

# REVUE DE STATISTIQUE APPLIQUÉE

C. CANS

C. LAVERGNE

## **De la régression logistique vers un modèle additif généralisé : un exemple d'application**

*Revue de statistique appliquée*, tome 43, n° 2 (1995), p. 77-90

[http://www.numdam.org/item?id=RSA\\_1995\\_\\_43\\_2\\_77\\_0](http://www.numdam.org/item?id=RSA_1995__43_2_77_0)

© Société française de statistique, 1995, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## DE LA RÉGRESSION LOGISTIQUE VERS UN MODÈLE ADDITIF GÉNÉRALISÉ : UN EXEMPLE D'APPLICATION

C. Cans (1), C. Lavergne (2)

(1) *Laboratoire de Génétique, Faculté de Médecine de Grenoble*  
38706 La Tronche Cedex France

(2) *IMAG - Laboratoire de Modélisation et Calcul*  
BP 53, 38041 Grenoble Cedex 9 France

### RÉSUMÉ

Nous présentons un problème de modélisation en régression logistique sur un exemple médical : comment expliquer la survenue d'un enfant malformé chez des parents «porteurs sains» d'une anomalie chromosomique de structure, à partir de la connaissance de mesures cytogénétiques? Après une modélisation de type régression logistique «classique» qui s'avère satisfaisante aux yeux des cytogénéticiens, on est amené à chercher si des liens non linéaires seraient plus appropriés. La technique récente des modèles additifs généralisés peut-elle s'avérer efficace dans cette problématique? Nous proposons ici une démarche empirique et exposons le gain obtenu sur la base de données étudiée.

*Mots-clés : Modèles linéaires et modèles additifs généralisés, Régression logistique, Conseil génétique, Translocations réciproques.*

### SUMMARY

On a large data collection of reciprocal translocations, logistic regression has been used to modelize the risk of viable unbalanced offspring. Preliminary results being interesting, the use of Generalized Additive Model is proposed. An improvement of the performances of the logistic model is obtained by taking into account the non linearity between the predictor and the independent covariates. An empirical strategy for its use is described, and applied to our medical data set.

*Keywords : Generalized Linear Model, Generalized Additive Model, Logistic regression, Genetic counseling, Reciprocal translocations.*

### 1. Introduction

Basée sur les techniques de lissage, la théorie des modèles additifs généralisés (*GAM* pour Generalized Additive Model) est introduite en 1986 par Hastie, T. et Tibshirani, R. ([13]) dans le but de pallier à l'hypothèse forte des modèles linéaires et modèles linéaires généralisés (*GLM* pour Generalized Linear Model) : le lien

«linéaire» entre une variable à expliquer et des prédicteurs. Dans un premier temps, on présente brièvement l'idée qui a prévalu à l'introduction du modèle additif avec son interprétation géométrique. On pourra se référer à [14] pour une lecture mathématique exhaustive ainsi que pour la très large bibliographie. Le mot «généralisés» de la terminologie «modèles additifs généralisés» devant être pris dans le même sens que dans la terminologie «modèles linéaires généralisés».

Une stratégie d'utilisation est alors présentée sur un exemple médical dans le cas de la régression logistique. L'objectif poursuivi est de proposer une démarche empirique de la modélisation utilisant les modèles additifs généralisés.

Les données sont du type présence/absence d'anomalie. Dans un premier temps, on utilise la modélisation type régression logistique du modèle linéaire généralisé; adoptant une démarche approfondie mais «classique»: une très large littérature existe sur ce sujet et quelques références sont citées au paragraphe 2.2. On se restreint à un objectif de modélisation: on ne discute pas du choix de la fonction de lien, fixée ici à la fonction de lien canonique (fonction logit); on n'utilise pas non plus de techniques de choix de modèles ([2]), ni celles des modèles graphiques ([20]).

Après l'exposé détaillé de la démarche de modélisation logistique en modèles linéaires généralisés; et après l'obtention d'un modèle issu d'un prédicteur linéaire, on recherche une amélioration de ce modèle grâce à un prédicteur additif mais non linéaire sur ces mêmes variables. Par la forme de la fonction non paramétrique du prédicteur, le modèle additif généralisé suggère la transformation de certaines variables explicatives.

Cet exemple d'application permet de montrer l'amélioration de l'ajustement de la régression logistique, après prise en compte de ces transformations. Si le gain en déviance est significatif, la qualité de ces deux prédicteurs additifs peut être comparée sur des résultats empiriques de classification ou sur une analyse des résidus: c'est ce qui est exposé dans la dernière partie sur l'exemple d'application.

### 1.1 Le modèle additif

Si l'on se restreint au modèle linéaire standard, étant donné une variable réponse  $Y$  et  $p$  régresseurs  $X_1, X_2, \dots, X_p$  observés sur  $n$  individus, on a l'écriture matricielle:

$$Y = X\beta + \varepsilon \quad (1)$$

$X$  est la matrice  $n.p$  des  $X_j$ .

Une possibilité pour pallier à l'hypothèse de linéarité est de proposer le modèle suivant:

$$Y = f(X) + \varepsilon \quad (2)$$

Le modèle additif est un cas particulier de celui-ci: tout en cherchant à supprimer l'hypothèse de linéarité, on conserve l'effet additif de chaque prédicteur. On définit

alors le modèle additif de la façon suivante :

$$E(Y/X) = \sum_{j=1}^p f_j(X_j) \quad (3)$$

où les fonctions  $f_j$  sont des fonctions inconnues. Si ce modèle est vrai, alors pour tout  $k$ ,  $E(Y - \sum_{j \neq k} f_j(X_j)/X_k) = f_k(X_k)$ . La procédure d'estimation de ces fonctions  $f_j$  est donc proposée par l'algorithme suivant (appelé «backfitting algorithm») :

- i) Initialisation :  $f_j^{(0)}$ , pour  $j = 1, \dots, p$
- ii) Pour  $j = 1, \dots, p$ ,

$$f_j^{(q)} = S_j^{(q)}(y - \sum_{k \neq j} f_k^{(q-1)}/X_j)$$

- iii) Arrêt de ii) si chaque fonction  $f_j^{(q)} = f_j^{(q-1)}$

La fonction  $S(y/x)$  représente un lissage d'une réponse  $y$  par rapport au prédicteur  $x$ . A chaque étape, la fonction  $f_j^{(q)}$  est donc estimée par une procédure d'interpolation non paramétrique. Les procédures usuellement utilisées seront des procédures splines ou des procédures d'estimateurs à noyau.

En pratique les  $f_j^{(q)}$  sont des vecteurs de  $\mathbb{R}^n$  ( $f_j^{(q)}(x_{1j}), \dots, f_j^{(q)}(x_{nj})$ ),  $x_{ij}$  étant la valeur prise par la variable  $X_j$  pour l'individu  $i$ .

### 1.2 Interprétation géométrique

On appelle  $\mathcal{H}_j$  l'espace de Hilbert des fonctions mesurables de la variable  $X_j$ , centrées et de carré intégrable, muni du produit scalaire

$$\langle \phi_j^1(X_j), \phi_j^2(X_j) \rangle = E(\phi_j^1(X_j)\phi_j^2(X_j)).$$

$\mathcal{H}_j$  peut être considéré comme un sous espace d'un espace de Hilbert  $\mathcal{H}$  des fonctions mesurables de  $X_1, \dots, X_p$ , centrées et de carrés intégrables. Notons alors  $\mathcal{H}^{add}$ , sous espace de  $\mathcal{H}$ , défini par  $\mathcal{H}^{add} = \mathcal{H}_1 + \dots + \mathcal{H}_p$ . Le problème d'optimisation est la minimisation de  $E(Y - g(X))^2$  avec  $g(X) = \sum_{j=1}^p f_j(X_j)$  élément de  $\mathcal{H}^{add}$ .

Sous l'hypothèse que  $\mathcal{H}^{add}$  est un sous espace fermé de  $\mathcal{H}$ , la solution à ce problème de minimisation existe et est unique (voir [3]).

En fait chaque fonction  $f_j(X_j)$  ne peut être déterminée de façon unique. Posons  $P_j$  le projecteur orthogonal de  $\mathcal{H}$  sur  $\mathcal{H}_j$ ; la solution  $g(X)$  peut alors être caractérisée

par  $Y - g(X) \perp \mathcal{H}^{add}$ .  $\mathcal{H}^{add}$  étant généré par les espaces  $\mathcal{H}_j$ , on peut écrire :

$$Y - g(X) \perp \mathcal{H}_j, \forall j \text{ (ou bien } P_j(Y - g(X)) = 0, \forall j)$$

Si les projecteurs  $P_j$  étaient identifiables, le «backfitting algorithm» s'écrirait alors :

$$f_j^{(q)}(X_j) = P_j \left\{ Y - \sum_{k \neq j}^p f_k^{(q-1)}(X_k) \right\} = E \left\{ Y - \sum_{k \neq j}^p f_k^{(q-1)}(X_k) / X_j \right\}$$

On peut proposer alors l'écriture matricielle suivante :

$$\begin{pmatrix} I & P_1 & P_1 & \dots & P_1 \\ P_2 & I & P_2 & \dots & P_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_p & P_p & P_p & \dots & I \end{pmatrix} \begin{pmatrix} f_1(X_1) \\ f_2(X_2) \\ \vdots \\ f_p(X_p) \end{pmatrix} = \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_p Y \end{pmatrix}$$

ou

$$P_f = QY$$

où l'on remplacera l'opérateur projecteur  $P_j$  (opérateur mathématique) par une estimation obtenue par un opérateur de lissage  $S_j$ , comme on le fait classiquement (voir par exemple [13]) en estimation fonctionnelle non paramétrique.

## 2. Un exemple médical de régression logistique

### 2.1 La question des médecins

Pour certaines maladies un diagnostic prénatal est envisageable permettant de prévoir ou de prévenir la survenue d'un enfant sévèrement malformé. C'est le cas des translocations réciproques qui constituent la grande majorité des anomalies chromosomiques de structure. Pour aider les généticiens dans leur activité de conseil prénatal, il est nécessaire de pouvoir fournir une estimation du risque de survenue d'un enfant malformé. Disposant à Grenoble d'un recensement très important de translocations (voir [9]) (base de données la plus volumineuse au monde), nous avons retenu 5564 individus issus d'un parent porteur sain d'une translocation.

De façon à prédire, pour chacun d'eux, le risque de survenue d'un enfant malformé, la modélisation logistique a permis de sélectionner 10 variables sur un total de 19 variables explicatives (voir paragraphe suivant). Parmi ces 10 variables, certaines étaient disponibles d'emblée dans notre base de données, c'est le cas des bras des points de cassure chromosomiques (*bras*), de l'origine maternelle ou paternelle de la translocation (*origine*) et des longueurs des segments chromosomiques (*csbandeR1*, *tsbandeR1*, *tsbandeR2*). D'autres variables ont dû être construites par regroupement de modalités, c'est le cas des chromosomes impliqués (*mc*), du mode de déséquilibre

le plus probable (*moseg*), et des variables décrivant la viabilité zygotique (*via47*, *lamin*, *sinusadj*). La description détaillée de cette base de données et de l'ensemble des variables disponibles peut être trouvée dans [10] et [5].

## 2.2 La régression logistique

Pour aborder ce problème de prédiction nous avons été conduits à utiliser la régression logistique comme outil de modélisation : la variable à expliquer est la variable dichotomique «malade / non-malade ». La probabilité d'être malade qui est estimée par cette modélisation sera considérée comme étant le risque de survenue d'un enfant malformé, pour un parent porteur d'une translocation.

Lors de la construction du modèle logistique, nous n'avons pas retenu de méthode pas à pas pour les raisons suivantes :

- cette méthode peut conduire à sélectionner des variables non pertinentes ([11]),
- il n'existe pas *a priori* de supériorité du pas à pas par rapport à un choix orienté et avisé des variables ([12], [15]),
- disposant des résultats préliminaires de l'analyse descriptive des données concernant les variables chromosomiques les plus pertinentes, et compte tenu de l'existence de certaines variables contraintes à rester dans le modèle, il nous a semblé plus opportun de ne pas procéder à une sélection de variables «à l'aveugle».

Nous avons respecté le principe de parcimonie; ceci de manière à gagner en précision dans l'estimation des paramètres et en simplification dans l'interprétation du modèle. Le nombre important de variables (16 quantitatives et 3 qualitatives) a nécessité dans un premier temps une analyse univariée pour préciser l'influence de chaque variable. On utilise le test du score univarié pour les variables quantitatives : 6 variables ont été supprimées soit en raison de leur absence d'influence, soit en raison d'une forte corrélation avec une autre variable qui sera conservée pour l'étape suivante. En fin de modélisation les variables supprimées pourront être reinjectées dans le modèle pour vérifier, à l'aide d'un test du score multivarié, qu'elles n'apportent effectivement aucune amélioration significative. Le test de Wald est utilisé pour les variables qualitatives : cette procédure permet de proposer des regroupements de modalités. Dans notre exemple 2 variables à 3 modalités ont été ramenées à 2 modalités; la troisième est ramenée de 8 à 5 modalités. Les détails de cette procédure sont exposés dans [6]. Dans un deuxième temps, on procède à une analyse multivariée en partant du modèle le plus complet : ici 10 variables quantitatives, 3 qualitatives; avec des interactions allant jusqu'à l'ordre 3. Avec le même test cité précédemment, et selon une stratégie similaire à celle proposée dans [12] ou [18], nous avons retenu parmi les meilleurs modèles «possibles», un modèle comportant les variables les plus influentes, le plus petit nombre de paramètres et la plus petite déviance.

Les données étudiées sont des données observées qui peuvent être de mauvaise qualité. Dans la mesure où ces données peuvent avoir une influence erronée sur l'estimation des paramètres, il est important de les repérer. Les observations aberrantes ou atypiques ont été détectées par l'analyse des résidus.

Plusieurs méthodes de calcul des résidus sont disponibles : résidus de la déviance, résidus de Pearson, résidus d'Anscombe (dans [1], le chapitre VI avec entre

autre deux exemples d'analyse des résidus en régression logistique, est entièrement consacré à ces notions). Bien que les résidus de la déviance présentent une distribution plus symétrique que les deux autres, nous avons constaté qu'il était aussi pertinent de travailler sur les résidus de Pearson étant donné la grande taille de notre échantillon.

Tout résidu standardisé supérieur à 2 a été considéré comme résidu d'une observation «suspecte» *a priori*. Toute observation «suspecte» a été vérifiée avec retour à la source de l'information (article ou laboratoire d'origine). Des corrections ont parfois été effectuées lorsqu'il s'agissait d'erreurs de saisie ou d'interprétation des données.

Pour les données vérifiées qui restaient aberrantes, nous avons utilisé le test par délétion : y a-t-il changement dans la valeur des paramètres et de la déviance selon que l'observation à valeur suspecte est incluse dans les données ou non (voir [19])?

Comme il n'y a pas eu de modification dans la valeur des paramètres ou de la déviance, nous avons opté pour le choix de garder ces observations dans les données, dans la mesure où ces valeurs apparaissant comme aberrantes pouvaient être nécessaires à une estimation non biaisée des paramètres ([16]).

Le modèle logistique proposé était-il raisonnable pour décrire les données? Si ce n'était pas le cas, les données auraient dû montrer une déviation systématique par rapport aux valeurs prédites, ou bien quelques données seulement ne pas coller du tout.

En régression logistique, les résidus n'ont pas une distribution bien connue, il s'agit donc de simuler des données de manière à obtenir une estimation de la distribution théorique de ces résidus. Selon Jennings ([16]), «chaque résidu aurait sa propre distribution, ce qui complique singulièrement les choses. Les composantes individuelles de la déviance, comme celles du  $\chi^2$ , ont une distribution à deux maximums. Dans tous les cas, les résidus n'ont aucune raison de suivre une distribution du  $\chi^2$  ou une distribution normale». Pour vérifier s'il existe ou non une déviation systématique de ces résidus, on utilise le «graphe des probabilités empiriques». C'est un graphique qui représente en abscisse une distribution théorique de résidus (résidus simulés) et en ordonnée les résidus observés obtenus par régression logistique : il doit être linéaire, et permet de détecter des écarts des données par rapport aux hypothèses du modèle logistique ([17]). La distribution théorique est obtenue en simulant  $K$  nouvelles variables à expliquer; chacune d'elles suit une loi binomiale  $B(1, p)$ ,  $p$  étant le vecteur des valeurs prédites obtenu à partir des données observées sous le modèle considéré. Sur ces  $K$  nouvelles variables simulées, on applique le même modèle de façon à obtenir un vecteur de résidus pour chacune d'elles.  $K$  distributions de résidus sont alors obtenues, et on prend la médiane de ces résidus pour obtenir une approximation de la distribution théorique des résidus pour ce modèle. Le graphe obtenu (voir figure 1) ne montre pas de déviation systématique des résidus.

En l'absence de biais d'échantillonnage, le choix du modèle peut donc être considéré satisfaisant, d'autant que les résultats des tests n'ont pas montré de contradiction avec les données ou la connaissance actuelle ([2]).

Enfin, s'agissant d'un problème médical de prédiction, il était important, à partir de la valeur estimée du risque d'être malade, de procéder à une classification des individus malades et non malades de la base de données. Plusieurs méthodes sont possibles pour obtenir un taux d'erreur réel des résultats de classification d'un modèle :

soit tester le modèle sur de nouvelles observations, soit construire le modèle à partir d'une sélection du fichier global et le tester sur le reste du fichier, ce qui s'apparente à la méthode de la validation croisée ([4]). Nous avons retenu la deuxième méthode pour obtenir ce taux d'erreur réel.

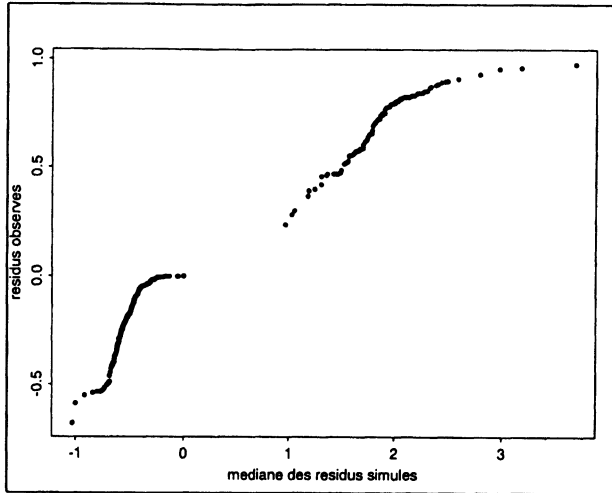


FIGURE 1 :

Graphe des résidus obtenus après simulation

Les résultats numériques de cette régression se sont avérés satisfaisants et exploitables pour les généticiens et ils sont détaillés dans [6]. Cette étape constitue un préalable indispensable à l'utilisation de la stratégie présentée ci-dessous et visant à l'amélioration du modèle. Ce n'est qu'après l'obtention d'un modèle ayant sélectionné les variables les plus pertinentes que les techniques des modèles additifs généralisés peuvent à notre avis être les plus efficaces.

### 3. Une stratégie en modèles additifs généralisés

Il s'agit, pour chaque variable explicative, de déterminer l'existence éventuelle d'une non linéarité, de proposer une transformation si la non linéarité existe, et de vérifier, après introduction de la variable transformée que l'on a amélioré les performances du modèle. La théorie des modèles additifs généralisés étant assez récente, il n'existe pas à notre connaissance dans la littérature d'étude de cas bien approfondie. Notre démarche a été totalement empirique. Schématiquement, nous distinguerons six étapes dans la stratégie que nous proposons. On rappelle que les modèles utilisés (aussi bien en glm qu'en gam) sont des modèles à loi binomiale avec fonction de lien canonique (logit). La notation «glm/  $Y$  sur  $X$ »(resp. «gam/  $Y$  sur  $X$ ») signifiera la régression de la variable  $Y$  sur la variable  $X$  au sens «glm/binomial-logit»(resp. «gam/binomial-logit»).



## 1) Existe-t-il une non-linéarité?

Une non linéarité sera retenue si la comparaison du modèle linéaire généralisé (glm/  $Y$  sur  $X_j$ ) et du modèle additif généralisé (gam/  $Y$  sur  $X_j$ ) montre un changement de déviance significatif. On conclura à l'existence d'une non linéarité pour un gain en déviance significatif à 0.1% . Dans le cas contraire la variable  $X_j$  restera sous forme linéaire simple.

2) Quelle est la forme de la fonction  $s_j(X_j)$  (fonction estimée de  $X_j$  obtenue après arrêt du «backfitting algorithm»)?

La forme du graphe de la fonction spline ( $s_j(X_j)$ ) avec les valeurs de la variable considérée ( $X_j$ ) peut suggérer une transformation non linéaire qui, dans certains cas, peut être polynomiale, logarithmique, inverse, ou autre. S'il existe une non-linéarité, on essaiera plusieurs transformations naturelles au vue du graphique.

## 3) Choix de la meilleure transformation ?

Les différentes transformations suggérées sont introduites dans un modèle glm/  $Y$  sur  $t(X_j)$ ,  $t$  fonction connue. Ces modèles n'étant pas emboîtés, il n'est pas possible ici d'utiliser un test pour le changement en déviance. On retient la transformation qui procure le gain en déviance maximum pour la plus petite perte en degré de liberté (ddl). Il apparaît intéressant d'explorer les transformations polynômiales, tout en restant vigilant quant au nombre de ddl perdus. La fonction spline de GAM provoque une perte de 3 ddl : on utilise en effet une fonction de spline cubique entraînant l'estimation de 3 paramètres supplémentaires. Pour un polynôme d'ordre 6, cette perte sera de 6 ddl. C'est pourquoi, lorsqu'un polynôme d'ordre élevé semble être la meilleure transformation, une analyse en composantes principales ayant pour variables la variable  $X_j$  et ses puissances successives est effectuée, et on retient les premières composantes principales associées aux axes de plus forte inertie. Cette procédure permet, lorsque le pourcentage d'inertie de l'ACP est satisfaisant, de réduire ainsi la perte en ddl.

4) Quelle est la forme de la fonction  $s()$  après transformation ?

Une transformation  $t$  ayant été retenue, on vérifie, par un modèle gam/  $Y$  sur  $t(X_j)$  la non significativité du gain en déviance par rapport au modèle glm/  $Y$  sur  $t(X_j)$ ; ce qui permet de s'arrêter dans la recherche d'une transformation satisfaisante.

## 5) On introduit toutes les variables transformées dans le modèle.

On introduit à nouveau les variables dans un modèle glm mais avec les transformations proposées. C'est la forme transformée de la variable qui est utilisée aussi dans les termes d'interactions. On considère alors le modèle gam sur le même prédicteur, et on compare les déviiances des deux modèles. On espère que le gain en déviance obtenu sera non significatif.

## 6) On vérifie que l'on a amélioré l'ajustement du modèle.

Pour vérifier l'amélioration due à la méthode GAM, on analyse l'adéquation du modèle aux données ainsi que les résultats de classification *a posteriori* (% de bien classés par le modèle, cf. application). Une étude des résidus permet de vérifier leur distribution, ainsi que la proportion de résidus élevés.

**4. Application sur la base de données**

Afin d'améliorer la clarté de la présentation des résultats numériques, nous noterons :

**ModGLM(1)** le modèle linéaire généralisé issu du paragraphe 2 :  $glm/ Y$  sur les variables  $X_j$  retenues.

**ModGAM(1)** le modèle additif généralisé  $gam/ Y$  sur les mêmes variables  $X_j$ .

**ModGLM(2)** le modèle linéaire généralisé issu du paragraphe 3 :  $glm/ Y$  sur les variables après transformations notées  $t(X_j)$ .

**ModGAM(2)** le modèle additif généralisé  $gam/Y$  sur les mêmes transformations  $t(X_j)$ .

Dix effets principaux (6 variables quantitatives et 4 variables qualitatives), et 3 termes d'interactions d'ordre 2 avaient été retenues pour le meilleur modèle proposé ModGLM(1). Dans le modèle ModGLM(2) quatre variables quantitatives sont introduites sous une forme transformée. Les transformations retenues sont les suivantes : - fonction inverse pour *csbandeR1* - fonction polynomiale du second degré pour *tsbandeR2* - fonction logarithmique pour *tsbandeR1* et pour *lamin*. (voir le paragraphe 2.1 pour la description de ces variables)

La procédure GAM n'a pas détecté de non-linéarité pour les 2 autres variables quantitatives (*via47* et *sinusadj*). Concernant les interactions, deux d'entre elles comprennent des variables quantitatives; pour celles-ci la procédure GAM a montré l'existence d'une non-linéarité. Les meilleurs résultats, en terme de gain en déviance, sont obtenus lorsque la variable quantitative est introduite sous sa forme transformée dans l'interaction (exemple : *origine \* t(tsbandeR1)*).

TABLEAU 1  
*Caractéristiques des trois modèles*  
 ModGLM(1), ModGAM(1) et ModGLM(2)

	ModGLM(1)	ModGAM(1)	ModGLM(2)
Nombre de paramètres	25	—	26
Nombre de paramètres significatifs	16	—	20
Déviance résiduelle	5303	5224	5261
ddl	5539	5521	5538
Valeurs du risque (min-max)	0.00 – 0.75	0.00 – 0.86	0.00 – 0.99
Résidus de Pearson > 4	16	—	12

Ces résultats confirment que la procédure GAM améliore l'ajustement du modèle ModGLM(1) en terme de déviance. Le modèle ModGLM(2) tente de se

rapprocher du modèle ModGAM(1) avec une perte négligeable en degré de liberté par rapport au modèle ModGLM(1). Ils confirment également le fait qu'une partie de la non-linéarité entre le prédicteur et les covariables a pu ainsi être éliminée. Le Tableau 2 montre la comparaison des déviations entre les 4 modèles.

TABLEAU 2 : Résultats de déviance

	ddl	déviance
Modèle ModGLM(1)	5539	5303
Modèle ModGAM(1)	5521	5224
	gain en déviance significatif à 0.1%	

	ddl	déviance
Modèle ModGLM(2)	5538	5261
Modèle ModGAM(2)	5518	5221
	gain en déviance non significatif à 0.1%	

A titre d'exemple, pour une des variables (*lamin* notée  $V_6$ ) les figures 2 et 3 montrent la forme de la fonction du prédicteur avant et après transformation de la variable ( $V_6 \text{Log}$  représente  $\log(K - \textit{lamin})$ ,  $K$  étant choisi  $> \textit{lamin}$ ).

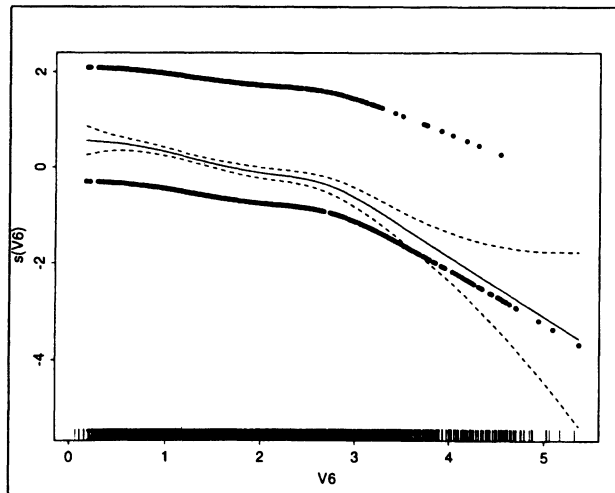


FIGURE 2 : Variable *lamin* avant transformation

La densité des observations est représentée au niveau de l'axe des abscisses par des tirets verticaux plus ou moins espacés. La courbe en trait plein représente

la fonction de lissage; elle est encadrée par les courbes (en pointillés) donnant son intervalle de confiance. Les points gras sont les observations «malade / non-malade» par rapport à la variable en question.

Comme l'objectif de la modélisation était l'estimation d'un risque de survenue d'un enfant malformé dans un but prédictif, nous avons aussi comparé les résultats de classification pour les deux modèles ModGLM(1) et ModGLM(2).

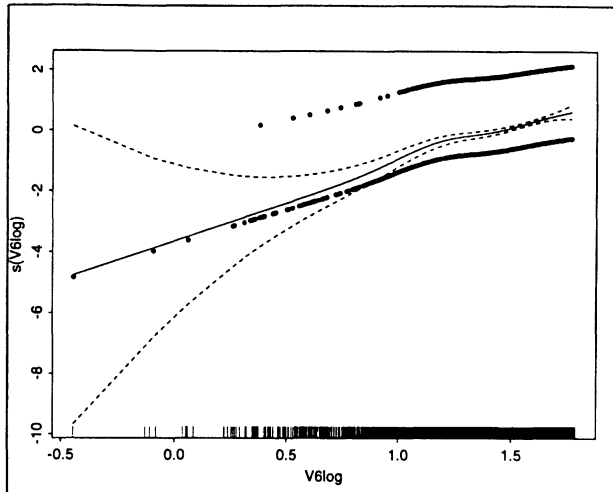


FIGURE 3 : Variable *lamin* après transformation

Si on classe les observations par ordre croissant de la valeur du risque estimé par le modèle, on remarque que la première observation à être mal classée (risque faible mais individu «malade») est la 294<sup>e</sup> pour le modèle ModGLM(1), et la 332<sup>e</sup> pour le modèle ModGLM(2). Ces observations ayant servi à la construction des modèles, ces résultats sont à considérer comme des taux d'erreurs apparents.

Nous avons alors calculé un taux d'erreur réel de classification, en choisissant des seuils de risque (voir tableau des résultats) pour la prédiction de la maladie, et en construisant des échantillons randomisés de validation du modèle. Pour chaque échantillon, 70 % de la population du fichier ont servi à l'estimation des paramètres du modèle, et les 30 % restants à sa validation.

Les tableaux 3 et 4 montrent que la proportion d'observations bien classées est assez proche avec l'un ou l'autre des modèles, mais que les résultats sont plus performants avec le modèle ModGLM(2) pour les risques faibles (< 1%). L'intervalle de confiance (I.C.) est calculé comme suit : pour chaque modèle 30 échantillons ont été tirés au hasard dans l'ensemble des données; l'I.C. est alors l'étendue du pourcentage de bien classés après exclusion de la valeur minimale et de la valeur maximale.

Ces résultats ont des conséquences pratiques importantes pour les généticiens, en raison d'une conduite à tenir éventuellement différente pour les risques très faibles.

TABLEAU 3

*Proportion d'observations «non malades» bien classées (taux d'erreur réel)*

Risque estimé	Bien classées avec ModGLM(1)			Bien classées avec ModGLM(2)		
	<i>n</i>	%	I.C.	<i>n</i>	%	I.C.
< 0.01	35	98.8	[93.8 – 100.0]	133	99.9	[98.0 – 100.0]
0.01 << 0.05	402	96.7	[94.5 – 98.2]	348	97.0	[95.5 – 98.3]
0.05 << 0.10	514	93.1	[88.6 – 96.5]	486	93.1	[87.9 – 95.7]

TABLEAU 4

*Proportion d'observations «malades » bien classées (taux d'erreur réel)*

Risque estimé	Bien classées avec ModGLM(1)			Bien classées avec ModGLM(2)		
	<i>n</i>	%	I.C.	<i>n</i>	%	I.C.
0.30 << 0.50	362	34.2	[30.6 – 39.1]	355	33.0	[28.8 – 39.6]
0.50 <	18	62.1	[42.9 – 83.4]	20	57.4	[37.5 – 75.0]

## 5. Discussion et conclusion

Sur cet exemple, et malgré l'utilisation d'une démarche empirique, nous avons vérifié l'amélioration de l'ajustement d'un modèle logistique lorsque la non-linéarité du lien entre certaines variables et le prédicteur peut être prise en compte.

La stratégie décrite peut être appliquée facilement à d'autres sujets utilisant une modélisation de type régression logistique ou toute autre modélisation appartenant à la famille des modèles linéaires généralisés. Des logiciels comme Splus ([8]) permettent d'avoir directement accès à la forme de la fonction  $s(X_j)$  dans un modèle additif généralisé. Ceci permet de suggérer de façon éclairée et non plus à l'aveugle des transformations de variables paraissant pertinentes.

Si parmi les objectifs recherchés celui de la prédiction est prédominant, on peut alors se limiter à l'estimation fournie directement par le modèle additif généralisé. Par contre lorsque l'interprétation de l'influence des variables est l'objectif prioritaire, les transformations suggérées peuvent aider à comprendre les mécanismes d'action de ces variables : on peut par exemple détecter un phénomène de rupture ou un phénomène de seuil.

## 6. Références Bibliographiques

- [1] ANTONIADIS, A., BERRUYER, J. et CARMONA, R. (1992) *Régression non linéaire et applications*. Economica, Paris.
- [2] BONNEU, M. et LAVERGNE, C. (1992) Bootstrap and Asymptotic prediction criterion estimate for binomial proportions in insemination data. *Biometrical Journal* 34 -1, p. 69-79
- [3] BREIMAN, L. and FRIEDMAN, J.H. (1985) Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Am. Statist. Assoc.* 80, p. 580-619.
- [4] BUNKE, O. and DROGE, B. (1984) Bootstrap and cross validation estimates of the prediction error for linear regression models. *Ann. Statistics* 13, p. 1400-1424.
- [5] CANS, C., COHEN, O. et coll. (1993) Human reciprocal translocations : is the unbalanced mode at birth predictable? *Human Genetic* 91, p. 228-232.
- [6] CANS, C. et coll. (1993) Logistic regression model to estimate the risk of viable unbalanced offspring in reciprocal translocations. *Human Genetic* 92, p. 598-604.
- [7] CANS, C. et coll. (1994) Application of G.A.M. in modelisation adverse outcome in reciprocal translocations. *Soumis à Genetic Epidemiology*.
- [8] CHAMBERS, J. and HASTIE, T. (1991) *Statistical models in S*. Chapman & Hall, New York.
- [9] COHEN, O. et coll. (1992) Human reciprocal translocations : a new computer system for genetic counseling. *Ann. Génét.* 35, p. 193-201.
- [10] COHEN, O. et coll. (1994) Viability thresholds for partial trisomies and monosomies. A study of 1159 viable unbalanced translocations. *Human Genetic* 93, p. 188-194.
- [11] FLACK, M. and CHANG, S. (1987) Frequency of selecting noise variable in subset regression analysis : a simulation study. *Am Statistician* 41, p. 84-86.
- [12] GREENLAND, S. (1989) Modeling and variable selection in epidemiologic analysis. *Am. J. Public Health* 79, p. 340-349.
- [13] HASTIE, T. and TIBSHIRANI, R. (1986) Generalized additive models (with discussion) *Statist. Sci.* 1, p. 297-318.
- [14] HASTIE, T. and TIBSHIRANI, R. (1990) *Generalized additive models*. Chapman & Hall, New York.
- [15] HOSMER, D.W. and LEMESHOW, S. (1989) *Applied logistic regression*. John Wiley & sons, New York.
- [16] JENNINGS, D.E., (1986) Outliers and residual distributions in logistic regression. *J. Am. Statist. Assoc.* 81, p. 987-990.
- [17] LANDWEHR, J., PREGIBON, D. and SHOEMAKER, A.C. (1984) Graphical methods for assessing logistic regression models. *J. Am. Statist. Assoc.* 79, p. 61-71.

- [18] MAC CULLAGH, P. and NELDER, J.A. (1989) *Generalized linear models*. Chapman & Hall, London.
- [19] PREGIBON, D., (1981) Logistic Regression Diagnostics. *Ann. Statistics* 9, p. 705-724.
- [20] WHITTAKER, J. (1989) *Graphical Models in Applied Multivariate Statistics*. John Wiley & sons, New York.