Longitudinal quantile regression in presence of informative drop-out through longitudinal-survival joint modeling

> Alessio Farcomeni Sapienza - University of Rome

> > Joint work with Sara Viviani

1

Longitudinal Data

- Longitudinal data arise when repeatedly measuring an outcome over time.
- Extremely interesting designs: increase sample size for analysis without having to find new subjects to study, possibility of evaluating the evolution of the outcome over time.

Longitudinal quantile Regression

- Modeling a conditional quantile rather than the conditional expectation of the outcome may be more appropriate in many situations.
- Median regression: useful for skewed outcomes, robust to the presence of outliers
- Predictors may have a different effect on lower or larger quantiles with respect to the center of the distribution.

Informative drop-out

- An ubiquitous problem in longitudinal studies is that subjects are lost at follow up.
- This event may be informative and bias estimates if ignored

Longitudinal QR with informative drop-out

- Despite the importance of informative missing data, there are only few approaches to QR in this setting.
- Lipsitz *et al.* (1997), Yi and He (2009): weighting by inverse probability of drop-out
- Bayesian approach by Yuan and Yin (2010)
- In the previous works, drop-out can occurr *only* at one of the observation times. This is a strong limit as our motivating example suggests.

A typical example: CD4 Data • 467 HIV infected patients randomized to didanosine (ddI) or zalcitabile (ddC). • 188 died during fup. • Longitudinal outcome is CD4 count, which is recorded at baseline, as well as (hopefully) 2, 6, 12 and 18 months thereafter.



Pros and Cons

SP and JM are very effective, but:

- limited to Gaussian error distributions and modeling of the conditional mean
- often limited to Gaussian random effects
- rigid structures for the relationship between the longitudinal and survival processes

We propose a general solution which is more flexible than SP and JM and can be used to model the mean or any quantile of the longitudinal outcome. A general MCEM can be used in all formulations of the model.

Set up

- (T_i, Δ_i) : time to event and censoring indicator, $i = 1, \ldots, n$
- Y_{it} : continuous outcome repeatedly observed at $t = 1, \ldots, t_i$; $i = 1, \ldots, n; t_i \leq T_i$.
- W_i, X_{it} : time fixed and time varying covariates, to be partitioned (e.g. $(X_{it1}, X_{it2}) = X_{it})$.

The JMQR

$$\begin{cases} Y_{it} = \beta'_1 X_{it1} + \beta_2 X_{it2} + u'_i X_{it3} + \epsilon_{it} \\ h_i(t|u_i) = h_0(t) \exp\{\gamma' W_i + \alpha_1 \beta'_2 X_{it2} + \alpha_2 u'_i X_{it3}\} \\ u_i \sim f(u_i|\Sigma) \end{cases}$$

where

- $f(\epsilon_{it})$ normal gives a model on the conditional mean of Y_{it}
- An ALD:

$$f(\epsilon_{it}) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho\left(\frac{\epsilon_{it}}{\sigma}\right)\right\},\,$$

 $\rho(u) = u\{\tau - I(u < 0)\}$, gives a model on the conditional τ quantile of Y_{it} .

Our contribution: the JMQR

- ALD is a simple and convenient parametric assumption for QR, pioneered in the longitudinal context by Geraci and Bottai (2007), Liu and Bottai (2009).
- The JM approach works when drop-out occurs at discrete or continuous time points
- The JMQR generalizes the model of Liu and Bottai (2009) to include informative drop-out
- Shared-parameter and JM are special cases: shared parameter when $\alpha_2 = 0$; JM when $\alpha_1 = \alpha_2$ and $X_{it1} = \emptyset$.

Random effects

- A natural distributional assumption is $u_i \sim MVN(0, \Sigma)$.
- Rizopoulos *et al.* (2008) shows that this working assumption is always OK asymptotically.
- When t_i is not large, or τ is far from 0.5, it may be better to use a multivariate T or a multivariate ALD.

Likelihood

• The joint distribution of time to event and longitudinal processes is:

$$f(T_i, \Delta_i, Y_i; \theta) = \int f(Y_i | u_i; \theta) f(T_i, \Delta_i | u_i; \theta) f(u_i | \Sigma) du_i,$$

where

$$f(T_i, \Delta_i \mid u_i; \theta) = f(T_i \mid u_i; \theta)^{\Delta_i} S(T_i \mid u_i; \theta)^{1 - \Delta_i}$$

= $h(T_i \mid \mathcal{T}_{it}, u_i; \theta)^{\Delta_i} S(T_i \mid \mathcal{T}_{it}, u_i; \theta).$

• The log-likelihood is then:

$$\ell(\theta) = \sum_{i} \log f(T_i, \Delta_i, Y_i; \theta).$$

• This is analytically intractable.

$$\begin{aligned} \mathcal{C} \text{omplete Likelihood} \\ \ell_{c}(\theta) &= \sum_{i} \log f(Y_{i}|u_{i};\theta) + \sum_{i} \log f(T_{i},\Delta_{i}|u_{i};\theta) + \sum_{i} \log f(u_{i}|\Sigma) \\ &= -\log \sigma \sum_{i} n_{i} - \sum_{i} \sum_{t=1}^{n_{i}} \rho \left(\frac{Y_{it} - \beta_{1}'X_{it1} - \beta_{2}'X_{it2} - u_{i}'X_{it3}}{\sigma} \right) \\ &+ \sum_{i} \Delta_{i} \log h_{0}(T_{i}) + \sum_{i} \gamma' W_{i} + \alpha_{1} \sum_{i} \beta_{2}'X_{iT_{i}2} + \alpha_{2} \sum_{i} u_{i}'X_{iT_{i}3} \\ &- \sum_{i} \int_{0}^{T_{i}} h_{0}(s) \exp\{\gamma' W_{i} + \alpha_{1}\beta_{2}'X_{is2} + \alpha_{2}u_{i}'X_{is3}\} ds \\ &+ \sum_{i} \log f(u_{i}|\Sigma). \end{aligned}$$

MCEM

We can obtain the MLE through an MCEM algorithm as follows:

- MCE-step: approximate the posterior for u_i through an Adaptive Rejection Metropolis Sampling. The number of samples for each *i* is chosen along the lines of Eichoff (2004) to guarantee an approximation error below a small threshold.
- M-step: average out the sampled values to estimate the complete log-likelihood. Compute the profile expected complete likelihood (next slide). Find values for the parameters such that it is increased through a one-step Nelder-Mead

Profile expected complete likelihood

- The number of parameters involved at the M step is too large due to $h_0(t)$.
- We obtain a profile expected complete likelihood by plug-in of a Nelson-Aalen type estimator

$$\widehat{h}_{0}(s) = \sum_{i=1}^{n} \frac{\Delta_{i}I(T_{i}=s)}{\frac{1}{B}\sum_{i:T_{i}\geq s}\sum_{b=1}^{B} \exp\{\gamma'W_{i} + \alpha_{1}\beta'_{2}X_{iT_{i}2} + \alpha_{2}v'_{ib}X_{iT_{i}3}\}},$$

where v_{ib} is sampled from the posterior of u_i .

• An implicit closed form expression can be found for σ as well.

Other inferential issues

- Standard errors, confidence intervals: non-parametric block bootstrap. Standard errors can be used to build Wald statistics for testing on the regression parameters.
- Testing, model choice: the likelihood can be directly approximated from the MCEM output, and can be used to check convergence, likelihood ratio testing, computation of information criteria.





Data: separate models

• Fixed Effects for Longitudinal Median Regression:

	estimate	std.err	p-value
Intercept	7.40	0.21	< 2e - 16
ddI	0.17	0.04	9e-5
ddI*Time	0.01	0.06	0.8648

• Survival for time to event gives a log HR for ddI of -0.33, p=0.25. You get the same result even after correction for CD4 count.

Data: JM for the median

- Longitudinal outcome: CD4 count.
- Fixed effects: ddI, ddI:Time
- Random effects: Patient ID, Time
- Survival outcome: time to death
- Baseline predictors: none
- Shared predictors: ddI
- Random effects: Patient ID, Time

(Last minute) Results

• Fixed Effects for Longitudinal Median Regression:

	estimate	std.err	p-value
Intercept	5.90	0.047	< 1e - 16
ddI	0.063	0.020	0.0016
ddI*Time	0.058	0.028	0.038

• Survival for time to event gives a log HR for ddI of -0.18, p=0.34.





Conclusions

- Informative drop-out may bias estimates of longitudinal parameters both in mean and quantile regression.
- The CD4 example suggests that this problem may be stronger for quantiles corresponding to a higher rate of events.
- In this work we propose a possible solution, generalizing shared-parameter and joint-models in different directions.

Further work

- The numerical results are last minute, as said.
- Simulation study
- Evaluation of sensitivity to informative drop-out
- Simoultaneous estimation of multiple quantiles