

Introduction au modèle linéaire généralisé

David Causeur

Laboratoire de Mathématiques Appliquées
Pôle d'Enseignement Supérieur et de Recherche Agronomique de Rennes
65, rue de St-Brieuc - CS 84215
35042 Rennes Cedex
email : david.causeur@agrocampus-rennes.fr

Table des matières

Table des matières	4
Liste des tableaux	6
Liste des figures	8
1 Introduction au modèle linéaire généralisé	9
Introduction au modèle linéaire généralisé	9
1.1 Introduction	9
1.1.1 Conditions d'utilisation du modèle linéaire	10
1.1.2 Quelques situations en dehors des limites du modèle linéaire	12
1.2 Les différents types de variables	16
1.2.1 Y est qualitative à deux modalités	18
1.2.2 Y est qualitative à plus de deux modalités	24
1.2.3 Y est une variable de comptage	25

1.3	Modèles statistiques	26
1.3.1	Prédicteur du modèle	26
1.3.2	Fonction de lien et prédicteurs linéaires	26
1.3.3	Y est qualitative à deux modalités	29
1.3.4	Y est qualitative à plus de deux modalités	33
1.3.5	Y est une variable de comptage	38
1.4	Estimation des paramètres	39
1.4.1	Première impression graphique	40
1.4.2	Vraisemblance d'un modèle	40
1.4.3	Maximisation de la vraisemblance	41
1.4.4	Propriétés des estimateurs du maximum de vraisemblance	44
1.5	Tests de validité et comparaison de modèles	45
1.5.1	Déviance du modèle	46
1.5.2	Test de la validité par rapport à un sous-modèle	49
1.5.3	Sélection pas à pas du meilleur modèle	51
1.5.4	Contributions individuelles à la déviance résiduelle	53
1.6	Modèle logit et discrimination	54

Liste des tableaux

1.1	Extrait de données d'intention d'achat. La promesse d'achat prend les valeurs 0 (lorsque le sondé achèterait le produit) ou N (lorsque le sondé n'achèterait pas le produit)	12
1.2	Nombre de plantules avec sporulations (dans des bacs de 7 plantules) pour différentes doses de fongicide et différentes races de mildiou inoculé.	14
1.3	Épaisseur de muscle (en mm) et génotype de carcasses de porcs.	16
1.4	Extrait de données concernant le comportement de clients d'un magasin vendant du vin.	18
1.5	Correspondance entre résultats détaillés et synthèse par dénombrement	23
1.6	Nombres d'agneaux ayant un gras coloré.	24
1.7	Ancienneté et état des poumons de 371 mineurs (McCullagh & Nelder, 1983).	36
1.8	Estimation par la méthode du maximum de vraisemblance dans le cas des données d'intention d'achat.	43
1.9	Coefficients estimés du modèle complet pour les données « tournesol »	48
1.10	Analyse de la déviance du modèle complet pour les données « tournesol »	48
1.11	Analyse de la déviance du modèle avec interaction pour les données « tournesol ».	51
1.12	Descripteurs sensoriels des sandwichs.	52

1.13 Sélection d'un sous-ensemble pertinent de descripteurs sensoriels. En italique, les probabilités critiques conduisant à considérer que le modèle n'est pas intéressant. En gras, la probabilité critique du modèle sélectionné. 53

Table des figures

1.1	Variations de la teneur en viande maigre en fonction d'une épaisseur de gras (haut) et en fonction du sexe de la carcasse (bas).	11
1.2	Lien entre les promesses d'achat et les notes hédoniques attribuées par 36 juges. La promesse d'achat prend les valeurs 0 (lorsque le sondé achèterait le produit) ou N (lorsque le sondé n'achèterait pas le produit). Notons que, pour éviter les superpositions de points, les valeurs de note globale représentées sur le graphique sont légèrement modifiées par rapport aux valeurs mesurées.	13
1.3	Proportion de plantules avec sporulations en fonction du logarithme de la dose (afin de visualiser toutes les mesures, le logarithme de la dose a été perturbé de manière aléatoire en chaque point).	15
1.4	Type génétique et épaisseur de muscle de carcasses de porcs.	17
1.5	Nombre de bouteilles examinées par un client en fonction du logarithme de son temps de séjour en magasin.	19
1.6	Types de variables.	20
1.7	Relation entre variance et espérance d'une loi conditionnelle dans le cas où Y est qualitative à deux modalités.	21
1.8	Relation entre promesse d'achat et note hédonique. Pour chaque valeur de la note hédonique, la proportion d'acheteurs potentiels du sandwich est représentée par un cercle. Pour éviter les superpositions de points, les valeurs de note globale représentées sur le graphique sont légèrement modifiées par rapport aux valeurs mesurées.	22

1.9	Prédicteurs du modèle logit	28
1.10	Relation entre le logodds de la promesse d'achat et la note hédonique.	30
1.11	Interprétation des paramètres de la surface de réponse. Dans tous les exemples, $\beta_0 = -5\beta_1$	31
1.12	Modèle logistique multinomial représentant le lien entre type génétique et épaisseur de muscle.	35
1.13	Ajustement du modèle à risques proportionnels aux données sur la maladie des poumons des mineurs.	38
1.14	Modèle log-linéaire pour les données de comportements de clients en magasin.	39
1.15	Comparaison de 2 modèles logit pour les données d'intention d'achat.	42
1.16	Vraisemblance du modèle logit. Exemple des données d'intention d'achat	43
1.17	Ajustement d'un modèle de régression logistique (courbe noire). Les courbes en gris sont les limites d'un intervalle de confiance d'estimation de niveau 95 %. Exemple des données d'intention d'achat.	44
1.18	Ajustement du modèle saturé, du modèle complet et du modèle nul sur les données « tournesol ».	47
1.19	Ajustement du modèle avec interaction (en haut) et du modèle sans interaction (en bas) sur les données « tournesol ».	50
1.20	Détection de données mal ajustées (les points désignés par les lettres « R » ou « S » selon la race du mildiou) par l'analyse des résidus de la déviance.	55
1.21	Discrimination logit. Exemple des données d'intention d'achat. Les cercles désignent les données correctement affectées.	56
1.22	Choix de la meilleure valeur du seuil	57

Chapitre 1

Introduction au modèle linéaire généralisé

Causeur, D.

Agrocampus Rennes

Laboratoire de Mathématiques Appliquées

65, rue de St-Brieuc, CS84215 - 35042 Rennes cedex

david.causeur@agrocampus-rennes.fr

1.1 Introduction

Un des objectifs de cette introduction à la modélisation linéaire généralisée est de faire l'inventaire de questions qu'il est important de se poser au moment d'entreprendre un traitement de données. Ces questions interviennent donc à un stade de la démarche statistique où la réflexion sur la mesure du phénomène étudié a abouti à la définition d'un ensemble de caractéristiques pertinentes, appelées les **variables** du problème. Par exemple, dans le contexte agricole et dans les problèmes de sélection végétale ou animale en particulier, il est classique de retrouver parmi ces variables des mesures de rendement, d'autres caractérisant la qualité de l'environnement ou encore des informations sur le génotype des plantes ou des animaux qui font l'objet de l'étude. On le voit, toutes les variables n'ont pas le même statut : en effet, certaines d'entre elles, en l'occurrence ici les mesures de rendement, sont la cible de l'étude au sens où l'objectif de l'analyse est de comprendre comment ou pourquoi

ces quantités varient. Selon les domaines d'application et les logiciels, la terminologie diffère pour définir ces variables : on choisit pour la suite le terme **variables à expliquer** plutôt que variables réponses, variables à prédire ou variables endogènes. Par opposition, d'autres variables, ici les informations d'ordre environnemental ou génétique, apparaissent comme des sources de variabilité potentielles de ces variables à expliquer. Par analogie avec le choix de vocabulaire justifié ci-dessus, on choisit d'appeler ces variables les **variables explicatives** plutôt que prédictrices ou exogènes.

L'étude du lien entre les variables à expliquer et les variables explicatives passe maintenant par le choix d'une méthode statistique adaptée à la nature de leurs relations. En fait, chacune de ces méthodes statistiques s'appuie sur une représentation mathématique de ces relations, que l'on appelle le **modèle statistique**. Pour qu'il permette de mieux comprendre comment les variables à expliquer sont liées aux variables explicatives, il faut que ce modèle soit d'une part un bon compte-rendu de la réalité et d'autre part une représentation simple, réduite à l'essentiel, de cette réalité. Dans de nombreuses situations, les très populaires modèles linéaires de régression et d'analyse de la variance présentent ces deux qualités : il ne trahissent pas la réalité et leur linéarité leur confère une simplicité très attractive.

1.1.1 Conditions d'utilisation du modèle linéaire

L'utilisation des modèles linéaires est limitée à certaines conditions que l'on peut décrire à partir des exemples représentés sur la figure 1.1. Les graphes de cette figure représentent les variations des mesures de teneur en viande maigre (en kg de muscle par quintal) de 344 carcasses de porc selon les mesures d'une épaisseur de gras (en mm) d'une part et selon le sexe de la carcasse d'autre part. La variable à expliquer commune à ces deux exemples est la teneur en viande maigre. La variable explicative quant à elle est, selon le cas, soit l'épaisseur de gras soit le sexe.

La 1^{ère} condition d'application du modèle linéaire porte sur le **type de la variable** à expliquer. En l'occurrence, l'utilisation du modèle linéaire suppose que cette variable soit mesurable sur une échelle continue, à l'instar de la teneur en viande maigre et de nombreux autres indicateurs de rendement.

La 2^{ème} condition porte sur la distribution des valeurs de la variable à expliquer pour une valeur donnée de la variable explicative. Cette distribution s'appelle la **loi conditionnelle** de la variable à expliquer connaissant la variable explicative. Comme le montrent les courbes de densité de probabilité représentées sur les graphes de la figure 1.1, pour une épaisseur de gras donnée ou pour un sexe donné, la teneur en viande maigre se répartit de manière symétrique autour d'une position centrale. Plus précisément, la normalité de la loi conditionnelle est un pré-requis à l'application du modèle linéaire. Bien évidemment, ces deux conditions d'application constituent un

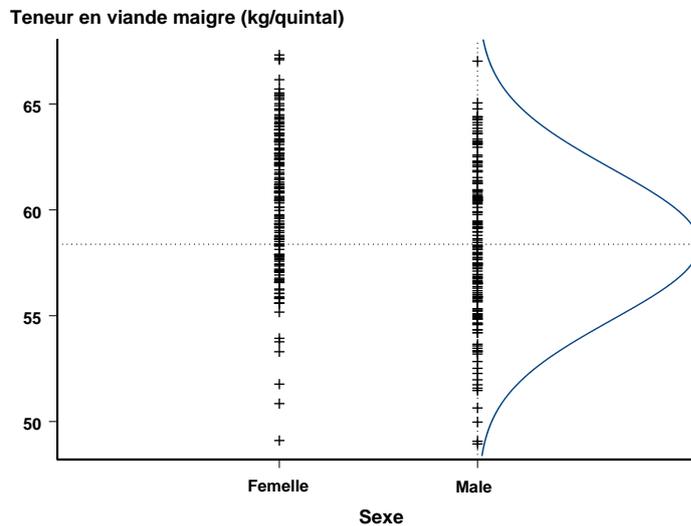
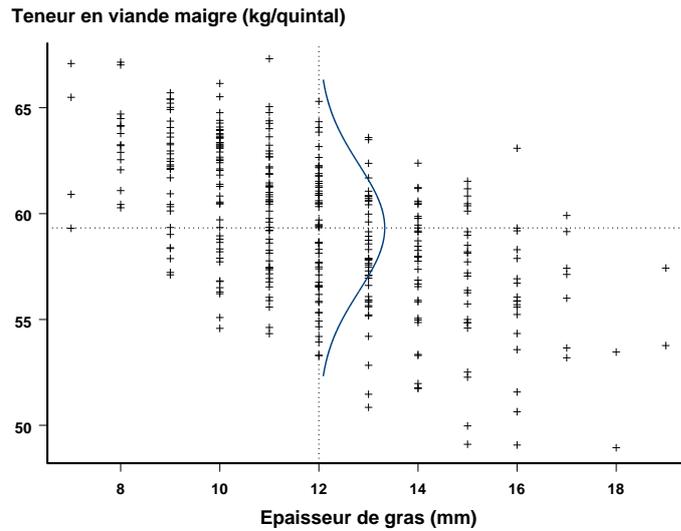


FIG. 1.1: Variations de la teneur en viande maigre en fonction d'une épaisseur de gras (haut) et en fonction du sexe de la carcasse (bas).

préalable à l'utilisation du modèle linéaire mais ne suffisent pas à spécifier entièrement le modèle. Par exemple, notons que l'utilisation du modèle linéaire suppose que l'écart-type de la loi conditionnelle soit le même, quelque soit la valeur de la variable explicative. Pour reprendre l'exemple de la teneur en viande maigre, ceci signifie que les dispersions des valeurs observées pour une épaisseur de gras donnée ne dépendent pas de la valeur de cette épaisseur ou que les dispersions intra-sexes sont les mêmes.

La 1^{ère} étape de l'analyse, déterminante dans le choix du modèle statistique, consiste donc à caractériser le type de la variable à expliquer et la forme de la loi conditionnelle.

1.1.2 Quelques situations en dehors des limites du modèle linéaire

On présente quelques situations courantes échappant au domaine d'utilisation classique du modèle linéaire. Afin d'apporter une réponse adaptée à chacune de ces situations, l'ensemble des modèles statistiques que nous envisageons dans la suite étend celui des modèles linéaires. Toutefois, afin de préserver toute la souplesse d'interprétation des modèles linéaires, l'élargissement des conditions d'utilisation se limitera à un cadre préservant une forme de linéarité du modèle : on parle alors de **modèle linéaire généralisé**.

Exemple : promesse d'achat d'un produit alimentaire. Le tableau 1.1 reproduit un extrait des résultats d'une dégustation de sandwiches par 36 juges. Cet extrait se limite à deux variables : d'une part, la promesse d'achat qui correspond à la réponse par oui (codée par la suite 1) ou non (codée 0) à la question « achèteriez-vous ce sandwich ? » et d'autre part, la note hédonique attribuée par le juge au produit sur une échelle entière allant de 1 à 10. L'objectif est ici d'étudier la relation entre la promesse d'achat, variable à expliquer, et la note hédonique, variable explicative.

	Promesse d'achat	Note globale
1	0	6
2	N	7
3	N	5
4	0	9
5	0	8
6	0	5
⋮	⋮	⋮

TAB. 1.1: Extrait de données d'intention d'achat. La promesse d'achat prend les valeurs 0 (lorsque le sondé achèterait le produit) ou N (lorsque le sondé n'achèterait pas le produit)

Le graphique de la figure 1.2 représente le nuage des valeurs observées des notes hédoniques et promesses d'achat. Comme dans le cas de la régression linéaire, ce type de graphique est traditionnellement utilisé, en

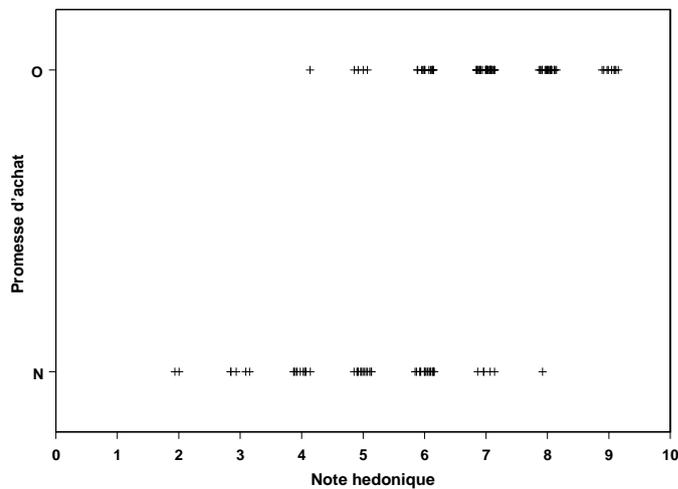


FIG. 1.2: Lien entre les promesses d'achat et les notes hédoniques attribuées par 36 juges. La promesse d'achat prend les valeurs 0 (lorsque le sondé achèterait le produit) ou N (lorsque le sondé n'achèterait pas le produit). Notons que, pour éviter les superpositions de points, les valeurs de note globale représentées sur le graphique sont légèrement modifiées par rapport aux valeurs mesurées.

préalable à l'ajustement, pour décrire l'évolution des valeurs de la variable à expliquer en fonction de celles de la variable explicative.

La forme du nuage de points, structuré en deux sous-nuages horizontaux, est liée à la nature de la variable à expliquer, binaire. Il en résulte aussi qu'à l'évidence, faire le postulat que, pour une note hédonique donnée, les valeurs de la variable à expliquer se répartissent selon une loi normale est en contradiction avec la réalité observée. □

Exemple : résistance du tournesol au mildiou. Les données dont un extrait est présenté dans le tableau 1.2 proviennent d'une expérience mise en place par l'équipe pathologie du tournesol, laboratoire INRA de pathologie et de mycologie de la station d'amélioration des plantes de Clermont-Ferrand (voir [?])¹. L'objectif de l'expérience est d'étudier l'effet d'un traitement fongicide sur 2 races de mildiou, dont l'une, notée R, présente une résistance au fongicide par mutation. L'expérience consiste à inoculer le champignon parasite à des groupes de 7 plantules et à mesurer le nombre de plantules présentant des signes d'infection, à savoir une sporulation.

1. Cet exemple sert aussi d'illustration au chapitre consacré au modèle linéaire généralisé du plan de formation à la statistique proposé par le département Biométrie et Intelligence Artificielle de l'INRA. (Voir [?])

Dans cet exemple, la variable à expliquer est la proportion de plantes avec sporulations par bac et les variables explicatives sont d'une part la race du mildiou inoculé et d'autre part la dose de fongicide ou plus judicieusement, comme nous le verrons plus loin, le logarithme de la dose.

Groupe	Race	Dose	Sporulations	Proportion de sporulations	Nombre de plantules
1	S	0,001	5	0.71	7
2	S	0,001	6	0.86	7
3	S	0,001	7	1.00	7
4	S	0,001	7	1.00	7
5	S	0,001	7	1.00	7
6	S	0,001	5	0.71	7
7	S	0,02	7	1.00	7
8	S	0,02	7	1.00	7
⋮	⋮	⋮	⋮	⋮	

TAB. 1.2: *Nombre de plantules avec sporulations (dans des bacs de 7 plantules) pour différentes doses de fongicide et différentes races de mildiou inoculé.*

Le graphique de la figure 1.3 montre l'évolution de la proportion de sporulations en fonction du logarithme de la dose de fongicide. Il apparaît assez nettement sur ce graphique que les évolutions sont très différentes pour les 2 races de mildiou. D'autre part, le graphique met également bien en évidence la non-pertinence des postulats du modèle linéaire, en particulier ceux concernant la normalité de la loi conditionnelle et l'homogénéité de sa variance. □

Exemple : lien entre génotype et épaisseur tissulaire. Dans de nombreux pays de l'Union Européenne, la teneur en viande maigre d'une carcasse de porc est évaluée à partir d'épaisseurs tissulaires mesurées en différents sites sur la carcasse. Les méthodes d'évaluation de la teneur en viande maigre ne tiennent en revanche pas compte du type génétique, exposant ainsi la prédiction à un biais entre les différents génotypes. Une manière de réduire le biais passe par la construction d'un modèle de prédiction du génotype par les épaisseurs tissulaires. La variable à expliquer est donc ici le génotype de la carcasse, variable qualitative nominale dont les modalités sont notées LWP (LargeWhite × Piétrain), LWLF (LargeWhite × Landrace Français), PAL (Pen-

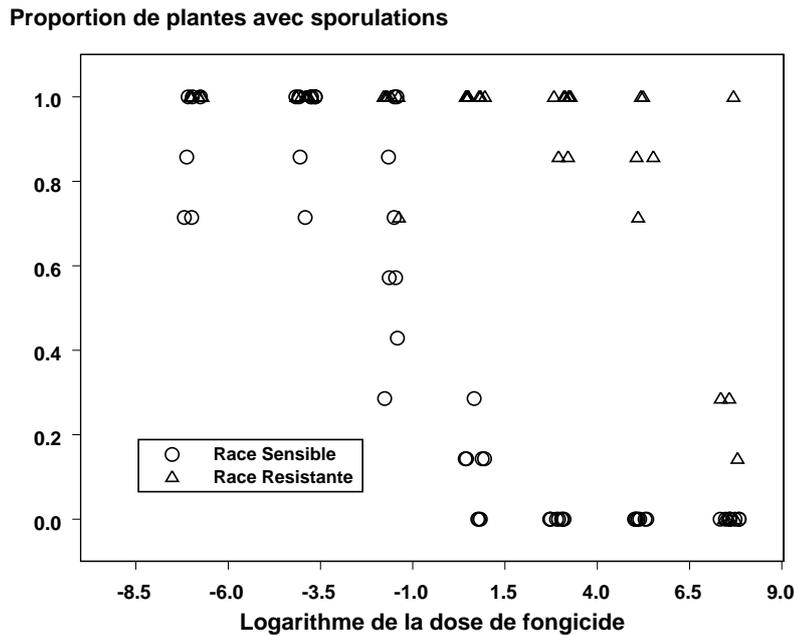


FIG. 1.3: *Proportion de plantules avec sporulations en fonction du logarithme de la dose (afin de visualiser toutes les mesures, le logarithme de la dose a été perturbé de manière aléatoire en chaque point).*

ar-Lan) et P (Piétrain). La variable explicative est une épaisseur de muscle (en mm). Un extrait de données est fourni dans le tableau 1.3 et le graphique de la figure 1.4 illustre ces données. Là encore, le nuage de points représenté sur le graphique de la figure 1.4 ne ressemble pas au nuage elliptique propice à l'utilisation du modèle linéaire. Ces données font l'objet d'une analyse détaillée dans le chapitre ?? □

Exemple : comportement de clients en magasin. Les chercheurs du département Marketing de Texas Tech University s'intéressent à l'effet des conditions d'ambiance en magasin sur le comportement des clients. Pour cela, ils mettent en place une expérience consistant à observer le comportement de clients d'un magasin commercialisant du vin, à travers la durée de présence dans le magasin, le nombre de fois qu'un client sort une bouteille du rayonnage, le nombre de bouteilles achetées ... dans différentes conditions de luminosité (douce ou forte) et de musique d'ambiance (classique ou pop). Un extrait des données est reproduit dans le tableau 1.4. Le graphique de la figure 1.5 illustre le lien entre le nombre de bouteilles examinées par un client et la durée de son passage dans le magasin. Parmi les enseignements à tirer de ce graphique, il apparaît à l'évidence que le nombre moyen de bouteilles examinées, mais aussi la dispersion de ce nombre, est d'autant plus grand que la durée de séjour en magasin est longue. C'est notamment cette dernière remarque qui va à l'encontre des hypothèses traditionnelles justifiant l'utilisation du modèle linéaire classique. □

	Épaisseur de muscle (en mm)	Génotype
1	51.48	LWP
2	52.24	LWP
3	57.74	PAL
4	56.98	PAL
5	57.36	LWP
6	56.61	LWP
7	61.16	LWP
8	45.97	LWLF
9	55.09	LWLF
10	54.71	LWLF
⋮	⋮	⋮

TAB. 1.3: *Épaisseur de muscle (en mm) et génotype de carcasses de porcs.*

1.2 Les différents types de variables

Dans la suite, on utilise les notations traditionnelles du modèle linéaire : Y désigne la variable à expliquer et $x = (x^{(1)}, x^{(2)}, \dots, x^{(p)})$ le vecteur des mesures des p variables explicatives. Le plus souvent, dans les exemples illustratifs, on se limitera au cas d'une seule variable explicative.

Comme on l'a vu, il est important dans un premier temps d'identifier le type de la variable à expliquer avant d'aller plus loin dans la démarche de modélisation. La typologie schématisée dans la figure 1.6 est relativement standard. Elle met en avant deux grandes familles de variables :

- les variables quantitatives, que l'on mesure par des valeurs numériques. La teneur en viande maigre d'une carcasse (en kg/quintal) et le nombre de cibles détruites par un fongicide sont deux exemples de variables quantitatives. Remarquons que les valeurs prises par la teneur en viande maigre se répartissent sur un intervalle continu : on dit que cette variable est **quantitative continue**. En revanche, les valeurs prises par le nombre de cibles détruites étant cantonnées aux entiers naturels, on parle de variable **quantitative**

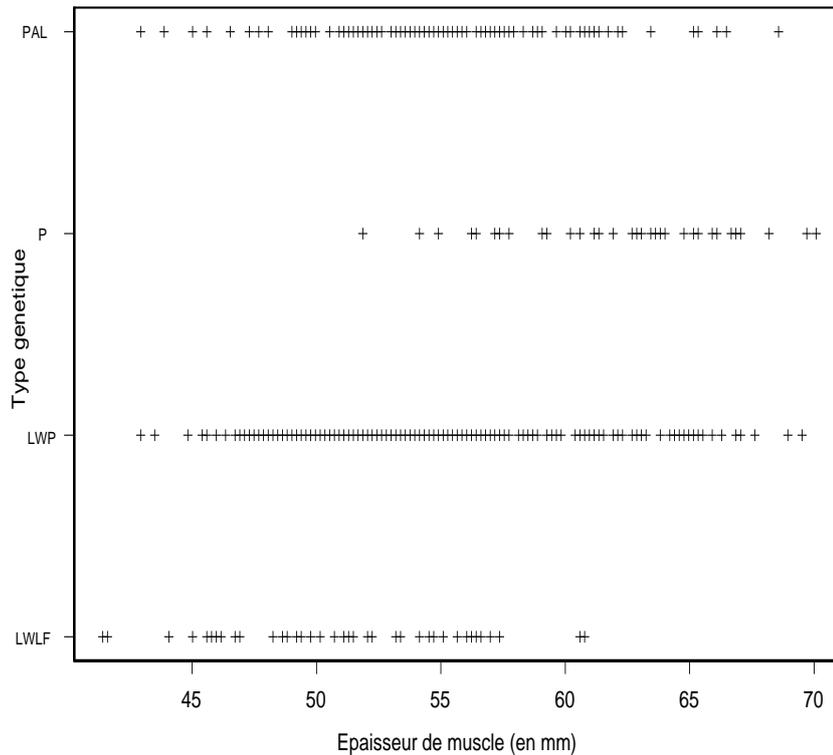


FIG. 1.4: *Type génétique et épaisseur de muscle de carcasses de porcs.*

discrète.

- les variables qualitatives, que l'on ne peut mesurer par des valeurs numériques. La couleur d'un produit alimentaire et son niveau de qualité (notée A, B, C, D) constituent deux exemples de variables qualitatives. Ici, les valeurs prises par la couleur ne peuvent a priori pas être rangées selon un ordre logique. On parle alors de **variable nominale** et on appelle **modalités** les valeurs de cette variable. En revanche, les valeurs prises par la qualité peuvent être rangées selon un ordre de qualité croissante ou décroissante. On parle alors de **variable ordinale** et on appelle **niveaux** les valeurs de cette variable.

Remarquons que, si l'examen de la nature de la variable à expliquer est une étape fondamentale dans le choix du modèle statistique, celui de la nature des variables explicatives est également très important. Lorsqu'une variable explicative est qualitative, on l'appelle aussi **facteur**. De même, lorsqu'elle est quantitative, on utilise parfois le terme **covariable**. Dans le cadre du modèle linéaire, lorsque toutes les variables explicatives sont

Client	Luminosité	Musique	Durée (minutes)	...	Nombre de bouteilles examinées	Nombre de bouteilles extraites	Nombre de bouteilles achetées
1	Douce	Classique	2.0	...	0	0	0
2	Douce	Classique	1.0	...	0	0	0
3	Douce	Classique	0.5	...	0	0	0
4	Douce	Classique	2.3	...	16	0	0
5	Douce	Classique	2.2	...	1	1	0
6	Douce	Classique	1.7	...	1	0	0
7	Douce	Classique	42.7	...	11	3	1
8	Douce	Classique	3.0	...	0	0	0
9	Douce	Classique	0.7	...	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

TAB. 1.4: Extrait de données concernant le comportement de clients d'un magasin vendant du vin.

quantitatives, on parle de **modèles de régression linéaire**. A l'opposé, si toutes les variables explicatives sont qualitatives, le modèle linéaire devient **modèle d'analyse de la variance**. L'extension que nous présentons par la suite ne porte que sur la nature de la variable à expliquer : pour chaque type de variable, on retrouve donc toute la gamme des modèles bien connus dans le cadre linéaire, de la régression à l'analyse de la variance.

Dans la suite, on passe en revue les différents types de variables à expliquer dans un ordre de complexité croissante.

1.2.1 Y est qualitative à deux modalités

Quoiqu'en apparence très basique, cette situation est en fait assez fréquente. Par exemple, les sciences animales et végétales sont de grandes consommatrices de ce type de modèle : la variable à expliquer peut alors être la réussite ou non d'un traitement, la présence ou non d'un état de grossesse, etc.

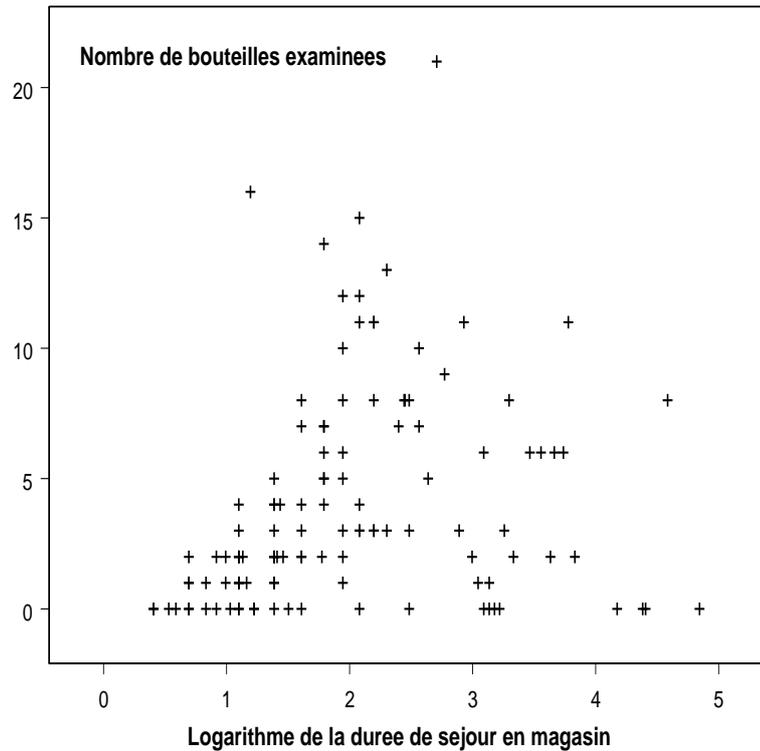
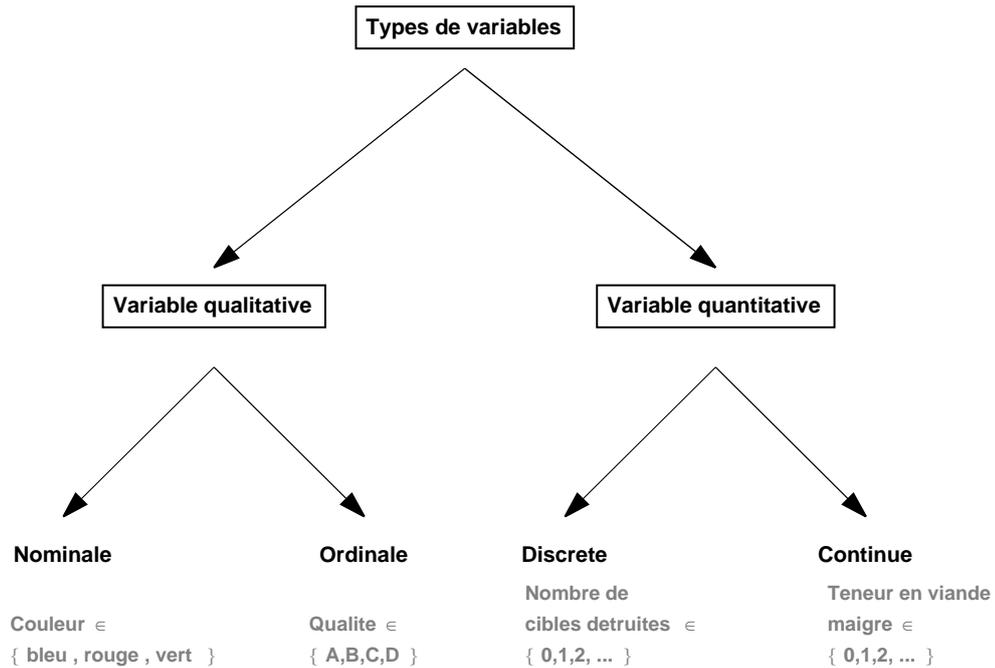


FIG. 1.5: *Nombre de bouteilles examinées par un client en fonction du logarithme de son temps de séjour en magasin.*

Dans la suite, par convention, les deux valeurs prises par la variable à expliquer sont 0 et 1. Dans de nombreuses situations, la valeur 1 code la modalité « positive ». Par exemple, si les états de la variable sont « réussite » et « échec », le 1 code la valeur « réussite ».

Exemple : promesse d'achat d'un produit alimentaire. Le tableau 1.1 illustre la forme générale sous laquelle se présentent les données par un extrait des résultats d'une dégustation de sandwiches par $n = 36$ juges. Ici, la variable à expliquer est la promesse d'achat qui correspond à la réponse par oui (codée par la suite 1) ou non (codée 0) à la question « achèteriez-vous ce sandwich ? » alors que la variable explicative est la note hédonique x . □

Pour une valeur donnée x des variables explicatives, la loi conditionnelle de Y est dite **loi de Bernoulli** et est

FIG. 1.6: *Types de variables.*

complètement définie par la relation suivante :

$$\mathbb{P}_x(Y = 1) = \pi(x),$$

où $\pi(x)$ est une fonction des variables explicatives telle que, pour tout x , $0 \leq \pi(x) \leq 1$.

Traditionnellement, $\pi(x)$ et $1 - \pi(x)$ s'interprètent comme des **Risques (mesures)** : par exemple, si Y est le résultat d'un traitement médical, ces deux quantités formalisent respectivement la probabilité de réussite codée 1 et le risque d'échec codé 0.

Notons que l'espérance et la variance de la loi conditionnelle se déduisent directement de $\pi(x)$:

$$\begin{aligned} \mathbb{E}_x(Y) &= \pi(x), \\ \text{Var}_x(Y) &= \pi(x)[1 - \pi(x)]. \end{aligned}$$

On déduit en particulier des relations précédentes une propriété intéressante de la loi conditionnelle lorsque Y est qualitative à 2 modalités : la variance de cette loi dépend de son espérance. La forme de la relation variance-espérance est décrite par le graphique de la figure 1.7. Il apparaît clairement sur ce graphique que, lorsque $\pi(x)$

est proche de 0 ou de 1, la variance de la loi conditionnelle est proche de 0, alors que celle-ci est maximale lorsque $\pi(x)$ est proche de 0.5.

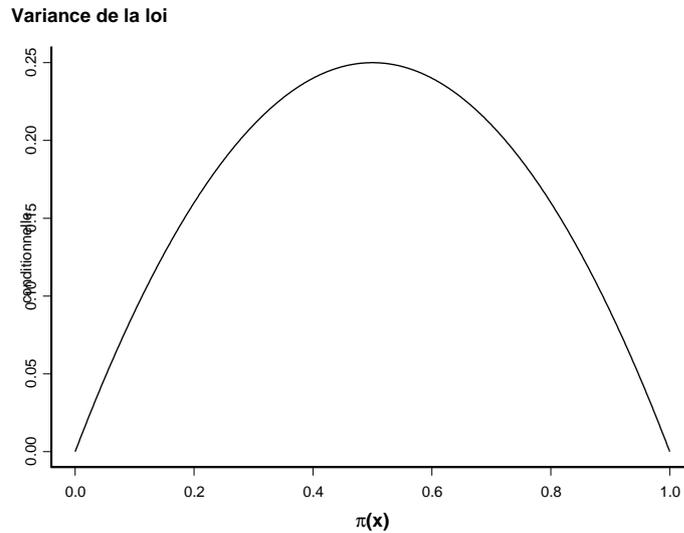


FIG. 1.7: Relation entre variance et espérance d'une loi conditionnelle dans le cas où Y est qualitative à deux modalités.

Exemple : promesse d'achat d'un produit alimentaire. Ici, $\pi(x)$ est la probabilité qu'un acheteur ayant attribué la note x au produit l'achète. La propriété énoncée plus haut se traduit concrètement sur le graphique de la figure 1.8 par le fait que :

- pour de faibles valeurs de x , $\pi(x)$ est presque nulle et effectivement la valeur de Y varie alors très peu de 0,
- pour des valeurs x autour de 6, $\pi(x)$ semble proche de 0.5 et alors les valeurs de Y se répartissent presque équitablement entre 0 et 1,
- pour des valeurs élevées de x , $\pi(x)$ est presque égal à 1 et alors la valeur de Y varie très peu de 1. \square

Lorsqu'il est possible de contrôler, en conditions expérimentales par exemple, les valeurs prises par les variables explicatives, alors on peut disposer de plusieurs valeurs de Y pour une même valeur de x . Dans ce cas, il est tentant de mener l'étude, non pas à partir des valeurs binaires de Y correspondant aux résultats individuels, mais plutôt à partir du dénombrement par valeur de x des résultats pour lesquels Y prend la valeur 1. En fait,

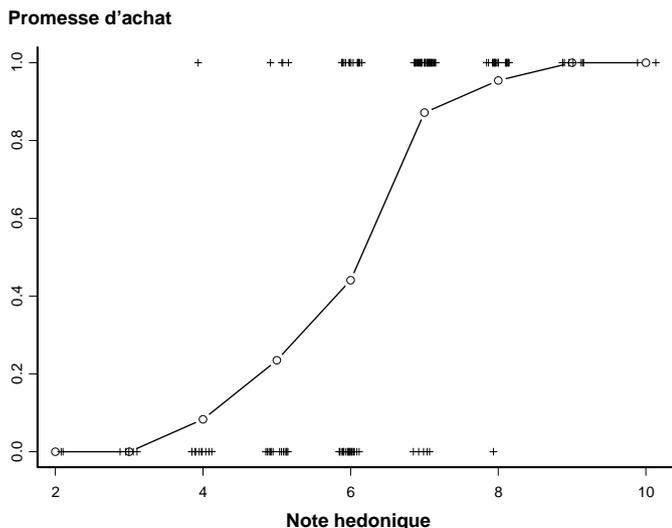


FIG. 1.8: Relation entre promesse d'achat et note hédonique. Pour chaque valeur de la note hédonique, la proportion d'acheteurs potentiels du sandwich est représentée par un cercle. Pour éviter les superpositions de points, les valeurs de note globale représentées sur le graphique sont légèrement modifiées par rapport aux valeurs mesurées.

lorsque les unités statistiques sont indépendantes, il est équivalent de travailler sur les données individuelles et sur les données ainsi groupées. Le tableau 1.5 illustre cette manière d'agréger les données : pour une valeur de x fixée pour laquelle on dispose de n_x répétitions de Y , Z désigne le nombre d'unités statistiques, parmi ces n_x répétitions, pour lesquelles Y prend la valeur 1.

Exemple : résistance du tournesol au mildiou. Dans cet exemple, on peut considérer que Y est qualitative à deux modalités, à savoir l'observation ou non d'une sporulation sur une plantule. Comme pour chaque valeur du couple (dose, race) on dispose de 7 mesures indépendantes de Y , il est aussi possible de considérer que la variable à expliquer est le nombre de plantules présentant une sporulation dans chaque bac ou encore la proportion de ces plantules. C'est ainsi que se présentent les données dans le tableau 1.2. \square

Pour une valeur x donnée des variables explicatives, si l'on peut considérer que les n_x mesures de Y sont indépendantes, alors la loi de Z est la **loi binomiale** notée de la manière suivante :

$$Z \sim \mathcal{B}[n_x, \pi(x)]. \quad (1.1)$$

	Données individuelles			Données groupées (agrégées)		
	x	Y		x	Z	n_x
1	0	1				
2	0	1				
3	0	0				
4	0	0		0	2	4
5	1	1	\Rightarrow			
6	1	0				
7	1	0		1	1	3
8	2	1				
9	2	1				
10	2	1		2	3	3
\vdots	\vdots	\vdots		\vdots	\vdots	

TAB. 1.5: Correspondance entre résultats détaillés et synthèse par dénombrement

Ici aussi, l'espérance et la variance de la loi conditionnelle se déduisent directement de $\pi(x)$:

$$\begin{aligned}\mathbb{E}_x(Z) &= n_x \pi(x), \\ \text{Var}_x(Z) &= n_x \pi(x) [1 - \pi(x)].\end{aligned}$$

A l'instar de la représentation des données « tournesol » sur le graphique de la figure ??, une variante toute aussi intuitive de l'agrégation des données consiste à résumer les informations individuelles par les proportions Z/n_x dont l'espérance et la variance conditionnelle sont données par :

$$\begin{aligned}\mathbb{E}_x\left(\frac{Z}{n_x}\right) &= \pi(x), \\ \text{Var}_x\left(\frac{Z}{n_x}\right) &= \frac{\pi(x)[1 - \pi(x)]}{n_x}.\end{aligned}$$

Exemple : l'essai gras coloré. Cet exemple est traité en détail dans le chapitre ??. Il concerne l'étude des effets de l'alimentation et du mode de logement sur la couleur du gras d'agneaux (voir Grenet, 1999). Dans

cette étude, la variable à expliquer est la coloration du gras considérée comme binaire (0 = absence de coloration, 1 = présence d'une coloration) et les variables explicatives sont les modes d'alimentation et de logement, toutes deux étant des variables qualitatives à 2 modalités. Le protocole expérimental prévoit la répartition de 80 agneaux en parts égales dans les quatre combinaisons possibles des modalités des variables explicatives. Les résultats sont présentés dans le tableau 1.6.

Alimentation	Logement	Nombre d'agneaux à gras coloré	Effectif total
A_1	L_1	10	$n_{11} = 20$
A_1	L_2	12	$n_{12} = 20$
A_2	L_1	15	$n_{21} = 20$
A_2	L_2	16	$n_{22} = 20$

TAB. 1.6: Nombres d'agneaux ayant un gras coloré.

Lorsque les variables explicatives sont des facteurs, on préfère, comme dans le cas linéaire, à la notation fonctionnelle (1.1) la notation indicée suivante : pour le i ème mode d'alimentation et le j ème mode de logement,

$$Z \sim \mathcal{B}[n_{ij}, \pi_{ij}]. \quad \square$$

Que l'on étudie la relation entre Y et x à partir des données individuelles ou des données groupées, l'objet de la modélisation est de mettre en avant une formulation réaliste de la fonction $\pi(x)$. Les nombres n_x , quant à eux, sont connus de l'expérimentateur et nécessaires à la modélisation.

1.2.2 Y est qualitative à plus de deux modalités

Cette situation généralise la précédente. Notons C_1, C_2, \dots, C_K les K modalités de la variable Y , alors la loi conditionnelle de Y est décrite par les relations suivantes :

$$\mathbb{P}_x(Y = C_k) = \pi^{(k)}(x), \quad k = 1, 2, \dots, K.$$

où, pour tout x , pour $k = 1, 2, \dots, K$, $0 \leq \pi^{(k)}(x) \leq 1$ et $\pi^{(1)}(x) + \pi^{(2)}(x) + \dots + \pi^{(K)}(x) = 1$.

Exemple : lien entre génotype et épaisseur tissulaire. Dans les données du tableau 1.3, la variable à expliquer est le génotype de la carcasse, variable qualitative nominale dont les modalités sont LWP (LargeWhite \times

Piétrain), LWLF (LargeWhite \times Landrace Français), PAL (Pen-ar-Lan) et P (Piétrain). \square

Comme dans le cas où Y est une variable qualitative à 2 modalités, les données peuvent également se présenter sous une forme groupée lorsque l'on dispose de plusieurs répétitions de Y pour une même valeur de x .

1.2.3 Y est une variable de comptage

Le dénombrement intervient dans de nombreuses situations de recueil d'information en sciences biologiques. En particulier, en écologie, l'étude de la dynamique des populations peut s'appuyer sur le comptage, en différents lieux et à différents instants, d'individus, animaux ou plantes, présentant les mêmes caractéristiques. La variable à expliquer est alors le nombre de ces individus et les variables explicatives sont par exemple des caractéristiques environnementales du lieu du comptage. Sous certaines hypothèses de répartition uniforme des individus dénombrés, hypothèses que nous ne discuterons pas ici, la variable à expliquer est distribuée selon une loi de Poisson :

$$Y \sim \mathcal{P}[\lambda(x)],$$

où $\lambda(x) \geq 0$ est appelée l'**intensité** de la loi de Poisson.

Comme dans les situations décrites plus haut, l'espérance et la variance de la loi conditionnelle sont ici aussi liées fonctionnellement :

$$\mathbb{E}_x(Y) = \text{Var}_x(Y) = \lambda(x).$$

Cette relation entre espérance et variance traduit en pratique le fait que la dispersion des effectifs dénombrés est d'autant plus grande que leur valeur moyenne est élevée.

Exemple : comportement de clients en magasin. Ici, la variable à expliquer est le nombre de bouteilles examinées par un client. Comme on le voit sur le graphique de la figure 1.5, la dispersion des nombres de bouteilles examinées est d'autant plus grande que leur nombre moyen est grand. Dans cette situation, il est judicieux de modéliser ce nombre par une loi de Poisson dont l'intensité est dépend de la durée du séjour du client en magasin. \square

1.3 Modèles statistiques

1.3.1 Prédicteur du modèle

Lorsque la variable Y est quantitative continue, la modélisation du lien entre la variable à expliquer et les variables explicatives par le modèle linéaire consiste essentiellement à décrire l'évolution de l'espérance de la loi conditionnelle en fonction des valeurs des variables explicatives :

$$\begin{aligned}\mathbb{E}_x(Y) &= \beta_0 + \beta_1 x^{(1)} + \beta_p x^{(p)}, \\ &= L(x; \beta)\end{aligned}$$

où β désigne l'ensemble des coefficients inconnus du modèle et $L(x; \beta)$ est communément appelé le **prédicteur linéaire**.

Dans le cas où Y est une variable de dénombrement, distribuée selon une loi de Poisson, appliquer la démarche précédente pose un premier problème : si, dans le modèle linéaire, $L(x; \beta)$ n'est en général pas l'objet de restrictions quant à sa plage de variations, l'espérance conditionnelle de Y est ici l'intensité de la loi de Poisson, ce qui impose que le prédicteur soit aussi positif. De même, lorsque Y est une variable qualitative à deux modalités, son espérance conditionnelle est une probabilité et le prédicteur doit en outre être inférieur à 1.

Enfin, lorsque Y est une variable qualitative à plus de deux modalités, la transposition directe de la démarche mise en œuvre dans le cas du modèle linéaire se heurte à un problème fondamental : la notion d'espérance conditionnelle n'a pas d'équivalent direct. Par conséquent, dans ce cas, la modélisation du lien porte directement sur les probabilités de chacune des modalités :

$$\mathbb{P}_x(Y = C_k) = \pi^{(k)}(x), \quad k = 1, \dots, K.$$

Ici, les fonctions $\pi^{(k)}$ que nous introduisons doivent naturellement satisfaire aux contraintes liées à la modélisation d'une probabilité, à savoir :

$$\begin{aligned}0 &\leq \pi^{(k)}(x) \leq 1, \text{ pour tout } x, \\ \pi^{(1)}(x) + \pi^{(2)}(x) + \dots + \pi^{(K)}(x) &= 1, \text{ pour tout } x.\end{aligned}$$

1.3.2 Fonction de lien et prédicteurs linéaires

Examinons dans un premier temps le cas où Y est une variable qualitative à deux modalités, 0 pour l'échec et 1 pour la réussite, et x est la seule variable explicative, supposée dans un premier temps continue. Une forme

très générale du modèle du lien entre Y et x est le suivant :

$$\mathbb{P}_x(Y = 1) = \pi(x; \beta),$$

où $\pi(x; \beta)$ est appelé le **prédicteur** du modèle.

Deux situations extrêmes et opposées peuvent illustrer, de manière schématique, l'étendue de la gamme des possibilités dans l'étude du lien entre x et Y . D'une part, l'absence totale d'influence de x sur la probabilité de réussite se traduit par $\mathbb{P}_x(Y = 1) = \pi$, où π est une constante. A l'opposé, une situation très simple d'une dépendance totale entre la probabilité de réussite et x est la suivante :

$$\mathbb{P}_x(Y = 1) = \begin{cases} 1 & \text{si } x < x_0 \\ 0 & \text{si } x \geq x_0 \end{cases}, \quad (1.2)$$

où x_0 est une valeur seuil au-delà de laquelle la probabilité de réussite est nulle. Ces deux situations extrêmes sont illustrées par le graphe de la figure 1.9.

De la même manière que l'utilisation du modèle de régression linéaire simple correspond au choix délibéré d'un prédicteur ayant la forme d'une droite, les prédicteurs que nous considérerons ici se limitent à un continuum de courbes allant de la fonction en escalier (1.2) à la fonction constante. Un exemple d'une telle courbe, par nature en forme de « S », est donné sur un graphique de la figure 1.9.

Le problème de la modélisation linéaire généralisée est de composer d'une part, avec cette forme sigmoïdale inhérente à la nature du problème et d'autre part, avec l'objectif d'une simplicité d'interprétation inspirée de la modélisation linéaire classique. Pour cela, le prédicteur $\pi(x; \beta)$ se définit de la manière suivante :

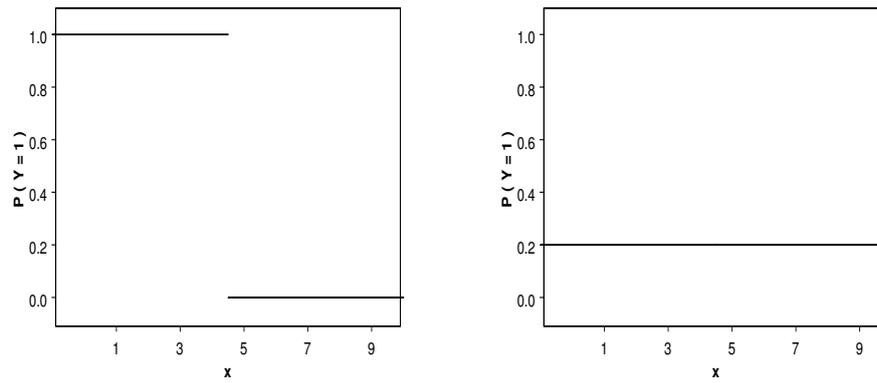
$$g[\pi(x; \beta)] = L(x; \beta), \quad (1.3)$$

où g est une fonction connue qui confère au prédicteur sa forme en « S » et $L(x; \beta) = \beta_0 + \beta_1 x$ est le **prédicteur linéaire** hérité de la modélisation linéaire. On appelle g la **fonction de lien** du modèle.

Dans le cas de plusieurs variables explicatives, la modélisation précédente se généralise naturellement par extension du prédicteur linéaire $L(x; \beta) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^p$. D'autre part, dans le cas où Y est une variable qualitative à K modalités, la modélisation décrite plus haut se généralise aussi par l'introduction de $K - 1$ prédicteurs linéaires $L(x; \beta^{(k)}) = \beta_0^{(k)} + \beta_1^{(k)} x^{(1)} + \beta_2^{(k)} x^{(2)} + \dots + \beta_p^{(k)} x^{(p)}$:

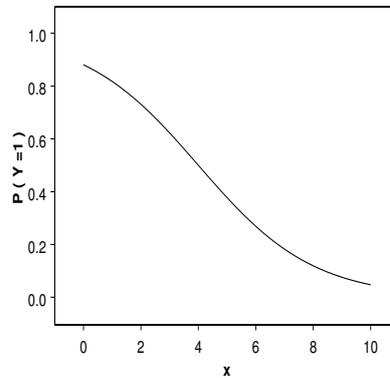
$$g[\pi^{(k)}(x)] = L(x; \beta^{(k)}), \quad k = 1, \dots, K - 1.$$

Remarquons aussi qu'à travers le prédicteur linéaire, qui fait la jonction entre la modélisation linéaire classique et la modélisation linéaire généralisée, il est très facile de définir un modèle d'analyse de variance pour une



$$\mathbb{P}_x(Y = 1) = \begin{cases} 1 & \text{si } x < x_0 \\ 0 & \text{si } x \geq x_0 \end{cases}$$

$$\mathbb{P}_x(Y = 1) = \pi$$



$$\mathbb{P}_x(Y = 1) = \pi(x; \beta)$$

FIG. 1.9: *Prédicteurs du modèle logit*

variable Y qualitative. Par exemple, si Y est qualitative à 2 modalités, le modèle d'analyse de la variance additif à deux facteurs prend la forme suivante :

$$g[\pi_{ij}] = \mu + \alpha_i + \beta_j,$$

avec les contraintes d'estimabilité habituelles sur les paramètres du modèle.

1.3.3 Y est qualitative à deux modalités

On se limite dans un premier temps au cas d'une variable à expliquer qualitative à deux modalités, codées par commodité 0 et 1. D'après l'expression (1.3), le choix du prédicteur n'est subordonné qu'au choix de la fonction de lien. Dans la suite, on se limite à la présentation de deux de ces prédicteurs : le modèle logit et modèle probit.

Le modèle logit

Le modèle logit correspond au choix suivant de la fonction de lien :

$$\begin{aligned} g(u) &= \ln\left(\frac{u}{1-u}\right), \\ &= \text{logit}(u). \end{aligned}$$

Par conséquent, le modèle logit définit la relation suivante entre $\mathbb{P}_x(Y = 1)$ et x :

$$\begin{aligned} \text{logit}(\mathbb{P}_x(Y = 1)) &= L(x; \beta), \\ \ln\left[\frac{\mathbb{P}_x(Y = 1)}{1 - \mathbb{P}_x(Y = 1)}\right] &= L(x; \beta). \end{aligned}$$

Cette modélisation met donc en avant le rapport entre la probabilité de réussite et le risque d'échec, que l'on désigne par le terme **cote** ou plus souvent par sa version anglo-américaine **odds** :

$$\text{odds}(x) = \frac{\mathbb{P}_x(Y = 1)}{1 - \mathbb{P}_x(Y = 1)}.$$

Dans le cas où Y est le résultat d'un traitement médical, si, pour une valeur de x donnée, la cote vaut 10, le traitement a alors 10 fois plus de chances de réussir que d'échouer. Choisir le modèle logit revient donc à opter pour un modèle linéaire du logarithme de la cote, aussi appelé **logodds** :

$$\begin{aligned} \ln[\text{odds}(x)] &= L(x; \beta), \\ \text{logodds}(x) &= L(x; \beta). \end{aligned}$$

Le choix de cette forme de fonction de lien est en partie motivé par des considérations d'ordre théorique qui donnent à la régression logistique le statut de modèle « canonique » ou « naturel » parmi tous les modèles possibles pour variables qualitatives. Cependant, l'exposé des arguments de théorie statistique qui sous-tendent cette propriété dépasse largement le cadre de cette introduction. Le lecteur désireux de parfaire sa connaissance théorique pourra assouvir sa curiosité en lisant l'ouvrage de [?] par exemple.

Exemple : promesse d'achat d'un produit alimentaire. Sur le graphique de la figure 1.8, le lien sigmoïdal entre la promesse d'achat et la note hédonique est matérialisé par la ligne brisée joignant les proportions d'acheteurs

calculées pour chaque note hédonique. Afin de confirmer la pertinence du lien logit, le graphique de la figure 1.10 représente le lien entre ces mêmes proportions après transformation logit et la note hédonique. Ce graphique met bien en avant la linéarité du lien entre le logodds et la note hédonique et permet même une évaluation visuelle des paramètres : $\beta_0 \approx -8$ et $\beta_1 \approx 1.5$. \square

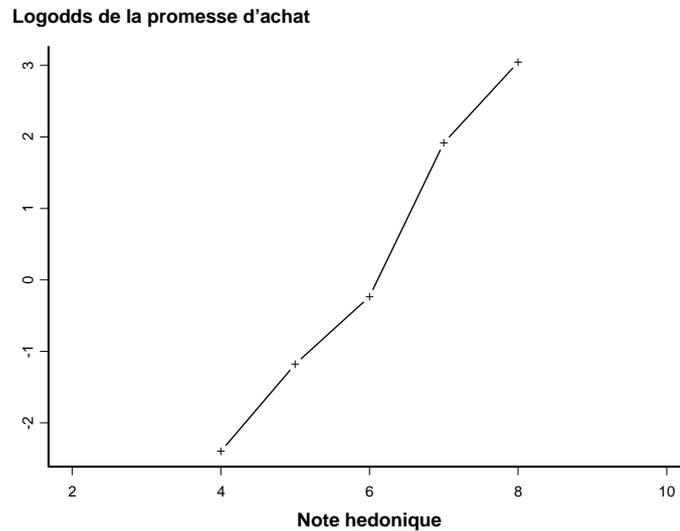


FIG. 1.10: Relation entre le logodds de la promesse d'achat et la note hédonique.

Interprétation des paramètres

L'introduction dans l'écriture du modèle d'une non-linéarité, par l'intermédiaire de la fonction de lien, rend l'interprétation des coefficients de la régression logistique moins immédiate que dans le cas de la régression linéaire. Dans le cas simple où x est la seule variable explicative et qu'elle est continue, il est cependant facile de vérifier que cette courbe de régression passe par le point d'abscisse $x = -\beta_0/\beta_1$ et d'ordonnée $\mathbb{P}_x(Y = 1) = 0.5$. Quelques calculs supplémentaires, ceux des dérivées premières et secondes de la surface de réponse, permettent aussi de voir, que ce point est également le seul point d'inflexion de la courbe de régression logistique. Enfin, il est intéressant de remarquer que le paramètre β_1 peut être vu comme un coefficient de pente au même titre que dans le cas de la régression linéaire. En effet, la fonction de lien étant une fonction croissante, le sens de variation de la courbe de régression logistique est également celui de $\beta_0 + \beta_1 x$: si $\beta_1 > 0$, alors la courbe de régression est croissante et inversement, si $\beta_1 < 0$, la courbe de régression est décroissante. D'autre part, plus la valeur absolue de β_1 est grande, plus la pente au point d'inflexion est grande. Pour reprendre les exemples extrêmes représentés par les graphes de la figure 1.9, la surface de réponse en escalier correspond au cas limite

où $|\beta_1| \rightarrow +\infty$, et, à l'opposé, la surface de réponse horizontale correspond au cas où $|\beta_1| = 0$. Les graphiques de la figure 1.11 illustrent ces propriétés de la surface de réponse par quelques exemples.

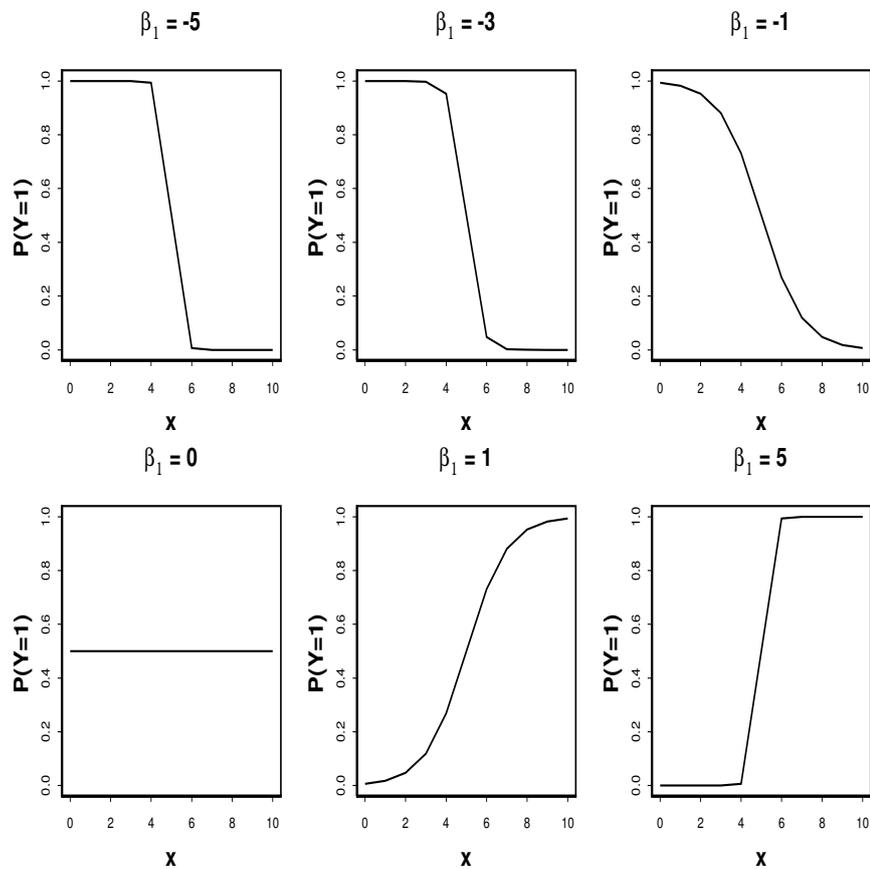


FIG. 1.11: *Interprétation des paramètres de la surface de réponse. Dans tous les exemples, $\beta_0 = -5\beta_1$.*

Odds ratio

L'analyse du lien entre Y et les variables explicatives est ramené à l'étude des variations de l'odds, ou du logodds, en fonction de ces variables explicatives. C'est dans cet objectif qu'il peut être intéressant d'étudier le rapport des odds entre deux situations caractérisées par des valeurs différentes des variables explicatives, aussi appelé risque relatif ou **odds ratio**, pour rester homogène dans le choix de la langue du spécialiste :

$$\begin{aligned} \text{OR}(x, x') &= \frac{\text{odds}(x)}{\text{odds}(x')} \\ &= \frac{\mathbb{P}_x(Y=1)\mathbb{P}_{x'}(Y=0)}{\mathbb{P}_x(Y=0)\mathbb{P}_{x'}(Y=1)}. \end{aligned}$$

Lorsque les variables explicatives sont quantitatives, il est fréquent de se limiter à l'examen d'odds ratio un peu particuliers, notés OR_j , pour lesquels x' est égal à x sauf pour une des variables explicatives, la j ème, où x' prend la valeur de x plus une unité. Il est facile de vérifier que, dans ce cas, $OR_j = \exp(\beta_j)$. Pour chaque variable, OR_j mesure donc l'impact d'une variation de une unité de la j ème variable (toutes choses égales par ailleurs) sur l'évolution de $\mathbb{P}_x(Y = 1)$. Par exemple, si Y désigne le résultat d'un traitement et si $OR_1 = 10$, alors on peut dire que lorsque la 1ère variable explicative augmente de 1 unité (toutes choses égales par ailleurs), les chances de réussite du traitement sont multipliées par 10. Dans la pratique, OR_j est donc aussi considéré comme un indicateur de l'influence de la j ème variable sur les variations de la variable à expliquer.

Lorsque les variables explicatives sont qualitatives, les odds ratio permettent de la même manière de comparer les risques entre différentes combinaisons des modalités de ces variables. Dans l'exemple introduit plus haut de la couleur du gras d'agneau, on distingue 4 combinaisons possibles des modalités des variables *Logement* et *Alimentation*. La comparaison des risques de coloration du gras entre deux de ces combinaisons de modalités peut donc s'appuyer sur le calcul du rapport des odds associés à chaque combinaison de modalités.

Le modèle probit

Toute la pertinence des arguments théoriques avancés pour le choix de la fonction de lien ne suffit parfois pas à convaincre les utilisateurs de ce type de modèle. Ainsi, dans de nombreux domaines d'applications et sous couvert d'arguments d'une pertinence aussi peu discutable, d'autres modèles ont émergé, associés à d'autres choix de fonctions de lien. Parce que nous ne prétendons pas à l'exhaustivité, nous avons choisi de ne présenter que l'un d'entre eux, très classique : le modèle probit.

Ce modèle est particulièrement naturel lorsque la variable binaire Y dont on peut observer les réalisations n'est que l'expression simplifiée d'une autre variable continue Y^* impossible à observer, parfois seulement conceptuelle. Par exemple, dans un contexte médical, une problématique classique est le classement d'un patient dans le système de catégories « malade » et « sain ». Bien entendu, ce classement n'est que l'expression simplifiée d'une variable quantifiant le niveau de santé d'un patient, variable qui n'est soit pas directement mesurable, soit la synthèse d'informations diverses concernant le patient.

Supposons donc que cette variable Y^* soit distribuée selon une loi normale d'espérance $\alpha_0 + \alpha_1 x$ et d'écart-type σ , où x est une variable explicative. La variable observable Y est donc définie de la manière suivante :

$$Y = \begin{cases} 1 & \text{si } Y^* \leq s \\ 0 & \text{si } Y^* > s \end{cases},$$

où s désigne une valeur seuil.

Dans ce contexte,

$$\begin{aligned}\mathbb{P}_x(Y = 1) &= \mathbb{P}_x(Y^* \leq s), \\ &= \mathbb{P}_x\left(\frac{Y^* - \alpha_0 - \alpha_1 x}{\sigma} \leq \frac{s - \alpha_0 - \alpha_1 x}{\sigma}\right), \\ &= \Phi\left(\frac{s - \alpha_0 - \alpha_1 x}{\sigma}\right),\end{aligned}$$

où Φ est la fonction de répartition de la loi normale centrée réduite. Par conséquent, si on pose $\beta_0 = (s - \alpha_0)/\sigma$ et $\beta_1 = -\alpha_1/\sigma$,

$$\begin{aligned}\mathbb{P}_x(Y = 1) &= \Phi(\beta_0 + \beta_1 x), \\ \Phi^{-1}[\mathbb{P}_x(Y = 1)] &= \beta_0 + \beta_1 x,\end{aligned}$$

où Φ^{-1} est la fonction de quantile de la loi normale centrée réduite.

Dans le cas d'une variable binaire qui correspond à l'expression simplifiée d'une variable distribuée selon une loi normale, une fonction de lien naturelle est donc la fonction Φ^{-1} .

De manière plus pragmatique, le choix d'une fonction de lien dépend beaucoup plus des habitudes inhérentes au domaine dans lequel on travaille et du logiciel de traitement de données que l'on a choisi que de justifications théoriques. Dans la grande majorité des cas, les ajustements selon un modèle ou un autre ne diffèrent d'ailleurs que très peu. Pour ces raisons, dans la suite, on se recentre sur le modèle logit mais son extension au modèle probit ou à tout autre modèle de régression pour variables binaires ne pose aucun nouveau problème conceptuel.

1.3.4 Y est qualitative à plus de deux modalités

Dans cette section, on suppose que Y est une variable à K modalités, notées C_1, C_2, \dots, C_K . Comme nous le mentionnons plus haut, il faut distinguer ici deux situations de nature différente : le cas où Y est qualitative nominale et le cas où Y est qualitative ordinale.

Y est qualitative nominale

Dans le cas où Y a 2 modalités, le fait que le modèle s'appuie sur une représentation linéaire du logodds impose d'une certaine manière la modalité 0 comme une valeur de référence. En effet, l'analyse de $\mathbb{P}_x(Y = 1)$ à travers l'odds est fondé sur une comparaison avec le risque que $Y = 0$. Par extension, nous allons ici aussi choisir une modalité de référence. Pour fixer les idées et les notations, on décide dans la suite que cette modalité de référence est C_K . En pratique, il n'est pas inutile de s'autoriser quelques minutes de réflexion sur le choix de cette modalité de référence. En effet, ce choix conditionne l'interprétation des résultats du traitement. En particulier, dans le cas où une modalité peut être considérée comme une valeur *témoin*, cette modalité s'impose naturellement comme une référence.

Le modèle se présente donc comme un ensemble de $K - 1$ modèles logit pour les $K - 1$ couples de modalités (C_j, C_K) , $j = 1, \dots, K - 1$:

$$\begin{aligned} \ln [\mathbb{P}_x(Y = C_1)/\mathbb{P}_x(Y = C_K)] &= L(x; \beta^{(1)}), \\ \ln [\mathbb{P}_x(Y = C_2)/\mathbb{P}_x(Y = C_K)] &= L(x; \beta^{(2)}), \\ &\vdots \\ \ln [\mathbb{P}_x(Y = C_{K-1})/\mathbb{P}_x(Y = C_K)] &= L(x; \beta^{(K-1)}). \end{aligned}$$

Ce modèle est traditionnellement appelé le **modèle logistique multinomial**.

La présentation suivante du modèle est équivalente et met en avant de manière plus directe les équations modélisant l'évolution des probabilités $\mathbb{P}_x(Y = C_j)$ en fonction de x :

$$\begin{aligned} \mathbb{P}_x(Y = C_1) &= \frac{\exp [L(x; \beta^{(1)})]}{1 + \exp [L(x; \beta^{(1)})] + \dots + \exp [L(x; \beta^{(K-1)})]}, \\ \mathbb{P}_x(Y = C_2) &= \frac{\exp [L(x; \beta^{(2)})]}{1 + \exp [L(x; \beta^{(1)})] + \dots + \exp [L(x; \beta^{(K-1)})]}, \\ &\vdots \\ \mathbb{P}_x(Y = C_{K-1}) &= \frac{\exp [L(x; \beta^{(K-1)})]}{1 + \exp [L(x; \beta^{(1)})] + \dots + \exp [L(x; \beta^{(K-1)})]}, \\ \mathbb{P}_x(Y = C_K) &= \frac{1}{1 + \exp [L(x; \beta^{(1)})] + \dots + \exp [L(x; \beta^{(K-1)})]}. \end{aligned}$$

Exemple : lien entre génotype et épaisseur tissulaire. Après ajustement des paramètres du modèle de régression logistique présenté ci-dessus, le graphique de la figure 1.12 montre l'évolution des probabilités des différents types génétiques en fonction de l'épaisseur de muscle. □

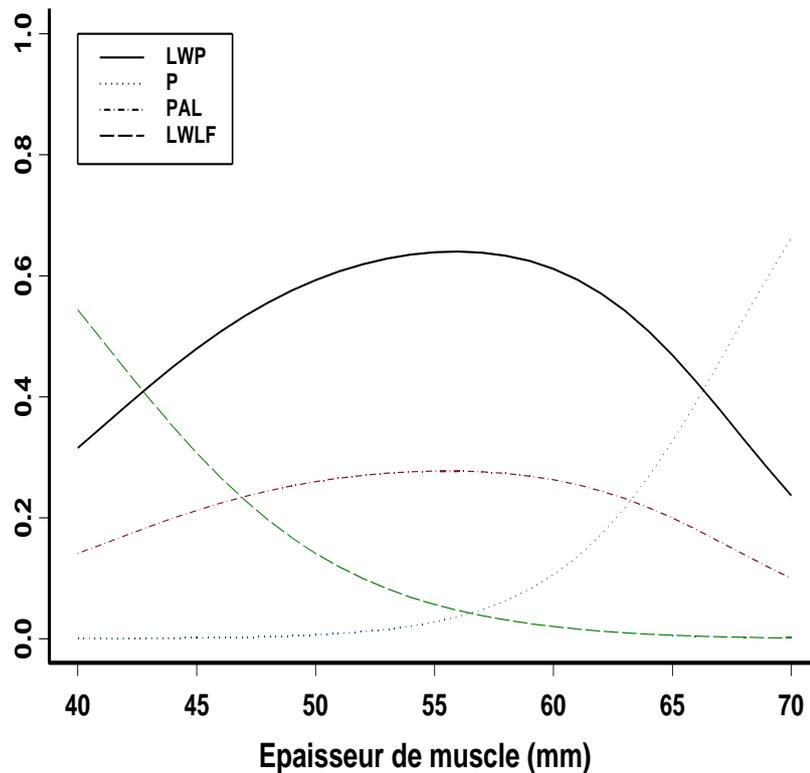


FIG. 1.12: *Modèle logistique multinomial représentant le lien entre type génétique et épaisseur de muscle.*

***Y* est qualitative ordinale**

On suppose ici qu'il existe un ordre naturel entre les modalités de la variable Y . Par convention, on considère que $C_1 \leq C_2 \leq \dots \leq C_K$. Il existe plusieurs manières de prendre en compte cet ordre dans la modélisation selon la nature des relations entre les modalités. On recense essentiellement 3 types de modèles :

- le modèle à **risques adjacents** : la cote d'une modalité est calculée par rapport à la modalité suivante.
- le modèle **séquentiel** : la cote d'une modalité est calculée par rapport aux modalités supérieures.
- le modèle **cumulatif** : le risque que la variable à expliquer soit inférieure à une modalité est calculé par rapport aux modalités supérieures.

Nous présentons ici de manière plus détaillé le modèle cumulatif :

$$\begin{aligned} \ln [\mathbb{P}_x(Y \leq C_1)/\mathbb{P}_x(Y > C_1)] &= L(x; \beta^{(1)}), \\ \ln [\mathbb{P}_x(Y \leq C_2)/\mathbb{P}_x(Y > C_2)] &= L(x; \beta^{(2)}), \\ &\vdots \\ \ln [\mathbb{P}_x(Y \leq C_{K-1})/\mathbb{P}_x(Y > C_{K-1})] &= L(x; \beta^{(K-1)}). \end{aligned}$$

Exemple : maladie du poumon chez les mineurs (McCullagh & Nelder, 1983). On s'intéresse à la relation entre l'état d'avancement d'une maladie des poumons touchant des mineurs et l'ancienneté de ces mineurs au travail. Le diagnostic de la gravité de la maladie se fonde sur l'examen d'une radio des poumons et conclut à l'état normal, moyennement atteint ou sévèrement atteint des poumons. Les données sont issues de l'examen de 371 mineurs et reproduites dans le tableau 1.7. Notons qu'elles se présentent sous la forme de données groupées. Ces données sont traitées en détail dans le chapitre ??.

Ancienneté (années)	État des poumons		
	Normal	Moyen	Sévère
5.8	98	0	0
15.0	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46.0	12	6	10
51.5	4	2	5

TAB. 1.7: Ancienneté et état des poumons de 371 mineurs (McCullagh & Nelder, 1983).

Dans cette situation, la variable à expliquer Y est l'état d'avancement de la maladie, variable qualitative à 3 modalités ordonnées, et la variable explicative est l'ancienneté x , variable quantitative continue. Le modèle

cumulatif s'écrit donc ici de la manière suivante :

$$\begin{aligned}\ln \left[\frac{\mathbb{P}_x(Y = \text{normal})}{\mathbb{P}_x(Y = \text{moyen}) + \mathbb{P}_x(Y = \text{sévère})} \right] &= \beta_0^{(1)} + \beta_1^{(1)}x, \\ \ln \left[\frac{\mathbb{P}_x(Y = \text{normal}) + \mathbb{P}_x(Y = \text{moyen})}{\mathbb{P}_x(Y = \text{sévère})} \right] &= \beta_0^{(2)} + \beta_1^{(2)}x.\end{aligned}$$

Il est important de remarquer que, sous cette forme, sans contrainte particulière sur $\beta_1^{(1)}$ et $\beta_1^{(2)}$, il est possible que les prédicteurs linéaires $\beta_0^{(1)} + \beta_1^{(1)}x$ et $\beta_0^{(2)} + \beta_1^{(2)}x$ se croisent et qu'alors, pour certaines valeurs de x , on ait :

$$\begin{aligned}\ln \left[\frac{\mathbb{P}_x(Y = \text{normal})}{\mathbb{P}_x(Y = \text{moyen}) + \mathbb{P}_x(Y = \text{sévère})} \right] &> \ln \left[\frac{\mathbb{P}_x(Y = \text{normal}) + \mathbb{P}_x(Y = \text{moyen})}{\mathbb{P}_x(Y = \text{sévère})} \right], \\ \frac{\mathbb{P}_x(Y = \text{normal})}{\mathbb{P}_x(Y = \text{moyen}) + \mathbb{P}_x(Y = \text{sévère})} &> \frac{\mathbb{P}_x(Y = \text{normal}) + \mathbb{P}_x(Y = \text{moyen})}{\mathbb{P}_x(Y = \text{sévère})}.\end{aligned}$$

Or, à l'évidence,

$$\frac{\mathbb{P}_x(Y = \text{normal}) + \mathbb{P}_x(Y = \text{moyen})}{\mathbb{P}_x(Y = \text{sévère})} \geq \frac{\mathbb{P}_x(Y = \text{normal}) + \mathbb{P}_x(Y = \text{moyen})}{\mathbb{P}_x(Y = \text{moyen}) + \mathbb{P}_x(Y = \text{sévère})}.$$

Par conséquent, sans contrainte particulière sur les coefficients de pente $\beta_1^{(1)}$ et $\beta_1^{(2)}$, il est possible que :

$$\begin{aligned}\frac{\mathbb{P}_x(Y = \text{normal})}{\mathbb{P}_x(Y = \text{moyen}) + \mathbb{P}_x(Y = \text{sévère})} &> \frac{\mathbb{P}_x(Y = \text{normal}) + \mathbb{P}_x(Y = \text{moyen})}{\mathbb{P}_x(Y = \text{moyen}) + \mathbb{P}_x(Y = \text{sévère})}, \\ \mathbb{P}_x(Y = \text{normal}) &> \mathbb{P}_x(Y = \text{normal}) + \mathbb{P}_x(Y = \text{moyen}),\end{aligned}$$

ce qui est manifestement fâcheux.

Une manière de contourner ce problème consiste à contraindre les coefficients de pente du modèle à être les mêmes pour les 2 prédicteurs linéaires : $\beta_1^{(1)} = \beta_1^{(2)} = \beta$. Dans certaines situations, cette contrainte de parallélisme entre les prédicteurs linéaires peut apparaître comme une solution radicale au problème d'intersection des prédicteurs linéaires qui peut prêter à discussions. Ici, elle conduit au graphique de la figure 1.13 représentant l'évolution des risques en fonction de l'ancienneté.

Sous cette nouvelle contrainte, notons que les logarithmes des rapports des risques relatifs pour deux valeurs x_1 et x_2 de l'ancienneté sont proportionnels à l'écart $x_1 - x_2$:

$$\begin{aligned}\ln \left[\frac{\mathbb{P}_{x_1}(Y = \text{normal})}{\mathbb{P}_{x_1}(Y = \text{moyen}) + \mathbb{P}_{x_1}(Y = \text{sévère})} \frac{\mathbb{P}_{x_2}(Y = \text{normal})}{\mathbb{P}_{x_2}(Y = \text{moyen}) + \mathbb{P}_{x_2}(Y = \text{sévère})} \right] &= \beta[x_1 - x_2], \\ \ln \left[\frac{\mathbb{P}_{x_1}(Y = \text{normal}) + \mathbb{P}_{x_1}(Y = \text{moyen})}{\mathbb{P}_{x_1}(Y = \text{sévère})} \frac{\mathbb{P}_{x_2}(Y = \text{normal}) + \mathbb{P}_{x_2}(Y = \text{moyen})}{\mathbb{P}_{x_2}(Y = \text{sévère})} \right] &= \beta[x_1 - x_2]. \square\end{aligned}$$

De manière générale, lorsque le modèle cumulatif est simplifié par l'hypothèse d'égalité des pentes des prédicteurs linéaires, on parle de **modèle à risques proportionnels**.

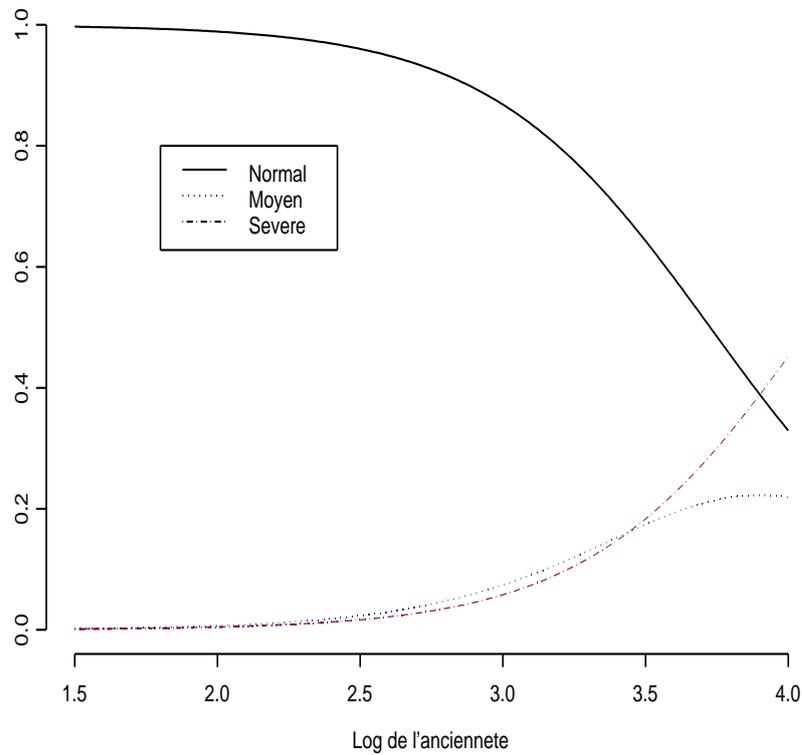


FIG. 1.13: Ajustement du modèle à risques proportionnels aux données sur la maladie des poumons des mineurs.

1.3.5 Y est une variable de comptage

Dans cette section, Y est distribuée selon une loi de Poisson dont on modélise l'intensité par l'intermédiaire de la fonction de lien logarithme, garantissant le fait que l'intensité soit positive :

$$\ln[\lambda(x; \beta)] = L(x; \beta).$$

Le modèle précédent est appelé **modèle log-linéaire**.

Comme nous l'avons mentionné dans le cadre du modèle pour variable qualitative à 2 modalités, ce choix de fonction de lien est également motivé par des arguments d'ordre théorique. D'ailleurs, de la même manière que la fonction de lien probit peut être préférée à la fonction de lien logit, la plupart des logiciels proposent ici aussi d'autre choix de fonction de lien que le logarithme.

Exemple : comportement de clients en magasin. Si Y désigne le nombre de bouteilles examinées par un client pendant son séjour au magasin et x le logarithme de la durée de ce séjour, alors on propose le modèle suivant, formalisant le lien entre Y et x :

$$\begin{aligned}\ln[\mathbb{E}_x(Y)] &= \beta_0 + \beta_1 x, \\ \ln[\lambda(x; \beta)] &= \beta_0 + \beta_1 x.\end{aligned}$$

Le graphique de la figure 1.14 illustre le lien entre le nombre de bouteilles examinées par un client et la durée de son séjour dans le magasin par l'ajustement du modèle précédent. \square

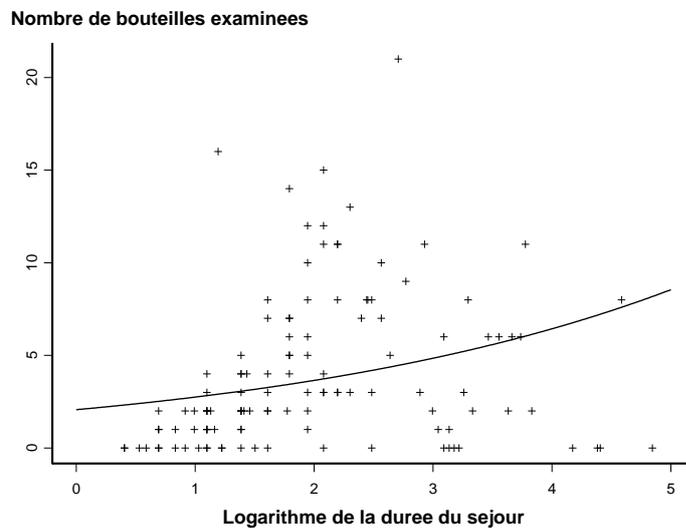


FIG. 1.14: *Modèle log-linéaire pour les données de comportements de clients en magasin.*

1.4 Estimation des paramètres

La procédure d'estimation des paramètres consiste à attribuer des valeurs aux paramètres du modèle, à partir de l'observation conjointe des variables explicatives et de la variable à expliquer. Comme dans le cas plus connu du modèle linéaire, ces paramètres sont ici aussi les coefficients de prédicteurs linéaires.

1.4.1 Première impression graphique

Dans les situations les plus simples, il est intéressant de faire précéder la phase d'estimation des paramètres du modèle par l'élaboration de quelques graphiques. Par exemple, les graphiques des figures 1.2 et 1.3 représentent chacun un nuage de points d'abscisse la valeur de la variable explicative et d'ordonnée la valeur de la variable à expliquer. Comme dans le cas de la régression linéaire, ce type de graphique est traditionnellement utilisé, préalablement à l'ajustement, pour décrire l'évolution des valeurs de Y en fonction de celles de x . Il peut par exemple suggérer d'utiliser un autre modèle que le modèle logit si le prédicteur sigmoïdal apparaît irréaliste.

1.4.2 Vraisemblance d'un modèle

Dans le cas du modèle linéaire, les estimateurs sont obtenus par minimisation d'un critère, celui des moindres carrés, défini comme une distance entre les données observées et le modèle ajusté. Dans le contexte du modèle linéaire généralisé, la procédure d'estimation repose également sur l'optimisation d'un critère, appelé **vraisemblance du modèle**, que l'on peut aussi décrire intuitivement comme une mesure d'adéquation entre les valeurs des paramètres et les données observées.

Illustrons cette notion de vraisemblance à partir de l'exemple des données d'intention d'achat représentées sur le graphique de la figure 1.2. Comme on l'a vu précédemment, le modèle adapté à la description du lien entre la promesse d'achat, Y codée 1 si le dégustateur est prêt à acheter le sandwich et 0 sinon, et la note hédonique, x variable quantitative, est le modèle logit défini comme suit :

$$\text{logit}[\mathbb{P}_x(Y = 1)] = \beta_0 + \beta_1 x,$$

ou encore :

$$\mathbb{P}_x(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Ce modèle permet en particulier de calculer, pour chaque valeur possible des paramètres (β_0, β_1) et chaque valeur x_i de la note hédonique dans l'échantillon, la probabilité que la promesse d'achat Y_i associée prenne la valeur $y_i = 0$ ou 1 :

$$\mathbb{P}_{x_i}(Y_i = y_i) = \begin{cases} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} & \text{si } y_i = 1 \\ 1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} & \text{si } y_i = 0. \end{cases}$$

Plus généralement, il est donc possible, pour chaque valeur possible des paramètres (β_0, β_1) , de calculer la probabilité d'une configuration de données $(x_i, y_i)_{i=1, \dots, n}$ quelconque :

$$\mathbb{P}_x(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) = \mathbb{P}_{x_1}(Y_1 = y_1) \mathbb{P}_{x_2}(Y_2 = y_2) \dots \mathbb{P}_{x_n}(Y_n = y_n).$$

En particulier, par ce mode de calcul, on peut associer à chaque valeur des paramètres (β_0, β_1) la probabilité, notée $\mathcal{V}(\beta)$, de la configuration de données telle qu'effectivement observée et représentée sur le graphique de la figure 1.2: $(x_1 = 6, y_1 = 1)$, $(x_2 = 7, y_2 = 0)$, $(x_3 = 5, y_3 = 0)$, $(x_4 = 9, y_4 = 1)$, ...

Illustrons l'intérêt de ce calcul : on a donné plus haut une estimation visuelle des paramètres β_0 et β_1 à partir du graphique de la figure 1.11. Rappelons que nous avons déduit de ce graphique que $\beta_0 \approx -8$ et $\beta_1 \approx 1.5$. On peut maintenant mesurer la pertinence de ces approximations par le calcul de $\mathcal{V}(-8, 1.5) = 2.88.10^{-28}$. La très faible valeur de cette probabilité ne doit pas surprendre : elle s'obtient par le produit de $n = 36$ probabilités, dont certaines proches de 0. Toutefois, pour se ramener à une échelle de valeurs moins extraordinaires, il est d'usage de retenir comme mesure de l'adéquation entre les valeurs des paramètres et les données le logarithme de $\mathcal{V}(\beta)$, ici -63.41 . Si maintenant on réalise le même calcul pour une autre valeur des paramètres, par exemple $\beta_0 = -4$, $\beta_1 = 1.5$, alors on obtient -210.70 . On en déduit que la première approximation visuelle, $\beta_0 = -8$ et $\beta_1 = 1.5$, est plus en adéquation avec les observations que la deuxième proposition ($\beta_0 = -4$, $\beta_1 = 1.5$), ce qui est confirmé par le graphique de la figure 1.15 représentant les modèles pour ces deux jeux de paramètres.

On appelle $\mathcal{V}(\beta)$ la vraisemblance du modèle. Le graphique de la figure 1.16 montre l'évolution de $\mathcal{V}(\beta)$ en fonction de β dans le cas des données d'intention d'achat.

1.4.3 Maximisation de la vraisemblance

Ayant choisi de mesurer l'adéquation d'un modèle aux données par la vraisemblance, il est naturel de calculer l'estimateur $\hat{\beta}$ de β par maximisation de la vraisemblance :

$$\hat{\beta} = \arg \max_{\beta} \mathcal{V}(\beta)$$

On dit alors que $\hat{\beta}$ est l'**estimateur du maximum de vraisemblance** de β .

A l'image de la représentation graphique de la figure 1.16, la vraisemblance du modèle est à l'évidence une fonction continue et même dérivable des paramètres, ce qui en théorie définit un cadre idéal à la recherche des points maximaux. Toutefois, la complexité analytique de cette vraisemblance rend le plus souvent impossible le calcul de formes explicites d'estimateurs. Les techniques de maximisation mises en œuvre dans les logiciels de traitement statistique des données s'appuient en fait sur des algorithmes itératifs convergeant vers les estimateurs du maximum de vraisemblance. Il faut donc être indulgent avec son logiciel préféré de traitement de données : si le modèle contient un trop grand nombre (l'appréciation du terme « trop » dépend du logiciel, de la performance de la machine, ...) de paramètres, il n'est pas rare que l'algorithme itératif renonce à donner une estimation de ces paramètres. La présentation des algorithmes d'estimation et de leurs propriétés dépassent les

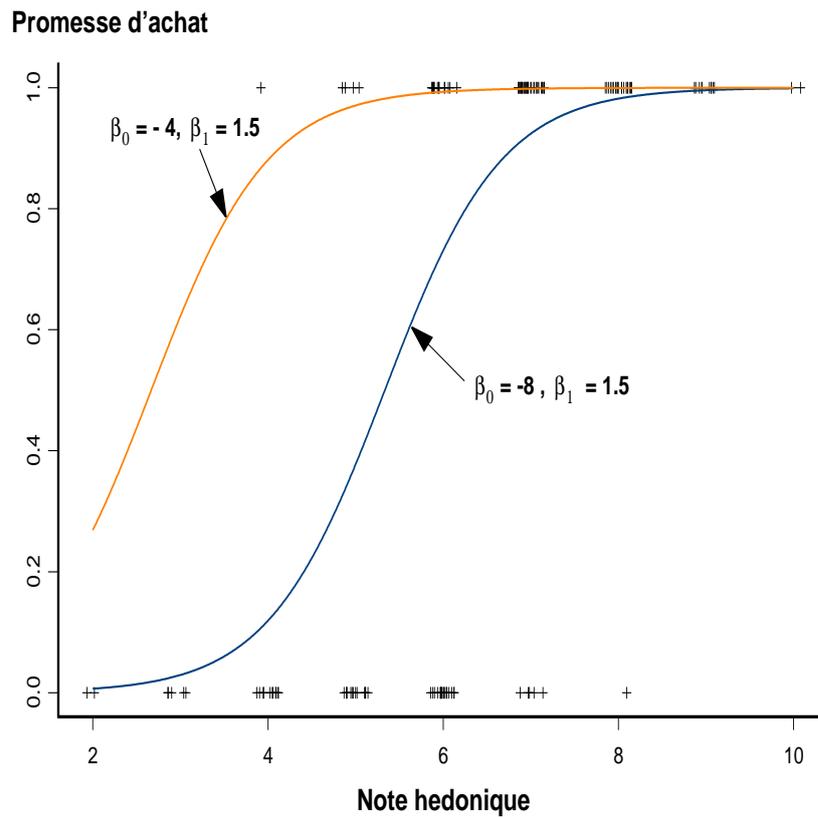


FIG. 1.15: Comparaison de 2 modèles logit pour les données d'intention d'achat.

objectifs de cette introduction. Mais, là aussi, le lecteur frustré par le manque d'ambition affiché par les auteurs pourra se reporter avantageusement à [?].

Le tableau 1.8 reproduit un extrait du listage des résultats de l'ajustement d'un modèle logit aux données d'intention d'achat. Outre l'estimation des coefficients, il faut noter qu'une information concernant la bonne marche de l'algorithme d'estimation est aussi proposée. Ici, cette information se limite au nombre d'étapes de l'algorithme avant convergence. Le graphique de la figure 1.17 présente le tracé de la surface de réponse ajustée par la méthode du maximum de vraisemblance, dans le cas des données d'intention d'achat.

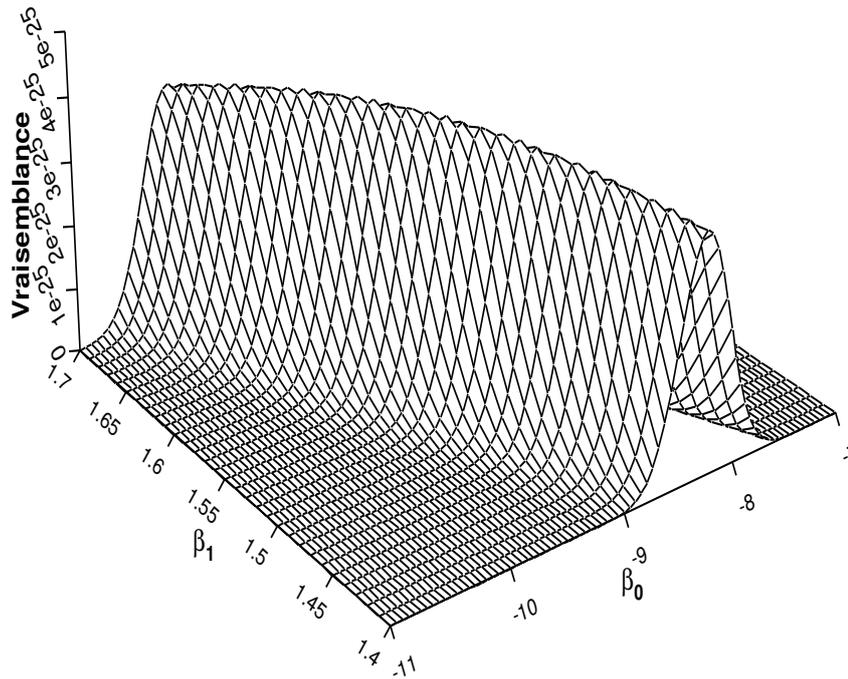


FIG. 1.16: *Vraisemblance du modèle logit. Exemple des données d'intention d'achat*

Coefficients:

	Value	Std. Error	t value
(Intercept)	-9.187662	1.6574553	-5.543234
globale	1.543768	0.2647332	5.831412

Number of Fisher Scoring Iterations: 5

TAB. 1.8: *Estimation par la méthode du maximum de vraisemblance dans le cas des données d'intention d'achat.*

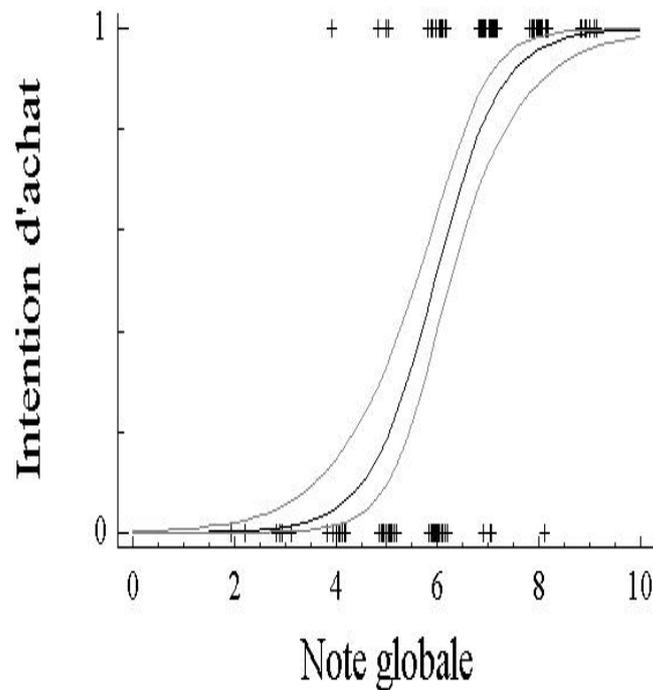


FIG. 1.17: Ajustement d'un modèle de régression logistique (courbe noire). Les courbes en gris sont les limites d'un intervalle de confiance d'estimation de niveau 95 %. Exemple des données d'intention d'achat.

1.4.4 Propriétés des estimateurs du maximum de vraisemblance

La méthode du maximum de vraisemblance n'est pas propre au modèle linéaire généralisé. C'est même d'une certaine manière la méthode d'estimation de référence s'il on en juge par la diversité des situations dans lesquelles elle est utilisée. Cette notoriété est pour une grande part liée au fait que, de manière générale, les estimateurs du maximum de vraisemblance ont de bonnes propriétés statistiques. Le lecteur avide de connaissances plus approfondies (celui auquel on fait référence depuis le début et qui a déjà lu courageusement et minutieusement citedobson et [?]) aura trouvé par exemple des réponses à ses questions sur la théorie du maximum de vraisemblance dans le livre de [?].

Dans le cas présent, on retient de cette théorie du maximum de vraisemblance que, lorsque la taille n de l'échan-

tillon est suffisamment grande, on peut affirmer que les estimateurs sont distribués approximativement selon une loi normale centrée sur les paramètres qu'ils estiment et dont il est possible de calculer la variance. Ceci explique que, dans le cas des données d'intention d'achat, le tableau 1.8 donne une estimation des écarts-types de $\hat{\beta}_0$ et $\hat{\beta}_1$. Les conséquences pratiques de ces calculs sont nombreuses : tests de la nullité des paramètres, calculs d'intervalles de confiance, etc. Ainsi, le listage du tableau 1.8 donne par exemple les valeurs de statistiques de Student dont on sait qu'elles sont utiles aux tests de nullité des paramètres. Le graphique de la figure 1.17 fournit également le tracé des bornes inférieures et supérieures de l'intervalle de confiance d'ajustement de niveau 95% ainsi calculé.

Dans le cas du modèle logit avec une seule variable explicative quantitative, on montre par exemple que :

$$\hat{\beta}_1 - \beta_1 \sim \mathcal{N} \left[0, \sqrt{\frac{1}{s_{\pi}^2 S_x^2}} \right] \quad (1.4)$$

où $s_{\pi}^2 = \sum_{i=1}^n \pi_i(1 - \pi_i)$, $S_x^2 = \sum_{i=1}^n p_i(x_i - m_1)^2$ est une variance pondérée des valeurs x_i et $m_1 = \sum_{i=1}^n p_i x_i$ est une moyenne pondérée des valeurs x_i . Les poids $(p_i)_{i=1, \dots, n}$, positifs et tels que $\sum_{i=1}^n p_i = 1$, sont ici définis de la manière suivante :

$$p_i = \frac{\pi_i(1 - \pi_i)}{\sum_{i=1}^n \pi_i(1 - \pi_i)}.$$

Ils donnent donc une importance plus grande aux points pour lesquels π_i est proche de 0.5, à savoir les points les plus proches de l'inflexion de la surface de réponse. Comme dans le cas de la régression linéaire simple, il découle de l'expression (1.4) que l'écart-type des estimateurs est d'autant plus faible, et donc la précision de l'estimation d'autant meilleure, que les valeurs x_i sont dispersées.

1.5 Tests de validité et comparaison de modèles

Une des questions récurrentes lorsque l'on a ajusté un modèle porte sur l'existence effective ou non d'une relation entre Y et les variables explicatives. Pour reprendre l'exemple simple du modèle logit avec une seule variable explicative quantitative, le problème soulevé peut se ramener à la question suivante : compte tenu de l'information dont on dispose, peut-on considérer que le prédicteur du modèle diffère significativement de la droite horizontale ? Dans un contexte statistique, on formalise traditionnellement ce type de problème par le test d'hypothèses suivant :

$$\begin{cases} H_0 & : \quad x \text{ n'a pas d'influence sur } Y \\ H_1 & : \quad x \text{ a une influence sur } Y \end{cases}$$

Il s'agit donc ici de comparer deux modèles emboîtés : le modèle avec coefficients de pente non-nuls (sous H_1), que l'on appelle aussi **modèle complet**, et le modèle avec tous les coefficients de pente nuls (sous H_0), que l'on appelle aussi **modèle nul**. Une stratégie intuitive consiste à comparer ces deux modèles sur la base d'une mesure de la qualité de leur ajustement aux données.

Exemple : résistance du tournesol au mildiou. Ici, la variable à expliquer est le taux de sporulations dans des bacs de 7 plantules et les variables explicatives sont d'une part le logarithme de la dose de fongicide, variable quantitative, et d'autre part la race du mildiou, variable qualitative à 2 modalités. Si $\pi_i(x)$ désigne la probabilité de sporulation pour un mildiou de race i ($i = 1$ pour la race résistante, $i = 2$ pour la race sensible) et une log-dose x , alors le modèle complet est le suivant :

$$\text{logit}[\pi_i(x)] = \mu + \alpha_i + [\beta + \delta_i]x, \quad (1.5)$$

où $\alpha_2 = \delta_2 = 0$. Notons que l'on peut préférer les contraintes suivantes : $\alpha_1 + \alpha_2 = \delta_1 + \delta_2 = 0$, même si elles ne traduisent pas le fait que la race sensible « S » est, d'une certaine manière, une modalité « témoin ». Le modèle nul fait état d'une absence de relations entre la probabilité de sporulation d'une part et la race du mildiou et la dose de fongicide d'autre part :

$$\text{logit}[\pi_i(x)] = \mu.$$

Remarquons que valider le modèle complet sur la base d'une comparaison avec le modèle nul est par nature une approche minimaliste puisqu'elle se fonde sur la mesure de ce que le modèle complet apporte par rapport à un sous-modèle trivial. Ici, grâce aux répétitions dont on dispose pour chaque valeur de la dose de fongicide, il est aussi possible de comparer le modèle complet à un sur-modèle, appelé **modèle saturé** dans lequel il y a autant de paramètres que de combinaisons dose \times race :

$$\text{logit}[\pi_i(x)] = \mu_i(x).$$

Par opposition avec la comparaison au modèle nul, cette nouvelle approche consiste à mesurer la diminution de la qualité de l'ajustement par le modèle complet par rapport au sur-modèle le plus paramétré. Le graphique de la figure 1.18 représente l'ajustement de ces trois modèles. \square

1.5.1 Déviance du modèle

Comme nous l'avons déjà mentionné auparavant, la vraisemblance $\mathcal{V}(\beta)$ constitue une mesure pertinente de l'adéquation entre un modèle de paramètre β et des données. Les vraisemblances maximisées \mathcal{V}_1 et \mathcal{V}_0 du

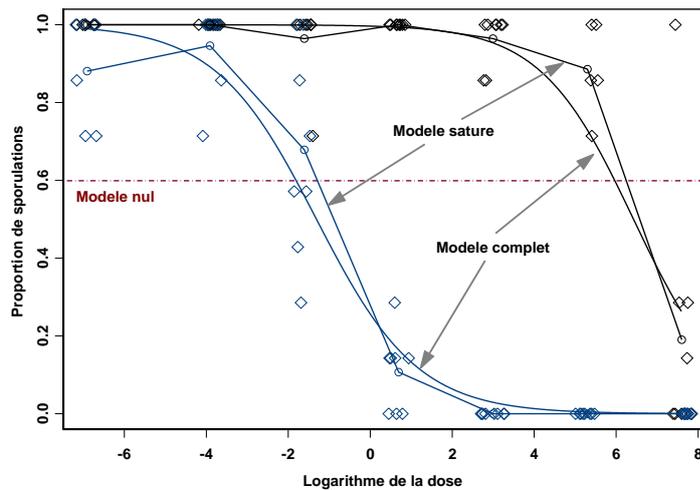


FIG. 1.18: Ajustement du modèle saturé, du modèle complet et du modèle nul sur les données « tournesol ».

modèle complet et du modèle nul respectivement sont donc des indicateurs naturels de la qualité de leur ajustement. On compare alors les modèles par le rapport entre les vraisemblances des deux modèles :

$$RV = \frac{\mathcal{V}_0}{\mathcal{V}_1}. \quad (1.6)$$

Plus précisément, il découle des propriétés des estimateurs du maximum de vraisemblance que, sous l'hypothèse nulle et pour de grandes valeurs de n , on peut considérer qu'approximativement $-2 \ln RV \sim \chi_{k-1}^2$, où k est le nombre de paramètres du modèle complet. Cette propriété permet la construction de la stratégie de test et l'évaluation des risques d'erreurs.

La quantité $-2 \ln RV$, que nous noterons \mathcal{D} , est appelée **déviante du modèle**. Lorsque cette déviante est proche de 0, la statistique RV est proche de 1 et donc le modèle complet n'est pas plus intéressant que le modèle nul. Inversement, une grande valeur de \mathcal{D} indique une statistique RV proche de 0, et consécutivement, incite à considérer que le modèle complet est plus intéressant que le modèle nul.

Par analogie avec la décomposition de la variance à l'origine des tests de validité des modèles de régression linéaire, on peut parler ici d'équation d'**analyse de la déviante** :

$$\begin{aligned} \mathcal{V}_0 &= \frac{\mathcal{V}_0}{\mathcal{V}_1} \mathcal{V}_1, \\ \underbrace{-2 \ln \mathcal{V}_0}_{\mathcal{D}_0} &= \underbrace{-2 \ln RV}_{\mathcal{D}} \underbrace{-2 \ln \mathcal{V}_1}_{\mathcal{D}_r}, \\ \mathcal{D}_0 &= \mathcal{D} + \mathcal{D}_r, \end{aligned}$$

où \mathcal{D}_0 est appelée déviance du modèle nul, ou **déviance totale**, et \mathcal{D}_r est appelée **déviance résiduelle**. L'interprétation de cette équation d'analyse de la déviance est analogue à celle de l'équation d'analyse de la variance d'un modèle de régression linéaire : la validité du modèle se lit dans la répartition de la déviance totale entre la déviance du modèle et la déviance résiduelle.

On peut également considérer qu'approximativement, lorsque la taille de l'échantillon est suffisamment grande, $\mathcal{D}_r \sim \chi_{n-k-1}^2$.

Exemple : résistance du tournesol au mildiou. Un extrait du listage des résultats de l'ajustement du modèle (1.5) aux données «tournesol» est fourni dans le tableau 1.9. Le tableau 1.10 reproduit une table d'analyse de la déviance du modèle complet. Comme le graphique de la figure 1.3 le laissait penser, la probabilité critique dans la table d'analyse de la déviance conduit à considérer que le modèle complet est informatif. \square

Coefficients				
Effets	Estimation	Écart-type	Student	Probabilité critique
μ	-1.065	0.213	-5.003	2.304e-6
α_1	6.720	0.803	8.37	2.796e-13
β	-0.812	0.089	-9.113	6.439e-15
δ_1	-0.0686	0.149	-0.460	6.464e-1

TAB. 1.9: Coefficients estimés du modèle complet pour les données « tournesol »

Analyse de la déviance			
	Déviance	Degré de liberté	Probabilité critique
Modèle nul	836.739	107	
Modèle complet	689.422	3	0
Résiduelle	147.316	104	

TAB. 1.10: Analyse de la déviance du modèle complet pour les données « tournesol »

1.5.2 Test de la validité par rapport à un sous-modèle

Lorsque le test précédent a conclu au fait que le modèle complet était intéressant, une analyse plus détaillée consiste à tester l'intérêt du modèle complet non plus par rapport au modèle nul mais par rapport à un sous-modèle. Supposons par exemple que le modèle complet, noté M_p soit fondé sur p variables explicatives $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ et qu'un sous-modèle particulièrement intéressant, noté M_q ne soit basé que sur q de ces p variables explicatives, par exemple $x^{(1)}, x^{(2)}, \dots, x^{(q)}$. On dit alors que M_q est un **sous-modèle** de M_p . La problématique peut alors se résumer en un test d'hypothèses :

$$\begin{cases} H_0 & : M_p \text{ est aussi intéressant que } M_q \\ H_1 & : M_p \text{ est plus intéressant que } M_q \end{cases}$$

En d'autres termes, sous l'hypothèse nulle, les coefficients de pente des variables $x^{(q+1)}, x^{(q+2)}, \dots, x^{(p)}$ dans le prédicteur linéaire de M_p sont non-nuls alors que sous l'hypothèse alternative, ils sont nuls.

Exemple : résistance du tournesol au mildiou. Si $\pi_i(x)$ désigne la probabilité de sporulation pour un mildiou de race i ($i = 1$ pour la race résistante, $i = 2$ pour la race sensible) et une log-dose x , alors le modèle complet, noté M_2 , est le suivant :

$$\text{logit}[\pi_i(x)] = \mu + \alpha_i + [\beta + \delta_i]x,$$

où $\alpha_2 = \delta_2 = 0$. Dans ce modèle, on suppose qu'il peut y avoir un effet de l'interaction entre la race et la dose de fongicide : en d'autres termes, le lien entre la probabilité de sporulation et la dose de fongicide prend une forme différente selon que la race du mildiou est sensible ou résistante. Un sous-modèle intéressant, que l'on note M_1 , est le modèle sans interaction obtenu par $\delta_1 = 0$:

$$\text{logit}[\pi_i(x)] = \mu + \alpha_i + \beta x.$$

Le modèle M_1 suppose lui que les prédicteurs pour les races sensible et résistante sont parallèles. Le test de la validité de M_2 par rapport à M_1 est donc un test de l'interaction dose \times race. Le graphique de la figure 1.19 représente l'ajustement de ces deux modèles. Le peu de différences entre les ajustements des deux modèles laisse entrevoir l'issue du test de comparaison de ceux-ci. \square

Comme pour le test de validité par rapport au modèle nul, on compare le modèle complet et son sous-modèle à partir du rapport entre leur vraisemblance, notées respectivement \mathcal{V}_p et \mathcal{V}_q :

$$\text{RV}_{p|q} = \frac{\mathcal{V}_q}{\mathcal{V}_p}.$$

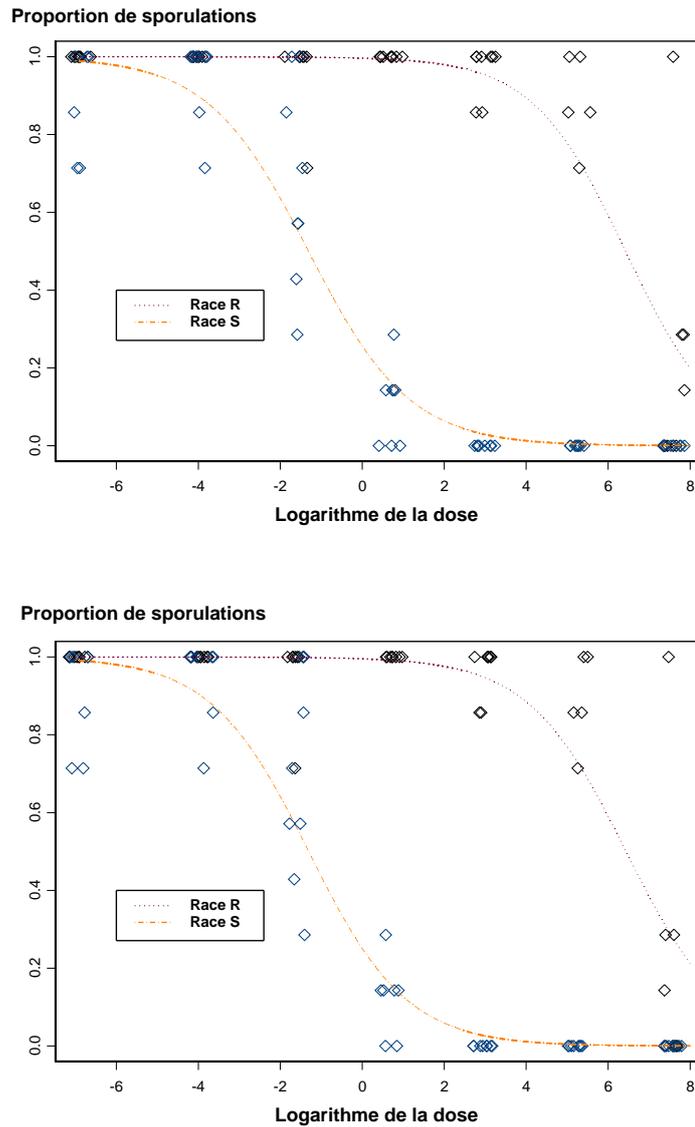


FIG. 1.19: Ajustement du modèle avec interaction (en haut) et du modèle sans interaction (en bas) sur les données « tournesol ».

Sous l'hypothèse nulle, et si la taille de l'échantillon est suffisamment grande, cette statistique est approximativement distribuée selon la loi χ^2_{p-q} , ce qui permet de construire la stratégie de décision et d'évaluer les risques d'erreurs.

L'équation d'analyse de la déviance prend ici la forme suivante :

$$\begin{aligned}\mathcal{V}_0 &= \frac{\mathcal{V}'_0}{\mathcal{V}'_q} \frac{\mathcal{V}'_q}{\mathcal{V}'_p} \mathcal{V}'_p, \\ \underbrace{-2 \ln \mathcal{V}'_0}_{\mathcal{D}_0} &= \underbrace{-2 \ln \text{RV}_q}_{\mathcal{D}_q} \underbrace{-2 \ln \text{RV}_{p|q}}_{\mathcal{D}_{p|q}} \underbrace{-2 \ln \mathcal{V}'_p}_{\mathcal{D}_r}, \\ \mathcal{D}_0 &= \mathcal{D}_q + \mathcal{D}_{p|q} + \mathcal{D}_r,\end{aligned}$$

où \mathcal{D}_q est la déviance du modèle M_q , $\mathcal{D}_{p|q}$ est la déviance mesurant le gain d'ajustement que permet le modèle M_p par rapport au modèle M_q . Notons que, par comparaison avec l'équation d'analyse de la déviance du modèle complet, on retrouve la déviance du modèle M_p par $\mathcal{D}_q + \mathcal{D}_{p|q}$. Remarquons aussi que ce type d'analyse de la déviance donne lieu à des tests séquentiels de validité des modèles : \mathcal{D}_q permet de tester la validité du sous-modèle et $\mathcal{D}_{p|q}$ la validité du modèle complet par rapport au sous-modèle. Dans ce cas, on parle d'analyse de la déviance de **type 1**.

Exemple : résistance du tournesol au mildiou. Le tableau 1.11 donne les éléments pour l'analyse de la déviance de type 1 du modèle complet. A l'évidence, il amène à conclure qu'il n'y a pas d'effet de l'interaction dose \times race. \square

Analyse de la déviance			
	Déviance	Degré de liberté	Probabilité critique
Modèle nul	836.739	107	
Modèle sans interaction	689.21	2	0
Modèle avec interaction	0.21	1	0.64
Résiduelle	147.316	104	

TAB. 1.11: *Analyse de la déviance du modèle avec interaction pour les données « tournesol ».*

1.5.3 Sélection pas à pas du meilleur modèle

La mise en œuvre de cette méthode dans un cas concret fait l'objet du chapitre ???. Cette méthode est pré Une application intéressante des tests de validité d'un modèle par rapport à un sous-modèle est la construction d'une procédure de choix d'un sous-ensemble pertinent de variables explicatives parmi les p variables disponibles

$x^{(1)}, x^{(2)}, \dots, x^{(p)}$. Une manière de considérer cette procédure est la suivante : on commence par comparer tous les modèles construits à partir d'une seule variable explicative sur la base de leur déviance. Parmi les modèles pour lesquels le test de validité conduit à penser qu'ils sont intéressants, on retient alors la variable explicative correspondant au modèle de plus grande déviance, donc s'ajustant le mieux. On opère maintenant de la même manière avec les $p - 1$ modèles à deux variables explicatives, dont la première déjà retenue. On reproduit l'opération jusqu'à ce que l'ajout d'une nouvelle variable ne conduise pas à un gain d'ajustement significatif, au sens du test d'analyse de la déviance.

Exemple : promesse d'achat d'un produit alimentaire. On s'intéresse ici à la mise en évidence de descripteurs sensoriels qui favoriseraient l'intention d'achat. En plus de la promesse d'achat et de la note hédonique, on dispose pour cela d'évaluations sensorielles des sandwiches, selon différents descripteurs, par les juges, sur une échelle de note allant de 0 à 10. La liste des descripteurs sensoriels est fourni dans le tableau 1.12.

On reporte dans le tableau 1.13 les probabilités critiques des tests séquentiels de la validité des modèles construits sur chacun des descripteurs sensoriels. La procédure pas à pas conduit à retenir comme descripteurs sensoriels influençant la promesse d'achat en premier lieu la qualité de la garniture en bouche, ensuite la qualité du pain en bouche et enfin la pertinence de l'association des ingrédients. \square

Variable	Definition
$x^{(1)}$	Impression visuelle
$x^{(2)}$	Appétance
$x^{(3)}$	Qualité du pain en bouche
$x^{(4)}$	Originalité du pain
$x^{(5)}$	Quantité de garniture
$x^{(6)}$	Qualité de la garniture en bouche
$x^{(7)}$	Originalité de la garniture
$x^{(8)}$	Pertinence de l'association des ingrédients de la garniture

TAB. 1.12: *Descripteurs sensoriels des sandwiches.*

Variable	Modèle à une variable explicative	Modèle à 2 variables explicatives dont $x^{(6)}$	Modèle à 3 variables explicatives dont $x^{(6)}$ et $x^{(3)}$	Modèle à 4 variables explicatives dont $x^{(6)}$, $x^{(3)}$ et $x^{(8)}$
$x^{(1)}$	1.7e-2	<i>1.4e-1</i>	<i>4.2e-1</i>	<i>5.6e-1</i>
$x^{(2)}$	1.2e-4	4.0e-2	<i>2.8e-1</i>	<i>3.7e-1</i>
$x^{(3)}$	4.9e-6	1.0e-3		
$x^{(4)}$	<i>7.8e-1</i>	<i>9.6e-1</i>	<i>2.7e-1</i>	<i>2.0e-1</i>
$x^{(5)}$	<i>6.1e-1</i>	<i>8.8e-1</i>	<i>8.6e-1</i>	<i>8.2e-1</i>
$x^{(6)}$	5.4e-10			
$x^{(7)}$	7.7e-3	4.6e-1	<i>3.3e-1</i>	<i>6.9e-1</i>
$x^{(8)}$	1.8e-7	1.1e-2	1.7e-1	

TAB. 1.13: Sélection d'un sous-ensemble pertinent de descripteurs sensoriels. En italique, les probabilités critiques conduisant à considérer que le modèle n'est pas intéressant. En gras, la probabilité critique du modèle sélectionné.

1.5.4 Contributions individuelles à la déviance résiduelle

Cette étape de l'analyse des données par le modèle linéaire généralisé peut être vue comme le pendant du traditionnel examen des résidus de l'ajustement lorsque l'on utilise le modèle linéaire classique. Elle apporte souvent un éclairage nouveau sur la qualité de l'ajustement, en particulier lorsque le test de validité amène à remettre en cause la pertinence du modèle. D'autre part, elle permet aussi un diagnostic individuel de l'adéquation d'une donnée à un modèle. Dans le cas du modèle linéaire classique, le calcul des résidus par l'écart entre les valeurs observées de Y et leur valeur ajustée par le modèle fait l'objet d'un consensus. En revanche, dans le cadre du modèle linéaire généralisé, on trouve plusieurs définitions des résidus d'ajustement. On se limite ici à la présentation d'un de ces types de résidus dont la définition est cohérente avec les arguments avancés pour le choix de la procédure d'estimation et de la stratégie du test de la validité du modèle.

Dans la suite, on illustre le calcul de ces résidus dans le cadre du modèle logit. Dans un premier temps, exami-

nous plus en détail l'expression de la déviance résiduelle d'un modèle de vraisemblance \mathcal{V} :

$$\begin{aligned} \mathcal{D}_r &= -2 \ln \mathcal{V}, \\ &= \sum_{i=1}^n \underbrace{-2 \ln [\hat{\mathbb{P}}_{x_i}(Y_i = y_i)]}_{d_i^2}, \end{aligned}$$

où $\hat{\mathbb{P}}_{x_i}(Y_i = y_i)$ est la probabilité, calculée par le modèle ajusté, que l'on observe la valeur y_i de la variable à expliquer lorsque la variable explicative prend la valeur x_i . Par conséquent, cette probabilité mesure le degré d'adéquation, pour la i ème donnée de l'échantillon, entre la valeur observée y_i de la variable à expliquer et sa valeur ajustée par le modèle. De plus, si on note $\tilde{\epsilon}_i = d_i$ lorsque y_i est plus petit que sa valeur ajustée et $\tilde{\epsilon}_i = -d_i$ lorsque y_i est plus grand que sa valeur ajustée, alors $\tilde{\epsilon}_i$ peut être vu comme une mesure de la contribution de la i ème donnée à la déviance résiduelle. Les valeurs $\tilde{\epsilon}_i$, $i = 1, \dots, n$, sont appelées **résidus de la déviance** et sont distribuées selon une loi normale centrée réduite, de sorte que la stratégie de détection de données mal ajustées est similaire à celle utilisée dans le cadre du modèle linéaire classique.

Exemple : résistance du tournesol au mildiou. Les graphiques de la figure 1.20 illustrent ici le calcul des résidus de la déviance.

1.6 Modèle logit et discrimination

Outre les contributions importantes à l'étude de l'influence de variables explicatives sur une variable binaire que permet le modèle logit, son utilisation s'inscrit souvent dans une problématique de discrimination. Il s'agit alors de construire des règles de décision basées sur la connaissance des variables explicatives et visant à affecter une donnée à la classe définie par $Y = 0$ ou à celle définie par $Y = 1$. Dans le contexte du modèle de régression logistique, la prédiction, telle qu'elle a été définie par exemple dans le cas du modèle de régression linéaire, se ramène à une affectation à l'une ou l'autre des classes définies par les valeurs de Y .

Supposons donc que nous ne connaissons d'un individu statistique que la valeur x_0 de la variable explicative, que donc la valeur Y_0 de Y ne soit pas observée, et que nous souhaitons, à partir de cette information, affecter cet individu soit à la classe définie par $Y_0 = 0$, soit à celle définie par $Y_0 = 1$. Le modèle logit nous permet dans un premier temps d'estimer la probabilité $\pi_0 = \mathbb{P}_{x_0}(Y_0 = 1)$:

$$g(\hat{\pi}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Le principe de la stratégie de discrimination est alors de considérer que si $\hat{\pi}_0$ dépasse un seuil s , alors la valeur

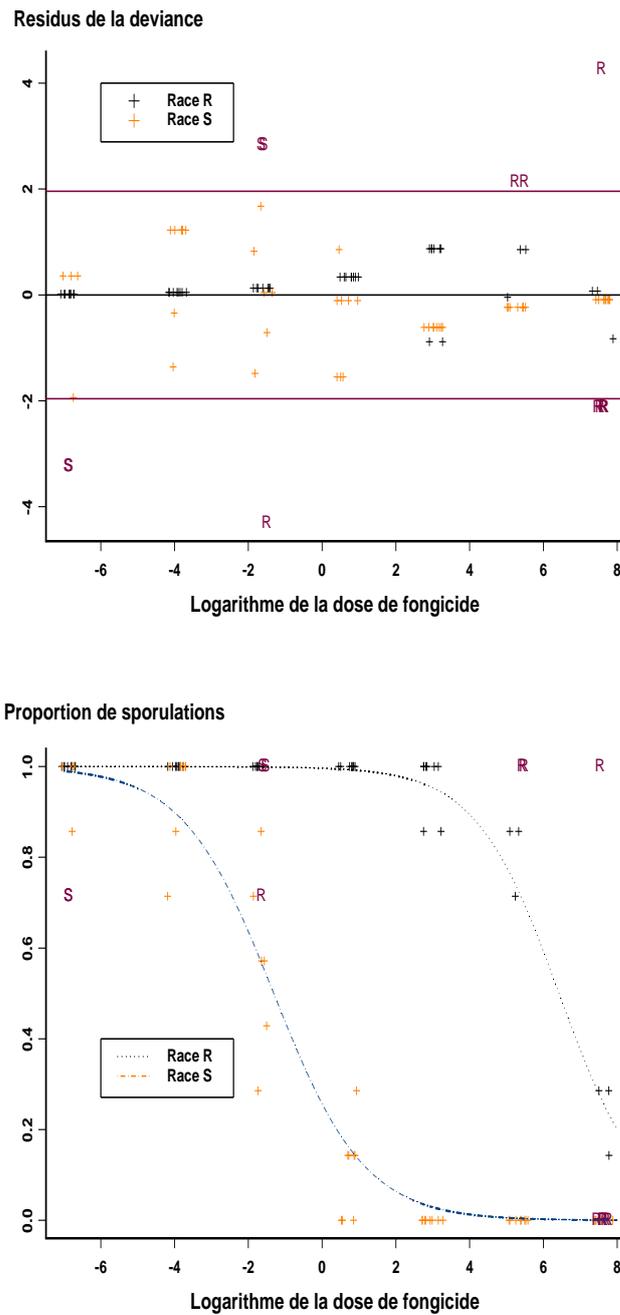


FIG. 1.20: Détection de données mal ajustées (les points désignés par les lettres « R » ou « S » selon la race du mildiou) par l'analyse des résidus de la déviance.

prédite \hat{Y}_0 de Y_0 est 1, sinon $\hat{Y}_0 = 0$. En d'autres termes :

$$\begin{cases} \text{si } \hat{\pi}_0 > s \text{ alors } \hat{Y}_0 = 1 \\ \text{si } \hat{\pi}_0 \leq s \text{ alors } \hat{Y}_0 = 0 \end{cases}$$

Exemple : promesse d'achat d'un produit alimentaire. Le graphique de la figure 1.21 illustre la stratégie logit de discrimination. Pour cet exemple, la valeur du seuil s est arbitrairement fixée à 0.3. Une manière simple de mesurer la pertinence de ce choix est de comptabiliser les erreurs d'affectations observées sur l'échantillon. Ces erreurs sont de deux types : soit une donnée a été affectée à la classe $Y = 0$ alors que la valeur observée de Y est 1, soit une donnée a été affectée à la classe $Y = 1$ alors que la valeur observée de Y est 0. Ces deux types d'erreurs sont recensées sur le graphique. En l'occurrence, pour ce choix particulier de seuil, le taux d'affectations correctes s'élève à 79.02% : parmi les personnes n'ayant pas l'intention d'acheter le produit, 55.36% ont été bien classées et parmi les personnes ayant l'intention d'acheter le produit, 94.25% ont été bien classées.

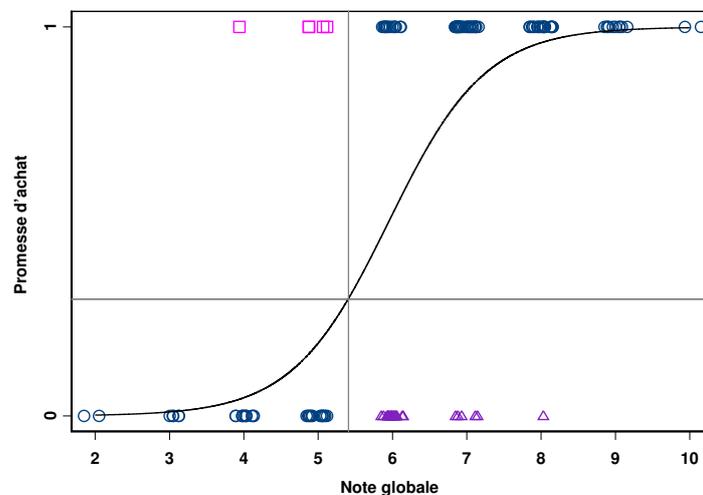


FIG. 1.21: *Discrimination logit. Exemple des données d'intention d'achat. Les cercles désignent les données correctement affectées.*

Ces taux d'affectations correctes servent de support à la détermination d'une valeur pertinente du seuil s . On choisit par exemple la valeur du seuil pour laquelle le taux d'affectations correctes global est le plus fort. Le graphique de la figure 1.22, appelé graphique de capacité, représente l'évolution des différents taux d'affectations correctes en fonction de s et permet donc un choix rapide de la meilleure valeur du seuil. □

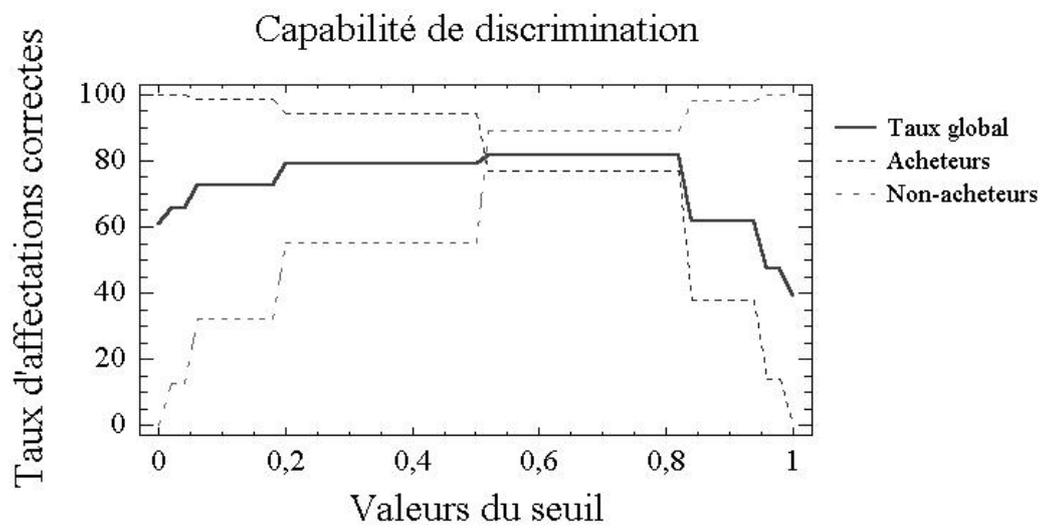


FIG. 1.22: Choix de la meilleure valeur du seuil

