BIOST 572: Final Talk

Cheng Zheng

Biostatistics, University of Washington

May 24th, 2012

The Paper

Appl. Statist. (1999)
48, *Part* 4, *pp.* 533–551

Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children

Patrick J. Heagerty
University of Washington, Seattle, USA

and Margaret S. Pepe
Fred Hutchinson Cancer Research Center, Seattle, USA

Motivational Example

- Data: Observational longitudinal study of obesity from birth to adulthood.
- Overall Goal: Build age-, gender-, height-specific growth charts (under 3 years) to diagnose growth abnormalities.
- Specific Aim: Estimate the reference range for age-, gender-, height-specific weight.
- A simple version: Estimate the reference range for age-, gender-specific BMI.
- Statistical Problem: Estimate covariate specific quantiles in a reference group.
- Other Applications: Regression using placement value/receiver operating characteristic (ROC) regression.

Data Display

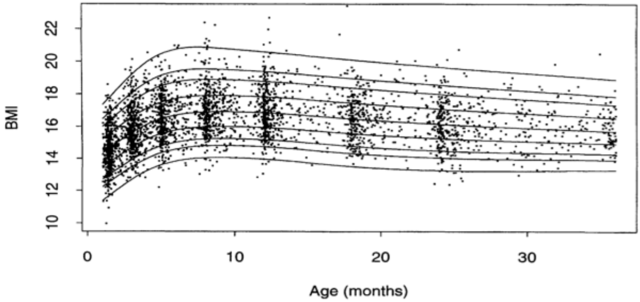


Figure: Estimated 1st, 5th, 10th, 25th,50th, 75th, 90th, 95th and 99th percentiles of BMI as a function of age

Previous Solutions: Bin and Smooth Estimation

- Bin and Smooth Quantiles (BSQ): Empirical quantiles for each narrow interval $X \pm \lambda$, then smoothed it (with splines). (Hamill et al., 1977)
- Bin and Smooth Parameters: Model $f(Y|X)$ by $\theta(X)$, estimate $\theta(X)$ for each narrow interval of $X \pm \lambda$, then smoothed it.
  Cole (1990) let $\theta(X) = \{\mu(X), \sigma(X)\}$ and assume $Y|X$ follows normal distribution with mean $\mu(X)$ and standard deviation $\sigma(X)$.

  $$Y^\alpha = \mu(X) + \sigma(X)z^\alpha,$$

  z^α is the αth quantile for standard normal distribution.
- Limitations: (1) Require large sample size; (2) Curse of dimensionality.

Previous Solutions: Parametric Models

- Idea: Specify a fully parametric model for $Y|X$ indexed by parameter $\boldsymbol{\theta}$, then estimate $\boldsymbol{\theta}$ via likelihood.
- LMS (Cole and Green 1992): Assume Y can be transformed to the standard normal random variable Z as follow:

$$Z = \frac{\{Y/M(X;\boldsymbol{\theta})\}^{L(X;\boldsymbol{\theta})} - 1}{L(X;\boldsymbol{\theta})S(X;\boldsymbol{\theta})},$$

  $M(X;\boldsymbol{\theta})$ is median response, $L(X;\boldsymbol{\theta})$ is Box-Cox power transformation term and $S(X;\boldsymbol{\theta})$ approximate variance.

$$Y^{\alpha}(X;\boldsymbol{\theta}) = M(X;\boldsymbol{\theta})\{1 + z^{\alpha}L(X;\boldsymbol{\theta})S(X;\boldsymbol{\theta})\}^{1/L(X;\boldsymbol{\theta})}$$

- Limitation: (1) The distribution assumption (transformed normal distribution); (2) Sensitivity of the transformation part $L(X)$ to outliers.

Previous Solutions: Nonparametric Models

- Idea: Directly estimate $Y^\alpha(X)$ without assuming certain distribution for Y.
- Quantile Regression (QR): Koenker and Bassett (1978) proposed an M-estimation to obtain $\hat{Y}^\alpha(X)$ that minimize

$$\sum_i \alpha\{Y_i - Y^\alpha(X)\}_+ + (1-\alpha)\{Y_i - Y^\alpha(X)\}_-,$$

  $x_+ = max(0, x)$ and $x_- = max(0, -x)$.
- Limitation: $\hat{Y}^\alpha(X)$ may not be monotone in α.

New method: Semiparametric Models (SM)

- Allow the shape depend on X and do not specify specific distribution for Y. Model $\mu(X)$ and $\sigma(X)$ parametrically to gain efficiency.
- General model

$$Y_i = \mu(X_i; \boldsymbol{\theta}) + \sigma(X_i; \boldsymbol{\theta})\varepsilon(X_i),$$

$\mu(X_i; \boldsymbol{\theta})$ is the location parameter, $\sigma(X_i; \boldsymbol{\theta})$ is the scale parametre, i.e. $\sqrt{Var(Y_i|X_i)}$, and $\varepsilon(X_i)$ is from baseline distribution with mean zero, unit varaince. Denote baseline distribution function by $F_0(z, X) = P(\varepsilon(X) \leq z|X)$, we have

$$Y^\alpha(X; \boldsymbol{\theta}, F_0) = \mu(X_i; \boldsymbol{\theta}) + \sigma(X_i; \boldsymbol{\theta})Z^\alpha(X),$$

$Z^\alpha(X)$ is the αth quantile of $\varepsilon(X)$, i.e.

$$F_0(Z^\alpha(X), X) = \alpha.$$

Quasi-likelihood

Using independent working correlation and use normal distribution as working model for $Y|X$, we obtain quasi-likelihood score equation as below:

$$
\begin{aligned}
0 &= \frac{\partial \mu(X;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{Y-\mu(X;\boldsymbol{\beta})}{Var(Y|X)} \\
0 &= \frac{\partial \sigma^2(X;\gamma)}{\partial \gamma} \frac{(Y-\mu(X;\boldsymbol{\beta}))^2-\sigma^2(X;\gamma)}{Var[(Y-\mu(X;\boldsymbol{\beta}))^2|X]} \\
&= \frac{\partial \sigma^2(X;\gamma)}{\partial \gamma} \frac{(Y-\mu(X;\boldsymbol{\beta}))^2-\sigma^2(X;\gamma)}{2Var(Y|X)^2}
\end{aligned}
$$

Model details

- We can use splines to model $\mu(X)$ and $\sigma(X)$. Let $\boldsymbol{\theta} = \{\beta_1, \cdots, \beta_p, \gamma_1, \cdots, \gamma_q\}$, $R(X) = \{R_1(X), \cdots, R_p(X)\}$ and $S(X) = \{S_1(X), \cdots, S_q(X)\}$ are pre-specified regression spline basis functions.

$$\mu(X) = \sum_{k=1}^{p} \beta_k R_k(X), \log\{\sigma(X)\} = \sum_{k=1}^{q} \gamma_k S_k(X)$$

The scores becomes

$$\begin{aligned} 0 &= \sum_i R(X_i)^T \frac{Y_i - \mu(X_i; \boldsymbol{\beta})}{\sigma^2(X_i; \gamma)} \\ 0 &= \sum_i S(X_i)^T \frac{(Y_i - \mu(X_i; \boldsymbol{\beta}))^2 - \sigma^2(X_i; \gamma)}{\sigma^2(X_i; \gamma)} \end{aligned}$$

Estimate Baseline Function

- Obtain the estimated residuals

$$\hat{e}_i(X_i) = \frac{Y_i - \mu(X_i; \hat{\beta})}{\sigma(X_i; \hat{\gamma})},$$

- Special case: F_0 does not depend on X.

$$\hat{F}_0(z, X) = n^{-1} \sum_{i=1}^{n} I(\hat{e}_i \leq z).$$

- Step function $\Rightarrow$ continuous function.

$$\hat{F}_0(z, X) = n^{-1} \sum_{i=1}^{n} K_{\lambda_2}(z, \hat{e}_i),$$

where $K_{\lambda_2}(z, \hat{e}_i) = \Phi\{(z - \hat{e}_i)/\lambda_2\}$. $\hat{F}_0(z, X)$ is monotonically increasing in z.

Estimate Baseline Function

- General case: F_0 does depend on X.

$$\hat{F}_0(z, X) = \frac{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i) I(\hat{e}_i \leq z)}{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i)},$$

  where $w_{\lambda_1}(X, X_i) = \phi((X - X_i)/\lambda_1)$.
- Continuous version:

$$\hat{F}_0(z, X; \lambda_1, \lambda_2) = \frac{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i) K_{\lambda_2}(z, \hat{e}_i)}{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i)},$$

- For λ_1 and λ_2, we can use either fixed value or allow them to be functions of X.
- Estimate αth quantile by $\mu(x, \hat{\beta}) + \sigma(x, \hat{\gamma}) \hat{F}_0^{-1}(\alpha, X)$.

Nonparametric kernel estimator (NKE) based on Y_i

- A nonparametric way will be use

$$\hat{F}(z, X) = \frac{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i) I(Y_i \leq z)}{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i)},$$

  then estimate αth quantile by $\hat{F}^{-1}(\alpha, X)$
- To estimate $\hat{F}(z, X)$, assuming a uniform distribution for X,

$$\begin{aligned} Bias &= \frac{1}{4}\ddot{F}(z, X)\lambda_1^2\sigma_W^2 \\ Var &= (n\lambda_1)^{-1} F(z, X) \int W(u)^2 du \end{aligned}$$

- Optimal weight is $\lambda_1^* = \left(\frac{n^{-1}F(z,X)\int W(u)^2 du}{\ddot{F}(z,X)^2\sigma_W^4}\right)^{1/5}$.
- Minimum mean square error (MSE) is in order of $n^{-2/5}$.

Multiple covariates

- Estimating equation part: Use two set of spline basis $R(X_1)$, $R(X_2)$ and $S(X_1)$, $S(X_2)$. Use their tensor product to generate the new basis $R(X) = R(X_1) \otimes R(X_2)$, $S(X) = [S(X_1), S(X_2)]$.
- Baseline estimator: Not approximate indicator function by continuous one. For kernel $W(X)$, can use any multivariate kernel, for example the tensor product of two univariate kernel $W_1(X_1) \times W_2(X_2)$. In application, they use $W(X) = W_1(X_1)$.

Simulation: Methods Comparing

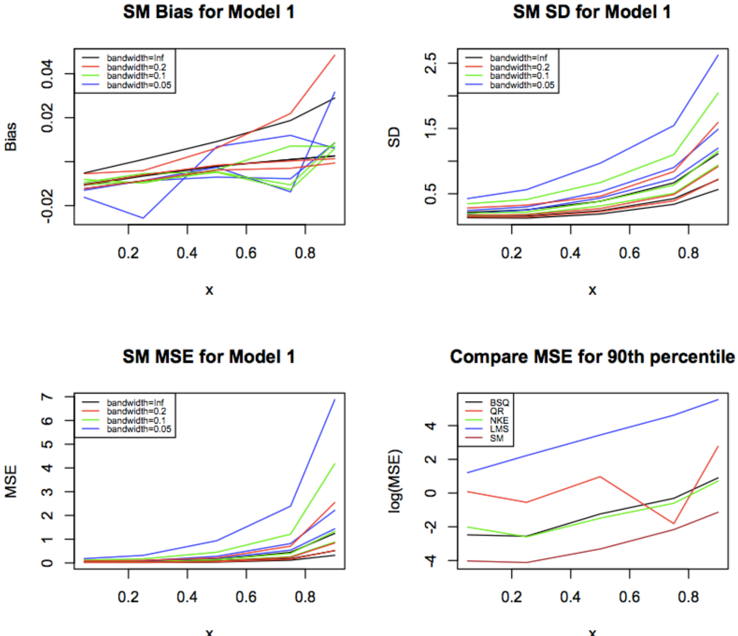
Assume univariate covariate X follows the standard uniform distribution.

- Methods: BSQ=Bin smoothed quantile, QR=Quantile regression, NKE=Nonparametric kernel smooth estimator, LMS=LMS parametric method, SM=Semiparametric method
- Model 1:

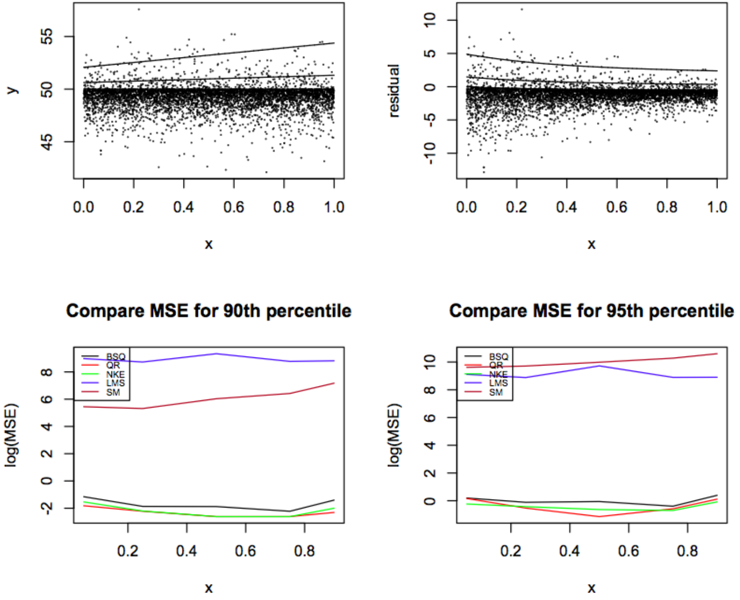
$$\mu(X) = 50 + 10X, \log(\sigma(X)) = 1 + 2X,$$

  and $\varepsilon(X)$ follows the standard normal distributions.
- Model 2: $(Y - 50)|X$ follows the mixture distribution $0.1 Exp(0.9 + X) + 0.9(-Exp(0.1 + X))$.
- All sample size is $n = 5000$ and bias, variance and MSE calculated from 1000 simulations.

Simulation Result: Model 1



Simulation Result: Model 2



Data Analysis

Can follow the procedures below

- Step 1: Fit the nonparametric quantile regression to see whether certain quantile $Y^\alpha(X)$ and/or distribution $F(z, X)$ change over X.
- Step 2: If step 1 no, use the nonparametric kernel smooth estimator.
- Step 3: If step 1 yes, fit a semiparametric quantile regression model and then check whether the residual distribution $F_0(z, X)$ change over X.
- Step 4: If step 3 no, fit a semiparametric quantile regression model assuming same baseline.
- Step 5: If step 3 yes, choose between the semiparametric quantile regression with kernel smooth and the nonparametric kernel smooth estimator.

Example

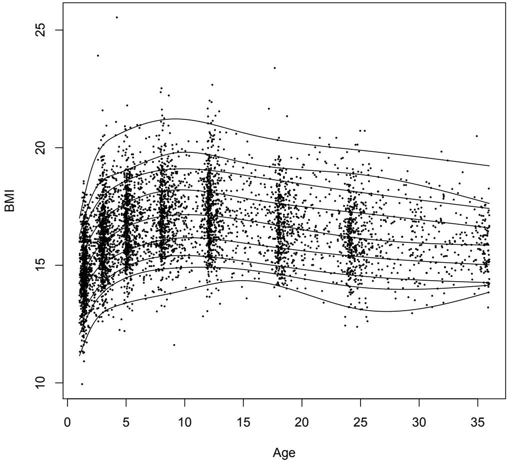


Figure: Fitted 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles from nonparametric quantile regression

Example

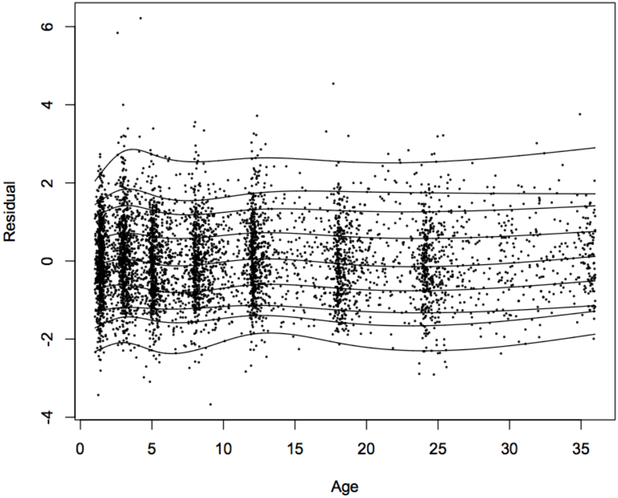


Figure: Fitted 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles from nonparametric quantile regression for residuals

Example

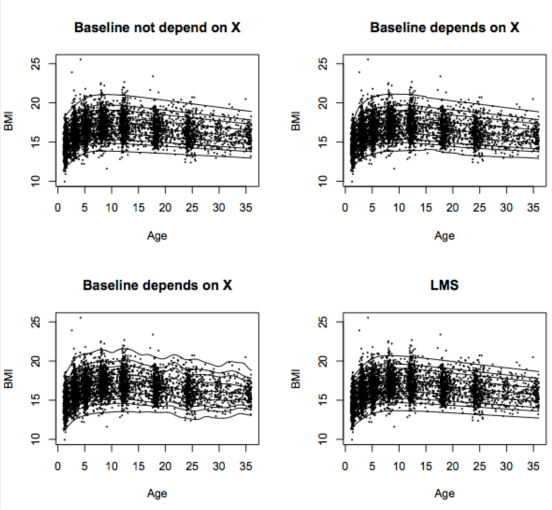


Figure: Fitted 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles from four methods