JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ABDELKRIM ZEGHNOUN

Association à court terme entre l'ozone et la survenue de la toux : utilisation de modèles linéaires généralisés et leurs extensions

Journal de la société française de statistique, tome 145, n° 3 (2004), p. 81-88

<http://www.numdam.org/item?id=JSFS_2004__145_3_81_0>

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) 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/

ASSOCIATION À COURT TERME ENTRE L'OZONE ET LA SURVENUE DE LA TOUX : UTILISATION DE MODÈLES LINÉAIRES GÉNÉRALISÉS ET LEURS EXTENSIONS

Abdelkrim ZEGHNOUN *

RÉSUMÉ

Dans cette étude, nous nous intéressons à la relation à court terme entre les concentrations journalières d'ozone et la survenue de la toux dans un panel d'enfants scolarisés à Armentières, Nord-Pas-de-Calais. Pour chaque enfant, la survenue de la toux, un jour donné, est probablement liée à la présence de la toux le ou les jours précédents. L'utilisation de modèles adaptés aux données corrélées s'avère donc nécessaire. Pour cela, nous avons proposé d'utiliser un modèle markovien de régression. Les résultats observés ont été ensuite comparés à ceux obtenus en utilisant un modèle marginal et un modèle à effets aléatoires. L'association entre l'ozone et la toux est globalement similaire quel que soit le modèle utilisé. L'odds ratio (OR), pour une augmentation de 10 $\mu g/m^3$ d'ozone, est d'environ 1,14 (IC 95 % = [1,01-1,28]).

ABSTRACT

In this study, we are interested in the short-term association between daily ozone concentrations and cough episode in a panel of schoolchildren in Armentières, Northern France. Because children who experienced cough symptoms on one day were probably more likely than the others to have cough symptoms on the next day, the past history of cough symptoms might significantly influence the present occurrence of cough symptoms. To take into account the correlation between observations, we propose to use a Markov regression model. The results of this model are then compared with those obtained by using the marginal model and the random-effects model. The results suggest that the association between ozone and children's cough episodes is similar whatever the model used. A 10 $\mu g/m^3$ increase in ozone concentrations is associated with a 14 % increase in cough symptoms (OR = 1.14, CI 95 % = [1.01-1.28]).

* Institut de Veille Sanitaire/Département Santé Environnement, 12 rue du Val d'Osne, 94415 Saint Maurice
a.zeghnoun@invs.sante.fr

1. Introduction

Les méthodes d'analyse usuelles, comme la régression linéaire, supposent que les observations sont indépendantes et normalement distribuées. Bien que les *Generalized linear models* (GLM) étendent ces modèles en termes de loi de probabilité, l'hypothèse d'indépendance des observations est maintenue. Lorsque la variable d'impact sanitaire est binaire, comme c'est le cas dans cette étude, la relation entre celle-ci et les indicateurs de pollution atmosphérique est quantifiée, le plus souvent, en utilisant un modèle de régression logistique marginal prenant en compte la corrélation des observations par la méthode des *Generalized estimation equations* (GEE) (Liang et Zeger, 1986; Zeger et Liang, 1986).

Dans cette étude, nous nous intéressons à la relation à court terme entre les concentrations journalières d'ozone et la survenue de la toux dans un panel d'enfants scolarisés à Armentières, Nord-Pas-de-Calais. Pour chaque enfant, la survenue de la toux, un jour donné, est probablement liée à la présence de la toux le ou les jours précédents. L'utilisation d'un modèle de régression pour données indépendantes n'est plus adaptée et le recours à des modèles pour données corrélées est nécessaire. Différentes approches peuvent être utilisées pour modéliser les données corrélées :

- l'approche conditionnelle consiste à modéliser l'espérance conditionnelle de la réponse en fonction des variables explicatives et des réponses antérieures,
- l'approche marginale consiste à modéliser l'espérance marginale de la réponse en fonction des variables explicatives, et prendre en compte, séparément, la corrélation par l'intermédiaire de la matrice des variances-covariances de la réponse,
- la dernière approche est basée sur les modèles à effets aléatoires qui supposent qu'il existe une variabilité naturelle de la réponse des individus due à des facteurs non mesurés. Cette variabilité est alors la source de la corrélation des observations répétées sur un même individu.

Dans ce travail, nous avons proposé d'utiliser un modèle markovien de régression ou modèle conditionnel (Zeger et Qaqish, 1988). Les résultats observés ont été ensuite comparés à ceux obtenus en utilisant un modèle marginal et un modèle à effets aléatoires.

2. Matériels et méthodes

2.1. Données

Ce travail a été réalisé sur des données extraites d'un panel d'enfants recrutés parmi les 110 enfants des classes de CM1 d'une école publique d'Armentières (Declercq et Macquet, 2000). À la fin de chaque journée, les enfants notaient sur un carnet de surveillance quotidien, la présence ou non de symptômes respiratoires, en particulier, la toux. Le recueil quotidien des données a été réalisé sur une période de 3 mois, du 1er avril au 30 juin 1996.

L'indicateur de pollution étudié était l'ozone mesuré par une station fixe implantée à 850 m de l'école à une hauteur de 2 mètres. À partir des mesures horaires, le maximum journalier des moyennes glissantes sur 8 heures a été construit et utilisé comme indicateur d'exposition. Les données météorologiques (température et humidité relative) ont été fournies par Météo France.

2.2. Analyse statistique

L'hypothèse d'indépendance des observations dans les GLM est très restrictive lorsqu'il s'agit de données longitudinales où les mesures répétées sur un même sujet sont vraisemblablement corrélées. Ignorer cette corrélation intra-sujet entraîne, en général, une estimation biaisée des variances des paramètres d'intérêt et une perte d'efficacité.

L'analyse de la relation à court terme entre les concentrations journalières d'ozone et la survenue de la toux, a eu recours à trois types de modèles : un modèle de régression logistique markovien, un modèle de régression logistique marginal et un modèle de régression logistique à effets aléatoires.

Pour illustrer ces trois modèles, soient $y_{i1}, \ldots, y_{ini}$ les observations répétées représentant les données de symptômes respiratoires (présence ou non de la toux) pour l'enfant i ($i = 1, \ldots, K$). Les données sont quotidiennes avec ni observations pour l'enfant i. Pour chaque jour j ($j = 1, \ldots, ni$) on observe la réponse y_{ij} et un vecteur de variables explicatives x_{ij} de taille p (température, ozone, etc.).

Dans un modèle de régression logistique markovien, la distribution conditionnelle de chaque réponse est une fonction explicite des réponses antérieures et des variables explicatives. Soit, $H_{ij} = \{y_{ij-1}, \ldots, y_{i1}; x_{ij}, x_{ij-1}, \ldots, x_{i1}\}$ l'histoire de l'enfant i au jour j. Soit également, $\mu_{ij} = E(y_{ij}|H_{ij})$ et $\mathrm{var}(y_{ij}|H_{ij})$ les deux premiers moments conditionnels. Lorsqu'un nombre limité, q, de réponses antérieures est spécifié dans le modèle, on parle de modèle de régression logistique markovien d'ordre q. Ce modèle se présente sous la forme :

$$\mu_{ij} = \frac{\exp\left(\sum_{r=1}^{q} \psi_r y_{ij-r} + x_{ij}^T \beta^{**}\right)}{1 + \exp\left(\sum_{r=1}^{q} \psi_r y_{ij-r} + x_{ij}^T \beta^{**}\right)}$$

avec comme variance conditionnelle $\mathrm{var}(y_{ij}|H_{ij}) = \mu_{ij}(1 - \mu_{ij})$. Les paramètres à estimer sont ψ et β^{**}.

Une composante du vecteur $\beta = (\beta_1, \ldots, \beta_p)'$ mesure l'effet moyen sur l'ensemble des enfants de la variable explicative correspondante.

En ce qui concerne le modèle de régression logistique marginal, il possède les propriétés suivantes :

- l'espérance marginale de la réponse, $E(y_{ij}) = \mu_{ij}$, est liée aux variables explicatives, x_{ij}, par :

$$\mu_{ij} = \frac{\exp(x_{ij}^T \beta)}{1 + \exp(x_{ij}^T \beta)},$$

- la variance marginale est liée à l'espérance marginale par la relation : $\text{var}(y_{ij}) = \mu_{ij}(1 - \mu_{ij})$,
- la corrélation entre y_{ij} et y_{ik}, $j < k$, est une fonction de l'espérance marginale et d'un vecteur de paramètres α : $\text{corr}(y_{ij}, y_{ik}) = Ri(\alpha)$, où $Ri(\alpha)$ est une matrice de corrélation de travail de dimension $ni \times ni$, α est un vecteur de paramètres associé à un modèle spécifique de Ri, et représente la dépendance entre les mesures consécutives.

Les composantes de β s'interprètent comme dans les GLM. Elles mesurent l'effet moyen de la variable explicative correspondante sur l'ensemble des enfants.

Le modèle à effets aléatoires a d'abord été développé dans le cadre du modèle linéaire (Harville, 1977; Laird et Ware, 1982). Dans ce modèle, on suppose qu'à chaque individu i correspond un paramètre spécifique b_i relié à des variables explicatives z_{ij}. Celles-ci sont souvent un sous-ensemble des variables explicatives x_{ij}. Conditionnellement à b_i, le modèle logistique à effets aléatoires – ou modèle mixte – s'écrit :

$$\mu_{ij} = \frac{\exp(x_{ij}^T \beta^* + z_{ij}^T b_i)}{1 + \exp(x_{ij}^T \beta^* + z_{ij}^T b_i)}$$

Le modèle est dit mixte car il associe des effets fixes β, identiques pour tous les individus, et des effets aléatoires b_i, spécifiques à chaque individu. De plus, conditionnellement aux effets aléatoires b_i, les observations de chaque individu i sont supposées indépendantes. La variation inter-individuelle des effets aléatoires b_i est modélisée en supposant qu'ils sont indépendants et distribués selon une loi de distribution commune. Habituellement, une distribution gaussienne $N(0, G)$ est adoptée où G est la matrice des variances-covariances des effets aléatoires. Lorsque $z_{ij} = 1$, le modèle ne fait intervenir qu'une ordonnée à l'origine aléatoire b_i unidimensionnelle, qui peut être interprétée comme représentant l'effet de variables non mesurées propres à l'enfant.

L'interprétation des paramètres de régression β est en général différente de celle d'un modèle marginal ou conditionnel. Dans un modèle à effets aléatoires, une composante de β mesure l'effet d'un changement dans les valeurs de la variable explicative correspondante sur la réponse d'un individu et non l'effet moyen de celle-ci sur l'ensemble des enfants.

L'analyse statistique a pris en compte l'effet de la tendance, des jours de semaine et l'effet à court terme de la température et de l'humidité relative moyennes. Sur la base de critères statistiques, seules les variables météorologiques ont été retenues dans le modèle. L'estimation des paramètres a été effectuée en utilisant le logiciel SAS (SAS, 1999).

3. Résultats

L'association à court terme entre l'ozone et la survenue de la toux est présentée pour les trois modèles dans le tableau 1.

En ce qui concerne le modèle de régression logistique markovien, l'analyse exploratoire de la dépendance temporelle entre les observations successives, utilisant le critère BIC (*Bayesian Information Criterion*) pour modèles emboîtés, a permis de suggérer un modèle markovien d'ordre deux, c'est-à-dire un modèle dans lequel la survenue de la toux chez un enfant, un jour donné, est statistiquement liée à la présence de la toux les 2 jours précédents. L'association entre l'ozone et la survenue de la toux est statistiquement significative avec un OR = 1,14 (IC 95 % = [1,01-1,28]) pour un accroissement de 10 $\mu g/m^3$ des concentrations d'ozone. Cette association n'était pas modifiée par l'utilisation d'un modèle markovien d'ordre supérieur à 2.

TABLEAU 1. — Coefficients de régression estimés pour la toux en utilisant les 3 types de modèles.

Y_{ij}	**Ordonnée à l'origine**	**Ozone**	**Température**	**Humidité relative**		
	Modèle logistique markovien d'ordre 2				Y_{ij-1}	Y_{ij-2}
Toux	−7.699	0.013	−0.027	0.043	2.933	2.129
	(0.892)	(0.006)	(0.028)	(0.009)	(0.203)	(0.213)
	Modèle logistique marginal*					
Toux	−6.034	0.011	−0.011	0.036		
	(1.112)	(0.006)	(0.039)	(0.011)		
	Modèle logistique à ordonnée à l'origine aléatoire				v^{2**}	σ^{2***}
Toux	−7.294	0.012	−0.045	0.044	3.455	0.634
	(1.661)	(0.008)	(0.048)	(0.017)	(0.673)	(0.013)

* matrice de corrélation de travail autorégressive; ** variance de l'ordonnée à l'origine aléatoire; *** variance des résidus.

Ozone, température et humidité relative au retard 0-2 jours (moyenne des concentrations du jour même et deux jours précédents); (.) Écart-type.

Concernant le modèle de régression logistique marginal, l'effet de l'exposition à l'ozone sur la survenue de la toux est relativement plus faible avec un OR =1,12 (IC 95 % = [0,92-1,26]) pour un accroissement de 10 $\mu g/m^3$ des concentrations d'ozone.

Enfin, pour le modèle à ordonnées à l'origine aléatoire, l'OR pour une augmentation de 10 $\mu g/m^3$ est de 1,13 (95 % = [0,96-1,32]).

4. Discussion

Dans cette étude, une association positive a été observée entre les concentrations d'ozone et la survenue de la toux chez des enfants d'âge scolaire. Dans le modèle logistique à effet aléatoire, la valeur élevée de la variance de l'ordonnée à l'origine aléatoire montre une hétérogénéité de la survenue de la toux chez les enfants. Cette hétérogénéité peut être expliquée par des facteurs extérieurs non mesurés tels que les conditions de vie. La comparaison des résultats des trois types de modèles a montré que l'association entre l'ozone et la survenue de la toux était globalement similaire quel que soit le modèle utilisé. Elle est cependant plus significative pour le modèle logistique markovien. L'OR estimé pour une augmentation de 10 $\mu g/m^3$ était d'environ 1,14 et est proche de ceux observés dans la littérature (Schwartz *et al.*, 1994; Romieu *et al.*, 1997).

En général, les coefficients de régression d'un modèle marginal (β) et d'un modèle à effets aléatoires (β^*) ont des interprétations différentes. Une composante du vecteur β mesure l'effet moyen sur l'ensemble de la population de la variable explicative correspondante alors qu'une composante du vecteur β^* mesure l'effet d'une variable explicative sur la réponse d'un individu. En raison de ces différentes interprétations, le modèle marginal est appelé *population-averaged model* et le modèle à effets aléatoires *cluster-specific model.* Lorsque la fonction de lien n'est pas linéaire, il n'y a pas de relation simple entre β et β^*. En revanche, lorsque le lien est linéaire, les paramètres β et β^* sont égaux. Neuhaus *et al.* (1991) ont montré que si, dans un modèle logistique avec ordonnée à l'origine aléatoire, la variance de l'ordonnée à l'origine aléatoire est strictement supérieure à zéro ($v^2 > 0$) alors $|\beta| < |\beta^*|$; l'égalité est obtenue si et seulement si $\beta^* = 0$. Ils ont montré également que la différence entre β et β^* augmente avec la variance de l'effet aléatoire. Les paramètres de régression β^* et la variance de l'effet aléatoire peuvent être estimés, dans le cas linéaire, en utilisant les GEE. L'estimateur de β^* est convergent même si la matrice des variances-covariances n'est pas correctement spécifiée. Pour d'autres fonctions de lien, la méthode du maximum de vraisemblance est généralement employée en supposant des effets aléatoires gaussiens.

Concernant le modèle conditionnel, les paramètres de régression (β^{**}) mesurent l'effet moyen des variables explicatives sur l'ensemble de la population conditionnellement aux réponses antérieures. Dans le cas linéaire, les paramètres de régression du modèle conditionnel ont la même interprétation que ceux du modèle marginal ou du modèle à effets aléatoires alors que, dans le cas non linéaire, des différences entre ces paramètres peuvent apparaître. La relation entre β et β^{**} a été discutée dans certains cas par Bishop *et al.* (1975) et Zeger et Liang (1992). En général, l'estimation des paramètres d'un modèle conditionnel peut être effectuée par la méthode du maximum de vraisemblance conditionnelle en traitant les réponses antérieures comme des variables explicatives additionnelles.

L'estimation des modèles marginaux nécessite peu d'hypothèses, particulièrement lorsque l'intérêt de l'étude porte sur les paramètres de régression β. En effet, une spécification correcte de l'espérance marginale suffit pour garantir la convergence de l'estimateur de β (Liang et Zeger, 1986). En revanche,

pour les modèles à effets aléatoires, l'estimation des paramètres nécessite la spécification de la distribution conditionnelle de la variable dépendante. La loi de l'effet aléatoire doit également être spécifiée. En effet, l'inférence sur les paramètres de régression β^* dépend de ce choix. Cependant, lorsque l'intérêt de l'analyse porte essentiellement sur l'estimation des paramètres de régression, le choix de la loi de l'effet aléatoire n'est pas décisif (Neuhaus *et al.*, 1992). Concernant les modèles conditionnels, la convergence des estimateurs du maximum de vraisemblance est fondée sur une spécification correcte de la distribution conditionnelle de la réponse ou au moins des deux premiers moments conditionnels. Dans le cas des modèles de quasi-vraisemblance, la convergence des estimateurs est assurée lorsque la moyenne conditionnelle est correctement spécifiée en fonction des variables explicatives et des réponses antérieures. Comme pour les données indépendantes, une fonction de variance de travail peut être utilisée, mais il est nécessaire de modéliser correctement la dépendance temporelle pour obtenir des estimateurs convergents.

Le choix entre modèle marginal, conditionnel, ou à effets aléatoires doit être guidé par la formulation du problème scientifique qui a justifié l'étude. Lorsque le raisonnement s'effectue à l'échelle de la population, comme c'est souvent le cas en épidémiologie, il semble plus naturel de quantifier les effets de facteurs de risque, tels que la pollution atmosphérique, à l'aide d'odds ratio marginaux ou conditionnels. En revanche, lorsque le raisonnement se fait au niveau des individus, les modèles à effets aléatoires trouvent un intérêt pratique. En ce qui concerne les modèles conditionnels, il est essentiel de vérifier la sensibilité des paramètres de régression β^{**} au modèle supposé pour la dépendance temporelle même lorsque l'intérêt de l'analyse porte essentiellement sur l'estimation des paramètres de régression β^{**}.

Références

BISHOP Y.M., FIENBERG S.E. and HOLLAND P.W. (1975). *Discrete Multivariate Analysis : Theory and practice.* MIT Press, Cambridge, MA.

DECLERCQ C. and MACQUET V. (2000). Effets à court terme de l'ozone sur la santé respiratoire d'enfants d'Armentières, Nord de la France. *Rev. Epidémiol. Santé Publique* **48** 37-43.

HARVILLE D. (1977). Maximum likelihood approach to variance component estimation and related problems. *J. Am. Statist. Assoc.* **72** 320-39.

LAIRD N. and WARE J.H. (1982). Random effects models for longitudinal data. *Biometrics* **38** 963-74.

LIANG K.Y. and ZEGER S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13-22.

NEUHAUS J.M., HAUCK W.W. and KALBFLEISCH J.D. (1992). The effect of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79** 755-62.

NEUHAUS J.M., KALBFLEISCH J.D. and HAUCK W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int. Stat. Rev.* **59** 25-35.

ROMIEU I., MENESES F., RUIZ S., HUERTA J., SIENRA J.J., WHITE M., ETZEL R. and HERNANDEZ M. (1997). Effects of intermittent ozone exposure on peak expiratory flow and respiratory symptoms among asthmatic children in Mexico city. *Arch. Environ. Health* **52** 368-76.

SAS Institute Inc. (1999). *SAS/STAT User's Guide - Version 8.* SAS Institute Inc, Cary, NC.

SCHWARTZ J., DOCKERY D.W., NEAS L.M., WYPIJ D., WARE J.H., SPENGLER J.D., KOUTRAKIS P., SPEIZER F.E. and FERRIS B.G. Jr. (1994). Acute effects of summer air pollution on respiratory symptom reporting in children. *Am. J. Respir. Crit. Care Med.* **150** 1234-42.

ZEGER S.L. and LIANG K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121-30.

ZEGER S.L. and LIANG K.Y. (1992). An overview of methods for the analysis of longitudinal data. *Statist. Med.* **11** 1825-39.

ZEGER S.L. and QAQISH B. (1988). Markov regression models for time series : a quasi-likelihood approach. *Biometrics* **44** 1019-31.