Model estimation and dynamic prediction for subject-specific event probability in joint modeling using longitudinal quantile regression

Ming Yang, Sheng Luo, Stacia DeSantis

Department of Biostatistics
The University of Texas School of Public Health

@JSM2016
August 2

Outline

Introduction

A motivating data

Joint models for longitudinal and survival data

Research questions

Statistical methods

Review of quantile regression

JM using longitudinal quantile regression

Dynamic prediction of event-free probabilities

Predictive performance

Simulation studies

Model inference

Dynamic prediction

Data application

Data and model

Results

Discussion & Conclusion

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 measure 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).

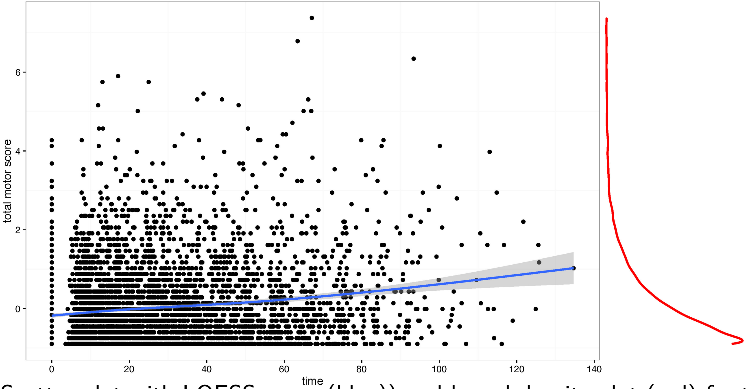


Figure: Scatter plot with LOESS curve (blue)) and kernel density plot (red) for total motor score from the study population (time unit: month; lower total motor score is better).

Joint models for longitudinal and survival 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 process
- Proportional hazards model (PHM) for the survival process
- Longitudinal outcome is treated as a time-dependent covariate in the survival submodel

Limitations of traditional JM

- LMM is sensitive to outliers and deviation of normality
- The normality of the error term 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 questions

1. How to deal with the non-normality in the data?
2. How to study the covariates effect on the higher/lower tail of the biomarkers?
3. How to make predictions of HD-free probability in the future? Dynamically?

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:

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

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

Parameter estimation from QR vs. mean regression

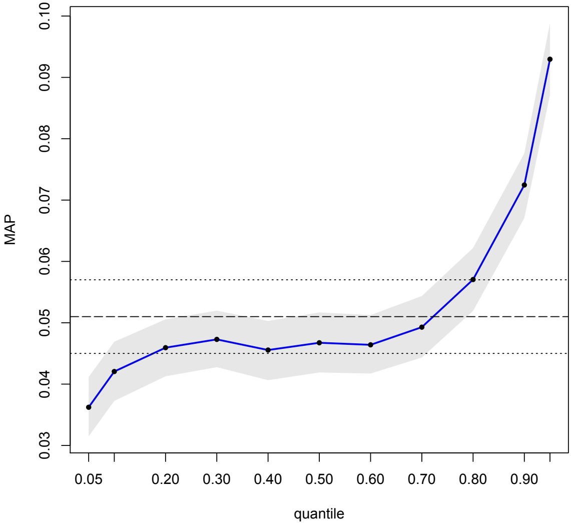


Figure: Quantile effect v.s. mean effect

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.

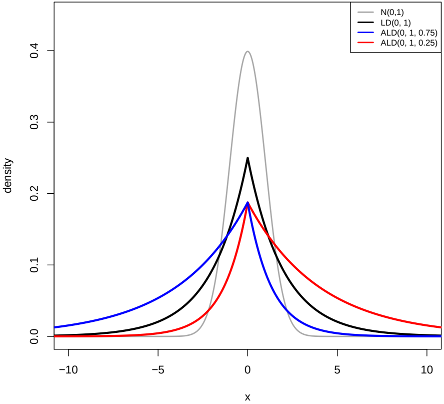


Figure: Asymmetric Laplace, Laplace, and normal distributions

Quantile regression joint models (QRJM)

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

where:

- $m_i(t)$: 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 prediction of future event-free probabilities

- 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

- Generate Monte Carlo samples 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 $p_i^{(k)}(m|t) = S_i[m|\mathcal{M}_i(m, u_i^{(k)}, \boldsymbol{\theta}^{(k)}); \boldsymbol{\theta}^{(k)}]S_i[t|\mathcal{M}_i(t, u_i^{(k)}, \boldsymbol{\theta}^{(k)}); \boldsymbol{\theta}^{(k)}]^{-1}$
- Approximate the true value using sample mean or median:

$$\hat{p}_i(m|t) = \frac{1}{K}\sum_{k=1}^{K} p_i^{(k)}(m|t). \tag{5}$$

Statistics to assess predictive performance

- Let $\hat{r}_i(t+\Delta t|t) = 1 - \hat{S}_i(t+\Delta t|\mathcal{Y}_i(t), \boldsymbol{u}_i; \boldsymbol{\theta}), i = 1, \cdots, N$.
- Then

$$\widehat{TPR}_t^{\Delta t}(p) = \frac{\sum_{i=1}^{N} \hat{r}_i(t+\Delta t|t) I(\hat{r}_i(t+\Delta t|t) \geq p)}{\sum_{i=1}^{N} \hat{r}_i(t+\Delta t|t)},$$

$$\widehat{FPR}_t^{\Delta t}(p) = \frac{\sum_{i=1}^{N} (1 - \hat{r}_i(t+\Delta t|t))\, I(\hat{r}_i(t+\Delta t|t) \geq p)}{\sum_{i=1}^{N} (1 - \hat{r}_i(t+\Delta t|t))}.$$

- We use the following three statistics as measures of predictive performance:

$$\widehat{AUC}_t^{\Delta t} = \int \widehat{TPR}_t^{\Delta t} \left\{ (\widehat{FPR}_t^{\Delta t})^{-1}(u) \right\} du,$$

$$\widehat{AARD}_t^{\Delta t} = \widehat{TPR}_t^{\Delta t}(\hat{\rho}) - \widehat{FPR}_t^{\Delta t}(\hat{\rho}),$$

$$\widehat{MRD}_t^{\Delta t} = \int_p \widehat{TPR}_t^{\Delta t}(p)dp - \int_p \widehat{FPR}_t^{\Delta t}(p)dp,$$

where $\hat{\rho} = \frac{\sum_{i=1}^{N} \hat{r}_i(t+\Delta t|t)}{N}$ is the average risk in the study population at time $t + \Delta t$.

Simulation study 1: model inference

- Simulate data from Equation (3) and consider the following three scenarios:
    1. Scenario 1: choose $\tau = 0.25$ for the ALD
    2. Scenario 2: choose $\tau = 0.5$ for the ALD (i.e, median equals 0)
    3. Scenario 3: random error follows standard normal distribution
- 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 2.
- Compare the bias, standard error (SE), mean square error (MSE), and coverage probability (CP) for QRJM and the standard JM (LMJM).

Simulation 1 results

Table: Simulation study: Inference results for data generated from ALD with $\tau = 0.25$ (Scenario 1).

	QRJM ($\tau = 0.25$, true model)				QRJM ($\tau = 0.5$)				LMJM			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP
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
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_{11}	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_{12}	0.016	0.083	0.013	0.950	0.014	0.112	0.042	0.970	0.078	0.128	0.052	0.920
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

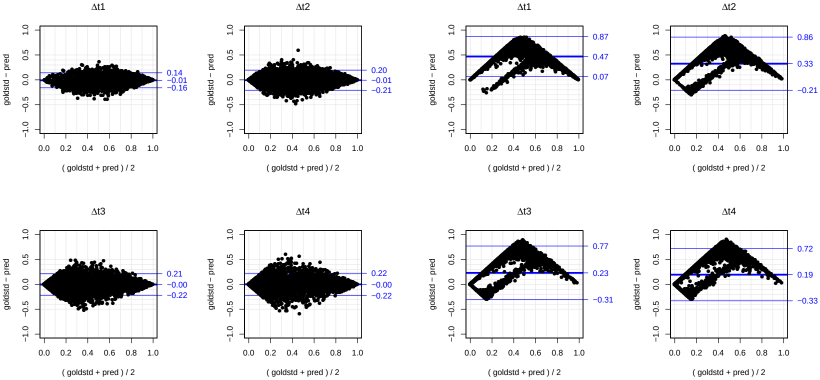
Table: Inference results for data generated from $N(0, 1)$ (Scenario 3).

	QRJM ($\tau = 0.5$)				LMJM (true model)			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP
α	-0.013	0.055	0.006	0.95	0.007	0.055	0.006	0.95
β_0	0.015	0.037	0.003	0.95	0.000	0.035	0.002	0.98
β_{11}	0.004	0.034	0.002	0.96	-0.003	0.033	0.002	0.95
β_{12}	0.013	0.050	0.005	0.95	0.006	0.049	0.005	0.95
γ_1	0.008	0.055	0.006	0.92	0.003	0.054	0.006	0.90
γ_2	0.015	0.055	0.007	0.92	0.010	0.054	0.006	0.92

Simulation 2: dynamic prediction

- 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 2 results: Bland Altman plot

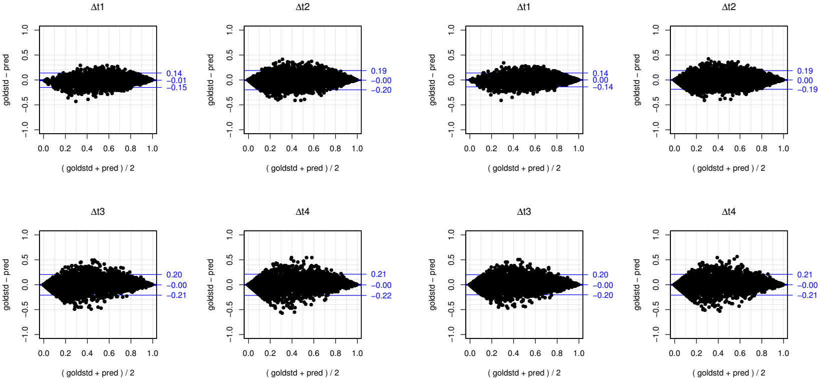


(a) QRJM with $\tau = 0.25$ (True model)

(b) LMJM

Figure: Bland-Altman plot (bias and 95% limits of agreement) of gold standard versus model predictions based on the first two longitudinal observations and four different prediction time intervals ($\Delta t_1 < \Delta t_2 < \Delta t_3 < \Delta t_4$) under Scenario 1.

Simulation 2 results: Bland Altman plot (Cont'd)



(a) QRJM with $\tau = 0.5$ (b) LMJM (True model)

Figure: Bland-Altman plot (bias and 95% limits of agreement) of gold standard versus model predictions based on the first two longitudinal observations and four different prediction time intervals ($\Delta t_1 < \Delta t_2 < \Delta t_3 < \Delta t_4$) under Scenario 3.

Simulation 2 results: summary table

Table: Simulation study: MSE and bias of the difference between predicted survival probability and the gold standard (Scenario 1).

		QRJM ($\tau = 0.25$)		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias	MSE	Bias
0.25 (subjects left: 48.1%)	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
	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 (subjects left: 34.6%)	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
	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 (subjects left: 22.8%)	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
	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:

$$\left\{\begin{array}{l} y_i(t) = m_i(t) + \varepsilon_{it} = \beta_0 + \beta_1 t + \beta_2 age_{0i} + u_{i1} + u_{i2} t + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ h(T_i|\mathcal{M}_i(T_i); \gamma, \alpha) = \sum_{k=1}^{3} \lambda_k I_k(T_i) \exp(\gamma_1 education_i + \gamma_2 I_{male_i} + \alpha m_i(T_i)) \end{array}\right.$$

- $y_i(t)$ 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 is 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.

Results

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

	total motor score		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
longitudinal 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-event 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)

Results (Cont'd)

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 & Conclusion

- 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.

Acknowledgement

Sheng Luo Ph.D. and Stacia DeSantis Ph.D. are co-advisers of Ming Yang's dissertation work at UTSPH.

We would like to acknowledge the Texas Advanced Computing Center (TACC) for providing high-performing computing resources.

Selected references

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 (Cont'd)

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.

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

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 modelss.
Biometrics, 69(1):206–213, 2013.

Thank you!

Questions?