Two-stage estimation and composite likelihood in the Poisson correlated Gamma-frailty model

Marta Fiocco, Hein Putter, Hans van Houwelingen

Leiden University Medical Center, The Netherlands

Warwick workshop on composite likelihood methods April
15-17 April 2008

Longitudinal count data | Poisson correlated Gamma frailty | Composite likelihood and two-stage estimations | Application

Outline

Two-stage estimation and composite likelihood in the Poisson correlated Gamma frailty model M. Fiocco

Recurrent event data

- In medical studies subjects can experience recurrent or repeated events
- This implies correlation of event times within an individual

Statistical models

- Two approaches:
  - Marginal approach: dependence between recurrent events is seen as nuisance
  - Frailty approach: model explicitly the correlation
- Model the event occurrences through counts
- Or number of events over period of time

Poisson correlated Gamma frailty model

- Vector of event counts $Y = (Y_1, \ldots, Y_T)$
- General formulation: $Z = (Z_1, \ldots, Z_T)$ multivariate gamma frailty
  - Z_t mean one, variance ξ
  - Correlation between Z_s and Z_s equals $\rho^{|s-t|}$
- $Y_1, \ldots, Y_T$ are assumed conditionally independent given the frailties, with

$$Y_t \mid Z_t \sim \mathrm{Po}(\mu_t Z_t) \ ,$$

with

$$\mu_t = \exp(\mathbf{x}_t^\top \beta) \ .$$

Properties of Poisson correlated Gamma frailty

- Marginal $Y_t \sim \mathrm{NB}(\mu_t, \theta)$
- $EY_t = \mu_t$; $\mathrm{var}\, Y_t = \mu_t + \mu_t^2 \xi$
- Full joint distribution can be derived in theory from Laplace transform

$$\mathrm{P}(Y_1 = y_1, \ldots, Y_T = y_T) = \left(\prod_{t=1}^{T} \frac{\mu_t^{y_t}}{y_t!} \right) \cdot \mathrm{E}\{Z_1^{y_1} \ldots Z_T^{y_T} \exp(-\boldsymbol{\mu}^\top Z)\} \ ,$$

but is intractable in practice

Estimation

- Henderson & Shimakura (2003) proposed to maximize the composite log-likelihood

$$\sum_{i=1}^{N} \sum_{1 \leq s \leq t \leq T} \log P(Y_s = y_s, Y_t = y_t)$$

jointly over all parameters $\boldsymbol{\beta}, \theta, \rho$

- Two problems occurred when we tried to implement this
  - Serious rounding errors for high counts
  - Maximization over large number of parameters
  - (No flexible software available)

Estimation

- A further rearrangement is possible when counts are very high
- In our case this arrangement does not work
- Normal approximation in case of high counts is not good!

Solutions

- Rounding errors:
  - Alternative multivariate Gamma distribution: to be used as frailty vector in a Poisson model for longitudinal count data.
  - The new Gamma distribution is based on renewal processes.
- Dimension problem: composite likelihood still entails a high-dimensional maximization problem
  - Two-stage approach

New proposal

Resulting distribution of Y

- Marginal still $\mathrm{NB}(\mu_t, \theta)$
- Full joint distribution still intractable
- Pairwise distribution:

$$
\begin{aligned}
P(Y_s = y_s, Y_t = y_t) &= \frac{\mu_s^{y_s}\mu_t^{y_t}}{y_s!y_t!} \times E(Z_s^{y_s} Z_t^{y_t} e^{\mu_s Z_s} e^{\mu_t Z_t}) \\
&= \sum_{k=0}^{y_s}\sum_{l=0}^{y_t} E\left(e^{-\mu_s X_s - \mu_t X_t - (\mu_s+\mu_t)X_0} \cdot \frac{X_s^k X_0^{y_s-k}}{k!(y_s-k)!} \frac{X_t^l X_0^{y_t-l}}{l!(y_t-l)!} \mu_1^{y_s}\mu_t^{y_t} \right) \\
&= \sum_{k=0}^{y_s}\sum_{l=0}^{y_t} P_{NB}(k; \mu_s(1-\rho_{st}), \xi) \cdot \mathrm{P}_{\mathrm{NB}}(l; \mu_t(1-\rho_{st}), \xi) \cdot \\
&\qquad \mathrm{P}_{\mathrm{NB}}(y_s + y_t - k - l; (\mu_s+\mu_t)\rho_{st}, \xi) \cdot \\
&\qquad \mathrm{P}_{\mathrm{Bin}}(y_s - k; y_s + y_t - k - l, \frac{\mu_s}{\mu_s+\mu_t})
\end{aligned}
$$

Advantages

- No rounding errors, because the terms contributing to the sum are all products of probabilities, hence between 0 and 1
- It is possible to generate data from the multivariate proposed Gamma distribution for all values of θ, not only for $\theta = \frac{q}{2}$ as in Henderson & Shimakura (2003)

Estimation for the Poisson-gamma mixed model

- Parameters: $\boldsymbol{\beta}, \theta, \rho$
- Full likelihood analysis requires the joint probability $\mathrm{P}(Y_{i1} = y_{i1}, \dots, Y_{iT} = y_{iT})$ which is intractable (no closed form)
- Alternative: Composite likelihood approach [Lindsay 1988] and two-stage estimation procedure
- First stage: Estimate $\eta = (\boldsymbol{\beta}, \theta)$, applying composite likelihood only using marginals
- Second stage: the estimated values $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}$ are used in the composite likelihood based on all pairwise time points for estimating the correlation parameter ρ

First stage

- Composite likelihood using marginals negative binomial
  - $Y_{it} \sim \mathrm{NB}(\mu_{it}, \theta)$, with mean $\mu_{it} = \exp(\mathbf{x}_{it}^\top \boldsymbol{\beta})$
- For fixed θ, the negative binomial distribution can be formulated as a GLM (with log–link).
- Can use `glm.nb` from MASS library in R
- Fits GLM for negative binomials by exploiting GLM-structure for fixed θ and adding maximization of log-likelihood with respect to θ
- Standard errors for $\eta = (\boldsymbol{\beta}, \theta)$ are found using a sandwich estimator

Second stage

- Estimate ρ using again composite likelihood based on all pairs of time points
- Composite log-likelihood contribution for subject i is

$$l_{2i}(\rho, \hat{\eta}) = \sum\sum_{1 \leq s \leq t \leq T} \log \mathrm{P}(Y_{is} = y_{is}, Y_{it} = y_{it})$$

- Total composite log-likelihood is sum over subjects
- Estimate $\hat{\rho}$ is found as solution to the composite score equations with the estimate $(\hat{\theta})$ from stage one plugged in:

$$\sum_{i=1}^{N} \frac{\partial l_{2i}(\rho, \hat{\eta})}{\partial \rho} = 0$$

- Advantage: only single parameter ρ to be estimated at this stage

Standard errors

- Standard errors for the estimates of the parameters of interest β, θ, and ρ can be obtained in two ways
  - Parametric bootstrap: feasible since it is possible to generate data from the proposed multivariate Gamma distribution
  - Asymptotic theory (apply sandwich methodology)
    - The asymptotic variance of $\hat{\rho}$ can also be estimated by a sandwich estimator, which also accounts for uncertainty of first stage estimator [Andersen 2004]
    - Actual formulas:

$$\begin{aligned}
\mathrm{var}(\hat{\eta}) &\approx \frac{1}{N} \cdot \mathbf{B}_1^{-1}\mathbf{M}_1\mathbf{B}_1^{-1} \\
\mathrm{var}(\hat{\rho}) &\approx \frac{1}{N} \cdot \Big[\mathbf{B}_2^{-1}\mathbf{M}_2\mathbf{B}_2^{-1} - 2\mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\mathbf{B}_1^{-1}\mathbf{M}_{12}\mathbf{B}_2^{-1} \\
&\qquad + \mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\mathbf{B}_1^{-1}\mathbf{M}_1\mathbf{B}_1^{-1}\mathbf{B}_{12}\mathbf{B}_2^{-1}\Big]
\end{aligned}$$

Application

- Data set consists of 65 patients used before in Henderson & Shimakura
- Counts in 12 successive intervals of equal length
- Estimate the baseline rate and the effect of the treatment
- Assumed same effect of the treatment for the 12 time points
- Counting model: $y_t \sim \mathrm{Po}((\beta_t + \gamma Z)Z_t)$,
- $Z_1, \ldots, Z_{12}$ multivariate serially correlated gamma-frailty vector
- $Z = 0$ or $Z = 1$ indicates treatment

Two–stage estimation procedure

- First-stage: the regression parameters β, γ and the overdispersion parameter θ are simultaneously estimated based on the marginal negative binomial distribution
- Estimation via glm with a negative binomial family
- Second-stage: estimate ρ based on the joint distribution of the pair

Compare one stage with two-stage estimation

- $\hat{\beta}$ and $\hat{\theta}$ obtained with the two–stage procedure were identical, up to five decimals, to those obtained in Henderson & Shimakura, even though both the underlying gamma frailty process and the estimation method differed.
- $\hat{\rho}$ = 0.847 with the two-stage procedure
- $\hat{\rho} = 0.849$ with Henderson & Shimakura's method
- Standard errors in the procedures were also quite similar

Simulation study

Compare one-stage vs two–stage composite likelihood

- Study the robustness of our procedure against different frailty vectors
- Results from both estimation procedures were remarkably similar
- The efficiency of our two–stage estimation was 99% compared to the one–stage composite likelihood procedure of Henderson & Shimakura
- Estimates appeared to be quite robust to misspecification of the particular multivariate frailty distribution generating the count data.

Summary

- We propose a new multivariate gamma distribution based on renewal processes
- The construction is based on the infinite divisibility property of the Gamma distribution
- The new multivariate gamma distribution has been used as a mixing distribution in a Poisson model for longitudinal count data
- Full likelihood is intractable, applied a composite likelihood and two-stage estimation procedure for estimating the parameters in the model
- Quantification of the loss of efficiency with respect to full likelihood requires further study

References

Andersen, E. W.
Composite likelihood and two-stage estimation in family studies.
Biostatistics **5**, 15 – 30, 2004.

Fiocco, M., Putter, H, and van Houwelingen H.
A new serially correlated gamma frailty process for longitudinal count data.
Submitted Biostatistics

Henderson, R. and Shimakura, S.
A serially correlated gamma frailty model for longitudinal count data.
Biometrika **90**, 355 – 366, 2003.

Lindsay, B. G.
Composite likelihood methods.
Contemporary Mathematics **80**, 221 – 239, 1988.