Rev. Epidém. et Santé Publ., 1999, 47, 593-604

Régression logistique *vs* autres modèles linéaires généralisés pour l'estimation de rapports de prévalences

Logistic regression vs *other generalized linear models to estimate prevalence rate ratios*

P. TRAISSAC, Y. MARTIN-PRÉVEL, F. DELPEUCH, B. MAIRE

Unité de Nutrition – IRD, Centre collaborateur de l'OMS, BP 5045, 911, Avenue Agropolis, 34032 Montpellier Cedex 01, France.
E.mail : traissac@mpl.ird.fr (*Tirés à part* : P. Traissac).

In cross-sectional studies, to quantify the association between a risk factor and a disease (possibly adjusted for confounders), in the framework of the multiplicative model, the more obvious effect measure is a prevalence rate ratio with an associated confidence interval. The validity of this confidence interval requires an unbiased estimator and an appropriate estimate of the variance. In numerous epidemiological studies however, routine use is made of odds ratios and logistic regression. As the odds ratio per se *is difficult to understand, prevalence odds ratios are often interpreted as prevalence rate ratios. But this latter approximation is valid only under the rare disease assumption. Moreover, in the logistic regression model, the variance of the estimates is based on the assumption of binomial variability, which is not always supported by the data; in the frequent case of overdispersion, this leads to under-estimation of the type I error rate. Yet, within the generalized linear model, it is easy to choose a link function other than the logit. For example, the log link (log-binomial model) is appropriate to directly estimate adjusted prevalence rate ratios. In case of overdispersion, it is also possible to achieve a better fit of the model, either by choosing another distribution in the exponential family or by estimating a dispersion parameter for the binomial distribution. Thus, there are no valid reasons for the systematic choice of odds ratio and of the logistic regression model to estimate prevalence rate ratios, unless the type of study imperatively requires their use.*

Cross-sectional study. Generalized linear model. Odds ratio. Prevalence rate ratio. Logistic regression.

Dans une étude transversale, pour quantifier l'association entre l'exposition à un facteur F et un état dichotomique M (éventuellement ajustée sur divers facteurs de confusion), dans le cadre du modèle multiplicatif, l'indice d'association le plus naturel est le rapport de prévalences, pour lequel le calcul d'un intervalle de confiance nécessite un estimateur sans biais et une estimation correcte de la variance. L'odds ratio et la régression logistique sont cependant très populaires dans la littérature épidémiologique. Mais l'odds ratio per se *étant difficilement interprétable, son estimation est souvent interprétée comme un rapport de prévalences, or ce n'est possible que si la condition de maladie rare est vérifiée. En outre, dans la régression logistique, le calcul de la variance d'estimation repose sur l'hypothèse de dispersion binomiale, qui n'est pas nécessairement vérifiée par les données; dans le cas, fréquent, de la surdispersion, cela conduit à une sous-estimation du risque de première espèce. Dans le cadre du modèle linéaire généralisé, il est aisé de choisir une autre fonction de lien que la fonction logit. Par exemple, le choix de la fonction de lien log (modèle log-binomial) permet d'obtenir directement des estimations de rapports de prévalences ajustés. De même, il est possible d'adapter le modèle aux données par le choix d'une distribution autre que binomiale dans la famille exponentielle ou par l'estimation d'un paramètre de dispersion. Mise à part une certaine pratique routinière, il*

n'existe donc pas de raison d'utiliser systématiquement l'odds ratio et la régression logistique lorsque leur emploi n'est pas imposé par le type d'étude.

Études transversales. Modèle linéaire généralisé. Odds ratio. Rapport de prévalences. Régression logistique.

INTRODUCTION

La quantification de la force de l'association entre l'exposition à un facteur de risque F et un état de santé dichotomique M (éventuellement ajustée sur divers facteurs de confusion) est une étape majeure dans les études épidémiologiques [1]. Dans le cadre d'une étude transversale, si l'on fait le choix du modèle multiplicatif pour quantifier l'association, les indices d'association les plus classiques sont le rapport de prévalences $RP = P_1/P_0$ et l'odds ratio de prévalences $ORP = [P_1/(1-P_1)] / [P_0/(1-P_0)]$, P_0 et P_1 désignant les prévalences de maladie respectivement dans le groupe non exposé et le groupe exposé. Ces définitions se généralisent aisément au cas d'un facteur de risque à plus de 2 catégories par le choix d'une catégorie de référence. Dans la suite, la présentation est faite à partir d'un facteur de risque dichotomique pour simplifier l'exposé.

Dans la majorité des cas, dans la mesure où l'étude porte seulement sur un échantillon représentatif et non sur la totalité des individus objets de l'étude, l'association observée sur l'échantillon est extrapolée à la population cible sous forme d'un intervalle de confiance se rapportant à l'indice d'association retenu. La validité de l'estimation par intervalle sur laquelle vont par la suite se baser les interprétations épidémiologiques (que ce soit dans une optique descriptive et/ou étiologique) nécessite, d'une part un estimateur sans biais et, d'autre part une estimation correcte de la variance.

Ici, on pose le problème de l'estimation de rapports de prévalences dans la mesure où, dans le cadre du modèle multiplicatif, c'est l'indice d'association le plus facilement interprétable. On discute les pratiques souvent rencontrées dans les publications, aussi bien pour l'estimation ponctuelle, objet de nombreux débats dans la littérature [2-7], que pour l'estimation par intervalle (variance des estimateurs). Des extensions possibles sont ensuite présentées dans le cadre du modèle linéaire généralisé.

LIMITES DE L'UTILISATION DE LA RÉGRESSION LOGISTIQUE POUR ESTIMER DES RAPPORTS DE PRÉVALENCES AJUSTÉS DANS LES ENQUÊTES TRANSVERSALES

Estimation ponctuelle d'un rapport de prévalences

Pour ce qui concerne l'estimation ponctuelle, on retient classiquement les biais de sélection, de classement et de confusion comme pouvant fausser la perception de l'association entre la maladie et le facteur de risque étudié [1]. Une autre source de biais, davantage proche du sens statistique du terme, peut être liée à un mauvais choix de l'estimateur (quantité calculée sur l'échantillon) vis-à-vis du paramètre à estimer dans la population. Dans le cadre d'une enquête transversale, dans la mesure où l'on peut estimer directement les prévalences de maladie dans les groupes exposés et non exposés, l'estimateur le plus naturel du rapport de prévalences dans la population est le rapport de prévalences observé sur l'échantillon soit $rp = p_1/p_0$. Cet estimateur est un estimateur sans biais de $RP=P_1/P_0$.

Estimation d'un rapport de prévalences par un odds ratio de prévalences

Dans la littérature épidémiologique, même dans le cas des études transversales où rien ne l'impose, une quantité souvent retenue pour quantifier l'association entre le facteur de risque étudié et la maladie au niveau de l'échantillon est l'odds ratio de prévalences $orp = [p_1/(1-p_1)]/[p_0/(1-p_0)]$. Cet estimateur est sans biais pour ORP. Une des raisons de la popularité de l'odds ratio dans la littérature épidémiologique est son caractère « tout terrain » puisque, contrairement à un rapport de risques, il est estimable même lorsque l'échantillonnage est stratifié sur M (études cas-témoins) : on peut en effet montrer que ORP peut se mettre sous la forme $ORP= [P_{E1}(1-P_{E1})]/[P_{E0}(1-P_{E0})]$ qui ne fait pas intervenir les probabilités conditionnelles de maladie (où P_{EO} et P_{EI}

désignent respectivement les fréquences d'exposition parmi les non-malades et les malades).

Cependant, l'odds ratio *per se* est difficilement interprétable directement comme indice d'association (certains auteurs le qualifiant même d'incompréhensible [2, 6]). De fait, les estimations ponctuelles ainsi obtenues sont souvent assimilées à des estimations de rapports de prévalences et interprétées comme telles : c'est-à-dire que *orp* est utilisé comme estimateur de RP. Or, la validité des estimations des rapports de prévalences obtenues de cette façon nécessite que la condition dite de « maladie rare » dans la population soit vérifiée, faute de quoi il existe un risque de biais important sur l'estimation. Dans le cas contraire, *orp* surestime la force de l'association (i.e. valeur davantage éloignée de 1) et cela d'autant plus qu'on s'éloigne de l'hypothèse de maladie rare. On pourra se référer à [1, 8] pour des détails sur la relation entre ORP et RP.

À titre d'illustration, le *tableau I* présente des résultats issus d'une étude transversale portant sur 170 enfants âgés de 24 à 36 mois d'une région rurale africaine. L'état de santé étudié est le « retard de croissance staturale » : un enfant est défini comme présentant un retard de croissance staturale si son indice taille-âge (exprimé en score d'écart-types par rapport à la médiane de la population de référence OMS/NCHS) est inférieur à -2 [9]. Le facteur de risque est le « petit poids de naissance », soit un poids de naissance $< 2\,500$ g [9]. Le rapport de prévalences observé sur l'échantillon est $rp = 18/(18 + 13)/30(30 + 109) = 0{,}58/0{,}22 = 2{,}69$, avec un intervalle de confiance à 0,95 : [1,74 – 2,69]. L'odds ratio de prévalences observé sur l'échantillon est $orp = [0{,}58/(1 - 0{,}58)]/[0{,}22/(1 - 0{,}22)] = 5{,}03$, avec un intervalle de confiance à 0,95 : [2,21-11,41]. Il est clair dans cet exemple que l'utilisation de l'odds ratio comme estimateur du rapport de prévalences pour étudier le lien entre petit poids de naissance et retard de taille conduirait à une surestimation considérable de la force de l'association du fait de la prévalence élevée de retard de taille dans la population étudiée.

TABLEAU I. — *Tableau croisé retard de croissance staturale x petit poids de naissance.*

	Retard de croissance staturale		
	Oui	Non	
Poids de naissance < 2 500 g	18	13	31
Poids de naissance > = 2 500 g	30	109	139
	48	122	170

RÉGRESSION LOGISTIQUE – ESTIMATION PONCTUELLE

De même que l'odds ratio est très souvent employé de façon routinière comme indice d'association, même dans le cas des études transversales, le modèle de régression logistique est classiquement employé pour estimer l'association entre F et M, ajustée sur divers facteurs de confusion. Ce modèle repose sur l'hypothèse d'une relation entre la probabilité P de maladie et les variables explicatives $X_1, X_2, \ldots, X_p$ (facteur de risque et/ou facteurs de confusion potentiels) de la forme :

$$P(M/X_1, \ldots, X_p) = \frac{e_0^{B\sum B_j X_j}}{1 + e_0^{B\sum B_j X_j}}$$

ou, ce qui est équivalent,

$$\log[P/(1 - P)] = B_0 + \Sigma B_j X_j$$

Il découle de la forme du modèle que, en l'absence d'interactions, si B_1 est le coefficient dans le modèle de X_1, variable dichotomique codant l'exposition au facteur, $\log(ORP) = B_1$. L'estimation des paramètres par maximum de vraisemblance permet ainsi d'estimer, à partir des données observées, les coefficients B_j et donc les odds ratios ajustés correspondants [10-12].

De même que pour l'estimation de rapports de prévalence bruts, les odds ratio ajustés, du fait de leur difficulté d'interprétation, sont souvent interprétés comme des rapports de risques ajustés (RP dans le cas des études transversales). Le même risque de biais que précédemment existe si la condition de maladie rare n'est pas vérifiée.

RÉGRESSION LOGISTIQUE — ESTIMATION PAR INTERVALLE

L'intervalle de confiance pour l'indice d'association, ajusté pour les éventuels facteurs de confusion, est déduit de l'intervalle de confiance du

paramètre correspondant dans le modèle de régression logistique. Le calcul de cet intervalle de confiance utilise la normalité asymptotique des estimateurs du maximum de vraisemblance et l'estimation de la variance obtenue dans le cadre du modèle de régression logistique. Cette dernière repose sur des hypothèses précises, en particulier que la loi binomiale soit un modèle adéquat pour décrire la variabilité aléatoire de la réponse (variable dichotomique codant la maladie). Si les données observées contredisent cette hypothèse (cas fréquent de la surdispersion) cela peut conduire à des intervalles de confiance erronément trop étroits et, dans les tests, à une sous-estimation du risque de première espèce. Les causes de surdispersion et les remèdes possibles seront abordés plus en détail dans la suite.

Une façon de remédier aux problèmes posés par l'utilisation de la régression logistique pour estimer des RP, aussi bien en termes de valeur ponctuelle que d'estimation par intervalle, est de se placer dans le cadre du modèle linéaire généralisé dont la régression logistique est un cas particulier.

MODÈLE LINÉAIRE GÉNÉRALISÉ

Notre propos n'est pas de faire une présentation détaillée du modèle linéaire généralisé. Le lecteur interessé pourra se reporter à [13-15]. Seuls les éléments essentiels pour la suite sont précisés ici.

LIMITES DU MODÈLE LINÉAIRE GÉNÉRAL

Un modèle linéaire général « classique » (dont l'analyse de la variance, la régression, l'analyse de covariance sont des cas particuliers), est de la forme :

$$Y_i = B_0 + \sum B_j X_{ij} + U_i$$

où Y_i est la variable réponse pour la i^e observation et $(X_{ij})_j = {}_{1,\ldots,p}$ le vecteur des variables explicatives (quantitatives et/ou qualitatives). Les termes aléatoires U_i sont supposés être des variables aléatoires normales, indépendantes, de moyenne nulle et de même variance : on suppose donc que chaque observation y_i est une réalisation d'une variable aléatoire gaussienne Y_i, d'espérance $\mu_i = B_0 + \sum B_j X_{ij}$ et de variance constante. Le vecteur des coefficients B_j est estimé par moindres carrés à partir des données observées.

Il est clair que, dans un certain nombre de cas, ce type de modèle n'est pas approprié. Par exemple :

— si l'hypothèse de normalité (qui implique notamment que Y soit une variable quantitative continue) n'est manifestement pas adaptée à décrire les données, par exemple des comptages ou des proportions ;

— si la valeur de la réponse Y_i est restreinte de fait à un certain intervalle (comme par exemple l'intervalle 0-1 pour une proportion) ; en effet, sans contraintes sur les coefficients B_j la valeur estimée par le modèle à partir de $B_0 + \sum B_j X_{ij}$ peut ne pas appartenir à l'intervalle ;

— si l'hypothèse de variance constante n'est pas adaptée (cas fréquent de données dont la variance augmente avec la moyenne).

ÉLÉMENTS D'UN MODÈLE LINÉAIRE GÉNÉRALISÉ

Le modèle linéaire généralisé est une extension du modèle linéaire général qui va pouvoir être appliqué à une plus grande variété de situations. Un modèle linéaire généralisé comporte les éléments suivants :

Fonction de lien

La partie linéaire est définie comme dans un modèle linéaire général par $B_0 + \sum B_j X_{ij}$. De même que dans un modèle linéaire général, les variables explicatives peuvent être qualitatives, quantitatives ou un mélange des deux. Une fonction g, dite fonction de lien (« link function »), définit la relation entre l'espérance de la loi et la partie linéaire par $g(\mu_i) = B_0 + \sum B_j X_{ij}$ ou, ce qui est équivalent, $\mu_i = g^{-1}(B_0 + \sum B_j X_{ij})$. Une des utilités de la fonction de lien est de garantir que les valeurs prédites restent dans l'intervalle des valeurs possibles pour la variable. Dans les applications épidémiologiques elle déterminera également, comme nous le verrons dans la suite, l'interprétation qui pourra être faite des paramètres du modèle comme indices d'association.

Distribution

Les observations sont supposées indépendantes (cf [16] pour le cas plus général) et être des réalisations d'une loi de probabilité appartenant à la famille exponentielle. Une des conséquences est que la variance dépend de l'espérance suivant une certaine fonction de variance V(μi), par Var $(Y_i) = \Phi V(\mu_i)/wi$, où Φ, paramètre de dispersion, est une

constante (connue ou à estimer selon le cas) et w_i un poids connu à l'avance pour chaque observation (souvent égal à 1). Un certain nombre de distributions de probabilités classiques relèvent de la famille exponentielle, par exemple la loi normale ($V(\mu) = 1$, $\Phi = \sigma^2$), la loi binomiale ($V(\mu) = \mu(1-\mu)$, $\Phi = 1$), la loi de Poisson ($V(\mu) = \mu$, $\Phi = 1$).

La combinaison de différents choix de distributions pour modéliser la variabilité aléatoire de la variable réponse autour de son espérance et de fonctions de lien pour modéliser la relation entre l'espérance et les variables explicatives, permet de définir de nombreux modèles linéaires généralisés adaptés à analyser divers types de données. La plupart des logiciels statistiques généralistes tels que SAS, S, STATA, Genstat, GLIM, etc. disposent de modules ou de procédures permettant de mettre en œuvre ce type de modèle en toute généralité (choix de la fonction de lien et de la distribution), comme par exemple PROC GENMOD dans SAS® [17].

Exemples de modèles linéaires généralisés

Le modèle linéaire général est un cas particulier de modèle linéaire généralisé, correspondant au choix de la distribution normale et de la fonction de lien identité, du fait que l'on pose directement $\mu_i = B_0 + \sum B_j X_{ij}$.

De même, la régression logistique correspond au choix de la loi binomiale $B(1, P_i)$ pour modéliser la variable aléatoire Y_i dont l'observation yi (présence/absence de maladie) est supposée être une réalisation. La fonction de lien est la fonction logit, qui relie l'espérance de la loi binomiale $\mu_i = P_i$ aux prédicteurs par la relation $\text{logit}(\mu_i) = \text{logit}(P_i) = \text{Log}[P_i/(1-P_i)] = B_0 + \sum B_j X_{ij}$. Parmi d'autres considérations, le choix de la fonction logit est dicté par la nécessité de prédire des valeurs de p qui restent dans l'intervalle [0, 1].

D'autres choix sont évidemment possibles comme nous le verrons dans la suite.

Estimation des paramètres et tests dans le modèle linéaire généralisé

Les paramètres B_j sont estimés par maximum de vraisemblance. Les estimations par intervalle et certains tests sur les paramètres du modèle (Chi-2 de Wald) sont basés sur la normalité asymptotique des estimateurs du maximum de vraisemblance. On définit la déviance (« scaled deviance ») associée à un modèle donné comme 2 fois la différence entre la log-vraisemblance du modèle et la log-vraisemblance du modèle saturé [18].

La comparaison de modèles emboîtés (i.e. par exemple pour tester l'intérêt de l'ajout d'une ou plusieurs variables explicatives) est réalisée par test du rapport de vraisemblance : sous l'hypothèse nulle, la différence des déviances (égale à la différence des log-vraisemblances) suit asymptotiquement une loi de chi-deux à k degrés de libertés (d.d.l.) si k est le nombre de paramètres supplémentaires du modèle le plus complexe.

L'adéquation du modèle aux données (« goodness of fit ») peut théoriquement être testée par comparaison du modèle retenu au modèle saturé en se basant sur la déviance ; en effet, sous certaines conditions, celle-ci suit asymptotiquement, sous l'hypothèse nulle, une loi de chi-deux à n-p d.d.l. où n est le nombre d'observations du tableau de données et p le nombre de paramètres du modèle. Toutefois, en sus du problème du choix du modèle saturé, dans bon nombre de cas l'approximation n'est pas très bonne [13], si bien qu'en règle générale on utilise peu souvent la valeur p associée. Néanmoins, un critère rudimentaire pour juger de l'ajustement est de faire le rapport de la déviance associée au modèle considéré à son nombre de d.d.l. L'adéquation du modèle est alors jugée d'autant meilleure que le rapport est proche de 1 (du fait que l'espérance d'une loi de chi-deux est égale à son nombre de d.d.l.) [14]. D'autres tests sont également possibles, de même que l'analyse des écarts individuels au modèle (résidus) : voir [13, 14] pour davantage de précisions.

UTILISATION DU MODÈLE LINÉAIRE GÉNÉRALISÉ POUR L'ESTIMATION DE RAPPORTS DE PRÉVALENCES

Fonction de lien

Dans le cadre du modèle linéaire généralisé, le choix de la fonction de lien est conditionné par l'interprétation qui peut être faite des paramètres du modèle en termes d'indices d'association.

Fonction de lien logit

Dans le cas de la régression logistique (lien logit, distribution binomiale) c'est le choix de la fonction de lien logit qui permet l'interprétation des paramètres en termes d'odds ratio. En s'intéressant, pour simplifier la présentation, à un modèle univarié du type $\text{logit}(P) = B_0 + B_1X_1$ en fonction de X_1, variable binaire codant l'exposition, on a en effet :

$\log(ORP) = \log(P_1/(1-P_1)) - \log(P_0/(1-P_0) = \text{logit}(P_1) - \text{logit}(P_0) = B_0 + B_1 \times 1 - (B_0 + B_1 \times 0) = B_1$ soit $ORP = \exp(B_1)$

Le même résultat est obtenu pour un OR ajusté dans le cas d'un modèle de type $\text{logit}(P) = B_0 + \sum B_jX_j$ comprenant une ou plusieurs covariables, dans la mesure où tous les termes associés aux autres covariables s'annulent.

Fonction de lien log

Si l'on désire estimer d'autres indices d'association, cela est parfois possible par le choix de la fonction de lien adéquate. Par exemple, la fonction log permet d'estimer des rapports de prévalence. En effet, dans le cadre d'un modèle linéaire généralisé, analogue à celui de régression logistique, mais dont la fonction de lien est la fonction log, soit $\log(P) = B_0 + B_1E$ par analogie avec l'exemple précédent, on a alors :

$\log(RP) = \log(P_1) - \log(P_0) = B_0 + B_1 \times 1 - (B_0 + B_1 \times 0) = B_1$

soit $RP = \exp(B_1)$.

On peut faire la même remarque que précédemment concernant la présence d'éventuelles covariables dans le modèle. Ce choix de la fonction de lien log permet donc d'estimer directement des RP en s'affranchissant de l'hypothèse de maladie rare. Ce modèle, correspondant au choix de la fonction de lien log et de la distribution binomiale, est parfois appelé modèle log-binomial. Il est recommandé par certains auteurs comme le modèle à retenir pour l'analyse de données dont la réponse est de type présence/absence lorsqu'on s'intéresse à estimer des RP [7, 19]. Néanmoins, certains problèmes peuvent se poser lors de la phase itérative d'estimation des paramètres : en effet, en l'absence de contraintes sur les paramètres B_j, du fait du modèle $\log(P_i) = B_0 + \sum B_jX_{ij}$ ou, ce qui est équivalent, $P_i = \exp(B_0 + \sum B_jX_{ij})$, rien ne garantit *a priori* que les valeurs de p prédites par le modèle ne soient pas supérieures à 1. Dans [19], il est démontré que cela n'est pas un argument valable contre l'utilisation de ce modèle au moins dans le cas où toutes les variables explicatives sont qualitatives. Lors de la recherche itérative des solutions des équations de vraisemblance, il arrive cependant que l'algorithme de recherche de la solution du maximum de vraisemblance s'arrête, du fait de valeurs estimées pour $B_0 + \sum B_jX_{ij}$ supérieures à zéro. D'après [19] et notre expérience, cela peut être résolu dans la majorité des cas par un choix adéquat de valeurs initiales pour les paramètres. Une autre solution peut être de fixer des contraintes sur les valeurs des paramètres lors du processus d'estimation.

Cette fonction de lien log est également classiquement utilisée dans le modèle dit « régression de Poisson » ou modèle log-Poisson [20], modèle linéaire généralisé de fonction de lien log et dont la distribution est la distribution de Poisson $P(\lambda_i)$. *A priori* adapté pour modéliser des données de comptage, davantage que des proportions, il peut néanmoins être adapté sous la forme :

$\log(\lambda_i) = \log(m_i) + B_0 + \sum B_jX_{ij}$

(présentation avec données regroupées par « covariate patterns »), où $\log(m_i)$ est un « offset », terme exclu du processus d'estimation. On a alors, de même que pour la régression logistique (modèle logit-binomial), une interprétation des paramètres comme logarithmes de rapports de prévalences (ajustés sur les covariables).

Autres fonctions de lien

D'autres fonctions de lien sont également utilisables, en fonction des modèles que l'on pense adéquats pour représenter les données observées et des interprétations possibles des paramètres.

Par exemple, le modèle de régression probit (fonction de lien égale à l'inverse de la fonction de répartition de la loi normale centrée réduite) donne des résultats très proches de ceux de la régression logistique en terme de valeurs prédites. Il est néanmoins moins utilisé du fait de l'absence d'interprétation claire des paramètres du modèle en termes épidémiologiques.

La fonction $\log(-\log(1-p))$, dite complémentaire log-log, donne également souvent des résultats proches de la fonction logit sans toujours d'interprétation claire des paramètres. Elle peut

néanmoins être utilisée pour estimer, sous certaines hypothèses, des rapports de taux d'incidence à partir d'enquêtes transversales [21] : en effet, sous l'hypothèse d'une incidence constante au cours du temps, la fonction de survie s'écrit $1 - P_i = \exp(-\lambda t_i)$, d'où $\log(-\log(1 - P_i)) = \log(t_i) + \log(\lambda)$. Un modèle linéaire généralisé de type $\log(-\log(1 - Pi)) = B_0 + \sum B_j X_{ij}$ peut donc permettre, dans certains cas, d'estimer des rapports de taux d'incidence.

Si on se place dans le cadre du modèle additif (ce qui dépasse le cadre de cette présentation) on peut remarquer que la fonction de lien identité (modèle de type $P_i = B_0 + \sum B_j X_{ij}$) permet d'interpréter les paramètres du modèle en termes de différences de risques [22]. Cette approche peut néanmoins poser, de façon analogue au modèle log-binomial, un certain nombre de problèmes (valeurs prédites par le modèle non restreintes *a priori* à l'intervalle [0-1]).

Distribution

Après avoir opté pour une certaine fonction de lien, on est amené, pour définir la partie aléatoire du modèle, à faire le choix d'une distribution de probabilité permettant de modéliser au mieux la variabilité présente dans les données. Si des informations concernant le mode de recueil des données ou des hypothèses sur les phénomènes étudiés peuvent guider ce choix, elles doivent être prises en compte, mais ce n'est pas toujours le cas. On se contentera alors de choisir, dans la famille exponentielle, la distribution qui permet de décrire le mieux possible les données observées.

Surdispersion

Dans les modèles basés sur la distribution binomiale, un cas fréquent de mauvaise adéquation du modèle est celui de la surdispersion que l'on définit *stricto-sensu* comme une situation où la variance de la réponse est supérieure à la variance attendue sous hypothèse binomiale [13, 23]. Dans la mesure où les calculs sont faits sous hypothèse binomiale, comme évoqué ci-dessus, une conséquence de la surdispersion est de fournir des estimations des écarts-types des paramètres erronément trop petits (résultant en des intervalles de confiance trop étroits pour les indices d'association et des tests ne respectant pas le risque nominal de première espèce). De façon théorique, les deux principales causes qui peuvent faire qu'une somme de variables aléatoires de Bernoulli $B(1, P_i)$ ne suit pas une loi binomiale, sont la non-indépendance entre les tirages ou des probabilités Pi non constantes [23].

Du point de vue pratique, divers éléments peuvent conduire à l'une ou l'autre des deux situations précédentes et donc à une surdispersion [13, 14, 23] : a) mauvaise spécification du modèle décrivant l'espérance de la loi en fonction des variables explicatives (mauvais choix de la fonction de lien, interactions et/ou covariables non prises en compte), b) non-indépendance entre les observations, c) « outliers ».

Il est à noter que les critères les plus couramment utilisés pour juger de l'adéquation du modèle aux données (tels ceux basés sur la déviance) ne permettent pas de distinguer les causes possibles dont l'identification sera donc en grande partie conditionnée par les informations dont on dispose.

Les remèdes à apporter dépendront du diagnostic porté : il faudra en particulier, avant tout, vérifier que des covariables et/ou des interactions importantes n'ont pas été omises dans le modèle (certains auteurs [23] utilisent le terme de surdispersion « apparente » pour désigner un écart à la variabilité binomiale dû à l'omission de variables explicatives importantes).

Prise en compte de la surdispersion dans la distribution

Si la prise en considération d'interactions ou de variables explicatives supplémentaires ne permet pas de résoudre le problème, dans le cadre du modèle linéaire généralisé il est possible de prendre en compte la surdispersion en modifiant la variance de la distribution qui définit la partie aléatoire du modèle. Passer d'un modèle log-binomial à un modèle log-Poisson peut par exemple améliorer l'adéquation du modèle aux données et fournir des intervalles de confiance et des tests reflétant davantage la réalité des données.

Il est également possible d'estimer le paramètre de dispersion Φ (*a priori* égal à 1 dans le cas de la loi binomiale et de la loi de Poisson) pour modéliser de façon plus satisfaisante la variabilité par intégration de ce paramètre dans la fonction de variance. Celle-ci devient alors pour la loi binomiale $V(\mu) = \Phi\mu(1-\mu)$ et $V(\mu) = \Phi\mu$ dans le

cas de la loi de Poisson. Les tests et les intervalles de confiance sur les paramètres obtenus sont ainsi corrigés pour mieux tenir compte de la variabilité présente dans les données. Le paramètre de dispersion Φ peut être estimé de différentes façons, la plus classique étant par le rapport de la déviance à son nombre de d.d.l.

EXEMPLES

Dans ce paragraphe nous présentons des exemples d'utilisation des modèles évoqués ci-dessus pour estimer des rapports de prévalences. Ces exemples, mis en œuvre sur un même jeu de données simulées, ont pour but essentiel d'attirer l'attention du lecteur sur certains points particuliers concernant l'estimation de rapports de prévalences, sans caractère de généralité concernant l'application des mêmes modèles à d'autres données (comme cela est fait par exemple dans [19] et [7]).

Données

Les données correspondent à un échantillon aléatoire (supposé issu d'une enquête transversale) de 360 individus : on dispose d'informations concernant le statut vis-à-vis de l'état étudié (malade : oui /non), de l'exposition au facteur E étudié (exposé : oui/non), facteur pris dichotomique pour la simplicité de présentation et la valeur prise pour deux autres covariables F et G respectivement à 3 et 2 modalités.

Modèles mis en œuvre

Dans les exemples suivants, on se propose d'estimer l'effet de l'exposition au facteur de risque binaire E, ajusté sur le potentiel facteur de confusion F supposé à 3 classes, en quantifiant si possible, lorsque le modèle le permet, la force de l'association par l'estimation de rapports de prévalences ajustés, dans le cadre de différents modèles linéaires généralisés. Dans chacun des modèles, les variables explicatives sont la variable indicatrice codant l'exposition au facteur E et les deux variables indicatrices codant les valeurs prises par les individus pour le facteur F. Avec le terme constant on a donc un modèle à 4 paramètres à comparer à un modèle saturé à 12 paramètres, soit 8 d.d.l. pour la déviance.

L'une des raisons souvent invoquée pour l'utilisation en routine de la régression logistique est sa disponibilité dans tous les logiciels statistiques courants. Or, depuis quelques années, la mise en œuvre du modèle linéaire généralisé est aisée du point de vue informatique. En effet, la plupart des logiciels statistiques généralistes (GLIM, SAS, S, STATA, Genstat, etc.) disposent, depuis plusieurs années, de procédures permettant de mettre en œuvre le modèle linéaire généralisé de la manière la plus générale (choix de la fonction de lien et de la distribution). Pour ce qui concerne les exemples, la mise en œuvre informatique des modèles est faite avec la procédure PROC GENMOD de SAS®, dont des éléments de syntaxe sont donnés en annexe.

Résultats et discussion

La prévalence de maladie sur l'échantillon est de 0,55. Le *tableau II* présente les résultats de l'ajustement des différents modèles linéaires généralisés. Pour ce qui concerne les paramètres estimés et leur éventuelle interprétation en termes d'indices d'association, seuls les résultats correspondant au facteur de risque étudié E sont donnés.

Du point de vue de la signification statistique, tous les modèles 1 à 4 basés sur la loi binomiale semblent équivalents pour ce qui concerne l'effet de l'exposition E (ajustée sur F) sur la prévalence de maladie : valeurs p conduisant au rejet de l'hypothèse nulle pour des seuils de l'ordre de 0,01 ou 0,005. Les modèles de fonction de lien probit et complémentaire log-log présentent le désavantage que le paramètre B_1 ne peut, dans ce cas, s'interpréter directement comme fonction simple d'un indice d'association. En revanche, les modèles logit-binomial (i.e. régression logistique) et log-binomial, permettent de donner des estimateurs du rapport de prévalences et qui sont, dans ce cas, très différents (surestimation de près de 50 % de la force de l'association avec la fonction de lien logit si l'on interprète l'odds ratio en termes de rapport de prévalences, ce qu'il n'y a évidemment pas lieu de faire dans ce cas). Comme on l'a dit plus haut, le modèle à retenir pour estimer un rapport de prévalences ajusté serait certainement basé sur la fonction de lien log. Néanmoins, de manière générale, dans le cas de l'exemple, les modèles basés sur la distribu-

TABLEAU II. — *Résultats de l'ajustement des différents modèles linéaires généralisés.*

Modèle	Fonction de lien	Distribution	Paramètre de dispersion Φ	Scaled-Deviance	d.d.l. (n-p)	Scaled Deviance / d.d.l	Estimateur de B_1	Écart-type de l'estimateur	Valeur p (H_0: $B_1 = 0^a$)	Estimateur de $\exp(B_1)^c$	I.C. 0,95 pour $\exp(B_1)^d$
1	Logit	binomiale	1	29,29	8	3,66	0,68	0,22	0,0019	1,97	1,28-3,02
2	Probit	binomiale	1	29,35	8	3,66	0,42	0,13	0,0019		
3	C. log log	binomiale	1	29,30	8	3,66	0,46	0,15	0,0019		
4	Log	binomiale	1	29,47	8	3,68	0,29	0,10	0,0025	1,34	1,11-1,62
5	Log	poisson	1	11,95	8	1,49	0,29	0,14	0,0405	1,34	1,01-1,78
1'	Logit	binomiale	3,66	8	8	1	0,68	0,42	0,1052	1,97	0,87-4,46
2'	Probit	binomiale	3,66	8	8	1	0,42	0,26	0,1050		
3'	C. log log	binomiale	3,66	8	8	1	0,46	0,28	0,1049		
4'	Log	binomiale	3,68	8	8	1	0,29	0,19	0,1149	1,34	0,93-1,92
5'	Log	poisson	1,49	8	8	1	0,29	0,18	0,0937	1,34	0,95-1,89

[a] – Chi-deux de Wald.

[b] – Effet du facteur de risque E, ajusté sur le facteur de confusion F.

[c] – Rapport de prévalences ou odds ratio.

[d] – I.C. de Wald

tion binomiale donnent un ajustement peu satisfaisant et ne semblent donc pas à retenir : valeurs de déviance relativement élevées correspondant à une surdispersion. On doit alors prioritairement se demander si celle-ci ne découle pas d'un modèle mal spécifié (omission d'interactions et/ou de covariables importantes). Dans le cas contraire, on peut envisager de modifier la partie aléatoire du modèle pour mieux décrire la variabilité de la réponse.

C'est ce qui est fait dans le cas du modèle log-Poisson qui semble donner un ajustement meilleur. L'estimation ponctuelle du rapport de prévalences est évidemment la même que pour le modèle log-binomial du fait que seule la distribution est différente et non la fonction lien. En revanche, la meilleure prise en compte de la variabilité effectivement présente dans les données (qui était sous-estimée par l'hypothèse binomiale) résulte en une variance d'estimation plus élevée que celle du modèle log-binomial : par suite, la valeur p correspondant à l'échantillon observé n'est plus très éloignée de 0,05 ; de même, l'intervalle de confiance à 0,95 pour le rapport de prévalence s'élargit et frôle la valeur 1.

Si le passage de la distribution binomiale à la distribution de Poisson est une façon de gérer la surdispersion, une autre possibilité est, pour tous les modèles, d'estimer à partir des données un paramètre de dispersion (par exemple basé sur la déviance). La prise en compte d'un paramètre de dispersion ne modifie évidemment pas les estimations ponctuelles du paramètre B_1 et les éventuelles interprétations en termes de rapport de prévalences. En revanche, dans la mesure où tous les modèles initiaux étaient surdispersés à des degrés divers, cette correction a, dans notre exemple, des effets importants sur les variances des estimateurs et donc les valeurs p et les intervalles de confiance. Quel que soit le modèle retenu pour décrire la relation entre la probabilité de maladie et l'exposition au facteur E (ajustée sur F), il faut accepter un risque de première espèce de 0,10 ou plus pour rejeter l'hypothèse nulle. De façon corollaire, pour les modèles d'intérêt (log-binomial et log-poisson), l'inter-

valle de confiance à 0,95 pour le rapport de prévalences recouvre la valeur 1.

Cet exemple est bien sûr volontairement caricatural ; il est vraisemblable que dans le cas de données réelles, l'effet d'une telle correction (par exemple pour tenir compte de corrélations entre les observations) ne sera pas aussi fort. Néanmoins, il est important de garder à l'esprit que, quel que soit le modèle utilisé, les résultats obtenus sont basés sur une formulation précise du modèle, aussi bien pour la partie déterministe que pour la partie aléatoire. Des écarts trop importants à ces hypothèses peuvent avoir des implications fortes sur les résultats. En particulier, cela peut être parfois le cas lorsqu'on utilise le modèle de régression logistique sans trop de discernement.

CONCLUSION

Pour l'estimation de rapports de prévalences dans le cas d'études transversales, le modèle linéaire généralisé est très souple : le choix de la fonction de lien (log dans notre cas) permet de choisir le type d'indice d'association que l'on souhaite estimer tandis que le choix de la distribution de probabilité (avec ou sans paramètre de dispersion) permet, dans un certain nombre de cas, de représenter au mieux la variabilité présente dans les données. Ces choix peuvent, comme on l'a vu sur un exemple, avoir des conséquences importantes sur les résultats.

Mise à part une certaine pratique routinière, il n'existe donc pas de raison d'utiliser systématiquement l'odds ratio et la régression logistique lorsque leur emploi n'est pas imposé par le type d'étude (cas des études transversales). En effet, aussi bien en termes de valeur centrale de l'intervalle de confiance que de dispersion, cette méthode va parfois ne pas donner des estimations adéquates des rapports de prévalences pour quantifier l'association entre l'exposition et l'état de santé étudiés.

Annexe

MISE EN ŒUVRE DU MODÈLE LINÉAIRE GÉNÉRALISÉ DANS SAS : PROC GENMOD.

Dans SAS®, au-delà des procédures dédiées au modèle linéaire général, dont PROC GLM [24] et PROC MIXED [17], de nombreuses procédures permettent de mettre en œuvre certains modèles particuliers relevant du modèle linéaire généralisé : par exemple PROC PROBIT pour la régression logistique et probit, PROC LOGISTIC et PROC CATMOD pour la régression logistique et les modèles apparentés [24]. Néanmoins, la procédure permettant de mettre en œuvre le modèle linéaire généralisé de la façon la plus générale est la procédure PROC GENMOD dont la première version date du début des années 1990 [17, 25]. Cette procédure dispose d'un choix étendu de fonctions de lien et de distributions ainsi que de possibilités de programmation pour spécifier des fonctions de lien ou des distributions particulières. Les dernières versions permettent également de mettre en œuvre des modèles prenant en compte d'éventuelles corrélations entre les observations (Generalized Estimating Equations). Pour ceux que la syntaxe SAS rebute, le module d'analyse interactive de données SAS/INSIGHT permet également de mettre en œuvre le modèle linéaire généralisé. Les exemples de syntaxe suivants supposent que les données sont contenues dans un tableau SAS de nom TABDON, les identifiants des variables étant ceux décrits plus haut. Ils ont pour but d'illustrer les principaux éléments de la syntaxe de PROC GENMOD et non d'en donner une présentation exhaustive. Exemple de syntaxe pour la mise en œuvre d'un modèle de régression logistique (i.e. logit-binomial) :

```
PROC GENMOD DATA=TABDON;
  CLASS E F ;
  MODEL Y/N = E F / LINK=LOGIT DIST=BIN TYPE1 TYPE3
                                    OBSTATS RESIDUALS;
RUN ;
```

L'instruction CLASS permet de déclarer les variables E et F comme qualitatives (codage disjonctif). Le paramétrage et donc le choix de la classe de référence pour le calcul des rapports de prévalences ou des odds ratios peut être modifié par rapport aux options par défaut. L'instruction MODEL déclare la (ou les) variables réponse et les variables explicatives : pour la distribution binomiale, dans le cas de données regroupées, la syntaxe est du type « nombre de réalisations de l'événement »/« nombre de tirages » = variables explicatives. Ces dernières peuvent être quantitatives et/ou qualitatives avec possibilité de spécifier des interactions d'ordre quelconque. Contrairement à ce qui est dit dans [3, 7] dans le cas de données binomiales, PROC GENMOD peut également analyser directement des données individuelles : si M est une variable dichotomique qui code la présence/absence de maladie pour chaque individu la syntaxe est alors du type M = variables explicatives. Les options principales de l'instruction MODEL sont LINK = et DIST =, elles permettent de spécifier respectivement la fonction de lien et la distribution. Pour un modèle log-binomial on aura ainsi par exemple :

```
PROC GENMOD DATA=TABDON;
  CLASS E F ;
  MODEL Y/N = E F / LINK=LOG DIST=BIN TYPE1 TYPE3
                                    OBSTATS RESIDUALS;
RUN;
```

Pour un modèle probit-binomial ou complémentaire log-log binomial il suffirait de préciser respectivement LINK=PROBIT OU LINK= CLOGLOG.

Pour ajuster sur les mêmes données un modèle log-Poisson, tel que décrit plus haut, en sus de l'utilisation de l'option DIST=POISSON pour spécifier la distribution, il est nécessaire de l'adapter par l'intermédiaire d'un "OFFSET" qui nécessite au préalable le calcul de la quantité $\log(m_i)$ pour chaque ligne du tableau des données :

```
DATA TABDON;
  SET TABDON ;
  LN=LOG(N) ;
RUN ;
PROC GENMOD DATA=TABDON;
  CLASS E F ;
  MODEL Y = E F / OFFSET=LN LINK=LOG DIST=POISSON
                  TYPE1 TYPE3 OBSTATS RESIDUALS;
RUN ;
```

Si nécessaire, un certain nombre d'options de l'instruction MODEL sont également disponibles pour l'estimation d'un paramètre de dispersion pour les divers modèles. Par exemple l'option DSCALE permet d'estimer le paramètre de dispersion à partir de la déviance comme fait dans les exemples présentés :

```
PROC GENMOD DATA=TABDON;
  CLASS E F ;
  MODEL Y/N = E F / DSCALE LINK=LOG DIST=BIN TYPE1
                    TYPE3 OBSTATS RESIDUALS;
RUN ;
```

permet d'ajuster un modèle log-binomial en tenant compte d'une éventuelle surdispersion.

Dans la mesure où la procédure GENMOD permet de mettre en œuvre un grand nombre de modèles différents, de fait, les sorties ne comportent pas toujours pour certains modèles, autant d'éléments utiles à l'interprétation que certaines procédures spécifiques, dédiées à un type de modèle particulier. Il est toujours possible de les améliorer avec des instructions de programmation et/ou l'utilisation du macro-langage intégré dans SAS. En particulier, en fonction du type de modèle linéaire généralisé ajusté, il peut être intéressant d'enrichir les sorties par certains calculs spécifiques (rapports de prévalences pour les modèles basés sur la fonction de lien log, odds ratio pour ceux utilisant la fonction logit). Par exemple, le code suivant réalise un tel calcul à la suite de l'ajustement d'un modèle log-binomial :

```
PROC GENMOD DATA=TABDON;
  CLASS E F ;
  MODEL Y/N = E F / LINK=LOG DIST=BIN TYPE1 TYPE3
                                OBSTATS RESIDUALS;
MAKE 'PARMEST' OUT=PARMEST;
RUN;
DATA PARMEST;
  SET PARMEST(WHERE=(PARM<>'SCALE'));
  LOW=ESTIMATE -1.96*STDERR;
  HIGH=ESTIMATE+1.96*STDERR;
  RP=EXP(ESTIMATE);
  LRP=EXP(LOW);
  URP=EXP(HIGH);
  FORMAT RP LRP URP 6.2;
RUN;

PROC PRINT DATA=PARMEST ;
  VAR PARM RP LRP URP;
  TITLE « RAPPORTS DE PREVALENCES AJUSTES »;
RUN;
```

RÉFÉRENCES

1. Bouyer J, Hémon D, Cordier S *et al. Épidémiologie. Principes et méthodes quantitatives.* Paris : Editions INSERM ; 1995.
2. Lee J. Odds ratio or relative risk for cross-sectional data? [letter] [see comments]. *Int J Epidemiol* 1994; 23: 201-3.
3. Lee J. Estimation of prevalence rate ratios from cross sectional data: a reply [letter]. *Int J Epidemiol* 1995; 24: 1066-67.
4. Osborn J, Cattaruzza MS. Odds ratio and relative risk for cross-sectional data [letter; comment]. *Int J Epidemiol* 1995; 24: 464-5.
5. Zocchetti C, Consonni D, Bertazzi PA. Estimation of prevalence rate ratios from cross-sectional data [letter; comment]. *Int J Epidemiol* 1995; 24: 1064-7.
6. Hughes K. Odds ratios in cross-sectional studies [letter; comment]. *Int J Epidemiol* 1995; 24: 463-8.
7. Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? *Occup Environ Med* 1998; 55: 272-7.
8. Zocchetti C, Consonni D, Bertazzi PA. Relationship between prevalence rate ratios and odds ratios in cross-sectional studies. *Int J Epidemiol* 1997; 26: 220-3.
9. OMS. *Utilisation et interprétation de l'anthropométrie.* Genève : OMS ; 1995.
10. Bouyer J. La régression logistique en épidémiologie. I. *Rev Epidemiol Sante Publique* 1991 ; 39 : 79-87.
11. Bouyer J. La régression logistique en épidémiologie. II. *Rev Epidemiol Sante Publique* 1991 ; 39 : 183-96.
12. Lemeshow S, Hosmer DW. *Applied Logistic Regression*: Wiley; 1989.
13. Mc Cullagh P, Nelder JA. *Generalized Linear Models.* Second Edition ed: Chapman & Hall; 1989.
14. Dobson AJ. *An introduction to generalized linear models.* London: Chapman & Hall; 1990.
15. Cook RJ. Generalized Linear Model. In: Armitage P, Colton T, editors. *Encyclopedia of Biostatistics*: Wiley; 1998. p. 1637-1650.
16. Richardson S. Développements récents de la biostatistique. *Rev Epidemiol Sante Publique* 1996; 44: 482-93.
17. SAS Institute. *SAS/STAT Software: changes and enhancements through release 6.12.* Cary N.C.: SAS Institute; 1997.
18. Simonoff JS. Logistic regression, categorical predictors and goodness-of-fit: it depends on who you ask. *The American Statistician* 1998; 52: 10-4.
19. Skov T, Deddens J, Petersen MR, Endahl L. Prevalence proportion ratios: estimation and hypothesis testing. *Int J Epidemiol* 1998; 27: 91-5.
20. Viel JF. La régression de Poisson en épidémiologie. *Rev Epidemiol Sante Publique* 1994; 42 (1): 79-87.
21. Martuzzi M, Elliott P. Estimating the incidence rate ratio in cross-sectional studies using a simple alternative to logistic regression. *Ann Epidemiol* 1998; 8: 52-5.

22. Kleinbaum DA, Kupper LL, Morgenstern H. *Epidemiologic research. Principles and quantitative methods.* New York: Van Nostrand Reinhold; 1982.

23. Dean CB. Overdispersion. In: Armitage P, Colton T, editors. *Encyclopedia of Biostatistics*: Wiley; 1998. p. 3226-3232.

24. SAS Institute. *SAS/STAT User's Guide, Version 6, Volume 1 &2*. Fourth Edition ed. Cary NC: SAS Institute Inc.; 1989.

25. SAS Institute. Categorical data analysis using the SAS system. Cary N.C.: SAS Institute; 1995.60@