Contents

1 Background

Longitudinal studies are ubiquitous in biostatistics context. As a typical example, in randomized clinical trials (RCTs), patients are randomly allocated into different treatment arms and followed over time. Repeated measurements of interest will then be taken from those patients at designed follow-up time points. One of the important features of longitudinal data is that repeated measurements from the same subject are more "similar" to each other compared to those measures from different subjects, i.e. within subject measures tend to be intercorrelated. This feature requires special statistical techniques to handle the correlation thus valid scientific inference can be drawn from the data. As discussed in (Diggle et al., 2002), there are mainly three methods we can use to analyze longitudinal data: marginal model, transition model and random effects model. Different models lead to different interpretations of the regression coefficients and the

choice of a model depends on objectives of the study, the source of correlation as well as the areas where they can be applied. In this work we will focus on applying random effects model to longitudinal data. A model that contains both random effects and fixed effects is called mixed effects model. The mixed effects model methodology, first introduced by R.A. Fisher (Fisher, 1919), is a statistical tool that is used across a wide variety of disciplines including biostatistical contexts. Mixed effects models are especially popular in researches involving repeated measurements or observations from multilevel (or hierarchical) structure where the correlation between observations is not negligible as discussed above.

Linear mixed model (LMM) is a widely used application of mixed effects methods. In brief, LMM models the expected value of the outcome and assumes observations from the same subject share a same latent variable, i.e. random effect, to account for the correlation among those observations. When conditional random effects, those observations are assumed to be independently distributed.

Despite the popularity of LMM, in many circumstance, i.e. when there exits outliers or skewness in the outcome, the normality assumption of the error term can not be satisfied (even after trying various ways of transformation) thus LMM is not appropriate to use. In other cases, the conditional mean may not be the primary interest and researchers may also interested the covariate effects on the lower/upper quantiles of the outcome. Quantile regression is a single method that can fit all above needs at once and it's becoming more and more popular in the statistical community in recent years. Compared with the ubiquitous mean regression (a.n.a. linear regression) models, quantile regression models provide a much more comprehensive and focused insight into the association between the variables by studying the conditional quantiles functions of the outcome, which may not be observed by looking only at conditional mean models (Koenker, 2005). In quantile regression, the regression coefficients (β) are functions of the quantile (τ), and their estimates vary according to different quantiles. Thus quartile regression provides a way to studying the heterogeneity of the outcome that is associated with the covariates (Koenker, 2005). Moreover, as mentioned above, quantile regression is more robust against outliers in the outcome compared with the mean regression, which is an immediate extension from the property of quantiles.

For a linear quantile regression model, (Koenker and Bassett Jr, 1978) introduced a method in estimating the conditional quantiles. As an introductory material, Koenker (Koenker and Hallock, 2001) briefly covers the fundamentals of quantile regression, parameter estimation techniques, inference, asymptotic theory, etc., his book (Koenker, 2005) provides more comprehensive and deeper introduction on quantile regression related topics. Yu and Moyeed (Yu and Moyeed, 2001) introduced the idea of Bayesian quantile regression

by modeling the error term using asymmetric Laplace distribution (ALD) followed the idea proposed in (Koenker and Bassett Jr, 1978). (Kozumi and Kobayashi, 2011) developed a Gibbs sampling algorithm for Bayesian quantile regression models, in which they used a location-scale mixture representation of the ALD. Many works have been done to extend the quantile regression method to accommodate longitudinal data. (Jung, 1996) developed a quansi-likehood method for median regression model for longitudinal data. (Geraci and Bottai, 2007) proposed to fit the quantile regression for longitudinal data based on ALD and the estimation is made by using Monte Carlo EM algorithm, later on (Liu and Bottai, 2009) followed the idea of (Geraci and Bottai, 2007) and extended the model from random intercept to include random regression coefficients as well. (Fu and Wang, 2012) proposed a working correlation model for quantile regression for longitudinal data, a induced smoothing method was used to make the inference of the estimators. Fully Bayesian techniques were also developed for fitting linear quantile mixed models, for example (Luo et al., 2012) used the similar idea as (Kozumi and Kobayashi, 2011) did in decomposing the error term as a location-scale mixture and developed a Gibbs sampling algorithm for quantile linear mixed model. The fully Bayesian method is appealing because it is easy to implement and to make inference, the uncertainty of the unknowns is taken into account, and it is flexible in the distribution of random effects. The detailed background about Bayesian quantile linear mixed model will be introduced in Section **??**.

Another concern that frequently arises from longitudinal study is that the subjects may be lost at follow up due to events. Such cases are called informative drop-outs (ID) when the event occurrences are associated with the longitudinal outcome. Simply ignore the missing mechanism can lead to biased estimates of the parameters in the longitudinal model. The informative drop-out problem has attracted much attention from the statisticians and a wide array of methods have been proposed to handle this issue (Diggle and Kenward, 1994) (Lipsitz et al., 1997) (Touloumi et al., 2003) (Yuan and Yin, 2010). For more information in this research area, there are several good review papers of modeling longitudinal data with non-ignorable drop-outs, including (Diggle et al., 2007) and (Hogan et al., 2004). *[To-do: put some review of previous work and their limitations and propose our model then state the advantage.]*

[Add background for topic two here.]

To sum up, this research project focuses on developing and applying new statistical methods in analyzing longitudinal data and will cover the following topics. First, a fully Bayesian framework in modeling conditional quantiles of the longitudinal outcome, incorporating with time-to-event process to account for the informative drop-outs issue. In this part, the linear quantile mixed model is used so that the within subject

correlation is accounted and the results also will be more robust to potential outlying observations. Since we used quantile regress instead of mean regression, a more comprehensive insight about the association between the outcome and covariates will be learned. In this model we can also directly gain a sense of the association between the longitudinal and time-to-event processes. Second, *[to-do]*.

2 Specific amis and hypothese

3 Data

4 Methods

$$\begin{cases} y_{it} = \boldsymbol{\beta}^\top \boldsymbol{X}_{it} + \boldsymbol{\delta}^\top \boldsymbol{H}_{it} + \boldsymbol{u}_i^\top \boldsymbol{Z}_{it} + \varepsilon_{it} = \widetilde{\tau}_{it} + \varepsilon_{it} \\ h(T_i|\mathcal{T}_{iT_i}, \boldsymbol{W}_i; \boldsymbol{\gamma}, \alpha_1, \alpha_2) = h_0(T_i) \exp(\boldsymbol{\gamma}^\top \boldsymbol{W}_i + \alpha_1 \boldsymbol{\delta}^\top \boldsymbol{H}_{iT_i} + \alpha_2 \boldsymbol{u}_i^\top \boldsymbol{Z}_{iT_i}) \end{cases} \tag{1}$$

5 Plan for simulation study

The simulation study is designed to check the validity of our propsed joint model in modeling the longitudinal outcome and the informative drop-out event time. Our focus of the simulation results lies on the accuracy of our estimation, i.e. bias, and the precision, i.e. standard deviation, of the samples from posterior distribution. Comparision will be made between our posprosed model against the model that simply ignores the underlying informative drop-out mechanism.

$$\begin{cases} y_{it} = \boldsymbol{\beta}^\top \boldsymbol{X}_{it} + \boldsymbol{\delta}^\top \boldsymbol{H}_{it} + \boldsymbol{u}_i^\top \boldsymbol{Z}_{it} + \varepsilon_{it} = \widetilde{\tau}_{it} + \varepsilon_{it} \\ h(T_i|\mathcal{T}_{iT_i}, \boldsymbol{W}_i; \boldsymbol{\gamma}, \alpha_1, \alpha_2) = h_0(T_i) \exp(\boldsymbol{\gamma}^\top \boldsymbol{W}_i + \alpha_1 \boldsymbol{\delta}^\top \boldsymbol{H}_{iT_i} + \alpha_2 \boldsymbol{u}_i^\top \boldsymbol{Z}_{iT_i}) \end{cases} \tag{2}$$

6 Simulation settings

Following (Farcomeni and Viviani, 2014), by varying the values of the association parameters α_1 and α_2 in our model (2), we will have four different settings of simulation study, which are:

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, $\boldsymbol{H}_{it}$ 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

In model (2) note that we have specific covariates $\boldsymbol{X}_{it}$ only for the longitudinal model and covariates $\boldsymbol{W}_{it}$ only for the survival model.

Under different combinations of $(\alpha_1,\ \alpha_2)$, for the regression coefficients we choose $\boldsymbol{\beta, \delta, \gamma} = (1, 1)^\top$, the covriates $\boldsymbol{Z}_{it} = (1, t)^\top$, $\boldsymbol{H}_{it} = (h_{i1}, h_{i2} * t)^\top$, $\boldsymbol{X}_{it} = (1, x_i)^\top$, and $\boldsymbol{W}_i = (w_{i1}, w_{i2})^\top$ with $h_{i1}, h_{i2}, x_i, w_{i1}$ and w_{i2} generated from independent standard normal distributions, and the random effects $\boldsymbol{u}_i$ from bivariate normal with mean 0, standard deviations equal 0.3 and correlation 0.16. We also fix $\sigma = 1$ and vary the quantile τ among $\{0.25, 0.5, 0.75\}$ for the ALD specification when simulating longitudinal data.

To simulate the survival time data, for simplicity, we fix $h_0(s) = 1$ and obtain the survival distribution as

$$S(t|\boldsymbol{u}_i, \boldsymbol{H}_{it}, \boldsymbol{W}_i) = \exp\left\{-\frac{e^{\alpha_1(\delta_1 H_{i1} + \delta_2 H_{i2t}) + \alpha_2(u_{i1} + u_{i2}t) + \boldsymbol{\gamma}^\top \boldsymbol{W}_i} - e^{\alpha_1 \delta_1 H_{i1} + \alpha_2 u_{i1} + \boldsymbol{\gamma}^\top \boldsymbol{W}_i}}{\alpha_2 u_{i2} + \alpha_1 \delta_2 h_i}\right\}$$

when $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$ and

$$S(t|\boldsymbol{u}_i, \boldsymbol{H}_{it}, \boldsymbol{W}_i) = \exp\{-te^{\boldsymbol{\gamma}^\top \boldsymbol{W}_i}\}$$

when $\alpha_1 = \alpha_2 = 0$. We then can obtain event time T_i by inverting above survival function after generating n random variates from standard uniform distribution. To obtain a censoring proportion around 25%, we choose the censoring time $C_i/5$ be distributed according to $beta(4, 1)$.

To simulate the longitudinal data, we draw them independently from the ALD for the $\tau-$th quantile, centered on

$$\boldsymbol{\beta}^\top \boldsymbol{X}_{it} + \boldsymbol{\delta}^\top \boldsymbol{H}_{it} + \boldsymbol{u}_i^\top \boldsymbol{Z}_{it},$$

and with dispersion parameter σ. We keep maximum six observations for each subject at follow-up time $t = (0, 0.25, 0.5, 0.75, 1, 3)$ respectively, after incorporating the drop-out information.

7 Result

τ=0.25												
			$\boldsymbol{\beta}$		$\boldsymbol{\delta}$		$\boldsymbol{\gamma}$		α_1		α_2	
n	α_1	α_2	bias	s.d.	bias	s.d.	bias	s.d.	bias	s.d.	bias	s.d.
250	0	0										
250	0	1										
250	1	0										
250	1	1										

Table 1: An example table.

A Appendix: simulation code and model files

A.1 R code to simulate data

```
library(LaplacesDemon)
library(MASS)

#################################################################
############ function to simulate survival time ###############
#################################################################
# survival function is given by: S(t)= exp(- exp(B) * (exp(A*t) - 1) ) / A)
sim_Ti = function(n=500, alpha, delta=c(1,1), gamma=c(1,1)){
  Time = numeric(n)
  S = runif(n) # survival probability
  H = matrix(rnorm(2*n), ncol=2)
  W = matrix(rnorm(2*n), ncol=2)
  # random effects
  U = mvrnorm(n, mu=c(0,0), Sigma=matrix(c(0.09, 0.09*0.16, 0.09*0.16, 0.09),
    nrow=2, byrow=T))
  attributes(U)[[2]]=NULL # remove 'dimnames' attribute
```

```
16
17   # calcualate survival time
18   if(alpha[1]==0 & alpha[2]==0) Time = - log(S) / exp(gamma %*% t(W))
19
20   else{
21     B = alpha[1] * delta[1] * H[,1] + alpha[2] * U[,1] + gamma %*% t(W)
22     A = alpha[2] * U[,2] + alpha[1] * delta[2] * H[,2]
23     Time = log(1-log(S)*A/exp(B))/A
24   }
25
26   Ti_id = which(!is.na(Time))
27   Time = Time[Ti_id][1:250] # true survival time: take the first 250 that are
      not NA
28   Ci = rbeta(250, 4, 1)*2 # censoring time
29   Ti = pmin(Time, Ci) # observed survival time: choose the smaller one
30   event = as.numeric(Time == Ti) # 1 for event, 0 for censor
31   U = U[Ti_id, ][1:250, ]
32   H = H[Ti_id, ][1:250, ]
33   W = W[Ti_id, ][1:250, ]
34
35   list(Ti=Ti, event=event, H=H, U=U, W=W)
36 }
37
38 ################################################################
39 ###### function to simulate longitudinal data #################
40 ################################################################
41 sim_longitudinal_data = function(survival_data=surdata, n=250, time=c(0, 0.25,
       0.5, 0.75, 1, 3), tau, sigma=1, beta=c(1,1), delta=c(1,1)){
42   # survival_data - data simulated from survival model
43   # n - # of subjects
44   # time - time points of observations
```

```
45   # tau - quantile
46   # sigma - scale parameter for ALD
47   time = time # at most # = length(time) observations per patient
48   y = matrix(NA, nrow=n, ncol=length(time)) # wide format
49   Ti = survival_data$Ti
50   U = survival_data$U # random effects
51   H = survival_data$H
52   X = cbind(1, rnorm(n))
53   count = sapply(Ti, function(x) sum(x > time)) # number of observations after
       drop-outs
54
55   for (i in 1:n){
56     for (j in 1:count[i]){
57       location = beta %*% X[i, ] + delta %*% c(H[i,1], H[i,2]*time[j]) + U[i,]
       %*% c(1, time[j])
58       y[i,j] = ralaplace(1, location, scale=sigma, kappa=tau)
59     }
60   }
61
62   list(y = y, X = X, J=count)
63 }
64
65 ################################################################
66 ###### function to simulate multiple joint data sets ##########
67 ################################################################
68 sim_multiple_data = function(N, sur_fun=sim_Ti, longi_fun=sim_longitudinal_
     data, alpha, tau){
69   # N - number of data sets to generate
70   # sur_fun - function to simulate survival data
71   # longi_fun - function to sumualte longitudinal data
72   # alpha - association parameters for JM
```

```
  # tau - quantile

  outdata = vector(mode='list', N)
  for (i in 1:N){
    sur_data = sur_fun(alpha=alpha)
    longi_data = longi_fun(sur_data, tau=tau)
    outdata[[i]] = list(survival_data=sur_data, longitudinal_data=longi_data)
  }
  outdata
}
```

A.2 JAGS model file

```
model{
  zero[1] <- 0
  zero[2] <- 0
  k1 <- (1-2*qt)/(qt*(1-qt))
  k2 <- 2/(qt*(1-qt))

  for (i in 1:I){
  # prior for random effects
  u[i, 1:2] ~ dmnorm(zero[], precision[,])

  # longitudinal process, BQR mixed model using ALD representation
    for (j in 1:J[i]){
      er[i,j] ~ dexp(sigma)
      mu[i,j] <- u[i,1] + u[i,2]*t[j] + inprod(X[i,], beta[]) + delta[1]*H[i
    ,1] + delta[2]*H[i,2]*t[j] + k1*er[i,j]
      prec[i,j] <- sigma/(k2*er[i,j])
      y[i,j] ~ dnorm(mu[i,j], prec[i,j])
    } #end of j loop

```

```
# survival process, baseline hazard is set to 1
A[i] <- alpha2*u[i,2] + alpha1*delta[2]*H[i,2]
B[i] <- alpha1*delta[1]*H[i,1] + alpha2*u[i,1] + inprod(gamma, W[i,])
S[i] <- exp(- exp(B[i])*(pow(exp(A[i]), Ti[i])-1)/A[i])
h[i] <- exp(inprod(gamma, W[i,]) + alpha1*(delta[1]*H[i,1] + delta[2]*H[i
  ,2]*Ti[i]) + alpha2*(u[i,1] + u[i,2]*Ti[i]))
L[i] <- pow(h[i], event[i])*S[i]/1.0E+08

# zero trick
phi[i] <- -log(L[i])
zeros[i] ~ dpois(phi[i])

}#end of i loop

precision[1:2,1:2] <- inverse(Sigma[,])
Sigma[1,1] <- 1
Sigma[1,2] <- rho*sig1
Sigma[2,1] <- Sigma[1,2]
Sigma[2,2] <- sig1*sig1

# priors for other parameters
alpha1 ~ dnorm(0, 0.001)
alpha2 ~ dnorm(0, 0.001)
beta[1] ~ dnorm(0, 0.001)
beta[2] ~ dnorm(0, 0.001)
delta[1] ~ dnorm(0, 0.001)
delta[2] ~ dnorm(0, 0.001)
gamma[1] ~ dnorm(0, 0.001)
gamma[2] ~ dnorm(0, 0.001)
sigma ~ dgamma(0.001, 0.001)
rho ~ dunif(-1, 1)
```

```
49    sig1 ~  dgamma(0.01, 0.01)
50 }
```

References

Peter Diggle and Michael G Kenward. Informative drop-out in longitudinal data analysis. *Applied statistics*, pages 49–93, 1994.

Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

Peter Diggle, Daniel Farewell, and Robin Henderson. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(5): 499–550, 2007.

Alessio Farcomeni and Sara Viviani. Longitudinal quantile regression in presence of informative drop-out through longitudinal-survival joint modeling. *arXiv preprint arXiv:1404.1175*, 2014.

Ronald A Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(02):399–433, 1919.

Liya Fu and You-Gan Wang. Quantile regression for longitudinal data with a working correlation model. *Computational Statistics & Data Analysis*, 56(8):2526–2538, 2012.

Marco Geraci and Matteo Bottai. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154, 2007.

Joseph W Hogan, Jason Roy, and Christina Korkontzelou. Handling drop-out in longitudinal studies. *Statistics in medicine*, 23(9):1455–1497, 2004.

Sin-Ho Jung. Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 91(433):251–257, 1996.

Roger Koenker. *Quantile regression*. Cambridge university press, 2005.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Kevin Hallock. Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56, 2001.

Hideo Kozumi and Genya Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578, 2011.

Stuart R Lipsitz, Garrett M Fitzmaurice, Geert Molenberghs, and Lue Ping Zhao. Quantile regression methods for longitudinal data with drop-outs: Application to cd4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46 (4):463–476, 1997.

Yuan Liu and Matteo Bottai. Mixed-effects models for conditional quantiles with longitudinal data. *The International Journal of Biostatistics*, 5(1), 2009.

Youxi Luo, Heng Lian, and Maozai Tian. Bayesian quantile regression for longitudinal data models. *Journal of Statistical Computation and Simulation*, 82(11):1635–1649, 2012.

G Touloumi, AG Babiker, MG Kenward, SJ Pocock, and JH Darbyshire. A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modelled with a time-to-event outcome. *Statistics in medicine*, 22(20):3161–3175, 2003.

Keming Yu and Rana A Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4): 437–447, 2001.

Ying Yuan and Guosheng Yin. Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, 66(1):105–114, 2010.