Model Estimation and Dynamic Prediction for Subject-Specific Event Probabilities in Joint Modeling Using Longitudinal Quantile Regression

(Proposal Defense)

Ming Yang M.S.

Deptartment of Biostatistics, UTSPH

April 16, 2015

Table of Contents

Background

- Two types of data: **longitudinal data** and **time-to-event data**.

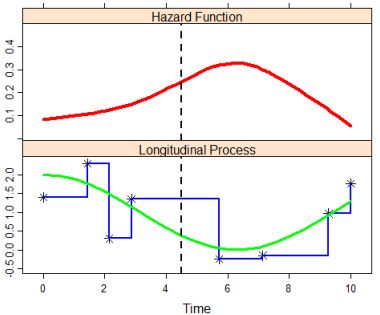

Figure: Correlation between time-to-event outcome and longitudinal outcome (Rizopoulos, 2014 online)

Background (Cont'd)

- The joint modeling method: handle two types of data simultaneously
- Linear mixed model (LMM) for longitudinal data
  - Normality assumption may be invalid
  - Sensitive to outliers
  - Modeling conditional mean is not very meaningful from clinical perspective
- Subject-specific predictions of event probabilities using joint modeling

Joint modeling: When to use it?

- When the focus is on the time-to-event outcome with **time-varying covariate** measured with error
- When the focus is on the longitudinal outcome and we would like to adjust for **non-random** drop-outs

Joint modeling: What is it?

$$\begin{cases} Y_{it} = m_i(t) + \varepsilon_{it} = \boldsymbol{X}_{it}^\top \boldsymbol{\beta} + \boldsymbol{Z}_{it}^\top \boldsymbol{u}_i + \varepsilon_{it}, \varepsilon_{it} \sim N(0, \sigma^2) \\ h(T_i | \mathcal{M}_{iT_i}, \boldsymbol{W}_i; \boldsymbol{\gamma}, \alpha_1, \alpha_2) = h_0(T_i) \exp(\boldsymbol{W}_i^\top \boldsymbol{\gamma} + \alpha m_i(T_i)) \end{cases} \quad (1)$$

- Y_{it}: the observed longitudinal outcome for ith subject at time t
- $T_i = \min(T_i^*, C_i)$: the event time for subject i, where T_i^* is the true underlying event time and C_i is the censoring time
- $m_i(t)$: the error-free longitudinal measure; $\mathcal{M}_{iT_i} = \{m_i(s) : 0 \leq s \leq T_i\}$
- $\boldsymbol{\beta}, \boldsymbol{\gamma}$: the fixed effects
- $\boldsymbol{u}_i$: a vector of random effects for subject i
- α: the parameter governing the strength of association

Joint modeling: How to do it?

- Bayesian method

  Complete likelihood function:

$$
\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{Y}, \boldsymbol{u}_i) &= \prod_{i=1}^{N} f(\boldsymbol{Y}_i|\boldsymbol{u}_i; \boldsymbol{\theta}) f(T_i, \Delta_i|\boldsymbol{u}_i; \boldsymbol{\theta}) f(\boldsymbol{u}_i; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{N} \prod_{t=1}^{n_i} f(Y_{it}|\boldsymbol{u}_i; \boldsymbol{\theta}) f(T_i, \Delta_i|\boldsymbol{u}_i; \boldsymbol{\theta}) f(\boldsymbol{u}_i; \boldsymbol{\theta}) \qquad (2)
\end{aligned}
$$

  Next: derive the full conditional of each parameter.

Quantile Regression (QR)

- The τth quantile of a random variable Y, where $\tau \in [0, 1]$

$$Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : Pr(Y \leq y) \geq \tau\} \tag{3}$$

- QR models

$$Q_{Y|X}(\tau) = X^\top \boldsymbol{\beta} \tag{4}$$

- Inference method:

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left[\rho_\tau (Y_i - X_i^\top \boldsymbol{\beta}) \right], \tag{5}$$

where $\rho_\tau(Y) = Y(\tau - I(Y < 0))$.

Quantile Regression (QR) (Cont'd)

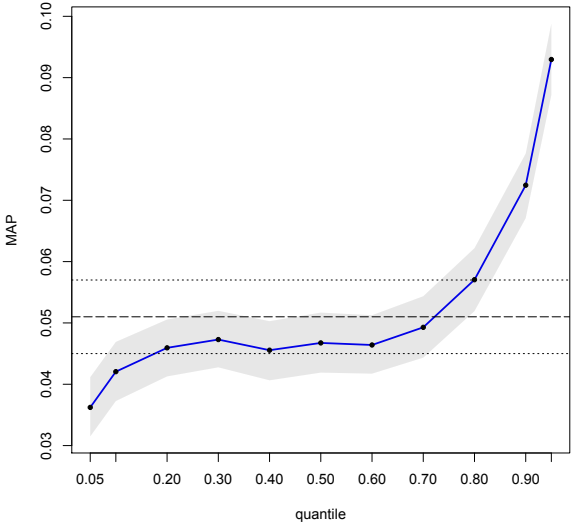

Figure: Quantile regression v.s. mean regression

Asymmetric Laplace distribution (ALD)

- Previous minimization problem can also be rephased as a maximum-likelihood problem by using ALD.
- An ALD is given by

$$f(Y|\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left[-\rho_\tau\left(\frac{Y-\mu}{\sigma}\right)\right], \tag{6}$$

where $\mu \in (-\infty, \infty)$ is the location parameter, σ is the scale parameter and $\tau \in (0, 1)$ is the skewness parameter.

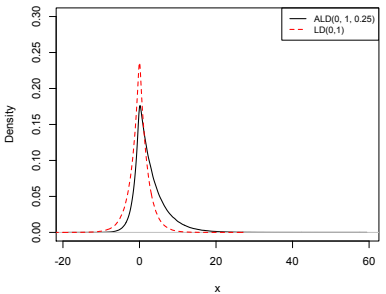

Figure: Asymmetric Laplace distribution and Laplace distribution

What have been done so far?

- ...
- Make predictions of survival probabilities under the traditional JM framework (Rizopoulos, 2011) (Taylor et al., 2013)
- Statistical inference of JM using quantile regression model for the longitudinal process (Farcomeni and Viviani, 2014)

Specific research aim 1

To develop a fully Bayesian method for subject-specific dynamic predictions of survival probabilities using quantile regression.

Specific research aim 2

To extend Aim 1 to recurrent events data and obtain statistical inference.

Specific research aim 3

To extend Aim 2 to obtain subject-specific predictions of recurrent events probabilities.

Public health significance

- Using quantile regression to focus on the **low or high tail** of the longitudinal outcome, which can be of greater interest clinically and more relevant to research questions
  - CD4 cell counts in HIV research (lower tail)
  - Study of low birth weight infants (lower tail)
  - hypertension in cardiovascular study (upper tail)
  - Prostate-Specific Antigen (PSA) levels in prostate cancer patients (upper tail)
- Subject-specific predictions and healthcare
  - The idea of ”personalized medicine”: to provide the right patient with the right drug at the right time
  - Targeted treatment will be more subject specific thus more effective

Public health significance (Cont'd)

- Medicine revolution: from reactive to preventive
- Enable the selection of optimal therapy
- Assess individual drug response: reduce adverse drug reactions

Medicine Today

Reactive, population-based, one-size-fits-all model of care

Personalized Medicine

Predictive, preventive, patient-centric model of care

Statistical methods

- JM using longitudinal quantile regression
- Subject-specific dynamic predictions
- Predictive performance of longitudinal biomarker

Longitudinal quantile regression

- The linear quantile mixed model (LQMM):

$$Q_{Y_{ij}|X_{ij},Z_{ij}}(\tau) = X_{ij}^{\top}\beta_{\tau} + Z_{ij}^{\top}u_i,\ i = 1, \cdots, N;\ j = 1, \cdots, n_i. \tag{7}$$

- Under $\varepsilon_{ij} \sim \text{ALD}(0, \sigma, \tau)$, $Y_{ij}|u_i \overset{iid}{\sim} \text{ALD}(X_{ij}^{\top}\beta + Z_{ij}^{\top}u_i, \sigma, \tau)$:

$$f(Y_{ij}|u_i; \beta_{\tau}, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left[-\rho_{\tau}\left(\frac{Y_{ij} - X_{ij}^{\top}\beta - Z_{ij}^{\top}u_i}{\sigma}\right)\right] \tag{8}$$

Longitudinal quantile regression (Cont'd)

- The location-scale mixture representation of the ALD (Kotz et al., 2001):

$$\varepsilon_{ij} = \kappa_1 e_{ij} + \kappa_2 \sqrt{\sigma e_{ij}} v_{ij}.$$

$$Y_{ij} = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\top \boldsymbol{u}_i + \kappa_1 e_{ij} + \kappa_2 \sqrt{\sigma e_{ij}} v_{ij}. \quad (9)$$

where

$$\kappa_1 = \frac{1-2\tau}{\tau(1-\tau)}, \kappa_2^2 = \frac{2}{\tau(1-\tau)},$$

and

$$v_{ij} \sim N(0,1), e_{ij} \sim \exp(1/\sigma).$$

Longitudinal quantile regression and JM

$$\begin{cases} Y_{it} = X_{it}^\top \boldsymbol{\beta} + H_{it}^\top \boldsymbol{\delta} + Z_{it}^\top \boldsymbol{u}_i + \varepsilon_{it}, \varepsilon_{it} \sim ALD(0, \sigma, \tau) \\ h(T_i | \mathcal{M}_{iT_i}, W_i; \boldsymbol{\gamma}, \alpha_1, \alpha_2) = h_0(T_i) \exp(W_i^\top \boldsymbol{\gamma} + \alpha_1 H_{iT_i}^\top \boldsymbol{\delta} + \alpha_2 Z_{iT_i}^\top \boldsymbol{u}_i) \end{cases} \tag{10}$$

Example:

- Y: Left ventricular ejection fraction (LVEF)
- T: Time to death
- X: intercept and age
- H: mildly dilated cardiomyopathy (MDCM) indicator $\times$ $(1 \quad t)$
- W: gender, New York Heart Association (HYHA) functional class
- Z: $(1 \quad t)$

Dynamic predictions of future event probabilities

- Notations:
  - $\mathcal{Y}_i(t) = \{Y_i(s), 0 \leq s \leq t\}$: complete history of observed longitudinal outcome for patient i up to time t
  - $\mathcal{D}_n = \{T_i, \Delta_i, \boldsymbol{Y}_i, i = 1, \cdots, n\}$: the training data
  - $p_i(m|t) = Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n; \boldsymbol{\theta})$: the probability that patient i is free of event up to time $m > t$, given he/she is free of event until time t.
- The predicted probability of no event until time m is then given by

$$\begin{aligned} & Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n; \boldsymbol{\theta}) \\ = & \int \frac{S_i[m|\mathcal{M}_i(m, u_i, \boldsymbol{\theta}); \boldsymbol{\theta}]}{S_i[t|\mathcal{M}_i(t, u_i, \boldsymbol{\theta}); \boldsymbol{\theta}]} Pr(u_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) du_i, \end{aligned} \tag{11}$$

Dynamic predictions of future event probabilities (Cont'd)

- Monte Carlo (MC) estimate of $p_i(m|t)$:
  - draw $\boldsymbol{\theta}^{(k)} \sim f(\boldsymbol{\theta}|\mathcal{D}_n)$;
  - draw $u_i^{(k)} \sim f(u_i|T_i^* > t, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(k)})$
  - compute $p_i^{(k)}(m|t) = S_i[m|\mathcal{M}_i(m, u_i^{(k)}, \boldsymbol{\theta}^{(k)}); \boldsymbol{\theta}^{(k)}]S_i[t|\mathcal{M}_i(t, u_i^{(k)}, \boldsymbol{\theta}^{(k)}); \boldsymbol{\theta}^{(k)}]^{-1}$
- Sample mean or median:

$$\hat{p}_i(m|t) = \frac{1}{K}\sum_{k=1}^{K} p_i^{(k)}(m|t), \tag{12}$$

Dynamic predictions of future event probabilities (Cont'd)

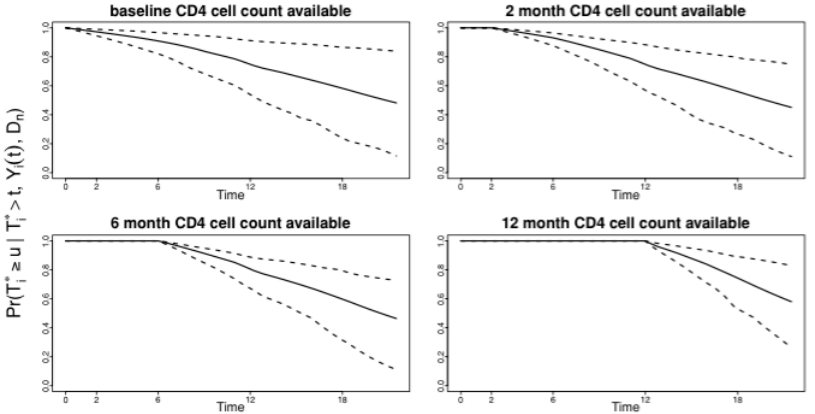

Figure: Example of subject-specific dynamic predictions for survival probabilities (Rizopoulos, 2011)

Predictive performance of the longitudinal biomarker

- Sensitivity

$$Pr[\mathcal{S}_i(k,t,\boldsymbol{c})|T_i^* > t, T_i^* \in (t, t+\Delta t]; \boldsymbol{\theta}] \tag{13}$$

- Specificity

$$Pr[\mathcal{F}_i(k,t,\boldsymbol{c})|T_i^* > t, T_i^* > t+\Delta t; \boldsymbol{\theta}] \tag{14}$$

- Notations:
  - $\mathcal{S}_i(k,t,\boldsymbol{c}) = \{Y_i(s) \leq c_s, k \leq s \leq t\}$ is defined as success (or event)
  - $\mathcal{F}_i(k,t,\boldsymbol{c}) = \mathbb{R}^{n(k,t)} \setminus \{Y_i(s) \leq c_s, k \leq s \leq t\}$ is defined as failure
  - $\boldsymbol{c}$ is a vector of threshold values and c_s is the threshold value at time s
  - $\mathbb{R}^n$ denotes the $n-$dimensional Euclidean space
  - $n(k,t)$ is the total number of longitudinal measurements in interval $[k,t]$

Predictive performance of the longitudinal biomarker (Cont'd)

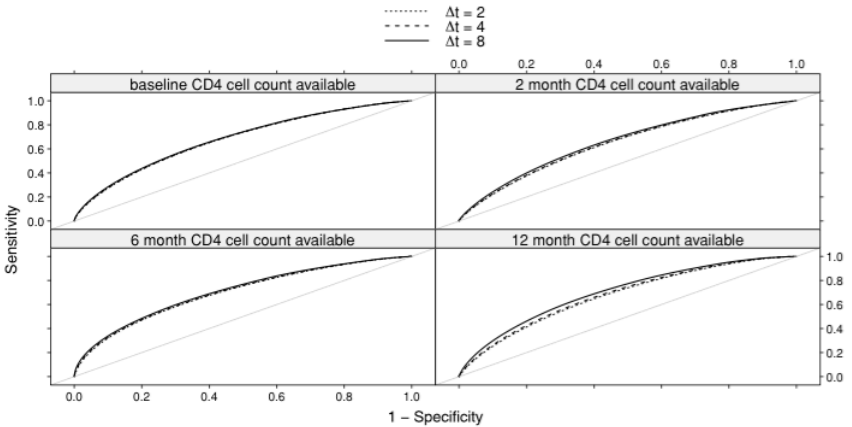

Figure: Example of ROCs for predictive performance (Rizopoulos, 2011)

Simulation studies

- **Simulation Study 1**: validity of proposed Bayesian inference method
  1. $(\alpha_1, \alpha_2) = (0, 0)$, the two models are indepdent with each other
  2. $(\alpha_1, \alpha_2) = (1, 0)$, the two models are related only through the observed heterogeneity in some covarites, i.e. H_{it} in our model
  3. $(\alpha_1, \alpha_2) = (0, 1)$, the two models are related only through the unobserved heterogeneity, i.e. the random effects
  4. $(\alpha_1, \alpha_2) = (1, 1)$, the depdence of the two models is explained by both observed and unobserved heterogeneity

Table: Bias and standard error of the parameter estimates from proposed fully Bayesian estimating method

			τ=0.25									
			β		δ		γ		α_1		α_2	
n	α_1	α_2	bias	s.d.	bias	s.d.	bias	s.d.	bias	s.d.	bias	s.d.
500	0	0										
500	1	0										
500	0	1										
500	1	1										

Simulation studies (Cont'd)

- **Simulation Study 2**: accuracy of the prediction method
- Statistics to be compared – Equation (11):

$$\frac{S_i[m|\mathcal{M}_i(m, u_i, \boldsymbol{\theta}); \boldsymbol{\theta}]}{S_i[t|\mathcal{M}_i(t, u_i, \boldsymbol{\theta}); \boldsymbol{\theta}]}.$$

- Compare the predicted values with the "gold standard", i.e. the simulated values

	$\Delta t = 2$	$\Delta t = 4$	$\Delta t = 6$
	bias(lower, upper)	bias(lower, upper)	bias(lower, upper)
t=2			
t=4			
t=8			
t=16			

Table: Summary table of comparing the predictive results from proposed method with gold standard

Simulation plans (Cont'd)

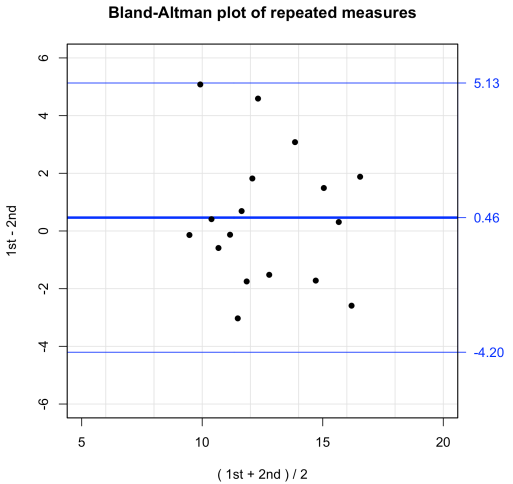

Figure: An example of Bland-Altman plot

Real data

- The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) (1994-2002): multi-center, randomized, double-blind, active-controlled clinical trial.
- 42,448 participants, 625 sites.
- 19% Hispanic patients, 46.8% women, 67 years old on average (with 35% aged $\geq$70 years), 36% of patients with diabetes, 47% are cardiovascular disease patients and 22% are smokers.
- Primary outcome: fatal coronary heart disease (CHD) or non-fatal myocardial infarction (MI)
- Secondary endpoints: cardiovascular events (stroke, heart failure (HF), CHD, etc.).

Description of real data (Cont'd)

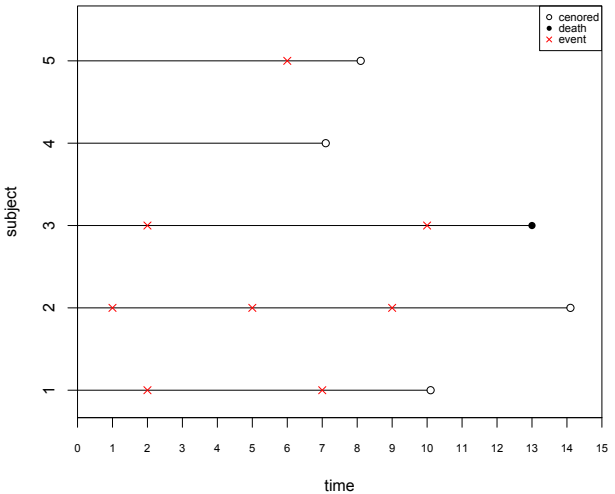

Figure: An example of recurrent events (e.g. stroke) data

Acknowledgement

- **Dissertation committee:**
  Stacia M. DeSantis, PhD (Chair & Academic Advisor)

  Sheng Luo, PhD (Dissertation Supervisor)

  David R. Lairson, PhD (Minor Adviosr)

  Xiaoming Liu, PhD (Breadth Advisor)

- **Externel reviewer:**
  Soeun Kim, PhD

Selected references

J Martin Bland and Douglas G Altman
Statistical methods for assessing agreement between two methods of clinical measurement.
The Lancet, 327(8476):307–310, 1986.

Barry R Davis, Jeffrey A Cutler, David J Gordon, Curt D Furberg, Jackson T Wright, William C Cushman, Richard H Grimm, John LaRosa, Paul K Whelton and H Mitchell Perry
Rationale and design for the antihypertensive and lipid lowering treatment to prevent heart attack trial (ALLHAT).
American Journal of Hypertension, 9(4):342–360, 1996.

Samuel Kotz, Tomasz Kozubowski and Krzysztof Podgorski
The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Exonomics, Engineering, and Finance.
Springer, 2001.

Selected references (Cont'd)

Roger Koenker
Quantile regression.
Cambridge university press, 2005.

Dimitris Rizopoulos
Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data.
Biometrics, 67(3):819–829, 2011.

Jeremy MG Taylor, Yongseok Park, Donna P Ankerst, Cecile Proust-Lima, Scott Williams, Larry Kestin, Kyoungwha Bae, Tom Pickles and Howard Sandler
Real-time individual predictions of prostate cancer recurrence using joint models.
Biometrics, 69(1):206–213, 2013.

Alessio Farcomeni and Sara Viviani
Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling.
Statistics in Medicine, 2014.