VHM 812/802: Analysis of Clustered Binary Data Exercise

Data

These human health data are in the public domain after having been analyzed in multiple statistical papers and books (Zeger and Karim (1991), *J. Amer. Statist. Assoc.* **86**, 79-86; Diggle et al. (2002), *Analysis of Longitudinal Data*, 2nd ed.). The dataset "xero.dta" comprises 275 Indonesian children, a subset of the cohort studied by Sommer et al. (1984), *Amer. J. Clin. Nutrition* **40**, 1090-1095. These preschool children were examined at up to six consecutive quarters for the presence of respiratory infection, yielding a total of 1200 observations. The subset of predictors included here are: gender; age; age at first visit (examination); presence/absence of xerophthalmia, an ocular manifestation of chronic vitamin A deficiency; visit no. and the corresponding season. We may assume that a question of primary interest is whether the prevalence of respiratory infection is higher among children who suffer xerophthalmia, and that it is also of interest to describe how the prevalence changes with age.

Variable	Description	Range
child	Child identification	(nominal)
resp	Respiratory disease indicator	0/1
age	Age (in months)	4-86
sex	Child's gender	0=boy; 1=girl
xero	Xerophthalmia indicator	0/1
time	Visit/examination number	1-6
seas	Season of visit	1-4; 1=summer, ..., 4=spring
age0	Age at first visit (in months)	4-80

Questions

Use the data and the information above to go through the analytical steps below. Note that all predictors should be considered of potential interest, even if including them all simultaneously may be impossible.

1. Identify the data structure, and if relevant represent it by a hierarchical diagram including both the units and the predictors of the study. For each hierarchical level of the model, determine the number of units and describe the replication within the level.
2. Use preliminary analyses (without accounting for the data structure) to make decisions about the modeling of predictors, in light of the research questions. Use standard model-building techniques to arrive at and fit a first model of interest.
3. Adjust the standard errors of the regression coefficients by using robust variance estimates. Which predictors are most strongly affected by this? -- try to explain your findings.
4. Would it be possible/sensible to adjust the analysis for the data structure by adding suitable fixed effects? Explain your answer.
5. Adjust the analysis for the data structure by fitting a suitable random effects (mixed) logistic model. What changes do you notice in estimates and SEs? -- try to explain your findings. Optionally, interpret the variance estimate in terms of an (approximate) ICC.
6. Alternatively, use GEE estimation to account for the data structure. Try at least a couple of different working correlation structures, and notice their impact on results. Compare with the results in previous questions, and draw conclusions.