-1cm

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU BEKR BELKAID TLEMCEN

FACULTE DES SCIENCES

DEPARTEMENT DE MATHEMATIQUES

MEMOIRE DE MAGISTERE

EN MATHEMATIQUES

THEME

"Regression Logistique : Théorie et Applications "

Présentée par :

Sebbagh Meriem

Soutenue à Tlemcen le 04 juin 2016 devant le jury composé de:

Abbas Ahmed	Président	M.C.A	Université Abou Bekr Belkaid Tlemcen
Mahdjoub Tewfik	Examinateur	M.C.A	Université Abou Bekr Belkaid Tlemcen
Dali-Sahi Majda	Examinateur	M.C.A	Université Abou Bekr Belkaid Tlemcen
Benmansour Djamel	Encadreur	M.C.A	Université Abou Bekr Belkaid Tlemcen

Année universitaire 2014/2015

Résumé

L'intérêt principal de ce mémoire est d'étudier les modèles décrivant les modalités prises par une ou plusieurs variables qualitatives. Consacrées notamment aux modèles dichotomiques simples, modèles probit et logit . Nous commencerons par présenter les principaux modèles , puis dans une seconde section nous nous intéresserons au problèmes de l'estimation des paramètres de ces modèles , notamment par la méthode du maximum de vraisemblance, dans une troisième section , nous étudierons la convergence des estimateurs du maximum de vraisemblance. Ensuite nous aborderons les tests de spécification des modèles ainsi que les différents problèmes d'inférence .

Enfin nous ferons une étude détaillée de cas intitulée : Estimation des risques d'hémopathies liées au diabète de type 2 chez la femme.

Le modèle dichotomique probit et logit admettent pour variable expliquée la probabilité

$pi = p(Yi = 1/xi) = F(\beta xi)$

Ou la fonction F(.) désigne une fonction de répartition. Toutefois on utilise généralement deux types de fonctions de répartition : Fonction de répartition de la loi logistique ou Fonction de répartition de la loi normale centrée et réduite.

A chacune de ces fonctions correspond un nom attribué au modèle obtenu : Modèle logit et modèle probit .

On cherche naturellement à estimer les composantes du vecteur b. La méthode la plus utilisée est la méthode du maximum de vraisemblance, il s'agit de maximiser la fonction de vraisemblance c'est-à-dire de résoudre l'équation $G(\beta) = 0$ *ou* G est le gradient de la log -vraisemblance . La méthode qui est notamment recommandée pour trouver la solution au problème dans un modèle dichotomique uni varié est la méthode d'optimisation de NEWTON RAPHSON (Méthode itérative car l'équation$G(\beta) = 0$, ne peut être résolue simplement) .

On cherche après à établir les propriétés asymptotiques des estimateurs du maximum de vraisemblance. Sous certaines conditions, l'estimateur du maximum de vraisemblance est convergent et suit asymptotiquement une loi normale.

Après avoir construit un modèle de prédiction, Nous évaluons son efficacité de différentes manières : Par la matrice de confusion , test de Hosmer-Lemeshow , courbe de

roc. . . etc.

Nous présenterons aussi les tests d'hypothèses sur les coefficients, puis nous envisagerons les principaux tests de spécification sur les modèles dichotomiques.

Dans la pratique on utilise le modèle logit à cause de sa simplicité par rapport au modèle probit (régression logistique)

La régression logistique permet de traiter le cas où la variable réponse est de type binaire (oui/non, malade/pas malade, etc. .....), et non pas continu comme dans le modèle de régression linéaire . On maintient quand même l'idée d'une relation linéaire entre la variable réponse et les prédicteurs.

Mots clés : Regression logistique, Probit , Logit, Estimateur du maximum de vraisemblance, Odds, Odd ratio

Summary

The main interest of this paper is to study the models describing the procedures taken by one or more categorical variables. Especially devoted to the simple dichotomous models, probit and logit models. We begin by presenting the models and then in a second section we focus on the problems of estimating the parameters of these models, including the method of maximum likelihood, in a third section, we will study the convergence of maximum estimators likelihood. Then we discuss the specification tests of the models and the different inference problems.

Finally we will make a detailed case study entitled: Estimated risk of hematologic-related type 2 diabetes in women.

The dichotomous probit and logit model to admit dependent variable probability

$pi = p(Yi = 1/xi) = F(\beta xi)$

Or the function F (.) Denotes a cumulative distribution function. However usually use two types of distribution functions: the law of logistic distribution function or the normal cumulative distribution function centered and reduced.

Each of these functions is a name given to the resulting model: logit and probit model.

Naturally we try to estimate the components of vector b. The most used method is the method of maximum likelihood, this is to maximize the likelihood function that is to say, to solve the equation $G(\beta) = 0$ where$G$ is the gradient of the log - likelihood. The method that is particularly recommended to find the solution in a univariate model is the dichotomous optimization Newton Raphson method (Iterative method because the equation $G(\beta) = 0$, can not be resolved simply).

We are looking after establishing the asymptotic properties of maximum likelihood estimators. Under certain conditions, the maximum likelihood estimator is consistent and asymptotically follows a normal distribution.

After building a prediction model, we evaluate its effectiveness in different ways: By the confusion matrix, Hosmer-Lemeshow test, roc curve ... etc.

We will also present the testing of hypotheses on the coefficients, then we will consider the key specification tests on dichotomous models.

In practice we use the logit model because of its simplicity compared to the probit model (logistic regression)

Logistic regression to handle the case where the response variable is binary (yes / no ill / not sick, etc .....), not continuous as in the linear regression model. It still maintains the idea of a linear relationship between the response variable and predictors

Keywords: Regression logistics, Probit, Logit, Maximum likelihood estimator, Odds, Odd ratio

ملخص

الاهتمام الرئيسي من هذه الورقة هو دراسة النماذج التي تصف الإجراءات التي اتخذتها واحد أو أكثر الفئوية المتغيرات. كرس خاصة إلى نماذج ثنائية التفرع، الاحتمالية وlogit نماذج بسيطة. نبدأ من خلال تقديم نماذج وبعد ذلك في القسم الثاني ونحن نركز على المشاكل تقدير المعلمات من هذه النماذج، بما في ذلك طريقة أقصى الاحتمالات، في القسم الثالث، سندرس التقارب القصوى المقدرات احتمال. تم نناقش اختبارات تحديد النماذج والمشاكل الاستدلال مختلفة.

وأخيرا سنقوم بعمل دراسة حالة مفصلة بعنوان: مخاطر المقدر للنوع 2 من مرض السكري المرتبطة الدموية لدى النساء.

والاحتمالية بين ثنيين ونموذج logit للاعتراف احتمال المتغير التابع

بي = ص (بي = 1 / الحادي عشر) = F (XIB)

أو وظيفة F (.) يدل على دالة التوزيع التراكمي. ولكن عادة ما تستخدم نوعين من الوظائف توزيع: قانون دالة التوزيع اللوجستية أو دالة التوزيع التراكمي الطبيعي تركز وتخفيضها.

كل من هذه الوظائف هو الاسم الذي يطلق على النموذج الناتج: اللوغارتميين مع نموذج الاحتمالية.

بطبيعة الحال نحن نحاول تقدير مكونات ناقلات ب. الأسلوب الأكثر استخداما هي طريقة أقصى الاحتمالات، وهذا هو تحقيق أقصى قدر من وظيفة احتمال وهذا يعني، من أجل حل المعادلة G (ب) = 0 حيث G هو التدرج من السجل - شيء محتمل. الطريقة التي يوصى بصفة خاصة لإيجاد الحل في نموذج وحيد المتغير هو الأمثل بين ثنيين طريقة نيوتن رافسون (طريقة تكرارية لأن المعادلة G (ب) = 0، لا يمكن أن تحل ببساطة).

ونحن نتطلع بعد تأسيس خصائص مقارب القصوى المقدرات احتمال. في ظل ظروف معينة، احتمال أقصى مقدر ثابت ومقارب يتبع التوزيع الطبيعي.

بعد بناء نموذج للتنبؤ، ونحن تقييم فعاليته بطرق مختلفة: من خلال مصفوفة الارتباك، هوسمر-Lemeshow الاختبار، منحنى ROC ... الخ

وسوف نقدم أيضا اختبار الفرضيات على معاملات، ثم سننظر اختبارات المواصفات الرئيسية على نماذج ثنائية التفرع.

في الممارسة العملية نستخدم نموذج logit بسبب بساطتها بالمقارنة مع النموذج الاحتمالية (الانحدار اللوجستي) الانحدار اللوجستي للتعامل مع الحالة التي يكون فيها متغير الاستجابة ثنائي (نعم / لا سوء / بست مريضة، الخ .....)، وليس مستمر كما هو الحال في نموذج الانحدار الخطي. لا يزال يحافظ على فكرة وجود علاقة خطية بين متغير الاستجابة وتنبؤ.

كلمات البحث: الخدمات اللوجستية الانحدار، الاحتمالية، Logit، والحد الأقصى احتمال مقدر، ونسبة احتمالات الغريب

Table des Matières

Introduction

Présentation générale

Bien souvent les données disponibles dans les études statistiques sont relatives à des variables qualitatives, exemple (la catégories socioprofessionnelle,le type d'études suivies,présence d'une maladie chez un dividu ou pas...).

Or les méthodes d'inférences traditionnelles ne permettent pas de modéliser ou d'étudier des caractères qualitatifs.

Les modèles Logit et Probit sont devenus d'un usage courant en économie et en science médicale ou sociale.

Il s'agit de modèles de regressions où la variable expliquée est qualitative (c'est à dire qu'elle ne peut prendre qu'un nombre fini de modalités).

1- Exemple: On peut vouloir modeliser, par exemple:

a- Le fait d'être ou non propriétaire de son logement $Y_i = 1$ si i est propriétaire, 0 si non).

b- Les comportements électoraux ($Y_i = 1$ si i a voté pour tel président et 0 pour l'autre).

c- Dans le cadre médical si l'individu est malade ($Y_i = 1$, 0 si non)

La liste des exemples est bien entendu très longue.

2- La variable à modeliser peut comprendre:

* deux modalités: on parlera de modèle dichotomique.

* Trois modalités (ou plus) non ordonnées, modèle polytomique non ordonné.

* Trois modalités ordonnées: modèle polytomique ordonné.

3- variables explicatives:

Les variables explicatives peuvent être discrètes ou continues exemple: on s'intéresse à l'explication de la migration d'individus ($Y_i = 1$ si l'individu a changé de région de résí-

dence et$Y_i = 0$ si non) en fonction des variables susceptibles d'influer sur le comportement des individus comme:

X_1 : Le salaire moyen dans la région de résidence (variable continue).

X_2 : L'activité de l'individu si le sujet est actif ($X_1 = 1$ et 0 si non).

Nous n'avons rien dit, pour l'instant, sur la forme mathématique de la relation qui relie la variable à modeliser Y_i (que nous supposerons dichotomique ie $Y_i = 0$ ou 1) et les variables explicatives X_{i_1}, X_{i_1}, ...X_{i_k}.

C'est le problème auquel nous allons nous interesser.

Le but de ce travail est d'étudier des méthodes spécifiques utilisées tenant compte par exemple de l'absence d'ordre naturel entre les modalités que peut prendre le caractère qualitatif.

Le plan de cet exposé est le suivant :

Nous allons nous intéresser au model le plus simple, à savoir le model dichotomique, dans lequel la variable expliquée ne peut prendre que deux modalités.

Puis nous commencerons par présenter les principaux modèles dichotomiques, et en particulier les modèles Logit et Probit. Puis dans une seconde section, nous nous intéresserons au problème de l'estimation des paramètres de ces modèles, notamment par la méthode du maximum de vraisemblance. Dans une troisième partie, nous étudierons la convergence des estimateurs de maximum de vraisemblance. Enfin dans une dernière section nous aborderons les tests de spécification de ces modèles.

Chapitre 1

Modèles Dichotomiques univariés :

On entend par modèle dichotomique un model statistique dans lequel la variable expliquée Y ne peut prendre que deux modalités (variable dichotomique). Il s'agit alors généralement d'expliquer la survenue ou la non survenue d'un événement.(Hurlin 7).

Notation : On considère un échantillon de N individus indicés $i = 1, \ldots N$. Pour chaque individu, on observe si un certain événement s'est réalisé et l'on note la variable codée associée à l'évènement. On pose $\forall i \in (1, N)$

$$Yi = \begin{cases} 1 & si\ l'evenement\ s'est\ réalisé\ pour\ l'individu\ i \\ 0 & si\ l'evenement\ ne\ s'est\ pas\ réalisé \quad pour\ l'individui \end{cases}$$

On remarque le choix du codage (0,1) qui est traditionnellement retenu pour les modèles dichotomiques. En effet celui-ci permet de définir la probabilité de survenue de l'événement comme l'espérance de la variable codée puisque :

$E(Y_i) = P(Y_i = 1) \times 1 + P(Y_i = 0) \times 0 = P(Y_i = 1) = pi$

L'objectif des modèles dichotomiques consiste alors à expliquer la survenue de l'évènement en fonction d'un certain nombre de caractéristiques observées pour les individus de l'échantillon.

1.1 Spécification linéaire des variables endogènes dichotomiques :

Supposons qu'on observe N observations $Y_i, \forall i = 1, ...N$, d'une variable endogène dichotomique , et parallèlement on observe K variables exogènes $X_i = (X_{i1}, ...X_{iK})$,et soit $\beta = (\beta_{1,}..., \beta_K)'$le vecteur des paramètres à estimer , $\forall i = 1, ...N$.(Hurlin 7).

Dans ce cas le modèle linéaire simple s'écrit

$$Y_i = \beta X_i \quad + \epsilon_i \quad \forall i = 1, ...N. \tag{1.1}$$

L'utilisation de méthodes d'estimation particulières s'avère indispensable pour ce type de modèles. En effet,pour mieux comprendre, appliquons une modélisation linéaire simple au cas d'une variable endogène dichotomique :

On pose

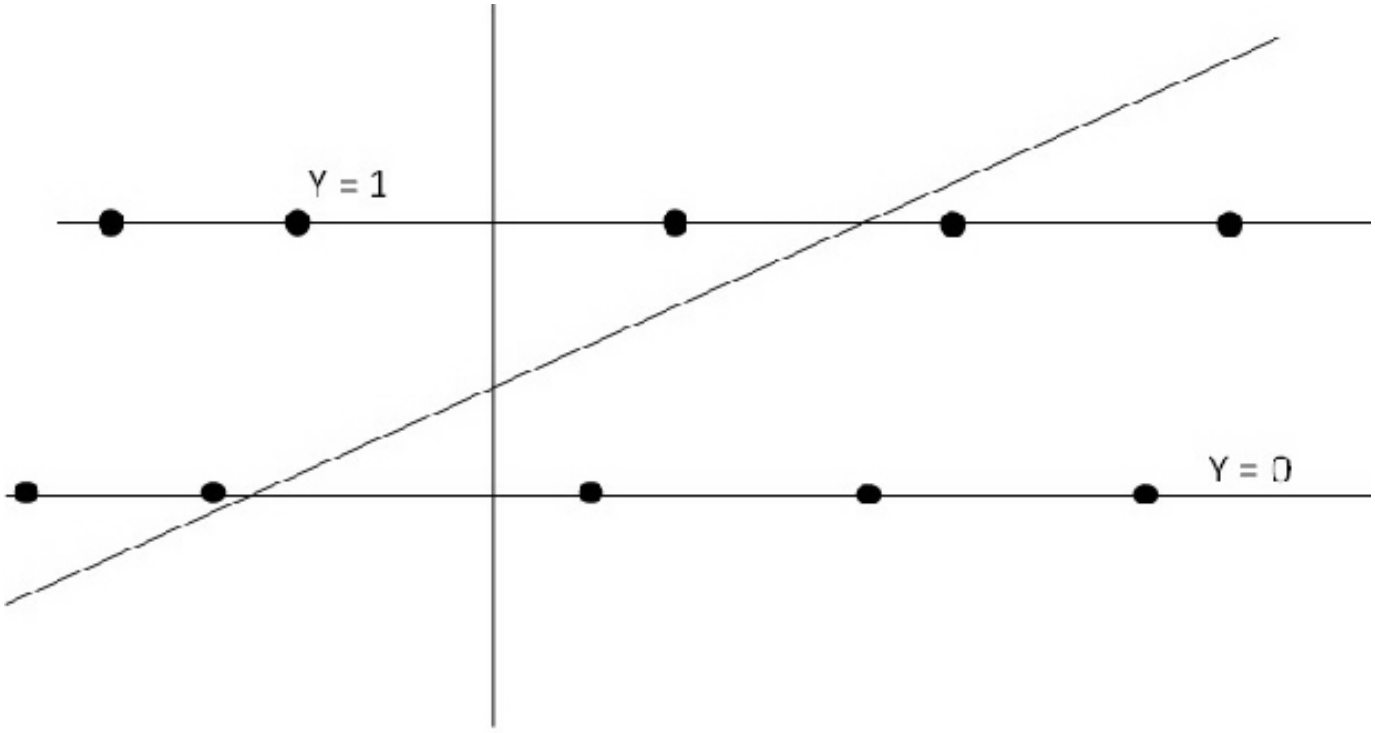
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \forall i = 1, ...N$$

$\beta_0, \beta_1 \in \mathbb{R}, X_i$ variable explicative

Premièrement une étude graphique montre que l'approximation linéaire n'est pas adaptée au problème posé, il suffit de se placer dans un repère(X_i, Y_i) et de reproduire les différents couples $(X_i, Y_i) \ \forall i = 1, ...N$ naturellement le nuage de points ainsi obtenu, se situe soit sur la droite $Y = 0$, soit sur la parallèle $Y = 1$

Ainsi comme on l'observe dans la figure:

Fig 1: Ajustement linéaire



Il est impossible d'ajuster de façon satisfaisante ,pour une seule droite ,le nuage de points associé à une variable dichotomique qui ,par nature ,est réparti sur deux droites parallèles .

Deuxiement, la spécification linéaire standard ne convient pas, aux variables dichotomiques, et plus généralement aux variables qualitatives, car elle pose un certain nombre de problèmes mathématiques:

1) Sachant que dans le cas d'une variables endogène y_i dichotomique, celle-ci ne peut prendre que les valeurs 0 ou 1, la spécification linéaire $[Y_i = \beta X_i + \varepsilon_i]$ implique que la perturbation ε_i ne peut prendre, elle aussi, que deux valeurs, conditionnellement au vecteur x_i:

$\varepsilon_i = 1 - \beta X_i$avec une probabilité de $p_i = P(Y_i = 1)$

$\varepsilon_i = -\beta X_i$ avec une probabilité de $(1 - p_i) = P(Y_i = 0)$

Ainsi la perturbation ε_i du modèle admet nécessairement une loi discrète, ce qui exclut en particulier l'hypothèse de normalité des résidus.

2) Lorsque l'on suppose que les résidus ε_i sont de moyenne nulle, la probabilité p_i associée à l'envènement $y_i = 1$ est alors déterminée de façon unique. En effet, écrivons

l'espérance des résidus.

$$E(\varepsilon_i) = p_i(1 - \beta X_i) - (1 - p_i)\beta X_i = p_i - \beta X_i = 0$$

On déduit immédiatement que:

$$p_i = \beta X_i = p(Y_i = 1)$$

Ainsi la quantité$\beta\ X_i$ correspond à une probabilité et doit par conséquent satisfaire un certain nombre de propriétes et en particulier appartenir à l'intervalle fermé $[0, 1]$

$$0 \leq \beta X_i \leq 1$$

Or rien n'assure que de telles conditions soient saitisfaites par l'estimateur des Moindres carrés utilisés dans le modèle linéaire $y_i = X_i\beta + \varepsilon_i$.

Si de telles contraintes ne sont pas assurées, le modèle

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad E(\varepsilon_i) = 0 \qquad \forall i = 1, ...N$$

n'a pas de sens

3) Enfin, même si l'on parvenait à assurer le fait que ces contraintes soient satisfaites par l'estimateur des Moindres carrés des paramétres du modèles linéaire, il n'en demeurerait pas moins une difficulté liée à la présence d'hétéroscédasticité. En effet, on constate immédiatement que, que dans le modèle $y_i = X_i\beta + \varepsilon_i$ la matrice de variance covariance des résidus varie entre les individus en fonction de leurs caractéristiques associées aux exogénes puisque:

$$\vee(\varepsilon_i) = \beta X_i(1 - \beta X_i)\ \forall i = 1, ...N$$

Pour démontrer ce résultat il suffit de considérer des résidus et de calculer la variance:

$$\begin{aligned} \vee(\varepsilon_i) &= E({\varepsilon_i}^2) = (1 - \beta X_i)^2 P(Y_i = 1) + (-\beta X_i)^2 P(Y_i = 0) \\ &= (1 - \beta X_i)^2 p_i + (-\beta X_i)^2 (1 - p_i) \end{aligned}$$

Sachant que d'après la relation précédente, on a $p_i = X_i\beta$ on en déduit que:

$$\begin{aligned} \vee(\varepsilon_i) &= (1-\beta X_i)^2\beta x_i + (-\beta X_i)^2(1-\beta x_i) \\ &= (1-\beta X_i)\beta X_i\left[(1-\beta X_i)+\beta X_i\right] \\ &= (1-\beta X_i)\beta X_i \end{aligned}$$

Or de plus ce problème d'hétéroscédasticité ne peut pas être résolu par une méthode d'estimation des Moindres carrés Généralisés tenant compte de la contrainte d'inégalité

$$0 \leq \beta X_i \leq 1 \qquad \forall i \in 1.....N,$$

puisque la matrice de variance covariance des perturbations dépend du vecteur β des paramètres à estimer dans la spécification linéaire, qui par nature sont supposés inconnus.

Pour ces raisons la spécification linéaire des variables dichotomique ,n'est jamais utilisée et l'on a recourt à des modèles Logit ou Probit que nous allons à présent étudier ,pour représenter ces variables.

1.1.1 Modélisation:

On peut l'interpréter à l'aide d'un exemple (Jacquot 8).

Exemple: migration d'un individu (changement de région de résidence).

Soit à modeliser

$$Y_i = \begin{cases} 1 \text{ si } i \text{ a changé de résidence} \\ 0 \text{ si non} \end{cases}$$

en fonction d'un vecteur de variables explicatives

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{bmatrix}$$

Nous venons de voir qu'il n'était pas possible de dire que Y_i était relié à X_i par une

liaison linéaire.

Pour nous sortir de ce problème, nous allon introduire le supplement de revenu auquel l'individu peut s'attendre s'il change de région de résidence que nous noteronsY_i^* . Ce supplément de revenu Y_i^* n'est pas une variable observable.(**variable latente**).

Nous supposerons.

a) que l'individu migre $\Leftrightarrow$ ce supplément de revenu auquel l'individu peut s'attendre en migrant est positif.

Si

$$Y_i = 1 \Longleftrightarrow Y_i^* > 0$$

b) que ce supplément de revenu auquel l'individu peut s'attendre en migrant est une variable inobservable certe mais continue, à la différence de y_i). Donc y_i^* est une variable linéaire des variables explicatives X_i.

On a donc

$$\begin{cases} Y_i^* = \beta_i X_i^t + \varepsilon_i \\ Y_i^* \text{ est inobservable} \\ \text{on observe} Y_i = I_{\left(Y_i^* > 0\right)} \end{cases}$$

Quelles hypothèses faire sur ε_i? On supposera que les ε_i sont iid et suivent une loi de fonction de répartition connue. En pratique:

- Si on suppose que les ε_i sont iid suivent une loi normale $N(0,1)$

i.e si

$$\Phi(X) = \int_{-\infty}^{X} \frac{1}{\sqrt{2\pi}} \exp(\frac{-z^2}{2}) dz$$

Alors le modèle est dit Probit

- Si on suppose que les ε_i sont iid et suivent une loi logistique

$$\Lambda(x) = \frac{1}{1+e^{-X}}$$

alors le modèle est dit Logit.

1.1.2 Identifiabilité de β et σ

Pourquoi impose t-on a priori une variance connue aux perturbations ?

Dans le modèle linéaire ordinaire , en effet ,on avait $\varepsilon_i \rightarrow N(0, \sigma^2)$.$\sigma^2$étant un paramètre à estimer .

Au lieu de faire l'hypothèse que les $\varepsilon_i$sont iid de fonction de répartition connue , on avait pu faire l'hypothèse que les $\frac{\varepsilon_i}{\sigma}$sont iid de fonction de répartitionconnue . σ étant un paramètre à estimer. Mais on se heurterait alors , comme nous allons le montrer , à un problème d'identification .

y_i prend la les valeurs 0 et 1

y_i suit donc , conditionnellement à x_i,une loi de bernoulli de paramètre :

$$\begin{aligned} p_i &= P\left(Y_i = 1\right) \\ &= P\left(Y_i^* > 0\right) \\ &= P(\beta_i X_i^t + \varepsilon_i > 0) \\ &= P(\varepsilon_i > -\beta X_i^t) \\ &= P(\frac{\varepsilon_i}{\sigma} > \frac{-\beta X_i^t}{\sigma}) \\ &= 1 - P(\frac{\varepsilon_i}{\sigma} < \frac{-\beta X_i^t}{\sigma}) \end{aligned}$$

Ainsi $y_i \sim B\left(p_i\right)$ avec

$$p_i = 1 - P(\frac{\varepsilon_i}{\sigma} < \frac{-\beta X_i^t}{\sigma})$$

Il découle de ce résultat que si (β_0, σ_0) est un couple de valeurs satisfaisant le modèle :

$$Y_i^* = \beta X_i^t + \varepsilon_i$$

Y_i^* non observable

$$Y_i = 1_{Y_i^* > 0)}$$

ε_i iid

Alors il en est de même de (β_1, σ_1) avec $\beta_1 = \alpha\ \beta_0$ et $\sigma_1 = \alpha\ \sigma_0$ quel que soit $\alpha \in R_+^*$

Deux valeurs différentes du vecteur des paramètres $\begin{pmatrix} \beta \\ \sigma \end{pmatrix}$ donnent le même modèle : β et σ ne sont pas identifiables . par contre , le rapport $\frac{\beta}{\sigma}$est identifiable .

Pour rendre β identifiable , nous l'avons vu , deux hypothèses de distributions sont utilisées pour les ε_i:

Les ε_i sont iid $N(0,1)$: modèle probit

Les $\varepsilon_i$sont iid suivent une loi logistique de fonction de répartition$\Lambda(x) = \frac{1}{1+e^{-x}}$ modèle logistique

Dans le cas du modèle logit . on montre aisément que $\sigma^2 = \frac{\pi^2}{3}$. Equivalents en pratique à ce facteur de proportionnalité près, les estimateurs obtenus avec un modèle probit sont assez voisins du modèle logit. L'estimation d'un modèle logit étant moins couteuse en ressources informatiques que celle d'un modèle probit .

La modélisation est donc la suivante :

LOGIT	*PROBIT*
$Y_i^* = \beta X_i^t + \varepsilon_i$	$Y_i^* = \beta X_i^t + \varepsilon_i$
Y_i^* non observable	Y_i^* non observable
$Y_i = 1_{(Y_i^*>0)}$	$Y_i = 1_{(Y_i^*>0)}$

ε_i iid de fonction de répartition

$$\Lambda(X) = \frac{1}{1+e^{-X}}$$

ε_i iid de fonction de répartition

$$\Phi(X) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Chapitre 2

Comparaison des modèles Logit et Probit:

(Hurlin 7).

Dï£¡finition 1 *(Hurlin 7.) Les modèles dichotomiques Probit et Logit admettent pour variable expliquée, non pas un codage quantitative associée à la réalisation d'un évenement, mais la*

probabilité d'apparition de cet évenement, conditionnellement aux variables éxogènes :

$$p_i = p(Y_i = 1/X_i) = F(\beta X_i^t) \ \forall_i = 1, ...N$$

Ou Fest une fonction de répartition .avec $x_i = (x_i^1, ...x_i^K)$,et $\beta = (\beta_{1,}..., \beta_K)$ le vecteur des paramètres à estimer $, \forall i = 1, ...N$. Tel que si

$$F \rightarrow N(0, 1)$$

on parle d'un modèle probit

Dï£¡finition 2 *Si*

$$F \rightarrow \Lambda(\beta X_i) = \frac{e^{\beta X_i^t}}{1 + e^{\beta X_{ii}^t}} \tag{2.1}$$

on parle d'un modèle logit.

Dans le cas du modèle logit , la fonction de répartition correspond à la fonction logistique

$$\Lambda(x) = \frac{1}{1+e^{-x}}$$

Dans le cas du modèle probit , la fonction de répartition correspond à la fonction de répartition de la loi normale centrée et réduite .

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

2.1 Propriétés des modèles Probit et Logit

(Hurlin 7.)

$\Lambda(x)$ et $\Phi(x)$ sont les fonctions de répartition de loi logistique et normale respectivement

les modèles Logit ont été introduits comme des approximations des modèles probit permettant des calculs plus simple.

La loi logistique de fonction de répartion $\Lambda(x)$ a pour moyenne 0, pour variance $\frac{\pi^2}{3}$, il est donc naturel de comparer à $\Phi(\omega)$ fonction de répartition de $N(0,1)$, la fonction $\Lambda_1(\omega)$ où

$$\Lambda_1(\omega) = \frac{1}{1+e^{(-\pi\omega/\sqrt{3})}} \tag{2.2}$$

$\Lambda(\omega)$, $\Lambda_1(\omega)$, $\Phi(\omega)$ sont sensiblement proches comme on p eut le constater à partir du tableau (.Amemiya 2.).

Tableau 0 : Comparaison des fonctions de répartition $\wedge(\omega)$,$\varnothing$ (ω) et $\wedge_1(\omega)$

ω	0	0.1	0.2	0.3	0.4	0.5	1	2	3
$\Phi(\omega$	0.5	0.5398	0.5793	0.61	0.65	0.69	0.84	0.97	0.99
$\Lambda(\omega)$	0.5	0.525	0.5498	0.5744	0.5987	0.6225	0.7311	0.8808	0.9526
$\Lambda_1(\omega)$	0.5	0.5452	0.5897	0.6328	0.6738	0.7124	0.8598	0.9741	0.9957

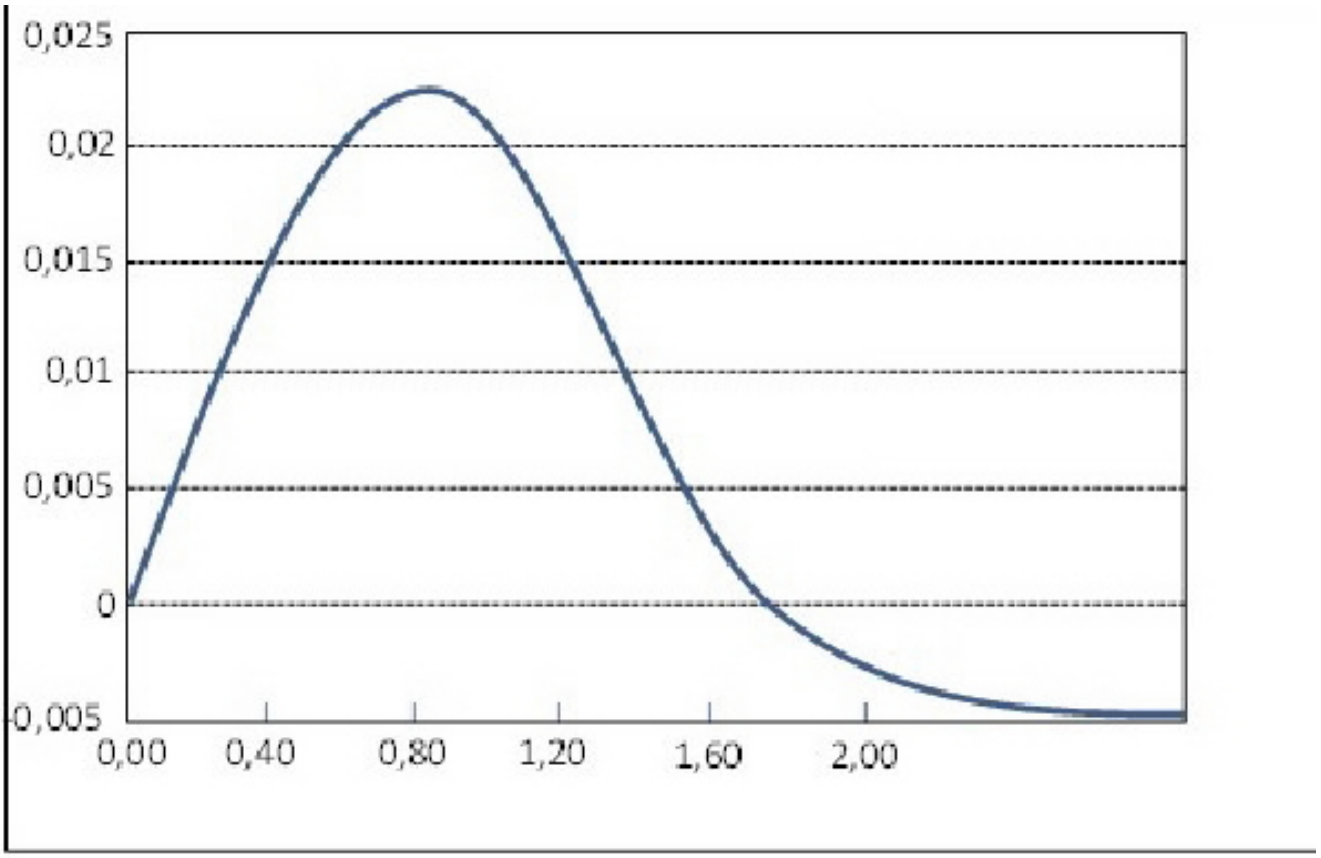
Même si les deux modèles se rapprochent ,il éxiste cependant certaines différences entre les modèles Probit et les modèles Logit . Nous évoquerons ici deux principales différences:

1) La loi logistique tend à attribuer aux évenements "extrèmes" une probabilité plus forte que la distribution normale.

2) Le modèle logit facilite l'interprétation des paramètres β associés aux variables explicatives x_i.

La figure ci-dessous donne en fonction de x la différence $\Lambda_1(x) - \Phi(x)$ des fonctions de répartition.

Fig 2.Différence des fonctions de répartitions $\Lambda_1(x) - \Phi(x)$



Au dela, de ces differences entre les lois logistiques et normales, il éxiste en effet certaines propiétés du modèle Logit qui sont particulièrement utiles pour simplifer les calculs. Tout d'abord si on note:

$p_i = P(Y_i = 1/Xi) = \wedge(\beta X_i^t)$ ($\wedge(\beta X_i^t)$ étant la fonction logistique (2.1)). On remarque que plusieurs égalités permettent de simplifier les calculs. Ils peuvents être établis comme suit:

$$p_i = \frac{e^{\beta X_i^t}}{1 + e^{\beta X_i^t}} \Longrightarrow e^{\beta X_i^t} = p_i(1 + e^{\beta X_i^t})$$

$$Log(\frac{p_i}{1-p_i}) = \beta X_i^t$$

$$e^{\beta xi} = \frac{p_i}{1-p_i}$$

En effet ,on sait que la probabilité p_i désigne la probabilité associée à l'évenement $y_i = 1$ et que la quantité $1-p_i$ désigne par conséquent la probabilité associée à l'évenement complémentaire $y_i = 0$.

Proposition 2.1 *(Hurlin 7) La quantité* $c_i = \frac{p_i}{1-p_i}$ *représente le rapport de la probabilité associée à l'évenement* $y_i = 1$ *à la probabilité de la non survenue de cet évenement, il s'agit de la cote (odds). Dans un modèle Logit cette cote correspond simplement à la quantité* $e^{\beta xi}$

$$c_i = \frac{p_i}{1-p_i} = e^{\beta xi} \tag{2.3}$$

Si ce rapport est égal à c_i pour l'individu i , cela signifie qu'il ya "c_i" fois plus de chances que l'évenement $y_i = 1$ se réalise. qu'il ne se réalise pas (c_i contre 1 dans le language usuel)

Au dèlà du simple calcul de la cote, on peut en outre chercher à mesurer les effets marginaux sur la cote.

Il s'agit alors de mesurer l'impact, pour le i^{eme} individu de variation de la j^{eme} variable explicative, notée x_{ij}, sur la cote.

Supposons que l'on considére une variation d'une unité de cette variable, et calculons alors la variation induite de la cote. En effet, étant donné la propriété du modèle logit, on peut alors facilement mesurer l'impact d'une variation d'une unité d'une des variables explicatives sur cette cote. En effet, si l'on note c la cote de l'évenement $y_i = 1, x_i = (x_{ik}; .....x_{ik})$ le vecteur des variables explicatives et $\beta = (\beta_1, ...\beta_k)'$ le vecteur des paramètres associés, on a:

$$c_i = \frac{p_i}{1-p_i} = \exp\left(\sum_{k=1}^{K} \beta_k x_{ik}\right) = \prod_{k=1}^{K} \exp(\beta_k x_{ik}) \tag{2.4}$$

On peut alors isoler la part de la cote imputable a une variable (X_{ij}) quelconque de la façon suivante.

Supposont que la variable X_{ij} augmente d'une unité, la nouvelle cote $\bar{c}_i$ est égale à :

$$\bar{c}_i = \exp\left[(x_{ij}+1)\beta_j\right] \prod_{\substack{k=1 \\ k\neq j}}^{K} \exp(x_{ik}\beta_j) = \exp(\beta_j) \prod_{k=1}^{K} \exp(x_{ik}\beta_k) \tag{2.5}$$

Proposition 2.2 *(.Hurlin 7)Dans un modèle logit, un accroissement d'une unité de la variable exogène X_{ij} ,(toutes choses égales par ailleurs), multiplie la valeur de la cote par* $\exp(\beta_j)$.

Si l'on note la cote initiale c_i et $\bar{c}_i$ la cote obtenue après variation de la j^{eme} variable explicative, on a:

$$\bar{c}_i = \exp(\beta_j) c_i \tag{2.6}$$

Un rapport entre la cote et ce que l'on appelle l'Odds-ratio (OR) est traité plus loin. On peut néanmoins faire remarquer que :

$$\begin{aligned} OR &= \exp(\beta_j) \\ &= \frac{\bar{c}_i}{c_i} \end{aligned} \tag{2.7}$$

Ce rapport correspondant à la probabilité de survenue de l'évenement $y_i = 1$ quand x_{ij}passe de l'etat x_{ij} à l'état $x_{ij} + 1$.

Toutefois, de façon plus générale, on calcule les effets marginaux non pas à partir de la cote mais directement à partir des probabilités associées à l'évenement de référence. On cherche ainsi à établir quelle est la variation de la probabilité de l'évenement $y_i = 1$, en cas de variation d'une des variables exogènes.

On considérera ici uniquement le cas de variables explicatives continues.

Dans ce cas pour de petites variations de la $j^{ème}$ variable explicative, on peut approximer la variation de probabilité p_i par la dérivée de celle ci par rapport à la variable explicative. On peut approximer la variation de probabilité p_i par la dérivée de celle ci par rapport à la variable X_{ij} :

$$\frac{\partial p_i}{\partial X_{ij}} = \frac{\partial F(\beta X_i^t)}{\partial X_{ij}} = \frac{\partial F(\beta X_i^t)}{\partial (\beta X_i^t)} \frac{\delta \beta X_i^t}{\delta X_{ij}} = \frac{\partial F(\beta X_i^t)}{\partial \beta X_i^t} \beta_j$$

puisque

$$\beta X_i = \sum_{k=1}^{K} \beta_k X_i$$

Proposition 2.3 *(Hurlin 7) Si l'on note f(.) la fonction de densité des résidus du modèle dichotomique, l'effet marginale associé à la j^{eme} variable explicative $x_i^{[j]}$ est défini par:*

$$\frac{\partial p_i}{\partial X_{ij}} = \beta_j f(\beta X_i) \tag{2.8}$$

Suivant que l'on considére un modèle probit ou un modèle logit, cette dérivée s'écrit comme suit:

$$\frac{\partial p_i}{\partial X_{ij}} = \beta_j \frac{e^{\beta X_i^t}}{(1 + e\beta X_i^t)^2} \qquad \textit{modèle logit} \tag{2.9}$$

$$\frac{\partial p_i}{\partial X_{ij}} = \frac{1}{\sqrt{2\pi}} \beta_j \exp\left[-\frac{1}{2}(\beta X_i^t)^2\right]. \qquad \textit{modèle probit} \tag{2.10}$$

Puisque par définition $f(.) > 0$, le signe de cette dérivée est donc identique à celui de β_j. Dès lors, l'augmentation d'une variable associeé à un coefficient positif induit une hausse de la probabilité de la réalisation de l'évenement $y_i = 1$.Inversement, la hausse d'une variable associée à un coefficient négatif induit une baisse de la probabilité de la réalisation de l'évenement $y_i = 1$.

Chapitre 3

Estimation des Paramètres par la Méthode du Maximum de Vraisemblance

Considérons le modèle suivant :

Dï£¡finition 3 *On considère un échantillon d'individus indicés* $i = 1, \ldots\ldots\ldots N$*.Pour chaque individus, on observe si un certain événement s'est réalisé et l'on note, la variable codée associée à l'événement comme suit:*

$$Yi = \begin{cases} 1 \quad tq \quad P(Y=1) \quad = \quad p_i \\ 0 \quad tq \quad P(Y_i=0) \quad = \quad 1-p_i \end{cases} \tag{3.1}$$

Où

$$X_i = (X_{i1}, \ldots\ldots\ldots X_{ik}) \ \forall i \in 1, \ldots\ldots N$$

désigne un vecteur de caractéristiques observables et ou $\beta = (\beta_1 \ldots\ldots \beta_K)' \in R^K$ *est un vecteur de paramètres inconnus*

On cherche naturellement à estimer les composantes du vecteur β par la méthode la plus utilisée lorsque la loi des perturbations est connue. Elle consiste en la méthode du maximum de vraisemblance.

3.1 Estimation par Maximum de Vraisemblance

Dans le cas du modèle dichotomique, la vraisemblance associée à l'observation y_i s'écrit sous la forme :

$$L(Yi, \beta) = p_i^{Yi}(1 - p_i)^{1-Y_i}$$

Définition 3.1 *(Hurlin 7) Pour un modèle dichotomique univarié simple, la vraisemblance associée à l'échantillon de taille N, noté* $y = (y_1, y_2, \ldots\ldots y_N)$ *s'écrit sous la forme :*

$$L(Y, \beta) = \prod_{i=1}^{N} p_i^{Y_i}(1 - p_i)^{1-Y_i} \tag{3.2}$$

et la fonction log-vraisemblance est comme suit

$$\begin{aligned} LogL(Y, \beta) &= Log \prod_{i=1}^{N} p_i^{Y_i}(1 - p_i)^{1-Y_i} = \sum_{i=1}^{N} Y_i Log\,(p_i) + (1 - Y_i) Log[1 - p_i)] \\ &= \sum_{i=1:y_i=1}^{N} Log\,(p_i) + \sum_{i=1:y_i=0}^{N} Log[1 - p_i)] \end{aligned} \tag{3.3}$$

Il ne reste plus alors qu'a spécifier la fonction de répartition pour obtenir la forme fonctionnelle de la vraisemblance.

Ainsi, $\forall(x_i\beta)$ dans le cas du modèle logit :

on a

$$\Lambda(\beta x_i). = \frac{e^{\beta x_i}}{1 + e^{x_i\beta}}$$

Alors que pour le modèle probit on a :

$$\phi(\beta X_i) = \int_{-\infty}^{\beta X_i} \frac{1}{\sqrt{2\pi}} e^{-z^2} dz$$

L'estimateur du maximum de vraisemblance des paramètres β est obtenu en maximisant soit la fonction de vraisemblance $L(Y, \beta)$ soit la fonction de log vraisemblance $LogL(Y, \beta)$.

En dérivant La Log vraisemblance par rapport aux éléments du vecteur β, de dimension $(K, 1)$, on obtient un vecteur noté $G(\beta)$ appelé vecteur du gradient .

Définition 3.2 *(Hurlin 7)L'estimateur* β *du maximum de vraisemblance du vecteur de*

paramètres $\beta \rightarrow \mathbb{R}^k$ dans un modèle dichotomique est défini par la résolution du système de K équations non linéaires en β :

$$\widehat{\beta} = \underset{\{\beta\}}{argmax}[LogL(Y,\beta)] \quad (3.4)$$

$$\Leftrightarrow \quad (\partial LogL(Y,\widehat{\beta}))/(\partial\widehat{\beta}) = \sum_{i=1}^{N} \frac{[Y_i - F(\ \widehat{\beta}X_i)]f(\widehat{\beta}X_i)}{(F(\widehat{\beta}X_i)[1 - F(\ \widehat{\beta}X_i)]} X_i^t = G(\widehat{\beta}) = 0 \quad (3.5)$$

(X_i^t est la transposée deX_i)

Ou $G(\beta)$ désigne le gradient associé à la log-vraisemblance évalué au point $\widehat{\beta}$

et $F(X)$ la fonction de répartition associée

Définition 3.3 *Dans le cas du modèle logit (3.5) devient :*

$$G(\widehat{\beta}) = \sum_{i=1}^{N}[Y_i - \Lambda(\ \widehat{\beta}X_i)]X_i^t = 0 \quad (3.6)$$

En effet on a : $\lambda(X) = \Lambda(X)(1 - \Lambda(X))$

Dans le cas du modèle probit (3.5) devient :

$$G(\widehat{\beta}) = \sum_{i=1}^{N} \frac{[y_i - \Phi(\widehat{\beta}X_i)]\phi(\ \widehat{\beta}X_i)}{(\Phi(\ \widehat{\beta}X_i)[1 - \Phi(\ \widehat{\beta}X_i)]} X_i^t = 0 \quad (3.7)$$

Remarque 3.1 *le système défini par l'équation (3.5) est non linéaire. L'estimateur $\widehat{\beta}$ ne peut être obtenu directement. Un algorithme d'optimisation numérique de la vraisemblance est donc nécessaire ,ces algorithmes se fondent a la fois sur le gradient mais aussi sur la matrice hessienne des dérivées secondes .*

3.1.1 Matrices Hessiennes et matrices d'information de Fisher :

Définition 3.4 *Pour un modèle dichotomique univarié, la matrice hessienne associée à la Log-vraisemblance d'un échantillon de taille N, noté $Y = (Y_1, \ldots\ldots, Y_N)$, s'écrit sous*

la forme

$$\begin{aligned} H(\beta) &= \frac{\partial^2 LogL(Y,\beta)}{\partial\beta\partial\beta'} \\ &= -\sum_{i=1}\left[\frac{Y_i}{F(X_i\beta)^2}+\frac{1-Y_i}{[1-F(X_i\beta)]^2}\right]f(X_i\beta)^2X_i^tX_i \\ &\quad +\sum_{i=1}^{N}\left[\frac{(Y_i-F(X_i\beta)}{F(X_i\beta)[1-F(X_i\beta)]}\right]f^t(X_i\beta).X_i^tX_i \end{aligned} \tag{3.8}$$

Ou $f\prime(.)$ désigne la dérivée de la fonction de densité $f(.)$ associée à $F(.)$.

en effet on a :

$$\begin{aligned} H(\beta) &= \frac{\partial}{\partial\beta}\left(\frac{\partial LogL(Y,\beta)}{\partial\beta}\right)=\frac{\partial}{\partial\beta}G(\beta)^t \\ &= \frac{\partial}{\partial\beta}\sum_{i=1}^{N}\frac{[Y_i-F(X_i\beta)]f(X_i\beta)}{(F(X_i\beta)[1-F(X_i\beta)]}X_i \\ &= \sum_{i=1}^{N}\frac{(F(X_i\beta)[1-F(X_i\beta)]}{(F(X_i\beta)^2[1-F(X_i\beta)]^2}\frac{\partial[Y_i-F(X_i\beta)]f(X_i\beta)}{\partial\beta}X_i \\ &\quad -\sum_{i=1}^{N}\frac{[Y_i-F(X_i\beta)]f(X_i\beta)}{(F(X_i\beta)^2[1-F(X_i\beta)]^2}\frac{\partial(F(X_i\beta)[1-F(X_i\beta)]}{\partial\beta}X_i \\ &= \sum_{i=1}^{N}\frac{-f^2(X_i\beta)+[Y_i-F(X_i\beta)]f^{'}(X_i\beta)}{(F(X_i\beta)[1-F(X_i\beta)]}X_i^tX_i \\ &\quad -\sum_{i=1}^{N}\frac{[Y_i-F(X_i\beta)]f(X_i\beta)}{(F(X_i\beta)^2[1-F(X_i\beta)]^2}\left[[1-F(X_i\beta)]f(X_i\beta)-F(X_i\beta)f(X_i\beta)\right]X_i^tX_i \\ &= -\sum_{i=1}^{N}\frac{f^2(X_i\beta)}{(F(X_i\beta)[1-F(X_i\beta)]}X_i^tX_i+\sum_{i=1}^{N}\frac{[Y_i-F(X_i\beta)]f\acute{(}X_i\beta)}{(F(X_i\beta)[1-F(X_i\beta)]}X_i^tX_i-\sum_{i=1}^{N}\frac{f^2(X_i\beta)[Y_i-F(X_i\beta}{(F^2(X_i\beta)[1-F(X_i} \end{aligned}$$

$$H(\beta)=-\sum_{i=1}\left[\frac{Y_i}{F(X_i\beta)^2}+\frac{1-Y_i}{[1-F(X_i\beta)]^2}\right]f^2(X_i\beta)X_i^tX_i+\sum_{i=1}^{N}\left[\frac{(Y_i-F(X_i\beta)}{F(X_i\beta)[1-F(X_i\beta)]}\right]f'(X_i\beta).X_iX_i$$

Notons que cette écriture n'est pas simplifiable, par contre l'espérance de la matrice hessienne est beaucoup plus simple. En effet l'égalité (3.8) et du fait que

$E(Y_i)=F(X_i\beta)$

on a alors :

$$
\begin{aligned}
E(H(\beta)) &= E\left[\frac{\partial^2 LogL(Y,\beta)}{\partial\beta\partial\beta'}\right] \\
&= -\sum_{i=1}\left[\frac{E(Y_i)}{F(X_i\beta)^2}+\frac{1-E(Y_i)}{[1-F(X_i\beta)]^2}\right]f(X_i\beta)^2X_i^tX_i+\sum_{i=1}^{N}\left[\frac{(E(Y_i)-F(X_i\beta)}{F(X_i\beta)[1-F(X_i\beta)]}\right]f'(X_i\beta).X \\
&= -\sum_{i=1}\left[\frac{F(X_i\beta)}{F(X_i\beta)^2}+\frac{1-F(X_i\beta)}{[1-F(X_i\beta)]^2}\right]f(X_i\beta)^2X_i^tX_i+\sum_{i=1}^{N}\left[\frac{(F(X_i\beta)-F(X_i\beta)}{F(X_i\beta)[1-F(X_i\beta)]}\right]f'(X_i\beta).X \\
&= -\sum_{i=1}\left[\frac{f^2(X_i\beta)X_i^tX_i}{F(X_i\beta)[1-F(X_i\beta)]}\right]
\end{aligned}
$$

On reconnait alors dans ce cas l'expression de l'opposé de la matrice d'information de Fisher

Définition 3.5 *(Hurlin 7). Pour un modèle dichotomique univarié, la matrice d'information de Fisher s'écrit sous la forme*

$$I(\beta) = -E\left[\frac{\partial^2 L(Y,\beta)}{\partial\beta\delta\beta^t}\right] = \sum_{i=1}^{N}\frac{f^2(X_i\beta)}{F(X_i\beta)[1-F(X_i\beta)]}X_i^tX_i \tag{3.9}$$

Dans le cas du modèle Logit on a :

$$I(\beta) = \sum_{i=1}^{N}\lambda(X_i\beta)X_i^tX_i = \sum_{i=1}^{N}\frac{\exp(X_i\beta)}{[1+\exp(X_i\beta)]^2}X_i^tX_i \tag{3.10}$$

Dans le cas du modèle Probit on a :

$$I(\beta) = \sum_{i=1}^{N}\frac{\phi^2(X_i\beta)}{\Phi(X_i\beta)[1-\Phi(X_i\beta)]}X_i^tX_i \tag{3.11}$$

En effet (3.10) est obtenue à l'aide de la relation :$\Lambda(X)\left[1-\Lambda(X)\right] = \lambda(X)$.($\Lambda$ indiquant la F r de la fonction logit et λ indiquant sa fonction de densité).

3.1.2 Unicité du maximum global de la fonction de log vraisemblance

Si l'on admet que le maximum global de $Log(Y,\beta)$ existe, la condition suffisante pour que ce maximum soit unique consiste à montrer que la fonction $LogL(Y,\beta)$ est concave. Etant

donnée l'écriture (3.3) de la log-vraisemblance, il suffit alors de monter que les fonctions $Log(\Lambda(X))$ et $Log[1-\Lambda(X)]$ sont concaves.

Dans le cas du modèle logit, les dérivées première et seconde de la fonction : $Log(F(X)) = Log(\Lambda(X))$ sont comme suit:

$$\begin{aligned}\frac{\partial Log(\Lambda(X))}{\partial X} &= \frac{1}{\Lambda(X)}\frac{\partial \Lambda(X)}{\partial X} \\ &= \frac{1+e^X}{e^X}\frac{e^X}{(1+e^X)^2} = \frac{1}{1+e^X}\end{aligned}$$

et

$$\frac{\partial^2 Log(\Lambda(X))}{\partial X^2} = \frac{\partial}{\partial X}\left(\frac{1}{1+e^X}\right) = \frac{-e^X}{(1+e^X)^2} < 0$$

et

$$\begin{aligned}\frac{\partial Log(1-\Lambda(X))}{\partial X} &= -\frac{1}{1-\Lambda(X)}\frac{\partial \Lambda(X)}{\partial X} \\ &= -\frac{1+e^X}{1}\frac{e^X}{(1+e^X)^2} = -\frac{e^X}{1+e^X} = -\Lambda(X)\end{aligned}$$

enfin

$$\frac{\partial^2 Log(1-\Lambda(X))}{\partial X^2} = -\frac{\partial \Lambda(X)}{\partial X} = \frac{-e^X}{(1+e^X)^2} < 0$$

Ainsi dans le cas du modèle logit les fonctions $Log(F(X))$ et $Log[1-F(X)]$ sont donc strictement concaves. Et donc la log-vraisemblance $LogL(Y,\beta)$ est elle même strictement concave. S'il existeun maximum à cette fonction en β , ce maximum est unique. Le même résultat peut être démontré dans le cas probit.

Proposition 4 *(Hurlin. 7): dans un modèle dichotomique univarié , la fonction de log-vraisemblance ($Log(Y,\beta)$) est strictement concave, ce qui garantit l'unicité du maximum de cette fonction.*

Dans la pratique ce résultat garantit la convergence des estimateurs du maximum de

vraisemblance vers la vraie valeur β_0 des paramètres , quelque soit le choix des conditions initiales et de l'algorithme d'optimisation utilisé

3.1.3 L'estimation dans la pratique :

Dans la pratique, les logiciels utilisent une procédure approchée pour obtenir une solution satisfaisante de la maximisation ci-dessus.

La procédure la plus connue est la méthode de Newton Raphson qui est la méthode itérative du gradient , elle s'appuie sur la relation suivante :

$$\widehat{\beta}_i = \widehat{\beta}_{i-1} - \left[\frac{\partial^2 LogL(Y,\beta)}{\partial\beta\delta\beta'}\right]^{-1}_{\beta=\widehat{\beta}_{i-1}} \left(\frac{\partial LogL(Y,\beta)}{\partial\beta}|_{\beta=\widehat{\beta}_{i-1}}\right) \tag{3.12}$$

et d'après (3.5) on a aussi

$$\widehat{\beta}_i = \widehat{\beta}_{i-1} - H(\widehat{\beta}_{i-1})^{-1} G(\widehat{\beta}_{i-1}) \tag{3.13}$$

β_i est la solution courante à l'étape i

Les itérations sont interrompues lorsque la différence entre deux vecteurs de solution successifs $\widehat{\beta}_i - \widehat{\beta}_{i-1}$ ou la variation du critère $LogL(Y,\widehat{\beta}_i) - LogL(Y,\widehat{\beta}_{i-1})$ est inférieure à un certain seuil fixé dans le programme.d'itération.$\frac{\delta^2 LogL(Y,\beta)}{\delta\beta\delta\beta\prime}$

Le dernier estimateur obtenu $\widehat{\beta}_i = \widehat{\beta}$ correspond alors à l'estimateur optimal du maximum de vraisemblance. Notons que cet suite $\widehat{\beta}_i$ converge bien vers l'estimateur du maximum de vraisemblance.

En effet on vérifie immédiatement que si la suite $\widehat{\beta}_i$ converge vers une limite $\widetilde{\beta}$, cette limite est forcément solution des équations de vraisemblance..

Si l'on pose $\widetilde{\beta} = lim_{i\to\infty}\widehat{\beta}_i$ alors on obtient en appliquant (3.13) :

$$\widetilde{\beta} = \widetilde{\beta} - H(\widetilde{\beta})^{-1} G(\widetilde{\beta}) \Leftrightarrow H(\widetilde{\beta})^{-1} G(\widetilde{\beta}) = 0$$

La matrice hessienne $H(\widetilde{\beta})$, matrice des dérivées partielles secondes de la vraisemblance notée $H(\beta)$, étant définie positive, on a bien $G(\widetilde{\beta}) = \dfrac{\partial LogL(Y,\widetilde{\beta})}{\partial\beta} = 0$

Par conséquent, si la suite $\widehat{\beta}_i$ des estimateurs obtenus par l'algorithme de Newton Raphson,convergent vers une quantité $\widetilde{\beta}$, cette quantité est solution des équations du

premier ordre du programme de maximisation de la vraisemblance. Autrement dit, si la suite $\widehat{\beta}_i$ converge, elle converge alors nécessairement vers l'estimateur du maximum de vraisemblance $\widehat{\beta}$ défini par la

condition :

$$G(\widetilde{\beta}) = \frac{\partial LogL(Y, \widetilde{\beta})}{\partial \beta} = 0$$

3.1.4 Exemple d'un modèle logistique :

i) **Présentation du modèle :**

- Soit Y une variable qui suit une loi de Bernoulli :

$Y \sim B(P(X))$ ou $P(X)$ représente $P(Y = 1/X)$

- soit $r = Y - P(X)$

On a

$$E(r) = E(Y - p(X)) = E(Y) - p(X) = 0$$

$$Var(r) = Var(Y) = p(X)(1 - p(X))$$

Soit la transformation $A : \mathbb{R} \to [0, 1]$ definie par :

$$A(z) = \frac{\exp(z)}{1 + \exp(z)}$$

On va utiliser la fonction de probabilité :

$$P(X) = A\left(X^t \beta\right) \text{ avec } \beta \in \mathbb{R}^p$$

Avec l'utilisation de la fonction inverse

$$Logit\ P = \log \frac{P}{1 - P}$$

On peut ecrire le modèle de regression comme :

$$LogitP\left(X\right)=X^{t}\beta$$

ii) **E-M-V du paramètre** β :

Pour estimer la vraie valeur β^{0} de β , on considère un échantillon de n observations :

$$\begin{aligned}\left\{\left(Y_{i},X_{i}\right)\right\}_{i\in\{1,...n\}} &\in \left(\{0,1\}\times\mathbb{R}^{p}\right)^{n}\\ Y_{i} &\sim B\left(A\left(X^{t}\beta^{0}\right)\right)\ (\ i=1....n)\end{aligned}$$

On considère le modèle : $Y=A\left(X\beta^{0}\right)+r$

avec

$$Y=\left(Y_{1},Y_{2}.....Y_{n}\right)^{t}\in\{0,1\}^{n}$$

$$X=\begin{pmatrix}X_{1}^{t}\\ .\\ .\\ X_{n}^{t}\end{pmatrix}\in\mathbb{R}^{nXp}$$

$$r=Y-A\left(X\beta^{0}\right)\Rightarrow E\left(r_{i}\right)=0$$

$$var\left(r_{i}\right)=S_{i}^{2}=A\left(X_{i}\beta^{0}\right)\left(1-A\left(X_{i}\beta^{0}\right)\right)$$

$$=\frac{\exp\left(X_{i}\beta^{0}\right)}{1+\exp\left(X_{i}\beta^{0}\right)}\times\frac{1+\exp\left(X_{i}\beta^{0}\right)-\exp\left(X_{i}\beta^{0}\right)}{1+\exp\left(X_{i}\beta^{0}\right)}$$

$$=\frac{\exp\left(X_{i}\beta^{0}\right)}{\left(1+\exp\left(X_{i}\beta^{0}\right)\right)^{2}}=A^{t}\left(X_{i}\beta^{0}\right)$$

On a

$$L\left(\beta\right)=\Pi_{i=i}^{n}P\left(Y=Y_{i}\right)$$

$$= \Pi_{i=i}^{n} A\left(X_i\beta^0\right)^{Y_i}\left(1 - A\left(X_i\beta^0\right)\right)^{1-Y_i}$$

$$= \Pi_{i=i}^{n}\left(\frac{A\left(X_i\beta\right)}{\left(1 - A\left(X_i\beta\right)\right)}\right)^{Y_i} \times \left(1 - A\left(X_i\beta\right)\right)$$

$$= \Pi_{i=i}^{n}\left(\frac{\frac{\exp(X_i\beta)}{\left(1+\exp\left(X_i\beta^0\right)\right)^2}}{\frac{1+\exp(X_i\beta)-\exp(X_i\beta)}{1+\exp(X_i\beta)}}\right)^{Y_i} \times \frac{1 + \exp\left(X_i\beta\right) - \exp\left(X_i\beta\right)}{1 + \exp\left(X_i\beta\right)}$$

$$= \Pi_{i=i}^{n}\frac{\left(\exp\left(X_i\beta\right)\right)^{Y_i}}{\left(1 + \exp\left(X_i\beta\right)\right)}$$

$$= \Pi_{i=i}^{n}\frac{\left(\exp\left(Y_iX_i\beta\right)\right)}{\left(1 + \exp\left(X_i\beta\right)\right)}$$

$$\begin{aligned}\log L\left(\beta\right) &= \Sigma_{i=1}^{n}\log\left(\exp\left(X_i\beta_i\right)Y_i\right) - \Sigma_{i=1}^{n}\log\left(1 + \exp\left(X_i\beta_i\right)\right)\\ &= Y^tX\beta - \Sigma_{i=1}^{n}\ln\left(1 + \exp\left(X_i\beta_i\right)\right)\end{aligned}$$

Pour déterminer le maximum de cette fonction, on considère le gradient :

$$\begin{aligned}\bigtriangledown \ln L\left(\beta\right) &= X^tY - \Sigma_{i=1}^{n}\frac{\left(\exp\left(X_i\beta\right)\right)}{\left(1 + \exp\left(X_i\beta\right)\right)}X_i\\ &= X^tY - X^tA\left(X\beta\right)\\ &= X^t(Y - A\left(X\beta\right))\end{aligned}$$

Une condition nécéssaire pour que $\hat{\beta}$ soit un estimateur du maximum de vraisemblance de β^0 :

$$X^t\left(Y - A\left(X\hat{\beta}\right)\right) = 0$$

- par la 2ème dérivée partielle :

$$\frac{\delta^2}{\delta\beta_s\beta_r}\ln L\left(\beta\right) = \frac{\delta}{\delta\beta_s}\left(\Sigma_{i=1}^{n}X_iY_i - \Sigma_{i=1}^{n}X_iA\left(X_i\beta\right)\right)$$

$$= -\Sigma_{i=1}^{n} X_i \frac{\delta}{\delta \beta_s} A\left(X_i\beta\right) = -\Sigma_{i=1}^{n} X_i X_i A^{'}\left(X_i\beta\right)$$

-On définit la matrice hessienne "$H_{l\ n\ 1}\left(\beta\right)$" $\in \mathbb{R}^{p\times p}$ par :

$$H_{LnL}\left(\beta\right) = -X^t D\left(\beta\right) X$$

où $D\left(\beta\right) = \left(d_{ij}\right) \in \mathbb{R}^{n\times n}$ est une matrice diagonale définie par:

$$d_{ij} = \begin{pmatrix} A^t\left(X_i\beta\right) & \text{si} \quad i = j \\ 0 & \text{sinon} \end{pmatrix}$$

On a $H_{LnL}\left(\beta\right)$ est définie negative pour chaque $\beta \in \mathbb{R}^p$, en effet

On a

$$u^t H_{LnL}\left(\beta\right) u = -u^t X^t D\left(\beta\right) X u = -\Sigma_{i=1}^{n}\left(X_i u\right)^2 A^t\left(X_i\beta\right)$$

On a

$$A^{'}\left(X_i\beta\right) \geq 0 \text{alors} \quad u^t H_{LnL}\left(\beta\right) u \leq 0 \ \forall u \in \mathbb{R}^p \text{ et } \beta \in \mathbb{R}^p$$

- donc

$$X^t\left(Y - A\left(X\hat{\beta}\right)\right) = 0 \quad \Rightarrow \hat{\beta} \text{ est } L'EMV$$

3.1.5 Lois et variance asymptotique de l'estimateur du maximum de vraisemblance

Nous allons nous intéresser à la loi asymptotique de l'estimateur du maximum de vraisemblance ainsi qu'à sa variance asymptotique.

Proposition 5 *voir (Hurlin 7):* ***O****n suppose que les variables explicatives données par le modèle sont des variables aléatoires continues ou des variales déterministes donc pour garantir la convergence et la normalité asymptotique des estimateurs deux approches sont retenues suivant la nature des variables explicatives :*

1) Si les variables explicatives sont aléatoires continues on suppose que les variables X_i sont indépendantes identiquement distribuées de même loi iid

2) Si les variables explicatives sont déterministes les conditions imposent alors au valeurs X_i d'être bornées :

$\exists\ m > 0 \quad et\ \ \exists\ M < \infty\ tq\ m\ \ <\mid X_{ik} \mid< M\ \ ,\ \forall k \in \mathbb{R}$

sous ces conditions sur les variables explicatives, l'estimateur du maximum de vraisemblance β est convergent et suit asymptotiquement une loi normale de moyenne égale à la vraie valeur β_0 des paramètres et de matrice de variance covariance égale à l'inverse de la matrice d'information de Fisher $I(\beta_0)$ évaluée au point β_0 :

$$\sqrt{N}(\beta - \beta_0) \underset{n\to\infty}{\overset{L}{\to}} N[0, I(\beta_0)^{-1}] \tag{3.14}$$

Avec

$$I(\beta_0) = -E\left(\frac{(\partial^2 LogL(Y,\beta)}{\partial\beta\partial\beta'}\right)_{\beta=\beta_0} = \sum_{i=1}^{N} \frac{f^2(X_i\beta_0)}{F(X_i\beta_0)[1-F(X_i\beta_0)]} X_i^t X_i \tag{3.15}$$

Démonstration :

si l'on note : $G(\beta) = \dfrac{\partial LogL(.)}{\partial\beta}$ le vecteur de gradient et $H(\beta) = \dfrac{\partial^2 LogL(.)}{\partial\beta\delta\beta^t}$ la matrice Hessienne, on sait que l'estimateur du maximum de vraisemblance satisfait à la condition : $G(\widehat{\beta}) = 0$. Considérons un développement limité à l'ordre 1 autour de la vraie valeur des paramètres β_0. En omettant les termes de degré supérieur à 2, il vient :

$$G(\widehat{\beta}) = G(\beta_0) + H(\beta_0)(\widehat{\beta} - \beta_0) = 0$$

En pré multipliant cette égalité par $H^{-1}(\beta_0)$, on obtient $(\widehat{\beta} - \beta_0) = -H(\beta_0)^{-1}G(\beta_0)$ ce qui peut se réécrire sous la forme :

$$\sqrt{N}(\widehat{\beta} - \beta_0) = -\left[\frac{1}{N}H(\beta_0)\right]^{-1}\ [\sqrt{N}\bar{g}(\beta_0)]$$

où le vecteur $\overline{g}(\beta_0)$ de dimension $(K, 1)$ est défini par :

$$\bar{g}(\beta_0) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} \partial logL(Y_i,\beta)/\partial\beta_1) \\ \vdots \\ \sum_{i=1}^{N} \partial logL(Y_i,\beta)/\partial\beta_{K-1}) \\ \sum_{i=1}^{N} \partial logL(Y_i,\beta)/\partial\beta_K) \end{pmatrix}$$

en supposant que chaque composante est iid ,

$$Z_j = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial logL(Y_i,\beta)}{\partial \beta_j}$$

On a

$$E\left(Z_j\right) = \frac{1}{N}\sum_{i=1}^{N}E(\frac{\partial logL(Y_i,\beta)}{\partial \beta_j})$$

Or

$$E(\frac{\partial logL(Y_i,\beta)}{\partial \beta_j}) = E\left[\sum\frac{\left(Y_i-F\right)f}{F\left(1-F\right)}\right]X_i$$

Ou

$$E\left[\sum_{i=1}^{N}\frac{[Y_i-F(X_i\beta)]f(X_i\beta)}{(F(X_i\beta)[1-F(X_i\beta)]}X_i\right]X_i = \left[\sum E\left(\frac{[Y_i-F(X_i\beta)]f(X_i\beta)}{(F(X_i\beta)[1-F(X_i\beta)]}X_i\right)\right] = 0$$

Car

$$E\left(Y_i\right) = p_i = F(X_i\beta)$$

D'ou

$$E\left[\bar{g}(\beta_0)\right] = \begin{pmatrix} 0 \\ . \\ . \\ 0 \end{pmatrix}$$

Et donc d'après le théorème central limite

$$Z_j = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial logL(Y_i,\beta)}{\partial \beta_{j1}} \rightarrow N\left(0,1\right)$$

$$\bar{g}(\beta_0) \rightarrow N\left(0,I_d\right)$$

• D'autre part on a d'après la loi des grands nombres

$$\frac{H\left(\beta_0\right)}{N} \rightarrow E\left(H\left(\beta_0\right)\right) = I_f\left(\beta_0\right)$$

Où $I_f\left(\beta_0\right)$ est l'information de Fischer évaluée au point β_0

en appliquant les deux théorèmes à la formule suivante :

$$\sqrt{N}(\beta - \beta_0) = -[\frac{1}{N}H(\beta_0)]^{-1}\ [\sqrt{N}\bar{g}(\beta_0)]$$

On obtient :

$\sqrt{N}.\left(\hat{\beta} - \beta_0\right)$ a une distribution normale de moyenne 0 et de matrice de variance covariance $-E\left(H\left(\beta_0\right)\right)$

Remarque 6 *(Hurlin 7).Concernant la matrice de variance covariance asympotique de* $\widehat{\beta}$*, notée* $V(\widehat{\beta})$=$I\left(\beta_0\right)^{-1}$*. Il est naturel que cette matrice de variance covariance dépende de la vraie valeur du paramètre* β_0 *qui est par définition inconnue. Dès lors, on retient généralement comme estimateur de la matrice de variance covariance asympotique la matrices* $I\left(\widehat{\beta}\right)^{-1}$ *dans laquelle la vraie valeur du paramètre* β_0 *a été remplacée par son estimateur* $\widehat{\beta}$*. On a ainsi :*

$$V(\widehat{\beta}) = I\left(\widehat{\beta}\right)^{-1} = \left[-E\left(\frac{(\partial^2 LogL(Y,\beta)}{\partial\beta\partial\beta^t}\right)_{\beta=\widehat{\beta}}\right]^{-1} \tag{3.15}$$

Chapitre 4

Méthodes d'estimation non Paramétriques :

Un des problèmes qui peut se poser lors de la phase d'estimation des paramètres des modèles dichotomiques par maximum de vraisemblance provient de l'hypothèse que l'on fait sur la

distribution des résidus du modèle. Considérons le modèle dichotomique suivant :

$$Yi = \left\{ \begin{array}{ll} 1 & si\ Y_i^* = \beta_0 X_i^t + \varepsilon_i \quad \geq 0 \\ 0 & \text{sin on} \end{array} \right\}$$

Ou ε_i est une perturbation iid $(0, {\sigma_\varepsilon}^2)$

Quand on cherche à estimer les paramètres β par maximum de vraisemblance, on on choisit une certaine distribution pour les termes ε_i. On

considère par exemple une distribution logistique dans le cas d'un modèle logit et une distribution normale dans le cas probit. Or, rien ne garantit a priori que cette distribution que

l'on utilise pour construire la vraisemblance de l'échantillon corresponde réellement à la "vraie" distribution des perturbations ε_i Naturellement, une erreur sur la distribution des termes ε_i

conduit alors nécessairement à une estimation du maximum de vraisemblance non efficace des paramètres β.

Une des solutions pour se prémunir contre ce risque de mauvaise spécification de

la loi des perturbations du modèle, consiste à s'afranchir de toute de hypothèse sur la distribution paramétrique des résidus dans la phase d'estimation des paramètres β. On parle alors de méthodes d'estimation non paramètriques. Nous ne présenterons ici que la méthode du score maximum.

:

4.1 Méthode Du Score Maximum

Dï£¡finition 7 *(Klein 9) L'estimateur du score maximum est obtenu par la maximisation, par rapport au vecteur $\beta \in R^k$ d'un critère constitué du nombre de fois où $X_i\beta > 0$ lorsque $Y_i = 1$ et du nombre de fois ou $X_i\beta < 0$ lorsque $Y_i = 0$*

$$\hat{\beta}_s = arg\ max\ \frac{1}{N}\sum_{i=1}^{N} I_{(Y_{i=1})}I_{(X_i\beta>0)} + I_{(Y_{i=0})}I_{(X_i\beta<0)}$$

Où I_A désigne la fonction indicatrice de l'ensemble A

Le critère du score maximum consiste alors a maximiser en β la fréquence empirique (le score) des événement $Y_i = 1$ lorsque $X_i\beta > 0$ et $Y_i = 0$ lorsque $X_i\beta < 0$.

L'idée générale de cette méthode est la suivante. On sait que la probabilité associée à l'événement $Y_i = 1$ est définie par $p_i = P(\varepsilon_i < X_i\beta) = F(X_i\blacksquare)$. En d'autres termes, on a $Y_i = 1$ quand l'inégalité $\varepsilon_i < X_i\beta$ est vérifiée. Si l'on considère à présent des valeurs de $\blacksquare_i$ suffisamment faibles relativement à $X_i\beta$, cette relation peut être approximée de la façon suivante : $X_i\beta - \varepsilon_i \simeq X_i\beta$ .Ainsi, on doit observer $Y_i = 1$ quand $X_i\beta$ est positif, si on dispose de la vraie valeur β_0 du vecteur β. Parallèlement, on doit observer $Y_i = 0$ quand $X_i\beta$ est négatif. En termes de probabilités on obtient les approximations suivantes :

$$\begin{aligned} P(Y_i &= 1) \simeq P(X_i\beta) > 0) \\ P(Y_i &= 0) \simeq P(X_i\beta) < 0) \end{aligned}$$

Le critère du score maximum consiste alors à maximiser en β la fréquence empirique (le score) des évements $(Y_i = 1)$ et $(X_i\beta > 0)$.

Une autre interprétation équivalente ,de la méthode du score est qu'elle compare le signe de la prédiction $X_i\beta$ avec celui de la variable transformée $\blacksquare i = 2Y_i - 1$qui prend la valeur -1 quand $Y_i = 0$ et la valeur 1quand $Y_i = 1$. On compare donc une valeur observée $\blacksquare i$ qui est positive quand l'événement $Y_i = 1$ se réalise avec la quantité $X_i\beta$, qui pour la vraie valeur β_0 du vecteur β, doit elle aussi être positive quand l'événement $Y_i = 1$ se réalise.

4.2 Tests de Spécification et Inférence

Introduction

Comment tester le modèle dichotomique ? Comment tester les paramètres de ce modèle ?

Autant de questions auxquelles nous allons à présent tacher de répondre. Nous commencerons par évoquer les tests d'hypothèse sur les coefficients, puis dans une seconde sous section nous envisagerons les principaux tests de spécification sur les modèles dichotomiques.

4.2.1 Tests d'hypothèse sur les paramètres

(Hurlin 7).

Les différentes méthodes d'estimation présentées précédemment conduisent à des estimateurs asymptotiquement normaux lorsque le nombre d'observations tend vers l'infini. Il est donc

facile d'utiliser ces divers estimateurs pour construire des procédures de tests dont certaines seront asymptotiquement équivalentes. Nous présenterons ici les principales procédures de test à partir de la méthode d'estimation du maximum de vraismeblance qui est la plus souvent utilisée. On retouve alors les tests les plus fréquents :

1. Test de Wald
2. Test du rapport des maxima de vraisemblance : LRT (Likelihood Ratio Test)
3. Test du score ou multiplicateur de Lagrange : LM (Lagrange Mulitplier)

On rappelle que ces trois tests sont asymptotiquement équivalents, ce qui implique qu'ils peuvent notamment se contredire sur des petits échantillons. De plus, leur distrib-

ution n'étant valide qu'asymptotiquement, il convient d'être prudent dans leur utilisation sur de petits échantillons.

On sait en outre que le test LRT est localement le plus puissant et que donc il devrait être a priori préféré. Nous n'envisagerons ici que le cas d'un test bidirectionnel sur un coefficient ou sur un ensemble de coefficients.(Ceci dans le but d'avoir des intervalles de confiances pour les OR que nous verrons plus loin.)

4.2.2 Test de Wald

On considère le test $H0 : \beta_j = a$ contre $H1 : \beta_j \neq a$ où β_j désigne la $j^{ième}$ composante du vecteur de paramètres $\beta = (\beta_1, .., \beta_K)' \in \mathbb{R}^K$ d'un modèle dichotomique.

On sait que la Statistique du Test de Wald associée au test bidirectionnel admet la loi suivante sous $H0$:(Proposition 4)

$$\left[\widehat{\beta_j} - a\right]' (\widehat{s}_{jj})^{-1} \left[\widehat{\beta_j} - a\right] = \frac{\left[\widehat{\beta_j} - a\right]^2}{(\widehat{s}_{jj})} \longrightarrow^{L}_{N\to\infty} \chi^2(1) \tag{4.1}$$

où $\widehat{s}_{jj}$ désigne l'estimateur de la variance de l'estimateur du $j^{jième}$ coefficient β_j.

Ainsi si l'on note $\chi^2_{95\%}(1)$ le quantile à 95% de la loi de $\chi^2(1)$, le test de Wald au seuil de 5% de l'hypothèse $H0$ consiste à accepter $H0$ si $\frac{\left[\widehat{\beta_j} - a\right]^2}{(\widehat{s}_{jj})}$ est inférieure à $\chi^2_{95\%}(1)$ et à la refuser sinon.

Remarque 8 *La plupart des logiciels (sauf SAS) ne proposent pas cette statistique de Wald, mais une statistique Z_j définie comme la racine carré de la précédente. Compte tenu du lien entre la loi normale centrée réduite et la loi du χ^2 à un degré de liberté, on a immédiatement sous $H0$:*

$$Z_j = \frac{\left[\widehat{\beta_j} - a\right]}{\sqrt{(\widehat{s}_{jj})}} \longrightarrow^{L}_{N\to\infty} N(0,1) \tag{4.2}$$

De même si l'on note $t_{\frac{\alpha}{2}}$ le quantile à $1 - \frac{\alpha}{2}$de la loi Normale $N(0,1)$e $\chi^2(1)$, le test de Wald au seuil de 5% de l'hypothèse $H0$ consiste à accepter $H0$ si $\left|\frac{\left[\widehat{\beta_j} - a\right]}{\sqrt{(\widehat{s}_{jj})}}\right|$ est inférieure à $t_{\frac{\alpha}{2}}$ et à la refuser sinon.

Dans ce cas , vu l'expression de $L'OR$ =exp($\widehat{\beta}_j$) (voir Formule (2.7)) , un intevalle de confiance au niveau $1-\frac{\alpha}{2}$ de l' OR quand X_i passe de l'etat X_i à l'état X_i+1.

est :

$$IC_{(1-\alpha)\%} = \exp(\left[\widehat{\beta}_j\right] \pm t_{\frac{\alpha}{2}}\sqrt{(\widehat{s}_{jj})} \tag{4.3}$$

4.2.3 Tests du rapport des maxima de vraisemblance

Dans le cas des modèles dichotomiques, on peut appliquer sans difficulté particulière la logique du test du rapport des maxima de vraisemblance. Ainsi, on estime le modèle non contraint et d'autre part le modèle contraint :

Soient $\widehat{\beta}_j$ et $\widehat{\beta}_j^{\,c}$ les deux estimateurs ainsi obtenus.La statistique LRT correspond alors simplement à la différence des log-vraisemblance.

On considère le test $H0 : \beta_j = a$ contre $H1 : \beta_j \neq a$.La statistique LRT , notée G ,admet la loi suivante sous $H0$:(Proposition 4)

$$G_j = -2\left[\log(L(Y,\widehat{\beta}_j) - \log(L(Y,\widehat{\beta}_j^{\,c})\right] \longrightarrow^{L}_{N\to\infty} \chi^2(1) \tag{4.4}$$

Notons que cette porcédure est asymptotiquement équivalente à celle d'un test de Wald.

Dans le cas d'un test portant sur plus d'un paramètre, on utilise la statistique suivante:

$$G = -2\left[\log(L(Y,\widehat{\beta}) - \log(L(Y,\widehat{\beta}^{c})\right] \longrightarrow^{L}_{N\to\infty} \chi^2(r) \tag{4.5}$$

où r désigne le nombre de restrictions imposées sur les paramètres, et où $\widehat{\beta}$ et $\widehat{\beta}^{c}$ désignent les estimateurs respectivement non contraint et contraint du vecteur complet β.

4.2.4 Test du score ou du multiplicateur de Lagrange

Le principe de ce test est le suivant. On sait que si l'hypothèse nulle est satisfaite, les deux estimateurs non contraint $\widehat{\beta}_j$et contraint $\widehat{\beta}_j^{\,c}$ doivent être relativement proches l'un de l'autre, et que donc la même propriété doit être vérifiée pour le vecteur des conditions du premier ordre de la maximisation de la log-vraisemblance.

La statistique LMj , notée SC du test du multiplicateur de Lagrange associée au test $H0 : \beta_j = a$ contre $H1 : \beta_j \neq a$,admet la loi suivante sous $H0$:(Proposition 4)

$$SC = \left(\frac{\partial \log(L(Y,\beta)}{\partial \beta^{'}} \mid_{\beta=\widehat{\beta^c}} \right)^{'} \widehat{I}^{-1} \left(\frac{\partial \log(L(Y,\beta)}{\partial \beta^{'}} \mid_{\beta=\widehat{\beta^c}} \right) \longrightarrow^{L}_{N\to\infty} \chi^2(1) \qquad ((4.6))$$

où $\widehat{\beta}_j$ et $\widehat{\beta}^c_j$ désignent les estimateurs respectivement non contraint et contraint de β_j.
L'estimateur $\widehat{I}$ de la matrice d'information de Fischer peut être obtenu par :

$$\widehat{I} = \sum_{i=1}^{N} \left(\frac{\partial \log(L(Y_i,\beta)}{\partial \beta^{'}} \mid_{\beta=\widehat{\beta^c}} \right) \left(\frac{\partial \log(L(Y,\beta)}{\partial \beta^{'}} \mid_{\beta=\widehat{\beta}^c} \right)^{'}$$

et où

$$\frac{\partial \log(L(Y,\beta)}{\partial \beta^{'}} \mid_{\beta=\widehat{\beta^c}} = \sum_{i=1}^{N} \left(\frac{\partial \log(L(Y_i,\beta)}{\partial \beta^{'}} \mid_{\beta=\widehat{\beta^c}} \right)$$

4.3 Qualité de l'ajustement d'un modèle logistique:

(Nakache .12)

Elle est basée sur la mesure de la concordance entre le vecteur $Y^{'} = (Y_1, Y_2, ...)$ des valeurs observées de la variable réponse Y et le vecteur $\hat{y}^{'} = (\hat{y}_1, \hat{y}_2, ...)$ (valeurs de la variable réponse estimées par le modèle .

On considère un modèle ajusté avec p variables indépendantes .(Pour la suite nous adopterons des notations simplifiées . Voir (.Nakache 12), pour le modèle logistique que nous détaillons comme suit ; Soit :

- J le nombre de vecteurs ou profils distincts $X^{'} = (X_1, X_2, ...X_p)$ observées :

-m_j le nombre de sujets avec $X = X_j$

-Y_j le nombre de sujets présentant $(Y = 1)$ parmi les sujets avec ($X = X_j$)

-$\hat{\pi}_j$ la probabilité $P(Y = 1/X_j)$ estimée par le modèle logistique .

On suppose J proche de n (cas le plus fréquent en pratique). Le nombre de sujets estimés pour le profil "j" est :

$$\hat{Y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(X_j)}}{1 + e^{\hat{g}(X_j)}} \qquad (4.7)$$

où $\hat{g}(X_j)$ est l'estimation du logit.

4.3.1 CHI-2 de Pearson et déviance résiduelle:

Le $CHI-2$ d'ajustement de Pearson a pour expression

$$\chi^2 = \Sigma_{j=1}^{J} r_j(Y_j, \hat{Y}_j)^2 \qquad où \;\; r_j = r(\; Y_j \,, \hat{Y}_j \;) = \frac{Y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} \tag{4.8}$$

La déviance résiduelle D est le rapport des vraisemblances du modèle saturé avec j paramètres et d'un modèle ajusté avec $(p+1)$ paramètres dont l'expression est la suivante :

$$D = \sum_{j=1}^{J} d_j^2(\; Y_j \,, \hat{Y}_j \;) \;\; \text{avec } d_j(\; Y_j \,, \hat{Y}_j \;) = \pm 2 \left[\left(Y_i \log\left(\frac{Y_j}{m_j\hat{\pi}_j} \right) + (m_j - Y_j) \log\left(\frac{m_j - Y_j}{m_j(1-\hat{\pi}_j)} \right) \right) \right]^{\frac{1}{2}} \tag{4.9}$$

Sous l'hypothèse (H_0 : modèle ajusté correct) , les mesures χ^2 et D suivent une distribution du χ^2 à $J-(p+1)$ degrés de liberté .

4.3.2 Test de Hosmer − Lemeshow

Le test de Hosmer - Lemeshow (6) relève à peu près de la même logique que le diagramme de fiabilité. A la différence qu'au lieu de se baser simplement sur une impression visuelle, on extrait du tableau de calcul un indicateur statistique qui permet de quantifier la qualité des estimations $\hat{\pi}_j$.Concrètement, nous procédons de la manière suivante

1- appliquons le modèles sur les données pour obtenir les estimations $\hat{\pi}_j \;\; (score)$

2- Trier les données selon le score croissant.

3- Subdiviser les données en G groupes en se basant sur les quantiles (exp: les quantiles d'ordre 4 correspondent aux quartiles, les quantiles d'ordre 10 aux déciles etc ...)

Les auteurs proposent prioritairement les déciles ($G = 10$ groupes g)

4- Dans chaque groupe g, d'effectifs mg, nous devons caluler plusieur quantités:

- m_{g_1}; le nombre de positifs observés.

- m_{g_0}; le nombre de négatifs observés.

-$\hat{m}_{g_1} = \sum_{\omega \in g} \hat{\pi}(\omega)$, la somme des scores des observations situées dans le groupe g. On la désigne comme la fréquence théorique des positifs dans le groupe;

- $\overline{\pi}_{g_1} = \dfrac{\widehat{m}_{g_1}}{m_g}$, la moyenne des scores observés dans le groupe g

- $\hat{m}_{g_0} = m_g - \widehat{m}_{g_1}$ la fréquence théorique des négatifs.

5- Nous calculons alors la statistique de Hosmer et Lemeshow en utilisant une des formules suivantes:

$$\widehat{C} = \sum_g \left[\frac{(mg_1 - \widehat{m}g_1)^2}{\widehat{m}g_1} + \frac{(mg_0 - \widehat{m}g_0)^2}{\widehat{m}g_0} \right] \tag{4.10}$$

$$= \sum_g \left[\frac{mg(mg_1 - \widehat{m}g_1)^2}{\widehat{m}g_1(mg - \widehat{m}g_1)} \right] \tag{4.11}$$

$$= \sum_g \frac{(mg_1 - \widehat{m}g_1)^2}{\widehat{m}g_1(mg - \overline{\pi}_{g_1})} \tag{4.12}$$

6- Lorsque le modèle est correcte H_0, la statistique $\widehat{C}$suit approximativement une loi du χ^2 à $G-2$) degrés de liberté.

7- Lorsque la probabilité critique du test (P value) est plus grande que le risque choisi, le modèle issu de la regression logistique est accepté.

8. Les réserves usuelles concernant ce type de test restent de mise ici. Il faudrait entres autres que tous les effectifs théoriques soient supérieurs à 5 dans toutes les cases du tableau. Si ce n'est pas le cas, on devrait pROCéder à des regroupements et corriger en conséquence les degrés de liberté. Mais il ne faut pas non plus s'arc-bouter à cette idée. Il s'agit d'un outil d'évaluation du classifeur, il donne avant tout une indication sur la qualité des $\hat{\pi}(\omega)$

4.3.3 Exemple: Acceptation de credit - Test de Hosmer Lemeshow: (Rakotomalala 14)

Penchons-nous sur des données un peu plus réalistes pour montrer l'intérêt de cette procédure. Dans le problème qui suit, nous souhaitons expliquer l'accord d'un prêt par un organisme de crédit à partir l'âge du référant, le revenu par tête dans le ménage, le fait d'être propriétaire de son habitation ou non, occuper une profession indépendante ou non,

le nombre de problèmes rencontrés avec sa banque. Nous disposons de $n = 100$ observations, avec $n+ = 73$ positifs.ayant suffisamment d'observations $n = 100$ pour que la subdivision en $G = 10$ groupes ne pose pas trop de problèmes.

Nous avons alors un premier tableau des scores suivants

Tableau 1 : Scores

1	y	Pi(Score)	17	0	0.5553
2	0	0.0001	18	0	0.5720
3	0	0.0044	19	0	0.5821
4	0	0.0195	20	0	0.6066
5	0	0.0337	21	1	0.6066
6	0	0.1110	22	0	0.6294
7	1	0.1345	23	0	06369
8	0	0.1424	24	0	0.6483
9	0	0.2100	25	0	0.6510
10	0	0.2600	26	1	0.6614
11	1	0.2828	27	1	0.6882
12	0	0.2876	28	1	0.7077
13	1	0.3377	29	1	0.7162
14	0	0.4765	30	1	0.7230
15	1	0.5148	32	1	0.7265
16	1	0.5551	33	1	0.7378

le tableau suivant résume les groupes et effectifs théorique et observés par décile

Tableau 2: Classes d'effectifs

groupe	décile	effectif	observés	théoriques	observés	théoriques
1	0.2871	10	2	1.1985	8	8.8015
2	0.6249	10	4	5.0945	6	4.9055
3	0.7344	10	6	6.7886	4	3.2114
4	07874	10	7	7.5886	3	2.4114
5	0.8146	10	7	8.0422	3	1.9578
6	0.8485	10	10	8.3373	0	1.6627
7	0.8775	10	10	8.6720	0	1.3280
8	0.8917	10	8	8.8564	2	1.1436
9	0.9101	10	10	9.0357	0	0.9643
10	1.0000	10	9	9.3864	1	0.6136

La feuille de calcul est construite comme suit (Tableau 1, l'affichage est limité aux 32 premières observations) : Tout d'abord, nous calculons les déciles. Le 1er décile est égal à 0.271, le 2nd à 0.6249. Nous vérifions le nombre d'observations dans chaque groupe, nous avons bien mg = 10, ■g puisque n = 100.

Dans chaque groupe, nous comptons le nombre de positifs et de négatifs. Pour le 1er groupe par exemple, nous avons $m11 = 2$ et $m10 = 10 - 2 = 8$.

Puis nous calculons les effectifs espérés en faisant la somme des scores dans le groupe. Pour le 1er groupe, nous avons

$\widehat{m}_{11} = 0.0001 + 0.0044 + 0.0195 + \cdots + 0.2828 = 1.1985$. Nous en déduisons

$\widehat{m}_{10} = 10 - 1.1985 = 8.8015$.

Il ne reste plus qu'à calculer la statistique de Hosmer et Lemeshow en utilisant une des formules ci-dessus (4.9) et (4.10) par exemple. Pour la première, nous avons

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

Pour la seconde,

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

Les degrés de liberté étant égales à $ddl = 10 - 2 = 8$, nous obtenons une p-value de 0.4503avec la loi du $\blacksquare^2$

La $P - value$ est supérieure au risque usuel de 5%. Le modèle est validé, il est compatible avec les données.

Chapitre 5

Évaluation de la régression

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

Comparer les valeurs observées de la variable dépendante $Y(\blacksquare)$ avec les prédictions $\hat{Y}(\blacksquare)$.

Comparer les vraies valeurs π avec celles prédites par le modèle $\hat{\pi}$. En effet, la régression logistique fournit une bonne approximation de cette quantité . Elle peut se révéler très utile lorsque nous souhaitons classer les individus selon leurs degrés de positivité ou introduire d'autres calculs ultérieurement (ex. intégrer les coûts de mauvais classement).

5.1 La matrice de confusion

La matrice de confusion est une autre procédure d'évaluation de la régression logistique , elle confronte toujours les valeurs observées de la variable dépendante avec celle qui sont prédites, puis comptabilise les bonnes et les mauvaises prédictions :

Exemple : nous avons construit un modèle de prédiction qui vise à expliquer le « faible poids d'un bébé $= Y_i$»(Bernard 3)

Les variables explicatives sont :

$X_1 =$ fumer (le fait de fumer ou pas pendant la grossesse)

$X_2 =$(historique de l'hypertension)

$X_3 =$âge de la mère

Tableau 3: Matrice de confusion

$\hat{Y}/Y$	$\widehat{(Oui)}$	$\widehat{(Non)}$	$Somme$
Oui	$a = 21$	$b = 39$	$a + b = 60$
Non	$c = 10$	$d = 120$	$c + d = 130$
$Somme$	$a + c = 31$	$b + d = 159$	$a + b + c + d = n = 190$

Le taux d'erreur est égal au nombre de mauvais classement rapporté à l'effectif total c.-à-d.:

$$\epsilon = \frac{b + c}{n} \tag{5.1}$$

Il estime la probabilité de mauvais classement du modèle.

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

$$\theta = \frac{a + d}{n} = 1 - \epsilon \tag{5.2}$$

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

$$Se = Sensibilité = TVP = rappel = \frac{a}{a + b} \tag{5.3}$$

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

$$\text{Pr}écision = \frac{a}{a + c} \tag{5.4}$$

La spécificité, à l'inverse de la sensibilité, indique la proportion de négatifs détectés

$$Sp = Spécificité = \frac{d}{c + d} \tag{5.5}$$

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

$$TFP = \frac{c}{c+d} = 1 - Sp \tag{5.6}$$

Remarque 9 *Un "bon" modèle doit présenter des valeurs faibles de taux d'erreur et de taux de faux positifs (proche de 0). Des valeurs élevées de sensibilité, précision et spécifcité (proche de 1).Le taux d'erreur est un indicateur symétrique, il donne la même importance aux faux positifs et aux faux négatifs*

La sensibilité et la précision sont asymétriques, elles accordent un rôle particulier aux positifs.

Enfin, en règle générale, lorsqu'on oriente l'apprentissage de manière à améliorer la sensibilité, on dégrade souvent la précision et la spécificité. Un modèle qui serait meilleur que les autres sur ces deux groupes de critères antinomiques est celui qu'il faut absolument retenir.

L'exemple du Tabeau 3 suggère :

Dans la matrice de confusion nous lisons que sur les données d'apprentissage, le modèle de prédiction réalise 10+39 = 49 de mauvaises prédiction le taux d'erreur en résubstitution est de

$49/190 = 27,78\%$.

Plusieurs indicateurs peuvent être déduits pour rendre compte de la concordance des valeurs observées et les valeurs prédites.

Le taux de succès correspond à la probabilité de bon classement du modèle

$\theta = (21 + 120)/190 = 74\%$ complémentaire du taux d'erreur $\varepsilon = 1 - \theta$

La sensibilité : $TV_P = 21/(21 + 39) = 35\%$

La précision = Précision= $V_{PP} = 21/(21 + 10) = 67\%$

La spécifité, indique la proportion de négatifs détectés

$S_P = 120/(10 + 120) = 92\%$

5.2 La courbe ROC:

(Ricco 14)

La courbe ROC est un outil très riche. Son champ d'application dépasse largement le cadre de l'apprentissage supervisé. Elle est par exemple trés utilisé en épidémiologie .

Pour nous; elle présente surtout des caractristiques trés intéressantes pour l'évaluation et la comparaison des performances des classificeurs.

Elle propose un outil graphique qui permet d'évaluer et de comparer globalement le comportement des classifieur

Enfin, on peut lui associer un indicateur synthétique, le critère AUC (aire sous la courbe, en anglais **area under curve**), que l'on sait interpréter.

La courbe ROC met en relation le taux de vrais positifs TVp (la sensibilité, le rappel) et le taux de faux positifs $TFp(TFp = 1 - TVp)$ dans un graphique nuage de points.

Dans la pratique, nous pROCédons de la manière suivante:

1/ Calculer le score $\hat{\pi}(\omega)$ de chaque individu à l'aide du modèle de prédiction.

2/ Triez le fichier selon un score décroissant.

3/ Chaque valeur du score peut être protentiellement un seuil s.

Pour toutes les observations dont le score est supérieur ou égal à s, les individus dans la partie haute du tableau nous pouvons comptabilser le nombre positif $n_+(s)$ et le nombre négatif $n_-(s)$

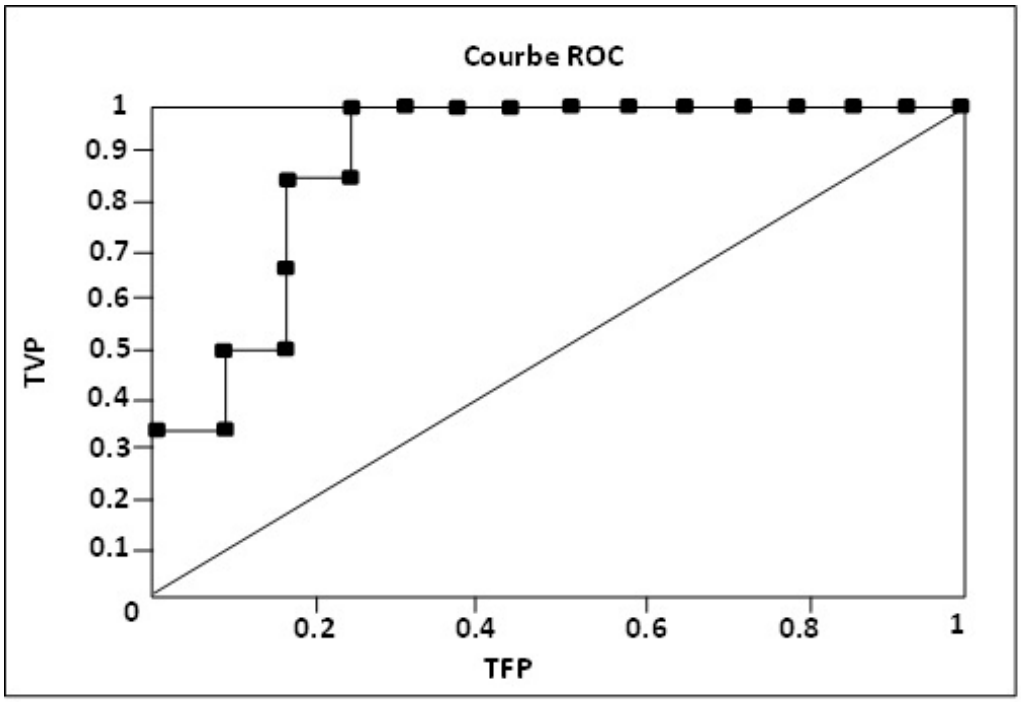
Nous en déduisons $TVp = \dfrac{n_+(s)}{n_+}$ et $TFp = \dfrac{n_-(s)}{n_-}$

4/ La courbe ROC correspond au graphique nuage de points qui relie les couples (TVp, TFp), le premier point est forcément $(0, 0)$, le dernier est $(1, 1)$.

Deux situation extrêmes peuvent survenir:

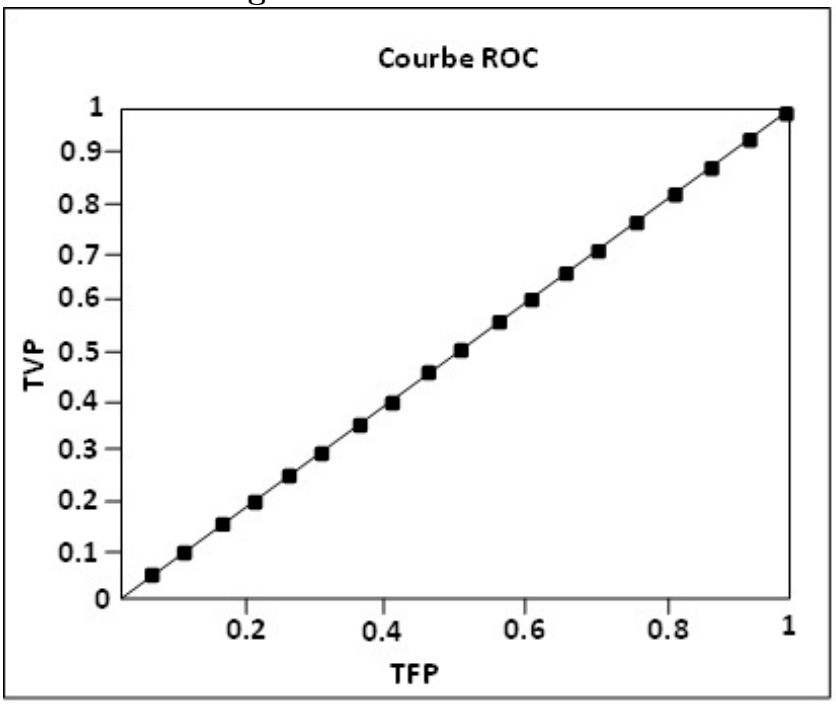
1) La discrimination est parfaite. Tous les positifs sont situés devant les négatifs, la courbe ROC est collée aux extrémités Ouest et Nord du repère.

Fig 3.Excellente discrimination



2) Les positifs et les négatifs sont mélangés - La courbe ROC se confond avec la premiére bissectrice.

Fig 4 : Aucune discrimination



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

5.2.1 Critère AUC

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

Tableau 4. Interprétation du critère AUC

Valeur de l'AUC	**Commentaire.**
AUC = 0.5	**Pas de discrimination.**
0.7 AUC < 0.8	**Discrimination acceptable**
0.8 AUC < 0.9	**Discrimination excellente**
AUC 0.9	**Discrimination exceptionnelle**

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

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

Exemple: fichier Coeur - Courbe ROC.(Rakotomalala 14)

Pour illustrer la construction de la courbe ROC, nous avons le fichier COEUR (Tableau 5)

Tableau 5 : Calcul Courbe ROC

Individu	Y+	Score (+)	Tvp	Tfn	Aire
			0	0	0
1	1	0,8789	0,16666667	0	0
2	1	0,8765	0,33333333	0	0
3	0	0,8584	0,33333333	0,07142857	0,023809524
4	1	0,5815	0,5	0,07142857	0
5	0	0,4057	0,5	0,14285714	0,035714286
6	1	0,3922	0,66666667	0,14285714	0
7	1	0,3782	0,83333333	0,14285714	0
8	0	0,3775	0,83333333	0,21428571	0,05952381
9	1	0,2134	1	0,21428571	0
10	0	0,1727	1	0,28571429	0,071428571
11	0	0,1382	1	0,35714286	0,071428571
12	0	0,1371	1	0,42857143	0,071428571
13	0	0,1244	1	0,5	0,071428571
14	0	0,1058	1	0,57142857	0,071428571
15	0	0,1037	1	0,64285714	0,071428571
16	0	0,0737	1	0,71428571	0,071428571
17	0	0,071	1	0,78571429	0,071428571
18	0	0,0584	1	0,85714286	0,071428571
19	0	0,0362	1	0,92857143	0,071428571
20	0	0,0164	1	1	0,071428571
					Somme=AUC=0,905

Nous savons qu'il y a $n_+ = 6$ positif et $n_- = 14$ négatifs dans ce fichier.

Nous avons calculé la colonne des scores $\hat{\pi}_i$, puis nous avons trié le Tableau 5 selon le score décroissant.

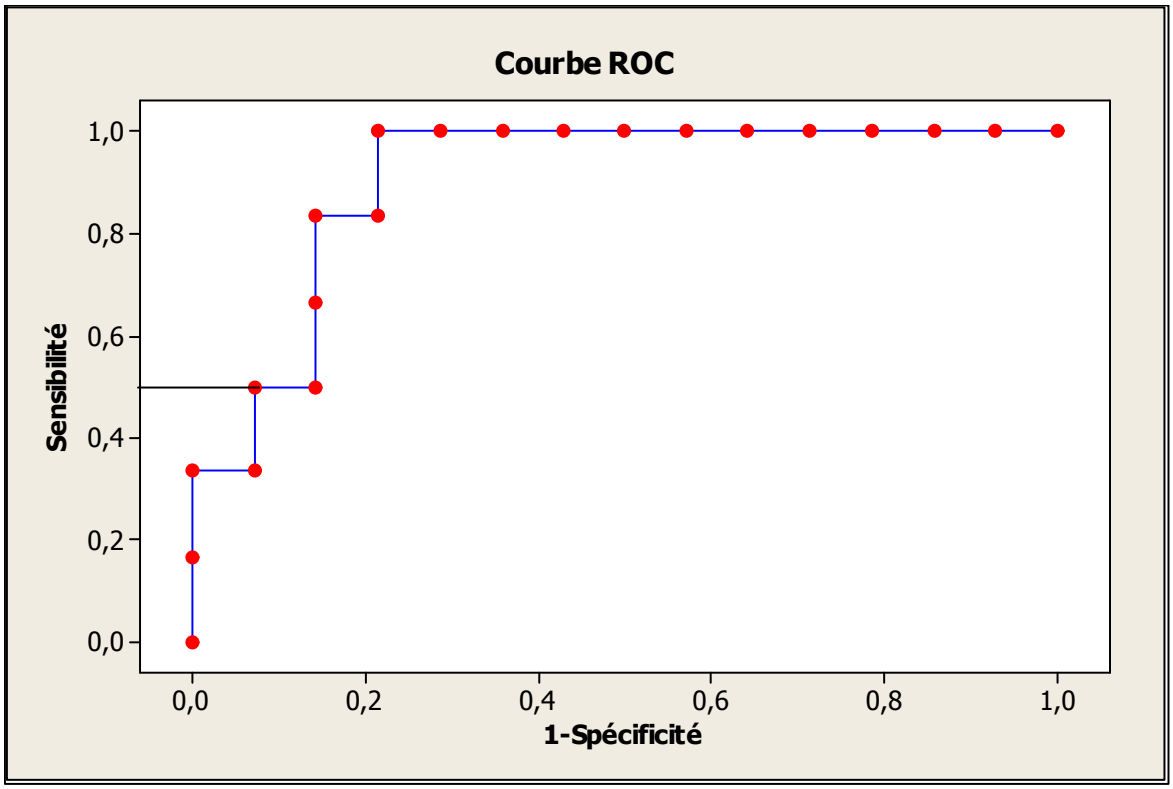
Nous insérons arbitrairement le couple $(0, 0)$.

- Il y a 1 individu ayant un score supérieur ou égal à 0.8789. Il est positif, soit $n_+(0.8789) = 1$ et $TVP1 = 1/6 = 0.1667$; par conséquent $n_-(0.8789) = 0$ et $TFP_1 = 0/14 = 0.00000$

- Prenons le cas de l'individu $N°4$ avec $n_+ : (0.5815) = 3$ et $TVP_4 = 3/6 = 0.5$; concernant les négatifs, nous avons $n_-(0.5815) = 1$ et $TFP4 = 1/14 = 0.0714$

- En procédant ainsi, nous obtenons l'ensemble des points. Il est d'usage d'ajouter la première bissectrice dans le graphique pour l'on se rende compte visuellement de l'écartement de la courbe ROC par rapport à la situation de référence. Notons dans ce cas une **Discrimination exceptionnelle avec un AUC=0.905.**

Fig 5: Courbe ROC



5.3 Les pseudo-R^2

(Ricco 14)

Une question cruciale est de pouvoir déterminer si le modèle obtenu est interessant ou non. Le premier à pouvoir trancher est l'expert. En son absence, il ne faut sutrout pas se lancer dans des considérations plus ou moin vaseuses. La seule attitude viable est de poser la question à quel classifieur de reference peut on se comparer ?.

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

Remarque 10 *Remarque : L'analogie avec le coefficient de détermination R^2 de la régression linéaire multiple est tout à fait intéressante. En effet, il est usuellement interprété comme la part de variance expliquée par le modèle. Mais il peut être également compris comme une confrontation entre les performances du modèle analysé (traduite par la somme des carrés des résidus :$SCR = \sum_{\omega}(Y - \hat{y})^2$) et celles du modèle par défaut réduite à la*

simple constante (dans ce cas, la constante est estimée par la moyenne de l'endogène $\overline{Y}$, la somme des carrés totaux correspond donc à la somme des carrés des résidus du modèle réduit à la simple constante $SCT = \sum_{\omega}(Y - \overline{Y})^2$. *Rappelons que* $R^2 = \dfrac{SCT - SCR}{SCT}$.

5.3.1 Estimation du paramétre a_0 et de la déviance du modèle trivial:

Le modèle trivial est réduit à la seul constante ie:

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

Nous ne tenons pas compte des variable X_j de fait:

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

On devine aisément l'estimateur $\hat{a}_0$ de la regression.

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

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

Et la $log - vraisemblance$ devient (d'après 3.3)

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

où $n_- = n - n_+$

$$l\ l_0 = \sum_i Y_i \ln(\hat{P}_i) + (1 - Y_i) \ln(1 - \hat{P}_i)$$

et donc

$$D_0 = -2LL_0.$$

* Reprenons l'exemple du fichier coeur (Tableau 5). Nous observons $n_+ = 6$ observations positifs parmi $n = 20$. Nous obtenons directement:

- Le nombre de négatifs $n_- = 20 - 6 = 14$

- La proportion de positif $\hat{P} = \frac{6}{20} = 0.3$

-L'estimation de la constante $\hat{a}_0 = \ln\left[\frac{n_+}{n_-}\right] = \ln\left[\frac{6}{14}\right] = -0.8473$

- La log vraisemblance $LL_0 = 20\ln(1 - 0.9) + 6\ln(\frac{0.3}{1-0.3}) = -12.217$

- La déviance $D_0 = -2LL_0 = -2(-12.217) = 24.4346$.

Quelques pseudo R^2

Tableau 6. Quelques pseudo-R^2 - Application au fichier COEUR

Indicateur	Formule	Valeur Min/ Max	Valeurs De R^2
R^2 de Mc Fadden	$R^2_{MF} = 1 - \frac{LLM}{LL0}$	$Min = 0$ si $LL_M = LLo$, on ne fait pas mieux que le modèle trivial $Max = 1$ si $LL_M = 0$, notre modèle est parfait.L'analogie avec le R^2 de la régression linéaire multiple est totale	$R^2_{MF} = 0.3199$
R^2 de Cox and Snell	$R^2_{CS} = 1 - (\frac{L_0}{L_M})^{\frac{2}{n}}$	$Min = 0.Max$ si $L_M = 1$, avec $max[R^2_{CS}] = 1 - L_0^{(\frac{2}{n})}$ L'indicateur n'est pas normalisé	$R^2_{CS} = 0.3235$
R^2 de Nagelkerke	$R^2_N = \frac{(R^2_{CS})}{max[R^2_{CS}]}$	$Min = 0$.$Max = 1$ C'est une simple normalisation de R^2 de Cox and Snell	$R^2_N = 0.4587$

Les pseudo-R2 résultent de l'opposition, sous différentes formes, de la vraisemblance du modèle étudié LM avec celle du modèle trivial L_0. Ils quantifient la contribution des descripteurs dans l'explication de la variable dépendante. Bref, il s'agit de véri er si notre modèle fait mieux que le modèle trivial c.-à-d. s'il présente une vraisemblance ou une log-vraisemblance plus favorable.

Plusieurs formes de pseudo-R2 sont proposés dans la littérature, nous en distinguons quelques uns (Tableau 6).

Les R^2 de Mac Fadden et de Nagelkerke sont les plus simples à appréhender : lorsque la régression ne sert à rien, les variables explicatives n'expliquent rien, l'indicateur vaut 0 ; lorsque la régression est parfaite, l'indicateur vaut 1. Notons que le R^2_{MF} de McFadden est le plus adapté à la régression logistique : il est le plus proche conceptuellement du coefficient de détermination de la

régression linéaire multiple ; il n'est pas sensible à des modifications de la proportion de positifs dans le fichier d'apprentissage.

Dans notre exemple, avec $R^2_{MF} = 0.3199$, il semble que notre modèle se démarque du modèle trivial.

On ne saurait pas dire en revanche si l'apport est significatif ou non, nous en saurons d'avantage lorsque nous aborderons l'évaluation statistique .

Chapitre 6

Pratique de la régression logistique binaire

Introduction

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

La régression logistique sait faire cela. Mais, à la différence d'autres méthodes, elle peut fournir en plus un indicateur de fiabilité de la prédiction avec une estimation de la probabilité $\widehat{\pi}(\omega^{'})$. Ainsi, lorsque $\widehat{\pi}$ est proche de 1 ou de 0, la prédiction est plutôt sûre ; lorsqu'elle prend une valeur intermédiaire, proche du seuil d'affectation $s(s = 0.5$ habituellement), la prédiction est moins assurée.

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

6.1 Lecture et interpretation des coefficients:

Dans certains domaines, l'explication est bien plus importante que la prediction. On souhaite comprendre les phénomènes de causalité, mettre à jour les relations de cause à effet. Bien entendu, les techniques statistiques n'ont vocation à répondre mécaniquement à des problèmes complexes. En revanche, elles ont pour rôle de donner aux experts les indicateurs adéquates pour qu'ils puissent se concentrer sur les informations importantes.

Par exemple, dans le domaine de santé, on cherche certes à détecter automatiquement une maladie particuliére, mais il est peut êre plus important que l'on comprenne pourquoi il l'a développe pour qu'on puisse l'anticiper.

Pour cela une étude approfondie des indicateurs s'impose, comme les risques relatifs(RR), odds, odds-ratio(OR).

6.1.1 Notations et Définitions.

Modèle unidimensionnel: une variable explicative:

1) variable binaire: on construit le tableau de contigence (y,x) qui fournit les différentes probabilités $p(Y/x)$:

Tableau 7: Tableau de contingence général

Y/X	$X=1$	$X=0$	$total$
$Y=1$	a	b	$a+b$
$Y=0$	c	d	$c+d$
$total$	$a+c$	$b+d$	$a+b+c+d$

A partir duquel on définit:

1/ Risque relatif: on appelle risque relatif le sucroît de chances d'être positif du groupe exposé par rapport au groupe témoin. On le note: RR

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

Soit par exemple le fichier COEUR,(Rakotomalala 14) (avoir une maladie cardiaque ou pas $Y = 1$ *ou* $Y = 0$) avec la variable explicative angine (avoir une angine ou pas $x = 1$ *ou* $x = 0$). Nous avons alors le tableau suivant :

Tableau 8 : Tableau de contingence :Croisement coeur vs. angine

Y/X	$X = 1$	$X = 0$	$total$
$Y = 1$	3	3	6
$Y = 0$	2	12	14
$total$	5	15	20

Nous l'interprétons de la manière suivante:

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

2/ Odds: L'odds ou rapport de chances ou cote (2.3) est défini comme un rapport de probabilité dans un groupe, par exemple dans le groupe exposé, (voir formule (2.3) il s'écrit:

$$odds(1) = \frac{P(+/1)}{P(-/1)}$$

$$= \frac{a/(a+c)}{c/(a+c)}$$

Dans l'exemple du tableau 8 on obtient:

$$odds(1) = \frac{3/5}{2/5} = 1.5$$

Dans le groupe de personnes ayant une angine de poitrine, on a 1.5 fois plus de chances d'avoir une maladie cardiaque que de ne pas en avoir.

Nous pouvons définir de la même manière odds (0).

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

3/ Odds-Ratio: l'odds ratio est égal au rapport entre l'odds du groupe exposé et l'odds du groupe témoin noté OR (voir formule 2.7)

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

Dans l'exemple du tableau 8 on obtient:

$$OR = \frac{3 \times 10}{3 \times 2} = 6$$

L'OR indique à peu près la même chose que le risque relatif, à savoir : dans le groupe exposé, on a 6 fois plus de chances d'avoir la maladie que dans le groupe témoin. Ils sont pratiquement égaux dès que $P(Y = 1/x = 1)$ et $P(Y = 1/x = 0)$ sont proches de zéro.

Relation entre odds - ratio et coefficients du modèle logistique :

$$\begin{aligned} Logit(X) &= \ln \frac{P(Y = 1/\times)}{P(Y = 0/\times)} = \beta_0 + \beta_1 X \\ pour\ X &= 1 \rightarrow Logit(1) = \beta_0 + \beta_1 \\ X &= 0 \rightarrow Logit(0) = \beta_0 \\ &\Rightarrow \ln(OR) = Logit(1) - Logit(0) = \beta \\ &\Longleftrightarrow OR = e^{\beta_1} \end{aligned}$$

Dans l'exemple croisant coeur et angine(Rakotomalala 14), l'odds ratio était égal à $OR = 6$. Si nous réalisons une regression logistique expliquant maladie du coeur avec angine comme seule variable explicative à l'aide du logiciel de traitement on obtient :

$LOGIT(X) = -1.3863 + 1.7918X$. Ainsi $\hat{\beta}_1 = 1.7918$ et donc $OR(angine) = e^{1.7918} \simeq 6$

Ainsi la regression logistique nous permet de mesurer directement le sucroît de risque associé à un facteur explicatif binaire.

Nous avons alors les indications suivantes :

Si $\hat{\beta}_j < 0 \Rightarrow OR < 1$, il ya diminution du risque

Si $\hat{\beta}_j > 0 \Rightarrow OR > 1$, il ya une augmentation du risque.

6.1.2 Intervalle de confiance de l'odds-ratio

Nous donné une formule générale de l'intervalle de confiance de l'odds-ratio (formule (4.3). Et comme la grande majorité des logiciels fournissent automatiquement ce type de résultat (avec un niveau de confiance fixé automatiquement à 95%). Nous allons détailler les calculs puisque nous disposons de l'estimation du coefficient et de son écart type.(voir (Rakotomalala 13) Pour un intervalle à 95%; le fractile de la loi normale utilisée est $\mu_{0.975} = 1.96$. Nous produisons les bornes de la manière suivante:

1) Borne min:

$$bm(\beta_1) = \hat{\beta}_1 - \mu_{0.975} \times \hat{\sigma}_{\hat{\beta}_1} \Rightarrow bm(OR) = e^{bm(\beta_1)}$$

et dans notre exemple on a

$$bm(\beta_1) = 1.791 - 1.96 \times 1.118 = -0.39$$

$$\Rightarrow bm(OR) = e^{-0.39} = 0.67.$$

2) Borne Max

$$bM(\beta_1) = \hat{\beta}_1 + \mu_{0.975} + \hat{\sigma}_{\hat{\beta}_1} \Rightarrow bM(OR) = e^{bM(\beta_1)}$$

$$= 1.791 + 1.96 \times 1.118 = 3.98$$

$$\Rightarrow bM(OR) = e^{3.98} = 53.68.$$

Dans cette section on a étudié seulement le cas unidimensionnel ie un seul prédicteur (une variable explicative binaire (ie $X_1 = 0$ ou 1))

Il éxiste encore plusieurs cas où le prédicteur peut être nominal-quantitatif, ordinal.et

d'autres modèles bidimensionnel où le modèle admet deux variables explicatives.

Pour le moment, on se contente de ce cas très particulier , à savoir un seul prédicteur binaire.

Dans le chapitre 4 , les tests sur un ou plusieurs coefficients ont été traités et qu'on récapitule comme suit :

6.1.3 Récapitulation

Pour tester

$$H_0 \left\{ \begin{array}{c} (\beta_j = 0) \\ contre\ \beta_j \neq 0 \end{array} \right\}$$

et $\beta = (\beta_1 ....\beta_k)$, on peut a priori envisager de trois maniéres différentes.(.Jacquot 1)

1- examiner si l'écart entre $\hat{\beta}$ et β_j^c est important cette idée est à la base du test de Wald déjà examiné plus haut, puisque la statique de Wald est une transformation simple de $\hat{\beta}_j - \beta_j$

Ce test s'appuie sur la normalité asymptotique de l'estimateurs du maximum de vraisemblance.

Le principal avantage est que les informations que l'on souhaite exploiter sont toutes disponibles à l'issue de l'estimation du modèle complet incluant l'ensemble des variables.

2- on peut aussi comparer $Log\ l\ (\hat{\beta})$ et $Log\ l\ (\beta_j^c)$, il ya peu de chances que H_0 soit vraie si l'écart entre ces 2 grandeurs est important. Cette idée est à la base du test du rapport des maximum de vraisemblance.

3- On peut enfin regarder si la pente de la log-vraisemblances est significativement differente de 0, cette idée est la base du test du score (on dit aussi test du multiplicateur de Lagrange)

à mesure

1. Donc la statistique du test de Wald est asociée au test bidirectionnel :

$$H_0 \left\{ \begin{array}{c} (\beta_j = 0) \\ \beta_j \neq 0 \end{array} \right.$$

$$\text{avec } \frac{\hat{\beta}_j}{\hat{\delta}_j} \frac{L}{n \to \infty} N(0,1) \text{ avec } \hat{\beta}_j \text{est l'estimateur de } \beta_j$$

Il s'agit de tester le rôle significatif d'une variable.

2. La statistique du test de vraisemblance est associée au test unidirectionnel

$$H_0 \begin{cases} (\beta_j = 0) \\ \beta_j \neq 0 \end{cases}$$

$$LRT = -2\left[Logl(g, \hat{\beta}_j) - Logl(Y, \hat{\beta}_j^c)\right] \frac{L}{n\to\infty} \chi^2(1)$$

Où $\hat{\beta}_j$ et $\hat{\beta}_j^c$ désignent respectivement les estimateurs non contraint et contraint de $\hat{\beta}_j$

3. La statistique de Lagrange est associée au test unidirectionnel $H_0 \begin{cases} (\beta_j = 0) \\ \beta_j \neq 0 \end{cases}$ qui admet la loi suivante sous H_0

$$LMj = (\frac{\delta Logl(Y,\beta)}{\delta\beta}/)'_{\beta=\hat{\beta}^c} \hat{I}^{-1} (\frac{\delta Logl(Y,\beta)}{\delta\beta}/)_{\beta=\hat{\beta}^c}$$

Chapitre 7

Analyse des interactions

On parle d'interaction lorsque l'effet d'une explicative sur la variable dépendante dépend du niveau (de la valeur) d'une autre variable explicative

Exemple :fumer est mauvais pour la santé être diabétique et fumer en même temps , c'est pire .

Il faut que l'on puisse décrire l'interaction sous la forme d'une nouvelle variable que la régression logistique saura prendre en compte .

On parle d'interaction d'ordre 1 lorsque l'on croise deux variables , d'ordre 2 lorsque l'on croise 3 variables . etc.. . . . . .

7.1 Interactions entre variables explicatives

Interaction par le produit de variables :On caractérise généralement l'interaction par le produit de deux (ou plusieurs) variables.

La signification n'est pas la même selon leur type. Lorsque les variables sont des indicatrices, soit parce qu'elles sont binaires par nature, soit parce qu'il s'agit d'une indicatrice de modalité d'une variable qualitative, le produit indique la conjonction des caractéristiques. Par exemple, si X_1 =fumeur et X_2 = diabétique la variable $Z = X1 * X2$ prend la valeur 1 lorsque l'on a affaire à un fumeur diabétique. Elle prend la valeur 0 lorsqu'il s'agit d'un fumeur qui n'est pas diabétique, ou d'un diabétique qui ne fume pas, ou lorsque la personne n'est ni fumeur, ni diabétique. L'insertion de la variable Z dans la régression permet de vérifier l'interaction. Si l'impact du tabac est constant que l'on soit diabétique

ou pas, le coefficient associé à Z ne devrait pas être significatif, dans le cas contraire, s'il est significativement diffèrent de 0, cela veut dire que l'impact du tabac n'est pas le même chez les diabétiques et les non diabétiques. On parle de modèle saturé lorsque l'on intégre toutes les interactions possibles dans la régression.

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

Le LOGIT s'écrit :

$$LOGIT = a_0 + a_1X_1 + a_2X_2 + a_3Z$$

$$= a_0 + a_1X_1 + a_2X_2 + a_3X_1.X_2$$

$$= a_0 + (a_1 + a_3X_2)X_1 + a_2X_2$$

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

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

$$\Delta Logit(\Delta X_1 = 1) = Logit(\Delta X_1 = 1) - Logit = a_1 + a_2X_3$$

De manière plus générale, la variation du Logit lorsque X_2 évolue de d unités s'écrit :

$$\Delta Logit(\Delta X_1 = d) = (a_1 + a_3X_2) \times d$$

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

Exemple 11 *(Rakotomalala 14) On cherche à déterminer les facteurs d'une maladie*

(ronflement)à partir d'un fichier comportant $n = 100$ *adultes.*

Les variables explicatives étudiées sont genre (homme = 1) et le tabac (fumeur = 1) nous réalisons la régression sur ces deux indicatrices dans un premier temps. Il semble au risque 10% qu'être un homme est propice a la maladie ,le tabac joue un rôle également.

Introduisons la variable Z= homme × tabac, Nous souhaitons savoir si la conjonction être un homme fumeur entraine une augmentation du risque d'avoir la maladie

Tableau 9: Ronflement = f (homme, tabac)

attiributes dans l'équation				
attribute	coef	Std-dev	Wald	Signif
constant	-1.903	-	-	-
homme	1.267	0.592	4.583	0.323
tabac	0.789	0.473	2.766	0.0963

Remarque : la lecture en termes de conjonctions en est une parmi les autres. Bien souvent, dans les études réelles les variables explicatives ne jouent pas même rôle. Dans notre exemple, on peut par exemple étudier l'effet du tabac (facteur de risque) sur l 'exposition à la maladie. Puis analyser si cet effet est diffèrent selon que l'on est un homme ou une femme. La variable « genre » (homme) est appelée variable modératrice.

Nous relançons la regréssion avec la troisième variable Z. nous constatons que :

La variable traduisant l'interaction est significative les hommes fumeurs sont exposés plus que les autres (ou, si nous sommes dans le schéma « facteur de risque vs effet moderateurs », le tabac ne joue pas un rôle différencié selon le genre).

Un autre résultat doit attirer notre attention, curieusement, les autres indicatrices ne sont plus significatives à 10%.

Cela laisse à penser que les variables ne pesent pas individuellemen sur le risque d'avoir la maladie.

Tableau 10: Ronflement= f (homme, tabac, homme × tabac)

attiributes dans l'équation				
attribute	coef	Std-dev	Wald	Signif
constant	-2.197	-	-	-
homme	1.5863	1.092	2.1102	0.1463
tabac	1.1856	1.2051	0.968	0.3252
homme*tabac	-0.479	1.313	0.1333	0.715

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

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

7.1.1 Stratégie pour explorer les intéractions

Les considérations de la section précédente nous amènent à un aspect très important de ce chapitre : les stratégies d'exploration des intéractions . il est évident que l'on ne pas s'appuyer sur des procédures purement mécaniques comme celles qui sont décrites dans le chapitre consacré à la sélection des variables . il faut tenir compte du rôle des variables dans differents niveaux d'interaction .

Un modèle est dit " hiérarchiquement bien formulé " (HBF) si toutes les intéractions d'ordre inférieur de l'interaction d'ordre le plus élevés sont présents . "

Exemple: Si l'interaction $X_1 * X_2 * X_3$ est présente dans la regression , nous devons y retrouver également les interactions d'ordre 1 c-a-d $X_1 * X_2, X_1 * X_3$et $X_2 * X_3$, mais aussi les interactions d'ordre 0 (les variables individuellement) c-a-d X_1 , X_2 et X_3. Cette contrainte doit être respectée lors du processus de selection des variables .

Deux situations envisageables :

1) Si $X_1 * X_2 * X_3$ est significatif , nous arrétons le processus de sélection des variables , toutes les autres interactions sont conservées.

2) Dans le cas contraire , nous pouvons la supprimer.Reste à définir une stratégie d'élimination parmi les multiples interactions du même ordre (ordre 1 concernant notre exemple), toujours en respectant la règle ci dessus :

a) Une première approche consiste à confronter le modèle complet incluant toutes les interactions d'ordre superieur

$Y = f(X_1, X_2, X_3, X_1 * X_2, X_1 * X_3, X_2 * X_3)$ avec celles ou elles sont absentes c-a-d $Y = f(X_1, X_2, X_3)$ en utilisant le test de wald .Si on accepte H_0, les coefficients associées aux termes d'interactions sont tous nuls , nous pouvons les supprimer en bloc . Dans le cas contraire, i e ,rejet de H_0,la situation se complique . Nous devons comparer le modèle complet avec un modèle n'incluant que certaines interactions .

Admettons que nous souhaitons évaluer le terme $X_2 * X_3$

b) Nous pouvons la confronter avec la regression .

$Y = f(X_1, X_2, X_3, X_1 * X_2, X_1 * X_3)$. Ce modèle est toujours HBF si l'on se réfère à la définition ci dessus . Apres il faut savoir interpréter correctement les coefficients .

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

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

Calcul de l'odds-ratio en présence d'interaction .

Le calcul de l'odds-ratio d'une variable dépend des valeurs des autres variables lorsqu'il ya des termes d'interaction dans la regression. Si l'estimation ponctuelle est assez simple à produire,

il en est tout autrement en ce qui concerne l'estimation par intervalle de confiance. Nous devons tenir compte des variances et covariances des coefficients pour obtenir la variance du log-odds .

7.1.2 Estimation ponctuelle

Prenons un exemple à deux variables $\{X_1, X_2\}$ pour fixer les idées le logit s'exprime de la manière suivante :

$$logit = a_0 + a_1X_1 + a_2X_2 + a_3X_1X_2$$

X_2 est binaire , nous souhaitons obtenir son odds-ratio , le logit pour $X_2 = 0$ s'écrit

$$logit\ (X_2 = 0) = a_0 + a_1X_1$$

pour $X_2 = 1$, il devient

$$logit\ (X_2 = 1) = a_0 + a_1X_1 + a_2 + a_3X_1$$

L'écart entre les logit , le log odds-ratio , est obtenu par différenciation

$$\begin{aligned} \Delta Logit(X_2) &= Logit(X_2 = 1) - Logit(X_2 = 0) \\ &= a_2 + a_3X_1 \end{aligned}$$

Ainsi, l'odds-ratio $OR(X_3) = e^{\Delta \log it(X_3)}$ dépend à la fois des coefficients a_2, a_3, mais aussi de la valeur de X_1. Nous ne pouvons plus nous contenter d'analyser uniquement le coefficient a_2 associé à la variable individuelle.

7.1.3 Interprétation des coefficients de la régression en présence d'interaction

L'obtention des odds-ratio est difficile pour les modèles avec interaction. Ils sont plus ou moins liés avec les coefficients de la régression, nous devons tenir compte des valeurs prises par les autres variables explicatives.

Dans le cas de régression à deux variables cependant, nous pouvons déduire les log odds-ratio à partir des coefficients.

Pour mieux nous expliquer, nous ferons tenir un rôle différent aux variables explicatives : l'une (X) sera le facteur de risque dont on veut étudier l'impact sur la variable dépendante,

généralement il s'agit d'une variable sur laquelle nous pouvons raisonnablement influer

(ex. fumer ou pas, le poids, etc.) ; l'autre (Z) sera la variable modératrice qui peut masquer ou exacerber cet impact, il s'agit le plus souvent d'une variable sur laquelle nous n'avons pas réellement prise (ex. l'âge, le sexe, etc.).

Exemple : Deux variables explicatives binaires **(**Rakotomalala 13)

Si on veut étudier par exemple le problème du ronflement (variable endogène Y) en fonction de : $X = tabac$ et $Z = genre$. On souhaite savoir si le tabac a une influence, et si elle est différente selon que l'on est un homme($Z = 1$) ou une femme ($Z = 0$: groupe de référence)

On a sur le tableau suivant les coefficients de la régression :

Tableau 11: Coefficients de la regression

coef	$\hat{a}$	p-value
$\hat{a}_0$	-2.1972	
$\hat{a}_x$	1.1856	0.3252
$\hat{a}_z$	1.5863	0.1463
$\hat{a}_{xz}$	-0.4794	0.7151

$ln[OR(femme)] = 1.1856 = \hat{a}_x$, le log odds-ratio associé au facteur de risque X dans le groupe de

référence correspond au coefficient du facteur de risque $\hat{a}_x$.

$ln[OR(homme)] = 0.7032 = \hat{a}_x + \hat{a}_{xz}$, le log odds-ratio dans le groupe des hommes correspond à la somme des coefficients associés au facteur de risque et au terme d'interaction.

7.1.4 Sélection des variables

On montre le principe de cette sélection sur l'exemple illustratif suivant. (Nakache 12)

On compare des modèles emboîtés en utilisant le test du rapport de vraisemblance . Les procédures en pas à pas descendant et ascendant sont déroulées à partir d'un exemple illustratif concernant trois variables explicatives binaires : X_1, X_2 et X_3et pour lesquelles on a considéré les 19 modèles qui suivent

Tableau 12 : Modèles

Modèle	Effets principaux	Interaction d'ordre 1	Interaction	log-vraisemblance
			d'ordre 2	$L_i = log\,[V(M_i)]$
M_1	$X_1X_2X_3$	$X_1 * X_2 \quad X_1 * X_3 \quad X_2 * X_3$	$X_1 * X_2 * X_3$	-350.7
M_2	$X_1X_2X_3$	$X_1 * X_2 \quad X_1 * X_3 \quad X_2 * X_3$		-350.8
M_3	$X_1X_2X_3$	$X_1 * X_2 \quad X_1 * X_3$		-353.2
M_4	$X_1X_2X_3$	$X_1 * X_2 \quad X_1 * X_3$		-355.9
M_5	$X_1X_2X_3$	$X_1 * X_3 \quad X_1 * X_3$		-351.1
M_6	$X_1X_2X_3$	$X_1 * X_2$		-357.7
M_7	$X_1X_2X_3$	$X_1 * X_3$		-352.3
M_8	$X_1X_2X_3$	$X_2 * X_3$		-356.0
M_9	X_1X_2	$X_1 * X_2$		-357.8
M_{10}	X_1X_3	$X_1 * X_3$		-369.6
M_{11}	X_2X_3	$X_2 * X_3$		-356.0
M_{12}	$X_1X_2X_3$			-357.7
M_{13}	X_1X_3			-372.2
M_{14}	X_1X_2			-357.1
M_{15}	X_2X_3			-357.5
M_{16}	X_1			-371.7
M_{17}	X_3			-372.0
M_{18}	X_2			-357.9
M_{19}	constante			-372.3

Procédure pas à pas descendant

On part du modèle saturé (modèle M1) en respectant le principe de hiérarchie (une interaction ou un effet principal ne peut être ôté du modèle que si les interactions d'ordre supérieur faisant intervenir cette interaction ou cette interaction ou cet effet principal sont déjà ôtées (du modèles).

On pose pour simplifier : $L_i = Log[V(M_i)]$ pour le modèle M_i .

Pas n°1 : Comparaison des modèles M1 et M2 en utilisant le test du rapport de vraisemblances :

$G = -2(L_2 - L_1) = 0,2$ étant non significatif, on retire l'interaction d'ordre 2 $(X_1 * X_2 * X_3)$et on garde donc le modèle M2 .

Pas n°2 : Comparaison M_2 à $M_3 \Rightarrow G = 4,8$ significatif

M_2 à $M_4 \Rightarrow G = 10.2$ significatif

M_2 à $M_5 \Rightarrow G = 0.5$ non significatif

on retire l'interaction d'ordre 1 $(X_1 * X_2)$ et on garde donc le modèle M_5

Pas n°3 : comparaisons M_5 à $M_7 \Rightarrow G = 2,4$ non significatif

M_5 à $M_8 \Rightarrow G = 9.8$ significatif

on retire l'interaction d'ordre 1 $(X_2 * X_3)$ et on garde donc le modèle M_7.

Pas n°4 : comparaisons $M_7\,(X_1, X_2, X_3, X_1 * X_3)$ à M_{10} $(X_1, X_3, X_1 * X_3) \Rightarrow G = 32.6$ significatif

$M_7\,(X_1, X_2, X_3, X_1 * X_3)$ à M_{12} $(X_1, X_2, X_3) \Rightarrow G = 8.8$ significatif

Conclusion: On retient le modèle $(X_1, X_2, X_3, X_1 * X_3)$.

Procédure pas à pas ascendant

On part du modèle sans covariable (variable explicative) en respectant le principe de hiérachie dans la sélection des variable en pas à pas ascendant

modèle $M_{19} \Rightarrow L_{19} = -372.3$

Pas n°1 : comparaisons M_{19} à $M_{18} \Rightarrow G = 28.8$ significatif

M_{19} à $M_{17} \Rightarrow G = 0.6$ non significatif

M_{19} à $M_{16} \Rightarrow G = 1.2$ non significatif

on introduit la variable X_2 et on garde donc le modèle $M_{18}(L_{18} = -357.9)$.

Pas n°2 : comaraisons M_{18} à $M_{15} \Rightarrow G = 0.8$ non significatif

M_{18} à $M_{14} \Rightarrow G = 1.6$ non significatif

Conclusion: On arrête la procédure pas à pas et on ne retient donc que la variable X_2 dans le modèle.

7.1.5 Tests de concordances

Des tests non paramétriques permettent de mesurer les capacités prévisionnelles du modèle retenu. . Le D de Somers, le Gamma de Goodman-Kruskal et Tau-a de Kendall sont des résumés du Tableau des paires concordantes et discordantes fournis en fin de traitement des données. Ces mesures sont, en général, comprises entre 0 et 1, où les valeurs les plus élevées indiquent que le modèle a de meilleures capacités de prévision.

Dï£¡finition 12 *Le D de Somers : est une mesure ordinale de l'association introduite par Somers (1962) . Elle peut être définie par le terme de kendall τ_a donnée par (Roger 14)*

$$\tau(X,Y) = E\left[(sign(X_i - X_j)\, sign(Y_i - Y_j))\right]$$

ou E est l'espérance mathématique.

D'ou la mesure de D Somers est définie par

$$D(Y/X) = \tau(X,Y)/\tau(X,X)$$

Dï£¡finition 13 *Le Gamma de Goodman - Kruskal (Newson . 14)*

L'estimation de Gamma dépend de deux quantités :

N_S : le nombre de paires concordantes

N_d : le nombre de paires discordantes

On a

$$G = \frac{N_S - N_d}{N_S + N_d}$$

Dï£¡finition 14 *Tau A de kendall (Mouchiroud 7) :Soit $(x_1, Y_1), (x_2, Y_2), ... (x_n, Y_n)$ un ensemble d'observations de variables jointes X et Y tel que les valeurs des (X_i) et (Y_i) sont uniques . les paires d'observations $(X_i, Y_i$ et (x_j, Y_j) sont dites concordantes si $X_i < x_j$ et $Y_i < Y_j$ ou si $X_i > x_j$ et $Y_i > Y_j$. Elle sont dites discordantes si $X_i < x_j$ et $Y_i > Y_j$ ou si $X_i > x_j$ et $Y_i < Y_j$. Dans le cas ou $X_i = x_j$ ou $Y_i = Y_j$, la paire n'est ni concordante ni discordante .Le tau A de Kendall est alors défini comme suit :*

Dï£¡finition 15

$$\tau = \frac{\textit{(nombre de paires concordantes)-(nombre de paire discordantes)}}{n^{\frac{1}{2}}.n.(n-1)}$$

Chapitre 8

Etude de Cas :Estimation des risques d'hémopathies liées au diabète de type 2 chez la femme

Introduction :

On sait que le diabète de type 2, est liée à un risque accru d'anémie, d'anomalie de la formule sanguine, et de cancers hématologiques comme la leucémie ou les lymphomes. Mais les raisons sont mal connues.

Ces anomalies concernent la taille, la forme et la concentration en hémoglobine des hématies. Ainsi que les anomalies de leucocytes concernant les polynucléaires, les lymphocytes, et, les anomalies des Plaquettes Sanguines. D'autres anomalies peuvent aussi être observées dans le sang. Dans ce travail nous tentons d'établir un lien entre les variations quantitative et qualitative de la formule sanguine qui exposent au diabète de type2.

8.1 Matériel et Méthode :

Cette étude a porté sur un échantillon de 1168 (Femmes) sujets dont 681 diabétiques et 487 témoins.

Il s'agit d'une étude cas témoins. Des analyses statistiques sont effectuées pour vérifier l'impact, la glycémie, sur le risque de développer des hémopathies liées au diabète de type2. Une étude logistique a été faite afin de déterminer un modèle prédictif du risque

de survenue d'hémopathie au cours du diabète de type 2 à l'aide des facteurs mesurés.

8.1.1 Résultats

Tableau 13 :Résultats de l'Étude du Modèle de Régression Logistique Simple

Prédicteur	Coeff	Z	P value	OR	IC min	IC max
Constante	-1,64987	-4,23	0			
c1nsg	0,465608	2,05	0,04	**1,59**	1,02	2,48
Héredité	0,726038	2,67	0,008	**2,07**	1,21	3,52
glyc	4,75434	8,68	0	**116,09**	39,7	339,48
M1n	-0,629001	-2,25	0,024	**0,53**	0,31	0,92
Basc	2,78957	6,63	0	**16,27**	7,13	37,1
HbFemme						
1	-0,0671517	-0,21	0,835	**0,94**	0,5	1,76
2	-2,34101	-3,57	0	**0,1**	0,03	0,35
VGM	-0,706665	-2,47	0,014	**0,49**	0,28	0,86
PL	-0,661871	-3,39	0,001	**0,52**	0,35	0,76
VPM	-2,25515	-2,75	0,006	**0,1**	0,02	0,52
VS2Femme	0,93847	2,61	0,009	**2,56**	1,26	5,17

Pour les monocytes leur élévation diminue le risque de survenue du diabète de 0.53 résultats significative

Pour les basophiles leur variation expose à un risque de 16.27 d'avoir le diabète de type2(significative)

Pour l'hémoglobine quelque soit la variation elle diminue le risque de survenue du diabète

Les variations des VGM diminuent le risque de survenue du diabète de type2 de 0.49 (significative)

Pour les plaquettes leur variation protège d'un coefficient de 0.52 (significative)

Pour les VPM leur variation protège d'un coefficient de 0.1 (significative)

Pour l'augmentation de la VS à 2 heures le risque de survenue du diabète de type2 est multiplié par un coefficient de 2,56 (significative)

8.1.2 Tests pour les termes avec plusieurs degrés de liberté

Tableau 14 : Tests pour les termes à plusieurs modalités

Terme	Khi deux	Dl	P-value
HbFemme	12,7965	2	0.002

Le Tableau 14 donne un résumé des facteurs à plus de deux modalités retenues effectivement dans le modèle logistique. On constate que tous les tests de nullité des coefficients associés individuellement au facteur Hbfemme (1) n'est pas significatif alors que le facteur Hbfemme (2) est très significatifs ($p<0.01$). On note aussi que le test de tous les coefficients simultanément égaux à zéro est très significatif (p-value <0.01). D'ou la décision de retenir la covariable Hbfemme.

8.1.3 Ajustement au modèle logistique

Tableau 15 :Tests d'Adéquation de l'Ajustement

Méthode	Khi deux	Dl	P-value
Pearson	206.62	239	0.936
Deviance	154.593	239	0.999
Hosmer-Lemeshow	12.289	8	0.139

Le Tableau 15 justifie le choix du modèle logistique. En effet aucun test d'adéquation n'est significatif.

8.1.4 Capacités prévisionnelles du modèle logistique

Tableau 16 :Mesures d'Association (entre la Variable Réponse et les Prévisions de Probabilité)

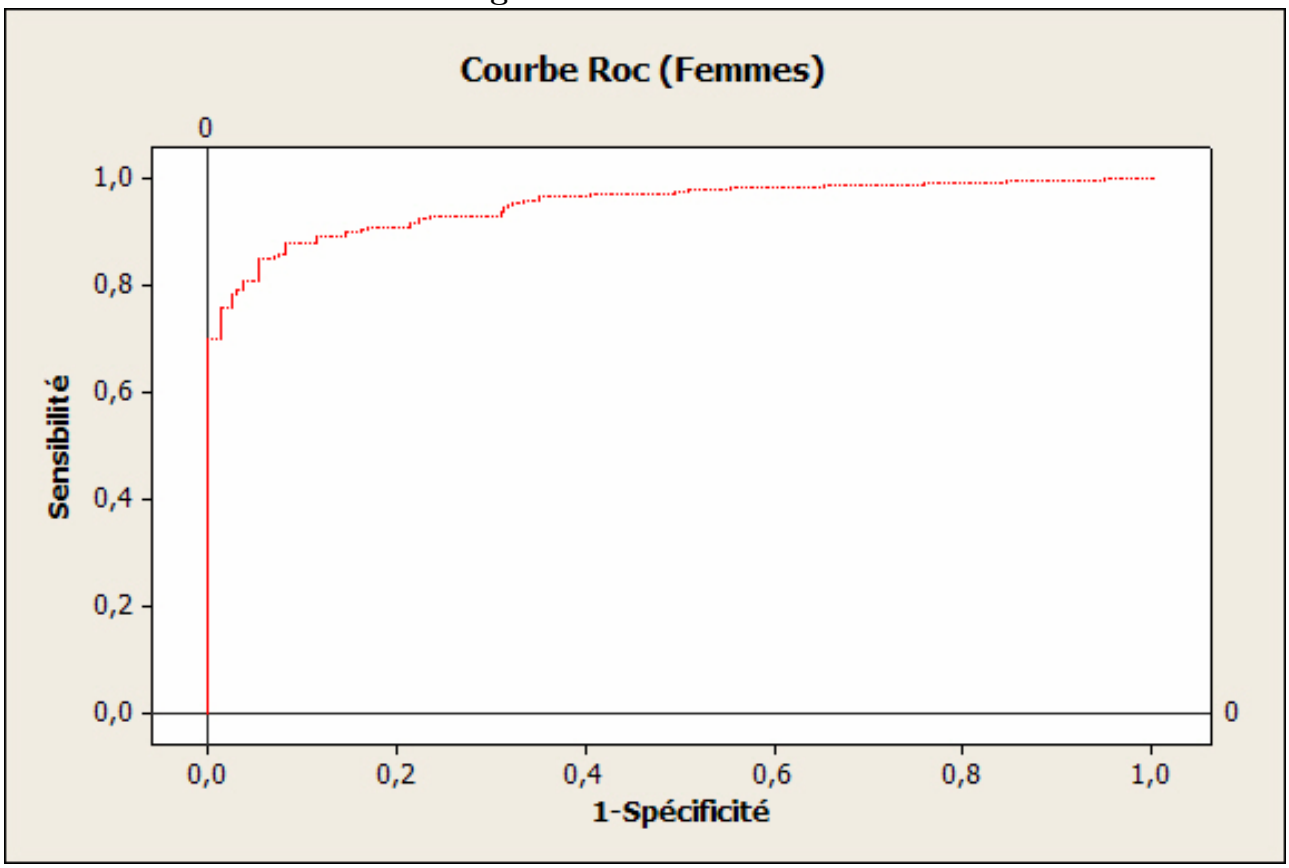
Le Tableau 16 indique les capacités prévisionnelles de ce modèle. On constate un très fort pourcentage de paires concordantes (93.1%). Le D de Somers, le Gamma de Goodman-Kruskal et Tau-a de Kendall sont des résumés du Tableau des paires concordantes et discordantes. Ces mesures sont, en général, comprises entre 0 et 1, où les

Paires	Nombre	Pourcentage	Mesures récapitulatives	
Concordant	77784	93,1	D de Somers	0,87
Discordant	5015	6	Gamma de Goodman-Kruskal	0,88
Ex aequo	746	0,9	Tau a de Kendall	0,42
Total	83545	100		

valeurs les plus élevées indiquent que le modèle a de meilleures capacités de prévision. Dans ce cas, les deux premières mesures valant 0,87 et 0,88, impliquent une très forte capacité de prévision. Le Tau-a de Kendall donne une capacité prévisionnelle relativement bonne.(0.42)

8.1.5 Courbe ROC

Fig 6: Courbe ROC



Notons dans ce cas une **Discrimination exceptionnelle avec un AUC:** =0,9532

Chapitre 9

Conclusion :

Historiquement l'étude des modèles décrivant les modalités prises par une ou plusieurs variables qualitatives date des années - Les travaux les plus marquants de cette époque sont ceux de Berkson consacrés notamment aux modèles dichotomiques simples (modèles logit et probit) . Les premières applications ont alors essentiellement été menées dans le domaine de la biologie , de la sociologie et de la psychologie . Ainsi , ce n'est finalement que récemment , que ces modèles ont été utilisés pour décrire des données économiques .

Nous avons pour notre part résumé l'essentiel des résultats et travaux sur la régression logistique binaire et notamment les travaux de Hurlin C (7),Rakotomalala R (14) ,Nakache , J P et Confais T (12) ..etc.

Evidemment la regression logistique polytomique et conditionnelle , restent des modèles très riches à developper et qu'on n'a pas pu faire dans le cadre de ce travail , vu la consistence de tous les resultats et informations contenus dedans.

Nous avons appliqué ce modèle au cas d'une étude cas témoins. On a pu vérifier l'impact, la glycémie, sur le risque de développer des hémopathies liées au diabète de type2. Une étude logistique a été faite afin de déterminer un modèle prédictif du risque de survenue d'hémopathie au cours du diabète de type 2 à l'aide des facteurs mesurés.

L'étude à permis de mettre en exergue un modèle prédictif à très fortes capacités prévisionnelles . L'AUC étant de **0**,9532 , notre modèle prédictif etant à "Discrimination exceptionnelle".

Nous souhaitons appliquer d'autres traitements similaires , notamment aux modèles de Cox (Modèles qui prennent en compte des données censurées (c'est à dire tenant compte

des temps d'observation individuels).

La bibliographie jointe démontre , par son volume , que les modèles logit et assimilés ont ces dernières années su séduire un nombre grandissant de statisticiens et de chercheurs par la simplicité de ses applications et interprétations.

Bibliographie

[1] Amemiya T.(1981)"Qualitative Response Models : A Survey" , Journal of Economic Litterature , 19(4) , 481-536

[2] Bernard ,P . -M . , " Analyse des tableaux de contingence en épidémiologie "

[3] Cornfield , (1951) propriétés mathématiques du risque (odds - ratio)

[4] Gourieroux C (1989) "Asympototic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models

[5] Greene W.H (1997) "Econometric Analysis " Londres , Prentice Hall

[6] Hosmer DW ,Lemeshow S ,Applid logistic regression 2000

[7] Hurlin Christophe 2002 (Econométrie des variables qualitatives)

[8] Jacquot A (Les modèles économétriques , logit -probit -tobit)

[9] Klein R . W et Spady R .W et Spady R.H (1993) "An Efficient Semi Parametric stimator for Binary Response Models " , Econometrica ,61 ,387 - 421

[10] Leblanc D (l'économétrie et l'étude des comportements)

[11] Mouchiroud D (2003) test d'ajustement

[12] Nakache , J P. Confais T , statistique Explicative Appliquée , Technip , 2003

[13] Newson Roger B , 2014 (interpretation of somers , D under four simple models)

[14] Rakotomalala Ricco (Pratique de la regression logistique) fev 2014

[15] Revue "prescrire diabète de type 2 n°278 déc 2006 p. 845 et 846

[16] Le guide médecin élaboré par la haute autorité de santé (HAS) : le diabète

Résumé

Dans ce mémoire on étudie les modèles dichotomiques simples, probit et logit . Nous présenterons les principaux modèles , puis dans une seconde section nous nous intéresserons à l'estimation des paramètres de ces modèles , notamment par la méthode du M V. Dans une troisième section , nous étudierons la convergence des estimateurs du M V. Ensuite nous aborderons les tests de spécification des modèles ainsi que les différents problèmes d'inférence . Enfin nous ferons une étude détaillée de cas intitulée : Estimation des risques d'hémopathies liées au diabète de type 2 chez la femme.

Le modèle dichotomique probit et logit admettent pour variable expliquée la probabilité

pi=p(yi=1/xi)=F(xib). Où la fonction F(.) désigne une fonction de répartition. Toutefois on utilise généralement deux types de fonctions de répartition : Fonction de répartition de la loi logistique ou Fonction de répartition de la loi normale centrée et réduite.

On cherche à estimer les composantes du vecteur b. La méthode la plus utilisée est la méthode du M V, c'est-à-dire de résoudre l'équation G(b)=0 ou G est le gradient de la log -vraisemblance . La méthode qui est recommandée solutionner ce problème dans un modèle dichotomique univarié est la méthode d'optimisation de NEWTON RAPHSON

On cherche après à établir les propriétés asymptotiques des estimateurs du M V. Sous certaines conditions, l'estimateur du M V est convergent et suit asymptotiquement une loi normale.

Après avoir construit un modèle de prédiction, Nous évaluons son efficacité de différentes manières : Par la matrice de confusion , test de Hosmer-Lemeshow , courbe de roc…etc.

Nous présenterons aussi les tests d'hypothèses sur les coefficients, puis nous envisagerons les principaux tests de spécification sur les modèles dichotomiques.

Mots clés : Regression logistique, Probit , Logit, Estimateur du maximum de vraisemblance (M V), Odds, Odd ratio

Summary

Dichotomous models simple, probit and logit is investigated in this thesis. We will present the main models, then in a second section we will look at the estimation of the parameters of these models, especially by the method of M L In a third section, we will explore the convergence of estimators of M L Then we'll discuss testing specification of the models as well as the different inference problems.

Finally we will make a detailed case study entitled: Estimation of the risk of haematological diseases related to of type 2 diabetes in women.

The dichotomous model probit and logit admit for explained variable probability

pi = p (yi = 1/xi) = F (xib). Where the function F (.) refers to a distribution function. However generally two types of distribution functions are used: distribution function of the logistic distribution or distribution function of the normal distribution centered and reduced.

We try to estimate the components of the vector b. The most used method is the method of M L, i.e. to solve the equation G (b) = 0 or G is the gradient of the log - likelihood. The method that is recommended solve this problem in a univariate dichotomous model is the NEWTON RAPHSON optimization method

After seeking to identify the asymptotic properties of the estimators of M L Under certain conditions, the estimator of M L is convergent and asymptotically follows a normal distribution.

After you construct a predictive model, we evaluate its effectiveness in different ways: by the confusion matrix, Hosmer-Lemeshow test, curved ROC... etc.

We will also present the tests of hypotheses on the coefficients, and then we will consider the main tests of specification on dichotomous models.

Keywords: Regression logistics, Probit, Logit, Maximum likelihood estimator (M L), Odds, Odd ratio

ملخص

نماذج ثنائية التفرع بسيطة، بروبيت واللوغاريتمية هو التحقيق في هذه الأطروحة. وسوف نقدم النماذج الرئيسية، ثم في قسم ثاني سوف ننظر في تقدير معالم هذه النماذج، لا سيما بواسطة الأسلوب الخامس م في مقطع ثالث، سوف نستكشف تقارب المقدرات الخامس م ثم سوف نناقش مواصفات اختبار النماذج، فضلا عن مشاكل استدلال مختلفة.

وأخيراً سوف نبذل دراسة حالة مفصلة بعنوان: تتصل بتقدير لخطر الإصابة بالأمراض المتصلة بالدم لمرض السكري من النوع 2 لدى النساء.اعترف اللوغاريتمية والتفرع الطراز بروبيت لشرح الاحتمال متغير

pI= p (يي = 1/الحادي عشر) = F (إكسيب). حيث الدالة F (..) يشير إلى دالة توزيع. ولكن عموما يستخدم نوعان من المهام التوزيع: دالة التوزيع التوزيع اللوجيستي أو دالة التوزيع للتوزيع العادي توسيط وخفض.

ونحن في محاولة لتقدير مكونات ب متجه. الأسلوب الأكثر استخداماً هو الأسلوب الخامس م، أي أن حل المعادلة ز (ب) = 0 أو ز هو تدرج السجل-احتمال. الأسلوب الذي يوصي بحل هذه المشكلة في وحيد المتغير التفرع النموذج هو الأسلوب الأمثل RAPHSON نيوتنبعد السعي لتحديد خصائص أدوات التقدير الخامس م مقارب تحت ظروف معينة، مقدر م الخامس متقاربة ومقارب يتبع توزيع الطبيعي.بعد أن تقوم بإنشاء نموذج تنبؤي، نقوم بتقييم فعاليتها بطرق مختلفة: بالمصفوفة الارتباك، اختبار هوسمر استخدم، منحنى ROC.. إلخ.سوف نقدم أيضا اختبارات الفرضيات في المعاملات ومن ثم سننظر في الاختبارات الرئيسية من مواصفات في نماذج ثنائية التفرع.

الكلمات الرئيسية: الانحدار اللوجستي، والاحتمالية، واللوغاريتمية، الحد الأقصى لاحتمال (م ت)، الصعاب، ونسبة غريبة مقدر