Long assignment 2

Stefan Velev

2022-12-08

Gee model

What we have want to do here is evaluate a marginal model with correlated and repeated ordinal responce variable. To that end we chose to implement a GEE model with local odds ratios `Touloumis et al. (2013)`. In a normal GEE framework we run into problems due to a lack of a convenient multivariate distribution for multinomial responses, as well as the sensitivity of the maximum likelihood method to misspecification of the association structure. This gave rise to a modification of the GEE method by `Liang and Zeger (1986)` in order to be able to account for multinomial responses `(Miller et al. 1993; Lipsitz et al. 1994;Williamson et al. 1995; Lumley 1996; Heagerty and Zeger 1996; Parsons et al. 2006)`. These approaches to GEE solve the same set of estimating equations as `Liang and Zeger (1986)` but diverge in the way they estimate or parametrize α, that is the vector describing the "working" assumption about the asossiation structure. `Touloumis et al. (2013)` demonstrate that the joint existence of the estimated marginal regression parameter vector and $\hat{\alpha}$ cannot be assured in existing approaches because the parametric space of the proposed parameterizations of the association structure depends on the marginal model specification. To address this issue,`Touloumis et al. (2013)` define α like a "nuisance" parameter vector that contains the marginalized local odds ratios structure. In essence these are the local odds ratios as if no covariates were recorded, and they employe a family of association models `(Goodman 1985)` to develop parsimonious and meaningful structures regardless of the response scale. This in practice makes their approach applicable both for ordinal and nominal multinomial responce variables without beeing restricted by the specification of the marginal model itself.
We conduct the GEE analysis in two steps. First we selected a the structure for the marginalized local odds ratios. The full specification is given by

$$\log \theta_{tjt'j'} = \phi^{(t,t')}(\mu_j^{t,t'} - \mu_{j+1}^{t,t'})(\mu_{j'}^{t,t'} - \mu_{j'+1}^{t,t'})$$

where $\{\mu_j^{t,t'}; j = 1 \dots J\}$ are the score parameters for the J response at the time pair $\{t, t'\}$ and $\phi^{(t,t')}$ is the intrinsic parameter. We nevertheless chose to select a uniform structure i.e. just a fixed ϕ given that when we ran both a categorically exchangible structure (i.e $\phi^{\{t,t'\}}$) and a time exchangible structure (i.e.$\phi(\mu_{j'} - \mu_{j'+1})$), they both gave estimates very close to contant. The final log local odds ratios look something like this

$$\log \theta_{tjt'j'} = \begin{pmatrix} 0 & 0 & \phi & \phi & \cdots & \phi & \phi \\ 0 & 0 & \phi & \phi & \cdots & \phi & \phi \\ \phi & \phi & 0 & 0 & \cdots & \phi & \phi \\ \phi & \phi & 0 & 0 & \cdots & \phi & \phi \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi & \phi & \phi & \phi & \cdots & 0 & 0 \\ \phi & \phi & \phi & \phi & \cdots & 0 & 0 \end{pmatrix}$$

Where the matrix is dimesions of the per subject observations and the 4 zero block on the diagonal comes up because at each time step we have both a left and a right ear. The side of the ear comes up non-signifacnt.

The next step is model selection which we do with the help of Wald test. We conducted a greedy model selection (both forward by starting with a minimal model and adding variables, and backwards by starting with a full model and removing variables) which fortunately converged. The resulting model was the same as in the model in the previous assignment with the difference having an additional $TIME^2$ term instead of just $TIME$. Given that performance was almost identical we decided to chose a more parsimonious model which is the same as the model we chose in the previous assignment which end up being $hearing_{\{discrete\}} = age \times TIME + age^2 + learning$

It is easy to see that young people are more likely to have Excellent or Normal hearing. Further as people age they are more likely to have Normal hearing and hearing loss becoming more and more likely the older they are. If it is the first time a person is performing this test they are slightly more likely to be categorized in a lower group, where it is "Excellent" is on average less likely, "Hearing loss" is on average more likely, and for "Normal" depends on the age the subject is first measured.

Mixed model

In this context we evaluate a Cumulative link mixed model with a random effect for the individual subject. First lets take a look at the Cumilative Link Model (CML), it is designed to take ordinal response data where , for each level j of the ordinal response,the cumulative probability of being in level j or lower is modeled. CLM models take the following general form:

$$G^{-1}[P(Y \leq j)] = a_j - X\beta$$

where X represents the model matrix, β the vector of true coefficients for each regressor as well as its intercept a_j, the threshold for level j,where $j = 1, \dots J$ for an ordinal variable with J levels, and G^{-1} is the link function. We can interpret a_j and G^{-1} is by considering Y as if it had come from a continious latent variable Y^*. The CLM is then equivalent to an ordinary least squares model where a_j represent cut of points in Y^* which separate the levels of Y and the link function as is the inverse cumilative density of Y^*. However when we have multiple observations per individual across time we violate the assumption of independence. In order to account for dependent observations a random effect can be added to the previous model. Cumulative link mixed models have the following general form:

$$G^{-1}[P(Y \leq j)] = a_j - (Z_{t[i]}u_t + X_i\beta)$$

where $u_t \sim \mathcal{N}(0, \sigma_2^2)$. In this notation, u_t represents the vector of coefficients corresponding to the group-level predictors $Z_{t[i]}$ for observation i in cluster t. This model has the added assumption that the random effects are Normally distributed and centered at zero. The random effect induces the correlation expected between observations in the same cluster and allows inferences to be made to the population from which the groups were sampled. It should be noted that model estimates can be unstable if there are a small number of observations within clusters or if there are few clusters from which to estimate within group correlation. In the case for our data we use the the logit, or log odds, link function, which is the inverse cumulative density function of a Logistic probability distribution. When using the logit link function, CLM models are more commonly referred to as proportional odds models

Here we again use a greedy approach and for the model selection criterion we use the AIC and come up with the following specification $hearing_{\{discrete\}} = age \times TIME + TIME + age^2 + learning$. As this is a subject specific model in order to get concrete interpretations we need to provide subject specific information (e.g. age), however we feel that the plot bellow sums up our results rather succinctly. We see that (as in the other model) as people age they are less likely to categorize as Excellent hearing, and more likely to have Normal hearing. Then as people approach their 50s they start having a positive probability of suffering from hearing loss. It is important to note that there is a difference between the interpretation of these two models. The marginal model is more useful when we are interested in results pertaining to the entire population, while mixed models are more suited to answering question regarding an individual in said population. The choice of model will therefore be motivated by the research question.