\chi^2(5)$, la p-value serait de 0.6565. La conclusion est la même qu'avec le résidu de Pearson, le retrait du profil n^o3 renforce la qualité du modèle.

9.4 Mesurer l'impact de chaque "covariate pattern" sur les coefficients

9.4.1 La distance de Cook

L'objectif de la distance de Cook est d'indiquer l'effet de la suppression d'un profil sur les paramètres c.-à-d. mesurer l'écart entre les vecteurs de coefficients selon que les observations associées à un profil

sont présentes ou non dans les données. Elle est très utilisée en régression linéaire multiple pour détecter les points influents.

La distance de Cook pour le profil m s'écrit [9] (page 173)

$$(\Delta\hat{a})_m = r_m^2 \frac{h_m}{(1-h_m)^2} = r_{sm}^2 \frac{h_m}{1-h_m} \tag{9.9}$$

La distance de Cook est désignée sous cette appellation dans le logiciel R. Nous montrons un exemple d'application dans la section suivante.

9.4.2 Les critères C et CBAR

L'objectif des critères C et CBAR ($\bar{C}$) est aussi de mesurer l'écart entre les vecteurs de paramètres suite à la suppression d'un profil. C et CBAR sont proposés sous ces noms dans le logiciel SAS. Un étude rapide des formules montre que l'indicateur C de SAS et la distance de Cook de R sont identiques.

Nous calculons C et CBAR de la manière suivante pour chaque profil m :

- Pour l'indicateur CBAR

$$\bar{C}_m = r_m^2 \frac{h_m}{1-h_m} \tag{9.10}$$

- Pour l'indicateur C, qui est ni plus ni moins que la Distance de Cook,

$$C_m = r_m^2 \frac{h_m}{(1-h_m)^2} \tag{9.11}$$

  Par rapport à CBAR, le critère C rend plus fort l'effet du levier à mesure que ce dernier augmente.

Application aux données HYPERTENSION

Nous disposons de toutes les informations nécessaires aux calculs lors de la présentation des différents types de résidus. Il ne nous reste plus qu'à compléter la feuille Excel (Figure 9.10) :

- A partir des résultats dans les différents tableaux ci-dessus (Figure 9.5 et 9.7), nous pouvons obtenir C_1 et $\bar{C}_1$ pour le premier individu, avec

$$\bar{C}_1 = (0.057)^2 \frac{0.481}{1-0.481} = 0.003$$

  et

$$C_1 = (0.057)^2 \frac{0.481}{(1-0.481)^2} = 0.006$$

- Manifestement, il y a des choses à dire sur les profils n^o3 et n^o7 : nous savons que le n^o3 pose problème parce qu'il est mal modélisé; le n^o7 pèse parce qu'il présente un levier élevé (on le voit bien, $C_7 > C_3$ alors que dans le même temps $\bar{C}_7 < \bar{C}_3$, l'indicateur C accentue le rôle du levier), il est de plus assez mal modélisé si l'on se réfère aux contributions au χ^2 de Pearson et à la déviance.

N° covariate	alcool	surpoids	h(m)	r.pears	CBAR	C
1	1	1	0.481	0.057	0.003	0.006
2	1	2	0.186	-0.716	0.117	0.144
3	1	3	0.499	1.364	1.855	3.703
4	2	1	0.301	-0.708	0.216	0.309
5	2	2	0.068	0.221	0.004	0.004
6	2	3	0.365	-1.239	0.882	1.389
7	3	1	0.576	1.077	1.575	3.715
8	3	2	0.172	-1.056	0.232	0.280
9	3	3	0.351	0.580	0.182	0.280

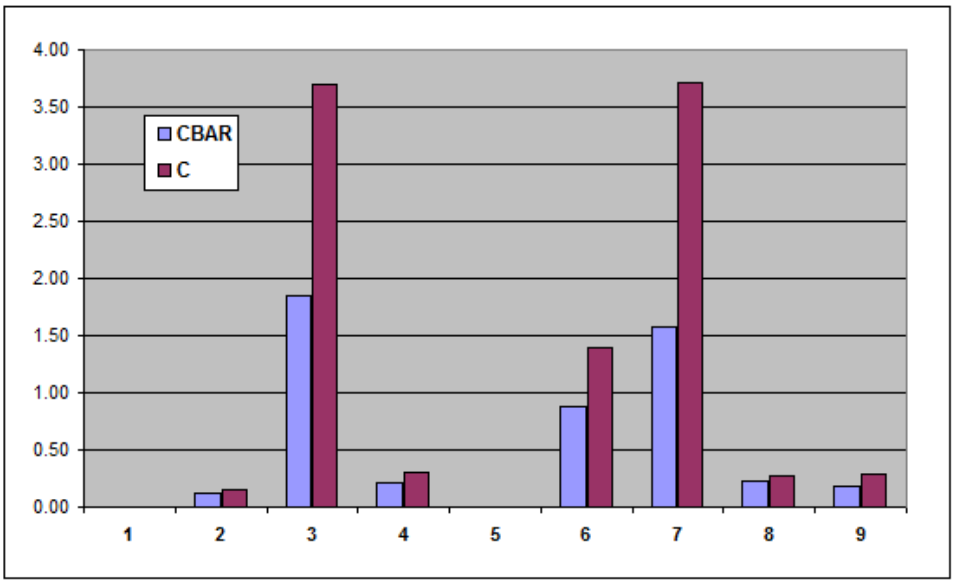


Fig. 9.10. Hypertension - Tableau de calcul des C et CBAR

- Agrémenter la présentation d'un graphique simple permet de détecter rapidement les profils à considérer avec attention.
- De manière générale, il est possible de définir toute une série de graphiques qui permettent de détecter visuellement les profils à étudier en priorité (un nuage de points entre $\hat{\pi}_m$ en abcisse et $\Delta\chi^2_m$ en ordonnée par exemple) [2].

9.4.3 Les critères DFBETA et DFBETAS

On sait avec les indicateurs C et $\bar{C}$ quels profils pèsent sur le vecteur des paramètres. Mais nous ne savons pas sur quel coefficient en particulier. L'intérêt pour nous est de pouvoir analyser les interactions entre les profils et les variables explicatives.

Les DFBETA et DFBETAS permettent de quantifier la modification d'un coefficient associée à une variable lorsque nous supprimons un profil (les observations relatives à un profil) des données :

2. Voir [9], pages 176 à 182 pour plusieurs propositions de graphiques. Certains sont particulièrement judicieux, notamment lorsque les auteurs proposent de rendre la taille des points proportionnels à $\Delta\hat{a}$.

– DFBETA mesure l'écart absolu du coefficient estimé avec ou sans le profil

$$DFBETA_{j,m} = (X'VX)^{-1}x'_m \frac{y_m - \hat{y}_m}{1 - h_m} \tag{9.12}$$

– DFBETAS mesure un écart normalisé par l'écart-type du coefficient estimé, il est surtout intéressant lorsque les variables sont mesurées sur des échelles différentes

$$DFBETAS_{j,m} = \frac{(X'VX)^{-1}x'_m}{\sqrt{(X'VX)^{-1}_j}} \frac{y_m - \hat{y}_m}{1 - h_m} = \frac{DFBETA_{j,m}}{\sqrt{(X'VX)^{-1}_j}} \tag{9.13}$$

Ici également, ces indicateurs ont été principalement développés dans le cadre de la régression linéaire. Il s'agit donc d'approximations pour la régression logistique. Nous le verrons cependant pour nos données, ils sont relativement précis.

Lorsque les variables explicatives sont (1) mesurées sur la même échelle (ex. mêmes unités), ou (2) directement des échelles de valeurs (cf. l'exemple Hypertension), ou (3) exclusivement des indicatrices, nous avons intérêt à utiliser directement le DFBETA. L'interprétation n'en sera que plus aisée. Dans le cas des données groupées, nous sommes souvent dans les situations (2) ou (3).

Comment interpréter la valeur d'un DFBETA relatif à un coefficient d'une variable explicative ? Si l'on supprime le profil m des données et que l'on estime le modèle sur les données restantes, le nouveau coefficient estimé pour la variable X_j s'écrira

$$\hat{a}_{j,(-m)} = \hat{a}_j - DFBETA_{j,m} \tag{9.14}$$

Nous disposons du nouveau coefficient *sans avoir à relancer explicitement l'estimation par le maximum de vraisemblance sur l'ensemble de données réduit.*

Application aux données HYPERTENSION

N°	alcool	surpoids	y(m)	y^(m)	h(m)	dfbeta.const	dfbeta.alcool	dfbeta.surpoids
1	1	1	16	15.8	0.481	0.029	-0.007	-0.006
2	1	2	11	12.9	0.186	-0.111	0.039	0.004
3	1	3	39	34.1	0.499	0.145	-0.140	0.124
4	2	1	20	22.5	0.301	-0.150	0.008	0.049
5	2	2	15	14.5	0.068	0.005	0.000	0.001
6	2	3	41	45.5	0.365	0.183	-0.025	-0.110
7	3	1	38	33.7	0.576	0.009	0.156	-0.106
8	3	2	15	17.5	0.172	0.100	-0.056	-0.012
9	3	3	33	31.5	0.351	-0.154	0.049	0.043

Fig. 9.11. Hypertension - Tableau de calcul des DFBETA

Toutes les informations nécessaires aux calculs ont été produites au fur et à mesure que nous avancions dans ce chapitre consacrée aux "covariate pattern". Nous produisons le tableau recensant les DFBETA pour chaque profil (Figure 9.11), nous détaillons le calcul pour le profil $n^o 7$:

$$(X'VX)^{-1} = \hat{\Sigma} = \begin{pmatrix} 0.15655 & -0.04066 & -0.03370 \\ -0.04066 & 0.01775 & 0.00288 \\ -0.03370 & 0.00288 & 0.01449 \end{pmatrix}$$

$$x_m = \begin{pmatrix} 1 & 3 & 1 \end{pmatrix}$$

$$(X'VX)^{-1} \times x'_m = \begin{pmatrix} 0.00088 \\ 0.01548 \\ -0.01055 \end{pmatrix}$$

$$\frac{y_m - \hat{y}_m}{1 - h_m} = \frac{38 - 33.7}{1 - 0.576} = 10.05410$$

$$(X'VX)^{-1} x'_m \frac{y_m - \hat{y}_m}{1 - h_m} = \begin{pmatrix} 0.00088 \\ 0.01548 \\ -0.01055 \end{pmatrix} \times 10.05410 = \begin{pmatrix} 0.009 \\ 0.156 \\ -0.106 \end{pmatrix}$$

Ce sont les valeurs que nous retrouvons pour le covariate pattern $n^o 7$ dans notre tableau récapitulatif (Figure 9.11, les valeurs ont été transposées en ligne).

Voyons maintenant comment lire ces informations. Lorsque nous retirons le profil $n^o 7$ de nos données,

- Le coefficient de ALCOOL va être diminué de 0.156 (Équation 9.14) ;
- Le coefficient de SURPOIDS va être augmenté de 0.106.
- Moralité : le profil $n^o 7$ a tendance à exacerber le rôle de l'ALCOOL et à atténuer le rôle du SURPOIDS dans la détermination du risque d'hypertension.
- A bien y regarder, on comprend le mécanisme. Il s'agit d'une population "d'alcolos maigrichons". L'ALCOOL prend une valeur élevée (ALCOOL = 3), SURPOIDS faible (SURPOIDS = 1), et il y a $f_7 = \frac{38}{63} = 60.3\%$ de positifs dans ce profil. Il n'est guère étonnant que le rôle de l'ALCOOL soit si décrié à partir de ce profil.

On remarquera par ailleurs que le profil $n^o 3$ joue exactement le rôle contraire. Ce sont des gros (SURPOIDS = 3) sobres (ALCOOL = 1), et il y a une majorité de positifs $f_3 = \frac{39}{55} = 70.9\%$. Les valeurs des DFBETA dans le tableau récapitulatif sont sans surprises, il vont dans le sens contraire de celles du profil $n^o 7$.

Régression sans le profil $n^o 7$

A titre de vérification, nous réalisons la régression sans les individus du profil $n^o 7$. Nous ne disposons plus que de $n = 336$ observations. Les résultats nous inspirent plusieurs commentaires (Figure 9.12 ; à mettre en parallèle avec les résultats de la régression sur la totalité des données, Figure 9.4) :

- Nous retrouvons bien les valeurs attendues des coefficients. En effet, pour la variable ALCOOL

$$\hat{a}_1 - DFBETA_{1,7} = 0.411 - 0.156 = 0.255 \approx 0.257 = \hat{a}_{1,(-7)}$$

- Comme nous pouvons le constater, l'approximation est réellement de bonne qualité. Les valeurs ne diffèrent qu'à la 3^{eme} décimale.

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.689	-	-	-
alcool	0.257	0.161	2.542	0.111
surpoids	0.692	0.138	25.192	0.000

Fig. 9.12. Hypertension - Résultats de la régression sans le profil n^o7

– De même pour la variable SURPOIDS

$$\hat{a}_2 - DFBETA_{2,7} = 0.584 - (-0.106) = 0.690 \approx 0.692 = \hat{a}_{2,(-7)}$$

– L'autre nouvelle importante est qu'**en retirant le profil n^o7, la variable ALCOOL devient non-significative à** 5%. Le rôle du profil n^o7 était considérable dans la régression incluant la totalité des données.

En conclusion de cette section, nous dirons que ces outils nous permettent de caractériser les profils d'individus en identifiant leur rôle dans la détermination des résultats de la régression. Ils s'avèrent particulièrement précieux lorsque nous souhaitons valider ou faire valider par un expert les résultats. Ils concourent à nous prémunir de *l'artefact statistique*, ce serpent masqué qui nous guette constamment lorsque nous traitons des données à l'aide de techniques avant tout numériques.

9.5 Sur-dispersion et sous-dispersion

Dans le modèle binomial, la variance de la variable dépendante est définie par $\sigma^2_Y = \pi(1-\pi)$. Lorsque nous travaillons sur des données individuelles, cette condition est naturellement satisfaite. Lorsque nous travaillons sur des données groupées, la variance devrait être $\sigma^2_{Y_m} = n_m\pi_m(1-\pi_m)$. Cette caractéristique peut ne pas être respectée pour plusieurs raisons [10] (page 89) : une variable explicative importante n'est pas présente ; un ou plusieurs groupes se démarquent fortement des autres ; les données sont organisées par blocs, autres que les covariate pattern ; ou tout simplement parce que le modèle ne convient pas.

On parle de sur-dispersion (resp. sous-dispersion) lorsque la variance σ_Y est plus grande (resp. plus petite) que prévue. La principale conséquence est une mauvaise estimation des écarts-type des paramètres [7] (page 90). Lorsqu'il y a sur-dispersion, les tests de Wald ont tendance à être exagérément significatifs car les écarts-type sont sous évalués (inversement pour la sous-dispersion). Le même commentaire est valable pour les tests de rapport de vraisemblance. Toute la statistique inférentielle est donc faussée.

Pour estimer la dispersion, on propose d'utiliser l'indicateur

$$\delta = \frac{D}{M-J-1} \tag{9.15}$$

où D est la déviance (section 9.3.2, on aurait pu utiliser la statistique de Pearson aussi) ; $M-J-1$ représente le degré de liberté.

Lorsque $\delta \approx 1$, tout va bien ; lorsque $\delta > 1$ (resp. $\delta < 1$), il y a sur-dispersion (resp. sous-dispersion).

Bonne nouvelle, il est possible de corriger les écarts-type estimés en introduisant le facteur δ comme suit :

$$\hat{\sigma}^*_{\hat{a}_j} = \hat{\sigma}_{\hat{a}_j} \times \sqrt{\delta} \tag{9.16}$$

Application aux données HYPERTENSION

delta	1.1782
Racine(delta)	1.0854

Attributes in the equation

Attribute	Coef.	Std-dev (originel)	Std-dev (corrigé)	Wald	Signif
constant	-1.673659	-		-	-
alcool	0.410675	0.1332	0.1446	8.0683	0.0045
surpoids	0.583889	0.1204	0.1307	19.9619	0.0000

Fig. 9.13. Hypertension - Correction de la sur-dispersion

Nous avions calculé la déviance pour les données HYPERTENSION plus haut, $D = 7.0690$, avec $M - J - 1 = 9 - 2 - 1 = 6$ degrés de liberté. Nous obtenons

$$\delta = \frac{7.0690}{6} = 1.1782$$

Il y a une légère sur-dispersion dans cette modélisation. Nous introduisons le facteur de correction $\sqrt{\delta} = \sqrt{1.178} = 1.0854$ dans l'estimation des écarts-type des coefficients et dans la définition des tests de significativité individuels (Figure 9.13). La correction des écarts-type est réelle. Mais la significativité des coefficients n'est pas modifiée par rapport au modèle originel (Figure 9.4).

10

Modifications de la règle d'affectation

10.1 Redressement pour les échantillons non représentatifs

A plusieurs reprises, nous avons évoqué le schéma d'échantillonnage rétrospectif dans ce document. A juste titre, la pratique est fréquente. On parle aussi de "données cas-témoin".

Explicitons l'idée. Souvent dans les études réelles, les positifs sont rares, voire très rares. Plutôt que d'utiliser un échantillonnage simple (représentatif) au risque de n'avoir que trop peu d'observations positives dans le fichier de données, on préfère souvent procéder différemment. On fixe le nombre d'observations n_+ positives à obtenir et on tire aléatoirement dans ce groupe ; on fait de même chez les négatifs pour avoir n_- individus. Souvent les effectifs sont sciemment équilibrés c.-à-d. $n_+ = n_-$, mais ce n'est pas une obligation.

Ca c'est la théorie. Dans la pratique, les positifs sont tellement rares qu'on prend ce qui vient. Puis on procède effectivement par échantillonnage chez les négatifs.

De fait, la proportion $\frac{n_+}{n}$ ne reflète plus la "vraie" proportion p des positifs dans la population. On dit que l'échantillon n'est pas représentatif. On suppose que nous pouvons connaître p par d'autres moyens, ou tout du moins nous pouvons faire des hypothèses crédibles sur sa véritable valeur.

Plusieurs questions se posent lorsque nous lançons les calculs sur un échantillon non représentatif :

- Est-ce que nous pouvons retrouver les "vrais" coefficients que l'on aurait estimé si nous avions travaillé sur un échantillon représentatif ?
- Quelle est la nature des corrections à introduire pour produire la prédiction $\hat{y}$ de la classe d'appartenance d'un individu ?
- Quelle est la nature des corrections à introduire lors du calcul de sa probabilité a posteriori $\hat{\pi}$ d'être positif ?
- Dans quel cadre pouvons-nous utiliser tels quels les résultats de la régression sans introduire de correction ? Cette question est très importante car l'obtention de p peut parfois poser problème. Est-ce que nous sommes totalement démunis dans ce cas ?

Dans ce chapitre, nous privilégions l'approche analytique parce que la régression logistique s'y prête à merveille. Pour certaines méthodes supervisées, ce n'est pas possible. On doit alors se tourner vers les approches empiriques, plus génériques, et adaptées à tous les contextes, que le score soit mal calibré ou

pas (les probabilités sont agglutinées autour de certaines valeurs), qu'il corresponde à une probabilité ou non (un score peut prendre des valeurs en dehors de $[0;1]$, qu'importe s'il arrive à ordonner les individus selon leur propension à être positif) [1].

10.1.1 Données

Nous utiliserons des données simulées pour illustrer ce chapitre. Nous voulons prédire les valeurs d'une variable binaire Y en fonction de deux prédictives continues X_1 et X_2. Nous disposons de 3 fichiers :

1. Un fichier d'apprentissage non représentatif avec $n_+ = 30$ et $n_- = 40$ (ANR70). Nous l'utiliserons pour construire le modèle de prédiction.
2. Un premier fichier test non représentatif avec toujours 30 positifs et 40 négatifs (TNR70). Il nous servira à montrer comment calculer le taux d'erreur sur un échantillon non représentatif.
3. Un second fichier test représentatif avec 10.000 positifs et 50.000 négatifs (TR60K). On considérera que la vraie prévalence est $p = \frac{1}{6} = 0.1667$.

Dans les études réelles, nous disposons de ANR70, éventuellement de TNR70, jamais de TR60K.

10.1.2 Correction du logit pour les échantillons non représentatifs

Correction du logit via le taux de sondage

On note C le logit obtenu sur les données d'apprentissage non représentatives, C^* celui que l'on obtiendrait si on travaillait sur un échantillon représentatif. Ils sont liés par la relation suivante ([9], pages 205 à 210 ; [23], pages 431 à 434 ; [2], pages 67 et 68 ; [3], pages 79 à 82)

$$C^* = -\ln \frac{\tau_+}{\tau_-} + C \tag{10.1}$$

où τ_+ (resp. τ_-) est le taux de sondage chez les positifs (resp. négatifs).

Comment pouvons nous ramener cette expression à la prévalence p? Mettons qu'il y a N observations dans la population, dont N_+ positifs. La prévalence est $p = \frac{N_+}{N}$. Le taux de sondage $\tau_+ = \frac{n_+}{N_+}$ correspond à la proportion d'individus que l'on a extrait dans le groupe des positifs. Nous pouvons nous ramener à la prévalence avec

$$\tau_+ = \frac{n_+}{N_+} = \frac{n_+}{p \times N}$$

Nous voyons autrement le rapport des taux de sondage

1. Voir R. Rakotomalala, *Redressement - Affectation optimale dans le cadre du tirage rétrospectif - Approches analytiques et empiriques*, `http://eric.univ-lyon2.fr/~ricco/cours/slides/affectation_optimale_et_redressement.pdf`

$$\begin{aligned}\ln\frac{\tau_+}{\tau_-} &= \ln\frac{n_+/(p\times N)}{n_-/[(1-p)\times N]}\\ &= \ln\frac{n_+\times(1-p)}{n_-\times p}\\ &= \ln\frac{n_+}{n_-} - \ln\frac{p}{1-p}\end{aligned}$$

Nous l'introduisons dans l'expression 10.1

$$C^* = -\ln\frac{n_+}{n_-} + \ln\frac{p}{1-p} + C \tag{10.2}$$

Commentons tout cela :

- En partant des résultats fournis pas les logiciels sur les données non représentatives, il suffit de connaître la prévalence p pour produire les coefficients corrigés.
- **La correction porte uniquement sur la constante**, les coefficients associées aux variables explicatives ne sont pas modifiés. On peut aller plus loin même : toute l'inférence statistique qui porte sur ces coefficients est valable (intervalle de confiance, test de significativité), il en est de même en ce qui concerne les interprétations (odds-ratio). C'est un résultat très important.
- **Dans les contextes où le principal objectif est de classer les observations selon leur degré de positivité** (scoring, construction de la courbe ROC, etc.), **les résultats obtenus sur les données non représentatives peuvent être utilisées tels quels**, sans correction. En effet, que l'on corrige ou pas, les individus seront ordonnés de la même manière.
- Il est possible d'obtenir les probabilités a posteriori corrigées $\hat{\pi}^*$ avec des calculs simples.

Application aux données simulées

Nous avons lancé l'apprentissage sur les données ANR70. Tanagra nous fournit les coefficients de la régression (Figure 10.1). Les coefficients $\hat{a}_1 = -1.315429 = \hat{a}_1^*$ et $\hat{a}_2 = 1.047243 = \hat{a}_2^*$ ne nécessitent pas de modifications. En revanche, pour la constante, nous calculons :

$$\begin{aligned}\hat{a}_0^* &= \hat{a}_0 - \ln\frac{n_+}{n_-} + \ln\frac{p}{1-p}\\ &= 3.127522 - \ln\frac{30}{40} + \ln\frac{1/6}{5/6}\\ &= 1.805766\end{aligned}$$

Finalement l'équation du logit corrigé s'écrit

$$C^* = 1.805766 - 1.315429X_1 + 1.047243X_2$$

A titre de vérification, nous avons lancé la régression sur l'échantillon représentatif TR60K. Bien évidemment, nous n'obtiendrons pas exactement les mêmes coefficients à cause des fluctuations d'échantillonnage (ce serait même suspect), mais au moins nous aurons un ordre d'idées sur les différences. C'est tout à fait édifiant.

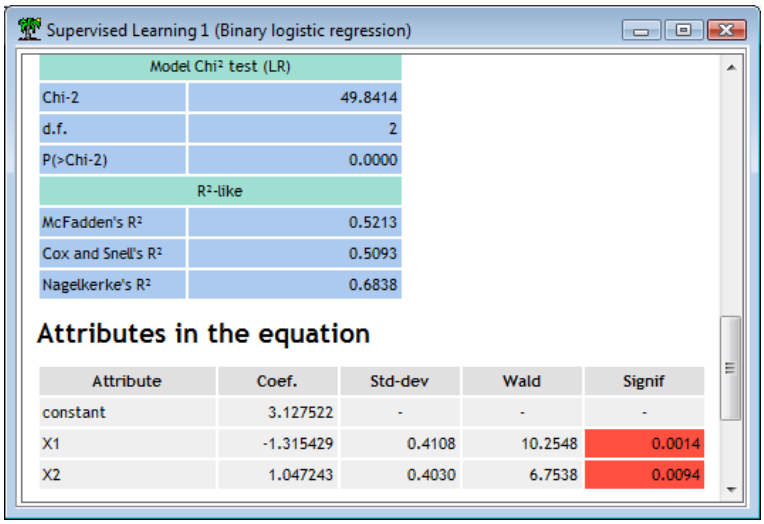


Fig. 10.1. Apprentissage - Échantillon non représentatif - $n = 70$ obs.

-	Ech. non représentatif		Ech. représentatif
Coef.	Non corrigé	Corrigé	-
a_0	3.127522	1.805766	1.359818
a_1	-1.315429	-	-1.267896
a_2	1.047243	-	1.16237

On notera principalement (1) que la constante calculée sur l'échantillon non représentatif est clairement surestimée ; (2) la correction va dans le bon sens ; (3) les coefficients associés aux variables sont (assez) similaires sans qu'il soit nécessaire d'introduire un ajustement.

Correction de la probabilité d'affectation

La correction tempère l'optimisme de la probabilité a posteriori $\hat{\pi}$ attribuée aux individus, dus à la sur-représentation des positifs dans le fichier d'apprentissage (par rapport à la prévalence réelle dans la population).

Prenons l'individu ($X_1 = 2.51$; $X_2 = 0.85$). Sans la correction, nous aurons

$$\hat{\pi} = \frac{1}{1 + e^{-(3.127522 - 1.315429 \times 2.51 + 1.047243 \times 0.85)}} = \frac{1}{1 + e^{-(0.7160)}} = 0.6717$$

Lorsque nous l'introduisons

$$\hat{\pi}^* = \frac{1}{1 + e^{-(1.805766 - 1.315429 \times 2.51 + 1.047243 \times 0.85)}} = \frac{1}{1 + e^{-(-0.6058)}} = 0.3530$$

L'ajustement n'est pas anodin. Sans, la probabilité d'être positif attribuée à l'individu serait exagérée.

On remarquera également que si l'on s'en tient au seuil usuel de 0.5, dans le 1^{er} cas, l'individu est classé positif, dans le 2^{nd}, négatif. Dans ce qui suit, nous allons étudier les implications de la correction sur la construction des prédictions $\hat{y}$.

10.1.3 Modification de la règle d'affectation pour le classement

Affectation basée sur le logit

La règle usuelle basée sur le logit est

$$\text{Si } C(\omega) > 0 \text{ Alors } \hat{y}(\omega) = + \text{ Sinon } \hat{y}(\omega) = -$$

Nous pouvons la transposer de deux manières équivalentes avec la correction du logit :

1. Nous utilisons le logit corrigé C^* et nous nous en tenons à la règle habituelle, soit

$$\text{Si } C^*(\omega) > 0 \text{ Alors } \hat{y}(\omega) = + \text{ Sinon } \hat{y}(\omega) = -$$

2. Nous utilisons le logit fourni par les logiciels sur les données non représentatives, mais nous ajustons le seuil d'affectation, soit

$$\text{Si } C(\omega) > \ln \frac{n_+}{n_-} - \ln \frac{p}{1-p} \text{ Alors } \hat{y}(\omega) = + \text{ Sinon } \hat{y}(\omega) = -$$

Quoiqu'il en soit, il faut utiliser une des deux procédures ci-dessus. Utiliser directement les sorties du logiciel, sans modifications, dégrade indûment les performances en classement comme nous allons le voir sur nos données. Nous allons appliquer les classifieurs corrigés et non corrigés sur le fichier test représentatif de 60.000 observations (TR60K).

Classement sans correction

C'est une erreur assez répandue. On utilise directement le classifieur proposé par le logiciel, sans se poser des questions sur le schéma d'échantillonnage. La règle de prédiction "brute" produite par Tanagra est la suivante (Figure 10.1)

$$\text{Si } 3.127522 - 1.315429X_1 + 1.047243X_2 > 0 \text{ alors } \hat{y} = + \text{ sinon } \hat{y} = -$$

Appliqué sur le fichier test de 60.000 observations, nous obtenons une matrice de confusion (Figure 10.2) avec un taux d'erreur de $\epsilon_{nc} = 0.1326$

Nombre de Y	Pred.NonCorr		
Y	pos	neg	Total
pos	8247	1753	10000
neg	6202	43798	50000
Total	14449	45551	60000

Err	0.1326
Précision	0.5708

Fig. 10.2. Évaluation du modèle **non corrigé** sur l'échantillon test représentatif (60.000 obs.)

Nombre de Y	Pred.Corri		
Y	pos	neg	Total
pos	6690	3310	10000
neg	1486	48514	50000
Total	8176	51824	60000

Err	0.0799
Rappel (Sensibilité)	0.6690
Précision	0.8182
TFP	0.0297
Spécificité	0.9703

Fig. 10.3. Évaluation du modèle **corrigé** sur l'échantillon test représentatif (60.000 obs.)

Classement avec ajustement du seuil d'affectation

Avec la même fonction logit, nous corrigeons le seuil. A la place de 0, nous utilisons

$$\ln \frac{n_+}{n_-} - \ln \frac{p}{1-p} = \ln \frac{30}{40} - \ln \frac{1}{5} = 1.3218$$

La règle de prédiction devient

$$\text{Si } 3.127522 - 1.315429X_1 + 1.047243X_2 > 1.3218 \text{ Alors } \hat{y} = + \text{ Sinon } \hat{y} = -$$

Appliqué au même fichier test, nous obtenons une autre matrice de confusion (Figure 10.3) avec un taux d'erreur autrement plus intéressant, $\epsilon_c = 0.0799$. Rien qu'en modifiant le seuil d'affectation, nous avons quasiment divisé par 2 (1.65 pour être précis) le taux d'erreur !

Comme le seuil est plus élevé, il y a moins d'individus classés positifs (8176 vs. 14449). L'amélioration porte essentiellement sur la précision (0.8182 vs. 0.5708).

Nous avons profité de la matrice de confusion pour calculer les autres indicateurs. Nous les utiliserons comme référence plus loin lorsqu'il s'agira de construire la matrice de confusion sur un échantillon non représentatif. Nous observons :

- Rappel = Sensibilité = $S_e = 0.6690$
- Précision = VPP = 0.8182
- Taux de faux positifs = TFP = 0.0297
- Spécificité = $S_p = 0.9703$

Affectation basée sur la probabilité a posteriori

Le même raisonnement peut être transposé à la règle d'affectation fondée sur la probabilité a posteriori. Nous pouvons la corriger comme nous l'avions montré plus haut (section 10.1.2) et utiliser le seuil usuel de 0.5. Nous pouvons aussi utiliser la probabilité fournie par le logiciel et modifier le seuil d'affectation. Voyons comment obtenir ce seuil en partant de la règle corrigée du logit

$$\begin{aligned} C &> \ln\frac{n_+}{n_-} - \ln\frac{p}{1-p} \\ -C &< \ln\frac{n_-}{n_+} + \ln\frac{p}{1-p} \\ e^{-C} &< \frac{n_+}{n_-} \times \frac{p}{1-p} \\ 1 + e^{-C} &< 1 + \frac{n_+}{n_-} \times \frac{p}{1-p} \\ \frac{1}{1+e^{-C}} &> \frac{1}{1+\frac{n_+}{n_-}\times\frac{p}{1-p}} \end{aligned}$$

La règle de prédiction devient

$$\text{Si } \hat{\pi} > \frac{1}{1+\frac{n_+}{n_-}\times\frac{p}{1-p}} \text{ Alors } \hat{y} = + \text{ Sinon } \hat{y} = -$$

De par sa construction, elle produit un classement totalement équivalent à celle basée sur le logit pour lequel nous avons ajusté le seuil d'affectation.

Remarque : le cas des échantillons équilibrés. Lorsque l'échantillon a été volontairement équilibré c.-à-d. $n_+ = n_-$, une pratique largement répandue, la règle est grandement simplifiée. Elle devient

$$\text{Si } \hat{\pi} > 1 - p \text{ Alors } \hat{y} = + \text{ Sinon } \hat{y} = -$$

10.1.4 Évaluation sur un échantillon non représentatif

Mesures dérivées de la matrice de confusion

Nous avons la chance de disposer d'un échantillon test représentatif. Nous pouvons évaluer les modèles sans se poser des questions sur la transposition des résultats dans la population.

Dans les études réelles, ce luxe est inaccessible. Le fichier test, s'il existe, est lui aussi non représentatif. Plusieurs questions se posent : est-ce que nous pouvons quand même élaborer la matrice de confusion dans ces conditions ? Y a-t-il des corrections à faire ? Sur tous les indicateurs ou sur quelques-uns seulement ?

A la première question, la réponse est oui. Rien ne nous empêche de construire la matrice de confusion. Nous disposons d'individus pour évaluer la prédiction, nous aurons tort de nous en priver. Après, selon les indicateurs, nous aurons besoin de la vraie prévalence p pour caler les estimations.

Nous appliquons le classifieur sur l'échantillon test non représentatif comportant 70 observations (TNR70). Nous obtenons une matrice de confusion, nous calculons directement les indicateurs habituels (Figure 10.4) :

Nombre de Y	Pred		
Y	pos	neg	Total
pos	19	11	30
neg	1	39	40
Total	20	50	70

Err	0.1714
Rappel (Sensibilité)	0.6333
Précison	0.9500
TFP	0.0250
Spécificité	0.9750

Fig. 10.4. Évaluation du modèle **corrigé** sur l'échantillon test non représentatif (70 obs.)

- Taux d'erreur = 0.1714
- Sensibilité = $S_e = 0.6333$
- Précision = VPP = 0.9500
- Taux de faux positifs = TFP = 0.0250
- Spécificité = $S_p = 0.9750$

Que faut-il en penser ? Nos références sont les valeurs obtenues sur l'échantillon représentatif (Figure 10.3), aux fluctuations d'échantillonnage près bien sûr. On se rend compte que certains indicateurs sont très loin du compte (taux d'erreur, précision), d'autres en revanche se rapprochent des "bonnes" valeurs (sensibilité, TFP, spécificité).

Et ce n'est pas étonnant. Tous les indicateurs correspondant à des profils lignes dans la matrice de confusion sont insensibles à la proportion des positifs dans le fichier test. Ainsi, la sensibilité, le taux de faux positifs et la spécificité peuvent être adoptés tels quels sans avoir à se poser des questions sur la représentativité de l'échantillon.

Les autres par contre (taux d'erreur, précision) doivent être corrigés en fonction de la prévalence $p = \frac{1}{6}$. Nous utilisons les expressions que nous avons mis en avant dans la section 2.1.2, lorsque nous ré-écrivons les différents indicateurs en fonction de la sensibilité et de la spécificité. Elles prennent toute leur saveur ici.

-	Ech. représentatif	Ech. non représentatif	
Indicateur	Pas d'ajustement	Sans ajustement	Avec Ajustement
Sensibilité (S_e)	0.6690	0.6333	-
TFP	0.0297	0.0250	-
Spécificité (S_p)	0.9703	0.9750	-
Taux d'erreur	0.0799	0.1714	$p(1-S_e)+(1-p)(1-S_p) = 0.0616$
Précison (VPP)	0.8182	0.9500	$\frac{p\times S_e}{p\times S_e+(1-p)\times(1-S_p)} = 0.8352$

Les valeurs obtenues sont autrement plus crédibles lorsque nous introduisons les ajustements pour le taux d'erreur et la précision.

Courbe ROC

La courbe ROC est un autre outil d'évaluation des classifieurs (section 2.5). Elle présente un double avantage dans le cadre des données non représentatives :

1. Elle repose uniquement sur l'ordonnancement des individus selon le score. Il n'est donc pas nécessaire de corriger le modèle avant de la construire. En effet, corriger la constante, c.-à-d. retrancher ou rajouter la même valeur pour tous les logit, ne modifiera en rien les positions relatives des individus.
2. Elle est construite à partir de la confrontation du taux de faux positifs ($1 - S_p$) et du taux de vrais positifs (S_e), deux profils lignes des matrices de confusions successives (pour chaque seuil d'affectation) utilisées pour produire les points qui la constituent. De fait, nous obtiendrons la même courbe ROC, qu'elle soit élaborée à partir d'un échantillon représentatif ou non. A aucun moment, nous n'avons besoin de la "vraie" prévalence p pour introduire une quelconque correction.

Ces deux propriétés font de la courbe ROC un outil extrêmement précieux (et populaire) dans les études réelles. Souvent, nous ne savons pas vraiment si le fichier manipulé est représentatif ou non. Obtenir des informations sur la vraie prévalence est parfois très difficile, voire impossible. La courbe ROC nous affranchit de ces contraintes.

Sur notre fichier de données, nous avons construit le modèle de prédiction sur les données d'apprentissage non représentatif (ANR70). Puis nous avons construit deux courbes ROC : l'une sur l'échantillon test non représentatif de 70 observations (TNR70) ; l'autre sur l'échantillon représentatif avec 60.000 observations (TR60K). Nous les avons placés dans le même repère (Figure 10.5).

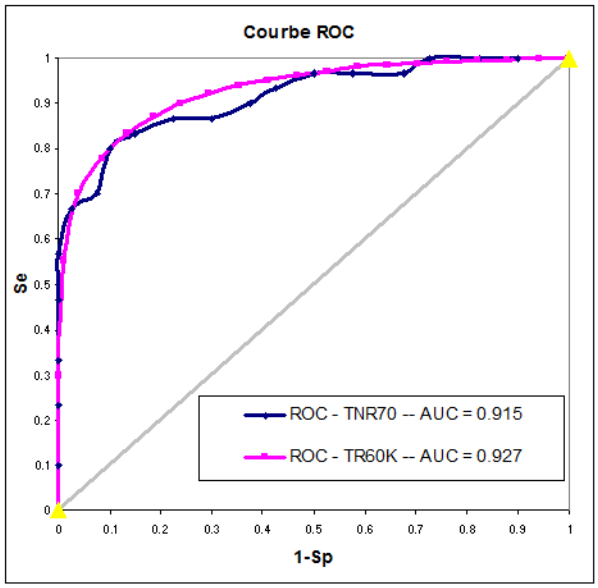


Fig. 10.5. Courbes ROC sur échantillon non-représentatif (TNR70) et représentatif (TR60K)

Les deux tracés sont très proches, ce qui accrédite l'idée avancée ci-dessus : quelle que soit la répartition des modalités de la variable dépendante dans le fichier de données, la courbe ROC reste imperturbable.

La courbe élaborée sur l'échantillon $TR60K$ est moins heurtée, mieux lissée, parce que les effectifs sont nettement plus élevés. Concernant l'aire sous la courbe, nous obtenons également des valeurs similaires (aux fluctuations d'échantillonnage près) avec AUC(TNR70) = 0.915 et AUC(TR60K) = 0.927.

10.2 Prise en compte des coûts de mauvais classement

10.2.1 Définir les coûts de mauvaise affectation

Tout le monde s'accorde à dire que l'intégration des coûts de mauvais classement est un aspect incontournable de la pratique du Data Mining. Diagnostiquer une maladie chez un patient sain ne produit pas les mêmes conséquences que de prédire la bonne santé chez un individu malade. Dans le premier cas, le patient sera soigné à tort, ou peut être demandera-t-on des analyses supplémentaires superflues ; dans le second cas, il ne sera pas soigné, au risque de voir son état se détériorer de manière irrémédiable. Pourtant, malgré son importance, le sujet est peu abordé, tant du point de vue théorique c.-à-d. comment intégrer les coûts dans l'évaluation des modèles (facile) et dans leur construction (un peu moins facile), que du point de vue pratique c.-à-d. comment les mettre en oeuvre dans les logiciels.

Une matrice de coûts de mauvais classement se présente sous la forme d'une matrice $c(k,l)$ avec, en ligne les valeurs observées de la variable à prédire, en colonne les valeurs prédites par les modèles : $c(k,l)$ est le coût associé à la prédiction $\hat{Y}(\omega) = y_l$ alors que la valeur observée est $Y(\omega) = y_k$. Usuellement, nous avons $c(k,l) > 0$ si $k \neq l$, mal classer induit un coût ; et $c(k,k) = 0$, bien classer ne coûte rien.

Mais cette première écriture est un peu restrictive. En réalité, bien classer entraîne souvent un gain, soit un coût négatif, nous écrirons plutôt $c(k,k) \leq 0$. Dans le domaine du crédit scoring par exemple, prédire la fiabilité d'un client qui s'avère l'être effectivement rapporte de l'argent à la banque : le montant des intérêts.

Quantifier les conséquences d'un bon ou mauvais classement appartient aux experts du domaine. Il n'est pas question pour nous statisticiens de s'immiscer dans cette phase. En revanche, nous devons la prendre en compte lors du processus d'extraction de connaissances.

L'intégration des coûts lors de l'évaluation ne pose pas de problèmes particuliers. Il s'agit de faire le produit terme à terme entre la matrice de coût et la matrice de confusion. Nous obtenons ainsi un " coût moyen de mauvais classement " (ou d'un gain moyen si nous multiplions le résultat par -1). Son interprétation n'est pas très aisée. Il vaut surtout pour comparer des modèles concurrents.

La prise en compte des coûts lors de l'élaboration du modèle de classement est moins connue. Nous étudierons une approche très simple, mais déjà efficace. Il s'agit d'estimer les paramètres $\hat{a}$ sans tenir compte des coûts, puis d'utiliser une règle d'affectation qui minimise le coût moyen lors du classement de nouveaux individus. Concrètement, on s'appuie sur les probabilités conditionnelles fournies par le modèle pour calculer la perte associée à chaque décision. On choisit la décision qui minimise la perte espérée. C'est une généralisation de la règle de classement classique qui cherche à minimiser le taux d'erreur. Le principal intérêt de cette correction par les coûts est que nous pouvons exploiter, sans modifications spécifiques, les résultats fournis par les logiciels courants.

Il existe d'autres techniques, plus ou moins sophistiquées, décrites dans la littérature. Nous citerons, entre autres [2] :

- L'intégration des coûts de mauvais classement dans le processus d'apprentissage. Peu de méthodes permettent cela. Nous citerons en particulier les arbres de décision qui peuvent utiliser explicitement la matrice de coûts lors du post-élagage.
- L'utilisation de systèmes de pondération d'individus. L'idée est de donner plus de poids aux individus "coûteux" de manière à orienter en priorité l'apprentissage vers leur bon classement.
- L'utilisation des schémas d'agrégation de modèles, basés sur des ré échantillonnages plus ou moins adaptatifs (bagging ou boosting). Même si elles sont pour la plupart performantes, elles présentent un inconvénient majeur : nous disposons d'une série de modèles, l'interprétation des résultats devient difficile, voire impossible.
- Ré-étiqueter les individus c.-à-d. modifier artificiellement les valeurs de la variable dépendante, toujours de manière à orienter l'apprentissage vers les individus à problème, ceux qui vont induire un coût élevé s'ils sont mal classés (ex. la méthode Metacost de Domingos 1999).

Pour intéressantes qu'elles soient, ces méthodes sont peu répandues, peu présentes dans les logiciels usuels [3]. Nous nous en tiendrons donc à la méthode très simple de correction de la règle d'affectation dans ce document.

10.2.2 Intégrer les coûts lors de l'évaluation

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$
$+$	α	β
$-$	γ	δ

Tableau 10.1. Matrice de coûts de mauvais classement pour un problème binaire

Dans le cadre de la prédiction binaire, nous allons simplifier l'écriture de la matrice de coûts (Tableau 10.1). Nous devons la prendre en compte lors de l'évaluation des classifieurs, en la mariant au mieux (et non pour le pire) avec la matrice de confusion. Le taux d'erreur qui ignore la structure de coûts n'est plus adapté dans ce contexte.

Le coût moyen de mauvaise affectation pour un modèle M est défini de la manière suivante :

$$\zeta(M) = \frac{1}{n}(a \times \alpha + b \times \beta + c \times \gamma + d \times \delta) \tag{10.3}$$

Son interprétation n'est pas toujours facile, d'autant que les coûts sont exprimés dans des unités imprécises (*qui oserait - tout du moins ouvertement - exprimer en euros le fait de classer un patient*

2. A propos des différentes méthodes, voir R. Rakotomalala, *Intégrer les coûts de mauvais classement en apprentissage supervisé*, `http://eric.univ-lyon2.fr/~ricco/cours/slides/couts_en_apprentissage_supervise.pdf`

3. Pour la prise en compte des coûts dans les logiciels R, Tanagra et Weka, voir `http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Cost_Sensitive_Learning.pdf`

sain chez les malades ?). Quoiqu'il en soit, cet indicateur intègre bien la structure de coûts, il permet de comparer les performances des différents modèles. C'est déjà pas mal.

Un exemple - L'attrition

Nous sommes dans un problème de détection automatique de clients faisant défection pour un fournisseur d'accès internet. On parle d'*attrition* (en anglais "churn")[4]. Les responsables de l'entreprise proposent d'utiliser la matrice de coûts suivante

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$
$+$	-1	10
$-$	5	0

- Laisser passer un client à la concurrence coûte $c(+,-) = 10$;
- Aller tarabuster, et lui donner de mauvaises idées, un client qui ne pensait pas à partir, $c(-,+) = 5$;
- Soigner à juste titre un client sur le point de partir "coûte" $c(+,+) = -1$;
- Laisser tranquille le gars bien installé, $c(-,-) = 0$.

Encore une fois, fixer les coûts est l'affaire des experts. Il n'appartient pas au data miner de se lancer dans des élucubrations sur le coût de telle ou telle configuration. Dans la pratique, on teste d'ailleurs différents scénarios de coûts.

Deux modèles de prédiction (M_1 et M_2) sont en concurrence. Nous voulons savoir quel est le meilleur. Nous disposons des matrices de confusion (Figure 10.6).

Modèle 1

Obs. vs. Pred.	Pred +	Pred -	Tot.
+	40	10	50
-	20	30	50
Tot.	60	40	100

Taux d'erreur 0.30

Modèle.2

Obs. vs. Pred.	Pred +	Pred -	Tot.
+	20	30	50
-	0	50	50
Tot.	20	80	100

Taux d'erreur 0.30

Matrice de coûts

Obs. vs. Pred.	Pred +	Pred -
+	-1	10
-	5	0

Cout.Moyen 1.6

Cout.Moyen 2.8

Fig. 10.6. Comparaison de deux classifieurs sans et avec prise en compte de la matrice de coûts

Si l'on s'en tient au taux d'erreur, les deux modèles sont équivalents, en effet

4. Bon, même si j'en meurs d'envie, je ne dirai pas à cause de qui j'ai été privé d'ADSL pendant 2 mois cet été, c'est comme si on me privait d'air... un vrai retour au moyen âge. Moralité, je suis allé à la concurrence bien sûr. Voilà un bel exemple d'attrition. Comme quoi le data mining fait partie intégrante de notre vie de tous les jours.

$$\epsilon(M_1) = \frac{20+10}{100} = 0.3$$
$$\epsilon(M_2) = \frac{0+30}{100} = 0.3$$

Mais lorsque l'on prend en compte la structure de coûts, le modèle M_1 se démarque nettement

$$\zeta(M_1) = \frac{1}{100}(40 \times (-1) + 10 \times 10 + 20 \times 5 + 30 \times 0) = 1.6$$
$$\zeta(M_2) = \frac{1}{100}(20 \times (-1) + 30 \times 10 + 0 \times 5 + 50 \times 0) = 2.8$$

Et ce n'est pas étonnant : il se trompe peu là où c'est le plus coûteux $c(+,-) = 10$; il classe à bon escient là où c'est le plus avantageux $c(+,+) = -1$. Avec cette structure de coûts, nous avons tout intérêt à choisir le modèle M_1 qui est nettement plus performant.

Le taux d'erreur est un cas particulier

A bien y regarder, on se rend compte que le taux d'erreur est un coût moyen de mauvais classement avec une matrice de coûts symétrique et unitaire (Tableau 10.2).

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$
$+$	0	1
$-$	1	0

Tableau 10.2. Matrice de coûs symétrique et unitaire

Reprenons l'exemple du modèle M_1 de la section précédente, nous obtenons

$$\zeta(M_1) = \frac{1}{100}(40 \times 0 + 10 \times 1 + 20 \times 1 + 30 \times 0) = 0.3 = \epsilon(M_1)$$

Le coût moyen de mauvais classement est une "vraie" généralisation. Il y a donc des hypothèses implicites dans le taux d'erreur : bien classer ne coûte rien, mais ne gagne rien non plus ; mal classer coûte 1, quelle que soit l'erreur.

10.2.3 Intégrer les coûts lors du classement

Maintenant que nous savons évaluer les classifieurs en intégrant la structure de coût, il reste un problème épineux : comment orienter l'apprentissage pour qu'il en tienne compte. L'objectif est de produire un classifieur qui minimisera, non plus le taux d'erreur, mais plutôt le coût moyen de mauvaise affectation. Comme nous le disions plus haut, il existe moult stratégies pour ce faire. Nous choisissons de présenter une approche très simple dans cette section. Nous procédons en deux temps :

1. Nous estimons les paramètres du logit en utilisant la régression logistique usuelle.
2. Lors du classement d'un nouvel individu ω, nous nous appuyons sur la probabilité estimée $\hat{\pi}(\omega)$ et la matrice $c(k,l)$ pour lui assigner la classe y_{l^*} qui minimise les coûts.

Cette stratégie est possible parce que la régression logistique fournit une estimation fiable (bien calibrée) de $\hat{\pi}(\omega)$. Ce n'est pas le cas de certaines les méthodes supervisées (ex. support vector machine, bayesien naïf).

Nous pouvons utiliser les logiciels habituels de régression logistique. C'est un avantage non négligeable. Nous verrons que malgré sa simplicité, elle est performante. Le classifieur ainsi défini se démarque nettement du modèle de référence, celui qui ignore les coûts.

Pour modifier la procédure d'affectation de la régression logistique, il nous faut revenir sur les fondamentaux et intégrer la structure de coûts dans la règle de Bayes décrite précédemment (section 1.1.3). Nous écrivons [3] (page 4)

$$y_{l^*} = \arg\min_l \zeta(y_l) \tag{10.4}$$

où $\zeta(y_l)$ est la perte moyenne associée à la prédiction $\hat{Y}(\omega) = y_l$,

$$\zeta(y_l) = \sum_{k=1}^{K} P(Y = y_k/X) \times c(k,l) \tag{10.5}$$

L'idée est finalement très sensée : nous choisissons la prédiction la moins coûteuse en moyenne.

Pour un classement binaire, la règle est simplifiée

$$\text{Si } \zeta(+) < \zeta(-) \text{ alors } Y = + \text{ sinon } Y = - \tag{10.6}$$

Remarque : règle d'affectation pour une matrice de coût symétrique et unitaire

Si nous utilisons une matrice de coût symétrique et unitaire (Tableau 10.2), le coût associé à une prédiction s'écrit

$$\begin{aligned}\zeta(y_l) &= \sum_{k\neq l} P(Y = y_k/X) \times 1 \\ &= \sum_{k\neq l} P(Y = y_k/X) \\ &= 1 - P(Y = y_l)\end{aligned}$$

Nous retrouvons une règle d'affectation que nous connaissons bien

$$\begin{aligned}y_{l^*} &= \arg\min_l [1 - P(Y = y_l/X)] \\ &= \arg\max_l P(Y = y_l/X)\end{aligned}$$

10.2.4 Classement d'un individu

Pour la matrice de coût de l'exemple "attrition" (section 10.2.2), l'individu ω présente les probabilités a posteriori $[\hat{P}(Y = +/X) = \hat{\pi} = 0.4\,;\ \hat{P}(Y = -/X) = 1 - \hat{\pi} = 0.6]$. Avec la règle usuelle (maximisation de la probabilité a posteriori), sans tenir compte des coûts, nous lui assignerons l'étiquette "-".

Voyons ce qu'il en est si nous intégrons le coût dans la prise de décision. Nous calculons la perte moyenne relative à chaque modalité

$$\zeta(+) = 0.4 \times (-1) + 0.6 \times 5 = 2.6$$
$$\zeta(-) = 0.4 \times 10 + 0.6 \times 0 = 4$$

La conclusion la moins coûteuse consiste à attribuer l'étiquette "+" finalement. La décision est inversée par rapport à la précédente.

10.2.5 Traitement du fichier COEUR

Reprenons notre fichier COEUR. Nous utilisons une matrice de coût de mauvaise d'affectation dont la structure est la suivante

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$
$+$	-5	10
$-$	2	0

Le gain est élevé lorsque nous soignons une personne malade $[c(+,+) = -5]$; la perte est particulièrement importante lorsque nous ne la diagnostiquons pas la maladie chez une personne en mauvaise santé $[c(+,-) = 10]$.

Performances du modèle non corrigé M

Dans un premier temps, nous prenons le modèle avec la règle d'affectation non corrigée. Après estimation des coefficients, construction de la colonne de $\hat{\pi}$ et de la colonne prédiction $\hat{y}$ (Figure 2.1), nous avions produit une matrice de confusion (Tableau 10.3), avec un taux d'erreur de $\epsilon = 0.20$.

En appliquant la matrice de coût, nous constatons un coût moyen de mauvais classement égal à

$$\zeta(M) = \frac{1}{20}(3 \times (-5) + 3 \times 10 + 1 \times 2 + 13 \times 0) = 0.85$$

Si l'on corrige la règle d'affectation lors de la prédiction, nous devrions obtenir de meilleures performances c.-à-d. un coût moyen plus faible. Vérifions cela.

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	Total
$+$	3	3	6
$-$	1	13	14
Total	4	16	20

Tableau 10.3. COEUR - Matrice de confusion - Modèle **sans** correction de la règle d'affectation

a0	a1	a2	a3
14.4937	-0.1256	-0.0636	1.7790

	X1	X2	X3							
	age	taux_max	angine	coeur	y	C(X)	PI	C(+)	C(-)	Prédiction
1	50	126	1	presence	1	1.9825	0.8789	-4.1526	8.7895	presence
2	49	126	0	presence	1	0.3291	0.5815	-2.0708	5.8155	presence
3	46	144	0	presence	1	-0.4381	0.3922	-0.7454	3.9220	presence
4	49	139	0	presence	1	-0.4972	0.3782	-0.6475	3.7821	presence
5	62	154	1	presence	1	-1.3048	0.2134	0.5065	2.1336	presence
6	35	156	1	presence	1	1.9601	0.8765	-4.1358	8.7655	presence
7	67	160	0	absence	0	-4.0933	0.0164	1.8851	0.1641	absence
8	65	140	0	absence	0	-2.5708	0.0710	1.5027	0.7104	absence
9	47	143	0	absence	0	-0.5001	0.3775	-0.6425	3.7751	presence
10	58	165	0	absence	0	-3.2804	0.0362	1.7463	0.3625	absence
11	57	115	1	absence	0	1.8022	0.8584	-4.0090	8.5842	presence
12	59	145	0	absence	0	-2.1348	0.1058	1.2597	1.0576	absence
13	44	175	0	absence	0	-2.1572	0.1037	1.2744	1.0366	absence
14	41	153	0	absence	0	-0.3820	0.4057	-0.8396	4.0565	presence
15	54	152	0	absence	0	-1.9516	0.1244	1.1293	1.2438	presence
16	52	169	0	absence	0	-2.7809	0.0584	1.5914	0.5837	absence
17	57	168	1	absence	0	-1.5665	0.1727	0.7910	1.7272	presence
18	50	158	0	absence	0	-1.8304	0.1382	1.0327	1.3819	presence
19	44	170	0	absence	0	-1.8394	0.1371	1.0401	1.3712	presence
20	49	171	0	absence	0	-2.5311	0.0737	1.4841	0.7371	absence

Matrice de coût

	presence	absence
presence	-5	10
absence	2	0

Matrice de confusion

Nombre de	Prédiction		
coeur	presence	absence	Total
presence	6		6
absence	7	7	14
Total	13	7	20

Taux d'erreu	0.35
Cout Moyen	-0.8

Fig. 10.7. COEUR - Prédiction et matrice de confusion - Modèle corrigé

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	Total
$+$	6	0	6
$-$	7	7	14
Total	13	7	20

Tableau 10.4. COEUR - Matrice de confusion - Modèle **avec** correction de la règle d'affectation

Performances du modèle corrigé Mc

La structure de la feuille Excel est exactement la même jusqu'à la construction des probabilités estimées $\hat{\pi}$ (Figure 10.7). Nous insérons deux colonnes supplémentaires, $\zeta(+)$ et $\zeta(-)$. La colonne $\hat{y}$ va tenir compte des coûts en utilisant la nouvelle règle de classement (Équation 10.6).

Voici le détail des opérations pour le 1^{er} individu :

- Ses coordonnées sont $X(1) = (\text{constante} = 1\,;\ \text{age} = 50\,;\ \text{taux max} = 126\,;\ \text{angine} = 1)$.
- Nous produisons son logit

$$C(X(1)) = 14.4937 - 0.1256 \times 50 - 0.0636 \times 126 + 1.7790 \times 1 = 1.9825$$

- La probabilité a posteriori estimée est

$$\hat{\pi}(1) = \frac{1}{1 + e^{-(1.9825)}} = 0.8789$$

– Calculons les pertes associées à chaque prédiction

$$\zeta(+) = 0.8789 \times (-5) + (1 - 0.8789) \times 2 = -4.1526$$
$$\zeta(-) = 0.8789 \times 10 + (1 - 0.8789) \times 0 = 8.7895$$

– La prédiction $\hat{y} = +$ est celle qui minimise la perte, nous assignons au premier individu l'étiquette positive (présence).

Nous faisons de même pour les autres individus de la base. Nous obtenons une matrice de confusion (Tableau 10.4) avec un taux d'erreur de 0.35. Mais peu importe cet indicateur en réalité, il faut évaluer le classifieur avec la structure de coûts. Le coût moyen de mauvais classement est égal à

$$\zeta(Mc) = \frac{1}{20}(6 \times (-5) + 0 \times 10 + 7 \times 2 + 7 \times 0) = -0.8$$

Le modèle Mc est nettement meilleur que M [$\zeta(Mc) = -0.8$ vs. $\zeta(M) = 0.85$]. Pourtant ils s'appuient sur les mêmes paramètres estimés $\hat{a}_j$. Conclusion : la règle d'affectation qui tient compte des coûts permet d'orienter la prédiction dans le sens de la réduction du coût moyen de mauvais classement. Les calculs supplémentaires demandés sont négligeables face à l'amélioration spectaculaire des performances.

11

Quelques éléments supplémentaires

11.1 L'écueil de la discrimination parfaite

L'étude attentive de la matrice hessienne (Equation 1.12) nous éclaire sur un des aspects sombres de la régression logistique : le plantage des logiciels lorsque la discrimination est parfaite c.-à-d. lorsque les positifs et les négatifs sont parfaitement séparables dans l'espace de représentation.

A priori, cette situation est idyllique. Un hyperplan séparateur permet de discriminer parfaitement les classes. L'analyse discriminante linéaire se promène littéralement dans cette configuration. Pas la régression logistique. La raison n'est pas dans les fondements de la méthode elle-même, nous devrions obtenir normalement une déviance égale à 0 (ou une vraisemblance égale à 1), mais plutôt dans la stratégie d'optimisation de la log-vraisemblance. L'algorithme de Newton-Raphson (Équation 1.10) a besoin de calculer la matrice hessienne, puis de l'inverser. Or, lorsque la discrimination est parfaite, pour tout individu ω, nous avons soit $\pi(\omega) = 1$, soit $\pi(\omega) = 0$. Tous les termes de la matrice H sont nuls et, de fait, elle n'est pas inversible. Le logiciel plante ! Certains arrivent à mettre en place des astuces pour s'en prémunir, d'autres non. Il n'en reste pas moins que toute la partie "statistique inférentielle" (tests, intervalle de confiance) n'est plus réalisable.

Prenons l'exemple des données de Tomassone et al. ([24], Figure 1, page 30). Manifestement les classes sont parfaitement discernables (Figure 11.1). Nous pouvons tracer une droite séparant les positifs des négatifs. Il apparaît clairement également que c'est la combinaison des 2 descripteurs (X_1 et X_2) qui permet de réaliser cette discrimination : un classifieur basé sur X_1 seul (resp. X_2 seul) ne pourrait pas réaliser une séparation parfaite.

Or, que nous disent nos logiciels préférés ? Tanagra, tout comme R, trouve une solution optimale sans planter. Mais pas le même vecteur $\hat{a}$! C'est normal, il y a une infinité de solutions ! La déviance du modèle est égale à $D_M = -2LL = 0$ (Figure 11.2, A). Le taux d'erreur en resubstitution, si nous le calculons, serait égal à 0. Le modèle est sans erreur. R annonce néanmoins qu'il y a eu des problèmes lors de l'optimisation, ce qui devrait nous inciter à la prudence.

Et en effet, le bilan est très décevant lorsque nous nous penchons sur la contribution des variables (Figure 11.2, B). Tanagra refuse de calculer les écarts-type des coefficients. R propose des valeurs fantaisistes. Les deux s'accordent à annoncer qu'aucune des deux variables n'est pertinente dans la discrimination avec des probabilités critiques (p-value) égales à 1. Or, nous savons pertinemment à la lumière du nuage de

Y	X1	X2
positive	0	3
positive	2	1
positive	4	5
positive	6	9
positive	8	7
negative	5	2
negative	7	0
negative	9	4
negative	11	8
negative	13	6

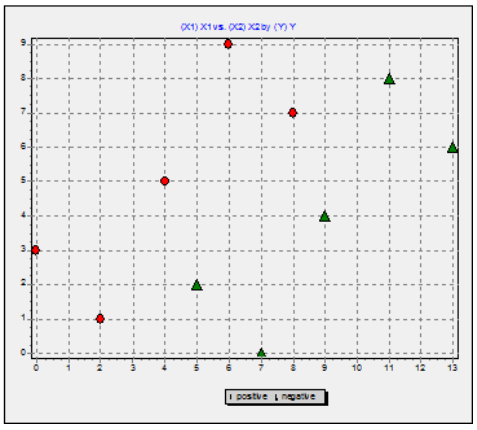


Fig. 11.1. Données Tomassone et al. - Discrimination parfaite

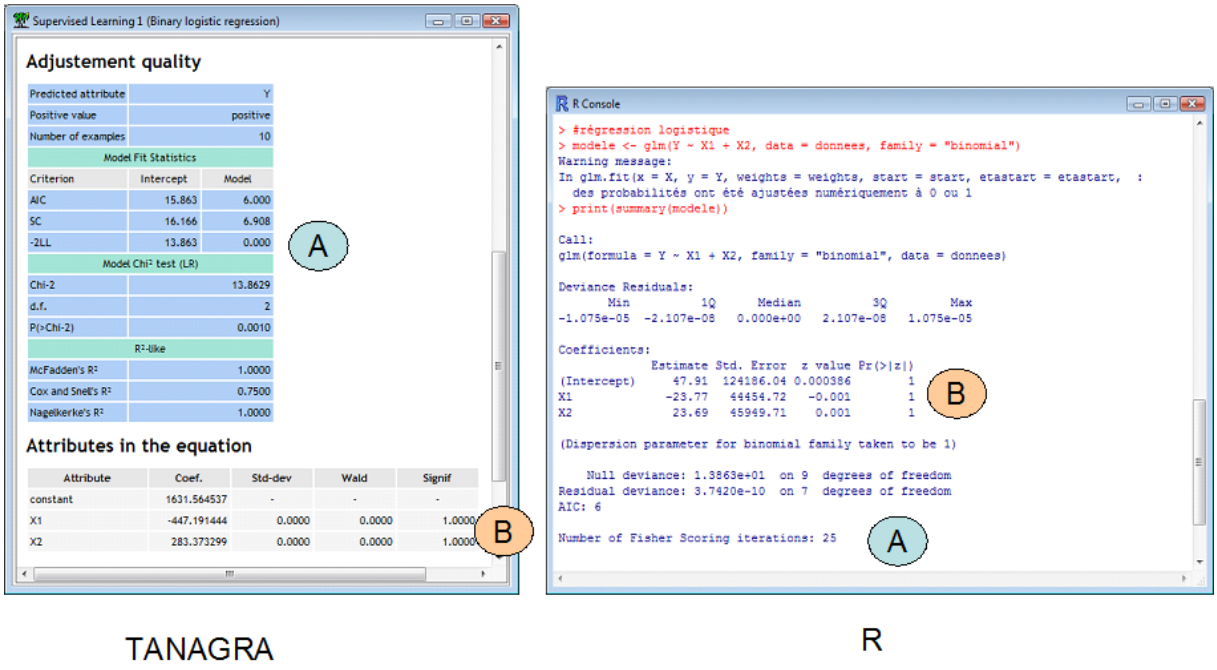


Fig. 11.2. Données Tomassone et al. - Résultats de la régression logistique

points (Figure 11.1) que c'est faux : les deux variables prises ensembles sont capables de produire un classifieur parfait.

Dans le même contexte, l'analyse discriminante produit les résultats adéquats : la discrimination est excellente, les deux variables y contribuent (Figure 11.3, A et B).

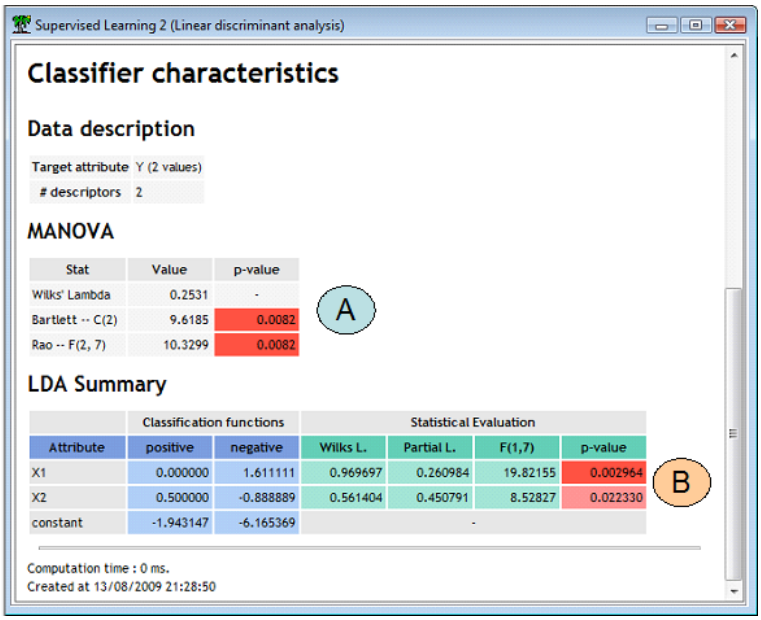


Fig. 11.3. Données Tomassone et al. - Résultats de l'analyse discriminante linéaire

11.2 Estimation des coefficients par les MCO pondérés

11.2.1 Quel intérêt ?

Il est possible de retrouver les résultats de la régression logistique à l'aide de la régression linéaire multiple. Il suffit de transformer la variable dépendante et de pondérer les individus [9] (pages 128 à 130). C'est plutôt une bonne nouvelle. En effet les programmes de régression linéaire sont largement répandus (il y en a par exemple dans le tableur Excel). Ils sont nettement plus performants en temps de traitement, un atout important lorsque nous traitons de grandes bases de données avec des centaines de milliers d'observations.

Mais ce n'est pas aussi simple. En effet, pour réaliser les calculs, l'algorithme des moindres carrés a besoin des $\hat{\pi}(\omega)$... fournis par la régression logistique. Dès lors une question se pose : pourquoi s'enquiquiner à estimer les paramètres à l'aide de la régression linéaire alors qu'il est nécessaire de passer par une étape préalable de calculs via la régression logistique ?

En temps normal, aucun effectivement. Les calculs supplémentaires ne sont absolument pas justifiés. En revanche, dans un contexte de sélection de variables, s'appuyer sur l'équivalence s'avère très avantageux :

1. nous estimons les paramètres de la régression logistique en incluant toutes les variables explicatives ;
2. puis, nous utilisons les résultats pour produire les probabilités prédites $\hat{\pi}(\omega)$ pour chaque observation ;
3. cette information acquise, nous pouvons l'introduire dans l'algorithme des moindres carrés ;

4. et utiliser les stratégies de sélection de variables propres à la régression linéaire, on cite souvent la méthode "branch and bound" [1] de Furnival et Wilson (1974) qui est capable de produire les k (paramétrable) "meilleurs" modèles à une variable, les k meilleurs modèles à 2 variables, etc. On utilise par la suite des critères tels que le C_p de Mallows pour choisir le bon modèle parmi ces candidats [9] (pages 131 à 135) [2].
5. Le meilleur sous-ensemble de variables ainsi détecté sera présenté à la régression logistique qui produira le modèle définitif.

Dans ce qui suit, nous décrivons les formules qui permettent d'obtenir les estimations $\hat{a}$ à partir des moindres carrés. Nous détaillons tout cela sur un exemple numérique en utilisant la fonction, on ne peut plus standard, DROITEREG d'Excel.

11.2.2 Équivalence entre la régression logistique et la régression linéaire

L'estimation $\hat{a}$ des paramètres de la fonction LOGIT peut être obtenue par la formule des moindres carrés généralisés. Faisons un petit retour sur la régression linéaire multiple avec une variable dépendante $Y \in \{0,1\}$, l'équation s'écrit

$$Y = a_0 + a_1 X_1 + \cdots + a_J X_J + \varepsilon \tag{11.1}$$

où ε est l'erreur du modèle. Voyons quelques propriétés :

$$\begin{aligned}
E[Y(\omega)] &= \pi(\omega) \qquad \text{Moyenne de Y} \equiv \text{probabilité de Y} \\
E[\varepsilon(\omega)] &= 0 \qquad \text{Par hypothèse des MCO} \\
V(\varepsilon(\omega)) &= V(Y(\omega)) \qquad \text{Par hypothèse, les } X \text{ sont non aléatoires, indépendants de } \varepsilon \\
&= E\{[Y(\omega) - E(Y(\omega))]^2\} \\
&= E(Y(\omega)^2) - E(Y(\omega))^2 \qquad Y^2 = Y \text{ puisque défini dans } \{0,1\} \\
&= \pi(\omega) - \pi(\omega)^2 \qquad \text{Il y a hétéroscédasticité}
\end{aligned}$$

Pour obtenir les bonnes estimations, nous devrons donc pondérer chaque individu par

$$\nu(\omega) = \frac{1}{\sqrt{\hat{\pi}(\omega) - \hat{\pi}(\omega)^2}} = \frac{1}{\sqrt{\hat{\pi}(\omega)(1 - \hat{\pi}(\omega))}} \tag{11.2}$$

Concernant la variable dépendante, pour qu'il y ait équivalence entre la régression logistique et la régression linéaire, nous devons utiliser la transformation suivante [9] (page 130)

1. Méhode de séparation et d'évaluation, voir par exemple D. de Werra, T. Liebling, J.F. Hêche, *Recherche opérationnelle pour ingénieurs - I*, Presses polytechniques et universitaires romandes, 2003 ; pages 340 à 346.
2. Pour une présentation plus détaillé du critère C_p de Mallows dans le cadre de la régression linéaire, voir Y. Dodge, V. Rousson, *Analyse de régression appliquée*, Dunod, 2004 ; pages 147 à 149.

$$z(\omega) = \ln\left(\frac{\hat{\pi}(\omega)}{1-\hat{\pi}(\omega)}\right) + \frac{y(\omega)-\hat{\pi}(\omega)}{\hat{\pi}(\omega)(1-\hat{\pi}(\omega))} = \hat{c}(x(\omega)) + \frac{y(\omega)-\hat{\pi}(\omega)}{\hat{\pi}(\omega)(1-\hat{\pi}(\omega))} \tag{11.3}$$

En passant à une notation matricielle, nous retrouvons l'expression de l'estimateur des moindres carrés généralisés $\hat{a}_{MCG}$ qui produit les mêmes paramètres que l'estimateur $\hat{a}_{MMV}$ du maximum de vraisemblance de la régression logistique ([11], pages 109 et 110)

$$\hat{a}_{MMV} = \hat{a}_{MCG} = (X'VX)^{-1}X'Vz \tag{11.4}$$

où

- X est la matrice des données, avec la constante en première colonne ;
- V est la matrice diagonale des $\hat{\pi}(\omega)(1-\hat{\pi}(\omega))$;
- $z = X\hat{a} + V^{-1}r$ est la transformation de la variable dépendante ;
- $r = y - \hat{\pi}$ est le vecteur des résidus.

A priori, si l'on veut mettre en oeuvre la méthode, il faudrait que l'on construise la variable z puis que l'on dispose d'un logiciel capable de prendre en compte le poids ν. Ils ne sont pas nombreux. En pratique, il s'avère que nous pouvons utiliser les logiciels usuels qui implémentent les moindres carrés ordinaires (MCO) en estimant les paramètres de la régression [3]

$$\frac{z}{\nu} = a_0\frac{1}{\nu} + a_1\frac{X_1}{\nu} + \cdots + a_J\frac{X_J}{\nu} \tag{11.5}$$

Enfin, la formulation ci-dessus nous fournit bien les estimations $\hat{a}$. Mais il faut introduire une autre correction pour obtenir une estimation correcte des écarts-type. On définit s^2 la variance estimée des résidus de la manière suivante [4]

$$s^2 = \frac{1}{n-J-1}\sum_{\omega}\nu^2(\omega) \times (y(\omega)-\hat{\pi}(\omega))^2 \tag{11.6}$$

Le rapport suivant assure l'équivalence entre les estimations des écarts-type :

$$\hat{\sigma}_{\hat{a}_j}(MMV) = \frac{\hat{\sigma}_{\hat{a}_j}(MCG)}{s} \tag{11.7}$$

3. Attention, il faudra spécifier une régression sans constante dans les logiciels. En effet, $\nu(\omega)$ est différent d'un individu à l'autre, le terme associé à a_0 n'est plus constant.

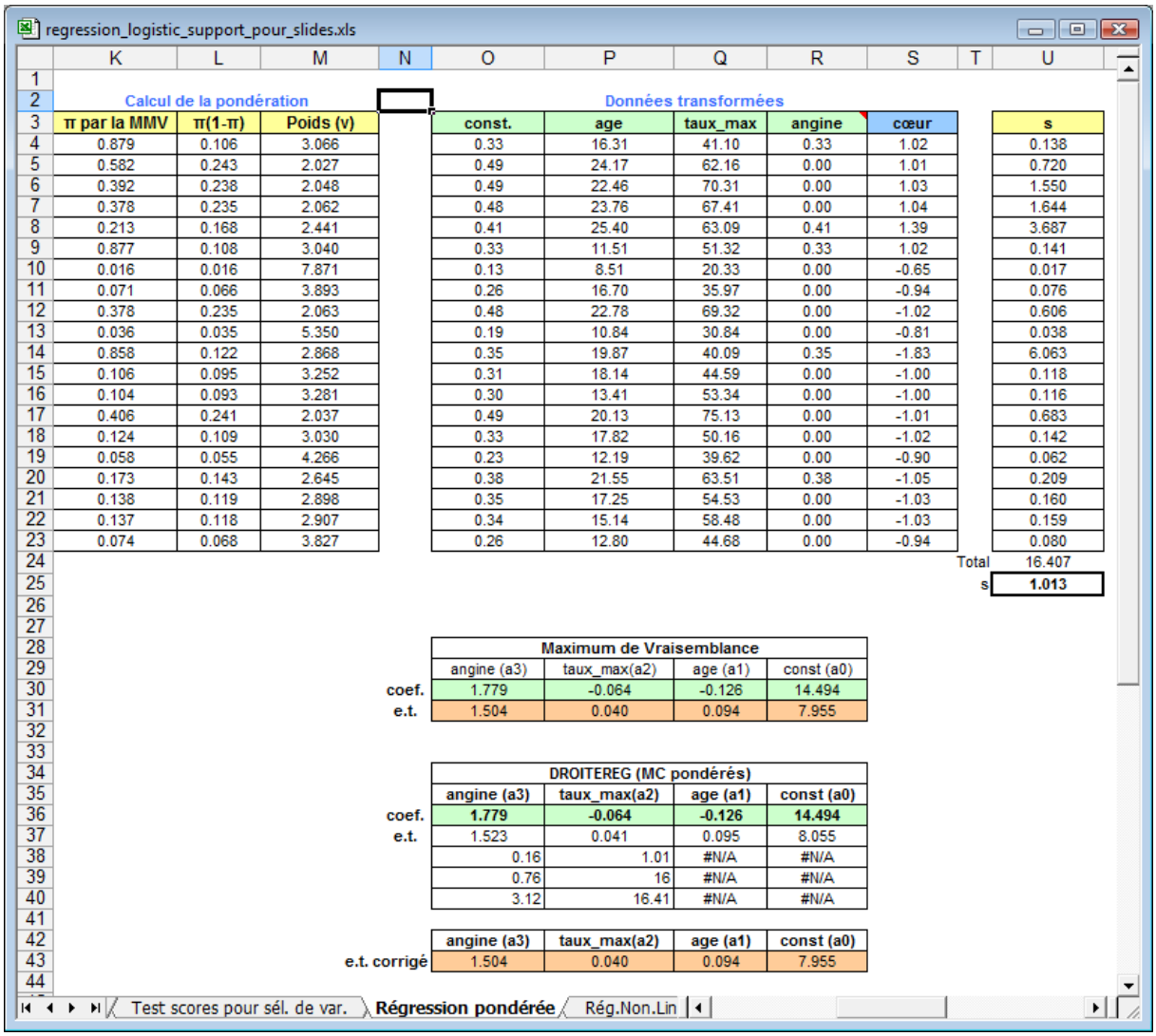
4. On ne manquera pas de noter la similitude avec la variance des erreurs en régression linéaire $\hat{\sigma}^2_\varepsilon = \frac{\text{somme des carrés des résidus}}{\text{degrés de liberté}}$.

11.2.3 Un exemple numérique avec la fonction DROITEREG

Reprenons le fichier COEUR ($n = 20$, Figure 0.1). Nous avions obtenus les paramètres $\hat{a}_{MMV}$ suivants avec la méthode du maximum de vraisemblance

-	angine a_3	taux max a_2	age a_1	a_0
Coef.	1.779	−0.064	−0.126	14.494
Ecart-type	1.504	0.040	0.094	7.955

Nous produisons les prédictions $\hat{\pi}$ à partir de ces coefficients. Nous pouvons réaliser les calculs pour obtenir $\hat{a}_{MCG}$ (Figure 11.4) :



regression_logistic_support_pour_slides.xls

Calcul de la pondération

π par la MMV	π(1-π)	Poids (v)
0.879	0.106	3.066
0.582	0.243	2.027
0.392	0.238	2.048
0.378	0.235	2.062
0.213	0.168	2.441
0.877	0.108	3.040
0.016	0.016	7.871
0.071	0.066	3.893
0.378	0.235	2.063
0.036	0.035	5.350
0.858	0.122	2.868
0.106	0.095	3.252
0.104	0.093	3.281
0.406	0.241	2.037
0.124	0.109	3.030
0.058	0.055	4.266
0.173	0.143	2.645
0.138	0.119	2.898
0.137	0.118	2.907
0.074	0.068	3.827

Données transformées

const.	age	taux_max	angine	cœur	s
0.33	16.31	41.10	0.33	1.02	0.138
0.49	24.17	62.16	0.00	1.01	0.720
0.49	22.46	70.31	0.00	1.03	1.550
0.48	23.76	67.41	0.00	1.04	1.644
0.41	25.40	63.09	0.41	1.39	3.687
0.33	11.51	51.32	0.33	1.02	0.141
0.13	8.51	20.33	0.00	-0.65	0.017
0.26	16.70	35.97	0.00	-0.94	0.076
0.48	22.78	69.32	0.00	-1.02	0.606
0.19	10.84	30.84	0.00	-0.81	0.038
0.35	19.87	40.09	0.35	-1.83	6.063
0.31	18.14	44.59	0.00	-1.00	0.118
0.30	13.41	53.34	0.00	-1.00	0.116
0.49	20.13	75.13	0.00	-1.01	0.683
0.33	17.82	50.16	0.00	-1.02	0.142
0.23	12.19	39.62	0.00	-0.90	0.062
0.38	21.55	63.51	0.38	-1.05	0.209
0.35	17.25	54.53	0.00	-1.03	0.160
0.34	15.14	58.48	0.00	-1.03	0.159
0.26	12.80	44.68	0.00	-0.94	0.080

Total 16.407

s 1.013

Maximum de Vraisemblance

	angine (a3)	taux_max(a2)	age (a1)	const (a0)
coef.	1.779	-0.064	-0.126	14.494
e.t.	1.504	0.040	0.094	7.955

DROITEREG (MC pondérés)

	angine (a3)	taux_max(a2)	age (a1)	const (a0)
coef.	1.779	-0.064	-0.126	14.494
e.t.	1.523	0.041	0.095	8.055
	0.16	1.01	#N/A	#N/A
	0.76	16	#N/A	#N/A
	3.12	16.41	#N/A	#N/A

	angine (a3)	taux_max(a2)	age (a1)	const (a0)
e.t. corrigé	1.504	0.040	0.094	7.955

Test scores pour sél. de var. | Régression pondérée | Rég.Non.Lin

Fig. 11.4. Calcul des coefficients via la régression linéaire pondérée

– Tout d'abord, nous avons la colonne de $\hat{\pi}$ obtenue à partir de la régression logistique. Par exemple, $\hat{\pi}(1) = 0.879$, $\hat{\pi}(2) = 0.582$, etc.

- Nous formons dans la colonne suivante $\hat{\pi}(1-\hat{\pi})$.
- Nous en déduisons le poids $\nu = \frac{1}{\sqrt{\hat{\pi}(1-\hat{\pi})}}$. Par exemple, $\nu(1) = \frac{1}{\sqrt{0.879(1-0.879)}} = 3.066$.
- Nous transformons maintenant les variables explicatives, de la constante (const., qui prenait systématiquement la valeur 1 initialement) à angine. Nous divisons simplement les valeurs initiales par le poids. Par exemple, $const(1) = \frac{1}{3.066} = 0.33$, $const(2) = \frac{1}{2.027} = 0.49$, etc. ; $age(1) = \frac{50}{3.066} = 16.31$, etc.
- Pour la variable dépendante, nous travaillons en deux temps : tout d'abord, nous construisons la variable z en utilisant la formule ci-dessus (Équation 11.3), puis nous la divisons par le poids ν. Pour le 1^{er} individu qui porte la valeur $y(1) = 1$, nous avons : $z(1) = \ln\left(\frac{0.879}{1-0.879}\right) + \frac{1-0.879}{0.879(1-0.879)} = 3.12$, puis $\frac{z(1)}{\nu(1)} = \frac{3.12}{3.066} = 1.02$.
- Nous pouvons lancer la régression via la fonction DROITEREG d'Excel. Attention, il faut demander une régression sans constante. Nous visualisons les résultats à partir de la ligne 35 dans la feuille Excel.
- Effectivement, les coefficients obtenus concordent avec ceux de la régression logistique $\hat{a}_{MCG} = \hat{a}_{MMV}$
- En revanche les écarts-type ne coïncident pas. Si l'on prend la variable *angine*, nous avons $\hat{\sigma}_{\hat{a}_3}(MCG) = 1.523$
- Il faut introduire la second correction (Équation 11.7). Pour cela, nous calculons la quantité s (Équation 11.6) (dernière colonne dans la feuille Excel)

$$s = \sqrt{\frac{1}{20-3-1} \times 16.407} = 1.013$$

- Nous pouvons corriger les écarts-type. Pour la variable angine, nous avons

$$\hat{\sigma}_{\hat{a}_3}(MMV) = \frac{1.523}{1.013} = 1.504$$

L'équivalence est totale.

11.3 Régression non-linéaire mais séparateur linéaire

La régression logistique est une régression non linéaire parce qu'elle utilise une fonction de transfert non linéaire (la fonction logistique). En revanche, elle induit bien une frontière linéaire entre les positifs et les négatifs dans l'espace de représentation. Ce sont là deux points de vues différents sur la même technique. Voyons ce qu'il en est sur un exemple.

Nous traitons le fichier COEUR (Figure 0.1). Nous prenons comme seules variables explicatives *age* et *taux max*. Ainsi, nous pourrons projeter les observations dans le plan (Figure 11.5). Pas besoin d'être grand clerc pour observer que nous avons la possibilité de tracer une droite pour séparer les positifs ($\diamondsuit$) des négatifs ($\triangle$).

Nous lançons une régression logistique à l'aide du logiciel Tanagra (Figure 11.6). Le LOGIT est une fonction linéaire des variables explicatives

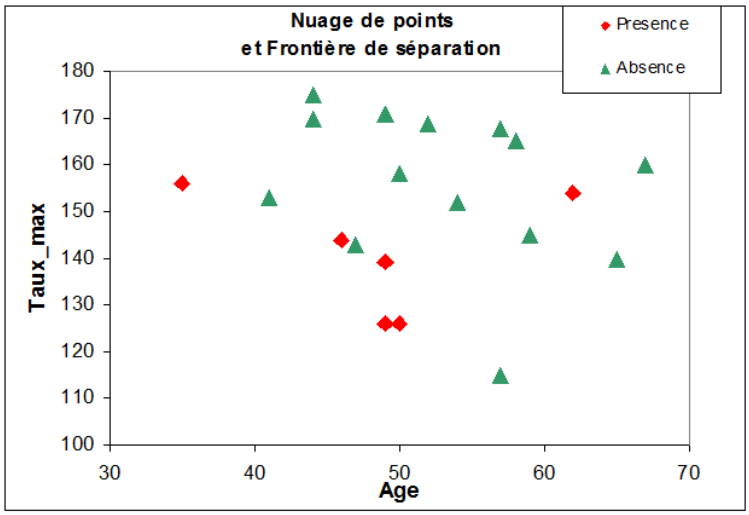


Fig. 11.5. Nuage de points (Age vs. Taux Max) selon Coeur

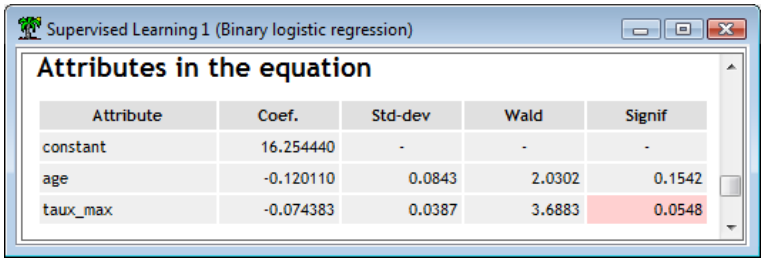


Fig. 11.6. Coefficients de la régression Coeur = f (age, taux max)

$$LOGIT = C(X) = 16.254 - 0.120 \times age - 0.074 \times taux\ max$$

La règle de décision usuelle est

$$\text{Si } C(X) > 0 \text{ Alors } \hat{Y} = + \text{ Sinon } \hat{Y} = -$$

Ainsi l'égalité $C(X) = 0$ définit la frontière séparant les positifs des négatifs

$$-0.120 \times age - 0.074 \times taux\ max + 16.254 = 0$$

Passons l'équation sous un forme explicite, nous obtenons une expression exploitable de la frontière

$$\begin{aligned} taux\ max &= \frac{-16.254}{-0.074} + \frac{0.120}{-0.074} \times age \\ &= 218.52 - 1.61 \times age \end{aligned}$$

Nous pouvons reporter cette droite dans le nuage de points. Nous visualisons la frontière utilisée par le classifieur pour distinguer les positifs des négatifs (Figure 11.7).

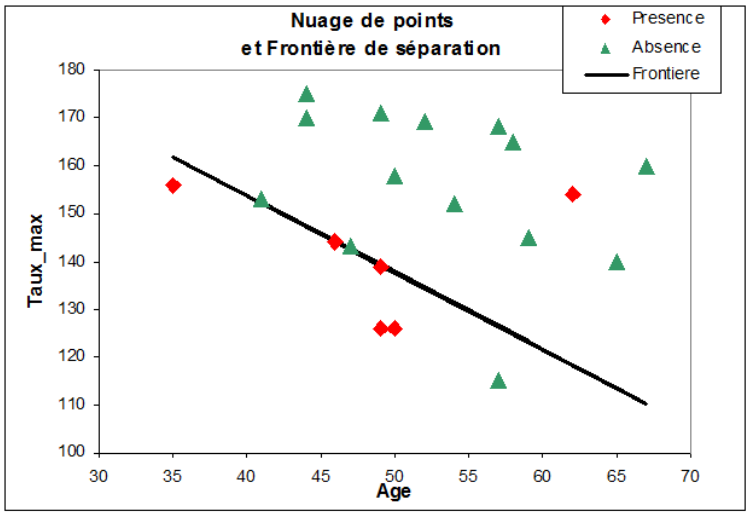


Fig. 11.7. Nuage de points (Age vs. Taux Max) - Frontière séparant les positifs et les négatifs

n°	age	taux_max	coeur	cœur	C(X)	π
1	50	126	presence	1	0.877	0.706
2	49	126	presence	1	0.997	0.730
3	46	144	presence	1	0.018	0.505
4	49	139	presence	1	0.030	0.507
5	62	154	presence	1	-2.647	0.066
6	35	156	presence	1	0.447	0.610
7	67	160	absence	0	-3.694	0.024
8	65	140	absence	0	-1.966	0.123
9	47	143	absence	0	-0.028	0.493
10	58	165	absence	0	-2.985	0.048
11	57	115	absence	0	0.854	0.701
12	59	145	absence	0	-1.618	0.166
13	44	175	absence	0	-2.048	0.114
14	41	153	absence	0	-0.051	0.487
15	54	152	absence	0	-1.538	0.177
16	52	169	absence	0	-2.562	0.072
17	57	168	absence	0	-3.088	0.044
18	50	158	absence	0	-1.504	0.182
19	44	170	absence	0	-1.676	0.158
20	49	171	absence	0	-2.351	0.087

Fig. 11.8. Coeur = f (age, taux max) - Tableau de calcul

Pour analyser finement ces résultats, nous donnons également le tableau des données complété des LOGIT et π prédits par le modèle (Figure 11.8). Quelques remarques viennent par rapport à la lecture croisée du graphique et du tableau de données :

- Certains individus sont bien classés mais à la lisière de la frontière. L'individu n^o3 avec (age = 46, taux max = 144) est "positif". Si on s'intéresse à son LOGIT, nous avons $\hat{C}(46, 144) = -0.120 \times 46 - 0.074 \times 144 + 16.254 = 0.018$ et $\hat{\pi} = \frac{1}{1+e^{-(0.018)}} = 0.0505$. On se rend compte effectivement qu'il est très proche de la frontière.
- D'autres sont bien classés de manière sûre c.-à-d. en étant très éloignés de la frontière. Considérons l'individu n^o17, avec (age = 57, taux max = 168), qui est "négatif". Son LOGIT est égal à

$\hat{C}(57, 168) = -0.120 \times 57 - 0.074 \times 168 + 16.254 = -3.088$ et $\hat{\pi} = \frac{1}{1+e^{-(-3.088)}} = 0.044$. Les résultats sont cohérents : il est très éloigné de la frontière, et la probabilité d'affectation associée est proche de 0 [5].

- D'autres enfin sont mal classés. Il y en a 2 du mauvais côté de la frontière dans notre exemple : un négatif noyé au milieu des positifs (individu n^o5) et inversement (n^o11) (Figure 11.7).
- Ce que confirme la matrice de confusion fournie par Tanagra (Figure 11.9).

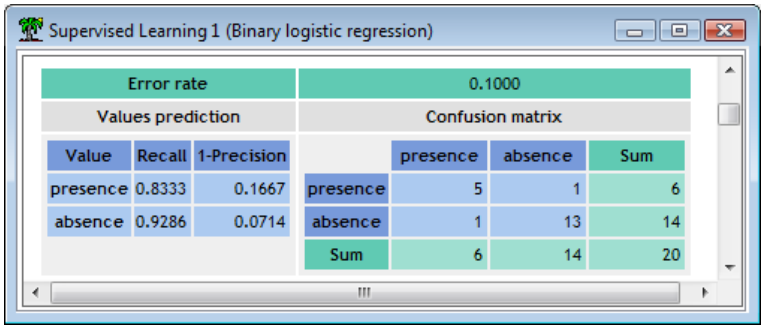


Fig. 11.9. Matrice de confusion de la régression Coeur = f (age, taux max)

5. Si c'était un positif bien classé, sa probabilité serait proche de 1

Partie III

La régression logistique multinomiale

12

Variable dépendante nominale - Principe et estimations

Lorsque la variable dépendante prend K ($K > 2$) modalités, nous sommes dans le cadre de la *régression logistique polytomique*. Dans ce partie, nous considérons qu'elle est nominale c.-à-d. il n'y a pas de relation d'ordre entre les modalités, ou tout du moins nous souhaitons ne pas en tenir compte si elle existe. On parle de *régression logistique multinomiale*. On peut la voir comme une forme de généralisation de la régression logistique binaire.

Nous devons répondre à plusieurs questions pour élaborer une stratégie d'apprentissage viable :

- Quelle forme de logit modéliser à l'aide d'une combinaison linéaire de variables, puisque nous devons rester dans le canevas de la régression linéaire généralisée ?
- Question corollaire : combien d'équations logit devons écrire ?
- Une fois le problème correctement posé, comment estimer les paramètres, étant entendu que nous passerons par la maximisation de la vraisemblance ?
- Question corollaire : comment s'écrit la (log)-vraisemblance ?
- Enfin, dernière question, comment évaluer la pertinence de la régression ? Nous traitons uniquement de validation en rapport direct avec les caractéristiques de la régression pour l'instant.

Pour l'heure, intéressons à la distribution de la variable dépendante Y.

12.1 La distribution multinomiale

L'objectif est de modéliser la probabilité d'appartenance d'un individu à une modalité y_k. Nous écrivons

$$\pi_k(\omega) = P(Y(\omega) = y_k / X(\omega)] \tag{12.1}$$

Avec la contrainte

$$\sum_k \pi_k(\omega) = 1$$

.

On s'appuie sur la loi multinomiale pour écrire la vraisemblance

$$L = \prod_{\omega} [\pi_1(\omega)]^{y_1(\omega)} \times \cdots \times [\pi_K(\omega)]^{y_K(\omega)} \tag{12.2}$$

Où

$$y_k(\omega) = \begin{cases} 1 \text{ si } Y(\omega) = y_k \\ 0 \text{ sinon} \end{cases}$$

Il s'agit bien d'une généralisation. En effet, nous retombons sur la loi binomiale si Y est binaire.

Quelle stratégie de modélisation utiliser pour parvenir à nos fins c.-à-d. obtenir des estimations de $\pi_k(\omega)$ à l'aide de la régression logistique ?

12.2 Écrire les logit par rapport à une modalité de référence

L'idée de la régression logistique multinomiale est de modéliser $(K-1)$ rapports de probabilités (odds). Nous prenons une modalité comme référence (en anglais, *baseline outcome*), et nous exprimons les logit par rapport à cette référence (*baseline category logits*) ([1], pages 307 à 317 ; [9], pages 260 à 287).

La catégorie de référence s'impose souvent naturellement au regard des données analysées : les non-malades vs. les différents type de maladies ; le produit phare du marché vs. les produits outsiders ; etc. Si ce n'est pas le cas, si toutes les modalités sont sur un pied d'égalité, nous pouvons choisir n'importe laquelle. Cela n'a aucune incidence sur les calculs, seule l'interprétation des coefficients est différente.

Par convention, nous décidons que la dernière catégorie Y_K sera la modalité de référence dans cette partie. Le logit pour la modalité y_k s'écrit

$$C_k = \ln \frac{\pi_k}{\pi_K} = a_{0,k} + a_{1,k}X_1 + \cdots + a_{J,k}X_J \tag{12.3}$$

Nous en déduisons les (K-1) probabilités a posteriori

$$\pi_k = \frac{e^{C_k}}{1 + \sum_{k=1}^{K-1} e^{C_k}} \tag{12.4}$$

La dernière probabilité π_K peut être obtenue directement ou par différenciation

$$\pi_K = \frac{1}{1 + \sum_{k=1}^{K-1} e^{C_k}} = 1 - \sum_{k=1}^{K-1} \pi_k \tag{12.5}$$

Pour un individus ω, les probabilités doivent vérifier la relation

$$\sum_{k=1}^{K} \pi_k(\omega) = 1$$

La règle d'affectation est conforme au schéma bayesien

$$Y(\omega) = y_{k^*} \Leftrightarrow y_{k^*} = \arg\max_k \pi_k(\omega) \tag{12.6}$$

12.3 Estimation des paramètres

12.3.1 Vecteur gradient et matrice hessienne

Pour estimer les $(K-1) \times (J+1)$ coefficients, nous devons optimiser la log-vraisemblance

$$LL = \sum_{\omega} y_1(\omega) \ln \pi_1(\omega) + \cdots + y_K(\omega) \ln \pi_K(\omega) \tag{12.7}$$

via l'algorithme de Newton-Raphson. Pour ce faire, nous avons besoin des expressions du vecteur gradient et de la matrice hessienne.

Le **vecteur gradient** G est de dimension $(K-1) * (J+1) \times 1$

$$G = \begin{pmatrix} G_1 \\ \vdots \\ G_{K-1} \end{pmatrix} \tag{12.8}$$

où G_k, relatif à la modalité y_k, est un vecteur de dimension $(J+1) \times 1$ dont la composante $n^o j$ s'écrit

$$g_{k,j} = \sum_{\omega} x_j(\omega) \times [y_k(\omega) - \pi_k(\omega)] \tag{12.9}$$

Concernant la **matrice hessienne**, elle est de dimension $(K-1) * (J+1) \times (K-1) * (J+1)$

$$H = \begin{pmatrix} H_{11} & \cdots & H_{1,K-1} \\ \vdots & & \vdots \\ H_{K-1,1} & \cdots & H_{K-1,K-1} \end{pmatrix} \tag{12.10}$$

$H_{i,j}$ est de dimension $(J+1) \times (J+1)$, définie par [1]

$$H_{i,j} = \sum_{\omega} \pi_i(\omega) \times [\delta_{i,j}(\omega) - \pi_j(\omega)] \times X(\omega) \times X'(\omega) \tag{12.11}$$

$X(\omega) = (1, X_1(\omega), \ldots, X_J(\omega))$ est le vecteur de description de l'observation ω, incluant la constante.

et

$$\delta_{i,j}(\omega) = \begin{cases} 1 \text{ si } i = j \\ 0 \text{ si } i \neq j \end{cases}$$

La matrice H est symétrique par blocs c.-à-d. $H_{i,j} = H_{j,i}$

1. A vrai dire, ces formules sont surtout mises en avant pour ceux qui souhaiteraient programmer la méthode. J'ai eu un mal fou à les retrouver pour les implémenter dans Tanagra, autant les détailler une fois pour toutes dans un document. Il ne sera pas question de les reproduire à la main dans Excel. Non, non, restons raisonnables.

12.3.2 Un exemple : prédiction de formule de crédit

Une enseigne de grande surface met à disposition de ses clients 3 formules de crédit revolver [2] (A, B et C). Le conseiller doit faire attention lorsqu'il est face au client. S'il met en avant une formule inadaptée, il risque de le décourager et de le voir partir. L'objectif de l'étude est de cibler, à partir de données que l'on peut facilement recueillir [l'âge (quantitatif), le sexe (binaire, 1 = homme) et le revenu par tête du ménage (quantitatif)], la formule que l'on doit proposer en priorité.

Nous disposons de $n = 30$ observations, les distributions sont équilibrées : $n_A = n_B = n_C = 10$. Pour bien détailler les étapes, à l'instar de ce que nous avions fait pour la régression binaire (section 1.4), nous montons une feuille Excel qui permet de produire la log-vraisemblance à partir des données et des paramètres, puis nous utilisons le solveur.

	const	age	sexe	rev.tete
a(A/C)	1.000	0.000	0.000	0.000
a(B/C)	1.000	0.000	0.000	0.000

age	sexe	rev.tete	type.credit	C(A/C)	C(B/C)	PI(A)	PI(B)	PI(C)	y.A	y.B	y.C	LL
29	0	7.09	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
27	0	2.25	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
24	1	4.47	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
18	1	4.66	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
21	0	6.47	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
30	1	9.18	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
26	0	8.96	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
23	1	5.68	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
22	1	7.39	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
20	1	9.93	A	1.00	1.00	0.42	0.42	0.16	1	0	0	-0.86
49	1	13.60	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
47	0	21.53	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
50	0	11.41	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
51	0	8.27	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
54	1	11.62	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
49	1	14.28	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
44	1	12.51	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
46	0	9.45	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
48	0	7.85	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
52	0	11.21	B	1.00	1.00	0.42	0.42	0.16	0	1	0	-0.86
39	0	6.50	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
37	1	7.84	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
30	0	7.40	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
29	0	10.10	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
27	1	8.43	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
34	1	9.19	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
30	1	12.31	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
21	1	14.13	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
52	1	10.04	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86
50	0	8.90	C	1.00	1.00	0.42	0.42	0.16	0	0	1	-1.86

LL	-35.860

-2 LL	71.720

Fig. 12.1. Régression multinomiale - Formule de crédit - Initialisation de la feuille Excel

Dans un premier temps, nous mettons arbitrairement les coefficients des variables à 0, et les constantes à 1. La modalité C est la catégorie de référence. Décrivons la feuille de calcul (Figure 12.1) :

– Dans la partie haute, nous distinguons les coefficients.

2. Au final, le client est toujours mort, c'est ça l'idée.

- Nous disposons du tableau de descripteurs (vert) et de la variable dépendante (bleu).
- Nous formons les logit. Pour le premier logit C_1 opposant A à C, pour la première observation $\omega = 1$

$$C_1(1) = 1.0 + 0.0 \times 29 + 0.0 \times 0 + 0.0 \times 7.09 = 1.00$$

  Nous faisons de même pour le second logit, nous obtenons $C_2(1) = 1.00$.
- Nous en déduisons la probabilité a posteriori

$$\hat{\pi}_1(1) = \frac{e^1}{1 + (e^{1.00} + e^{1.00})} = 0.42$$

- Pour les autres probabilités, nous avons $\hat{\pi}_2(1) = 0.42$ et $\hat{\pi}_3(1) = 1 - (0.42 + 0.42) = 0.16$
- Dans les 3 colonnes qui suivent, nous avons les indicatrices de modalités de la variable dépendante Y.
- Ainsi, nous pouvons former la fraction de la log-vraisemblance associée au premier individus $\omega = 1$

$$LL(1) = 1 \times \ln(0.42) + 0 \times \ln(0.42) + 0 \times \ln(0.16) = -0.86$$

- Et la log-vraisemblance

$$LL = -0.86 + \cdots + (-1.86) = -35.860$$

- La déviance à ce stade est égale à

$$D = -2 \times LL = 71.720$$

Nous lançons le solveur d'Excel [3]. Nous souhaitons maximiser la vraisemblance (ou minimiser la déviance, c'est équivalent). Les cellules variables correspondent aux coefficients de la régression. La feuille prend une autre tournure (Figure 12.2) :

- La log-vraisemblance optimisée est maintenant égale à

$$LL = -9.191$$

  La déviance

$$D = -2 \times (-9.191) = 18.382$$

- Nous avons les deux équations logit

$$C_1 = 21.165 - 0.471 \times age + 0.170 \times sexe - 0.935 \times rev.tete$$
$$C_2 = -26.328 + 0.286 \times age - 4.966 \times sexe + 1.566 \times rev.tete$$

3. Étonnamment, il faut le lancer 2 fois avant de parvenir à une solution stable définitive.

	const	age	sexe	rev.tete
a(A/C)	21.165	-0.471	0.170	-0.935
a(B/C)	-26.328	0.286	-4.966	1.566

age	sexe	rev.tete	type.credit	C(A/C)	C(B/C)	Pl(A)	Pl(B)	Pl(C)	y.A	y.B	y.C	LL
29	0	7.09	A	0.87	-6.93	0.70	0.00	0.30	1	0	0	-0.35
27	0	2.25	A	6.33	-15.08	1.00	0.00	0.00	1	0	0	0.00
24	1	4.47	A	5.84	-17.43	1.00	0.00	0.00	1	0	0	0.00
18	1	4.66	A	8.49	-18.85	1.00	0.00	0.00	1	0	0	0.00
21	0	6.47	A	5.22	-10.19	0.99	0.00	0.01	1	0	0	-0.01
30	1	9.18	A	-1.39	-8.33	0.20	0.00	0.80	1	0	0	-1.61
26	0	8.96	A	0.53	-4.86	0.63	0.00	0.37	1	0	0	-0.47
23	1	5.68	A	5.18	-15.82	0.99	0.00	0.01	1	0	0	-0.01
22	1	7.39	A	4.06	-13.43	0.98	0.00	0.02	1	0	0	-0.02
20	1	9.93	A	2.62	-10.02	0.93	0.00	0.07	1	0	0	-0.07
49	1	13.60	B	-14.48	4.03	0.00	0.98	0.02	0	1	0	-0.02
47	0	21.53	B	-21.12	20.84	0.00	1.00	0.00	0	1	0	0.00
50	0	11.41	B	-13.07	5.85	0.00	1.00	0.00	0	1	0	0.00
51	0	8.27	B	-10.61	1.22	0.00	0.77	0.23	0	1	0	-0.26
54	1	11.62	B	-14.98	2.36	0.00	0.91	0.09	0	1	0	-0.09
49	1	14.28	B	-15.11	5.09	0.00	0.99	0.01	0	1	0	-0.01
44	1	12.51	B	-11.10	0.89	0.00	0.71	0.29	0	1	0	-0.34
46	0	9.45	B	-9.35	1.64	0.00	0.84	0.16	0	1	0	-0.18
48	0	7.85	B	-8.80	-0.30	0.00	0.43	0.57	0	1	0	-0.85
52	0	11.21	B	-13.83	6.11	0.00	1.00	0.00	0	1	0	0.00
39	0	6.50	C	-3.30	-4.99	0.04	0.01	0.96	0	0	1	-0.04
37	1	7.84	C	-3.44	-8.43	0.03	0.00	0.97	0	0	1	-0.03
30	0	7.40	C	0.11	-6.15	0.53	0.00	0.47	0	0	1	-0.75
29	0	10.10	C	-1.95	-2.21	0.11	0.09	0.80	0	0	1	-0.22
27	1	8.43	C	0.73	-10.37	0.67	0.00	0.33	0	0	1	-1.12
34	1	9.19	C	-3.28	-7.17	0.04	0.00	0.96	0	0	1	-0.04
30	1	12.31	C	-4.32	-3.43	0.01	0.03	0.96	0	0	1	-0.04
21	1	14.13	C	-1.78	-3.16	0.14	0.04	0.83	0	0	1	-0.19
52	1	10.04	C	-12.56	-0.69	0.00	0.33	0.67	0	0	1	-0.41
50	0	8.90	C	-10.72	1.92	0.00	0.87	0.13	0	0	1	-2.06

LL	-9.191
-2 LL	18.382

Fig. 12.2. Formule de crédit - Après optimisation de la log-vraisemblance via le solveur

- Une première lecture rapide des coefficients estimées - nous reviendrons plus loin sur les interprétations - nous donne les indications suivantes (tout ceci sous réserve de la significativité des coefficients) :
  - Plus le client est âgé, moins il est enclin à prendre le crédit A (par rapport au C). Ou autrement, les personnes qui prennent le crédit A sont moins âgés que ceux qui prennent le C. Les hommes ont plus tendance à prendre A (par rapport à C). Enfin, un revenu par tête plus élevé dans le ménage n'incite pas à prendre A (par rapport à C). Bref, la principale idée à retenir est que toute la lecture doit se faire par rapport à la modalité de référence C.
  - Pour la seconde opposition B vs. C, nous constatons à contrario que une augmentation de l'âge incite à prendre B (par rapport à C) ; il semble que les femmes ont plus de chances de prendre la formule B (par rapport à C) ; plus son revenu est élevé, plus le client se dirigera volontiers vers B (par rapport à C).
- Concernant le logit et la probabilité a posteriori, voici le détail des calculs pour le premier individu $\omega = 1$
  - Pour le premier logit

$$C_1(1) = 21.165 - 0.471 \times 29 + 0.170 \times 1 - 0.935 \times 7.09 = 0.87$$

– La probabilité d'affectation à la première modalité

$$\hat{\pi}_1(1) = \frac{e^{0.87}}{1 + e^{0.87} + e^{-6.93}} = 0.70$$

– Pour les 3 modalités, nous avons

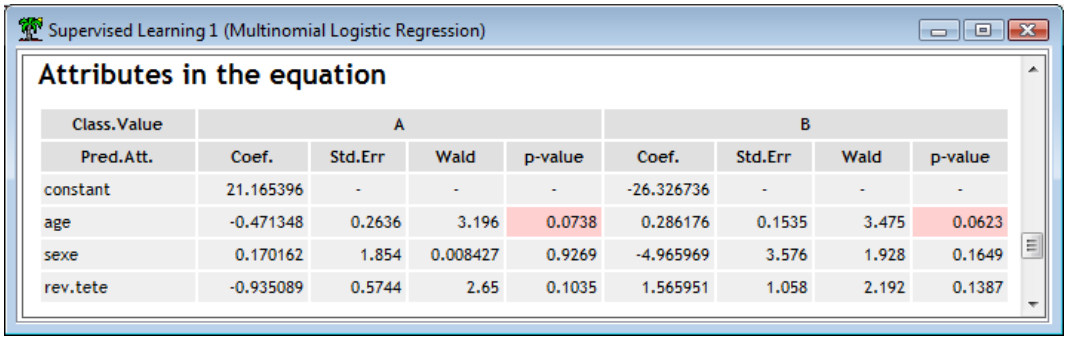
$$\hat{\pi}(1) = (0.70; 0.00; 0.30)$$

– La fraction de la log-vraisemblance associée[4] s'écrit

$$LL(1) = 1 \times \ln 0.70 + 0 \times \ln 0.00 + 0 \times 0.30 = -0.35$$

12.3.3 Estimation des coefficients avec Tanagra et R (packages nnet et VGAM)

La régression multinomiale est disponible via le composant **Multinomial Logistic Regression** dans **Tanagra**. Il s'utilise comme n'importe quel composant d'apprentissage supervisé. Il faut bien entendu que les explicatives soient numériques, quantitatives ou qualitatives codées 0/1. Pour notre exemple "Formules de Crédit", nous retrouvons les coefficients estimés à l'aide du tableur. Tanagra prend automatiquement la dernière modalité "C" comme référence. C'est exactement ce qu'il fallait dans notre configuration (Figure 12.3).



Class.Value	A				B			
Pred.Att.	Coef.	Std.Err	Wald	p-value	Coef.	Std.Err	Wald	p-value
constant	21.165396	-	-	-	-26.326736	-	-	-
age	-0.471348	0.2636	3.196	0.0738	0.286176	0.1535	3.475	0.0623
sexe	0.170162	1.854	0.008427	0.9269	-4.965969	3.576	1.928	0.1649
rev.tete	-0.935089	0.5744	2.65	0.1035	1.565951	1.058	2.192	0.1387

Fig. 12.3. Formule de crédit - Estimation des coefficients avec **Tanagra**

Deux outils (entres autres, il est impossible de tous les connaître) sont disponibles pour estimer les paramètres de la régression logistique multinomiale dans R. La première est la fonction **multinom** du package **nnet** (Figure 12.4). Il faut abaisser fortement les seuils de tolérance pour obtenir un résultat précis, conformes à ceux produits par les autres logiciels. L'immense avantage de "multinom" est *qu'elle sait nous fournir la matrice hessienne*. Nous en aurons l'usage lorsqu'il s'agira de mettre en place les tests de significativité (chapitre 14).

La seconde fonction est **vglm** du package **VGAM** (Figure 12.5). Elle ne fait guère plus que la précédente concernant la régression multinomiale. Son intérêt réside surtout dans l'intégration de la régression polytomique (nominale ou ordinale) dans un environnement unique. Nous l'utiliserons plus intensivement lorsqu'il s'agira de traiter la régression à variable dépendante ordinale (partie IV).

4. ($\ln 0.00$) devrait produire une erreur. La valeur est en réalité très petite mais non nulle. L'affichage est arrondi à 2 chiffres après la virgule.

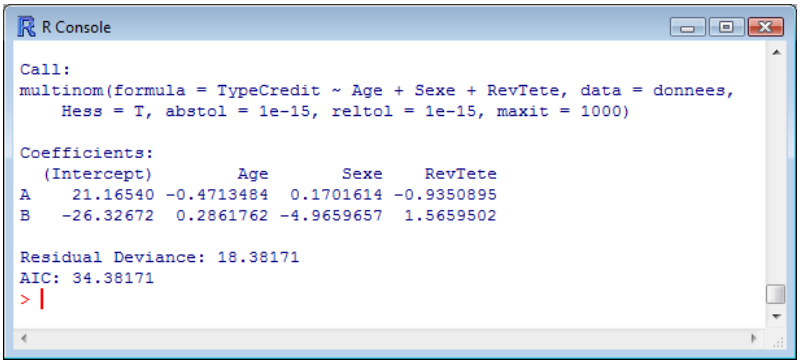


Fig. 12.4. Formule de crédit - Estimation des coefficients avec **R** (**multinom** de **nnet**)

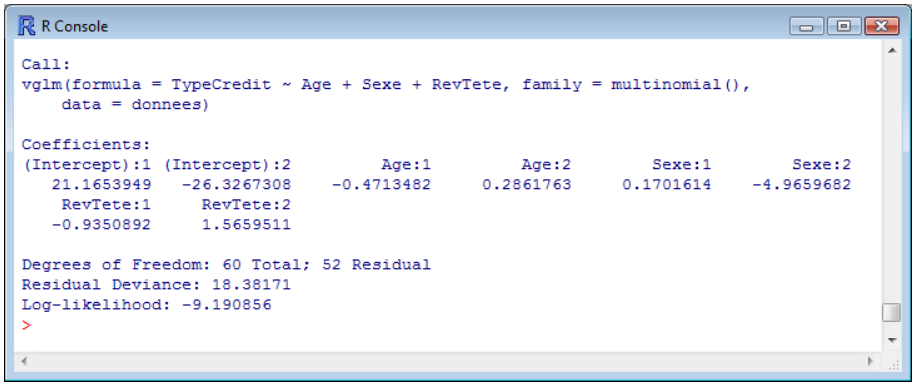


Fig. 12.5. Formule de crédit - Estimation des coefficients avec **R** (**vglm** de **VGAM**)

12.3.4 Modifier la modalité de référence

Le choix de la modalité de référence pèse essentiellement sur la lecture des coefficients. Dans notre exemple "Formule de Crédit", nous aimerions savoir quel pourrait être le rapport entre les modalités A et B, en l'état ce n'est pas possible parce que C est la référence. Est-ce que nous sommes condamnés à relancer la régression en modifiant explicitement la référence ?

Non. En toute généralité, si y_K est la modalité de référence, nous montrons dans cette section qu'il est possible d'opposer deux modalités quelconques y_i et y_j sans avoir à relancer les calculs. En effet

$$\begin{aligned} logit_{i,j} &= \ln \frac{\pi_i}{\pi_j} \\ &= \ln \frac{\pi_i/\pi_K}{\pi_j/\pi_K} \\ &= \ln \frac{\pi_i}{\pi_K} - \ln \frac{\pi_j}{\pi_K} \\ &= C_i - C_j \end{aligned}$$

Par simple différenciation, nous obtenons le logit (logarithme de l'odds) entre 2 modalités quelconques de la variable dépendante. Le choix initial de la modalité de référence n'est pas restrictif.

Formule de crédit

Essayons de caractériser la modalité A par rapport à B dans notre exemple des formules de crédit. Par différenciation des logit, nous obtenons

$$\begin{aligned} logit_{A,B} &= C_1 - C_2 \\ &= (21.165 + 26.328) + (-0.471 - 0.286) \times age + (0.170 + 4.966) \times sexe + (-0.935 - 1.566) \times rev.tete \\ &= 47.493 - 0.758 \times age + 5.136 \times sexe - 2.501 \times rev.tete \end{aligned}$$

Tous les effets sont exacerbés dans (A vs. B) par rapport (A vs. C). Les coefficients conservent leur signes, mais sont plus élevés en valeur absolue : plus l'âge augmente, moins les clients choisissent A (par rapport à B) ; les hommes sont plus enclins à prendre la formule A (par rapport à B) ; et les revenus élevés les dissuadent de prendre A (par rapport à B).

12.4 Significativité globale de la régression

Pour évaluer la qualité globale de la régression, nous le savons maintenant, nous devons mesurer les performances du modèle trivial réduit uniquement aux constantes. Il y en a $K-1$ dans notre configuration. A l'instar de la régression binaire, nous pouvons (1) produire directement l'estimation des constantes sans passer par une optimisation de la vraisemblance, (2) en déduire la valeur de la log-vraisemblance, (3) que l'on comparera avec celle du modèle à évaluer. Nous pourrons dégager 2 indicateurs : le test du rapport de vraisemblance ; le pseudo-R^2 de Mc Fadden.

12.4.1 Modèle trivial : estimations et log-vraisemblance

Les effectifs des modalités de la variable dépendante suffisent pour produire les estimations des constantes dans le modèle trivial, en l'occurrence

$$\hat{a}_{0,k} = \ln \frac{n_k}{n_K} \tag{12.12}$$

Puisque dans le modèle trivial, la prévalence constatée dans l'échantillon $\hat{p}_k = \frac{n_k}{n}$ est l'estimateur de la probabilité a posteriori $\hat{\pi}_k$, nous pouvons écrire facilement la log-vraisemblance

$$\begin{aligned} LL_0 &= \sum_{\omega} \sum_{k} y_k(\omega) \ln(\hat{\pi}_k) \\ &= \sum_{\omega} \sum_{k} y_k(\omega) \ln(\hat{p}_k) \\ &= \sum_{k} n_k \ln \frac{n_k}{n} \end{aligned}$$

Application aux données "Formule de crédit"

Dans notre exemple "Formule de crédit", les classes sont équilibrées, la log-vraisemblance du modèle trivial est très facile à produire

$$\begin{aligned} LL_0 &= 10 \ln \frac{10}{30} + 10 \ln \frac{10}{30} + 10 \ln \frac{10}{30} \\ &= 30 \ln \frac{10}{30} \\ &= -32.958 \end{aligned}$$

La déviance du modèle trivial est égale à

$$D_0 = -2 \times LL_0 = -2 \times (-32.958) = 65.917$$

12.4.2 Pseudo-R^2 de McFadden

Notons LL_M la vraisemblance du modèle étudié, le pseudo-R^2 de McFadden est défini de la même manière que pour la régression binaire, à savoir

$$R^2_{MF} = 1 - \frac{LL_M}{LL_0}$$

Le pseudo-R^2 de McFadden varie entre 0 (modèle pas meilleur que le trivial) et 1 (modèle parfait).

Concernant les données "Formule de crédit", nous obtenons

$$R^2_{MF} = 1 - \frac{LL_M}{LL_0} = 1 - \frac{(-9.191)}{(-32.958)} = 0.721$$

Le modèle semble bon. Nous verrons dans la section suivante s'il est globalement significatif.

12.4.3 Test du rapport de vraisemblance

Le test de rapport de vraisemblance consiste à comparer 2 déviances. Pour l'évaluation globale il s'agit de confronter celles du modèle étudié et du modèle trivial.

La statistique du test s'écrit

$$LR = D_0 - D_M \tag{12.13}$$

Elle suit une loi du χ^2, reste à déterminer les degrés de liberté.

Les degrés de liberté des modèles à opposer s'écrivent

$$ddl_M = n - [(K-1) \times (J+1)]$$
$$ddl_0 = n - (K-1)$$

Nous obtenons ceux du rapport de vraisemblance par différenciation, ils correspondent à l'écart entre le nombre de paramètres estimés dans les deux modèles

$$ddl = ddl_0 - ddl_M = (K-1) \times J \tag{12.14}$$

La région critique du test au risque α correspond aux grandes valeurs de la statistique de test c.-à-d.

$$LR > \chi^2_{1-\alpha}(ddl)$$

Nous pouvons aussi décider via la p-value. Si elle est plus petite que α, le modèle est globalement significatif.

Application aux données "Formule de crédit"

La déviance du classifieur étudié est $D_M = 18.382$, celui du modèle trivial $D_0 = 65.917$. Nous formons

$$LR = 65.917 - 18.382 = 47.535$$

Avec la loi du χ^2 à $ddl = (3-1) \times 3 = 6$ degrés de liberté, nous obtenons une p-value inférieure à 0.0001. Le modèle est globalement très significatif.

12.4.4 Les résultats fournis par les logiciels

Tanagra fournit ces valeurs (D_M, D_0, R^2_{MF}, test du rapport de vraisemblance) dans le tableau d'évaluation globale de la régression (Figure 12.6).

Le calcul n'est pas directement réalisé avec **multinom** de R. Le plus simple est d'estimer explicitement le modèle trivial pour obtenir la déviance $D_0 = 65.91674$ (Figure 12.7). Nous pouvons reproduire les calculs ci-dessus pour obtenir les indicateurs adéquats.

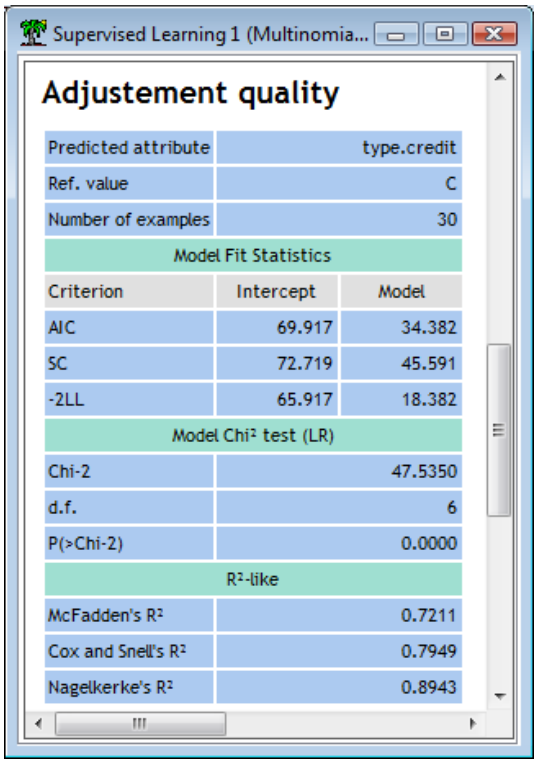


Fig. 12.6. Formule de crédit - Evaluation globale de la régression - Tanagra

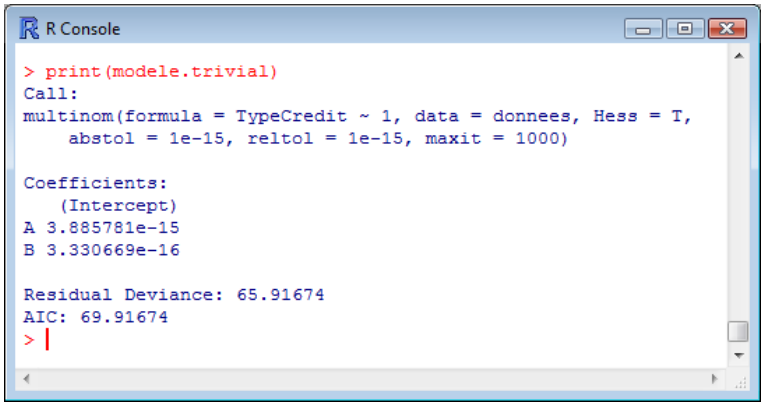


Fig. 12.7. Formule de crédit - Modèle trivial - R

13

Évaluation des classifieurs pour Y à $(K > 2)$ modalités nominales

De nouveau, nous travaillons principalement avec les prédictions $\hat{y}$ et les probabilités prédites $\hat{\pi}$ fournies par la régression dans ce chapitre. Les outils présentés dépassent donc le simple cadre de la régression logistique. Ils sont applicables pour tous types de classifieurs, pourvu qu'ils sachent fournir ces quantités.

La majorité des indicateurs de performances sont extraits de la matrice de confusion qui est une généralisation à K $(K > 2)$ modalités de celle présentée dans le cadre binaire. Il y a quand même une petite particularité. Dans le classement binaire, une des catégories revêtait une importance accrue par rapport à l'autre (positif vs. négatif). Certains ratios en tenaient compte (sensibilité, précision, etc.). Dans le cadre multi-classes [1], les modalités de Y sont mises sur un même pied d'égalité. Cela ne pose aucun problème pour certains (taux d'erreur). D'autres en revanche, ceux qui s'appuient sur le schéma "une catégorie contre les autres", doivent procéder à (une sorte de) moyenne sur l'ensemble des catégories pour parvenir à un indicateur caractérisant le comportement global du modèle (micro-averaging, macro-averaging pour la combinaison rappel et précision).

Bien entendu, les informations obtenues seront d'autant plus fiables que nous travaillons sur un fichier test n'ayant pas participé à l'estimation des paramètres du modèle.

13.1 Classement d'un individu

Pour classer un nouvel individu ω, nous calculons les probabilités a posteriori prédites $\hat{\pi}_k(\omega)$ pour chaque modalité de la variable dépendante. En accord avec la règle bayesienne,

$$\hat{Y}(\omega) = y_{k^*} \Leftrightarrow y_{k^*} = \arg\max_k \hat{\pi}_k(\omega)$$

Reprenons le premier individu du tableau de données "Formules de crédit" (Figure 12.2). Il est décrit par $(age = 29\,;\ sexe = 0\,;\ rev.tete = 7.09)$. Nous avions calculé les deux logit $C_1 = 0.87$ et $C_2 = -6.93$. Nous en avions déduit $\hat{\pi}_1 = 0.70$, $\hat{\pi}_2 = 0.00$ et $\hat{\pi}_3 = 0.30$. La prédiction du modèle est donc $\hat{Y}(1) = y_1 = A$ puisque c'est la modalité qui maximise la probabilité d'appartenance au groupe.

Obs. x Pred	y_1	...	y_l	...	y_K	Total
y_1	n_{11}	...	n_{1l}	...	n_{1K}	$n_{1.} = n_1$
...						...
y_k	n_{k1}	...	n_{kl}	...	n_{kK}	$n_{k.} = n_k$
...						...
y_K	n_{K1}	...	n_{Kl}	...	n_{KK}	$n_{K.} = n_K$
Total	$n_{.1}$	...	$n_{.l}$	...	$n_{.K}$	n

Tableau 13.1. Matrice de confusion pour un apprentissage multi-classes $(K > 2)$

13.2 Matrice de confusion et taux d'erreur

La matrice de confusion confronte les valeurs observées de Y sur l'échantillon et les valeurs prédites par le modèle. Nous avons un tableau de contingence (Tableau 13.1). Les effectifs de la case (k, l) est égal au nombre d'individus appartenant à la catégorie y_k qui ont été affectés à y_l

$$n_{kl} = \#\{\omega, Y(\omega) = y_k \text{ et } \hat{Y}(\omega) = y_l\}$$

Le taux d'erreur est l'estimation de la probabilité de mal classer, il correspond au rapport entre le nombre total d'observations mal classées et l'effectif total dans la fichier

$$\epsilon = \frac{\sum_k \sum_{l \neq k} n_{kl}}{n} = 1 - \frac{\sum_k n_{kk}}{n} \tag{13.1}$$

Si le modèle classe parfaitement les observations, nous avons $\epsilon = 0$. L'autre référence est le taux d'erreur du classifieur par défaut. Celui qui n'utilise pas les informations en provenance des explicatives. Nous avions vu plus haut comment le définir et comment en déduire un indicateur d'intérêt du modèle (section 2.1.4).

Le taux de succès θ est toujours le complément à 1 du taux d'erreur, il indique la probabilité de bien classer

$$\theta = 1 - \epsilon = \frac{\sum_k n_{kk}}{n}$$

Matrice de confusion pour les données "Formules de crédit"

Nous avons rajouté la colonne prédiction dans notre feuille Excel (Figure 13.1). Nous avons pu former la matrice de confusion. Nous avons mis en vert (pour ceux qui ont un moniteur couleur) les prédictions correctes, sur la diagonale principale de la matrice ; en rose saumon les mauvaises prédictions, hors diagonale. Nous obtenons le taux d'erreur et le taux de succès

1. Terme couramment utilisé en apprentissage automatique pour indiquer que Y prend plus de 2 modalités.

	const	age	sexe	rev.tete
a(A/C)	21.165	-0.471	0.170	-0.935
a(B/C)	-26.328	0.286	-4.966	1.566

age	sexe	rev.tete	type.credit	C(A/C)	C(B/C)	Pl(A)	Pl(B)	Pl(C)	Prédiction
29	0	7.09	A	0.87	-6.93	0.70	0.00	0.30	A
27	0	2.25	A	6.33	-15.08	1.00	0.00	0.00	A
24	1	4.47	A	5.84	-17.43	1.00	0.00	0.00	A
18	1	4.66	A	8.49	-18.85	1.00	0.00	0.00	A
21	0	6.47	A	5.22	-10.19	0.99	0.00	0.01	A
30	1	9.18	A	-1.39	-8.33	0.20	0.00	0.80	C
26	0	8.96	A	0.53	-4.86	0.63	0.00	0.37	A
23	1	5.68	A	5.18	-15.82	0.99	0.00	0.01	A
22	1	7.39	A	4.06	-13.43	0.98	0.00	0.02	A
20	1	9.93	A	2.62	-10.02	0.93	0.00	0.07	A
49	1	13.60	B	-14.48	4.03	0.00	0.98	0.02	B
47	0	21.53	B	-21.12	20.84	0.00	1.00	0.00	B
50	0	11.41	B	-13.07	5.85	0.00	1.00	0.00	B
51	0	8.27	B	-10.61	1.22	0.00	0.77	0.23	B
54	1	11.62	B	-14.98	2.36	0.00	0.91	0.09	B
49	1	14.28	B	-15.11	5.09	0.00	0.99	0.01	B
44	1	12.51	B	-11.10	0.89	0.00	0.71	0.29	B
46	0	9.45	B	-9.35	1.64	0.00	0.84	0.16	B
48	0	7.85	B	-8.80	-0.30	0.00	0.43	0.57	C
52	0	11.21	B	-13.83	6.11	0.00	1.00	0.00	B
39	0	6.50	C	-3.30	-4.99	0.04	0.01	0.96	C
37	1	7.84	C	-3.44	-8.43	0.03	0.00	0.97	C
30	0	7.40	C	0.11	-6.15	0.53	0.00	0.47	A
29	0	10.10	C	-1.95	-2.21	0.11	0.09	0.80	C
27	1	8.43	C	0.73	-10.37	0.67	0.00	0.33	A
34	1	9.19	C	-3.28	-7.17	0.04	0.00	0.96	C
30	1	12.31	C	-4.32	-3.43	0.01	0.03	0.96	C
21	1	14.13	C	-1.78	-3.16	0.14	0.04	0.83	C
52	1	10.04	C	-12.56	-0.69	0.00	0.33	0.67	C
50	0	8.90	C	-10.72	1.92	0.00	0.87	0.13	B

Matrice de confusion

Nombre de type	Prédi			
type.credit	A	B	C	Total
A	9		1	10
B		9	1	10
C	2	1	7	10
Total	11	10	9	30

Taux d'erreur	0.167
Taux de succès	0.833

Fig. 13.1. Formule de crédit - Construction de la matrice de confusion

$$\epsilon = \frac{0+0+1+0+1+2+1}{30} = \frac{5}{30} = 0.167 \tag{13.2}$$

$$\theta = \frac{9+9+7}{30} = 1 - \frac{5}{30} = 0.833 \tag{13.3}$$

Pour un nouveau gogo (un client dans la terminologie des crédits revolvers) qui se présente au guichet des emprunts, il y a 83.3% de chances qu'on l'aiguille vers la formule appropriée si nous utilisons le modèle issu de la régression logistique.

13.3 Indicateurs synthétiques pour le rappel et la précision

13.3.1 Rappel et précision par catégorie

Le rappel r et la précision p sont des indicateurs très populaires car leurs interprétations sont simples à appréhender. Le premier indique la capacité du modèle à retrouver les positifs, le second, la capacité à les prédire (désigner) avec justesse. Nous pouvons les associer aux catégories dans le cadre multi-classes, pour le rappel de y_k

$$r_k = \frac{n_{kk}}{n_{k.}} \tag{13.4}$$

et la précision (*accuracy* en anglais)

$$a_k = \frac{n_{kk}}{n_{.k}} \tag{13.5}$$

Pour notre exemple "Formule de crédit", nous aurions ainsi pour chaque modalité de la variable à prédire (Figure 13.1)

Catégorie	Rappel	Précision
A	$r_A = \frac{9}{10} = 0.9$	$a_A = \frac{9}{11} = 0.8182$
B	$r_B = \frac{9}{10} = 0.9$	$a_B = \frac{9}{10} = 0.9$
C	$r_C = \frac{7}{10} = 0.7$	$a_C = \frac{7}{9} = 0.7778$

Les informations sont précieuses. Nous pouvons caractériser la prédiction pour chaque classe. Nous notons dans notre fichier que la modalité C est moins bien détectée que les autres, et lorsque nous la prédisons, la précision est moindre.

Cela est intéressant, mais manipuler simultanément plusieurs indicateurs est toujours délicat. Il nous faut un indicateur synthétique pour quantifier les performances globales du modèle. Dans le cadre de la recherche d'information, plus précisément la catégorisation automatique de textes, des propositions ont été faites [2].

13.3.2 Microaveraging et macroaveraging

La *microaveraging* (micro-moyenne) est une moyenne pondérée où les catégories pèsent selon leur effectif dans le tableau de contingence. On accorde le même poids aux observations. Il est produit directement via la matrice de confusion (Tableau 13.1).

La *macroaveraging* (macro-moyenne) est une moyenne non-pondérée où l'on accorde le même poids aux catégories. Nous pouvons le produire directement via les rappels et précisions obtenues pour les catégories.

Lorsque les prévalences des modalités de la variable dépendante sont très différentes, ces deux ratios peuvent diverger assez fortement. A nous de choisir le bon selon les objectifs de l'étude. La micro-moyenne met l'accent sur les modalités fréquentes, la macro-moyenne accorde plus d'importance à celles qui sont peu fréquentes.

	Microaveraging	Macroaveraging
Rappel	$\mu_r = \frac{\sum_{k=1}^{K} n_{kk}}{\sum_{k=1}^{K} n_{k.}}$	$\rho_r = \frac{\sum_{k=1}^{K} r_k}{K}$
Précision	$\mu_a = \frac{\sum_{k=1}^{K} n_{kk}}{\sum_{k=1}^{K} n_{.k}}$	$\rho_a = \frac{\sum_{k=1}^{K} a_k}{K}$

Tableau 13.2. Microaveraging et macroaveraging

Les définitions numériques sont résumées dans le tableau 13.2. Nous noterons que les micro-moyennes pour le rappel et la précision produiront la même valeur : le taux de succès. Leur intérêt est très limité.

Appliquées sur le fichier "Formules de crédit", nous obtenons

2. F. Sebastiani, *Text Categorization*, in A. Zanasi (ed.), Text Mining ans its Applications, WIT Press, 2004.

	Microaveraging	Macroaveraging
Rappel	$\mu_r = \frac{9+9+7}{10+10+10} = 0.8333$	$\rho_r = \frac{0.9+0.9+0.7}{3} = 0.8333$
Précision	$\mu_a = \frac{9+9+7}{11+10+9} = 0.8333$	$\rho_a = \frac{0.8182+0.9+0.7778}{3} = 0.8320$

La situation est un peu particulière. En effet, les prévalences des catégories sont strictement identiques dans notre fichier ($n_k = \frac{n}{K},\ \forall k$). C'est pour cela que $\mu_r = \rho_r$.

13.4 Taux d'erreur et échantillon non représentatif

Lorsque l'échantillon de test n'est pas représentatif, le taux d'erreur n'est pas transposable à la population. Il ne correspond pas à la probabilité de mauvais classement du modèle. Comme dans le cadre binaire (section 10.1.4), il nous faut le corriger en utilisant les "vraies" prévalences p_k des catégories dans la population.

La formulation est simple, il faut généraliser l'expression mis en avant dans la section 2.1.2,

$$\epsilon = \sum_{k=1}^{K} p_k \times (1 - r_k) \tag{13.6}$$

Exemple : Formules de crédit. Mettons que notre échantillon a été volontairement équilibré par l'analyste. Nous savons par ailleurs que les "vraies" proportions des formules demandées dans la population est ($p_A = 0.15; p_B = 0.25; p_C = 0.6$). Pour obtenir le véritable taux d'erreur du classifieur, nous formons à partir des valeurs fournies par la matrice de confusion (Figure 13.1)

$$\begin{aligned}\epsilon &= 0.15 \times (1 - \frac{9}{10}) + 0.25 \times (1 - \frac{9}{10}) + 0.60 \times (1 - \frac{7}{10}) \\ &= 0.15 \times \frac{1}{10} + 0.25 \times \frac{1}{10} + 0.6 \times \frac{3}{10} \\ &= 0.22\end{aligned}$$

Par rapport au taux d'erreur mesuré sans précautions particulières sur un fichier volontairement équilibré (0.167), nous constatons que la "vraie" probabilité de se tromper avec le modèle serait plutôt de 0.22. La valeur est plus élevée parce que la pondération ($p_C = 0.6$) met l'accent sur la catégorie la moins bien reconnue ($r_C = 0.7$) dans notre exemple.

Remarque : Lorsque l'échantillon est représentatif, nous pouvons estimer p_k par $\hat{p}_k = \frac{n_{k.}}{n}$, voyons ce qu'il advient de l'expression ci-dessus (équation 13.6)

$$\begin{aligned}\epsilon &= \sum_k \hat{p}_k \times (1 - r_k) \\ &= \sum_k \frac{n_{k.}}{n} \times (1 - \frac{n_{kk}}{n_{k.}}) \\ &= \sum_k \left(\frac{n_{k.}}{n} - \frac{n_{kk}}{n}\right) \\ &= \sum_k \frac{n_{k.}}{n} - \sum_k \frac{n_{kk}}{n} \\ &= 1 - \sum_k \frac{n_{kk}}{n}\end{aligned}$$

Nous avons la forme usuelle du taux d'erreur (équation 13.1).

13.5 Intégrer les coûts de mauvais classement

L'intégration des coûts de mauvais classement a beaucoup été étudiée dans le cadre binaire, notamment lors de l'évaluation (section 10.2.2). Pour une variable dépendante à K modalités, le coût moyen de mauvais classement s'écrit

$$\zeta(M) = \frac{1}{n} \sum_{k=1}^{K} \sum_{l=1}^{K} n_{kl} \times c(k,l) \tag{13.7}$$

où $c(k,l)$ est le coût associé à la prédiction y_l alors que la vraie classe d'appartenance de l'individu est y_k (section 10.2.1).

Exemple : Formules de crédit. Mettons que la matrice de coûts de mauvais classement s'écrit comme suit dans notre problème d'affectation automatique de formules de crédit [3]

	$\hat{y}_A$	$\hat{y}_B$	$\hat{y}_C$
y_A	-5	3	10
y_B	4	-6	10
y_C	0	8	-1

En faisant la somme des produits croisés entre la matrice de confusion (Figure 13.1) et cette matrice de coûts, nous obtenons

$$\zeta(M) = \frac{1}{30}\left[9 \times (-5) + 0 \times 3 + \cdots + 1 \times 8 + 7 \times (-1)\right] = -2.6$$

Remarque : Dans le cadre multi-classes également, si nous utilisons une matrice de coûts symétrique et unitaire ($c(k,l) = 1,\ k \neq l\,;\ c(k,k) = 0$), nous retrouvons le taux d'erreur.

3. Les chiffres ont été mis un peu au hasard, il s'agit d'un simple exemple illustratif. Pour une définition un peu plus circonstanciée des coûts dans un problème réel, voir J.H. Chauchat, R. Rakotomalala, M. Carloz, C. Pelletier, *Targeting Customer Groups using Gain and Cost Matrix : a Marketing Application*, `http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/WS-Proceedings/w10/chauchat_workshop.pdf` ; voir aussi un de nos tutoriel relatif au concours *Data Mining Cup - 2007*, `http://tutoriels-data-mining.blogspot.com/2009/01/cots-de-mauvais-classement-en.html`

14

Tester les coefficients de la régression multinomiale

Les tests sur les coefficients consistent avant tout à éprouver leur significativité. Par rapport à la régression binaire, l'analyse est plus compliquée car nous pouvons multiplier les possibilités : tester la nullité de q coefficients dans un logit, dans un ensemble de logit ou dans les $K-1$ logit. Les conséquences ne sont pas les mêmes. Si une variable n'est pas significative dans l'ensemble des logit, nous pouvons l'exclure de l'étude. Si elle est significative dans un logit au moins, son rôle est avéré dans la caractérisation d'une des modalités de la variable dépendante. La variable ne peut pas être exclue.

Autre aspect intéressant, nous pouvons être amenés à tester l'égalité des coefficients pour plusieurs (ou l'ensemble des) équations logit. Cela ne préjuge en rien de leur significativité. Si l'hypothèse est vérifiée, on dira simplement que la variable joue un rôle identique dans la caractérisation des différentes modalités de la variable dépendante.

Comme pour la régression binaire, nous disposons de deux outils pour réaliser les tests. La statistique du rapport de vraisemblance correspond toujours à la comparaison des déviances des régressions sous H_0 et H_1. Elle suit une loi du χ^2 sous l'hypothèse nulle. Les degrés de liberté sont obtenus par différenciation du nombre de paramètres estimés. Pour rappel, dans notre exemple "Formules de crédit", la déviance du modèle complet, celui où tous les coefficients sont estimés, est $D_M = 18.382$ avec un degré de liberté de $ddl = 30 - 2 \times 8 = 22$.

La statistique de Wald exploite la normalité asymptotique des estimateurs du maximum de vraisemblance. Nous devons au préalable calculer la matrice de variance de covariance des coefficients qui est un peu plus complexe puisque nous en manipulons simultanément $(K-1) \times (J+1)$. La statistique suit une loi du χ^2, le nombre de degrés de liberté est égal au nombre de contraintes que l'on pose sur les coefficients sous l'hypothèse nulle. Cela apparaît clairement lorsque nous nous pencherons sur l'écriture généralisée.

Enfin, les commentaires émis sur ces tests précédemment (section 3.4) restent valables : le test du rapport de vraisemblance est plus puissant, il détecte mieux l'hypothèse alternative lorsqu'elle est vraie, il est préférable sur les petits effectifs ; le test de Wald est très conservateur ; les deux se rejoignent lorsque le nombre d'observations devient élevé.

14.1 Estimation de la matrice de variance covariance

La matrice de variance covariance est une pièce essentielle de la statistique inférentielle. Concernant la régression logistique, elle nous permettra de mettre en place les tests de Wald. Nous pourrions en tirer parti également pour la production des intervalles de confiance des coefficients et des prédictions.

La matrice de variance covariance $\hat{\Sigma}$ correspond à l'inverse de la matrice hessienne. Elle est aussi symétrique par blocs. Il faut bien faire attention pour discerner les informations importantes qu'elles comportent : nous avons la variance des coefficients pour chaque équation logit, les covariances entre coefficients de la même équation logit, et les covariances des coefficients relatives à des équations logit différentes. On peut s'y perdre rapidement.

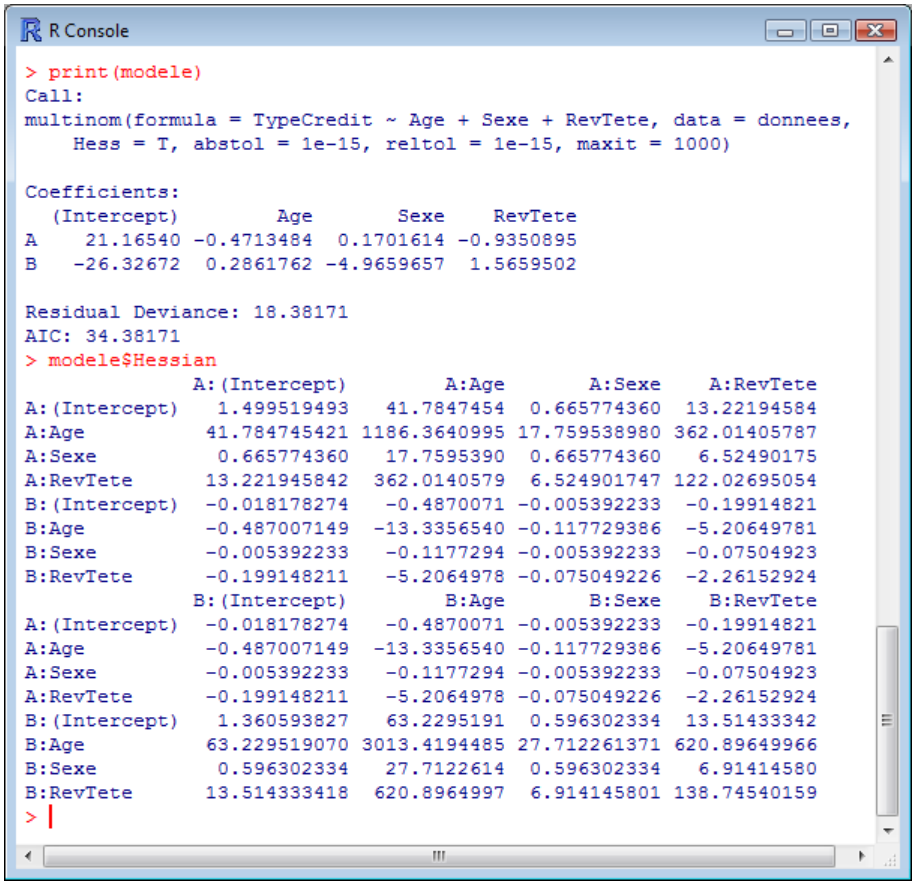


Fig. 14.1. Formule de crédit - Obtention de la matrice hessienne avec **multinom** de **R**

Pour le fichier "Formules de crédit", la matrice hessienne est accessible via un des champs de l'objet fourni par la fonction **multinom** du package **nnet** de **R** (Figure 14.1). Elle est de taille $[(K-1)*(J+1)\times(K-1)*(J+1)]$, soit 8×8. Nous calculons son inverse (Figure 14.2). Essayons d'y discerner les informations importantes :

Matrice Hessienne

	A:(Intercept	A:Age	A:Sexe	A:RevTete	B:(Intercept	B:Age	B:Sexe	B:RevTete
A:(Intercept)	1.500	41.785	0.666	13.222	-0.018	-0.487	-0.005	-0.199
A:Age	41.785	1186.364	17.760	362.014	-0.487	-13.336	-0.118	-5.206
A:Sexe	0.666	17.760	0.666	6.525	-0.005	-0.118	-0.005	-0.075
A:RevTete	13.222	362.014	6.525	122.027	-0.199	-5.206	-0.075	-2.262
B:(Intercept)	-0.018	-0.487	-0.005	-0.199	1.361	63.230	0.596	13.514
B:Age	-0.487	-13.336	-0.118	-5.206	63.230	3013.419	27.712	620.896
B:Sexe	-0.005	-0.118	-0.005	-0.075	0.596	27.712	0.596	6.914
B:RevTete	-0.199	-5.206	-0.075	-2.262	13.514	620.896	6.914	138.745

Matrice de variance covariance des coefficients

	A:(Intercept	A:Age	A:Sexe	A:RevTete	B:(Intercept	B:Age	B:Sexe	B:RevTete
A:(Intercept)	**117.545**	-2.649	1.946	-4.984	-1.964	0.035	-0.015	0.023
A:Age	-2.649	**0.069**	0.005	0.081	0.028	0.000	0.000	-0.001
A:Sexe	1.946	0.005	**3.435**	-0.409	-0.462	0.007	-0.004	0.011
A:RevTete	-4.984	0.081	-0.409	**0.329**	0.184	-0.003	0.004	-0.002
B:(Intercept)	-1.964	0.028	-0.462	0.184	**220.268**	-2.008	41.744	-14.548
B:Age	0.035	0.000	0.007	-0.003	-2.008	**0.023**	-0.314	0.106
B:Sexe	-0.015	0.000	-0.004	0.004	41.744	-0.314	**12.739**	-3.296
B:RevTete	0.023	-0.001	0.011	-0.002	-14.548	0.106	-3.296	**1.114**

Ecart type des coefficients

	A:(Intercept	A:Age	A:Sexe	A:RevTete	B:(Intercept	B:Age	B:Sexe	B:RevTete
A:(Intercept)	**10.842**							
A:Age		**0.263**						
A:Sexe			**1.853**					
A:RevTete				**0.574**				
B:(Intercept)					**14.841**			
B:Age						**0.153**		
B:Sexe							**3.569**	
B:RevTete								**1.055**

Fig. 14.2. Formule de crédit - Calcul de la matrice de variance covariance, inverse de la matrice Hessienne

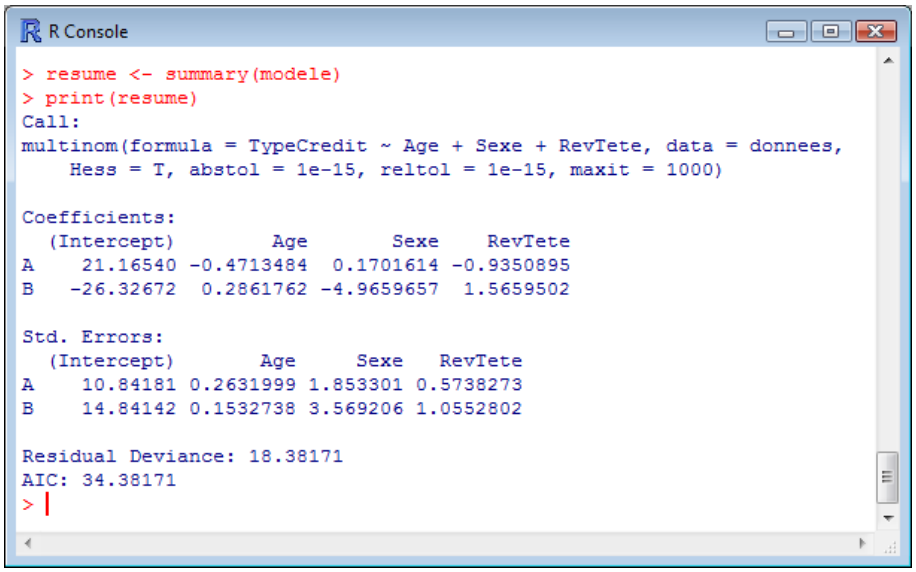


Fig. 14.3. Formule de crédit - Coefficients et écarts-type des coefficients avec R

– Les variances des coefficients, pour chaque équation logit sont lues sur la diagonale principale de la matrice. En prenant la racine carrée, nous obtenons les écarts-type fournis par les logiciels : ceux de **multinom** (Figure 14.3) ; ou ceux de Tanagra (Figure 12.3) [1].

1. Les estimations sont très légèrement différentes, c'est normal puisque les techniques d'optimisation utilisées ne sont pas les mêmes.

- Dans les blocs situés sur la diagonale principale ($H_{k,k}$), nous avons les covariances des coefficients intra-logit. Ex. $\widehat{cov}(\hat{a}_{1,age};\hat{a}_{1,sexe}) = 0.005$
- Dans les blocs hors diagonale ($H_{k,l}, k \neq l$), nous avons les covariances des coefficients inter-logit. Ex. $\widehat{cov}(\hat{a}_{1,age};\hat{a}_{2,rev.tete}) = -0.001$, qui est différent de $\widehat{cov}(\hat{a}_{1,rev.tete};\hat{a}_{2,age}) = -0.003$

Nous sommes maintenant parés pour réaliser tous les tests que l'on veut. Nous fixons le risque de première espèce à 10% pour tous les exemples traités.

14.2 Significativité d'un coefficient dans un logit

L'hypothèse nulle de ce test s'écrit

$$H_0 : a_{j,k} = 0$$

Un coefficient dans un des logit est-il significatif ? Si la réponse est non, il ne l'est pas, nous pouvons supprimer la variable associée dans le logit concerné. Nous ne pouvons rien conclure en revanche concernant les autres logit. Nous ne pouvons donc pas exclure la variable de l'étude.

14.2.1 Test du rapport de vraisemblance

Pour ce test, il s'agit d'optimiser la vraisemblance en forçant $a_{j,k} = 0$. Nous obtenons le modèle contraint (modèle sous H_0), d'en extraire la déviance D_{H_0}, que l'on comparera à celle du modèle complet D_M. La statistique de test

$$LR = D_{H_0} - D_M$$

suit une loi du χ^2 à 1 degré de liberté.

Dans notre exemple **"Formule de crédit"**, nous souhaitons tester la significativité du coefficient associé à la variable *rev.tete* dans la première équation logit (A vs. C). Nous lançons le solveur dans Excel, après avoir fixé sa valeur à 0 et en l'excluant des cellules variables pour l'optimisation. Nous obtenons un nouveau jeu de coefficients et $D_{H_0} = 24.839$, avec un degré de liberté $ddl = 30 - 7 = 23$ (Figure 14.4). Nous en déduisons la statistique de test

$$LR = 24.839 - 18.382 = 6.457$$

Avec un χ^2 à 1 degré de liberté, la probabilité critique est p-value = 0.0110. Nous concluons que la variable est significative au risque 10%.

	const	age	sexe	rev.tete
a(A/C)	10.468	-0.354	-1.108	0.000
a(B/C)	-24.728	0.254	-4.923	1.561

Deviance (H0)	24.839
ddl	23

Déviance (Modèle complet)	18.382
ddl	22

LR	6.457
ddl	1
p-value	0.0110

Fig. 14.4. Test du rapport de vraisemblance - Tester la significativité de *rev.tete* dans le 1^{er} logit

14.2.2 Test de Wald

La statistique de Wald est formé par le rapport entre le carré du coefficient et sa variance,

$$W_{k,j} = \frac{\hat{a}_{k,j}^2}{\hat{\sigma}_{\hat{a}_{k,j}}}$$

Elle suit une loi du χ^2 à 1 degré de liberté.

Toujours concernant *rev.tete* dans le premier logit, nous formons à partir des résultats glanés tout au long de ce chapitre (coefficient, figure 14.3; variance, figure14.2)

$$W_{1,rev.tete} = \frac{(-0.935)^2}{0.329} = 2.655$$

Avec un χ^2 à 1 degré de liberté, nous avons une p-value $= 0.103$. Nous sommes à la lisière de la région critique. Il n'en reste pas moins qu'au risque 10%, nous ne pouvons pas rejeter l'hypothèse nulle.

Encore une fois, le test de Wald s'avère conservateur en comparaison du test du rapport de vraisemblance où l'hypothèse nulle était clairement rejetée.

14.3 Significativité d'un coefficient dans tous les logit

L'hypothèse nulle du test s'écrit

$$H_0 : a_{k,j} = 0, \ \forall k$$

Il va plus loin que le précédent. Il cherche à savoir si les coefficients d'une variable explicative sont simultanément nuls dans l'ensemble des logit. Si les données sont compatibles avec H_0, nous pouvons la retirer du modèle.

	const	age	sexe	rev.tete
a(A/C)	10.323	-0.349	-1.084	0.000
a(B/C)	-11.060	0.257	-0.379	0.000

Deviance (H0)	30.874
ddl	24

Déviance (Modèle complet)	18.382
ddl	22

LR	12.492
ddl	2
p-value	0.0019

Fig. 14.5. Test du rapport de vraisemblance - Tester la significativité de *rev.tete* dans l'ensemble des logit

14.3.1 Test du rapport de vraisemblance

Le principe est toujours le même, nous calculons la déviance du modèle contraint et nous la comparons à celle du modèle complet. La statistique suit une loi du χ^2 à $(K-1)$ degrés de liberté.

Nous souhaitons savoir si les coefficients de *rev.tete* sont simultanément nuls dans toutes les équations logit. Nous fixons les cellules appropriés à 0 dans la feuille Excel, nous lançons le solveur en les excluant des cellules variables. La déviance du nouveau modèle est $D_{H_0} = 30.874$ avec des degrés de liberté $ddl = 24$ (Figure 14.5). La statistique est égale à

$$LR = 30.874 - 18.382 = 12.492$$

Avec un $\chi^2(2)$, nous avons une p-value $= 0.0019$. Nous rejetons l'hypothèse nulle, les coefficients ne sont pas simultanément nuls dans l'ensemble des logit.

14.3.2 Test de Wald

La statistique de test suit une loi du χ^2 à $(K-1)$ degrés de liberté sous H_0, elle s'écrit

$$W_j = \hat{a}_j' \hat{\Sigma}_j^{-1} \hat{a}_j$$

$\hat{a}_j$ est le vecteur des coefficients à évaluer, de dimension $(K-1) \times 1$; $\hat{\Sigma}_j$ est leur matrice de variance covariance. Tout l'enjeu est de savoir lire correctement la matrice de variance covariance globale $\hat{\Sigma}$ pour y "piocher" les valeurs de $\hat{\Sigma}_j$.

Pour notre exemple *rev.tete*,

$$\hat{a}_{rev.tete} = \begin{pmatrix} -0.935 \\ 1.566 \end{pmatrix}$$

et, en piochant dans la matrice de variance covariance (Figure 14.2),

$$\hat{\Sigma}_{rev.tete} = \begin{pmatrix} 0.329 & -0.002 \\ -0.002 & 1.114 \end{pmatrix}$$

Nous formons

$$\begin{aligned} W_{rev.tete} &= \hat{a}'_{rev.tete}\hat{\Sigma}^{-1}_{rev.tete}\hat{a}_{rev.tete} \\ &= \begin{pmatrix} -0.935 & 1.566 \end{pmatrix} \begin{pmatrix} 0.329 & -0.002 \\ -0.002 & 1.114 \end{pmatrix}^{-1} \begin{pmatrix} -0.935 \\ 1.566 \end{pmatrix} \\ &= \begin{pmatrix} -0.935 & 1.566 \end{pmatrix} \begin{pmatrix} 3.037 & 0.004 \\ 0.004 & 0.898 \end{pmatrix} \begin{pmatrix} -0.935 \\ 1.566 \end{pmatrix} \\ &= 4.845 \end{aligned}$$

Avec un $\chi^2(2)$, nous avons une p-value de 0.089. Nous rejetons l'hypothèse nulle au risque 10%.

Ce résultat doit nous interpeller. En effet, testés individuellement dans chaque équation logit, les coefficients de *rev.tete* ne sont pas significatifs, comme en attestent les résultats fournis par Tanagra (Figure 12.3). En revanche, testés simultanément, nous rejetons l'hypothèse nulle. Un test simultané ne peut pas être réduit en une succession de tests individuels.

14.4 Test d'égalité d'un coefficient dans tous les logit

Nous souhaitons savoir si les coefficients d'une variable X_j sont identiques d'un logit à l'autre. L'hypothèse nulle s'écrit

$$H_0 : a_{1,j} = \cdots = a_{K-1,j}$$

Lorsqu'elle est compatible avec les données, cela veut dire que la variable a le même impact dans tous les logit. Il n'est pas question en revanche de la supprimer de la régression si elle est par ailleurs significative : son impact est le même, mais il n'est pas nul.

14.4.1 Test du rapport de vraisemblance

Définir le modèle contraint dans les logiciels de statistique n'est pas très facile. Le couple tableur-solveur se révèle redoutable dans ce contexte. Nous souhaitons savoir si le coefficient de *rev.tete* est identique d'un logit à l'autre dans le fichier "Formule de crédit". L'astuce est relativement simple. Nous introduisons un des coefficients parmi les cellules variables du solveur. Pour les autres, nous forçons l'égalité. Prenons un exemple concret dans notre feuille Excel (Figure 14.6) : nous incluons la cellule de *rev.tete* du premier logit parmi les cellules variables du solveur (en **H3**), pour le second coefficient (en

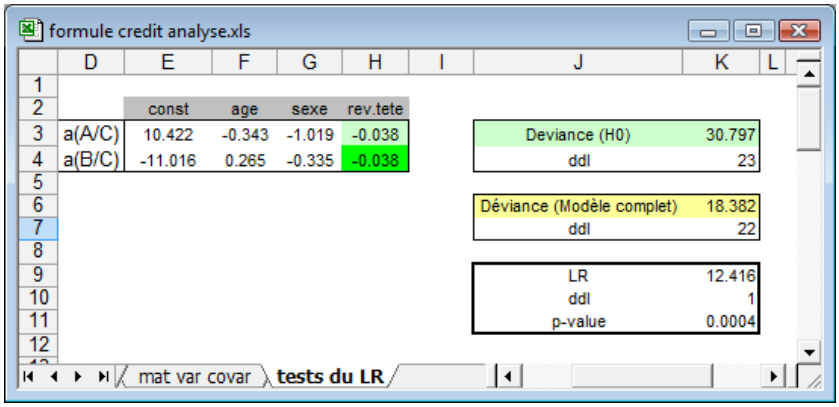


Fig. 14.6. Test du rapport de vraisemblance - Égalité des coefficients de *rev.tete* dans l'ensemble des logit

H4), nous introduisons simplement la formule =**H3**. Ainsi, lorsque nous lançons l'optimisation de la log-vraisemblance, cette contrainte est bien prise en compte[2].

La déviance du modèle contraint est $D_{H_0} = 30.797$, avec un degré de liberté égal à $ddl = 30 - 7 = 23$, le 8^e coefficient étant simplement déduit du 7^e. La statistique du test est égal à

$$LR = D_{H_0} - D_M = 30.797 - 18.382 = 12.416$$

Avec un χ^2 à $23 - 22 = 1$ degré le liberté, la p-value $= 0.0004$. Nous rejetons l'égalité des coefficients dans l'ensemble des logit.

14.4.2 Test de Wald - Calcul direct

Partons directement sur notre exemple pour expliciter la démarche. L'hypothèse nulle du test H_0 : $a_{1,rev.tete} = a_{2,rev.tete}$ peut s'écrire $d_{rev.tete} = a_{1,rev.tete} - a_{2,rev.tete} = 0$. La statistique de test est

$$\hat{d}_{rev.tete} = \hat{a}_{1,rev.tete} - \hat{a}_{2,rev.tete}$$

Elle d'espérance nulle sous H_0, et de variance [9] (page 268)

$$\hat{V}(\hat{d}_{rev.tete}) = \hat{V}(\hat{a}_{1,rev.tete}) + \hat{V}(\hat{a}_{2,rev.tete}) - 2 \times \widehat{COV}(\hat{a}_{1,rev.tete}, \hat{a}_{2,rev.tete})$$

Sous H_0,

$$\frac{\hat{d}^2_{rev.tete}}{\hat{V}(\hat{d}_{rev.tete})}$$

suit une loi du χ^2 à 1 degré de liberté.

2. Une autre stratégie aurait été de mettre tous les coefficients en cellules variables, puis d'ajouter la contrainte **H3=H4**

Introduisons les valeurs numériques :

$$\hat{d}_{rev.tete} = -0.935 - 1.566 = -2.501$$

Il faut aller à la pêche dans la matrice de variance covariance pour obtenir la variance (Figure 14.2) de $\hat{d}_{rev.tete}$

$$\hat{V}(\hat{d}_{rev.tete}) = 0.329 + 1.114 - 2 \times (-0.002) = 1.446$$

Il ne reste plus qu'à former le rapport

$$\frac{\hat{d}^2_{rev.tete}}{\hat{V}(\hat{d}_{rev.tete})} = \frac{(-2.501)^2}{1.446} = 4.326$$

Avec un $\chi^2(1)$, la p-value est 0.0375.

Nous rejetons l'égalité des coefficients de *rev.tete* dans l'ensemble des logit.

14.4.3 Test de Wald - Calcul générique

Lorsque le nombre d'équations logit est supérieur à 2, l'affaire devient plus compliquée. Il paraît plus judicieux de passer par l'écriture générique des tests (section 3.3.6). La gageure est d'écrire correctement la matrice M.

Pour l'exemple qui nous concerne ($H_0 : a_{1,rev.tete} = a_{2,rev.tete} \Leftrightarrow a_{1,rev.tete} - a_{2,rev.tete} = 0$), M est un matrice avec $m = 1$ ligne et $(K-1) \times (J+1) = 2 \times 4 = 8$ colonnes. Elle s'écrit comme suit

$$M = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

Avec

$$\hat{a}' = \begin{pmatrix} 21.165 & -0.471 & 0.170 & -0.935 & -26.328 & 0.286 & -4.966 & 1.566 \end{pmatrix}$$

Nous formons la statistique de test conformément à l'équation 3.7 dans une feuille Excel (Figure 14.7), nous obtenons la statistique de test

$$W_{(M)} = \hat{a}' M' \times \left[M \hat{\Sigma} M' \right]^{-1} \times M\hat{a} \tag{14.1}$$

$$= (-2.501) \times 0.692 \times (-2.501) \tag{14.2}$$

$$= 4.236 \tag{14.3}$$

Nous retrouvons exactement la même valeur qu'avec l'approche directe. Heureusement, le contraire eut été inquiétant. Bien évidemment, les conclusions sont identiques.

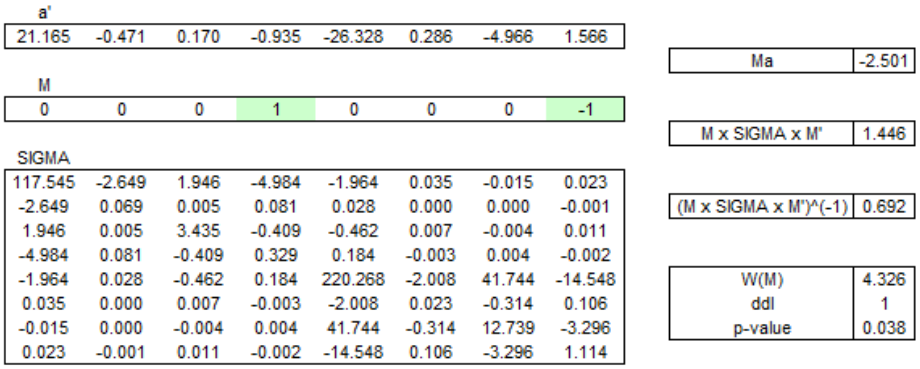


Fig. 14.7. Test de Wald - Approche générique - Égalité des coefficients de *rev.tete* dans l'ensemble des logit

14.5 Interprétation des coefficients - Les odds-ratio

Le charme de la régression logistique repose en partie sur les interprétations des coefficients sous forme de (log) odds-ratio. Voyons si cette propriété est préservée dans la régression multinomiale, et si c'est le cas, comment lire les coefficients des variables dans les logit. En effet, la nouveauté est qu'une même variable peut être présente plusieurs fois, avec des valeurs différentes, dans $K-1$ équations.

Pour illustrer notre propos, nous utiliserons le fichier BRAND, il s'agit de prédire le choix de marques de $n = 735$ clients à partir de leur genre (sexe $= 1 \rightarrow$ femme). Nous nous focaliserons principalement sur les variables binaires dans cette section. L'interprétation est liée à la présence/absence du caractère. La transposition aux variables quantitatives ne pose pas de problème particulier. L'interprétation est relative à l'augmentation d'une unité de l'explicative, comme nous avions pu le mettre en exergue dans la régression binaire.

14.5.1 Calcul de l'odds-ratio via le tableau de contingence

La variable dépendante BRAND prend 3 modalités : "petit prix" ($y_1 = 1$), "enseigne" ($y_2 = 2$) et "référence" ($y_3 = 3$). Il s'agit d'expliquer le choix des clients, *en les caractérisant par rapport à la marque de référence*. Voyons dans un premier temps comment calculer les odds-ratio à partir d'un tableau de contingence (Figure 14.8).

Tous les calculs doivent être organisés par rapport à la modalité de référence y_3. Concernant les odds :

- Les femmes ont $\frac{1}{0.80} = 1.24$ fois plus de chances de choisir la référence que la marque "petit prix", en effet

$$odds(1/3;1) = \frac{115}{143} = 0.80$$

- Elles sont 1.45 fois plus de chances de choisir la marque "enseigne" que la référence, car

$$odds(2/3;1) = \frac{208}{143} = 1.45$$

Nombre de brand	femme		
brand	1	0	Total
M__PetitPrix	115	92	207
M_Enseigne	208	99	307
M_Reference	143	78	221
Total	466	269	735

Odds (par rapport à référence)

M__PetitPrix	0.80	1.18
M_Enseigne	1.45	1.27

Odds-ratio (référence)

M__PetitPrix	0.68
M_Enseigne	1.15

Fig. 14.8. Fichier BRAND - Calcul des odds-ratio à partir d'un tableau de contingence

– Nous pouvons faire de même du côté des hommes. Ainsi, nous constatons qu'ils ont 1.18 fois plus de chances de choisir la marque "petit prix" (par rapport à la référence)

$$odds(1/3;0) = \frac{92}{78} = 1.18$$

Il faut faire le rapport des odds pour obtenir les **odds-ratio**, nous aurons

$$OR(1/3) = \frac{odds(1/3;1)}{odds(1/3;0)} = \frac{0.80}{1.18} = 0.68$$

Les femmes ont $\frac{1}{0.68} = 1.47$ fois plus de chances de choisir la marque de référence (par rapport à "petit prix") que les hommes. La lecture n'est pas très aisée. En clair, placées devant l'alternative "petit prix" - "référence", les femmes ont plus tendance à choisir la marque de référence que les hommes.

De même

$$OR(2/3) = \frac{odds(2/3;1)}{odds(2/3;0)} = \frac{1.45}{1.27} = 1.15$$

Les femmes ont tendance à préférer la marque enseigne à la référence par rapport aux hommes.

L'enjeu maintenant est de pouvoir retrouver ces coefficients avec la régression logistique.

14.5.2 Obtention des odds-ratio via la régression logistique

Nous avons lancé la régression logistique BRAND = f(sexe) dans Tanagra. Nous obtenons les équations logit (Figure 14.9)

$$C_1 = C(1/3) = 0.16508 - 0.38299 \times femme$$
$$C_2 = C(2/3) = 0.238411 + 0.13628 \times femme$$

Attributes in the equation

Class.Value	M__PetitPrix				M_Enseigne			
Pred.Att.	Coef.	Std.Err	Wald	p-value	Coef.	Std.Err	Wald	p-value
constant	0.16508	-	-	-	0.238411	-	-	-
femme	**-0.382992**	0.1984	3.725	0.0536	**0.136282**	0.1863	0.5349	0.4646

Fig. 14.9. Fichier BRAND - Coefficients de la régression logistique $brand = f(femme)$

Si nous prenons les exponentielles des coefficients associés à la variable $sexe = femme$

$$e^{\hat{a}_{1,femme}} = e^{-0.38299} = 0.68 = OR(1/3)$$
$$e^{\hat{a}_{2,femme}} = e^{0.13628} = 1.15 = OR(2/3)$$

Nous retrouvons les odds-ratio calculés à partir du tableau de contingence.

En conclusion, nous dirons :

- Les interprétations en termes de surcroît de risque (log odds-ratio) des coefficients de la régression logistique restent valables dans le cadre multinomial.
- Mais ils sont comptabilisés par rapport à la catégorie de référence. Il ne faut jamais l'oublier. Si nous souhaitons la modifier, il faut procéder par différenciation des logit (section 12.3.4). Les nouveaux coefficients se liront en relation avec la nouvelle référence.
- Avec les résultats de la régression, nous savons si les odds-ratio sont significatifs ou pas. Dans notre exemple, au risque 10%, nous avons que $OR(1/3)$ est significativement différent de 1 parce que $\hat{a}_{1,femme}$ est significativement différent de 0 (p-value = 0.0536) ; pas $OR(2/3)$ (p-value = 0.4646).
- Pour les autres types de variables explicatives (nominale à + de 2 modalités, ordinale, quantitative), les interprétations vues pour la régression logistique binaire restent valables, elles doivent être lues simplement par rapport à la catégorie de référence toujours.
- Les exponentielles des constantes se lisent comme des odds de la modalité complémentaire de la variable explicative binaire. Pour notre exemple, nous avons les odds chez les hommes (femme = 0)

$$e^{\hat{a}_{1,const}} = e^{0.16508} = 1.18 = odds(1/3;0)$$
$$e^{\hat{a}_{2,const}} = e^{0.238411} = 1.27 = odds(2/3;0)$$

15

S'appuyer sur des régression binaires séparées

La régression logistique binaire propose une série d'outils pour diagnostiquer, valider, explorer des solutions (analyse des résidus, sélection de variables, etc.). Ils pourraient être transposés sans aucune difficulté à la régression multinomiale. Pourtant, curieusement, ils ne sont pas implémentés dans les logiciels usuels. Ne serait-ce que la sélection de variables. Nous devrions pouvoir évaluer la pertinence des explicatives dans l'ensemble des logit pour les retirer unes à unes pour un processus backward basé par sur le test de Wald. L'idée est simple, sa réalisation également, pourtant nous la retrouvons pas dans les logiciels les plus répandus [9] (page 277).

Dans ce contexte, on se demande s'il n'est pas possible de décomposer la régression multinomiale en une série de régressions binaires indépendantes où l'on opposerait chaque modalité ($k = 1, \ldots, K-1$) de Y à la modalité de référence y_K [1] (page 310). Bien entendu, nous n'obtiendrons pas les mêmes résultats (coefficients). Le tout est de cerner jusqu'à quel point ils seront différents [1].

L'avantage de passer par cette solution est de pouvoir ainsi bénéficier des outils sus-mentionnés implémentés dans la très grande majorité des logiciels de statistique. Après il faut savoir quoi faire des résultats. En procédant à une sélection de variables dans chaque régression binaire, il est tout à fait possible que nous nous retrouvons avec des équations logit comportant des sous-ensembles solutions très dissemblables. De même, une observation peut être atypique pour une équation logit, mais pas pour les autres. Il faut savoir interpréter correctement ces éléments sans perdre de vue que nous souhaitons valider le modèle global expliquant simultanément les K valeurs de Y [9] (page 279).

On sait que décomposer la régression multinomiale en $K-1$ régressions binaires est moins efficace. Elle le sera d'autant moins que la prévalence de la catégorie de référence est faible. En l'absence de contraintes fortes sur les interprétations, nous avons intérêt à choisir une modalité de référence qui soit la plus fréquente dans la population, celle dont la prévalence $p_k = P(Y = y_k)$ est la plus élevée [1] (page 312). De manière générale, il apparaît que les coefficients obtenus via les deux stratégies sont assez proches [9] (page 278).

1. Cette situation n'est pas sans rappeler les problèmes posés par les méthodes binaires par essence en apprentissage automatique (ex. les support vector machine). Pour traiter les variables dépendantes multi-classes, des stratégies ont été développées pour combiner les prédicteurs binaires : une modalité contre les autres "1 vs. all", traitement par paires "1 vs. 1", etc. Voir S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, Elsevier, 2009 ; pages 127 et 128.

Le mieux est de le vérifier sur un exemple. Nous reprenons le fichier BRAND auquel nous avons adjoint la variable *age*. La régression s'écrit $brand = f(femme, age)$. Nous avons réalisé la régression multinomiale sur les $n = 735$ observations. Puis nous avons isolé les clients portant les modalités "Petit prix" (resp. "Enseigne") et "Reference". Le fichier comporte 428 (resp. 528) observations. Puis nous avons lancé les régressions binaires. Voici le code R correspondant

```
#régression logistique multinomiale - vgam
modele <- vglm(brand ~ femme + age, data = donnees, family =  multinomial())
print(modele)
#décomposition en régréssions individuelles
#1 vs. 3
donnees.1 <- donnees[(donnees$brand == "M__PetitPrix" | donnees$brand == "M_Reference"),]
donnees.1$brand <- as.factor(unclass(donnees.1$brand))
modele.1 <- glm(brand ~ femme + age, data = donnees.1, family = binomial)
print(modele.1)
#2 vs. 3
donnees.2 <- donnees[(donnees$brand == "M_Enseigne" | donnees$brand == "M_Reference"),]
donnees.2$brand <- as.factor(unclass(donnees.2$brand))
modele.2 <- glm(brand ~ femme + age, data = donnees.2, family = binomial())
print(modele.2)
```

Après réorganisation des signes, nous pouvons comparer les coefficients produits de la régression multinomiale et les régressions binaires (Tableau 15.1).

logit	Petit prix vs. Référence		Enseigne vs. Référence	
Régressions	Reg.Multinomiale	Reg.Binaire	Reg.Multinomiale	Reg.Binaire
constante	22.72	19.43	10.95	11.38
femme	−0.47	−0.39	0.06	0.04
age	−0.69	−0.59	−0.32	−0.33

Tableau 15.1. Coefficients de la régression multinomiale et des régressions binaires

Indéniablement, il y a une similitude entre les coefficients. Mais les écarts entre les valeurs sont néanmoins sensibles, du moins en ce qui concerne notre exemple.

Enfin, il reste un problème épineux : comment exploiter ces modèles en prédiction ? Nous n'avons plus la garantie que $\sum_k \hat{\pi}_k(\omega) = 1$. Il faut définir une stratégie appropriée pour combiner les $\hat{\pi}$ ou les $\hat{y}$ fournis par les $K - 1$ classifieurs binaires. Il n'y a pas de solution bien établie à vrai dire.

Partie IV

La régression logistique polytomique ordinale

16

Variable dépendante ordinale (1) - LOGITS adjacents

17

Variable dépendante ordinale (2) - ODDS-RATIO cumulatifs

A

Gestion des versions

La première version (version 1.0) de ce fascicule a été finalisée et mis en ligne le 13 septembre 2009. Il comprend 10 chapitres :

1. Régression Logistique Binaire - Principe et estimation
2. Évaluation de la régression
3. Tests de significativité des coefficients
4. Prédiction et intervalle de prédiction
5. Lecture et interprétation des coefficients
6. Analyse des interactions
7. La sélection de variables
8. Diagnostic de la régression logistique
9. "Covariate pattern" et statistiques associées
10. Redressement pour les échantillons non-représentatifs
11. Quelques éléments supplémentaires

Les parties dédiées à la régression multinomiale et à la régression polytomique ordinale ne sont pas commencées. Ce sera l'objet de la version 2.xx de ce document.

A.1 Version 1.1

Pour la version 1.1, le chapitre 10 a été remanié. Il intègre l'ancienne partie consacrée au redressement pour les échantillons non-représentatifs, et une nouvelle section consacrée à la prise en compte des coûts de mauvaise affectation. Le thème générique est la modification de la règle d'affectation dans des circonstances particulières.

La version a été mise en ligne le 16 septembre 2009.

A.2 Version 2.0

L'écriture de la partie III consacrée à la régression logistique multinomiale est la principale évolution dans la version 2.0. Elle comporte 4 chapitres

1. Variable dépendante nominale - Principe et estimations
2. Évaluation des classifieurs pour Y à K ($K > 2$) modalités
3. Tester les coefficients de la régression logistique multinomiale
4. S'appuyer sur des régression binaires séparées

La version a été mise en ligne le 22 septembre 2009.

B

Fichiers de données relatifs à ce fascicule

Pour que tout un chacun puisse reproduire à l'identique les exemples illustratifs, il faut que les données et les logiciels soient accessibles librement. C'est une règle à laquelle que j'astreindrai toujours. C'est valable pour les documents destinés à l'enseignement. Mais ça devrait l'être également pour les publications scientifiques.

S'agissant de ce fascicule de cours, les fichiers de données sont accessibles à l'adresse suivante `http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.zip`. L'archive comporte une série de fichiers XLS (Excel). Le plus souvent un fichier est associé à un chapitre.

Les logiciels Tanagra (1.4.32) et R (2.9.0) sont accessibles via leur site de distribution respectifs. Ainsi, le lecteur pourra reprendre pas à pas les exemples qui émaillent ce document. La compréhension des techniques n'en sera que meilleure.

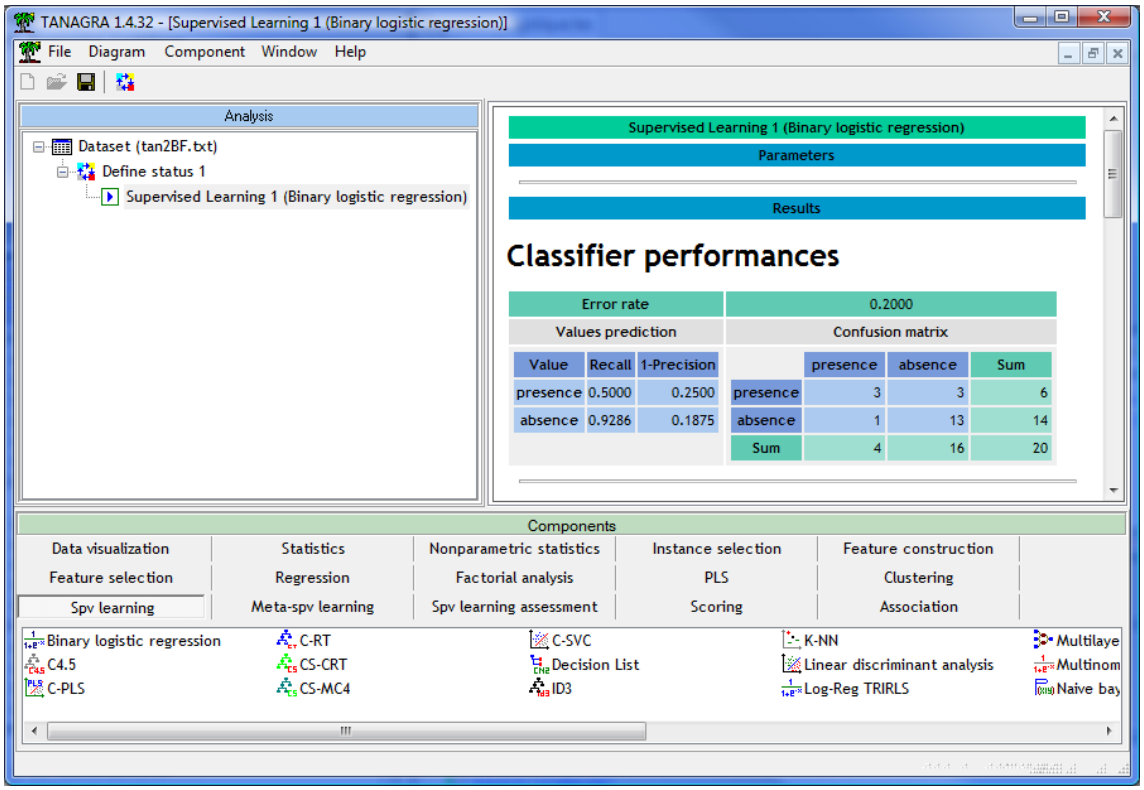
C

La régression logistique avec le logiciel TANAGRA

TANAGRA (`http://eric.univ-lyon2.fr/~ricco/tanagra/`) est un logiciel de data mining, statistique et analyse de données *open source*, totalement gratuit. La première version a été mise en ligne en Janvier 2004. La régression logistique a été implémentée dès la première version, elle a été constamment améliorée en termes de précision et de robustesse. Plus récemment, la régression logistique multinomiale a été programmée.

C.1 Lecture des résultats - Régression logistique binaire

La régression logistique binaire se situe dans l'onglet SPV LEARNING de la palette de composants, en compagnie des autres techniques d'apprentissage supervisé. Il n'y a pas de paramètres associées à la méthode. Voici l'interface générale du logiciel lorsque l'on met en oeuvre la régression logistique.



La fenêtre de résultats est subdivisée en plusieurs parties. Détaillons-les.

La matrice de confusion. Elle est automatiquement calculée sur la totalité des données disponibles (Figure C.1). Le taux d'erreur en resubstitution est affiché. Nous disposons aussi du rappel et de (1-précision) pour chaque modalité de la variable à prédire. Si on souhaite subdiviser les données en deux parties, construire le modèle sur la partie apprentissage et valider sur la partie test, une procédure commune à toutes les méthodes supervisées est proposée (cf. Tutoriels - `http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_TOW_Predefined_Test_Set.pdf`)

Classifier performances

Error rate	0.2000
Values prediction	Confusion matrix

Value	Recall	1-Precision
presence	0.5000	0.2500
absence	0.9286	0.1875

	presence	absence	Sum
presence	3	3	6
absence	1	13	14
Sum	4	16	20

Fig. C.1. Tanagra - Matrice de confusion

L'évaluation globale. Elle affiche les principaux indicateurs de significativité globale de la régression (Figure C.2). Les critères AIC (Akaike), BIC (SC pour Schwartz) et -2LL (Déviance) du modèle trivial (Intercept) et étudié (Model) sont confronté dans "Model Fit Statistics". Ensuite, nous avons le test du rapport de vraisemblance, avec LR (χ^2), le degré de liberté et la p-value. Enfin, plusieurs pseudo-R^2 sont proposés.

Paramètres estimés et odds-ratio. Dernière partie de la fenêtre, nous obtenons les paramètres estimés, les écarts-type, la statistique de Wald et la p-value du test de significativité individuelle. Un second tableau affiche les odds-ratio ($OR = e^{\hat{a}_j}$) et leur intervalle de confiance au niveau 95% (Figure C.3).

C.2 Sélection de variables

Les composants de sélection de variables pour la régression logistique se situent dans l'onglet FEATURE SELECTION. Leur utilisation "normale" consiste à les positionner juste après DEFINE STATUS qui indique la variable dépendante et les variables explicatives candidates. Ils filtrent automatiquement les explicatives. Nous pouvons brancher directement en aval la régression logistique (Figure C.4). Attention, si aucune explicative n'a été sélectionnée, la régression envoie un message d'erreur.

Nous pouvons techniquement brancher toute méthode d'apprentissage supervisé en aval. Après il faut savoir ce que l'on fait. Brancher une technique d'induction d'arbres de décision après avoir filtré les variables avec une procédure basée sur le test de Wald n'est peut être pas la stratégie la plus cohérente qui soit [1].

1. C'est même de la bêtise pour être honnête. Les biais de représentation et d'apprentissage ne sont absolument pas les mêmes. C'est comme napper de chantilly un rôti de veau, c'est peut être joli, mais sûrement infect.

Classifier characteristics

Data description

Target attribute	coeur (2 values)
# descriptors	3

Adjustement quality

Predicted attribute		coeur
Positive value		presence
Number of examples		20
Model Fit Statistics		
Criterion	Intercept	Model
AIC	26.435	24.618
SC	27.430	28.601
-2LL	24.435	16.618
Model Chi² test (LR)		
Chi-2		7.8169
d.f.		3
P(>Chi-2)		0.0500
R²-like		
McFadden's R²		0.3199
Cox and Snell's R²		0.3235
Nagelkerke's R²		0.4587

Fig. C.2. Tanagra - Évaluation globale de la régression

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	14.493790	-	-	-
age	-0.125634	0.0938	1.7936	0.1805
taux_max	-0.063560	0.0404	2.4694	0.1161
angine	1.779013	1.5046	1.3981	0.2370

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
age	0.8819	0.7338	1.0600
taux_max	0.9384	0.8669	1.0158
angine	5.9240	0.3104	113.0593

Fig. C.3. Tanagra - Coefficients estimés et odds-ratio

La description détaillée des sorties des composants est disponible dans la section consacrée à la sélection de variables (section 7.3.1, page 123).

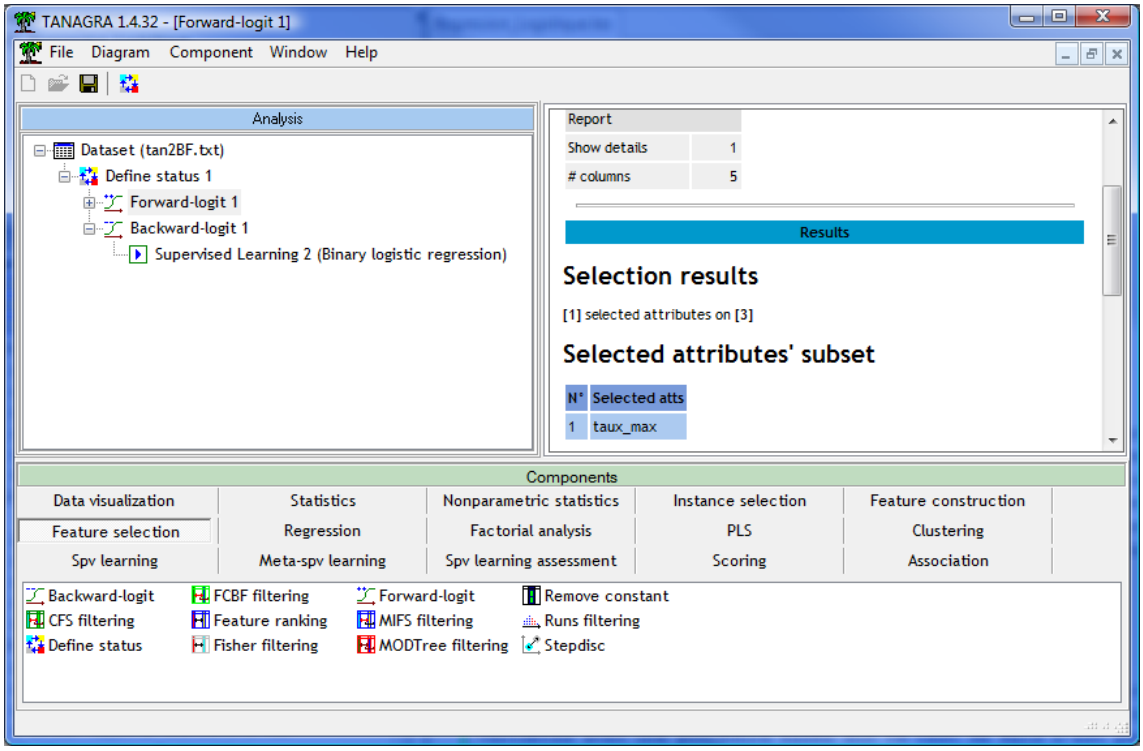


Fig. C.4. Tanagra - Sélection de variables - Enchaînements usuels

C.3 Didacticiels

Bien entendu, la régression logistique, méthode populaire s'il en fut, est très présente dans les didacticiels recensés sur notre site dédié `http://tutoriels-data-mining.blogspot.com`. Elle apparaît dans les comparaisons de méthodes, le scoring et la construction de la courbe de gain (gain chart ou lift curve), la construction de la courbe ROC,... Pour les consulter, le plus simple est d'explorer la section consacrée à la régression logistique ou de faire une recherche par mots-clés sur le site.

D

La régression logistique avec le logiciel R

Plus le temps passe, plus j'apprécie R. C'est pour cela que j'écris maintenant des tutoriels qui lui sont dédiés [1]. Bien sûr, il reste l'apprentissage du langage de programmation qui est une vraie barrière à l'entrée pour les réfractaires à l'idée de taper des instructions (ah bon ? et on fait quoi avec la souris monsieur ?), quant à les enchaîner n'en parlons même pas. R répond à un type de besoin différent de celui de Tanagra. Pour ma part, j'utilise les deux outils simultanément pour mes enseignements, avec le sacro-saint tableur bien entendu. On ne doit pas être dépendant d'un logiciel. Un scientifique se doit de contrôler les formules, croiser les références, recouper les résultats proposés par différents logiciels.

D.1 La régression logistique avec la commande `glm()`

D.1.1 `glm()`

La commande `glm()` implémente la régression linéaire généralisée. La régression logistique en est une déclinaison. Il suffit de spécifier la distribution de l'erreur avec l'option `family`.

L'affichage initial est assez succinct. Tout l'intérêt de R est que nous pouvons accéder à un certain nombre de champs internes dont la liste est obtenue avec `attributes()`. Elle est longue. Par exemple, le champ `fitted.values` nous donne accès aux $\hat{\pi}$ (Figure D.1).

D.1.2 `summary` de `glm()`

La commande `summary()` permet d'obtenir de plus amples informations sur la régression. L'affichage est déjà plus riche, avec les significativités individuelles des coefficients. Mais surtout, l'objet propose d'autres champs encore. Nous pouvons accéder à la matrice de variance covariance des coefficients entre autres (Figure D.2).

1. Comme j'en écris pour d'autres logiciels libres d'ailleurs : Knime, Orange, RapidMiner, Weka, ...

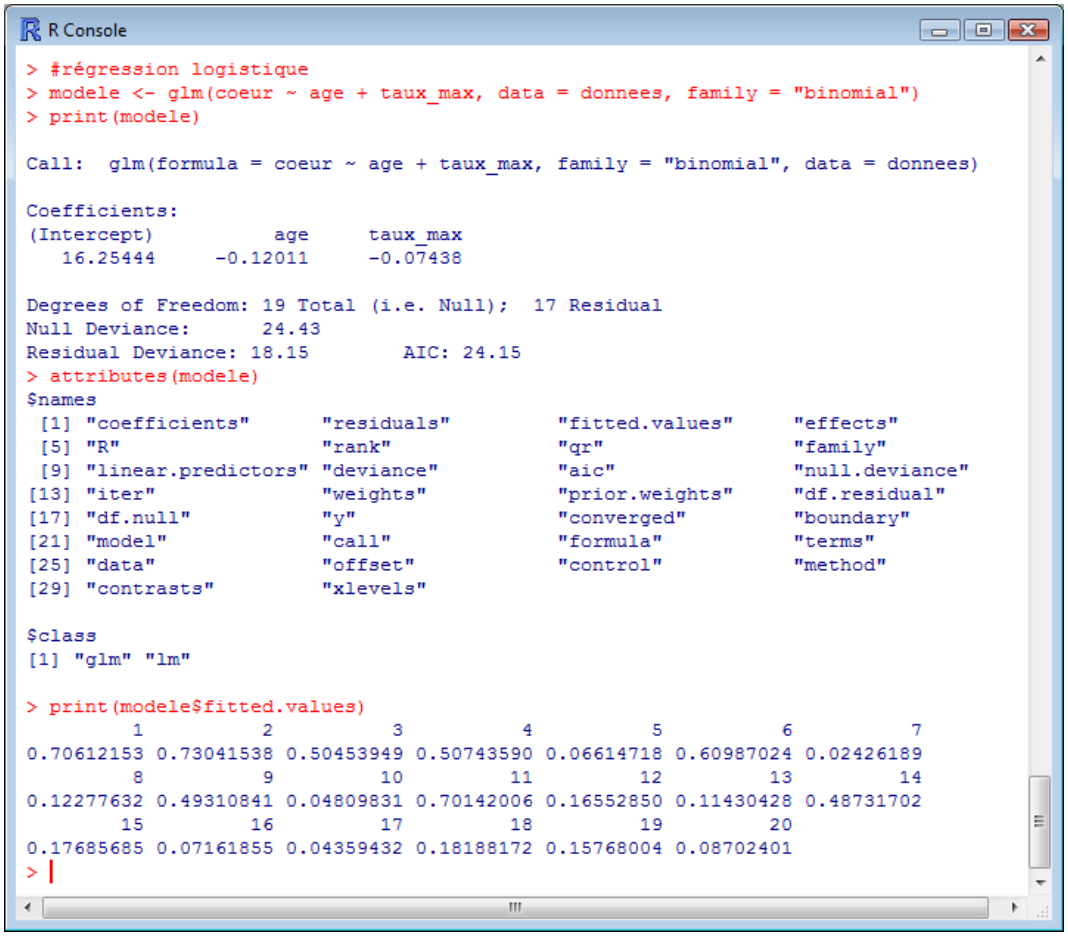


Fig. D.1. Logiciel R - Commande `glm()` et champs de l'objet associé

D.1.3 D'autres fonctions applicables sur l'objet `glm()`

Des fonctions peuvent s'appliquer sur un objet généré par la commande `glm()`. Nous avions vu la sélection de variable avec la commande `stepAIC` (section 7.2). Nous pourrions citer également la commande `influence.measures()`[2] qui produit les principaux indicateurs de l'analyse des résidus. Il en existe sûrement d'autres qui m'ont échappé, les possibilités sont immenses.

D.2 La régression logistique avec la commande `lrm()` du package `Design`

La commande `lrm()` du package `Design` implémente aussi la régression logistique binaire (et ordinale avec les odds proportionnels). En vérité, elle présente très peu d'avantages par rapport à `glm()`. Sauf en

2. En reproduisant les calculs, je me suis rendu compte que R ne fournit pas les dfbetas que j'ai calculé sous Excel, qui sont les mêmes que ceux de SAS et SPSS. J'y ai vraiment passé beaucoup de temps. J'avoue ne pas avoir pu reconstituer la formule utilisée par R. Pour ceux qui savent, un petit e-mail serait vraiment le bienvenu. Merci !

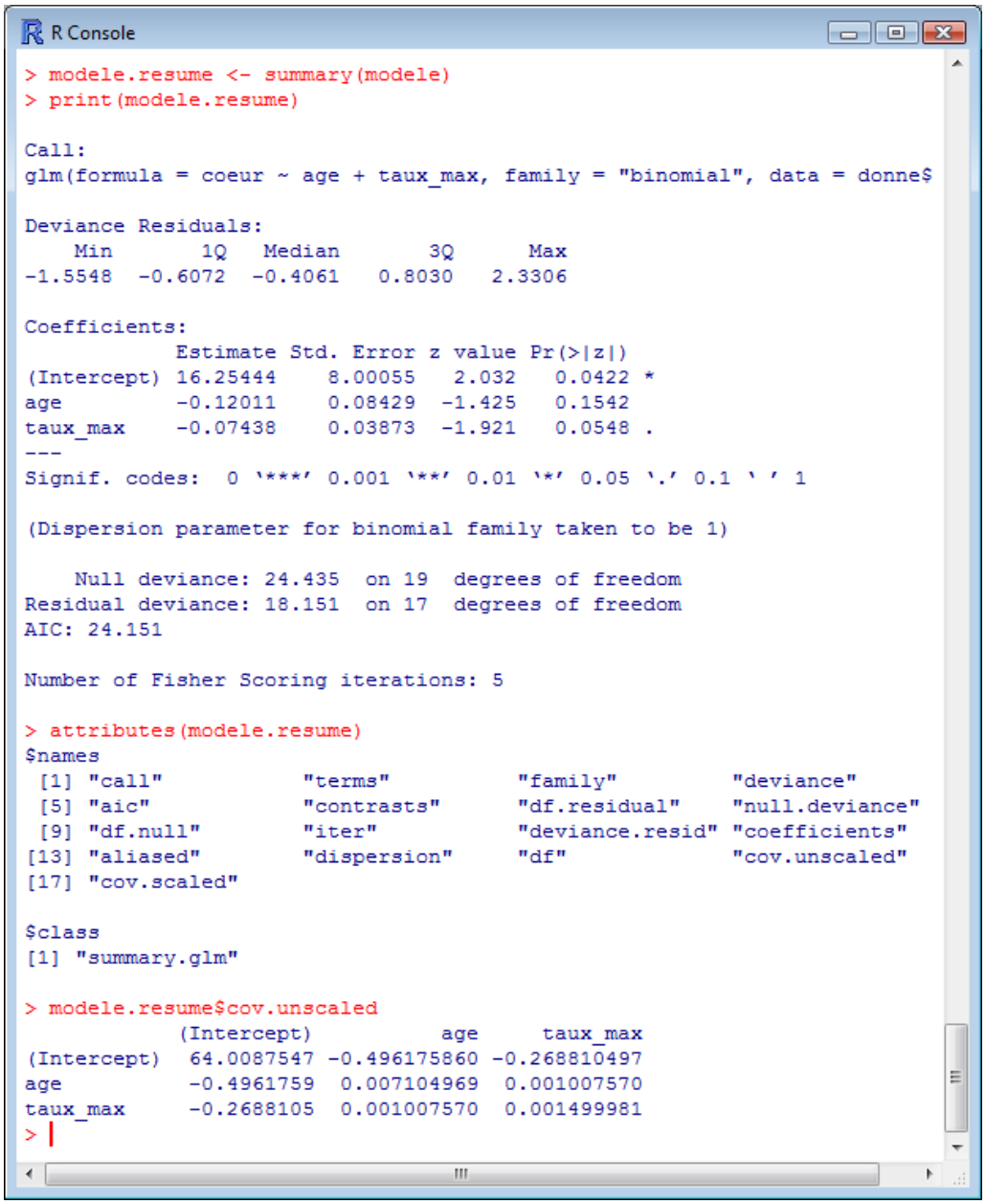


Fig. D.2. Logiciel R - Résumé de `glm()` et champs de l'objet associé

ce qui concerne la construction des résidus partiels. Toutes les combinaisons sont immédiatement fournies. Bien sûr, nous pourrions les reconstituer facilement en utilisant les fonctions spécialisées adéquates (loess, etc.), mais les obtenir facilement sans manipulations ésotériques reste un atout fort (voir section 8.2.4).

Littérature

1. A. Agresti, *Categorical Data Analysis*, Chapter 4, "Models for Binary Response Variables", pages 79-129, Wiley, 1990.
2. M. Bardos, *Analyse discriminante - Application au risque et scoring financier*, Chapitre 3, "Discrimination logistique", pages 61-79, Dunod, 2001.
3. G. Celeux, J.P. Nakache, *Analyse Discriminante sur Variables Qualitatives*, Polytechnica, 1994.
4. J. Jaccard, *Intercation Effects in Logistic Regression*, Series : Quantitative Applications in the Social Sciences, $n^0$135, Sage Publications, 2001.
5. D. Garson, *Logistic Regression*, `http://www2.chass.ncsu.edu/garson/PA765/logistic.htm`
6. R. Giraud, *Econométrie*, Collection "Que sais-je", n^o1423, PUF, 1993.
7. P.L. Gonzales, "Modèles à réponses dichotomiques", in *Modèles statistiques pour données qualitatives*, Droesbeke, Lejeune et Saporta Editeurs, Chapitre 6, pages 99-136, Technip, 2005.
8. T. Hastie, R. Tibshirani, J. Friedman, *The elements of Statistical Learning - Data Mining, Inference and Prediction*, Springer, 2001.
9. D.W. Hosmer, S. Lemeshow, *Applied Logistic Regression*, Second Edition, Wiley, 2000.
10. S. Menard, *Applied Logistic Regression Analysis (Second Edition)*, Series : Quantitative Applications in the Social Sciences, $n^0$106, Sage Publications, 2002.
11. J.P. Nakache, J. Confais, *Statistique Explicative Appliquée*, Partie 2, "Modèle Logistique", pages 77-168, Technip, 2003.
12. A.A. O'Connell, *Logistic Regression Models for Ordinal Response Variables*, Series : Quantitative Applications in the Social Sciences, $n^0$146, Sage Publications, 2006.
13. R. Rakotomalala, *Apprentissage Supervisé*, `http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`
14. R. Rakotomalala, *Régression logistique - Une approche pour rendre calculable* $P(Y/X)$, `http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`
15. R. Rakotomalala, *Régression logistique polytomique - Variable dépendante à* K $(K > 2)$ *modalités*, `http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`
16. R. Rakotomalala, *Normalisation des scores - Proposer une estimation fiable de* $P(Y = +/X)$ *dans un problème de discrimination*, `http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`
17. R. Rakotomalala, *Estimation de l'erreur de prédiction - Les techniques de ré-échantillonnage*, `http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`

18. R. Rakotomalala, *Comparaison de populations - Tests non paramétriques*, http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html#tests_non_parametriques

19. R. Rakotomalala, *Courbe ROC (Receiving Operating Characteristics - Une autre manière d'évaluer un modèle de prédiction*, http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html

20. R. Rakotomalala, *Étude des dépendances, Variables qualitatives - Tableau de contingence et mesures d'association*, http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html#mesures_association

21. G. Saporta, *Probabilités, Analyse de données et Statistique*, Section 18.6, "Régresion logistique binaire (deux groupes)", pages 475-480, Technip, 2006.

22. A. Slavkovic, *STAT 504 - Analysis of discrete data*, http://www.stat.psu.edu/online/development/stat504/06_logreg/01_logreg_intro.htm

23. M. Tenenhaus, *Statistique - Méthodes pour décrire, expliquer et prévoir*, Chapitre 11, "La régression logistique binaire", pages 387-460 ; Chapitre 12, "Régression logistique multinomiale : réponses polytomique et ordinale", pages 461-499, Dunod, 2007.

24. R. Tomassone, M. Danzart, J.J. Daudin, J.P. Masson, *Discrimination et classement*, Chapitre 6, pages 91-103, Masson, 1988.

25. Wikipedia, *Régression Logistique*, http://fr.wikipedia.org/wiki/Régression_logistique