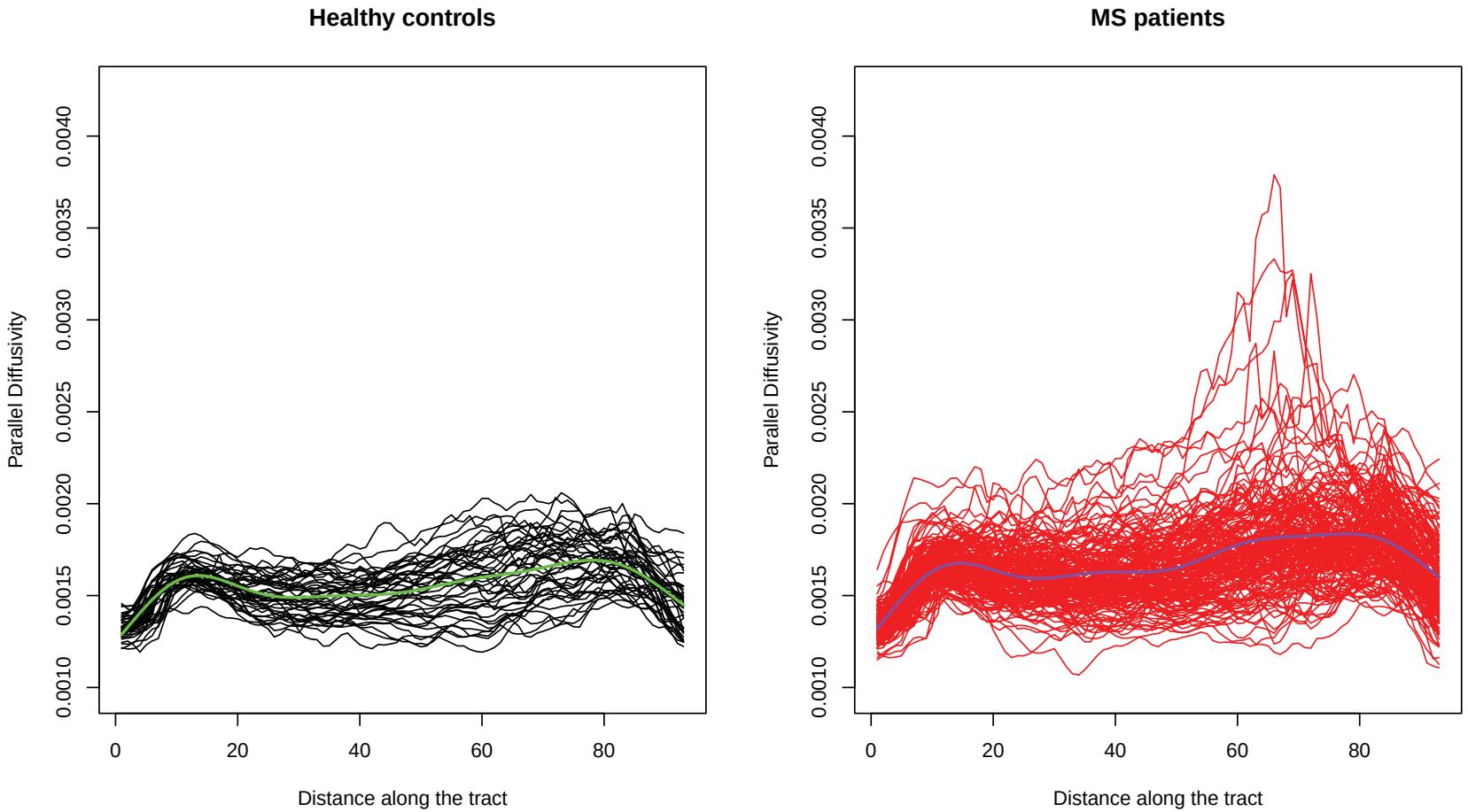
Simultaneous quantile regression for censored and dependent data

Brian Reich and Luke Smith

Mean regression versus quantile regression

- Most methods for studying covariate effects focus on mean effects.
- In quantile regression, the covariates are allowed to affect all quantiles, including:
  - Center (0.5 quantile)
  - Tail (0.95 quantile)
  - Spread (0.75 quantile - 0.25 quantile).
- This provides a more comprehensive and robust analysis.

DTI output along the corpus callosum

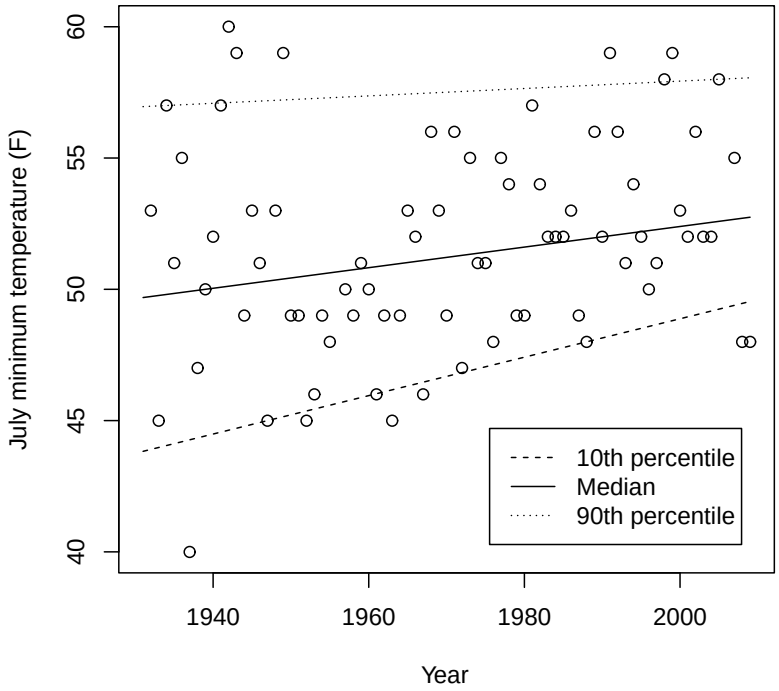


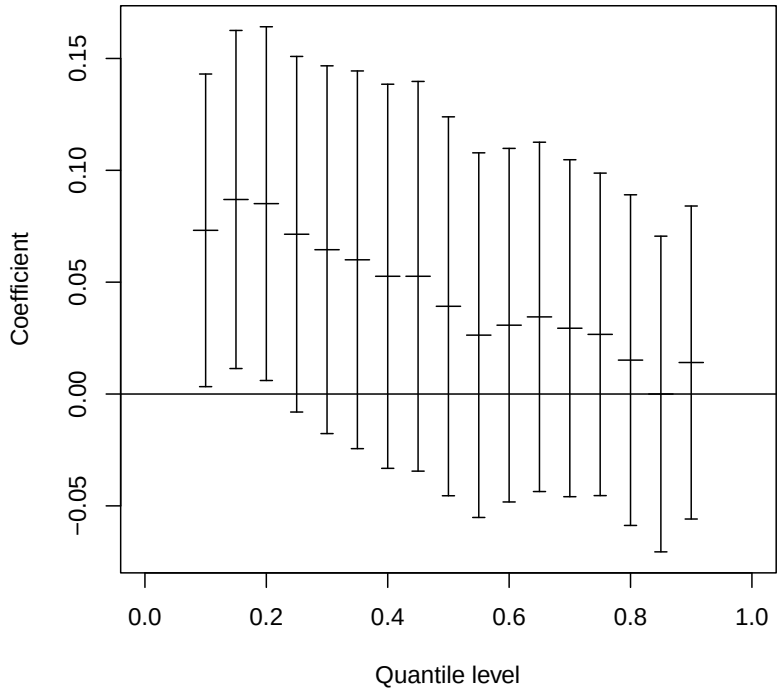
Climate change versus global warming

- Global warming refers to an increase in mean temperature.
- Climate change is more general, and refers to changes in the distribution of climate variables.
- This includes changes in the mean, increased variability, and severity of extreme events.
- All of these changes can be studied using quantile regression.

July minimum temperature data by year

Let Y_t be the value in year t.





Here the τ^{th} quantile is modeled as $\beta_0(\tau) + \beta_1(\tau)t$.

Quantile regression

- Mean regression: $E(Y) = \mathbf{X}^T\boldsymbol{\beta}$.
- Median regression: $\text{Median}(Y) = \mathbf{X}^T\boldsymbol{\beta}$.
- Quantile regression: $Q(\tau) = \mathbf{X}^T\boldsymbol{\beta}(\tau)$, where $Q(\tau)$ is the τ^{th} quantile of Y so that $P[Y < Q(\tau)] = \tau \in [0, 1]$.
- $\boldsymbol{\beta}(\tau)$ gives the covariate effects on the τ^{th} quantile.
- With $\tau = 0.5$, this is median regression.
- With $\tau = 0.99$, this allows us to test for the effects of covariates on the magnitude of extreme values.

Classical estimation methods

- **Mean regression**: $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \sum (Y_i - \mathsf{X}_i^T \boldsymbol{\beta})^2$.
- **Median regression**: $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \sum |Y_i - \mathsf{X}_i^T \boldsymbol{\beta}|$.
- **Quantile regression**: $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \sum \rho_\tau(Y_i - \mathsf{X}_i^T \boldsymbol{\beta})$ where

$$\rho_\tau(e) = \begin{cases} -(1-\tau)e & e < 0 \\ \tau e & e \geq 0 \end{cases}$$

- Typically quantile regression is implemented separately for different quantile levels, τ.

Summary of the classical approach to QR

Advantages:

- Provides a comprehensive and interpretable analysis.
- Semiparametric and robust estimator of β.
- Computationally efficient.

Disadvantages:

- Does not allow for prediction (no model, crossing quantiles).
- Does not borrow information across quantile levels.
- Difficult to incorporate missing data, censoring, dependence, etc.

Our general approach

- We take a model-based approach to quantile regression.
- Rather than analyzing quantile levels one-at-a-time, we analyze them simultaneously.
- The quantile function (inverse CDF) is modeled using basis functions, centered on a parametric error distribution.
- To illustrate the generality of the approach, we apply it to:
  1. Censored survival data with several covariates
  2. A longitudinal hypertension study.

Defining the quantile function

- The linear quantile function is $Q(\tau|X) = \sum_{j=0}^{p} X_j \beta_j(\tau)$, where $\beta_0(\tau)$,...,$\beta_p(\tau)$ are functions of τ.
- If $\beta_0(\tau) = \beta_0 + \sigma\Phi^{-1}(\tau)$ and $\beta_j(\tau) = \beta_j$, then

$$Y|X \sim \mathrm{N}\left(\sum_{j=0}^{p} X_j\beta_j, \sigma\right).$$

- If $\beta_j(\tau) = \beta_j + \alpha_j\Phi^{-1}(\tau)$ then

$$Y|X \sim \mathrm{N}\left(\sum_{j=0}^{p} X_j\beta_j, \sum_{j=0}^{p} X_j\alpha_j\right).$$

- Replacing the Gaussian CDF with another parametric quantile function q_0 gives a general location-scale family.

Defining the quantile function

- $Q(\tau|\mathrm{X})$ must be increasing in τ for every X.
- We model the derivative piece-wise over $L > 1$ intervals separated by breakpoints $0 = \kappa_0 < \kappa_1 < ... < \kappa_{L-1} < \kappa_L = 1$,

$$\frac{dq(\tau|\mathrm{X})}{d\tau} = \sum_{l=1}^{L} \mathrm{I}\left(\kappa_{l-1} < \tau \leq \kappa_l\right)(\mathrm{X}^T\boldsymbol{\alpha}_l)\frac{dq_0(\tau)}{d\tau}.$$

- The covariate effects for $\tau \in [\kappa_{l-1}, \kappa_l]$ are determined by $\boldsymbol{\alpha}_l = (\alpha_{l0}, ..., \alpha_{lp})^T$.
- The derivative is positive for all τ if and only if $\mathrm{X}^T\boldsymbol{\alpha}_l > 0$ for all $l > 0$ and X.

Monotonicity constraints

Recall: We must have $X^T \boldsymbol{\alpha}_l > 0$ for all $l > 0$ and for all X.

- This cannot possibly hold for all $X \in \mathcal{R}^p$, so we assume $X_0 = 1$ and $X_j \in [-1, 1]$.
- Then $X^T \boldsymbol{\alpha}_l$ is minimized (over X) at

$$WC(\boldsymbol{\alpha}_l) = \alpha_{l0} - \sum_{j=1}^{p} |\alpha_{lj}|.$$

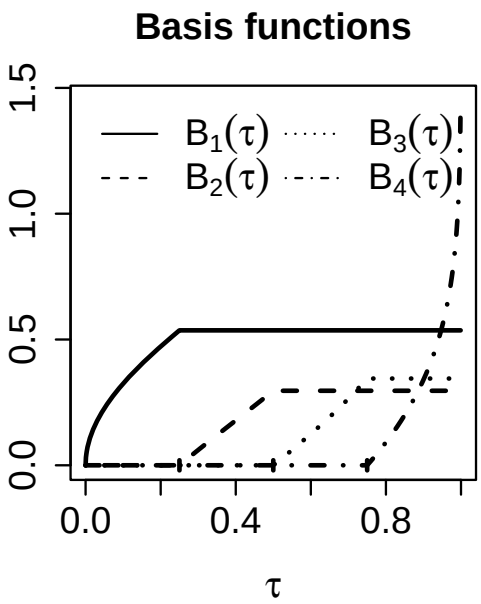
- To satisfy this criteria, we build the prior using latent unconstrained coefficients $\boldsymbol{\alpha}_l^* = (\alpha_{l0}^*, ..., \alpha_{lp}^*)$, and set

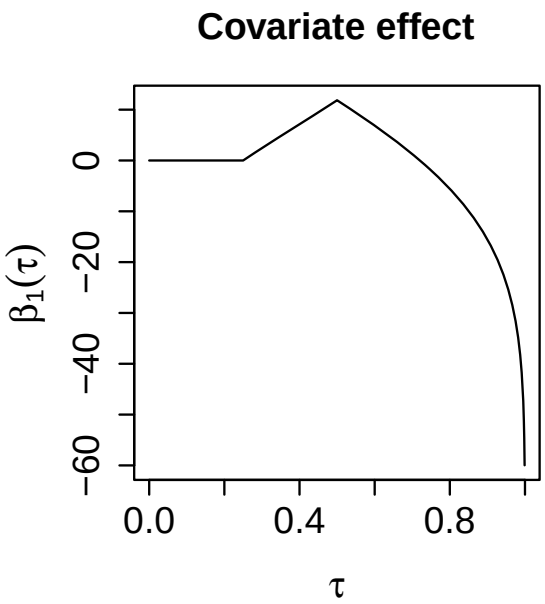
$$\alpha_{lj} = \begin{cases} \alpha_{lj}^*, & WC(\boldsymbol{\alpha}_l^*) > \epsilon \\ \epsilon I(j = 0), & WC(\boldsymbol{\alpha}_l^*) < \epsilon, \end{cases}$$

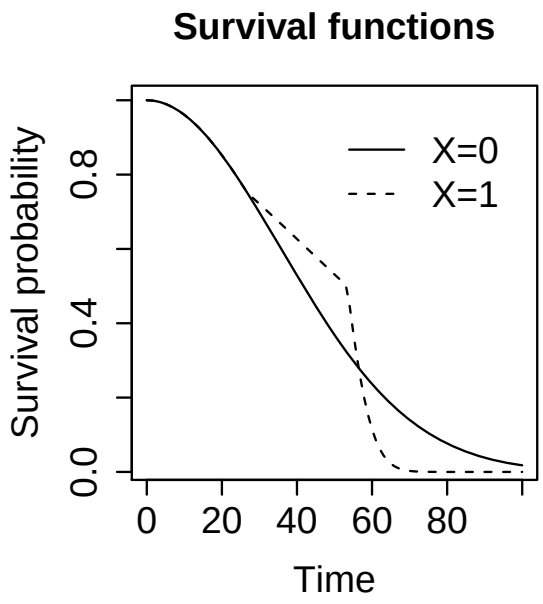
The resulting quantile function

- The quantile function for covariate j is $\beta_j(\tau) = \alpha_{0j} + \sum_{l=1}^{L} B_l(\tau)\alpha_{lj}$.
- This is a linear combination of fixed basis functions B_l with coefficients α_{lj}.
- The basis functions are integrals of q_0's derivative over different intervals.
- The next slide takes q_0 to be Weibull and $L = 4$.
- There are $p = 1$ covariates with $(\alpha_{00}, ..., \alpha_{0L}) = (0, 50, 50, 50, 50)$ and$(\alpha_{10}, ..., \alpha_{1L}) = (0, 0, 40, -40, -40)$.

Weibull example with crossing survival curves







Model flexibility

- Although these restrictions may seem prohibitive, the model is quite flexible as illustrated by the following theorem.
- Let $\tilde{\boldsymbol{\beta}}(\tau)$ be any valid set of quantile functions and $q_0(\tau)$ be any valid base quantile function.
- **Theorem 1**: There exists ϵ and $\{\alpha_{lj}^*\}$ so that $\boldsymbol{\beta}(\tau) = \tilde{\boldsymbol{\beta}}(\tau)$ at the interior breakpoints $\tau \in \{\kappa_1, ..., \kappa_{L-1}\}$.
- Therefore, by adding enough knots any quantile process can be approximated.

The likelihood

- The likelihood has a relatively simple closed form.

$$f(y|\mathbf{X}, \boldsymbol{\alpha}^*) = \frac{1}{\mathbf{X}^T\boldsymbol{\alpha}_l} f_0\left[\frac{z - \mathbf{X}^T\boldsymbol{\alpha}_0 - I(l > 1)\mathbf{X}^T\boldsymbol{\alpha}_l q_0(\kappa_l)}{\mathbf{X}^T\boldsymbol{\alpha}_l}\right],$$

  where $Q(\kappa_{l-1}|\mathbf{X}) < y < Q(\kappa_l|\mathbf{X})$ and f_0 is the density corresponding to q_0.

- There are discontinuities at the interior breakpoints κ_l.

- Discontinuous densities are not uncommon in Bayesian nonparametrics.

- Averaging over the unknown parameters gives a smooth posterior mean.

Bayesian implementation

- Unlike classical quantile regression, we have a fully-specified likelihood.
- This makes adapting the model to realistic settings fairly straight-forward.
- For example, for censored survival data the censored likelihood is (δ_i is the censoring indicator)

$$\prod_{i=1}^{n} f[y_i|\mathrm{X}_i,\boldsymbol{\alpha}]^{1-\delta_i}\left\{1-F[y_i|\mathrm{X}_i,\boldsymbol{\alpha}]\right\}^{\delta_i}.$$

- We use MCMC, but maximum likelihood would likely be effective as well.

Variable selection

- The model is hard to fit with several covariates, especially when there are a lot of basis functions.
- Our solution is to (stochastically) classify the covariates into three groups:
  - No effect with $\beta_j(\tau) = 0$.
  - Additive effect with $\beta_j(\tau) = \alpha_{0j}$.
  - Non-additive effect with $\beta_j(\tau)$ varying by τ.
- In our experience, we find that many covariates can be added as additive effects and this greatly reduces model complexity.

Stochastic search variable selection prior

The quantile function for covariate j is

$$\beta_j(\tau) = \alpha_{0j} + \sum_{l=1}^{L} B_l(\tau)\alpha_{lj}.$$

- The prior is $\alpha^*_{lj} \overset{iid}{\sim} \mathrm{N}(0, \sigma_j^2)$ for $l > 0$.
- If $\sigma_j = 0$, then $\beta_j(\tau) = \alpha^*_{0j}$.
- Our prior for σ_j has mass at zero.
- Standard Bayesian variable selection methods apply for the location parameters α^*_{0j}.

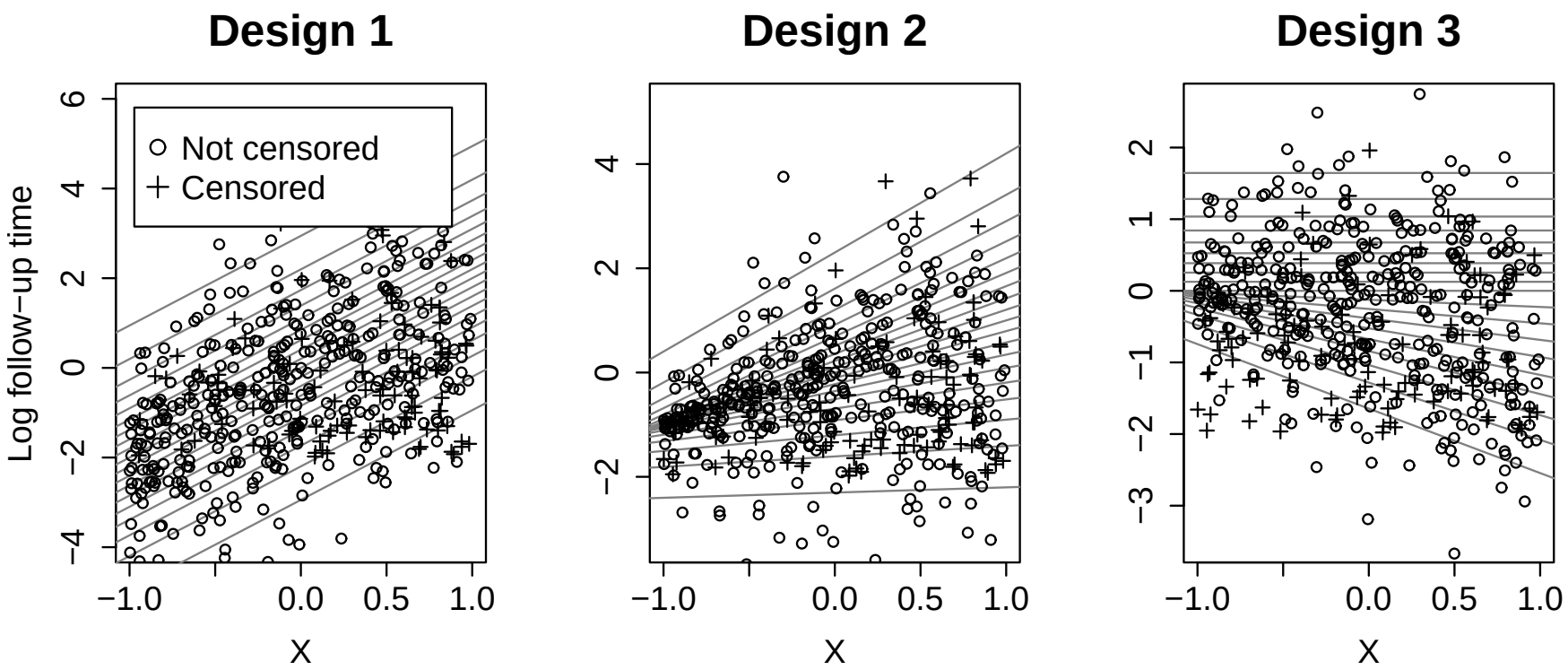
Simulation study for censored survival data

- We consider three simulation designs, each with $n = 250$ observations and $p = 1$ covariates.
- We consider three simulation designs:

1. $\beta_0(\tau) = \log[\tau/(1-\tau)]$
   $\beta_1(\tau) = 2$
2. $\beta_0(\tau) = \text{sign}(0.5-\tau)\log(1-2|0.5-\tau|)$
   $\beta_1(\tau) = 2\tau$
3. $\beta_0(\tau) = \Phi^{-1}(\tau)$
   $\beta_1(\tau) = 2\min\{\tau - 0.5, 0\}$

- The covariates are generated Unif(-1,1) and censoring is generated uniformly to give 20-30% censoring.

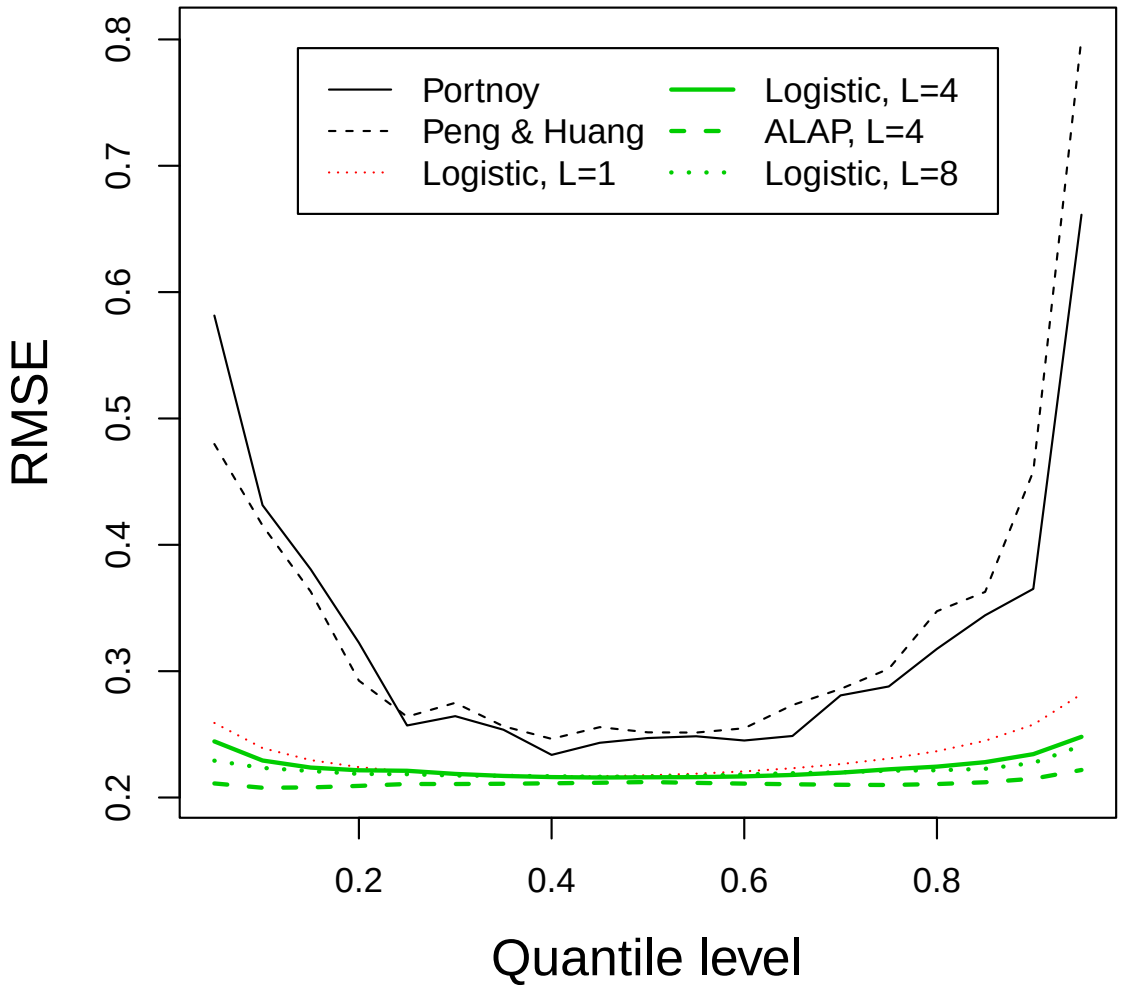
Simulation designs



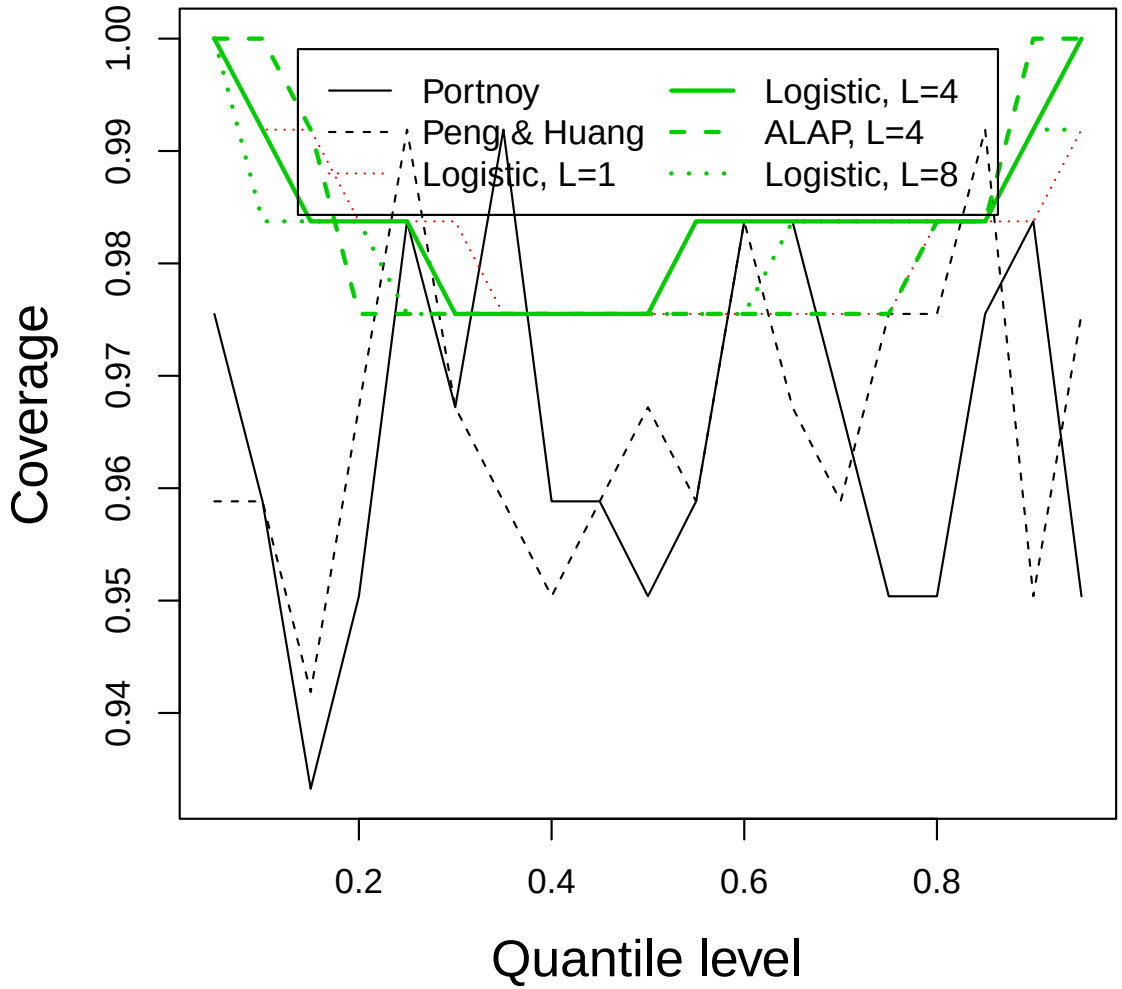
Comparisons

- Methods are compared using RMSE and coverage for $\beta_1(\tau)$.
- We compare with:
  1. The frequentist procedures of Portnoy (2003) and Peng and Haung (2008) implemented in `quantreg` package in `R`.
  2. The parametric heterodskedastic logistic model with $L = 1$ and $q_0(\tau) = \log[\tau/(1-\tau)]$.
  3. Our model with logistic q_0 with $L = 4$ and $L = 8$
  4. Our model with asymmetric Laplace q_0 with $L = 4$.

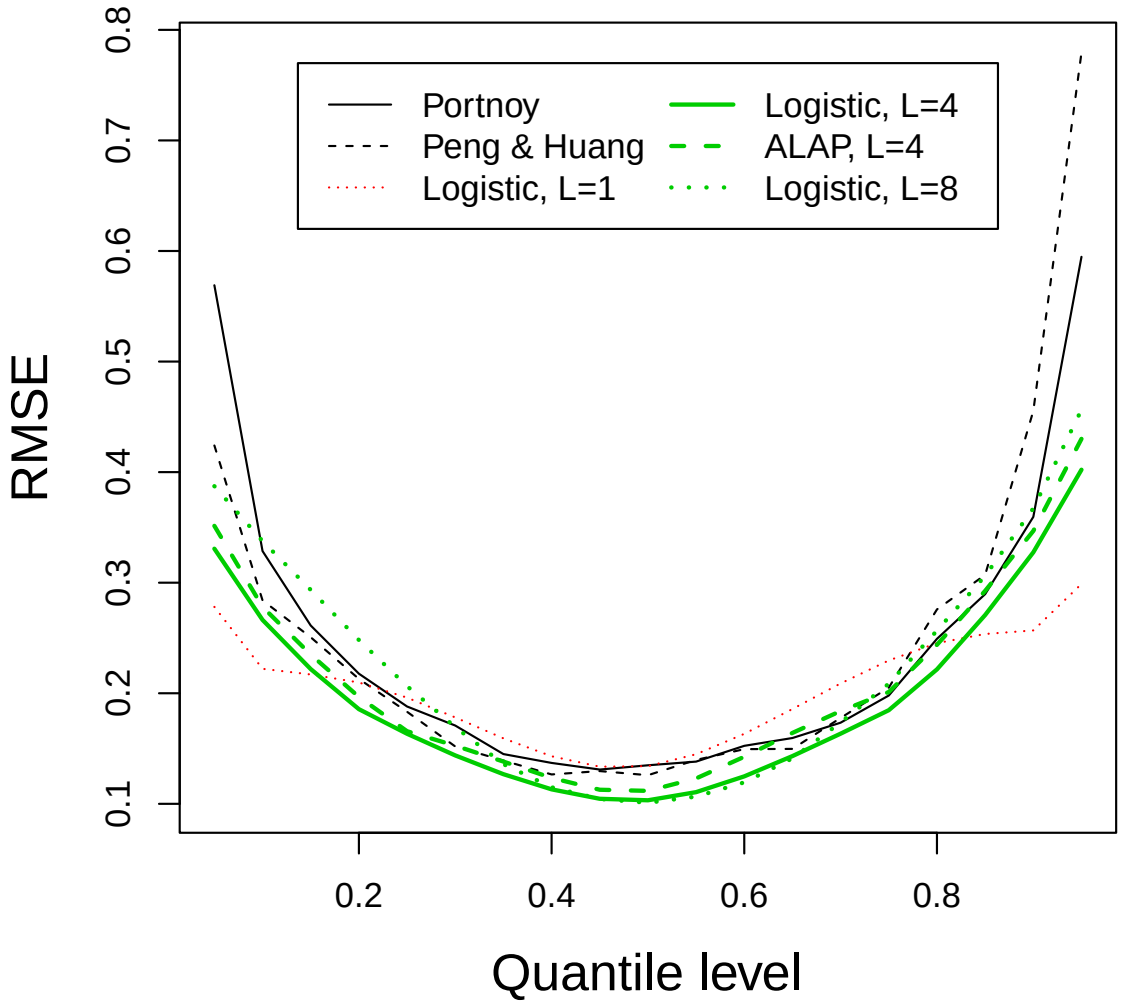
Design 1: RMSE



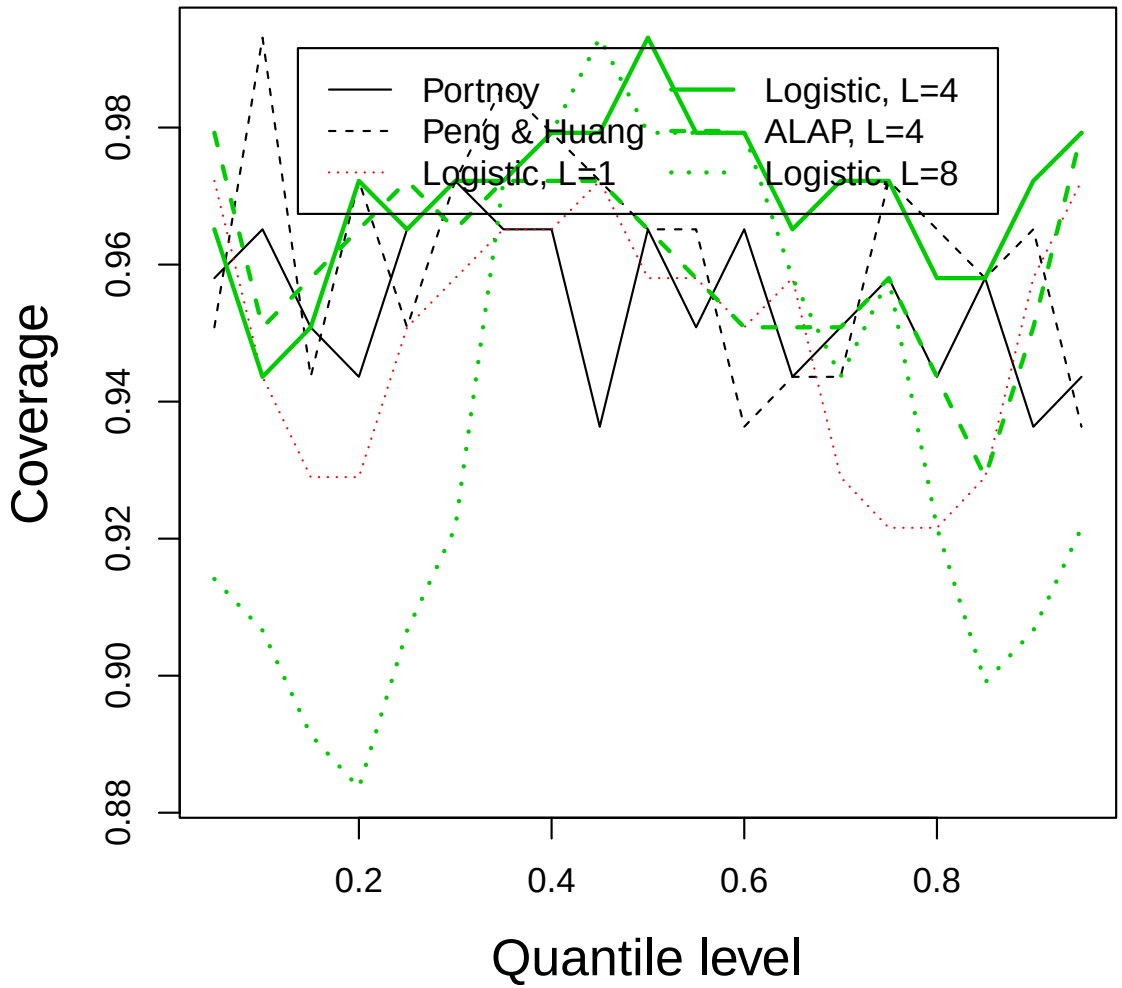
Design 1: Coverage of 95% intervals



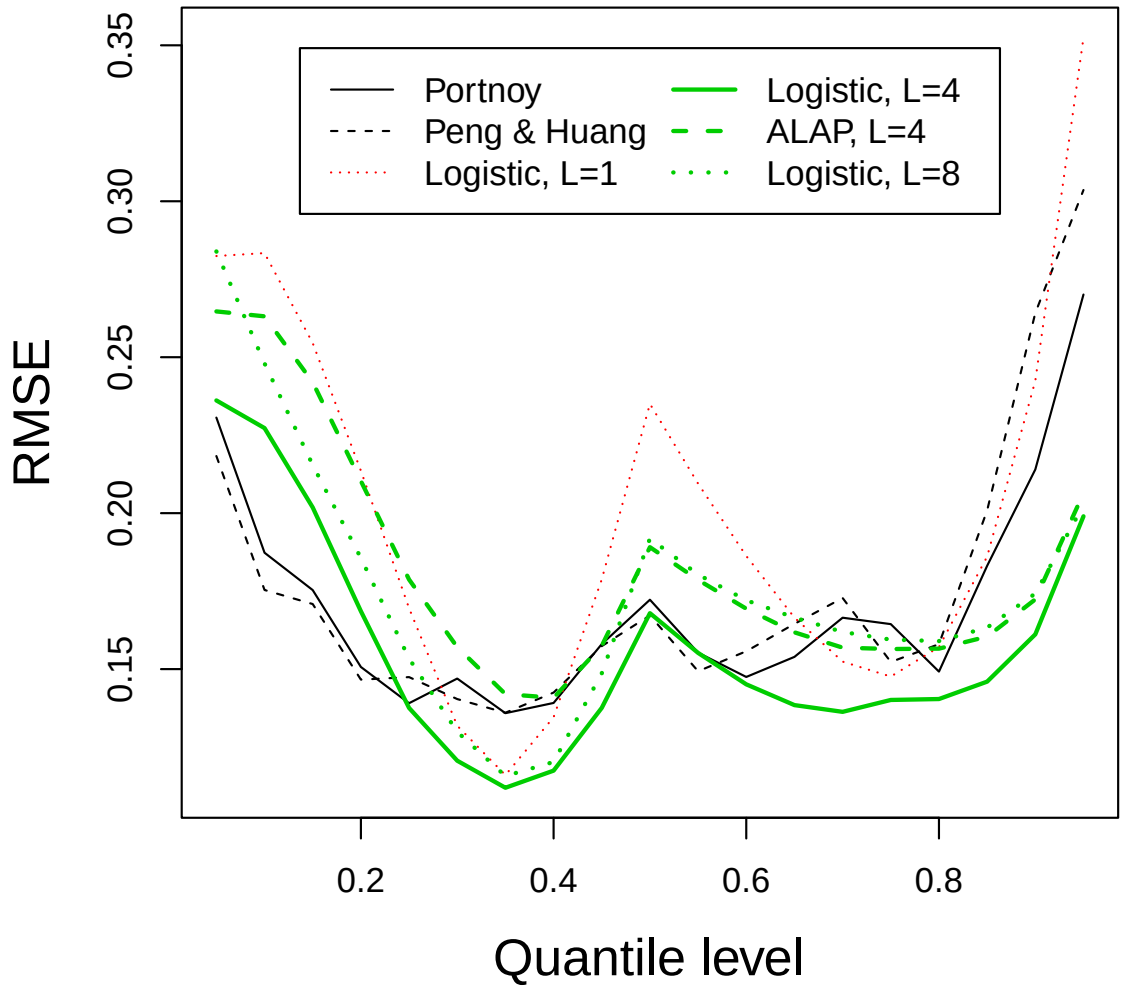
Design 2: RMSE



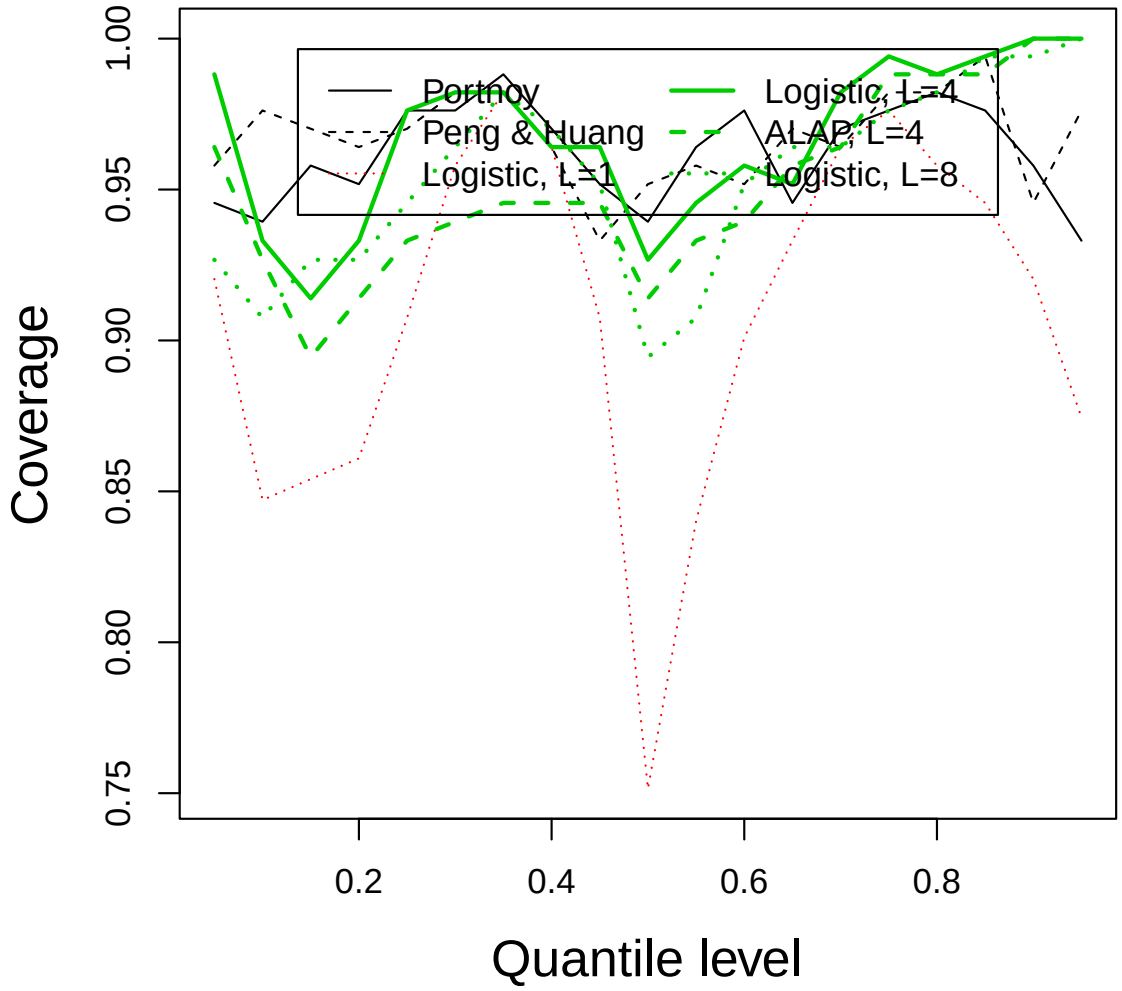
Design 2: Coverage of 95% intervals



Design 3: RMSE



Design 3: Coverage of 95% intervals



Summary

- When the linear regression model is correct, the Bayesian model gives a huge improvement.
- In other cases, the MSE is similar in the center of the distribution where data are abundant.
- The Bayesian approach improves estimation in the tails where data are sparse.
- The Bayesian approach gives tighter interval by borrowing strength across quantile levels (not shown).
- The results were similar in multiple regression.

Data analysis

- We analyze the UIS drug treatment study data available in the `quantreg` package in `R`
- The response is log time until relapse, and there are $n = 575$ observations with complete data.
- We use $p = 8$ predictors, including
  1. number of previous drug treatments (NDT)
  2. IV drug use (IV; "Yes" = 1, "No" = -1)
  3. treatment (TRT; "long" = 1, "short" = -1)
  4. compliance fraction ($FRAC$)

Model comparisons

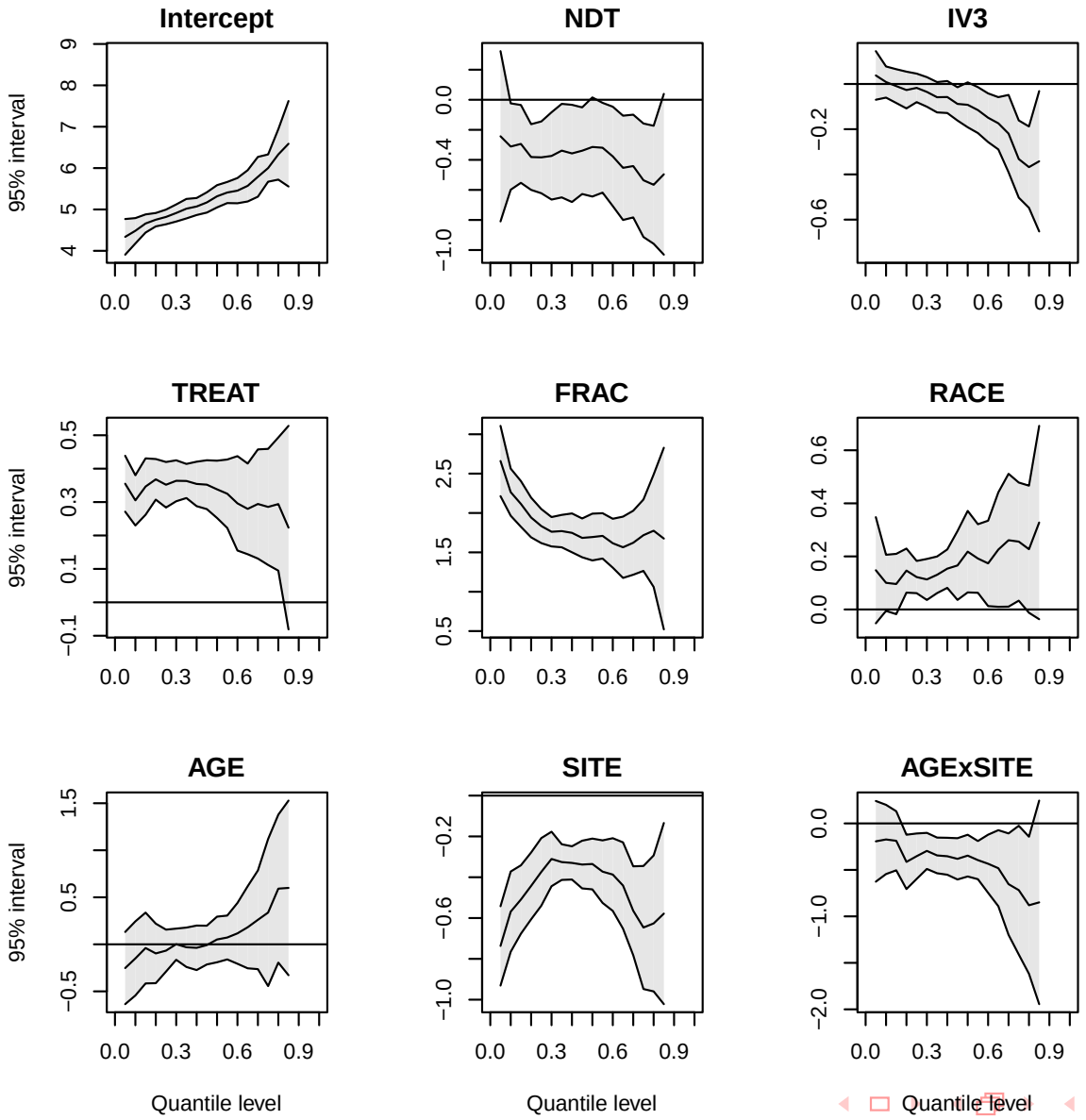
- We compare model-based fits using the log pseudo maximum likelihood statistic.

Base quantile function	$L = 1$	$L = 4$	$L = 8$	$L = 12$
Logistic	-195.8	-188.6	-186.9	-187.6
Asymmetric Laplace	-188.5	-185.7	-185.1	-186.3
T	-195.3	-186.4	-185.7	-187.9

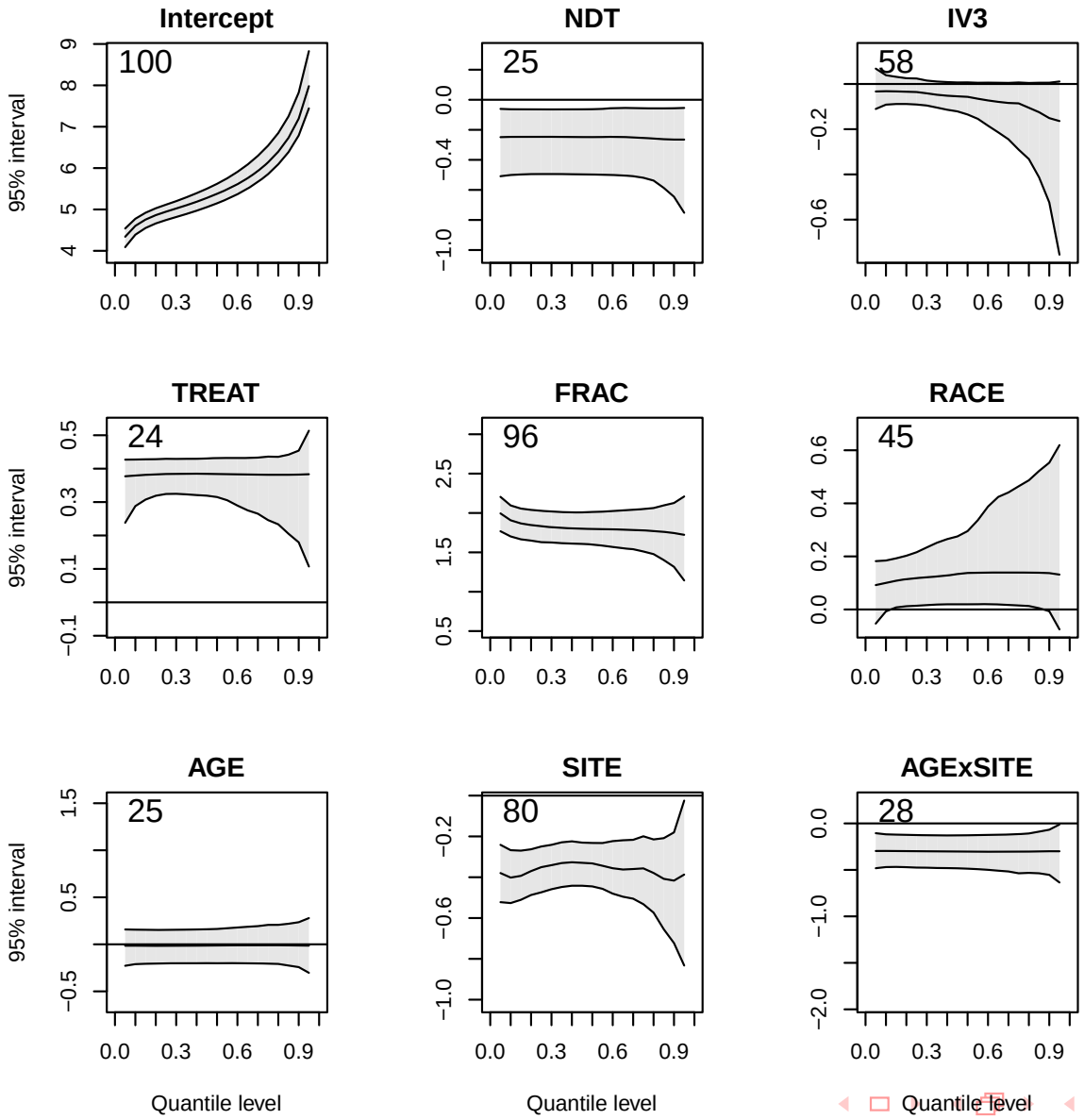
- Larger values are preferred.

- We present the results assuming the asymmetric Laplace base distribution and $L = 8$ basis functions.

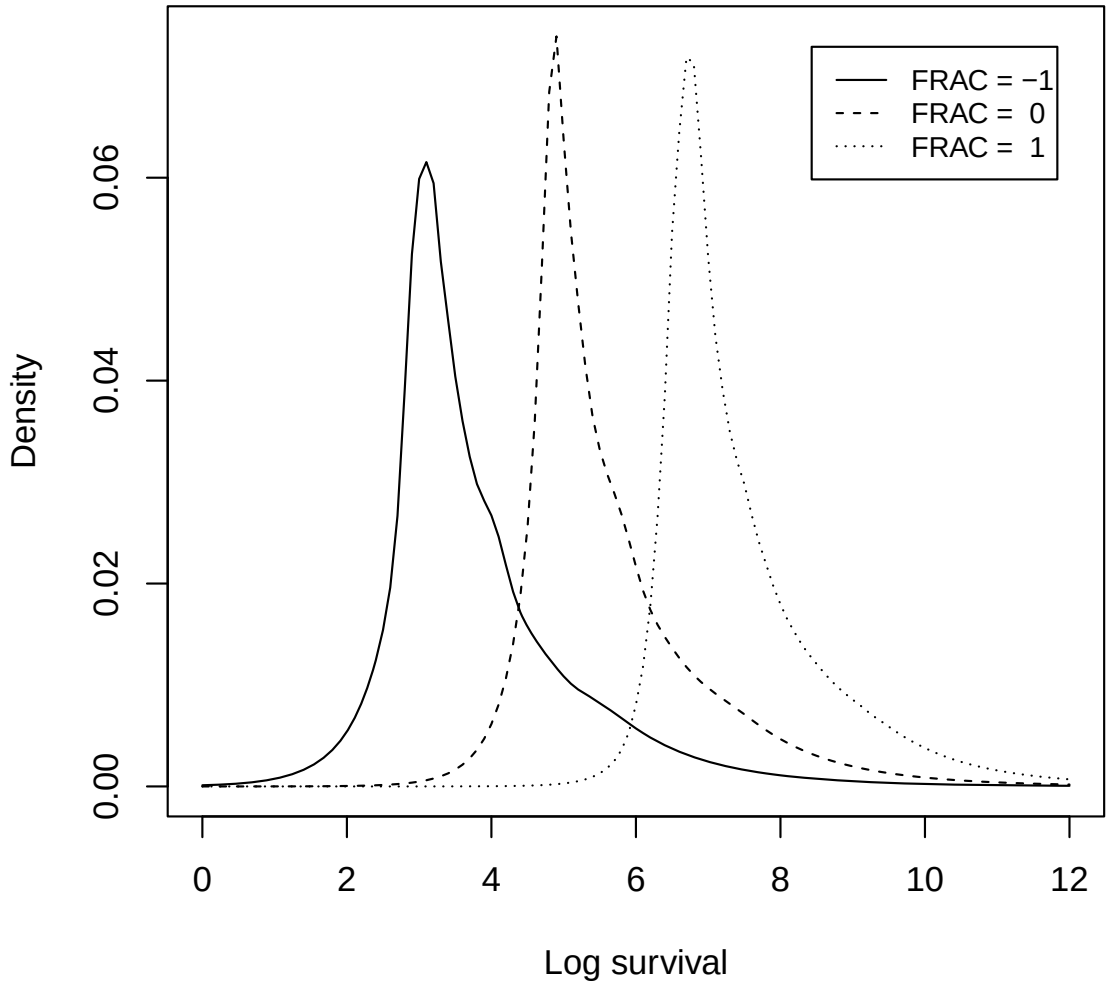
Posterior 95% intervals for Portnoy's method



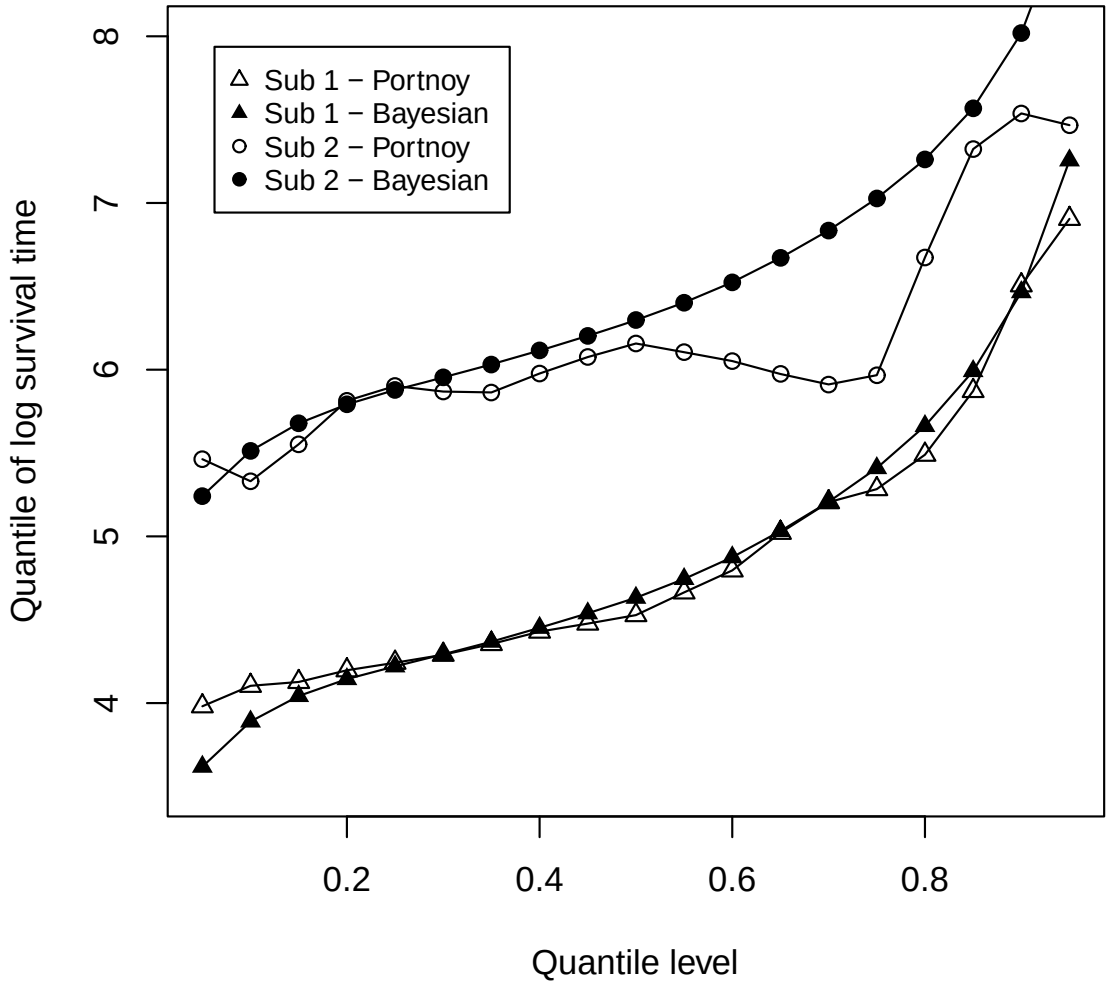
Posterior 95% intervals for our method



Estimated density by compliance level



Predictive model for two subjects



QR for longitudinal data

- Let Y_{ij} be response j for subject i, and X_{ij} be the associated covariate vector (could include functions of time).
- We would like to estimate the quantile function while accounting for autocorrelation.
- As before, the quantile function is

$$Q(\tau|\mathsf{X}_i) = \sum_{j=0}^{p} X_{ij}\beta_j(\tau)$$

and the covariate functions are

$$\beta_j(\tau) = \alpha_{0j} + \sum_{l=1}^{L} B_l(\tau)\alpha_{lj}.$$

The usual Gaussian mixed effects model

$$Y_i \sim X_i\beta + Z_i\gamma_i + E_i$$

- Y_i is the vector of data for subject i.
- X_i and Z_i are covariate matrices.
- β is the population fixed effect.
- $\gamma_i \sim N(0, \Delta)$ is a random effect.
- $E_i \sim N(0, \Omega)$ is error.

Equivalent representation: $Y_i \sim N(X_i\beta, \Sigma_i)$, where $\Sigma_i = Z_i\Delta Z_i^T + \Omega$ has diagonal elements $\{\sigma_{ij}\}$.

Extending to quantile regression

We want to:

1. Preserve $Q(\tau|\mathrm{X})$ as the marginal population quantile function.
2. Account for within-subject dependence via Σ_i.
3. Allow for random effects γ_i to study subject-specific deviation.

Extending to quantile regression

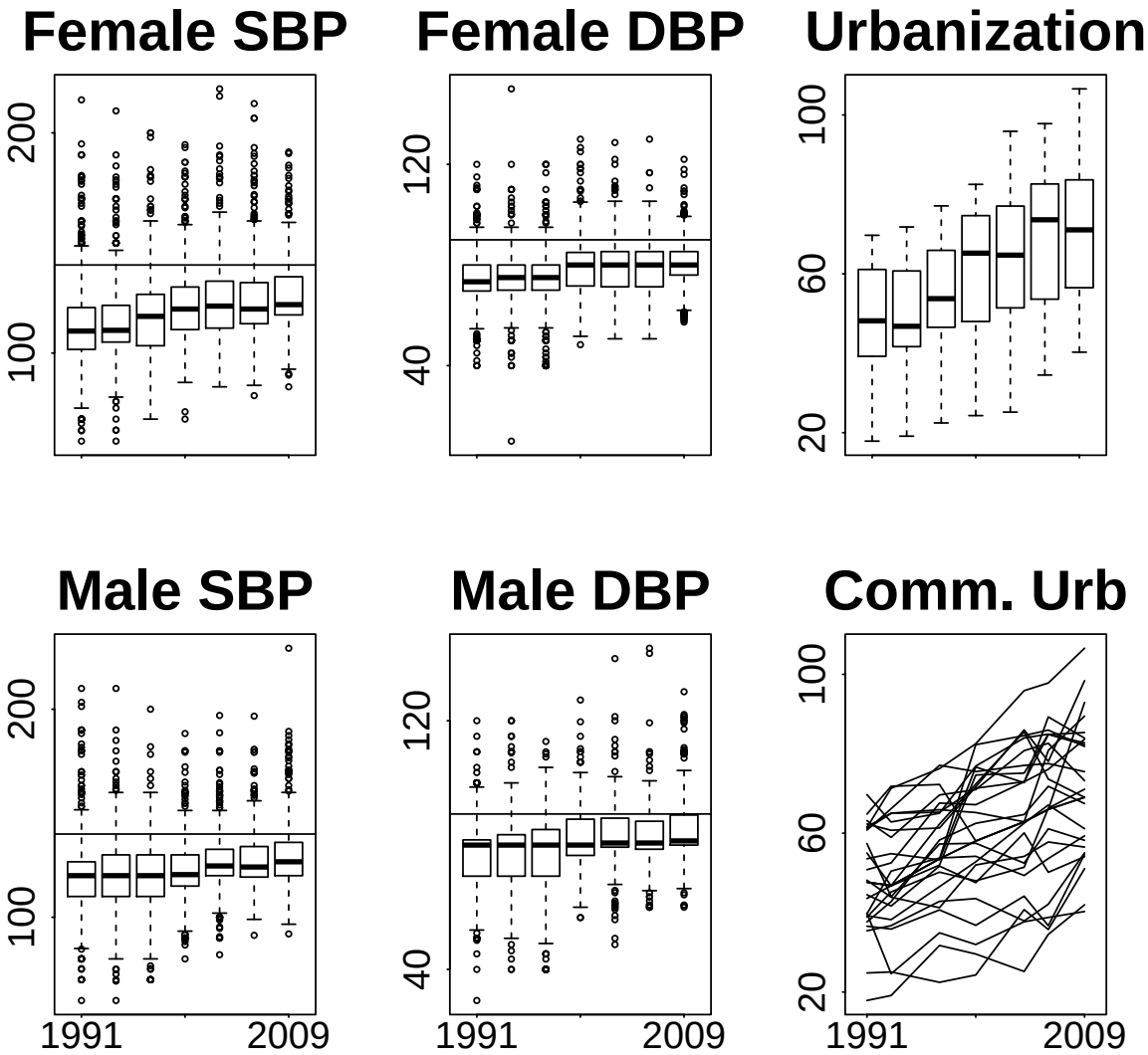
We achieve these modeling objectives using a Gaussian copula:

- $Y_{ij} = Q(U_{ij}|\mathrm{X}_{ij})$, where $U_{ij} = \Phi(V_{ij}/\sigma_{ij})$
- $\mathrm{V}_i \sim \mathrm{N}(\mathrm{Z}_i\gamma_i, \Omega)$ and $\gamma \sim \mathrm{N}(0, \Delta)$.
- For identification, we restrict Ω to be a correlation matrix.

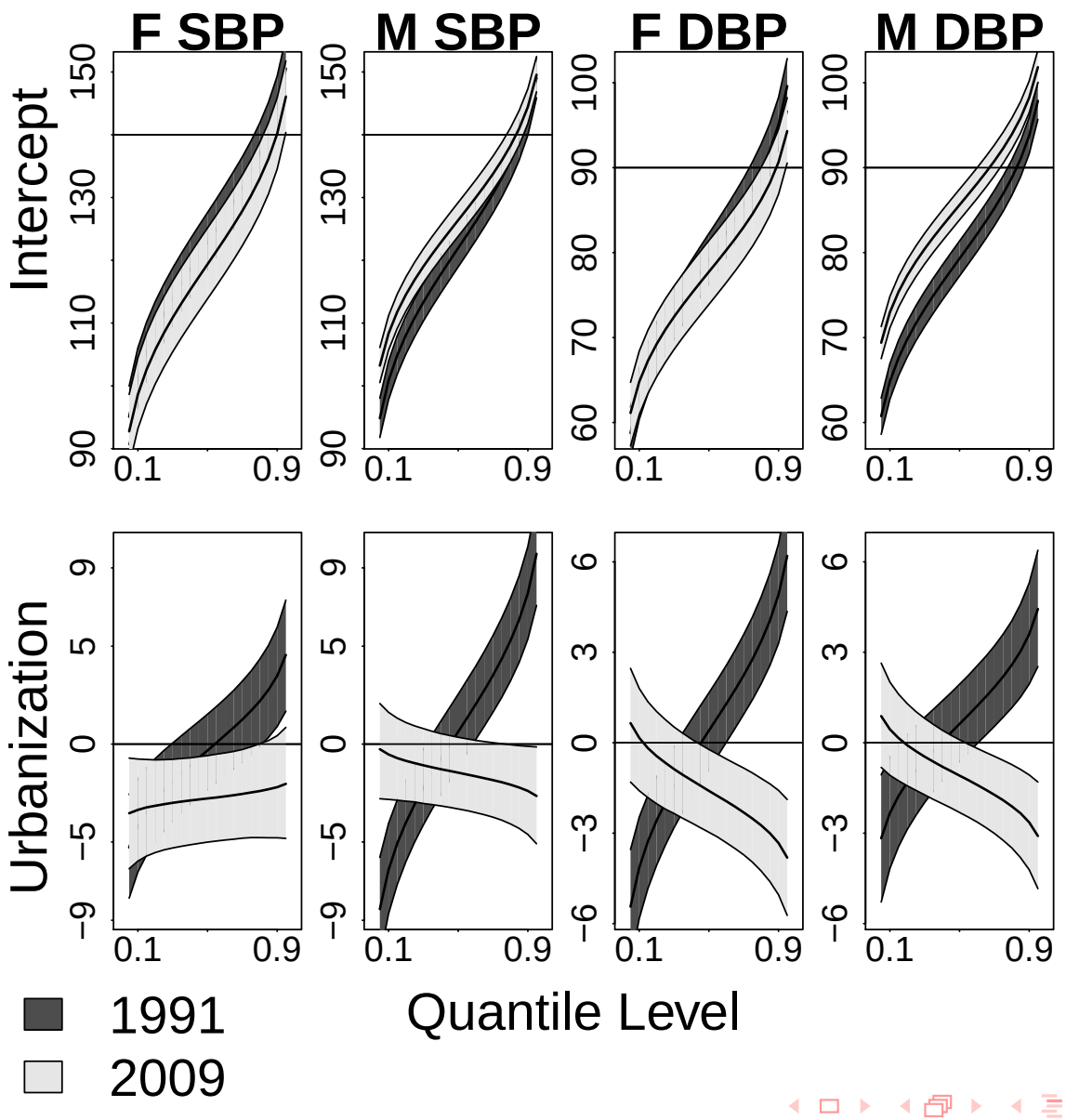
The China Health and Nutrition Survey

- Longitudinal survey collected from 1991-2009 to study the effects of urbanization and human health.
- The outcome is the subject's blood pressure, which is measured on seven occasions.
- The covariates of interest are the year, urbanicity of the subject's community, and the interaction.
- Many other covariates are included such as age, gender, etc.
- The primary scientific interest is in the marginal quantile function Q.

Exploratory analysis



Results


F SBP
M SBP
F DBP
M DBP
Intercept
Urbanization
Quantile Level
1991
2009

Summary

- We have proposed new approaches for Bayesian quantile regression.
- The methods are very flexible, and perform well compared to classic approaches.
- The methods are available in a new `R` package entitled `BSquare`!!!
- References:
  - Smith LB, Gordon-Larsen P, Reich BJ, Fuentes M. QR for mixed models. *In revision*.
  - Reich BJ, Smith LB (2014). Bayesian QR for censored data. *Biometrics*.
  - Reich BJ (2012). Spatiotemporal QR for detecting distributional changes in environmental processes. *JRSS-C*.
- Support: NIH - 5R01ES014843-02, NSF - 1107046.