Bayesian quantile regression joint models: inference and dynamic predictions

Ming Yang, M.S.

Department of Biostatistics
The University of Texas School of Public Health

November 29, 2016

Outline

Background
- Motivation
- Joint models for longitudinal and time-to-event data
- Research Aims

Proposed Journal Articles
- Journal Article 1
  - Methods
  - Simulation studies
  - Application to PREDICT-HD data
  - Discussion
- Journal Article 2
  - Background
  - Methods
  - Simulation
  - Application to ARIC data
  - Discussion
- Journal Article 3
  - Background
  - Methods
  - Simulation
  - Application to ARIC data
  - Discussion

Acknowledgement

References

A motivating data

- A prospective observational study designed to detect early neurobiological predictors of Huntington's Disease (PREDICT-HD; ClinicalTrials.gov number NCT00051324)
- Data: 1078 participants, median follow-up time: 61 months, 40 longitudinal biomarkers, time to HD onset and other demographic information
- Primary focus: to investigate the association between longitudinal biomarkers and the risk of HD onset
- More extreme values in longitudinal biomaker(s) are associated with higher risk of HD onset
- Many of the longitudinal biomakers are skewed

PREDICT-HD study: skewed longitudinal biomarker

Total Motor Score (TMS), a commonly used rating criteria of body motion abilities based on the Unified Huntington Disease Rating Scale (UHDRS).

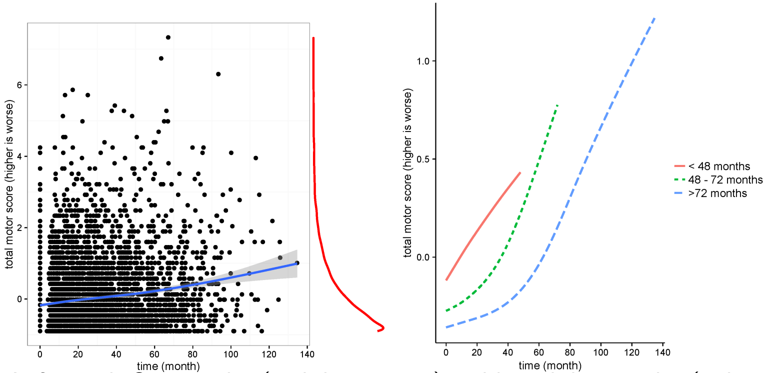


Figure: Left panel: Scatter plot (with loess curve) and kernel density plot (right side) for total motor score from the study population (time unit: month; lower total motor score is better); right panel: Mean total motor score values over time.

Joint models for longitudinal and time-to-event data

- Traditional joint models (JM)

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^\top(t)\beta + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \overset{iid}{\sim} N(0, \sigma^2) \\ h(t|\mathcal{M}_i(t), \boldsymbol{W}_i; \gamma, \alpha) = h_0(t)\exp(\boldsymbol{W}_i^\top\gamma + \alpha m_i(t)) \end{cases}$$

- Linear mixed model (LMM) for the longitudinal outcome
- Cox proportional hazards model (PHM) for the time-to-event outcome
- Longitudinal outcome is treated as a time-dependent covariate in the time-to-event submodel

Limitations of traditional JM

- LMM is sensitive to outliers and deviation of normality
- The normality assumption cannot be satisfied in many cases (even after applying various outcome transformations)
- LMM models only the conditional mean of the outcome – not very meaningful from clinical perspective in some cases

Research aims

- **Aim 1**: To build a new JM framework for longitudinal and **survival data** that is more robust against non-normal data and to develop fully Bayesian inference and dynamic prediction algorithms for the proposed JM
- **Aim 2**: To extend the new JM to study longitudinal data and **recurrent events**, develop a Bayesian method for model inference
- **Aim 3**: To make dynamic predictions of **recurrent events** risk from the JM developed in Aim 2.

Journal Article 1

Bayesian quantile regression joint models: inference and dynamic predictions

Statistical methods

- JM using longitudinal quantile regression
- Subject-specific dynamic predictions

Quantile Regression (QR)

- QR models

$$Q_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{X}^\top \beta_\tau, \tag{1}$$

where the τth quantile of a random variable Y, $\tau \in [0, 1]$, is defined as

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

- Regression parameters are estimated as:

$$\hat{\beta}_\tau = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left[\rho_\tau(Y_i - \boldsymbol{X}_i^\top \beta_\tau)\right], \tag{2}$$

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

Parameter estimation from QR vs. mean regression

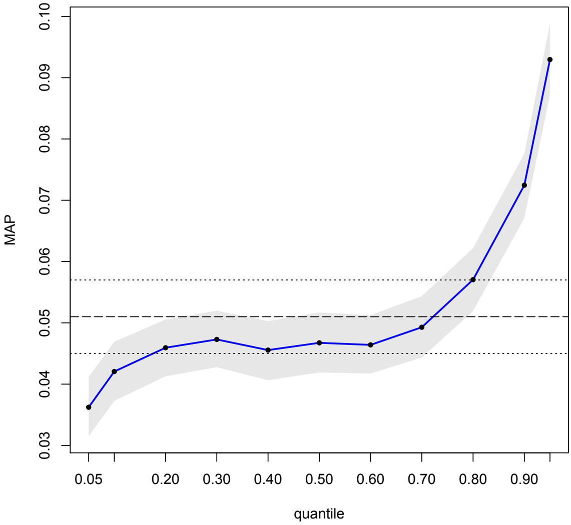


Figure: Quantile effect v.s. mean effect

Longitudinal quantile regression

- The linear quantile mixed model (LQMM):

$$\begin{cases} Y_i(t) = \boldsymbol{X}_i^\top(t)\beta_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \ i = 1, \cdots, N; \ t = 1, \cdots, n_i, \\ Q_{Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{u}_i}(\tau) = \boldsymbol{X}_i^\top(t)\beta_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i \end{cases}$$

- Assume asymmetric Laplace distribution (ALD) of the random error, i.e. $\varepsilon_i(t) \overset{iid}{\sim} \text{ALD}(0, \sigma, \tau)$:

$$f(\varepsilon_i(t)|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left[-\rho_\tau\left(\frac{\varepsilon_i(t)}{\sigma}\right)\right];$$

- Then $Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{u}_i \overset{iid}{\sim} \text{ALD}(\boldsymbol{X}_i^\top(t)\beta + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i, \sigma, \tau)$:

$$f(Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{u}_i; \beta, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left[-\rho_\tau\left(\frac{Y_i(t) - \boldsymbol{X}_i^\top(t)\beta - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i}{\sigma}\right)\right].$$

ALD vs. LD vs. Normal

In ALD(μ, σ, τ), $\mu \in (-\infty, \infty)$ is the location parameter, σ is the scale parameter and $\tau \in (0, 1)$ is the parameter that control the skewness of the distribution.

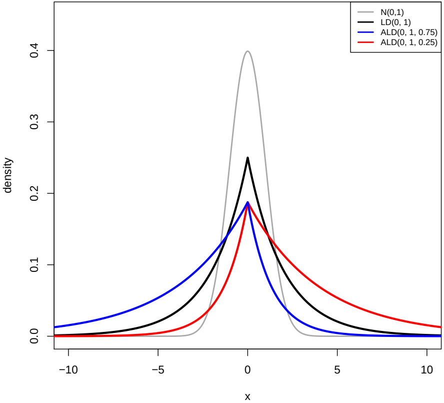


Figure: Asymmetric Laplace, Laplace, and normal distributions

Quantile regression joint models (QRJM)

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^\top(t)\beta_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ h(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i; \gamma_\tau, \alpha_\tau) = h_0(T_i)\exp(\boldsymbol{W}_i^\top\gamma_\tau + \alpha_\tau(\boldsymbol{X}_i^\top(T_i)\beta_\tau + \boldsymbol{Z}_i^\top(T_i)\boldsymbol{u}_i)) \end{cases} \tag{3}$$

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

Dynamic predictions of future event-free probability

- The predicted probability of no event until time m given no event until time t $(t < m)$ is given by

$$\begin{aligned} & Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_N; \boldsymbol{\theta}) \\ = & \int \frac{S_i[m|\mathcal{M}_i(m, u_i, \boldsymbol{\theta}); \boldsymbol{\theta}]}{S_i[t|\mathcal{M}_i(t, u_i, \boldsymbol{\theta}); \boldsymbol{\theta}]} Pr(u_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) du_i, \end{aligned} \tag{4}$$

- Notations:
  - $p_i(m|t) = Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n; \boldsymbol{\theta})$: the probability that patient i is free of event up to time $m > t$, given he/she is free of event until time t
  - $\mathcal{Y}_i(t) = \{Y_i(s), 0 \leq s \leq t\}$: 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

Estimation of the predicted probability

- A Monte Carlo (MC) approximation of $p_i(m|t)$ can be obtained using the following procedure:
  1. Draw $\boldsymbol{\theta}^{(p)} \sim Pr(\boldsymbol{\theta}|\mathcal{D}_N)$ for $p = 1, \cdots, P$;
  2. For each $\boldsymbol{\theta}^{(p)}$, draw $\boldsymbol{u}_i^{(q)} \sim f(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(p)})$ for $q = 1, \cdots, Q$ and compute

$$p_i^{(p)}(m|t) = \frac{1}{Q}\sum_{q=1}^{Q} S_i[m|\mathcal{M}_i(m, \boldsymbol{u}_i^{(q)}, \boldsymbol{\theta}^{(p)}); \boldsymbol{\theta}^{(p)}] S_i[t|\mathcal{M}_i(t, \boldsymbol{u}_i^{(q)}, \boldsymbol{\theta}^{(p)}); \boldsymbol{\theta}^{(p)}]^{-1};$$

  3. Approximate $p_i(m|t)$ by $\hat{p}_i(m|t) = \frac{1}{P}\sum_{p=1}^{P} p_i^{(p)}(m|t)$ after collecting all P samples of $p_i(m|t)^{(p)}$.

Predictive accuracy

- Let $\hat{r}_i(t+\Delta t|t) = 1 - \hat{p}_i(t+\Delta t|t), i = 1, \cdots, N$, i.e. the event risk.
- $$\widehat{TPR}_t^{\Delta t}(c) = \frac{\sum_{i=1}^{N} \hat{r}_i(t+\Delta t|t) I(\hat{r}_i(t+\Delta t|t) \geq c)}{\sum_{i=1}^{N} \hat{r}_i(t+\Delta t|t)},$$
  $$\widehat{FPR}_t^{\Delta t}(c) = \frac{\sum_{i=1}^{N} (1 - \hat{r}_i(t+\Delta t|t)) I(\hat{r}_i(t+\Delta t|t) \geq c)}{\sum_{i=1}^{N} (1 - \hat{r}_i(t+\Delta t|t))}.$$
- We use the following three statistics as measures of predictive performance (higher is better):
  - AUC: Area Under (the ROC) Curve
  - AARD: Above Average Risk Difference
  - MRD: Mean Risk Difference

Simulation study I: model inference

- Simulate data from QRJM model and consider the following scenarios:
  1. Scenario 1: random errors follow ALD(0, 1, $\tau = 0.25$) (right-skewed);
  2. Scenario 2: random errors follow a standard normal distribution (symmetric about 0).
- For each scenario, simulate 200 data sets with $N = 600$ in each.
- Among the 600 subjects, randomly select 500 as the training data used to fit the model, and use the remaining 100 subjects as the testing data to make out-of-sample predictions in Simulation II.
- Compare the bias, standard error (SE), mean squared error (MSE), and coverage probability (CP) for QRJM and the standard JM (LMJM).

Simulation I results I

Table: Simulation results in Simulation study I Scenario 1 in which random errors are generated from ALD(0, 1, $\tau = 0.25$).

	QRJM ($\tau = 0.25$)				QRJM ($\tau = 0.5$)				LMJM			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP
Coefficients for longitudinal process												
β_0	-0.003	0.080	0.014	0.930	1.659	0.129	2.807	0.020	2.702	0.146	7.350	0.000
β_1	0.015	0.068	0.010	0.950	0.024	0.105	0.043	0.890	0.080	0.116	0.052	0.860
β_2	0.016	0.083	0.013	0.950	0.014	0.112	0.042	0.970	0.078	0.128	0.052	0.920
Coefficients for survival process												
γ_1	0.005	0.055	0.006	0.940	0.008	0.057	0.006	0.960	0.009	0.058	0.007	0.960
γ_2	0.006	0.055	0.006	0.930	0.010	0.056	0.007	0.910	0.010	0.058	0.007	0.940
α	-0.004	0.078	0.010	0.970	-0.051	0.119	0.070	0.930	-0.087	0.103	0.040	0.800

Simulation I results II

Table: Simulation result in Simulation study I Scenario 2 in which random errors are generated from $\mathcal{N}(0, 1)$.

	QRJM ($\tau = 0.5$)				LMJM			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP
Coefficients for longitudinal process								
β_0	0.015	0.037	0.003	0.950	0.000	0.035	0.002	0.980
β_1	0.004	0.034	0.002	0.960	−0.003	0.033	0.002	0.950
β_2	0.013	0.050	0.005	0.950	0.006	0.049	0.005	0.950
Coefficients for survival process								
γ_1	0.008	0.055	0.006	0.920	0.003	0.054	0.006	0.900
γ_2	0.015	0.055	0.007	0.920	0.010	0.054	0.006	0.920
α	−0.013	0.055	0.006	0.950	0.007	0.055	0.006	0.950

Simulation II: dynamic predictions

- Use the 100 subjects as testing data and make out-of-sample predictions
- Compare the predicted values with the true simulated values ("gold standard")
- Use different combinations of $(t, \Delta t)$ for prediction to mimic the real-world situation

Simulation II results: summary table

Table: Simulation result in Simulation study II Scenario 1: MSE and bias of the difference between predicted survival probability and the gold standard.

		QRJM ($\tau = 0.25$)		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias	MSE	Bias
0.25	0.25	0.006	0.009	0.137	-0.330	0.244	-0.462
	1	0.010	0.007	0.111	-0.267	0.177	-0.343
(subjects left: 48.1%)	2	0.012	0.003	0.083	-0.197	0.126	-0.249
	3	0.013	0.000	0.072	-0.168	0.107	-0.210
0.5	0.25	0.007	0.009	0.130	-0.317	0.219	-0.439
	1	0.015	0.000	0.144	-0.321	0.221	-0.408
(subjects left: 34.6%)	2	0.017	-0.015	0.121	-0.259	0.174	-0.319
	3	0.018	-0.023	0.109	-0.228	0.153	-0.278
0.75	0.25	0.009	0.005	0.125	-0.301	0.189	-0.401
	1	0.023	-0.007	0.174	-0.356	0.253	-0.447
(subjects left: 22.8%)	2	0.025	-0.033	0.159	-0.310	0.218	-0.375
	3	0.027	-0.046	0.148	-0.282	0.197	-0.336

Data application

- Split the 1078 study participants into two parts: a first sub-cohort of 800 participants is used to draw statistical inference for the unknown parameters; the remainder is used as test data for predictions of HD-free probability.
- We consider the following joint models for our data analysis:

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_{it} = \beta_0 + \beta_1 t + \beta_2 age_{0i} + u_{i1} + u_{i2} t + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ h(T_i|\mathcal{M}_i(T_i); \gamma, \alpha) = \lambda(T_i) \exp(\gamma_1 education_i + \gamma_2 I_{male_i} + \alpha m_i(T_i)) \end{cases}$$

- $y_i(t)$ represents one of the longitudinal biomarkers
- age_0 is the baseline age at the enrollment.
- Specify a piecewise constant baseline hazard function with three time intervals, where λ_k stands for the hazard rate for time interval $[t_k, t_{k+1})$ and $I_k(t) = 1$ if $t \in [t_k, t_{k+1})$ and 0 otherwise.

Data analysis results I

Table: PREDICT-HD data analysis: Parameter estimation and 95% credible interval from QRJM at three different quantiles with TMS as the longitudinal biomarker.

	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
		longitudinal TMS process	
int.	-0.760 (-0.903, -0.628)	-0.525 (-0.699, -0.359)	-0.249 (-0.469, -0.035)
time (month)	0.019 (0.015, 0.023)	0.020 (0.016, 0.024)	0.022 (0.018, 0.026)
age_0	0.004 (0.001, 0.008)	0.005 (0.001, 0.010)	0.006 (0.001, 0.012)
		time to HD onset process	
assoct.	1.526 (1.321, 1.745)	1.300 (1.148, 1.459)	1.080 (0.968, 1.192)
eduyr	-0.083 (-0.115, -0.052)	-0.112 (-0.142, -0.082)	-0.128 (-0.157, -0.101)
male	0.317 (-0.037, 0.654)	0.360 (-0.020, 0.708)	0.317 (-0.010, 0.647)

Data analysis results II

Table: PREDICT-HD data analysis: AUC, AARD and MRD of the predictions of HD-free probability from QRJM and AUC from LMJM with TMS as the longitudinal biomarker.

t	Δt	AUC (τ)			AARD (τ)			MRD (τ)			AUC(LMJM)
(month)		0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	
12	12	0.647	0.683	0.738	0.213	0.261	0.356	0.010	0.020	0.059	0.679
	24	0.668	0.702	0.753	0.244	0.290	0.379	0.028	0.054	0.128	0.695
	36	0.685	0.714	0.760	0.273	0.311	0.391	0.054	0.091	0.170	0.693
24	12	0.836	0.857	0.864	0.539	0.575	0.577	0.168	0.218	0.285	0.855
	24	0.852	0.872	0.873	0.566	0.598	0.583	0.285	0.361	0.404	0.878
	36	0.866	0.877	0.872	0.581	0.599	0.575	0.368	0.420	0.430	0.836
48	12	0.875	0.878	0.868	0.583	0.598	0.589	0.326	0.320	0.303	0.669
	24	0.875	0.883	0.874	0.578	0.602	0.598	0.390	0.401	0.379	0.769
	36	0.877	0.887	0.879	0.589	0.614	0.599	0.417	0.439	0.417	0.774

Discussion

- The proposed JM provides a way to explore the covariates effect across the whole distribution span of the outcome variable. This becomes especially important when either the lower or higher quantile of the outcome becomes more relevant to the clinical interest.
- Our proposed algorithm performs well in recovering the truth in inference and in making predictions of future survival probabilities.
- The best predictive performance from our model outperforms that from the LMJM when data are highly skewed.
- Our novel application of JM in making personalized dynamic predictions of survival probability finds practical importance in many clinical applications.
- Predictive accuracy criteria and/or other model selection methods or method(s), e.g. Bayesian model averaging, to incorporate multiple regression results from different quantiles into a single prediction solution can be helpful in selecting the "best" quantile in prediction.

Journal Article 2

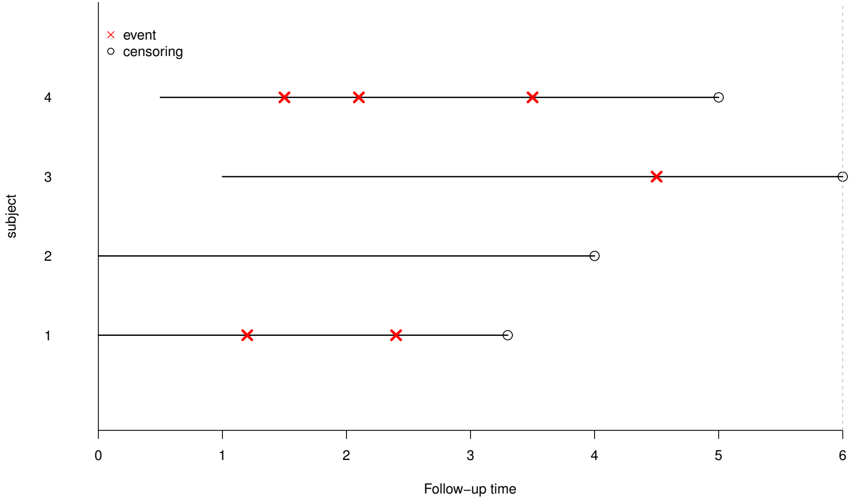
Bayesian Quantile Regression Joint Models of Longitudinal and Recurrent Event Data

Background I

- Recurrent events are commonly encountered in longitudinal biomedical studies
- JM: to simultaneously model the repeated instances of the continuous and time-to-event sequences and to examine the association between the two processes
- JM of longitudinal data and recurrent events have received less attention
- No work has considered incorporating QR model in JM of longitudinal and recurrent event data so far

Background II

A simple demo of recurrent events data



The QRJM

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^\top(t)\beta_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ r_i(t|\mathcal{M}_i(t), \boldsymbol{W}_i; \gamma_\tau, \alpha_\tau) = r_{i0}(t)\exp(\boldsymbol{W}_i^\top\gamma_\tau + \alpha_\tau(\boldsymbol{X}_i^\top(t)\beta_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i)) \end{cases} \tag{5}$$

- $r_{i0}(\cdot)$ is the subject-specific baseline intensity function
- $m_i(t)$ is the true underlying longitudinal outcome at time t and is estimated using an LQMM
- $\mathcal{M}_i(t) = \{m_i(s) : 0 \leq s \leq t\}$ is the true longitudinal process up to time t

Likelihood function of Recurrent events

Likelihood function for recurrent event data:

$$\begin{aligned}\ell_i(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}) &= \prod_{k=1}^{m_i} \left[r_i(T_{ik}; \boldsymbol{\theta} | \mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}} \exp\left(- \int_{T_{ik-1}}^{T_{ik}} r_i(s; \boldsymbol{\theta} | \mathcal{M}_i(s), \boldsymbol{W}_i) ds \right) \right] \\ &= \prod_{k=1}^{m_i} \left[r_i(T_{ik}; \boldsymbol{\theta} | \mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}} \right] \exp\left(- \int_{0}^{T_{im_i}} r_i(s; \boldsymbol{\theta} | \mathcal{M}_i(s), \boldsymbol{W}_i) ds \right),\end{aligned}$$

- Let C_i be the censoring time for subject i
- m_i is the total number of events observed within C_i
- T_{ik} is the kth observed event time, where $k = 0, \cdots, m_i$ $(T_{i0} = 0)$
- $\Delta_{ik} = I(T_{ik} < C_i)$ is the event indicator for kth event

Complete likelihood and Bayesian inference

- Complete likelihood function for subject i

$$L_i(\boldsymbol{\theta}; \boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(C_i), \boldsymbol{u}_i) = \ell_i(\mathcal{Y}_i(C_i); \boldsymbol{\theta}|\boldsymbol{u}_i)\ell_i(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}|\boldsymbol{u}_i)f(\boldsymbol{u}_i|\boldsymbol{\Sigma}) \qquad (6)$$

- Posterior distributions

$$f(\boldsymbol{\theta}|\boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}, \boldsymbol{u}) \propto \prod_{i=1}^{N} L_i(\boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(C_i), \boldsymbol{u}_i; \boldsymbol{\theta})f(\boldsymbol{\theta}) \qquad (7)$$

$$f(\boldsymbol{\theta}) = \pi(\beta)\pi(\gamma)\pi(\alpha)\pi(\sigma)\pi(\boldsymbol{\Sigma})$$

Simulation study

- Simulate data from (5), in which the baseline intensity is set to be constant 1
- Consider different error distributions:
  - Scenario 1: ALD(0, 1, $\tau = 0.25$) (right-skewed);
  - Scenario 2: Standard normal distribution;
- For each scenario, simulate 200 data sets with $N = 250$ or 500 in each.
- Compare bias, standard error (SE), mean square error (MSE), and coverage probability (CP) for QRJM and the standard JM (LMJM).

Simulation results I

Table: Simulation result for Scenario 1 in which random error is generated from ALD$(0, 1, \tau = 0.25)$.

		QRJM ($\tau = 0.25$)				QRJM ($\tau = 0.50$)				LMJM			
		Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP
	Coefficients for longitudinal process												
	β_1	0.014	0.091	0.008	0.980	0.025	0.102	0.012	0.960	0.036	0.112	0.013	0.955
	β_2	−0.002	0.164	0.029	0.920	0.007	0.174	0.031	0.930	0.022	0.182	0.034	0.955
	β_3	0.033	0.068	0.005	0.940	0.046	0.083	0.009	0.890	0.058	0.095	0.012	0.890
$n = 250$	σ	−0.000	0.031	0.001	0.950	−0.321	0.021	0.103	0.000	–	–	–	–
	Coefficients for recurrent event process												
	γ	0.001	0.073	0.005	0.955	0.002	0.078	0.005	0.970	0.004	0.081	0.007	0.935
	r_0	0.032	0.134	0.018	0.945	−0.786	0.055	0.622	0.000	−0.915	0.032	0.838	0.000
	α	−0.007	0.071	0.005	0.950	−0.028	0.080	0.008	0.905	−0.030	0.090	0.009	0.920
	Coefficients for longitudinal process												
	β_1	−0.001	0.064	0.004	0.920	0.009	0.071	0.006	0.920	0.010	0.078	0.007	0.930
	β_2	−0.003	0.116	0.011	0.970	0.011	0.121	0.012	0.980	0.006	0.126	0.013	0.955
	β_3	0.020	0.048	0.003	0.950	0.026	0.058	0.004	0.950	0.029	0.067	0.005	0.935
$n = 500$	σ	0.001	0.022	0.001	0.970	−0.320	0.015	0.103	0.000	–	–	–	–
	Coefficients for recurrent event process												
	γ	0.007	0.052	0.004	0.920	0.007	0.056	0.004	0.920	0.007	0.058	0.004	0.915
	r_0	−0.017	0.093	0.008	0.940	−0.810	0.036	0.657	0.000	−0.929	0.020	0.863	0.000
	α	0.003	0.051	0.003	0.950	−0.001	0.059	0.004	0.940	0.004	0.068	0.004	0.940

Simulation results II

Table: Simulation result for Scenario 2 in which random error is generated from $\mathcal{N}(0,1)$.

		LMJM				QRJM ($\tau = 0.5$)			
		Bias	SE	MSE	CP	Bias	SE	MSE	CP
$n = 250$	Coefficients for longitudinal process								
	β_1	−0.015	0.076	0.005	0.950	−0.010	0.076	0.006	0.960
	β_2	−0.002	0.148	0.026	0.920	0.000	0.149	0.027	0.910
	β_3	0.004	0.038	0.001	0.970	0.003	0.038	0.002	0.920
	σ	0.009	0.047	0.002	0.960	—	—	—	—
	Coefficients for recurrent event process								
	γ	0.002	0.054	0.003	0.960	-0.009	0.053	0.003	0.930
	r_0	0.014	0.090	0.009	0.940	0.046	0.091	0.011	0.875
	α	0.010	0.048	0.002	0.930	-0.022	0.047	0.003	0.875
$n = 500$	Coefficients for longitudinal process								
	β_1	−0.006	0.053	0.003	0.920	0.000	0.054	0.003	0.930
	β_2	0.001	0.106	0.012	0.930	0.006	0.106	0.012	0.940
	β_3	0.010	0.026	0.001	0.920	0.009	0.027	0.001	0.920
	σ	0.003		0.000	0.960	—	—	—	—
	Coefficients for recurrent event process								
	γ	0.003	0.038	0.002	0.940	−0.007	0.037	0.002	0.930
	r_0	−0.009	0.063	0.005	0.910	0.022	0.063	0.006	0.900
	α	0.009	0.034	0.002	0.890	−0.014	0.033	0.002	0.850

Data application

- Data from the Atherosclerosis Risk in Communities Study (ARIC).
- Longitudinal outcome: systolic blood pressure (SBP); event outcome: coronary heart disease (CHD). Higher SBP leads to higher risk of CHD recurrences (Wattanakit et al., 2005; Rodriguez et al., 2014)
- Study cohort: 657 participants; 115, 31, and 17 patients experienced 1, 2 or ≥ 3 CHD events.
- Consider the following QRJM:

$$\begin{cases} sbp_i(t) = m_i(t) + \varepsilon_i(t) = \beta_0 + \beta_1 age_{0i} + \beta_2 chol_i + \beta_3 I_{med_i} + \beta_4 t + u_{i1} + u_{i2} t + \varepsilon_i(t) \\ r_i(t|\mathcal{M}_i(t); \gamma, \alpha) = r_0(t) v_i \exp(\gamma_1 I_{male_i} + \gamma_3 I_{smoke_i} + \gamma_4 I_{diabetes_i} + \alpha m_i(t)) \end{cases}$$

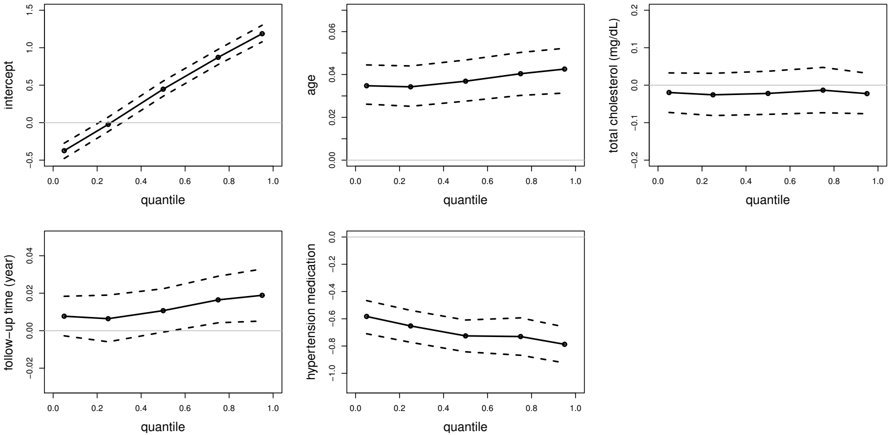
- $r_0(t)$: piecewise constant baseline intensity function with three intervals
- v_i is the frailty term that accounts for the correlation among the multiple event times within the same subject

Data analysis results I

Table: ARIC data analysis: Parameter estimation and 95% credible interval (in parenthesis) from QRJM at three quantiles.

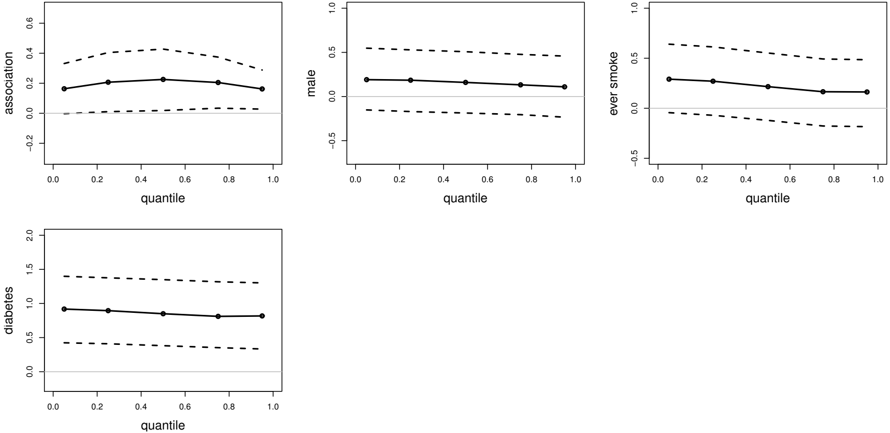
	$\tau = 0.05$	$\tau = 0.50$	$\tau = 0.95$
	longitudinal SBP process		
Intercept	−0.374 (−0.478, −0.274)	0.447 (0.352, 0.554)	1.187 (1.079, 1.300)
Age_0	0.035 (0.026, 0.044)	0.037 (0.028, 0.047)	0.043 (0.031, 0.052)
Total cholesterol (mg/dL)	−0.020 (−0.073, 0.033)	−0.022 (−0.078, 0.037)	−0.022 (−0.076, 0.032)
Hypertension medicine	−0.583 (−0.710, −0.467)	−0.725 (−0.842, −0.609)	−0.787 (−0.924, −0.660)
Follow-up time (yr)	0.008 (−0.003, 0.018)	0.011 (−0.001, 0.022)	0.019 (0.005, 0.033)
	recurrent CHD events process		
Association	0.163 (−0.003, 0.332)	0.226 (0.019, 0.428)	0.162 (0.028, 0.288)
Male	0.191 (−0.152, 0.548)	0.160 (−0.187, 0.507)	0.110 (−0.234, 0.458)
Ever smoke	0.291 (−0.044, 0.641)	0.216 (−0.121, 0.552)	0.163 (−0.184, 0.485)
Diabetes	0.918 (0.424, 1.399)	0.850 (0.381, 1.349)	0.818 (0.333, 1.301)

Data analysis results II



(a) Parameters in the longitudinal SBP process

Data analysis results III



(b) Parameters in the recurrent CHD events process

Figure: ARIC data analysis: Posterior mean (solid line) and point-wise 95% credible interval (dashed lines) of parameter estimation against different quantiles.

Discussion

- Our work on QRJM that uses an LQMM for the longitudinal process provides a more flexible way for simultaneously modeling conditional quantile of a longitudinal outcome and the risk of event recurrences.
- In the application of ARIC data our results reveal some findings that can not be observed using linear regression based method.
- Our novel extension of traditional JM finds practical importance in many clinical fields: cancer recurrences, hospital readmissions, etc.
- Other modeling format consideration: (i) nonlinear QR (Koenker and Park, 1996); (ii) accelerated failure time model when the proportionality assumption is violated.

Journal Article 3

Bayesian Quantile Regression Joint Models: Dynamic Predictions of Recurrent Event Probability

Background

- Disease recurrence is one of the important clinical outcomes in longitudinal biomedical studies.
- Accurate predictions of disease probability plays an important role in disease intervention and prevention.
- The JM framework offers a novel way of making such personalized dynamic predictions of future event probability (Rizopoulos, 2011; Taylor et al., 2013).
- Little work has been done on the dynamic predictions of event recurrences under the JM framework as far, especially the QR based JM

Predictions of Event-Free Probability

The predicted event-free probability (1−risk) at time m ($m > t$) given previous event times and longitudinal measurements up to time t is:

$$p_i(m|t) = Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}),$$

where $\mathcal{T}_{it-} = \{T_{ik} : 1 \leq k \leq K_i, T_{iK_i} < t\}$ are the recurrent times before time t.

With further derivation:

$$p_i(m|t) \quad = \quad \int \frac{Pr(T_{iK_i+1} \geq m | \mathcal{M}_i(m, \boldsymbol{u}_i; \boldsymbol{\theta}), \mathcal{T}_{it-}; \boldsymbol{\theta})}{Pr(T_{iK_i+1} > t | \mathcal{M}_i(t, \boldsymbol{u}_i; \boldsymbol{\theta}), \mathcal{T}_{it-}; \boldsymbol{\theta})} \cdot Pr(\boldsymbol{u}_i | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \qquad (8)$$

We approximate it by its posterior mean:

$$\begin{aligned} E_{\boldsymbol{\theta}|\mathcal{D}_N}[p_i(m|t)] &= Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t)) \\ &= \int Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_N) d\boldsymbol{\theta}. \end{aligned}$$

Estimation of the prediction

A Monte Carlo (MC) estimation of $p_i(m|t)$ can be obtained using the following procedure:

1. Draw $\boldsymbol{\theta}^{(p)}$ from the posterior distributions $Pr(\boldsymbol{\theta}|\mathcal{D}_N)$ for $p = 1, \cdots, P$;
2. For each of the P draws of $\boldsymbol{\theta}^{(p)}$, make Q draws of $\boldsymbol{u}_i^{(q)}, q = 1, \cdots, Q$, from the posterior distribution of random effects $Pr(\boldsymbol{u}_i|\mathcal{D}_N, \boldsymbol{\theta}^{(p)})$ and approximate $p_i(m|t)^{(p)}$ by

$$\frac{1}{Q}\sum_{q=1}^{Q}\frac{Pr(T_{iK_i+1} \geq m|\mathcal{M}_i(m, \boldsymbol{u}_i^{(q)}; \boldsymbol{\theta}^{(p)}), \mathcal{T}_{it-}; \boldsymbol{\theta}^{(p)})}{Pr(T_{iK_i+1} > t|\mathcal{M}_i(t, \boldsymbol{u}_i^{(q)}; \boldsymbol{\theta}^{(p)}), \mathcal{T}_{it-}; \boldsymbol{\theta}^{(p)})};$$

3. Approximate $p_i(m|t)$ by $\frac{1}{P}\sum_{p=1}^{P} p_i(m|t)^{(p)}$.

Simulation study

- Simulate data from following JM:

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 t + u_i + \varepsilon_i(t) \\ r_i(t|W_i; \gamma, \alpha) = r_{0i}(t) \exp(\gamma W_i + \alpha m_i(t)) \end{cases} \quad (9)$$

- A maximum of six observations for each subject at follow-up times $t = 0$, 0.25, 0.5, 0.75, 1.0, and 1.25. Also limit a maximum of five recurrent events for each subject.
- Consider two scenarios in error distribution: (i) $ALD(0, 1, 0.25)$; (ii) $\mathcal{N}(0, 1)$. 200 data sets for each scenario with $N = 500$ in each.
- Split the sample in to two parts: 400 (80%) are used to draw model inference and the rest 100 subjects are used to make out-of-sample dynamic predictions of event-free probability.

Simulation results I

Table: Simulation result for Scenario 1: MSE and bias of the difference between predicted event-free probability and the gold standard.

		QRJM ($\tau = 0.25$)		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias	MSE	Bias
0.25	0.25	0.028	0.001	0.035	0.067	0.033	0.023
	0.50	0.035	−0.006	0.045	0.079	0.043	0.024
	1.00	0.037	−0.021	0.048	0.074	0.046	0.015
0.5	0.25	0.022	0.002	0.029	0.067	0.026	0.011
	0.50	0.029	−0.005	0.039	0.078	0.036	0.007
	1.00	0.033	−0.018	0.043	0.077	0.040	−0.005
1.00	0.25	0.018	0.011	0.025	0.078	0.020	0.019
	0.50	0.023	−0.005	0.033	0.079	0.022	−0.004
	1.00	0.026	−0.016	0.036	0.078	0.022	−0.009

Simulation results II

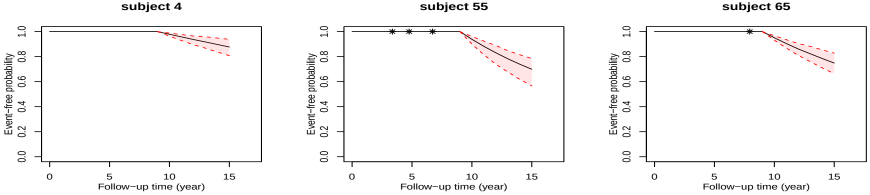
Table: Simulation result for Scenario 2: MSE and bias of the difference between predicted event-free probability and the gold standard.

		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias
0.25	0.25	0.015	−0.003	0.014	−0.001
	0.50	0.019	−0.007	0.018	−0.003
	1.00	0.020	−0.014	0.019	−0.010
0.5	0.25	0.012	0.001	0.011	0.002
	0.50	0.015	−0.004	0.014	−0.002
	1.00	0.016	−0.009	0.014	−0.007
1.00	0.25	0.009	0.006	0.008	0.005
	0.50	0.010	−0.004	0.009	−0.003
	1.00	0.010	−0.010	0.010	−0.009

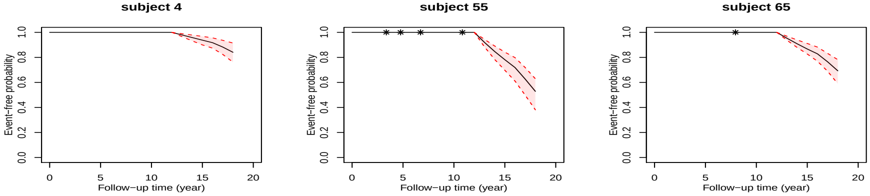
Data application

- Data from the Atherosclerosis Risk in Communities Study (ARIC).
- Longitudinal outcome: systolic blood pressure (SBP); recurrent events: coronary heart disease (CHD).
- Select 80% of the study cohort (i.e. 526 subjects) to draw parameter inference, and make predictions of CHD event probability for the rest 131 individuals.

Dynamic predictions of CHD risk I

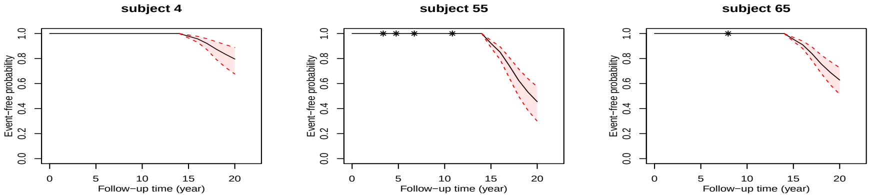


(a) Predictions based on follow-up time $t = 9$



(b) Predictions based on follow-up time $t = 12$

Dynamic predictions of CHD risk II



(c) Predictions based on follow-up time $t = 14$

Figure: ARIC data analysis: Dynamic predictions of CHD event-free probability, based on various follow-up time and prediction time window, with 95% credible interval from QRJM at $\tau = 0.5$ for selected subjects ($*$ indicates CHD event).

Dynamic predictions of CHD risk III

Table: ARIC data analysis: AUC, AARD and MRD of the predictions of CHD event-free probability from QRJM and AUC from LMJM.

t	Δt	AUC (τ)			AARD (τ)			MRD (τ)			AUC (LMJM)
(year)		0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	
9	1	0.726	0.713	0.712	0.357	0.327	0.327	0.035	0.032	0.034	0.717
	2	0.685	0.671	0.670	0.286	0.255	0.255	0.028	0.024	0.025	0.676
	3	0.669	0.654	0.654	0.257	0.227	0.228	0.027	0.022	0.023	0.659
12	1	0.770	0.756	0.754	0.434	0.402	0.400	0.056	0.053	0.053	0.761
	2	0.721	0.703	0.703	0.345	0.303	0.304	0.044	0.039	0.039	0.710
	3	0.699	0.680	0.680	0.307	0.266	0.267	0.042	0.035	0.035	0.687
14	1	0.797	0.784	0.784	0.487	0.463	0.464	0.071	0.068	0.069	0.789
	2	0.748	0.731	0.732	0.394	0.355	0.357	0.059	0.054	0.054	0.738
	3	0.714	0.695	0.697	0.331	0.288	0.293	0.059	0.049	0.050	0.704

Discussion

- The idea of personalized dynamic predictions of recurrent event risk finds its practical importance in disease control and prevention.
- Our novel extension of traditional JM with LQMM adds more flexibility to the modeling framework and allows us to investigate specific subgroup of patients of interest.
- The current version of QRJM uses LQMM and Cox PHM for the longitudinal and recurrent event processes respectively. However, other functional forms for both outcomes can also be considered to extend the proposed method.
- The best predictive performance from our model outperforms that from the LMJM. Selection of quantile in prediction or how to combine prediction results from different quantiles can be a topic for future work.

Acknowledgement

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

Sheng Luo, PhD (Dissertation Supervisor)

David R. Lairson, PhD (Minor Advisor)

Xiaoming Liu, PhD (Breadth Advisor)

External reviewer:
Soeun Kim, PhD

Thanks to all my friends and colleagues at UTSPH and MDACC!

Thanks to the Texas Advanced Computing Center (TACC) for providing high-performing computing resources.

рахмат
danke
謝謝
ngiyabonga
شكراً جزيلاً
Баярлалаа
mersi
kia ora
barka
welalin
tack
spas
teşekkür ederim
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
nanni
nandri
kiitos
dankie
dhanyavad
mauruuru
köszönöm
thank you
gracias
хвала
asante
manana
obrigada
tenki
murakoze
enkosi
bedankt
bayarlalaa
gràçie
hvala
dziękuję
akun
dankon
aciu
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
obrigado
mèsi
didi madloba
sağolun
sukriya
kop khun krap
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
najis tuke
kam sah hamnida
rahmat
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
mercé
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci

Selected references I

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

Paulsen, J. S., Long, J. D., Ross, C. A., Harrington, D. L., Erwin, C. J., Williams, J. K., Westervelt, H. J., Johnson, H. J., Aylward, E. H., Zhang, Y., et al.
Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study.
The Lancet Neurology, 13:1193–1201, 2014.

Kotz, S., Kozubowski, T., and Podgorski, K.
The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Exonomics, Engineering, and Finance.
Springer, 2001.

Selected references II

Plummer, M. et al.
JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.
In Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), pages 20–22, March 2003.

Koenker, R.
Quantile Regression.
Cambridge University Press, 2005.

Farcomeni, A. and Viviani, S.
Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint models.
Statistics in Medicine, 2015;34(7):1199âĂŞ1213.

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

Selected references III

Efendi, A., Molenberghs, G., Njagi, E. N., and Dendale, P.
A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation.
Biometrical Journal, 55(4):572–588, 2013.

Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H.
Real-time individual predictions of prostate cancer recurrence using joint models.
Biometrics, 69(1):206–213, 2013.

Wattanakit K, Folsom AR, Chambless LE, Nieto FJ.
Risk factors for cardiovascular event recurrence in the Atherosclerosis Risk in Communities (ARIC) study.
American Heart Journal. 2005;149(4):606–612.