New Approaches for Quantile Regression

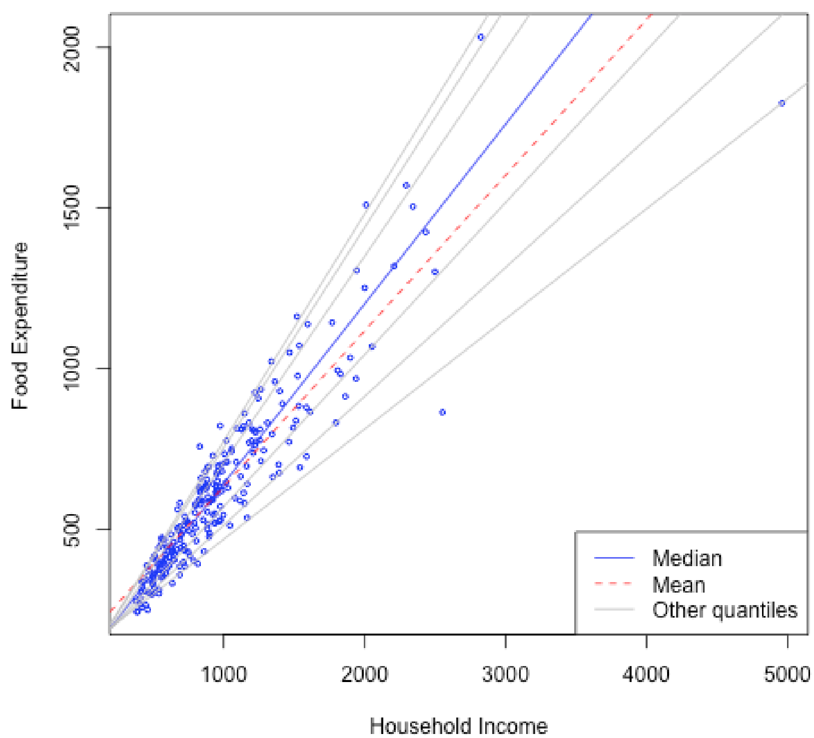
Ph. D. Dissertation Proposal

Minzhao Liu
Supervisor: Dr. Mike Daniels. Department of Statistics, University of Florida

Outline

- Introduction and Review
- Chapter 1: Bayesian Quantile Regression Using a Mixture of Polya Trees Priors
- Chapter 2: Quantile Regression in the Presence of Monotone Missingness with Sensitivity Analysis

Why Quantile Regression



- Engel data on food expenditure vs household income for a sample of 235 19th century working class Belgian households.
- τ: 5%, 10%, 25%, 75%, 90%, 95%
- Median regression
- Mean regression
- Increasing trend from mean regression
- More info from QR
  - Slope change
  - Skewness
- Less sensitive to heterogeneity and outliers

Introduction of Quantile Regression

Quantile (unconditional)

$$Q_Y(\tau) = \inf\{y : F(y) \geq \tau\},$$

Quantile Regression (conditional with covariates)

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}'\beta(\tau).$$

Quantile Regression vs Mean Regression

1. More information about the relationship of covariates and responses
2. Slope may vary for different quantiles
3. Can focus on certain quantiles as estimates of interest
4. More complete description of the conditional distribution

Traditional Frequentist Methods

- R package `quantreg` (Koenker, 2012)
- Using simplex for linear programming problems mentioned in Koenker et al. (1978)

$$\beta(\tau) = \arg\min_b \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i{}'b)$$

- No distributional assumptions
- Fast using linear programming
- Asymptotic inference may not be accurate for small sample sizes
- Easy to generalize:
  - Random effects
  - L_1 , L_2 penalties

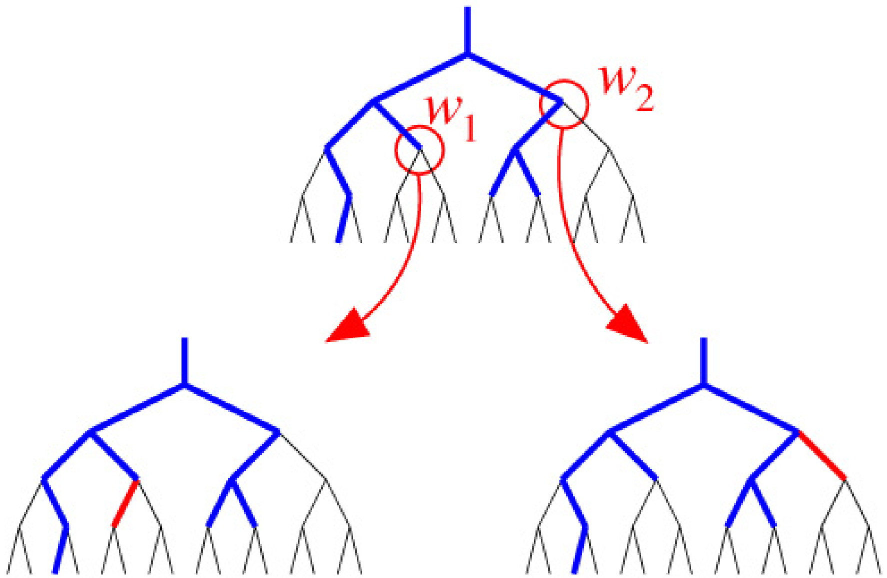
Bayesian Methods

- Walker & Mallick (1999): diffuse finite Polya Tree in a generalized linear mixed model
- Yu & Moyeed (2001): asymmetric Laplace distribution (ALD) for QR under Bayesian framework
- Hanson & Johnson (2002): mixture of Polya tree prior for median regression on survival time in AFT model
- Kottas & Krnjajic (2009): semi-parametric QR models using mixtures of DP for the error distribution
- Reich et al. (2010): an infinite mixture of two Gaussian densities for error
- Kozumi & Kobayashi (2011): developed a simple and efficient Gibbs sampling algorithm for fitting quantile regression based on a location-scale mixture representation of ALD
- Sanchez et al. (2013) proposed efficient and easy EM algorithm to obtain MLE for ALD settings from the hierarchical representation of ALD

Common Issues

- Single quantile regression each time
- Densities have their restrictive mode at the quantile of interest, which is not appropriate when extreme quantiles are being investigated
- Quantile lines monotonicity constraints
- Joint inference is poor in borrowing information through single quantile regressions
- Not coherent to pool from every individual quantile regression, because the sampling distribution of Y for τ_1 is usually different from that under quantile τ_2 since they are assuming different error distribution under two different quantile regressions (Tokdar & Kadane, 2011)

Chapter 1: Bayesian Quantile Regression Using Polya Trees Priors



Intuition

Consider heterogeneous linear regression model from He et al. (1998) :

$$y_i = \mathbf{x}_i\beta + (\mathbf{x_i}\gamma)\epsilon_i$$

The τ^{th} quantile regression parameters can be represented as

$$\beta(\tau) = \beta + F_\epsilon^{-1}(\tau)\gamma$$

- Homogeneous ($\gamma = (1, \mathbf{0})$): parallel vs Heterogeneous ($\gamma \neq (1, \mathbf{0})$): non-parallel
- Estimate $\beta, \gamma, F_\epsilon^{-1}(\tau) |\ \mathbf{Y}$, then $\beta(\tau) |\ \mathbf{Y}$
- Use mixture of Polya Tree priors to nonparametrically estimate $F_\epsilon^{-1}(\tau)$
- Closed form for predictive quantile regression parameters
- Exact inference through MCMC and fewer assumptions
- Avoid crossing quantile curves and simultaneously multiple QR in **ONE** model

Polya Tree

- Freedman (1963); Fabius (1964); Ferguson (1974), Lavine (1992); Lavine (1994)
- Advantage over Dirichlet process:
  - absolutely continuous with probability 1
  - easily tractable
  - Dirichlet process is just a special case of Polya Tree

Basic Notation

- $E = \{0, 1\}$
- E^m as the m-fold product of E
- $E^0 = \varnothing$
- $E^* = \cup_0^\infty E^m$
- Ω be a separable measurable space
- $\Pi_0 = \Omega$
- $\Pi = \{\Pi_m : m = 0, 1, \dots\}$ be a separating binary tree of partitions of Ω
- $B_\varnothing = \Omega$
- $\forall \epsilon = \epsilon_1 \cdots \epsilon_m \in E^*$, $B_{\epsilon 0}$ and $B_{\epsilon 1}$ are the two partition of B_ϵ.

Definition

Polya Tree:

A random probability measure G on $(\Omega, \mathcal{F})$ is said to have a Polya tree distribution, or a Polya tree prior with parameter $(\Pi, \mathcal{A})$, written as

$$G|\Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A}),$$

if there exists nonnegative number $\mathcal{A} = \left\{\alpha_\epsilon, \epsilon \in E^*\right\}$ and random vectors $\mathcal{Y} = \{Y_\epsilon : \epsilon \in E^*\}$ such that the following hold:

- All the random variables in $\mathcal{Y}$ are independent;
- $Y_\epsilon = (Y_{\epsilon 0}, Y_{\epsilon 1}) \sim \text{Dirichlet}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1}), \forall \epsilon \in E^*$;
- $\forall m = 1, 2, \ldots,$ and $\forall \epsilon \in E^*, G(B_{\epsilon_1, \ldots, \epsilon_m}) = \prod_{j=1}^{m} Y_{\epsilon_1 \cdots \epsilon_j}$.

Polya Tree Parameters: $\mathcal{A}$, Π

A Polya tree is centered around a pre-specified distribution G_0 (the baseline measure)

Weights $\mathcal{A}$

$\mathcal{A}$ determines how much G can deviate from G_0.

- Berger & Guglielmi (2001) considered $\alpha_{\epsilon_1,\ldots,\epsilon_m} = c\rho(m)$. In general, any $\rho(m)$ such that $\sum_{m=1}^{\infty} \rho(m)^{-1} < \infty$ guarantees G to be absolutely continuous.
- We adopt $\alpha_{\epsilon_1,\ldots,\epsilon_m} = cm^2$.

Partition parameter Π

- Canonical way of constructing a Polya Tree distribution G centering on G_0
- $B_0 = G_0^{-1}([0, 1/2]), B_1 = G_0^{-1}((1/2, 1])$
- $G(B_0) = G(B_1) = 1/2$
- $\forall \epsilon \in E^*$, choose $B_{\epsilon 0}$ and $B_{\epsilon 1}$ to satisfy $G(B_{\epsilon 0}|B_\epsilon) = G(B_{\epsilon 1}|B_\epsilon) = 1/2$

Density Function

Suppose $F = E(G), G|\Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A})$, where G_0 is the baseline measure. Then, using the canonical construction, $F = G_0$, the density function is

$$f(y) = \left[\prod_{j=1}^{m} \frac{\alpha_{\epsilon_1, \ldots, \epsilon_j}(y)}{\alpha_{\epsilon_1, \ldots, \epsilon_{j-1}, 0}(y) + \alpha_{\epsilon_1, \ldots, \epsilon_{j-1}, 1}(y)} \right] 2^m g_0(y)$$

where g_0 is the pdf of G_0.

When using the canonical construction with no data, $\alpha_{\epsilon_0} = \alpha_{\epsilon_1}$, above equation simplifies to

$$f(y) = g_0(y).$$

Conjugacy

- If $y_1, \ldots, y_n | G \sim G$
- $G|\Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A})$,
- then

$$G|y_1, \ldots, y_n, \Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A}^*)$$

where in $\mathcal{A}^*, \forall \epsilon \in E^*$,

$$\alpha_\epsilon^* = \alpha_\epsilon + n_\epsilon(y_1, \ldots, y_n),$$

where $n_\epsilon(y_1, \ldots, y_n)$ indicates the count of how many samples of $y_1, \ldots, y_n$ fall in B_ϵ.

Mixture of Polya Trees

- The behavior of a single Polya tree highly depends on how the partition is specified.
- A random probability measure G_θ is said to be a mixture of Polya tree if there exists a random variable θ with distribution h_θ, and Polya tree parameters $(\Pi^\theta, \mathcal{A}^\theta)$ such that

$$G_\theta|\theta \sim PT(\Pi^\theta, \mathcal{A}^\theta)$$

Example: Suppose $G_0 = \mathrm{N}(\mu, \sigma^2)$ is the baseline measure. For $\epsilon \in E^*, \alpha_{\epsilon_m} = cm^2$, $\theta = (\mu, \sigma, c)$ is the mixing index and the distribution on $\Theta = (\mu, \sigma, c)$ is the mixing distribution.

- With the mixture of Polya tree, the influence of the partition is lessened

Finite Polya Tree

- In practice, a finite M level Polya Tree is usually adopted to approximate the full Polya tree, in which, only up to M levels are updated.
- The rule of thumb for choosing M is to set $M = \log_2 n$, where n is the sample size Hanson & Johnson (2002)

Predictive Error Density

- Suppose G_θ is the baseline measure, $g_0(y)$ is the density function.
- Recall the posterior of $G|y_1, \ldots, y_n$ is

$$G|y_1, \ldots, y_n, \Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A}^*)$$

where in $\mathcal{A}^*, \forall \epsilon \in E^*$,

$$\alpha_\epsilon^* = \alpha_\epsilon + n_\epsilon(y_1, \ldots, y_n),$$

where $n_\epsilon(y_1, \ldots, y_n)$ indicates the count of how many samples of $y_1, \ldots, y_n$ fall in B_ϵ.

- The predictive density function of $Y|y_1, \ldots, y_n, \theta$, marginalizing out G, is

$$f_Y^\theta(y|y_1, \ldots, y_n) = \lim_{m \to \infty} \left(\prod_{j=2}^{m} \frac{cj^2 + n_{\epsilon_1 \cdots \epsilon_j(x)}(y_1, \ldots, y_n)}{2cj^2 + n_{\epsilon_1 \cdots \epsilon_{j-1}(x)}(y_1, \ldots, y_n)} \right) 2^{m-1} g_0(y),$$

- If we restrict the first level weight as $\alpha_0 = \alpha_1 = 1$, then we only need to update levels beyond the first level.

Predictive Cumulative Density Function

Based on the predictive density function of a finite Polya tree distribution, the predictive cumulative density function is

$$F_Y^{\theta,M}(y|y_1,\ldots,y_n) = \sum_{i=1}^{N-1} P_i + P_N\left(G_\theta(y)2^M - (N-1)\right),$$

where

$$P_i = \frac{1}{2}\left(\prod_{j=2}^{M} \frac{cj^2 + n_{j,\lceil i2^{j-M}\rceil}(y_1,\ldots,y_n)}{2cj^2 + n_{j-1,\lceil i2^{j-1-M}\rceil}(y_1,\ldots,y_n)}\right) \text{ and}$$
$$N = \left[2^M G_\theta(y) + 1\right],$$

in which $n_{j,\lceil i2^{j-M}\rceil}(y_1,\ldots,y_n)$ denotes the number of observations $y_1,\ldots,y_n$ in the $\lceil i2^{j-M}\rceil$ slot at level j, $\lceil\cdot\rceil$ is the ceiling function, and $[\cdot]$ is the floor function.

Predictive Error Quantiles

- The posterior predictive quantile of finite Polya tree distribution is

$$Q^{\theta,M}_{Y|y_1,\ldots,y_n}(\tau) = G_\theta^{-1}\left(\frac{\tau - \sum_{i=1}^{N} P_i + NP_N}{2^M P_N}\right),$$

  where N satisfies $\sum_{i=1}^{N-1} P_i < \tau \leq \sum_{i=1}^{N} P_i$.

- The explicit form for quantile regression coefficients becomes:

$$\beta(\tau) = \beta + \gamma G_\theta^{-1}\left(\frac{\tau - \sum_{i=1}^{N} P_i + NP_N}{2^M P_N}\right),$$

- Greatly facilitate computations

Fully Bayesian Quantile Regression with Mixture of Polya Tree Priors

The full Bayesian specification of quantile regression is given as follows,

$$\begin{aligned} y_i &= \mathbf{x_i}'\beta + (\mathbf{x_i}'\gamma)\epsilon_i, i = 1, \ldots, n \\ \epsilon_i | G_\theta &\sim G_\theta \\ G_\theta | \Pi^\theta, \mathcal{A}^\theta &\sim PT(\Pi^\theta, \mathcal{A}^\theta) \\ \theta = (\sigma, c) &\sim \pi_\theta(\theta) \\ \beta &\sim \pi_\beta(\beta) \\ \gamma &\sim \pi_\gamma(\gamma). \end{aligned}$$

In order to not confound the location parameter, ϵ_i or G is set to have median 0 by fixing $\alpha_0 = \alpha_1 = 1$. For the similar reason, the first component of γ is fixed at 1.

Posterior Distribution of β, γ, σ, c

$$L(\beta, \gamma, \sigma, c | \mathbf{Y}) \propto p(\mathbf{Y} | \beta, \gamma, \sigma, c)\pi_\beta(\beta)\pi_\gamma(\gamma)\pi_\sigma(\sigma)\pi_c(c)$$

Priors

σ, c using diffuse gamma prior:

$$\pi(\sigma) \sim \Gamma(1/2, 1/2),$$
$$\pi(c) \sim \Gamma(1/2, 1/2).$$

β, γ using Spike and Slab Priors

- Shrink toward zero
- Do variable selection on both quantile regression parameters and heterogeneity parameters
- Improve efficiency
- Use continuous spike and slab priors on each component of (β, γ) (George & McCulloch, 1993)

$$\pi_\beta(\beta_j) = \delta_{\beta_j}\phi(\beta_j; 0, s_j^2\sigma_{\beta_j}^2) + (1-\delta_{\beta_j})\phi(\beta_j; \beta_j^p, \sigma_{\beta_j}^2),$$
$$\delta_{\beta_j} \sim \text{Bernoulli}(\pi_{\beta_j}),$$

Computation Details

- R package *bqrpt*
- Posterior samples of $(\beta, \gamma, \sigma, c \mid \mathbf{Y})$
- Thinning
- Adaptive Metropolis-Hasting algorithm
  - For good MCMC mixing performance, we adjust the acceptance rate of the adaptive Metropolis-Hasting algorithm to around 0.2 for sampling
  - Tuning parameters are increased(decreased) by multiplying(dividing) $\delta(l) = \exp(\min(0.01, l^{-1/2}))$ when current acceptance proportion is larger(smaller) than target optimal acceptance rate for every 100 iterations during burn-in period, where l is the number of current batches of 100 iterations

Simulation

- *RQ*: rq function in (Koenker, 2012) (frequentist quantile regression method)
- *FBQR*: flexible Bayesian quantile regression (Reich et al. 2010)
- *PT*: Polya trees with normal diffuse priors
- *PTSS*: Polya trees with spike and slab priors
- Models:

$$y_i = 1 + x_i + (1 + \alpha x_i)\epsilon_i$$

  where M1: $\epsilon_i \sim N(0,1)$, M2: t_3, M3: $0.5N(-2,1) + 0.5N(2,1)$, M4: $0.8N(0,1) + 0.2N(3,3)$
- Compare for both homogeneous ($\alpha = 0$) (M1-M4) and heterogeneous ($\alpha = 0.2$) models (M1H-M4H)
- $n = 200$
- 100 data sets
- $x_i \sim \text{Uniform}(0,4)$
- [M5:] $y_i | R_i = 1 \sim 2 + x_{i1} + \epsilon_{1i}, y_i | R_i = 0 \sim -2 - x_{i1} + \epsilon_{1i}$, $\epsilon_{1i} \sim N(0,1)$

Evaluation Methods

- **MSE**

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_j(\tau) - \beta_j(\tau))^2,$$

  - N is the number of simulations
  - $\beta_j(\tau)$ is the j^{th} component of the true quantile regression parameters
  - $\hat{\beta}_j(\tau)$ is the j^{th} component of estimated quantile regression parameters
  - We use the posterior mean as estimated parameters.

- **Monte Carlo standard errors (MCSE)** are used to evaluate the *significance* of the differences between methods,

$$\text{MCSE} = \hat{\text{sd}}(\text{Bias}^2)/\sqrt{N},$$

  - $\hat{\text{sd}}$ is the sample standard deviation
  - $\text{Bias} = \hat{\beta}_j(\tau) - \beta_j(\tau)$.

Simulation Summary

- M1 and M1H: PT and PTSS better
- M2-M4, M2H-M4H, error away from Polya tree baseline measure, FBQR dominates (simulated models coincide with the models in the FBQR approach)
- PT and PTSS are also competitive (M3 and M3H with $\tau = 50\%$ and M4 with $\tau = 90\%$)
- M5: heterogeneity from the mixture of distributions. The deficit in 90% quantile is offset by much smaller bias in 50% quantile regression.
- RQ performs poorly
- PT is not impacted by lack of unimodality and heterogeneity and provides more information for the relationship between responses and covariates.
- Less information is available for our approach to detect the shape at a particular extreme percentile of the distribution since there are few observations at extreme quantiles.
- PT and PTSS can fit simultaneously multiple QR and provide coherent information about the error distribution.
- No crossing QR curves
- Expect to see advantages when dimension of responses is two or more.

Summary

- Bayesian approach for simultaneous linear quantile regression by introducing mixture of Polya tree priors and estimating heterogeneity parameters.
- Marginalizing the predictive density function of the Polya tree distribution, quantiles of interest are obtained in closed form by inverting the predictive cumulative distribution.
- Exact posterior inference can be made via MCMC.
- Quantile lines cannot cross since quantiles are estimated through density estimation.
- The simulations show advantages of our method in some cases especially when the error is multimodal and highly skewed.

Future Work

- Further research includes quantile regression for correlated data by modelling error as a mixture of multivariate Polya tree distribution
- Our approach allows for quantile regression with missing data under ignorability by adding a data augmentation step.
- It might be possible to use a slightly more complex baseline distribution in Polya tree adaptively to improve the estimation.

Chapter 2: Quantile Regression in the Presence of Monotone Missingness with Sensitivity Analysis

- Wei et al. (2012) proposed a multiple imputation method for quantile regression model when there are some covariates missing at random (MAR).
- Bottai & Zhen (2013) illustrated an imputation method using estimated conditional quantiles of missing outcomes given observed data.
- Yuan & Yin (2010) introduced a fully parametric Bayesian quantile regression approach for longitudinal data with non-ignorable missing data.
- When there are many possible dropout time, Roy (2003) proposed to group them by latent classes.
- Roy & Daniels (2008) extended Roy (2003) to generalized linear models and proposed a pattern mixture model for data with non-ignorable dropout, borrowing ideas from Heagerty (1999).

Missing Data Mechanism

- Missing data mechanism:

$$p(\mathbf{r}|\mathbf{y}, \mathbf{x}, \phi(\omega))$$

- Missing Complete At Random (MCAR)

$$p(\mathbf{r}|y_{obs}, y_{mis}, \mathbf{x}, \phi) = p(\mathbf{r}|\mathbf{x}, \phi).$$

- Missing At Random (MAR)

$$p(\mathbf{r}|y_{obs}, y_{mis}, \mathbf{x}, \phi) = p(\mathbf{r}|y_{obs}, \mathbf{x}, \phi).$$

- Missing Not At Random (MNAR), for $y_{mis} \neq y_{mis}{}'$,

$$p(\mathbf{r}|y_{obs}, y_{mis}, \mathbf{x}, \phi) \neq p(\mathbf{r}|y_{obs}, y_{mis}{}', \mathbf{x}, \phi).$$

Notation

- Under monotone dropout, WOLOG, denote $S_i \in \{1, 2, \ldots, J\}$ to be the number of observed $Y_{ij}'s$ for subject i,
- $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})^T$ to be the full data response vector for subject i,
- J is the maximum follow up time.
- We assume Y_{i1} is always observed.
- We are interested in the τ-th marginal quantile regression coefficients $\gamma_j = (\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jp})^T$,

$$Pr(Y_{ij} \leq \mathbf{x}_i^T \gamma_j) = \tau, \text{ for } j = 1, \ldots, J,$$

where $\mathbf{x}_i$ is a $p \times 1$ vector of covariates for subject i.

- Let

$$p_k(Y) = p(Y|S = k), \quad p_{\geq k}(Y) = p(Y|S \geq k)$$

be the densities of response $\mathbf{Y}$ given follow-up time $S = k$ and $S \geq k$. And Pr_k be the corresponding probability given $S = k$.

Pattern Mixture Model

- Mixture models factor the joint distribution of response and missingness as

$$p(\mathbf{y}, \mathbf{S}|\mathbf{x}, \omega) = p(\mathbf{y}|\mathbf{S}, \mathbf{x}, \omega)p(\mathbf{S}|\mathbf{x}, \omega).$$

- The full-data response distribution is given by

$$p(\mathbf{y}|\mathbf{x}, \omega) = \sum_{S \in \mathcal{S}} p(\mathbf{y}|\mathbf{S}, \mathbf{x}, \theta)p(\mathbf{S}|\mathbf{x}, \phi),$$

  where $\mathcal{S}$ is the sample space for dropout time S and the parameter vector ω is partitioned as (θ, ϕ).

- Furthermore, the conditional distribution of response within patterns can be decomposed as

$$P(Y_{obs}, Y_{mis}|\mathbf{S}, \theta) = P(Y_{mis}|Y_{obs}, \mathbf{S}, \theta_E)Pr(Y_{obs}|\mathbf{S}, \theta_{y,O}),$$

- θ_E: extrapolation distribution

- $\theta_{y,O}$: distribution of observed responses

Model Settings

- Multivariate normal distributions within pattern
- The marginal quantile regression models as:

$$Pr(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j) = \tau,$$

where γ_j is the τ^{th} quantile regression coefficients of interest for component j.

$$
\begin{aligned}
&p_k(y_{i1}) = N(\Delta_{i1} + \mathbf{x}_{i1}^T \beta_1^{(k)}, \sigma_1^{(k)}), k = 1, \ldots, J, \\
&p_k(y_{ij}|\mathbf{y}_{ij^-}) = \begin{cases} \mathrm{N}\left(\Delta_{ij} + \mathbf{x}_{ij}^T \mathbf{h}_j^{(k)} + \mathbf{y}_{ij^-}^T \beta_{y,j-1}^{(k)}, \sigma_j^{(k)}\right), & k < j; \\ \mathrm{N}\left(\Delta_{ij} + \mathbf{y}_{ij^-}^T \beta_{y,j-1}^{(\geq j)}, \sigma_j^{(\geq j)}\right), & k \geq j; \end{cases}, \text{ for } 2 \leq j \leq J, \\
&S_{ij} = k | \ \mathbf{x}_{ij} \sim \text{Multinomial}(1, \phi),
\end{aligned}
$$

Δ

Δ_{ij} are functions of $\tau, \mathbf{x}_{ij}, \beta, \mathbf{h}, \sigma, \gamma_j, \phi$ and are determined by the marginal quantile regressions,

$$\tau = Pr(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j) = \sum_{k=1}^{J} \phi_k Pr_k(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j) \text{ for } j = 1,$$

and

$$\begin{aligned}
\tau &= Pr(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j) = \sum_{k=1}^{J} \phi_k Pr_k(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j) \\
&= \sum_{k=1}^{J} \phi_k \int \cdots \int Pr_k(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j | \mathbf{y}_{ij^-}) p_k(y_{i(j-1)} | \mathbf{y}_{i(j-1)^-}) \\
&\quad \cdots p_k(y_{i2} | y_{i1}) p_k(y_{i1}) dy_{i(j-1)} \cdots dy_{i1}. \text{ for } j = 2, \ldots, J.
\end{aligned}$$

Intuition

- Embed the marginal quantile regressions directly in the model through constraints in the likelihood of pattern mixture models
- The mixture model allows the marginal quantile regression coefficients to differ by quantiles. Otherwise, the quantile lines would be parallel to each other.
- The mixture model also allows sensitivity analysis.
- For identifiability of the observed data distribution, we apply the following constraints,

$$\sum_{k=1}^{J} \beta_{l1}^{(k)} = 0, l = 1, \ldots, p,$$

Missing Data Mechanism and Sensitivity Analysis

- Mixture models are not identified due to insufficient information provided by observed data.
- Specific forms of missingness are needed to induce constraints to identify the distributions for incomplete patterns, in particular, the extrapolation distribution
- In mixture models , MAR holds (Molenberghs et al. 1998; Wang & Daniels, 2011) if and only if, for each $j \geq 2$ and $k < j$:

$$p_k(y_j | y_1, \ldots, y_{j-1}) = p_{\geq j}(y_j | y_1, \ldots, y_{j-1}).$$

- When $2 \leq j \leq J$ and $k < j$, Y_j is not observed, thus $\mathbf{h}_j^{(k)}$ and $\sigma_j^{(k)}$, $\beta_{y,j-1}^{(k)} = (\beta_{y_1,j}^{(k)}, \ldots, \beta_{y_{j-1},j-1}^{(k)})^T$ can not be identified from the observed data.

Sensitivity Analysis

$$\log \sigma_j^{(k)} = \log \sigma_j^{(\geq j)} + \delta_j^{(k)},$$
$$\beta_{y,j-1}^{(k)} = \beta_{y,j-1}^{(\geq j)} + \eta_{j-1}^{(k)},$$

where $\eta_{j-1}^{(k)} = \left(\eta_{y_1,j-1}^{(k)}, \ldots, \eta_{y_{j-1},j-1}^{(k)}\right)$ for $k < j$. Then $\xi_s = (\mathbf{h}_j^{(k)}, \eta_{j-1}^{(k)}, \delta_j^{(k)})$ is a set of sensitivity parameters (Daniels & Hogan, 2008), where $k < j, 2 \leq j \leq J$.

- $\xi_s = \xi_{s0} = \mathbf{0}$, MAR holds.
- ξ_s is fixed at $\xi_s \neq \xi_{s0}$, MNAR.
- We can vary ξ_s around $\mathbf{0}$ to examine the impact of different MNAR mechanisms.

- For Bayesian, put priors on (ξ_s, ξ_m):

$$p(\xi_s, \xi_m) = p(\xi_s)p(\xi_m),$$

where $\xi_m = \left(\gamma_j, \beta_{y,j-1}^{(\geq j)}, \alpha_j^{(\geq j)}, \phi\right)$

- Sensitivity analysis can be done by putting point mass priors on ξ_s
- MAR with no uncertainty: $p(\xi_s = \mathbf{0}) \equiv 1$.
- MAR with uncertainty: $\mathrm{E}(\xi_s) = \xi_{s0} = \mathbf{0}$ with $\mathrm{Var}(\xi_s) \neq \mathbf{0}$.
- MNAR with no uncertainty, $\mathrm{E}(\xi_s) = \delta_\xi$, where $\delta_\xi \neq \mathbf{0}$ and $\mathrm{Var}(\xi_s) = \mathbf{0}$.
- MNAR with uncertainty, $\mathrm{E}(\xi_s) = \delta_\xi$, where $\delta_\xi \neq \mathbf{0}$ and $\mathrm{Var}(\xi_s) \neq \mathbf{0}$.

Calculation of Δ_{ij} ($j = 1$)

Δ_{ij} depends on subject-specific covariates $\mathbf{x}_{ij}$, thus Δ_{ij} needs to be calculated for each subject. We now illustrate how to calculate Δ_{ij} given all the other parameters $\xi = (\xi_m, \xi_s)$.

Δ_{i1} : Expand equation :

$$\tau = \sum_{k=1}^{J} \phi_k \Phi\left(\frac{\mathbf{x}_{i1}^T \gamma_1 - \Delta_{i1} - \mathbf{x}_{i1}^T \beta_1^{(k)}}{\sigma_1^{(k)}} \right),$$

where Φ is the standard normal CDF. Because the above equation is continuous and monotone in Δ_{i1}, it can be solved by a standard numerical root-finding method (e.g. bisection method) with minimal difficulty.

Calculation of $\Delta_{ij}, 2 \leq j \leq J$

Lemma:

$$\int \Phi\left(\frac{x-b}{a}\right) d\Phi(x;\mu,\sigma) = \begin{cases} 1 - \Phi\left(\dfrac{b-\mu}{\sigma} \Big/ \sqrt{\dfrac{a^2}{\sigma^2}+1}\right) & a > 0, \\ \Phi\left(\dfrac{b-\mu}{\sigma} \Big/ \sqrt{\dfrac{a^2}{\sigma^2}+1}\right) & a < 0, \end{cases}$$

Recursively for the first multiple integral, apply lemma once to obtain:

$$\begin{aligned} Pr_1(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j) &= \int \cdots \int Pr_1(Y_{ij} \leq \mathbf{x}_{ij}^T \gamma_j | \mathbf{x}_{ij}, \mathbf{Y}_{ij^-}) \\ &\quad dF_1(Y_{i(j-1)} | \mathbf{x}_{ij}, \mathbf{Y}_{i(j-1)^-}) \cdots dF_1(Y_{i1} | \mathbf{x}_{ij}), \\ &= \int \cdots \int \Phi\left(\frac{Y_{i(j-2)} - b^*}{a^*}\right) dF_1(Y_{i(j-2)} | \mathbf{x}_{ij}, \mathbf{Y}_{i(j-2)^-}) \cdots dF_1(Y_{i1} | \mathbf{x}_{ij}). \end{aligned}$$

Then, by recursively applying lemma $(j-1)$ times, each multiple integral in equation can be simplified to single normal CDF.

MLE

The observed data likelihood for an individual i with follow-up time $S_i = k$ is

$$\begin{aligned} L_i(\xi|\mathbf{y}_i, S_i = k) &= \phi_k p_k(y_k|y_1, \dots, y_{k-1}) p_k(y_{k-1}|y_1, \dots, y_{k-2}) \cdots p_k(y_1), \\ &= \phi_k p_{\geq k}(y_k|y_1, \dots, y_{k-1}) p_{\geq k-1}(y_{k-1}|y_1, \dots, y_{k-2}) \cdots p_k(y_1), \end{aligned}$$

- Use the bootstrap to construct confidence interval and make inferences.

Goodness of Fit Check

- Check QQ plots of fitted residuals

$$\hat{\epsilon}_{ij} = \begin{cases} (y_{ij} - \hat{\Delta}_{ij} - \mathbf{x}_{\mathbf{ij}}^{\mathbf{T}} \hat{\beta}_{\mathbf{1}}^{(\mathbf{k})})/\hat{\sigma}_1^{(k)}, & j = 1 \\ (y_{ij} - \hat{\Delta}_{ij} - \mathbf{y}_{\mathbf{ij}^-}^{\mathbf{T}} \hat{\beta}_{\mathbf{y,j-1}}^{(\geq \mathbf{j})})/\hat{\sigma}_j^{(\geq j)}, & j > 1 \end{cases}.$$

Curse of Dimension

Each pattern $S = k$ has its own set of SP $\xi_s^{(k)}$. However, to keep the number of SP at a manageable level, we assume ξ_s does not depend on pattern.

Real Data Analysis: Tours

- Weights were recorded at baseline (Y_0), 6 months (Y_1) and 18 months (Y_2).
- We are interested in how the distributions of weights at six months and eighteen months change with covariates.
- The regressors of interest include **AGE**, **RACE** (black and white) and **weight at baseline** (Y_0).
- Weights at the six months (Y_1) were always observed and 13 out of 224 observations (6%) were missing at 18 months (Y_2).
- The **AGE** covariate was scaled to 0 to 5 with every increment representing 5 years.
- We fitted regression models for bivariate responses $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ for quantiles (10%, 30%, 50%, 70%, 90%).
- We ran 1000 bootstrap samples to obtain 95% confidence intervals.

Results

- For weights of participants at six months, weights of whites are generally 4kg lower than those of blacks for all quantiles, and the coefficients of race are negative and significant.
- Weights of participants are not affected by age since the coefficients are not significant. Differences in quantiles are reflected by the intercept.
- Coefficients of baseline weight show a strong relationship with weights after 6 months.
- For weights at 18 months after baseline, we have similar results.
- Weights at 18 months still have a strong relationship with baseline weights.
- However, whites do not weigh significantly less than blacks at 18 months unlike at 6 months.

Sensitivity Analysis

We also did a sensitivity analysis based on an assumption of MNAR.

- Based on previous studies of pattern of weight regain after lifestyle treatment (Wadden et al. 2001; Perri et al. 2008), we assume that

$$E(Y_2 - Y_1 | R = 0) = 3.6\text{kg},$$

  which corresponds to 0.3kg regain per month after finishing the initial 6-month program.

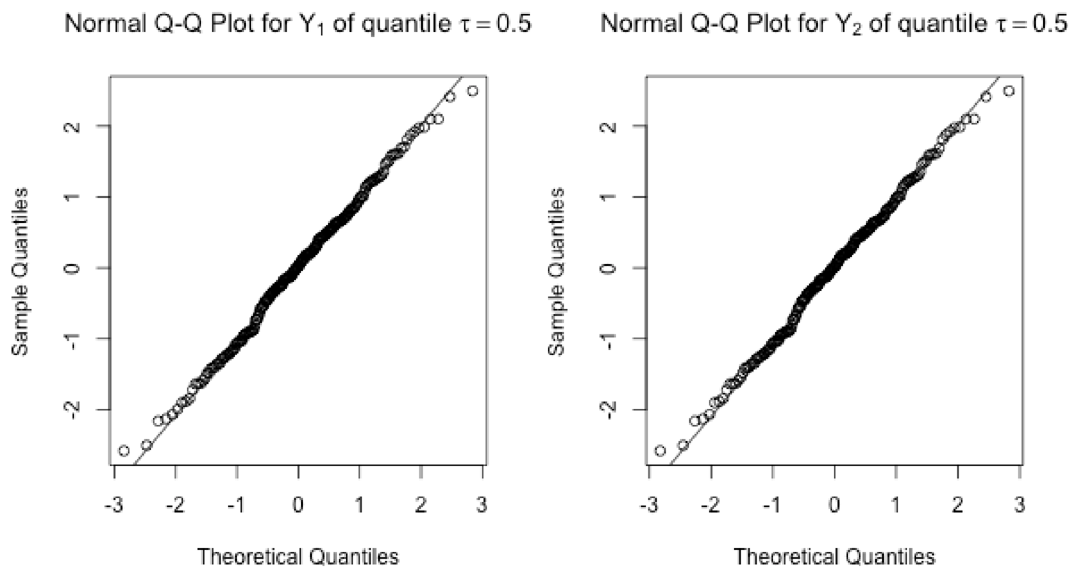
- We incorporate the sensitivity parameters in the distribution of $Y_2 | Y_1, R = 0$ via the following restriction:

$$\Delta_{i2} + \mathbf{x}_{i2}^T \mathbf{h}_2^{(1)} + E(y_{i1} | R = 0)(\beta_{y,1}^{(1)} + \eta_1^{(1)} - 1) = 3.6\text{kg}.$$

Results

- There are not large differences for estimates for Y_2 under MNAR vs MAR.
- This is partly due to the low proportion of missing data in this study.

Goodness of Fit Check



- We also checked the goodness of fit via QQ plots on the fitted residuals for each quantile regression fit.
- The QQ plots showed minimal evidence against the assumption that the residuals were normally distributed; thus we were confident with the conclusion of our quantile regression models.

Summary

- Developed a marginal quantile regression model for data with monotone missingness.
- Used a pattern mixture model to jointly model the full data response and missingness.
- Estimate marginal quantile regression coefficients instead of conditional on random effects
- Allows non-parallel quantile lines over different quantiles via the mixture distribution
- Allows for sensitivity analysis which is essential for the analysis of missing data (NAS 2010).
- Allows the missingness to be non-ignorable.
- Recursive integration simplifies computation and can be implemented in high dimensions.

Future Work

- Sequential multivariate normal distribution for each component in the PMM might be too restrictive
- Simulation results showed that the mis-specification of the error term did have an impact on the extreme quantile regression inferences.
- Working on replacing it with a non-parametric model, for example, a Dirichlet process mixture of normals.