Quantile regression for dependent data using a working odds ratios matrix

D. Bossoli[1*], M. Bottai[2]

[1] *Department of Statistical Sciences, University of Padua, Padova ; bossoli@stat.unipd.it.*
[2] *Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm; matteo.bottai@ki.se*
[*] *Presenting author*

Keywords. *Quantile regression; Dependent data; Generalized estimating equations; Working odds ratios matrix*

1 Abstract

Dependent data arises frequently in applied research. A common solution to adjusting for correlation among observations within clusters is to use generalized estimating equations. When the interest lies on quantiles of the conditional distribution of an outcome variable of interest, the working correlation matrix is no longer computed from the regression residuals but instead from the residuals' sign. Because correlation between binary variables is bounded, we propose an unconstrained alternative, the odds ratio. In addition to computational advantages, odds ratios allow flexible modeling. Different working structures can be estimated easily by appropriate logistic regression models. Simulations show similar results to generalized estimating equations applied to regression for the mean. We illustrate the proposed method with data from a randomized trial on cognitive behavior therapy for treatment of obsessive compulsive disorder.

2 Introduction and Method

Longitudinal and clustered data represent two frequent analytical situations in which observations within a cluster are not independent. In these cases, the statistical assumption about independent observations of traditional regression models is violated. Ignoring the dependency between observations within the same cluster generally leads to unbiased estimators but wrong standard errors. Generalized estimating equation (GEE) is a population-averaged (or marginal) method used to account for

the dependency induced by the clustering. The dependency between observations within the same cluster is modeled through a correlation matrix, which is usually considered the same for all clusters. In the literature this matrix is called working correlation matrix, because estimated parameters and their standard errors are correct even if the correlation matrix is misspecified.

Let $\{y_{ij}, x_{ij}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, T$ be the longitudinal data set, where $y \in \mathbb{R}$ is the response variable and $x \in \mathbb{R}^P$ is the covariate vector.
Consider the set of repeated measurement of the i-th individual, denoted by $y_i = (y_{11}, \ldots, y_{1T})$, and its design matrix $x_i = (x_{11}, \ldots, x_{1T})$. Marginal quantiles can be obtained solving the following estimating equation: (**?**)

$$\tilde{U}_Q(\beta) = \sum_{i=1}^{n} x_i^T W_i^{-1}(\eta) \tilde{\psi}_\tau(\epsilon_i) = 0,$$

where $\tilde{\psi}_\tau(\epsilon_i) = \left(1 - \Phi\left(\frac{y_{i1} - x_{i1}^T\beta}{r_{i1}}\right), \ldots, 1 - \Phi\left(\frac{y_{iT} - x_{iT}^T\beta}{r_{iT}}\right)\right)^T$, $\Phi(\cdot)$ is the standard normal cumulative distribution, $r_{ik} = (x_{ik}^T \Omega x_{ik})^{1/2}$, Ω is a smoothing parameter and $W_i(\eta)$ is the working correlation matrix. The main difference between mean and quantile GEE is related to the estimation of the working correlation matrix $W_i(\eta)$. In the quantile approach, it is estimated from the regression residuals' signs. However, correlation is not a good measure of dependency between binary variables because it is bounded by their marginal frequencies. An unconstrained alternative is given by the odds ratio. Let $S_{it} = I(y_{it} \leq x_{it}^T \beta_\tau)$ be the residual sign of the i-th individual at time t and $S_t = (S_{1t}, \ldots, S_{nt})$ be the set of residual signs at time t. The odds ratio between S_z and S_u is obtained by

$$\eta_{zu} = \frac{P(S_z = 1, S_u = 1)/P(S_z = 0, S_u = 0)}{P(S_z = 0, S_u = 1)/P(S_z = 1, S_u = 0)}.$$

Let $\mathcal{A} = \{(S_z, S_u),\ z = 1, \ldots, T,\ u = 1, \ldots, T,\ z > u\}$ be the set of pairwise comparisons required to estimate all the odds ratios of the working correlation matrix $W_i(\eta)$. Consider the augmented dataset (V_z, V_u, z, u, c), where $(V_z, V_u) = \{(S_z, S_u) \in \mathcal{A}\}$ and $c = 1, \ldots, \binom{T}{2}$ indicates the pairwise comparison between V_z and V_z . For any working structure of $W_i(\eta)$, the respective set of odds ratios can be estimated simultaneously through an appropriate choice of the linear predictor in a logistic regression of V_z on V_u, $V_z | V_u \sim Be(\mu)$:

- Exchangeable: $logit(\mu) = \alpha + \eta V_u$;
- Toeplitz: $logit(\mu) = \alpha + \sum_{i=1}^{T-1} \eta_i I_{z-u=i} V_u + \sum_{i=1}^{T-1} I_{z-u=i}$;
- Unstructured: $logit(\mu) = \alpha + \sum_{i=1}^{\binom{T}{2}} \eta_i I_{c=i} V_u + \sum_{i=1}^{\binom{T}{2}} I_{c=i}$.

We illustrate the proposed method with data from a randomized trial on cognitive behavior therapy for treatment of obsessive compulsive disorder.

References

Fu, L. & Wang, Y.G. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis*, **8**, 2526–2538.