PRATIQUE DE LA REGRESSION LOGISTIQUE

Pierre-Louis GONZALEZ

Quelques recommandations concernant la mise en œuvre d'une régression logistique

On peut distinguer deux usages de tels modèles : **expliquer la variance du phénomène afin de le prévoir au mieux**,

ou de façon plus ambitieuse souhaiter **utiliser le modèle pour dégager des processus explicatifs, des relations causales.**

Les difficultés liées à la colinéarité ou la non exogénéité prennent beaucoup plus d'importance dans cette optique.

1 Variables explicatives , colonnes explicatives

Variables explicatives:

- Les variables quantitatives sont représentées chacune par une colonne de données.

- Les variables qualitatives sont elles représentées par plusieurs colonnes: Ainsi la CSP d'un individu est caractérisée par plusieurs colonnes explicatives : « être cadre supérieur », « être agriculteur »

- Le fait de retenir ou non une variable explicative doit être réglé à l'aide de tests prenant en compte l'ensemble des colonnes la définissant. Il sera en particulier délicat d'utiliser les procédures pas à pas proposées par les logiciels si l'on a défini soi même les indicatrices des modalités d'une variable qualitative. **En revanche l'utilisation de l'instruction Class dans la procédure LOGISTIC de SAS permet de résoudre correctement ce problème.**

2 Représentation d'une variable explicative

- Dans le cas d'une **variable quantitative**, diverses représentations sont possibles suivant que l'on souhaite ou non la faire intervenir de façon linéaire dans le modèle. S'il est clair qu'une relation non linéaire existe avec la variable de réponse (l'âge d'un automobiliste n'explique pas de façon linéaire ses capacités de bon conducteur pour un assureur) on procède à un découpage en classes

- Dans le cas d'une **variable qualitative**, on introduit autant de colonnes indicatrices qu'il y a de modalités. Il est toujours possible en fonction d'analyses préalables ou en fonction des résultats d'une première modélisation de regrouper certaines modalités. Les colonnes introduites pour représenter une variable qualitative, **ne sont pas indépendantes** puisque leur somme vaut 1 quel que soit l'individu i.

- Le remède consiste à imposer une contrainte sur les coefficients associés aux modalités, par exemple la somme des coefficients est égale à 0. (C'est l'option par défaut de la procédure LOGISTIC)

  On peut aussi contraindre une des modalités à avoir un coefficient égal à 0. On considère alors qu'elle représente une situation de référence, par rapport à laquelle on mesure des déviations, des différences. (Option PARAM=GLM dans la procédure LOGISTIC)

- Mathématiquement, le choix de cette situation de référence n'a généralement que peu d'importance. Un changement a pour effets une translation des coefficients (le profil qu'ils dessinent reste inchangé) et une modification des écarts-types des estimateurs. **Bien évidemment le nombre de coefficients significativement différents de zéro peut changer. Ceci montre clairement que pour juger de l'apport d'une variable explicative qualitative on ne peut pas utiliser le nombre de coefficients significativement différents de zéro.**

- **Comme en analyse de la variance, on ne peut se contenter de modèles purement additifs.**

- Considérons par exemple une étude concernant le fait de faire de la couture (variable de réponse Y). Les variables explicatives dont on dispose sont au nombre de deux : le sexe et l'âge découpé en 3 classes : moins de 40 ans, 40 à 60 ans et plus de 60 ans.

  Si l'on écrit le modèle sous la forme :

  $$Y = \beta_0 + \beta_1 I_{femme} + \beta_2 I_{âge<40} + \beta_3 I_{âge>60} + \varepsilon$$

  Cela revient à supposer que l'on a des effets additifs.

Or on sait que les hommes font très rarement de la couture, quel que soit leur âge.

L'étude est donc réalisable en modélisant l'interaction entre l'âge et le sexe à l'aide du modèle défini par :

$$Y = \beta_0 + \beta_1 I_{\text{âge}<40} I_{\text{homme}} + \beta_2 I_{\text{âge}>60} I_{\text{homme}} + \beta_3 I_{\text{âge}<40} I_{\text{femme}} + \beta_4 I_{\text{âge}40\text{-}60} I_{\text{femme}} + \beta_5 I_{\text{âge}>60} I_{\text{femme}} + \varepsilon$$

3 Problèmes de pondération

- Quelle que soit l'étude considérée, se pose un problème concernant la pondération des observations. Deux questions sont à envisager :

- **1. Que doit-on faire si une des modalités de la variable de réponse est sur ou sous-représentée dans l'échantillon dont on dispose ?**

- **2. Que doit on faire si certaines modalités des variables explicatives sont sur ou sous-représentées dans l'échantillon dont on dispose ?**

- Une propriété mathématique du modèle logit permet de répondre à la première question :

- **Le modèle logit (mais pas le modèle probit) possède la propriété que les estimateurs des paramètres de pente (c'est à dire, des paramètres relatifs aux variables explicatives) sont invariants à une sur-représentation des individus ayant la caractéristique définie par la variable de réponse. Seule la constante du modèle est affectée par la sur-représentation .**

- On rencontre souvent ce phénomène dans les études médicales. Lorsque l'on s'intéresse à une maladie rare, on travaille en général avec un fort sur-échantillonnage des sujets malades.

- **La propriété citée indique donc que si le modèle logistique est vrai dans la population, peu importe que l'on sur-représente dans l'échantillon l'une des modalités de la variable de réponse.**

Plus précisément, on a le théorème suivant :

Théorème :

Soit D *l'événement {un membre de la population est malade}.*

Soit $\pi(x) = P(D/x)$ *la probabilité d'être malade conditionnellement à x dans la population.*

On suppose que $\pi(x)$ *suit un modèle logistique dans la population, i.e.* $\pi(x) = \dfrac{1}{1+\exp(-\alpha-\beta x)}$

Soit S *l'événement {l'individu est échantillonné}*

Soit $\rho_0 = P(S/D)$ *et* $\rho_1 = P(S/D^c)$

Alors $P(D/S,x)$ *suit également une loi logistique, avec le même paramètre d'effet* β *mais avec une ordonnée à l'origine* $\alpha^* = \alpha + \log(\dfrac{\rho_0}{\rho_1})$

En effet, par la formule de Bayes $\mu(x) = P(D/S,x) = \dfrac{P(S/D,x)P(D/x)}{P(S/x)}$

Mais $P(S/D,x) = \rho_0, P(D/x) = \pi(x),$

$P(S/x) = P(S/D,x)P(D/x) + P(S/D^c,x)P(D^c/x) = \rho_0\pi(x) + \rho_1(1-\pi(x))$

D'où $\mu(x) = \dfrac{\rho_0\pi(x)}{\rho_0\pi(x)+\rho_1(1-\pi(x))} = \dfrac{\rho_0\exp(\alpha+\beta x)}{\rho_0\exp(\alpha+\beta x)+\rho_1} = \dfrac{\frac{\rho_0}{\rho_1}\exp(\alpha+\beta x)}{\frac{\rho_0}{\rho_1}\exp(\alpha+\beta x)+1}$

Ce qui achève la démonstration.

- Il est par contre beaucoup plus difficile de répondre à la deuxième question. **En effet les estimateurs du maximum de vraisemblance pondérés et non pondérés sont différents car la vraisemblance dépend des effectifs des cases.**

- Rappelons enfin qu'il est toujours préférable d'utiliser des **pondérations normalisées**. Ainsi les résultats obtenus concernant les tests sont utilisables.

4 Le problème de la non convergence

- Ceci se produit chaque fois que pour une strate aucun individu, ou tous les individus sont concernés par la pratique étudiée. Une **analyse préalable consistant à croiser les deux modalités de la variable de réponse avec toutes les modalités des variables explicatives** permet de détecter une telle situation.

- On dispose alors de deux possibilités pour remédier à ce problème :

  - soit exclure la sous-population concernée. On travaille sur un sous-échantillon.

  - soit regrouper cette sous-population avec une strate voisine, de sorte que la fréquence de la pratique cesse d'être nulle ou égale à 100%. On conserve alors l'échantillon complet.

- On peut néanmoins après avoir résolu les problèmes de non convergence disposer d'estimations fragiles.

- C'est le cas lorsque l'une des modalités de la variable de réponse est très peu représentée dans certaines des modalités des variables explicatives.

- Ceci se détecte préalablement selon la méthode déjà exposée et au niveau du modèle par la **présence d'écarts-types des estimateurs très élevés.**

- **Autres cas de non convergence**

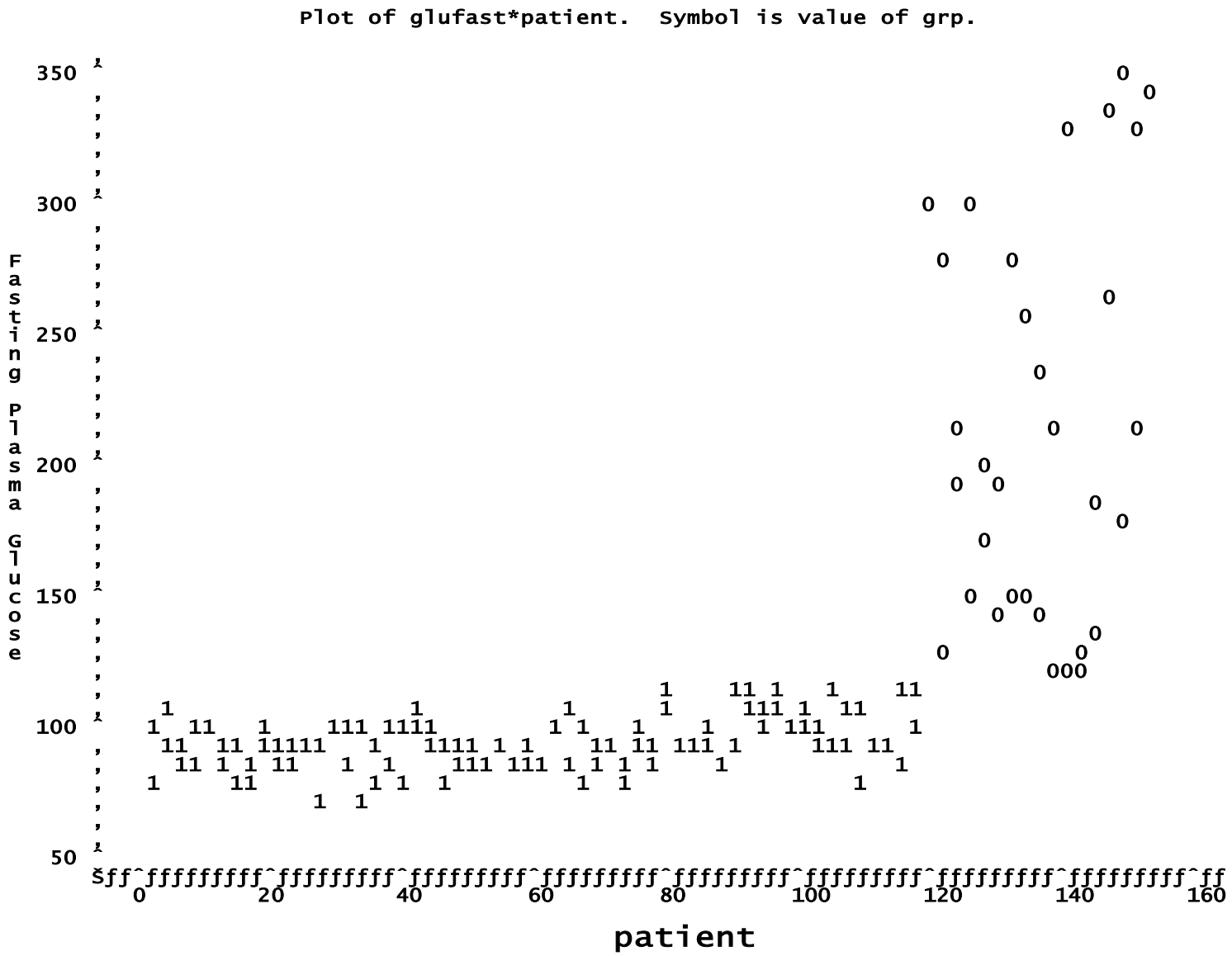
- **séparation complète des deux groupes** : l'estimateur au sens du maximum de vraisemblance n'existe pas. Il existe un vecteur de pseudo-estimations qui affecte correctement tous les individus à leur groupe. La configuration des données conduit à des estimations infinies, sans unicité de l'estimateur. Lors des itérations, la probabilité prévue pour chaque individu d'appartenir à son groupe réel tend rapidement vers 1 et le log de la vraisemblance tend vers 0.

- **Séparation quasi-complète des deux groupes :**

  Comme précédemment, l'estimateur au sens du maximum de vraisemblance n'existe pas. Il existe un vecteur de pseudo-estimations qui affecte correctement la plupart des individus à leur groupe. La configuration des données conduit aussi à des estimations infinies sans unicité de l'estimateur. Par contre, lors des itérations le log de la vraisemblance ne diminue pas et ne tend pas vers 0 comme dans le cas de la séparation complète.

- Notons que des cas de séparation complète ou quasi-complète risquent de se rencontrer dans le cas de petits échantillons.

- Remarquons que dans les deux cas particuliers évoqués, l'analyse discriminante est performante.


Plot of glufast*patient. Symbol is value of grp.
Fasting Plasma Glucose
350
300
250
200
150
100
50
0
20
40
60
80
100
120
140
160
patient

Séparation complète des données

The LOGISTIC Procedure

Model Information

Data Set	WORK.DIABET3
Response Variable	grp
Number of Response Levels	2
Number of Observations	145
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	grp	Total Frequency
1	0	33
2	1	112

Model Convergence Status

Complete separation of data points detected.

WARNING: The maximum likelihood estimate does not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

- **Références bibliographiques**

- Agresti, A., ***An Introduction to Categorical Data Analysis***, Wiley, New York, 1996.
- Agresti, A.,***Categorical Data Analysis***, Wiley, New York, 1990.
- Andersen E.B., ***The Statistical Analysis of Categorical Data***, Springer verlag, Berlin, 1991.
- Chap, T.L., ***Applied Categorical Data Analysis,*** Wiley, New York, 1998.
- Droesbeke J.-J., Lejeune M., Saporta G., Éditeurs, ***Modèles statistiques pour données qualitatives***, Technip, Paris 2005

- Fahrmeir, L., & Tutz, G., ***Multivariate Statistical Modeling Based on Generalized Linear Models***, Springer verlag, New York, 1994

- Gourieroux, C., ***Econométrie des Variables Qualitatives***, Economica, Paris 1989

- Hosmer, D., Lemeshow, S., ***Applied logistic regression***, second edition, Wiley, New York 2000

- Jobson, J.D., ***Applied Multivariate Data Analysis ; Volume II : Categorical and Multivariate Methods***, Springer verlag, New York, 1992

- Leblanc, D., Lollivier, S., Marpsat, M. ,& Verger, D., ***L'Econométrie et l'étude des comportements***, Série des documents de travail « Méthodologie statistique » N° 0001, INSEE, Paris, 2000.

- Kleinbaum, D.G., ***Logistic Regression : A Self-Learning text***, Springer Verlag, New York, 1994.

- Kleinbaum, D.G., Kupper L.L., Keith, E.M., & Nizam, A., ***Applied Regression Analysis and Other Multivariable Methods,*** Duxbury Press, Pacific Grove, CA, 1998

- Lebart, L., Morineau A., & Piron M., ***Statistique exploratoire multidimensionnelle***, Dunod, Paris, 1995

- Lloyd, C.J., ***Statistical Analysis of Categorical Data***, Wiley, New York, 1999.

- Nakache J.P., Confais J., ***Statistique explicative appliquée***, Dunod Paris 2003

- Santner, T.J., & Duffy, D.E., ***The Statistical Analysis of Discrete Data***, Springer verlag, New York, 1989

- SAS, ***Logistic Regression Examples Using the SAS System***, SAS Institute Inc, Cary, N.C., 1995

- Stokes, M.E., Davis, C.S., & Koch, G.G., ***Categorical Data Analysis using the SAS System***, SAS Institute Inc, Cary, NC, 2000.

- Tenenhaus, M., La régression logistique, ***MAD*** numéro3 ,p.21 à 39, 1992

- Thomas A., ***Économétrie des variables qualitatives***, Dunod Paris 2000

- Tuffery, S., ***Data mining et statistique décisionnelle*** Technip Paris 2012