Epidémiol. santé anim., 1997, 31-32

GENERALIZED LINEAR MIXED MODELS APPLIED TO LARGE-SCALE SALMONELLA DATA FOR DANISH BROILER FLOCKS

Chriél M.[1], Stryhn H.[2], Dauphin G.[3]

Dans cet article, nous présentons une étude épidémiologique rétrospective sur les facteurs de risque associés à Salmonella typhimurium dans des élevages de poulets danois. Les données sur l'occurrence de salmonelles dans 3.500 élevages (en 1994) ont été extraites à partir d'une large base de données Ante Mortem. Des informations sur chaque élevage ont été collectées par le vétérinaire avant l'abattage. D'autres informations supplémentaires sur les stocks parentaux, les éclosoirs et sur la fabrication des aliments ont été également recueillies. Les données sont analysées à l'aide des modèles linéaires généralisés mixtes (GLMM) avec un ou plusieurs effets aléatoires.

Nous avons essayé d'autres modèles similaires aux GLMM : le modèle logistique binomial (EGRET), des modèles utilisant la quasi-vraisemblance (Macro SAS glimmix, MLn), des modèles utilisant l'approche GEE et enfin l'approche bayesienne avec la méthode d'échantillonnage de Gibbs. Nous avons comparé les résultats et tiré des conclusions sur les avantages et les inconvénients de ces différents modèles.

INTRODUCTION

In Denmark the occurrence of different types of salmonella in chicken has been a cause of concern of the public health and of the producers in later years. In 1989 the Danish Poultry Council initiated a programme to monitor the occurrence of salmonella which led to the establishment of the country-wide ante-mortem database with data collected by the ante-mortem veterinarians, the slaughterhouses and the National Veterinary Laboratory. A detailed description of the Danish production and control system is given in Angen et al. (1996), a study on Salmonella enterica infections in the years 1992-93.

The present paper deals with occurrence of Salmonella typhimurium in 1994, and our main emphasis here is on illustration and comparison of statistical models with random effects accessible for practical analysis. The data comprise after some data reduction 3356 broiler flocks, the epidemiological units. From each flock a sample was taken to determine its salmonella status. Further information about each flock and the production facilities have been collected by the veterinarians prior to slaughtering, and also information is available about the parent stocks, the hatcheries and the feed mills. The flocks were housed in 631 production units (houses) distributed among 294 farmers. The data structure suggests therefore a hierarchical model with (at least) two levels of random effects, the houses and the farmers. However, our analysis showed only the variation between farmers to be of importance, and we present here — in order to include a wider range of statistical approaches — results for models with only one random component.

The risk factors of the study were in the main the same as in Angen et al. (1996); we refer to this paper for details and restrict ourselves here to those variables that turned out significant in the analysis. For each flock the variables were recorded as follows: the occurrence of Salmonella typhimurium in the parent flock (0=absent, 1=present), the hatchery (1,2,3), the use of (unspecified) medicine (0=not recorded, 1=no, 2=yes), the season of hatching (1=Jan–Mar, 2=Apr–Jun, 3=Jul–Sep, 4=Oct–Dec), and the sampling procedure which was altered by the National Veterinary Laboratory during the study period (0=old, 1=new). In brevity, in autumn 1994 a sampling scheme involving 60 faecal samples from each flock replaced the previous ante-mortem analysis of 16 chicks per flock, the recent method being cheaper and more sensitive to low prevalent Salmonella typhimurium infection.

STATISTICAL MODELS AND METHODS

The response y_i for each flock (i) is the salmonella status, infected $(y_i = 1)$ or non-infected $(y_i = 0)$,

[1] The Royal Veterinary and Agricultural University, Department of Animal Science and Animal Health and Department of Mathematics and Physics, Thorvaldsensvej 40, DK-1871 Copenhagen, Denmark
[2] The Danish Veterinary Laboratory, B\"{u}lowsvej 27, DK-1790 Copenhagen V, Denmark
[3] Ecole Nationale Veterinaire de Nantes (E.N.V.N.), 44087 Nantes Cedex 03, France

as determined by the sampling procedure. In generalized linear models the probability $p_i = P(y_i = 1)$ is related to the explanatory variables via the link function, which in our context is (taken as) the logit function, that is, $\text{logit}(p) = \ln(p/(1-p))$. If X is the design matrix (an $(n \times p)$-matrix where n is the number of observations and p is the number of parameters), and β is the parameter vector for the explanatory variables, the relationship takes the form

$$\text{logit}(p_i) = \sum_{j=0}^{p} x_{ij}\beta_j, \quad \text{or in condensed notation: } \text{logit}(p) = X\beta\,. \tag{1}$$

An additional assumption of independent outcomes y_i leads to a standard logistic regression model for which maximum likelihood (ML) estimation is feasible. Also, we can incorporate an overdispersion parameter into the model to compensate for the fact that the variation in the data may be larger than the binomial variation, and estimate parameters by the quasi-likelihood method (McCullagh & Nelder, 1989).

Overdispersion is often caused by clustering, and if potential clusters are known it is usually of interest to model their effect more explicitly as random effects, in a generalized linear mixed model (GLMM). There is an exhaustive and rapidly developing literature on GLMM's; some recent reviews are Breslow & Clayton (1992) and Lee & Nelder (1996). In this paper we compare five models of GLMM type. All analyses can be carried out using statistical software packages.

Addition of a random effect of farmers to model (1) yields

$$\text{logit}(p_i) = \sum_{j=0}^{p} x_{ij}\beta_j + \sigma\, u_{\text{farm}(i)}\,, \tag{2}$$

where $\sigma > 0$ and $u_1, \ldots, u_{294}$ are random variables describing the effects of individual farmers. Usually these are assumed to be Gaussian $N(0,1)$. ML estimation is computationally demanding because the likelihood function involves an integration over the random effects which must be approximated numerically. Therefore it has been suggested to replace the Gaussian distribution by a discrete distribution, namely the binomial distribution. The program EGRET (SERC, 1989) gives a ML analysis of this model.

Alternatively, analysis of model (2) can be based on a weighted least squares principle, where the equations used often are given a pseudo- or quasilikelihood motivation. Many variants — and acronyms — exist and there seems to be no clear answer to which one is preferable, neither on theoretical or practical grounds. In practice, the results are often quite similar. We have used the PQL procedure of Breslow & Clayton (1992) which has been implemented in the SAS macro glimmix (Littell et al., 1996).

Random effects models like (2) are sometimes termed subject-specific because they contain random terms for each "individual" (farmer) of the population, as opposed to marginal or population-averaged models for probabilities p_i averaged over the population. The difference between the two types of models can maybe most easily be understood by the different interpretation of the regression coefficient β for an explanatory variable, e.g. the use of medicine. In the subject-specific model β describes the effect of medicine on the probability of salmonella infection for individual flocks and farmers, whereas in the population-averaged model it describes the effect on the salmonella infection rate for the entire population. Which of the two approaches is more appropriate depends on the context. We have analysed the present data by the (marginal) MQL approach using the MLn program developed by Goldstein and co-workers (Goldstein, 1995). It should be noted that this program, as well as the glimmix macro, allow for both MQL and PQL estimation.

A related marginal approach based on generalized estimating equations (GEE; Liang & Zeger, 1986) allows for dependence within clusters without assuming it of a specific form as in (2). It has recently been implemented in the SAS system (vers. 6.12) and is also available in the statistical package S–Plus.

Finally, we have tried a Bayesian approach, where parameters are not treated as unknown constants but as themselves random variables, involving Gibbs sampling from the posterior distribution (Gilks et al., 1993) using the BUGS and CODA programs.

RESULTS

Table 1 below gives parameter estimates with standard errors for the same modelling of explanatory variables in the five models described above. For comparison all analyses are without an additional overdispersion parameter (available in SAS–glimmix and in MLn), except for GEE where its estimated value was very close to one. Estimates in SAS–glimmix indicated underdispersion which we tend to regard as an artifact. Noninformative prior distributions were used for the Bayesian model.

The interaction between the factors Hatchery and Medicine caused estimation problems with all methods except GEE due to two cells in their cross-table without cases. Parameter estimates on the logistic scale were negative and large, and the standard errors were clearly useless and have been omitted. In the EGRET and MLn programs adjustments to the convergence criteria were necessary for convergence,

and the glimmix macro failed to converge unless large negative values were forced into the corresponding cells using the offset option.

We focus here on comparing the methods and leave the epidemiological interpretation, to be reported elsewhere, largely for the reader. In the table, regression parameter estimates are in good agreement between the methods. On the other hand, the estimates of the random effect differ somewhat. The GEE method is clearly less influenced by the cells without cases than the other methods. Also the standard errors are generally slightly larger, indicating a minor loss of precision (Liang & Zeger, 1986).

Our experience of the reliability of the estimation procedures is that the GEE and Bayesian procedures are more stable than the others. We found, however, the recent version of glimmix (vers. 6.12) more stable than previous ones. For even larger data sets the MLn program and GEE method seem most promising. As to the flexibility of the models and software, the MLn and BUGS programs can handle hierarchical models with several random effects. Properly speaking, this applies to the glimmix macro also, however at the present data size the computational demands seem disencouraging. The GEE method and the implementation of ML estimation in EGRET are limited to a single random component. A final remark is that a comparison of the methods should also include statistical testing, e.g. accuracy and power of tests.

Table 1: Parameter estimates ± standard errors for different statistical analyses of salmonella data (see text). Values for Bayesian analysis are means and standard deviations from a Gibbs sample of size 2000 after a burn-in of 500 iterations. ([1] glimmix macro, [2] genmod procedure.)

		Statistical method and program used				
Effect	Level	ML EGRET	PQL SAS[1]	MQL MLn	GEE SAS[2]	Bayesian BUGS
Intercept		-2.26 ± 0.16	-2.15 ± 0.15	-2.11 ± 0.15	-2.13 ± 0.16	-2.27 ± 0.16
Season	1	-0.11 ± 0.16	-0.11 ± 0.16	-0.11 ± 0.16	-0.10 ± 0.18	-0.12 ± 0.16
	2	-0.25 ± 0.17	-0.24 ± 0.16	-0.24 ± 0.16	-0.23 ± 0.20	-0.25 ± 0.15
	3	-1.17 ± 0.31	-1.14 ± 0.31	-1.13 ± 0.30	-1.09 ± 0.30	-1.17 ± 0.30
Sampling	1	2.61 ± 0.32	2.53 ± 0.31	2.49 ± 0.31	2.47 ± 0.30	2.63 ± 0.31
Medicine	0	-0.70 ± 0.31	-0.68 ± 0.31	-0.67 ± 0.31	-0.69 ± 0.35	-0.74 ± 0.29
	2	-1.47 ± 0.34	-1.43 ± 0.33	-1.41 ± 0.34	-1.49 ± 0.42	-1.54 ± 0.34
Parent-Typh	1	0.86 ± 0.18	0.83 ± 0.17	0.82 ± 0.17	0.82 ± 0.20	0.87 ± 0.18
Hatchery	1	0.59 ± 0.17	0.56 ± 0.17	0.55 ± 0.17	0.57 ± 0.18	0.58 ± 0.17
	2	-2.03 ± 0.33	-1.98 ± 0.33	-1.97 ± 0.33	-1.95 ± 0.48	-2.11 ± 0.33
Med ×ParT	(0,1)	0.12 ± 0.32	0.11 ± 0.32	0.11 ± 0.32	0.11 ± 0.37	0.10 ± 0.32
	(2,1)	1.48 ± 0.58	1.44 ± 0.57	1.42 ± 0.58	1.49 ± 0.66	1.52 ± 0.60
Med ×Hat	(0,1)	1.18 ± 0.38	1.14 ± 0.37	1.13 ± 0.37	1.14 ± 0.45	1.24 ± 0.36
	(0,2)	-16.2	-10.6	-15.8	-3.58 ± 0.82	-78.6
	(2,1)	-0.25 ± 0.57	-0.24 ± 0.56	-0.24 ± 0.57	-0.23 ± 0.65	-0.24 ± 0.58
	(2,2)	-14.9	-9.4	-14.9	-2.36 ± 0.65	-78.7
σ		0.61 ± 0.10	0.34	0.35 ± 0.10	—	0.63 ± 0.11

REFERENCES

Angen, Ø., Skov, M. N., Chriél, M., Agger, J. F., Bisgaard, M., 1996. A retrospective study on salmonella infection in Danish broiler flocks. Prev. Vet. Med. 26, 223–237.

Breslow, N. E., Clayton, D. G., 1992. Approximate inference in generalized linear mixed models. J. Am. Statist. Ass. 88, 9–25.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N., McNeil, A., Sharples, L., Kirby, A., 1993. Modelling complexity: applications of Gibbs sampling in medicine. J. R. Statist. Soc. B 55, 39–102.

Goldstein, H., 1995. Multilevel Statistical Models. Arnold, London.

Lee, Y., Nelder, J. A., 1996. Hierarchical generalized linear models. J. R. Statist. Soc. B 58, 619–678.

Liang, K.-Y., Zeger, S. L., 1986. Longitudinal data analysis using generalized linear models. Biometrika, 13–22.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., 1996. SAS Systems for Mixed Models. Cary, NC: SAS Institute Inc.

McCullagh, P., Nelder, J. A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall, London.

SERC, 1989. EGRET's User Manual. Seattle.