Analyse discriminante

Plan du cours

- Généralités
- Analyse factorielle discriminante
- Analyse discriminante décisionnelle
- Régression logistique binaire

Bibliographie:

- Lebart, L., Morineau A. et Piron, M. Statistique exploratoire multidimensionnelle. 3eme édition, DUNOD 2000.
- Celeux, G. et Nakache, J.P. : Analyse discriminante sur variables qualitatives. POLYTECHNICA 1994.
- Bardos, M. Analyse discriminante : Application au risque et au scoring financier. DUNOD 2001.
- Saporta, G. Probabilités, analyse des données et statistique. TECHNIP 2006.
- Tenenhauss, M. Statistique : Méthodes pour décrire, expliquer et prévoir. DUNOD 2007.
- Volle, M. Analyse des données. ECONOMICA.
- Gomis, F.K. : Contribution sur l'analyse discriminante. Mémoire de master 2

1 Généralités

L'analyse discriminante est une technique visant à séparer au mieux des groupes prédéfinis et à affecter de nouveaux individus à ces groupes. Pour ce faire, elle utilise deux approches :
- L'approche *descriptive ou explicative* a pour but de mettre en évidence les différences entre des groupes prédéfinis et de visualiser ces groupes. Cette approche, dite encore *analyse factorielle discriminante* , cherche à identifier les variables ou combinaisons linéaires de variables les plus pertinentes pour séparer les groupes ; puis à affecter, grâce à des règles géométriques, chaque individu au groupe dont le centre de gravité est le plus proche de cet individu.
- L'approche *prédictive ou décisionnelle*, quant à elle, permet de déterminer le groupe d'appartenance d'un nouvel individu, connaissant ses valeurs sur les prédicteurs (variables explicatives). Cette approche, appelée aussi *analyse discriminante bayésienne*, utilise des méthodes probabilistes pour construire une règle d'affectation des individus à l'un des groupes déjà définis. Cette règle doit être utilisée dans le futur avec le minimum d'erreur possible.

L'analyse discriminante peut aussi être considérée comme une extension du problème de la régression au cas où la variable dépendante est qualitative. Ainsi, dans le cas de deux groupes elle peut être assimilée à la régression linéaire multiple.

L'analyse discriminante se rattache au champ plus vaste de la reconnaissance des formes. Par ses objectifs, elle s'apparente également aux réseaux neuronaux, sujet très à la mode en recherche informatique. Selon le domaine d'application, elle a un but décisionnel ou descriptif. Par exemple, dans le traitement automatique du courrier, la reconnaissance des codes postaux a un but décisionnel. Tandis que pour un référendum où il y a deux groupes de votants "oui" et "non", on cherche à caractériser chaque groupe ; le but est donc descriptif ici.

L'analyse discriminante s'applique sur des données consistant en n observations réparties en k classes (ou groupes déjà définies) et décrites par p variables explicatives. Elle est utilisée dans beaucoup de domaines :

- Médecine : aide au pronostic et au diagnostic
- Banque : crédit-scoring
- Assurance : prévision de risques
- Marketing: sélection de clients potentiels
- Militaire : détection de cible
- etc.

2 Analyse factorielle discriminante (AFD)

Considérons une population de n individus partitionnée en k classes (ou groupes) à l'aide d'une variable qualitative Y. Chaque individu est décrit par p variables numériques $X_1, \cdots, X_p$ appelées descripteurs.
L'analyse factorielle discriminante consiste à rechercher des combinaisons linéaires des p variables explicatives séparant au mieux les k classes au sens de la dispersion, i.e. des combinaisons linéaires dont la variance provient plus des différences entre classes que des différences des individus à l'intérieur d'une même classe.
Chaque groupe $h = 1, \cdots, k$ définit une sous-population composée de n_h individus de sorte que $n = \sum_{h=1}^{k} n_h$.
Supposons que :
- x_{ijh} = valeur de la variable X_j observée sur le i^{eme} individu du groupe h
- $\bar{x}_{jh}$ = moyenne de la variable X_j dans le groupe h; $\quad \bar{x}_{jh} = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{ijh}$
- $\bar{x}_j$ = moyenne de la variable X_j dans la population globale , $\bar{x}_j = \frac{1}{n} \sum_{h=1}^{k} \sum_{i=1}^{n_h} x_{ijh}$.

Alors l'équation d'analyse de la variance pour chaque variable X_j s'écrit :

$$\sum_{h=1}^{k} \sum_{i=1}^{n_h} (x_{ijh} - \bar{x}_j)^2 = \sum_{h=1}^{k} \sum_{i=1}^{n_h} (x_{ijh} - \bar{x}_{j,h})^2 + \sum_{h=1}^{k} n_h (\bar{x}_{jh} - \bar{x}_j)^2.$$

Pour le vecteur $X = (X_1, \cdots, X_p)$, la décomposition de la variance s'écrit avec des matrices. En effet, notons :
$x_{ih} = (x_{i1h}, \cdots, x_{iph})'$ le vecteur formé par les valeurs prises par l'individu i du groupe h sur l'ensemble des p variables ;
$\bar{x}_h = (\bar{x}_{1h}, \cdots, \bar{x}_{ph})'$ le vecteur formé par les moyennes des p variables dans le groupe h ;
$\bar{x} = (\bar{x}_1, \cdots, \bar{x}_p)'$ le vecteur formé par les moyennes des variables dans la population globale.
Alors on a

$$\underbrace{\sum_{h=1}^{k} \sum_{i=1}^{n_h} (x_{ih} - \bar{x})(x_{ih} - \bar{x})'}_{T} = \underbrace{\sum_{h=1}^{k} \sum_{i=1}^{n_h} (x_{ih} - \bar{x}_h)(x_{ih} - \bar{x}_h)'}_{W} + \underbrace{\sum_{h=1}^{k} n_h (\bar{x}_h - \bar{x})(\bar{x}_h - \bar{x})'}_{B},$$

où T est la matrice de variance totale, W la matrice de variance intra-classe et B la matrice de variance inter-classe.
L'analyse factorielle discriminante cherche une première variable Z_1, comme étant une combinaison linéaire des variables X_j **centrées** et tel que son rapport de corrélation avec la variable qualitative (de groupe) Y soit maximum, i.e.

$$Z_1 = \sum_{j=1}^{p} a_{1j} X_j, \; a_1 = (a_{11}, \cdots, a_{1p})' \in \mathbb{R}^p \text{ et } \eta^2(Z_1, Y) \text{ maximal},$$

puis une deuxième variable Z_2 non corrélée à Z_1 telle que $\eta^2(Z_2, Y)$ maximum, ainsi de suite...
Les variables $Z_1, \cdots, Z_m,\ m \geq 1$ sont appelées variables discriminantes et on a : $m \leq \min(k-1, p)$.

Détermination de la première variable discriminante Z_1

La variable $Z_1 = \sum_{j=1}^{p} a_{1j} X_j = Xa_1$ prend la valeur $z_{1ih} = a_1' x_{ih}$ pour l'individu i du groupe h.
Sa moyenne dans le groupe h est $\bar{z}_{1h} = a_1' \bar{x}_h$.
Sa moyenne dans l'échantillon global est $\bar{z}_1 = a_1' \bar{x}$.
Alors la décomposition de la variance de Z_1 donne :

$$\begin{aligned}
\sum_{h=1}^{k}\sum_{i=1}^{n_h}(z_{1ih} - \bar{z}_1)^2 &= \sum_{h=1}^{k}\sum_{i=1}^{n_h}(z_{1ih} - \bar{z}_{1h})^2 + \sum_{h=1}^{k} n_h(\bar{z}_{1h} - \bar{z}_1)^2 \\
\underbrace{\sum_{h=1}^{k}\sum_{i=1}^{n_h} a_1'(x_{ih} - \bar{x})(x_{ih} - \bar{x})' a_1}_{a_1' T a_1} &= \underbrace{\sum_{h=1}^{k}\sum_{i=1}^{n_h} a_1'(x_{ih} - \bar{x_h})(x_{ih} - \bar{x_h})' a_1}_{a_1' W a_1} + \underbrace{\sum_{h=1}^{k} n_h a_1'(\bar{x_h} - \bar{x})(\bar{x_h} - \bar{x})' a_1}_{a_1' B a_1.}
\end{aligned}$$

On définit le **pouvoir discriminant** de la variable Z_1 comme le carré du rapport de corrélation de la variable Z_1 avec la variable de groupe Y :

$$\eta^2(Z_1, Y) = \frac{a_1' B a_1}{a_1' T a_1}.$$

$\eta^2(Z_1, Y)$ est maximal si le vecteur a_1 maximise la fonction $f(u) = \frac{u'Bu}{u'Tu},\ u \in \mathbb{R}^p$.

$$\frac{\partial}{\partial u} f(u) = \frac{2Bu(u'Tu) - 2Tu(u'Bu)}{(u'Tu)^2}.$$

$\frac{\partial}{\partial u} f(u) = 0 \Rightarrow 2Bu(u'Tu) - 2Tu(u'Bu) = 0 \Rightarrow Bu = (\frac{u'Bu}{u'Tu})Tu.$
Donc si a_1 maximise $f(u)$ alors

$$Ba_1 = \left(\frac{a_1' B a_1}{a_1' T a_1}\right) T a_1 \Leftrightarrow Ba_1 =: \lambda_1 T a_1 \Leftrightarrow T^{-1} B a_1 = \lambda_1 a_1.$$

Donc a_1 est un vecteur propre de la matrice $T^{-1}B$ associée à la plus grande valeur propre $\lambda_1 = \frac{a_1' B a_1}{a_1' T a_1} =: \eta_1^2$.
$\sqrt{\lambda_1} = \eta_1$ est la première corrélation canonique.
De manière successive, on obtient les autres variables discriminantes $Z_m = Xa_m$, où a_m est un vecteur propre de $T^{-1}B$ associée à la valeur propre $\lambda_m = \eta^2(Z_m, Y) =: \eta_m^2$. Les valeurs propres sont rangées par ordre décroissant : $\lambda_1 \geq \cdots \geq \lambda_m$.

Choix du nombre de variables discriminantes à retenir

On supposera que le vecteur $X = (X_1, \cdots, X_p)'$ des variables prédictrices suit une loi multinormale dans chaque groupe. Ainsi on pourra tester l'hypothèse de nullité des q derniers rapports de corrélation, i.e. l'hypothèse

$$H_0 : \eta_{k-q}^2 = \eta_{k-q-1}^2 = \cdots = \eta_{k-1}^2 = 0,$$

k étant le nombre de groupes. On utilise la statistique de test

$$\Lambda_q = \prod_{m=k-q}^{k-1} (1 - \eta_m^2).$$

On rejette H_0 si Λ_q est très petit.
Pour mesurer le **pouvoir discriminant global** des p variables, on utilise la statistique suivante appelée **lambda de Wilks**

$$\Lambda = \Lambda_{k-1} = \prod_{m=1}^{k-1} (1 - \eta_m^2).$$

Plus Λ est petit, plus les variables sont globalement discriminantes. Λ peut aussi être utilisée pour tester l'égalité entre les moyennes des différents groupes.

Règle d'affectation géométrique : (Mahalanobis-Fisher)

Définition 1 *Soit M une matrice symétrique définie positive, on définit la distance entre deux vecteurs x, y par rapport à M par :*

$$d^2(x, y) = (x - y)'M(x - y).$$

On dit que M est la métrique associée à cette distance d.
La distance de Mahalanobis est définie par la métrique W^{-1} (inverse de la matrice intraclasse).

On cherche à affecter un nouvel individu à l'un des k groupes déjà séparés par les axes discriminants en utilisant la distance de Mahalanobis définie par la matrice inverse W^{-1}. Soit $g_1, \cdots, g_k$ les centres de gravité respectifs des k groupes. Pour classer un individu e, la règle d'affectation géométrique consiste à calculer la distance entre e et chaque centre de gravité g_h et à affecter l'individu e selon la distance la plus proche.
La distance d'un individu e par rapport au centre de gravité g_h du groupe h est donnée par

$$d^2(e, g_h) = (e - g_h)'W^{-1}(e - g_h) = e'W^{-1}e + g_h'W^{-1}g_h - 2e'W^{-1}g_h.$$

L'individu e sera alors affecté au groupe h_0 si $g_{h_0}'W^{-1}g_{h_0} - 2e'W^{-1}g_{h_0}$ est minimum.
Remarque. L'utilisation de cette règle d'affectation peut conduire à des erreurs lorsque les dispersions intra-groupes sont très différentes. Dans ce cas, on pourra utiliser la matrice intra-classe de chaque groupe W_h au lieu de la matrice intra-classe globale W pour améliorer la règle d'affectation. On peut aussi apprécier la qualité de cette règle d'affectation en calculant, pour modalité h de la variable Y, le pourcentage de "bien classés", i.e. le pourcentage d'individus de la classe h qui restent dans la classe h, lorsqu'on applique cette règle.

3 Analyse discriminante décisionnelle (ADD)

On considère une population divisée en k groupes selon un critère (ou variable) qualitatif Y.
$X_1, \cdots, X_p$ sont les descripteurs ou variables prédictrices.
Soit $q_h = P(Y = h)$: la probabilité *a priori* d'appartenance au groupe h, $h = 1, \cdots, k$.
Etant donnée une nouvelle observation $x = (x_1, \cdots, x_p)$ du vecteur $X = (X_1, \cdots, X_p)$, l'analyse discriminante décisionnelle cherche à calculer la probabilité d'appartenance de cette observation aux différents groupes déjà définis ; puis à affecter l'observation au groupe le plus probable. Cette probabilité conditionnelle *a posteriori* est donnée par :

$$p_h(x) = P(Y = h/X = x), \quad \forall h = 1, \cdots, k.$$

Si la distribution de probabilité du vecteur $X = (X_1, \cdots, X_p)$ admet une densité $f_h(x)$ dans chaque groupe h, alors on peut utiliser la formule de Bayes pour calculer les probabilités *a priori* $p_h(x)$ des groupes, connaissant x. On a pour tout $h = 1, \cdots, k$

$$p_h(x) = \frac{P(Y = h)P(X = x/Y = h)}{\sum_{h=1}^{k} P(Y = h)P(X = x/Y = h)} = \frac{q_h f_h(x)}{\sum_{h=1}^{k} q_h f_h(x)}.$$

Hypothèse fondamentale
Dans chaque groupe h la distribution du vecteur X suit une loi multinormale de moyenne μ_h et de matrice de covariance Σ_h. Sa densité de probabilité est alors donnée par :

$$f_h(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_h|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_h)'\Sigma_h^{-1}(x - \mu_h)\right\}, \quad |\Sigma_h| = det(\Sigma_h).$$

Pour réaliser l'analyse discriminante bayésienne, on distinguera deux cas de figure :
I. Cas de matrices de covariance égales:

Pour chaque groupe h, $X = (X_1, \cdots, X_p) \sim N(\mu_h, \Sigma)$, $\Sigma = \Sigma_1 = \cdots = \Sigma_k$. Alors la probabilité *a posteriori* est défini par :

$$p_h(x) = \frac{\exp\left\{-\frac{1}{2}(x-\mu_h)'\Sigma^{-1}(x-\mu_h)\right\} q_h}{\sum_{h=1}^{k} \exp\left\{-\frac{1}{2}(x-\mu_h)'\Sigma^{-1}(x-\mu_h) q_h\right\}}$$

qui s'écrit

$$p_h(x) = \frac{exp\left\{-\frac{1}{2}\mu_h'\Sigma^{-1}\mu_h + \mu_h\Sigma^{-1}x + ln(q_h)\right\}}{\sum_{h=1}^{k} exp\left\{-\frac{1}{2}\mu_h'\Sigma^{-1}\mu_h + \mu_h\Sigma^{-1}x + ln(q_h)\right\}}.$$

Pour déterminer le groupe d'appartenance d'un individu sur lequel on a observé le vecteur $x = (x_1, \cdots, x_p)$, on définit les fonctions discriminantes linéaires en x,

$$g_h(x) = -\frac{1}{2}\bar{x}_h' S^{-1}\bar{x}_h + \bar{x}_h' S^{-1}x + ln(q_h), \quad h = 1, \cdots, k$$

où $\bar{x}_h$ est l'estimation de la moyenne μ_h du groupe h et $S = \frac{1}{n-k}W$ est l'estimation de Σ, avec W la matrice de variance intraclasse, n la taille de l'échantillon. La probabilité a posteriori $p_h(x)$ est alors estimé par :

$$\hat{p}_h(x) = \frac{\exp[g_h(x)]}{\sum_{l=1}^{k} \exp[g_l(x)]}.$$

Ainsi, l'individu x sera affecté au groupe h si $g_h(x)$ est maximum.

Remarque: Si les probabilités a priori q_h sont égales, alors la règle d'affectation bayesienne est équivalente à la règle géométrique. En effet, si $q_1 = \cdots = q_k$ alors la fonction discriminante linéaire $g_h(x)$ se réduit à :

$$g_h(x) = -\frac{1}{2}\bar{x}_h' S^{-1}\bar{x}_h + \bar{x}_h' S^{-1}x.$$

On peut montrer que maximiser $g_h(x)$ en fonction du groupe h, revient à minimiser la fonction $u \mapsto u'W^{-1}u - 2x'W^{-1}u$ dans l'ensemble des centres de gravité $\{g_h\}$ des differents groupes ; ce qui correspond à la règle d'affectation géométrique.

II. Cas de matrices de covariance inégales:

Ici on suppose que dans chaque groupe h le vecteur $X = (X_1, \cdots, X_p) \sim N(\mu_h, \Sigma_h)$, les Σ_h sont différentes. La probabilité a posteriori s'écrit:

$$p_h(x) = \frac{\exp\left\{-\frac{1}{2}(x-\mu_h)'\Sigma_h^{-1}(x-\mu_h) + ln(q_h) - \frac{1}{2}ln|\Sigma_h|\right\}}{\sum_{h=1}^{k} \exp\left\{-\frac{1}{2}(x-\mu_h)'\Sigma_h^{-1}(x-\mu_h) + ln(q_h) - \frac{1}{2}ln|\Sigma_h|\right\}}.$$

En remplaçant μ_h par $\bar{x}_h$ et la matrice de covariance Σ_h par $S_h = \frac{1}{n_h - 1}\sum_{h=1}^{n_h}(x_{ih} - \bar{x}_h)(x_{ih} - \bar{x}_h)'$, on peut définir les fonctions discriminantes quadratiques en x par :

$$g_{qh}(x) = -\frac{1}{2}(x-\bar{x}_h)'S_h^{-1}(x-\bar{x}_h) + ln(q_h) - \frac{1}{2}ln|S_h|, \quad h = 1, \cdots, k.$$

D'où une estimation de la probabilité a posteriori pour le groupe h

$$\hat{p}_h(x) = \frac{\exp[g_{qh}(x)]}{\sum_{l=1}^{k} \exp[g_{ql}(x)]}.$$

Un individu x sera affecté au groupe h si $g_{qh}(x)$ est maximum.

4 Régression logistique binaire

La régression logistique binaire est une méthode statistique qui permet d'étudier la liaison entre une variable binaire Y et un ensemble de variables explicatives quantitatives et/ou qualitatives, $X = (X_1, \cdots, X_p)$, $p \geq 1$. Il s'agit généralement d'expliquer et de prévoir la survenue ou non d'un évènement à partir de l'observation

des variables explicatives. La variable dépendante Y prend deux modalités souvent notées "0" et "1", i.e Y est une variable dichotomique.

$$Y = \begin{cases} 1 & \text{si l'évènement est survenu} \\ 0 & \text{sinon.} \end{cases}$$

Pour un individu sur lequel on a observé les valeurs $x = (x_1, \cdots, x_p)$, la probabilité de survenue de l'évènement est définie par une probabilité conditionnelle

$$P(Y = 1/X = x) = P(Y = 1/x) = \pi(x).$$

Dans le modèle de régression logistique, cette probabilité $\pi(x)$ est donnée par :

$$\pi(x) = \frac{\exp(S(x))}{1 + \exp(S(x))}, \qquad \text{Hypothèse fondamentale} \tag{1}$$

où $S(x) = \sum_{j=1}^{p} \beta_j x_j + \beta_0$ est une combinaison linéaire des x_j, appelée fonction score. Ce modèle peut aussi s'écrire sous la forme :

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = S(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2}$$

Dans le cas, où il n y a qu'une seule variable explicative $x = x_1$, la probabilité $\pi(x)$ s'écrit :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \qquad \text{(RLBS)}. \tag{3}$$

Il s'agit du modèle de régression logistique binaire simple (RLBS).

Odds et Odds-ratio

On appelle *Odds* d'un évènement pour une modalité x fixée le rapport entre la probabilité de cet évènement et la probabilité de son contraire conditionnellement à x, i.e,

$$Odds(x) = \frac{P(Y = 1/x)}{P(Y = 0/x)} = \frac{\pi(x)}{1 - \pi(x)}.$$

L'Odds exprime un rapport de chances. Par exemple, si un individu présente un $Odds$ de 2, alors il y a 2 fois plus de chances que l'évènement survienne en lui qu'il ne survienne pas.

On appelle *Odds-ratio* d'un évènement pour une modalité $x + 1$ par rapport à une modalité x, le rapport noté

$$OR = \frac{Odds(x + 1)}{Odds(x)}. \tag{4}$$

Exemple : Une étude concernant les facteurs de risque de l'hypertension a été réalisée auprès d'un échantillon composé de fumeurs et de non fumeurs. La variable explicative X (facteur) est le fait de fumer,

$$X = \begin{cases} 1 & \text{si l'individu fume} \\ 0 & \text{sinon.} \end{cases}$$

La variable à expliquer Y est le fait d'avoir l'hypertension,

$$Y = \begin{cases} 1 & \text{si l'individu a l'hypertension} \\ 0 & \text{sinon.} \end{cases}$$

Le tableau suivant représente les données recueillies:

Y/X	1	0	$Total$
1	28	13	41
0	271	368	639
$Total$	299	381	680

Le risque relatif de l'hypertension est défini par le rapport

$$RR = \frac{P(Y = 1/X = 1)}{P(Y = 1/X = 0)} = \frac{28/299}{13/381} = 2,74.$$

Cela signifie que la probabilité d'avoir une hypertension est de 2,74 fois plus élevé chez les fumeurs que chez les non-fumeurs.
Le risque relatif n'est utilisé que si l'échantillon est représentatif. On préfère alors l'Odds-ratio qui est plus stable. Dans l'exemple ci-dessus, l'Odds de l'hypertension chez les fumeurs est donné par

$$Odds(Fumeur) = \frac{P(Y = 1/X = 1)}{P(Y = 0/X = 1)} = \frac{28}{271}.$$

L'Odds de l'hypertension chez les non-fumeurs est donné par

$$Odds(Non - Fumeur) = \frac{P(Y = 1/X = 0)}{P(Y = 0/X = 0)} = \frac{13}{368}.$$

L'Odds-ratio de l'hypertension relativement au facteur X (fait de fumer) est donc

$$OR = \frac{Odds(Fumeur)}{Odds(Non - Fumeur)} = \frac{28/271}{13/368} = 2,92.$$

Cela veut dire que le risque d'avoir une hypertension est de 2,92 fois plus éléve chez les fumeurs que chez les non-fumeurs.
Remarques :
- Un Odds-ratio égal à 1 signifie que les variables X et Y sont indépendantes.
- Un Odds-ratio supérieur strictement à 1 signifie que le facteur X a un effet sur la survenue de l'évènement.

4.1 Estimation du modèle

On dispose de données observées sur un échantillon de n individus indépendamment tirés au hasard dans une population. Notons par (x_i, y_i) les données recueillies sur l'individu i, alors la probabilité d'observer la valeur y_i sachant que $X = x_i$ est donnée par

$$P(Y = y_i/X = x_i) = \begin{cases} \pi(x_i) & \text{si } y_i = 1 \\ 1 - \pi(x_i) & \text{si } y_i = 0. \end{cases}$$

Cette probabilité peut être ré-écrit comme suit :

$$P(Y = y_i/X = x_i) = [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}.$$

Estimer le modèle revient à estimer le vecteur de coefficients $\beta = (\beta_0, \cdots, \beta_p)$. de la fonction score $S(x)$. Pour ce faire, on utilise la méthode du maximum de vraisemblance. On appelle vraisemblance d'un échantillon de n données la probabilité d'observer ces n données à partir d'un tirage dans la population étudiée. Cette probabilité est calculée en fonction des paramètres inconnues du modèle étudié ; en vertu de l'indépendance ici, elle s'écrit :

$$\ell(\beta) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}.$$

En pratique, on travaillera avec le logarithme de la vraisemblance défini par :

$$L(\beta) = \log \ell(\beta) = \sum_{i=1}^{n} \left[y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i)) \right].$$

L'estimateur du maximum de vraisemblance (EMV) $\widehat{\beta}$ est le vecteur qui maximise la log-vraisemblance $L(\beta)$. C'est un estimateur convergent, asymptotiquement normal et de variance minimale. On a

$$\frac{\widehat{\beta} - \beta}{\sqrt{I_n^{-1}(\beta)}} \longrightarrow N(0, I_{p+1}),$$

avec $I_n(\beta) = -E[\frac{\partial^2}{\partial \beta^2} \log \ell(\beta)]$ est la quantité d'information de Fisher et I_{p+1} est la matrice identité. Cette dernière propriété nous permet de faire de l'inférence statistique sur les coefficients du modèle (intervalle de confiance, tests, etc.).

4.2 Tests sur les coefficients du modèle

Pour étudier la significativité d'une variable X_j, on peut tester la nullité du coefficient β_j correspondant, i.e l'hypothèse $H_0 : \beta_j = 0$. Il existe trois types de test pour vérifier cette hypothèse.

Test de Wald

On pose

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

La statistique utilisée est $W = \left(\frac{\widehat{\beta}_j}{\sigma_j}\right)^2$, où $\sigma_j^2 = Var(\widehat{\beta}_j)$. Sous H_0, $W \sim \chi_1^2$. On rejette H_0 au risque α de se tromper si $W > \chi_1^2(1-\alpha)$ (fractile d'ordre $1-\alpha$ de la loi du chi-deux à un degré de liberté) ou de manière équivalente si la $p-value$, $p = P(\chi_1^2 > W) \leq \alpha$.

Test du rapport de vraisemblance

On pose

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

Ce test consiste à comparer la vraisemblance du modèle sous l'hypothèse H_0, notée $\ell(\widehat{\beta}^c)$ à la vraisemblance du modèle complet, notée $\ell(\widehat{\beta})$. Il repose sur la statistique

$$LR = -2\log\left[\frac{\ell(\widehat{\beta}^c)}{\ell(\widehat{\beta})}\right],$$

où $\widehat{\beta}^c$ est l'estimateur du maximum de vraisemblance sous l'hypothèse (ou la containte) H_0. Sous H_0, $LR \sim \chi_1^2$. On rejetera H_0 si la $p-value$, $p = P(\chi_1^2 > LR) \leq \alpha$.

Test du score ou du multiplicateur de Lagrange

Définition 2 *On appelle vecteur score U le gradient du logarithme de la vraisemblance, i.e.*

$$U(\beta) = \frac{\partial}{\partial\beta}\log\ell(\beta).$$

Pour tester les hypothèses

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

la statistique du score est définie par

$$S = U'(\widehat{\beta}^c)I_n^{-1}(\widehat{\beta}^c)U(\widehat{\beta}^c),$$

où $I_n^{-1}(\widehat{\beta}^c)$ est la matrice de covariance de $\widehat{\beta}$ sous l'hypothèse (ou contrainte) H_0. Si H_0 est vraie, $S \sim \chi_1^2$. Donc on rejetera H_0 au risque α de se tromper si la $p-value$, $p = P(\chi_1^2 > S) \leq \alpha$.

Remarque.

- Lorsque la taille de l'échantillon est élevée, ces trois tests sont équivalents.
- Pour des tailles d'échantillon réduites, le test de rapport de vraisemblance est préférable aux deux autres.

Intervalle de confiance des coefficients β_j

Soit $\widehat{\beta}_j$ un estimateur convergent de β_j. D'après la normalité asymptotique de l'EMV, on a

$$\widehat{\beta}_j \sim N(\beta_j, \sigma_j^2), \quad \text{avec } \sigma_j^2 = var(\widehat{\beta}_j).$$

D'où l'intervalle de confiance de niveau $1-\alpha$ pour β_j :

$$\left[\widehat{\beta}_j - u_{1-\alpha/2}\widehat{\sigma_j}, \widehat{\beta}_j + u_{1-\alpha/2}\widehat{\sigma_j}\right],$$

où $u_{1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ de la loi normale standard.
Dans le cas du modèle de RLBS défini en (3), on peut établir que l'Odds-ratio $OR = \exp(\beta_1)$. On en déduit les bornes de l'intervalle de confiance pour l'Odds-ratio,

$$\exp\left[\widehat{\beta}_1 - u_{1-\alpha/2}\widehat{\sigma_j}, \widehat{\beta}_1 + u_{1-\alpha/2}\widehat{\sigma_j}\right].$$

Pour une variable explicative continue, on interprétera la variation relative de l'Odds exprimé en pourcentage par :

$$100\left(\frac{Odds(x+1) - Odds(x)}{Odds(x)}\right) = 100(\exp(\beta_1) - 1).$$

4.3 Evaluation de la qualité du modèle

On définit ci-dessous quelques indicateurs pour mesurer la qualité du modèle estimé. Le modèle étudié peut s'écrire comme suit :

$$y_i = \pi_i + \varepsilon_i,$$

où ε_i est un résidu aléatoire tel que : $E(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \pi_i(1-\pi_i)$. La valeur de y_i estimée par le modèle est $\hat{y}_i = \hat{E}(y_i) = \hat{\pi}_i$.
- On appelle **résidu de Pearson** la quantité

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

- On en déduit le **Khi-deux de Pearson** qui est défini par

$$X^2 = \sum_{i=1}^{n} r_i^2.$$

- On définit la **déviance** (ou déviance résiduelle) par

$$D = -2L(\hat{\beta}),$$

où $L(\hat{\beta})$ est le logarithme de la vraisemblance du modèle étudié.
Remarque. Le Khi-deux de Pearson et la déviance sont généralement assez proches. Pour des données individuelles, on les comparera au nombre de degrés de liberté. Si leurs ratios respectifs par rapport au nombre de degrés de liberté sont proches de 1, alors l'ajustement du modèle étudié aux données est globalement satisfaisant.

-**Test de Hosmer et Lemeshow** : Il est utilisé pour évaluer l'adéquation du modèle étudié avec les données si celles-ci sont regroupées en classes. Le plus souvent les données sont rangées par valeurs décroissantes de $\hat{\pi}_i$ et partagées en environ 10 classes à l'aide des déciles, mais en ne séparant pas les *ex aequo*. Soit $C_1, \cdots, C_g$ une partition des données en g classes. Notons :
m_j : l'effectif de la classe C_j
$m_{j1} = \sum_{i \in C_j} y_i$: la fréquence théorique de l'évènement $[Y = 1]$ dans la classe C_j
$m_{j0} = m_j - m_{j1}$: la fréquence théorique de l'évènement $[Y = 0]$ dans la classe C_j
$\hat{m}_{j1} = \sum_{i \in C_j} \hat{\pi}_i$: la fréquence observée de l'évènement $[Y = 1]$ dans la classe C_j
$\hat{m}_{j0} = m_j - \hat{m}_{j1}$: la fréquence observée de l'évènement $[Y = 0]$ dans la classe C_j.
On construit un test du Khi-deux pour comparer les fréquences observées et théoriques :

$$\chi^2 = \sum_{j=1}^{n}\left(\frac{(\hat{m}_{j1} - m_{j1})^2}{m_{j1}} + \frac{(\hat{m}_{j0} - m_{j0})^2}{m_{j0}}\right).$$

Lorsque le modèle étudié est adéquat, la statistique χ^2 suit approximativement une loi du khi-deux à $g-2$ degrés de liberté. Le modèle est rejeté si la $p - value$ est inférieure à 0.05.

- **Pseudo** R^2: Pour évaluer la force de la liaison entre la variable Y et le prédicteur X, on dispose de plusieurs indicateurs appelés Pseudo-R^2 et analogues au coefficient de détermination R^2 en régression linéaire multiple. En régression logistique l'indicateur le plus utilisé est le Pseudo-R^2 de Mac Fadden défini par

$$R^2_{MF} = 1 - \frac{L(Constante, X)}{L(Constante)},$$

où $L(Constante, X)$ est le logarithme de la vraisemblance du modèle complet et $L(Constante)$ celui du modèle avec la constante uniquement.

4.4 Pouvoir discriminant du modèle

On peut utiliser le modèle logistique binaire pour affecter un individu à l'une des classes définies par la variable réponse Y. La règle d'affectation peut être définie en fixant un seuil c de la manière suivante :
- si $\hat{\pi}_i \geq c$, l'individu i est affecté à la classe $[Y = 1]$
- si $\hat{\pi}_i < c$, l'individu i est affecté à la classe $[Y = 0]$.
Considérons par exemple le tableau de classement suivant basé sur un seuil c :

	Malade ou non		
Résultat du test	$[Y = 1]$(Malade)	$[Y = 0]$(Non-malade)	Total
$(Y = 1)$(Positif)	$n_{11}(c)$	$n_{10}(c)$	$n_1(c)$
$(Y = 0)$(Négatif)	$n_{01}(c)$	$n_{00}(c)$	$n_0(c)$
Total	n_1	n_0	n

$n_{11}(c)$ est le nombre d'individus malades avec un test positif ; on les appelle vrais positifs
$n_{00}(c)$ est le nombre d'individus non-malades avec un test négatif ; on les appelle vrais négatifs
$n_{10}(c)$ est le nombre d'individus non-malades avec un test positif ; on les appelle faux positifs
$n_{01}(c)$ est le nombre d'individus malades avec un test négatif ; on les appelle faux négatifs.
On appelle sensibilité la proportion d'individus malades avec un test positif,

$$\text{Sensibilité} = \frac{n_{11}(c)}{n_1}.$$

On appelle spécificité la proportion d'individus non-malades avec un test négatif,

$$\text{Spécificité} = \frac{n_{00}(c)}{n_0}.$$

Ces deux indices nous renseignent sur le pouvoir discriminatoire du modèle logistique lorsqu'on l'utilise pour la prévision. La courbe représentant la sensibilité en fonction de 1- la spécificité est appelée courbe de ROC. Elle permet aussi d'apprécier le pouvoir discriminant du modèle.