# Quantile Function on Scalar Regression Analysis for Distributional Data

Hojin Yang*, Veera Baladandayuthapani
Arvind Rao, and Jeffrey S. Morris

Department of Statistics
Pusan National University

May 28, 2021

# Outline

# Glioblastoma Multiforme (GBM)

- Most common and aggressive form of brain cancer
- No current prevention approaches, and poor outcomes
  - Median survival 12mo, 3-5% 5yr survival
- Exhibits heterogeneous physiological and morphological features as it proliferates
- Investigating these heterogeneities and relating them to clinical/genetic outcomes can lead to the development of personalized treatment strategies.

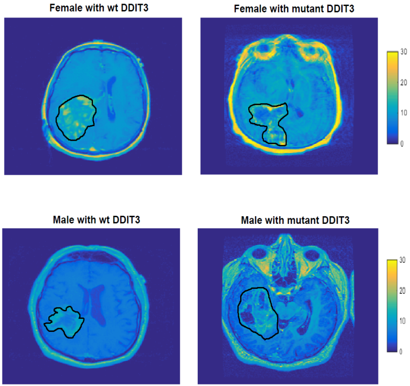
# Glioblastoma Multiforme (GBM)

- Most common and aggressive form of brain cancer
- No current prevention approaches, and poor outcomes
  - Median survival 12mo, 3-5% 5yr survival
- Exhibits heterogeneous physiological and morphological features as it proliferates
- Investigating these heterogeneities and relating them to clinical/genetic outcomes can lead to the development of personalized treatment strategies.

**Our Goal:**

Assess how variability in tumor image intensities is associated with demographic, clinical, and genetic factors

# Glioblastoma Images




- Presurgical T1-weighted post-contrast MRI images from GBM patients
- **Radiomics**: compute features summarizing tumor image characteristics and relate to clinical outcomes.
- *Histogram features*: Summaries computed from pixel intensity distributions (e.g. mean, variance, skewness, Q05, Q95)

# Modeling Distributions

The typical approach is to extract pre-chosen feature and fit separate regression analyses to each selected feature, which has some major drawbacks:

- Multiple testing problems
- May miss distributional differences not contained in pre-chosen summaries.

# Modeling Distributions

The typical approach is to extract pre-chosen feature and fit separate regression analyses to each selected feature, which has some major drawbacks:

- Multiple testing problems
- May miss distributional differences not contained in pre-chosen summaries.

**Alternative Approach**

Instead of just modeling the extracted summaries, model the entire distribution of pixel intensities (as functional data).

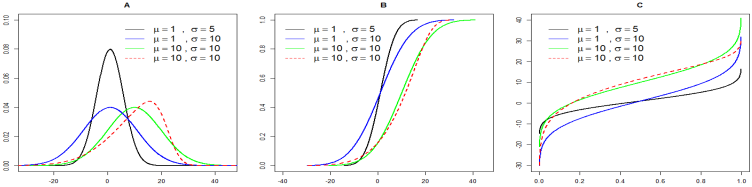
# Distributional Data

We consider regression problem for $i = 1, \ldots, n$ subjects.

- Random copies $(X_1, F_1), \ldots, (X_n, F_n)$ of $(X, F)$
  - Predictor: $X_i = (x_{i1}, \ldots, x_{iA})$. i.e. $X_i \in \mathbb{R}^A$
  - Outcome: $F_i(y)$ for $y \in \mathbb{R}$
- A challenge is that $F_i(y)$ is not actually observed.
- Observed data: $(X_1, Y_{11}, \ldots, Y_{1m_1}), \ldots, (X_n, Y_{n1}, \ldots, Y_{nm_n})$
- $Y_{i1}, \ldots, Y_{im_i}$ are samples from $F_i$.

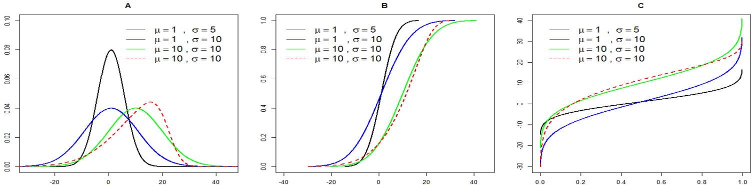
# Modeling Distributions

- Several choices to represent pixel intensity distributions: density, cumulative distribution, or quantile functions.



# Modeling Distributions

- Several choices to represent pixel intensity distributions: density, cumulative distribution, or quantile functions.




- We choose to use the quantile function.
  The quantile function of $Y$ on $p \in [0, 1]$, is defined as

## Definition of the quantile function

$$Q_Y(p) = F_Y^{-1}(p) = \inf\left(y : F_Y(y) \geq p\right),$$

where $p = F_Y(y)$ is the proportion less than or equal to $y$.

# Properties of Quantile Functions

Quantile functions have properties that make them useful here:

- Defined on a fixed domain, $p \in \mathcal{P} \subset [0, 1]$

# Properties of Quantile Functions

Quantile functions have properties that make them useful here:

- Defined on a fixed domain, $p \in \mathcal{P} \subset [0, 1]$
- Straightforward to compute empirical estimates without choice of smoothing parameters

**eDF**

Let $Y_{(1)} \leq \cdots \leq Y_{(m)}$ be order statistics from a sample of size $m$. For $p \in [1/(m+1), m/(m+1)]$, the eQF is given by

$$\widehat{Q}_Y(p) = (1-w)Y_{([(m+1)p])} + wY_{([(m+1)p]+1)},$$

where $w$ is a weight such that $(m+1)p = [(m+1)p] + w$.

# Properties of Quantile Functions

Quantile functions have properties that make them useful here:

- Defined on a fixed domain, $p \in \mathcal{P} \subset [0, 1]$
- Straightforward to compute empirical estimates without choice of smoothing parameters
- Straightforward formulas to calculate distributional moments

## Distributional Moments

$$\begin{aligned} \mu_Y &= \mathsf{E}(Y) = \int_0^1 Q_Y(p)dp \\ \sigma_Y^2 &= \mathsf{Var}(Y) = \int_0^1 \left(Q_Y(p) - \mu_Y\right)^2 dp \\ \xi_Y &= \mathsf{Skew}(Y) = \int_0^1 \left(Q_Y(p) - \mu_Y\right)^3 / \sigma_Y^3 dp \end{aligned}$$

# Quantile functional regression

**Approach:** Regress eQF as functional response on covariates.

# Quantile functional regression

**Approach:** Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $\widehat{Q}_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.

# Quantile functional regression

**Approach:** Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $\widehat{Q}_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.
2. Regress $\widehat{Q}_i(p)$ on covariates $x_{ia}, a = 1, \ldots, A$, each with regression coefficients $\beta_a(p)$ defined on $p \in \mathcal{P} \subset [0, 1]$.

## Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

# Quantile functional regression

**Approach:** Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $\widehat{Q}_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.
2. Regress $\widehat{Q}_i(p)$ on covariates $x_{ia}, a = 1, \ldots, A$, each with regression coefficients $\beta_a(p)$ defined on $p \in \mathcal{P} \subset [0, 1]$.

**Quantile Functional Regression Model**

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

3. Test for significantly associated covariates: $H_0 : \beta_a(p) \equiv 0$.

# Quantile functional regression

**Approach:** Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $\widehat{Q}_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.
2. Regress $\widehat{Q}_i(p)$ on covariates $x_{ia}, a = 1, \ldots, A$, each with regression coefficients $\beta_a(p)$ defined on $p \in \mathcal{P} \subset [0, 1]$.

**Quantile Functional Regression Model**

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

3. Test for significantly associated covariates: $H_0 : \beta_a(p) \equiv 0$.
4. Characterize the significant distributional differences e.g. range of $p$, mean, variance, skewness

# Quantile Functional Regression

## Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

**Naive approach:** compute independent regressions for each $p$

- fail to borrow strength over $p \rightarrow$ wiggly, inefficient $\widehat{\beta}_a(p)$.
- ignore correlation over $p$ in $E_i(p) \rightarrow$ loss of inferential power.

# Quantile Functional Regression

## Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

**Naive approach:** compute independent regressions for each $p$

- fail to borrow strength over $p \rightarrow$ wiggly, inefficient $\widehat{\beta}_a(p)$.
- ignore correlation over $p$ in $E_i(p) \rightarrow$ loss of inferential power.

**Functional regression approach:** Use *basis function* representations to account for correlation (Morris 2015)

- $\beta_a(p)$ regularized via L1/L2 penalization of basis coefficients.
- Basis functions induce correlation across $p$ in Cov$\{E_i(p)\}$.
- Common bases: splines, PC, Fourier bases, wavelets

# Beta Cumulative Distribution Functions

We consider **basis functions** for the quantile function, $Q(p)$.

$$\tilde{Q}_K(p) = \sum_{k=1}^{K} \psi_k(p) q_k^*$$

- Define $\psi_k(p) = \int_0^p \frac{\Gamma(K+2)}{\Gamma(k+1)\Gamma(K-k+1)} z^k (1-z)^{K-k} dz$
  - i.e., $\psi_k(p) = P(Z \leq p)$, where $Z \sim$ beta$(k+1, K-k+1)$

# Basis Transform Modeling Approach

## Data Space Model

$$Q_i(p) = X_i^T B(p) + E_i(p),$$

where $B(p) = (\beta_1(p), \ldots, \beta_A(p))^T$ and $E_i(p)$ is a noise process.

1. Compute basis coefficients

## Computing Coefficients

Let $\widehat{\boldsymbol{Q}}_{\boldsymbol{i}} = [\widehat{Q}_i(p_1), \ldots, \widehat{Q}_i(p_{m_i})]$ with $p_j = j/(m_i + 1)$
Let $\boldsymbol{\Psi_i}$ be $K \times m_i$ matrix with elements $\boldsymbol{\psi_i}(k, j) = \psi_k(p_j)$
Basis coefficients: $\widehat{\boldsymbol{q}}_{\boldsymbol{i}}^{*} = \widehat{\boldsymbol{Q}}_{\boldsymbol{i}}\boldsymbol{\Psi}_{\boldsymbol{i}}^{*}$ where $\boldsymbol{\Psi}_{\boldsymbol{i}}^{*} = \boldsymbol{\Psi}_{\boldsymbol{i}}^{\boldsymbol{T}}(\boldsymbol{\Psi}_{\boldsymbol{i}}\boldsymbol{\Psi}_{\boldsymbol{i}}^{\boldsymbol{T}})^{-1}$.

# Basis Transform Modeling Approach

## Data Space Model

$$Q_i(p) = X_i^T B(p) + E_i(p),$$

where $B(p) = (\beta_1(p), \ldots, \beta_A(p))^T$ and $E_i(p)$ is a noise process.

1. Compute basis coefficients
2. Fit projected space model

## Projected Space Model

$$\widehat{\boldsymbol{q}}_{\boldsymbol{i}}^{*} = X_i^T \boldsymbol{B}^* + \boldsymbol{E}_{\boldsymbol{i}}^{*}$$

where $\widehat{\boldsymbol{q}}_{\boldsymbol{i}}^{*} = (\hat{q}_{i1}^*, \ldots, \hat{q}_{iK}^*)^T$, $\widehat{Q}_i(p) = \sum_{k=1}^K \hat{q}_{ik}^* \psi_k(p)$, $\beta_a(p) = \sum_{k=1}^K B_{ak}^* \psi_k(p)$, $E_i(p) = \sum_{k=1}^K E_{ik}^* \psi_k(p)$, and $\boldsymbol{E}_{\boldsymbol{i}}^{*} \sim \text{MVN}(0, \Sigma^*)$ where $\Sigma^*$ is $K \times K$ covariance matrix.

# Basis Transform Modeling Approach

## Data Space Model

$$Q_i(p) = X_i^T B(p) + E_i(p),$$

where $B(p) = (\beta_1(p), \ldots, \beta_A(p))^T$ and $E_i(p)$ is a noise process.

1. Compute basis coefficients
2. Fit projected space model
3. Transform results back to data space for inference

## Transform Results to Data Space

$\beta_a(p) = \sum_{k=1}^{K} B^*_{ak}\psi_k(p)$, and then perform desired inference.

# Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.

# Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on $B^*_{ak}$ to regularize $\beta_a(p)$. (spike Gaussian-slab)

**Sparsity prior on $B^*_{ak}$**

$$B^*_{ak} \sim \gamma_{ak} N(0, \tau^2_{ak}) + (1-\gamma_{ak}) I_0$$
$$\gamma_{ak} \sim \text{Bernoulli}(\pi_{ak}),$$

# Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on $B^*_{ak}$ to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.

## Vague proper prior

$$\sigma_k^2 \sim \text{inverse-gamma}(\nu_0/2, \nu_0/2).$$

# Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on $B^*_{ak}$ to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
- We fit the project space model by using Markov chain Monte Carlo (MCMC).

# Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on $B^*_{ak}$ to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
- We fit the project space model by using Markov chain Monte Carlo (MCMC).
  - Complete conditional for $B^*_{ak}$ is mixture of $I_0$ and Gaussian.

## Posterior Sampling

$\boldsymbol{B^*_{ak} \sim \alpha_{ak} N(\mu_{ak}, v_{ak}) + (1-\alpha_{ak}) I_0}$

where $\mu_{ak} = \widehat{B}^*_{ak}(1 + S_{ak}/\tau_{ak})^{-1}$, $S_{ak} = (\sum_{i=1}^{n} x_{ia}/\sigma^2_k)^{-1}$, $v_{ak} = S_{ak}(1 + S_{ak}/\tau_{ak})^{-1}$, and $\alpha_{ak} = \mathrm{P}(\gamma_{ak} = 1 | Q^*_{.k}, B^*_{ak}, \sigma^2_k)$

# Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on $B^*_{ak}$ to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
- We fit the project space model by using Markov chain Monte Carlo (MCMC).
  - Complete conditional for $B^*_{ak}$ is mixture of $I_0$ and Gaussian.
  - Covariance parameters have conjugate complete conditionals.

## Posterior Sampling

$$\sigma_k^2 \sim \text{Inverse Gamma}\{(\nu_0 + n)/2, (\nu_0 + \|\widehat{q}_{.k} - \boldsymbol{X} B^*_{.k}\|^2)/2\}$$

# Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on $B^*_{ak}$ to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
- We fit the project space model by using Markov chain Monte Carlo (MCMC).
  - Complete conditional for $B^*_{ak}$ is mixture of $I_0$ and Gaussian.
  - Covariance parameters have conjugate complete conditionals.
- Posterior samples transformed back to original data space to get posterior samples of $\beta_a(p)$ on desired grid of $p$.

# Simulation

Figure: Four population groups in the simulation.

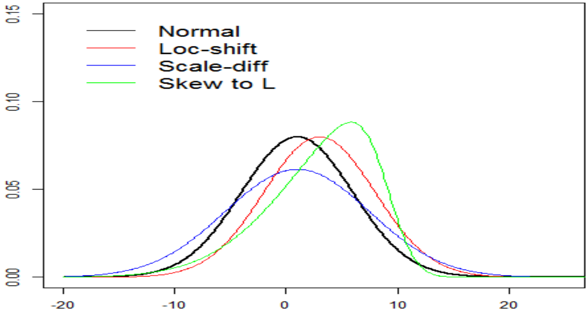
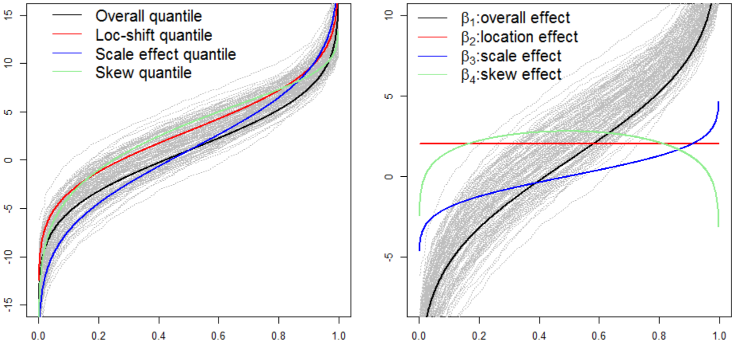


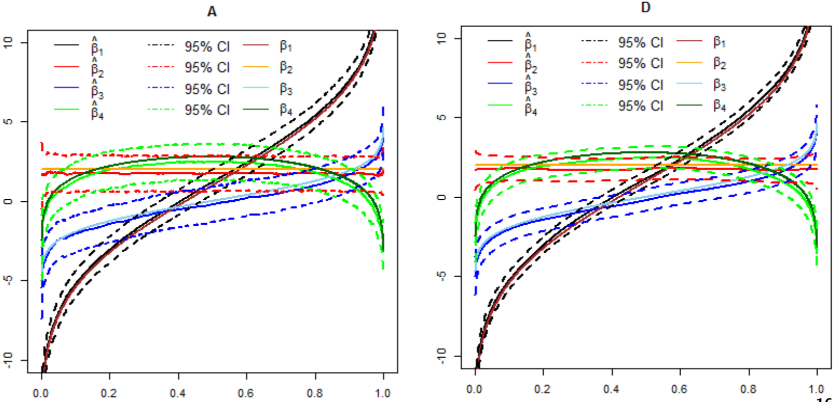
Figure: **Simulated Data**. $\beta_a(p)$ are location, scale, and skewness shifts.




- $Q_{ij}(p) = Q_{0j}(p) + \epsilon_{ij}(p)$
- $Y_{ij1} = Q_{ij}(u_1), \ldots, Y_{ijm_{ij}} = Q_{ij}(u_{m_{ij}})$, where $u_l \sim U(0,1)$, $m_{ij} = 1024$, $X_{ij} = (1, e_j)$, and $e_j$ is standard basis in $\mathbb{R}^3$.

# Simulation Results

Figure: Results of the simulation: estimations and $95\%$ joint CI (A=Naive *one-p-at-a-time* method; D=*our method* with regularization)



# Simulation Results

Table: Area and coverage for the joint 95% credible intervals.

| **Type** | **A** (naive) | **B** (PCA) | **C** (no reg.) | **D** (regularized) |
|---|---|---|---|---|
| $\beta_1(p)$ | 1.603 (1.000) | 1.092 (0.999) | 1.186 (1.000) | 1.069 (1.000) |
| $\beta_2(p)$ | 2.246 (1.000) | 1.551 (1.000) | 1.706 (1.000) | 1.465 (1.000) |
| $\beta_3(p)$ | 2.242 (1.000) | 1.599 (1.000) | 1.717 (1.000) | 1.457 (1.000) |
| $\beta_4(p)$ | 2.281 (1.000) | 1.583 (1.000) | 1.651 (1.000) | 1.499 (1.000) |

Table: Probability scores for differences in mean, variance, and skewness.

| **True** | $\mathbf{H}_0$ | **A** | **B** | **C** | **D** | **E** (feature) |
|---|---|---|---|---|---|---|
| $\mu_1 = \mu_3$ | $\mu_1 = \mu_3$ | 0.001 | 0.193 | 0.211 | 0.217 | 0.205 |
| $\sigma_1 \neq \sigma_3$ | $\sigma_1 = \sigma_3$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| $\xi_1 = \xi_3$ | $\xi_1 = \xi_3$ | 0.374 | 0.498 | 0.488 | 0.479 | 0.389 |
| $\mu_2 = \mu_4$ | $\mu_2 = \mu_4$ | 0.001 | 0.447 | 0.465 | 0.445 | 0.438 |
| $\sigma_2 = \sigma_4$ | $\sigma_2 = \sigma_4$ | 0.002 | 0.420 | 0.334 | 0.331 | 0.187 |
| $\xi_2 \neq \xi_4$ | $\xi_2 = \xi_4$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

# GBM Data Analysis

**Response:** T1 MRI images from 64 patients in glioblastoma (GBM) study, $Y_{ij}$=intensity of pixel $j$ from subject $i$, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, with $m_i$ ranging from $371$ to $3421$.

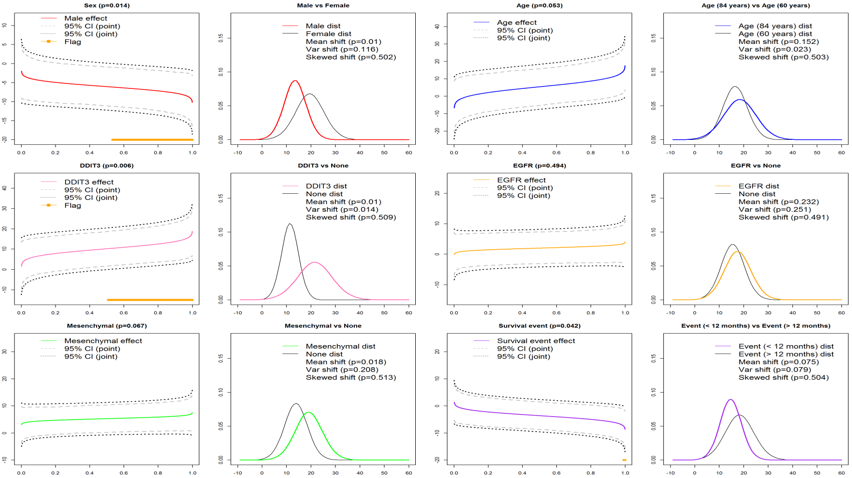
**Covariates:**

- **Demographic variables:** *sex* (21 F/43M) & *age* (56.5yr)
- **GBM subtype:** *mesenchymal* (30 mes./34 other)
- **Clinical outcome:** *survival* $(> 12\text{m}/< 12\text{m})$
- **Genetic alterations:** *DDIT3*(6m/58wt) & *EGFR*(24m/58wt)

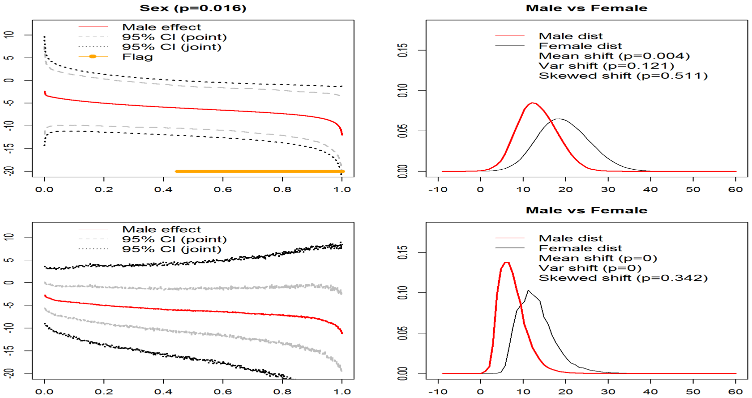
## Model

$$
\begin{aligned}
Q_i(p|X_i) = &\beta_0(p) + x_{\text{sex},i}\beta_{\text{sex}}(p) + x_{\text{age},i}\beta_{\text{age}}(p) + x_{\text{surv,i}}\beta_{\text{surv}}(p) \\
&+ x_{\text{Mes},i}\beta_{\text{Mes}}(p) + x_{\text{DDIT3},i}\beta_{\text{DDIT3}}(p) \\
&+ x_{\text{EGFR},i}\beta_{\text{EGFR}}(p) + E_i(p).
\end{aligned}
$$

# Full Results

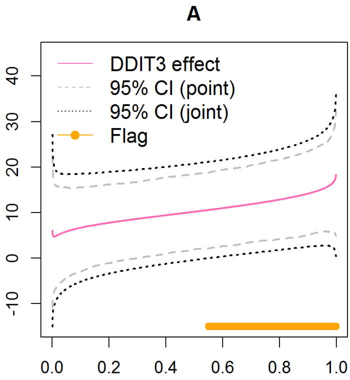


# GBM Results




- $P_{\text{sex},\mu} = 0.004,\ P_{\text{sex},\sigma^2} = 0.121,\ P_{\text{sex},\xi} = 0.51$

# GBM Results




- $P_{\text{DDIT3},\mu} = 0.008, \; P_{\text{DDIT3},\sigma^2} = 0.023, \; P_{\text{DDIT3},\xi} = 0.468$

# Summary

- General approach to regress distributions on covariates
- Useful in many settings without missing any information and insights on distributions
- Our framework yields global and local tests that adjust for multiple testing
  - Greater power than naive *one-p-at-a-time* approach
  - No power loss compared with feature extraction
- Applications of interest (future work)
  - Various types of imaging data
  - Climate change data
  - Activity data/wearable computing

# Reference

1. **Yang, H.**, Baladandayuthapani V., and Morris, J.S. (2020), “Quantile Function on Scalar Regression Analysis for Distributional Data”, *Journal of American Statistical Association*, 115, 90-106.
2. Morris, J. S. (2015), “Functional Regression”, *Annual Review of Statistics and Its Applicaion*, 2, 321-359.
3. Just, N. (2014), “Improving Tumour Heterogeneity MRI Assessment with Histograms”, *British Journal of Cancer*, 111, 2205-2213.

# Thank you.