BIOST 572: Introduction Talk

Cheng Zheng

Biostatistics, University of Washington

April 12th, 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 year) to diagnose growth abnomalities.
- 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.

Application beyond growth curve

- Regression using standardized covariates (placement value).
- ROC regression

Data Display

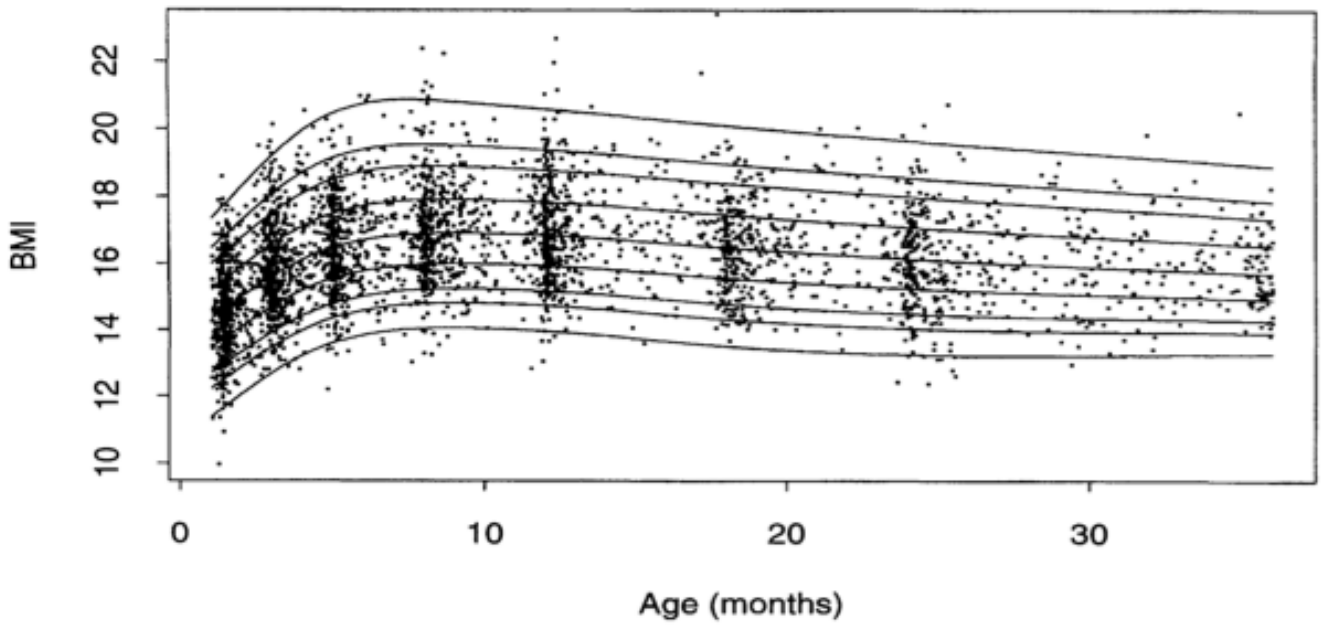

Figure: Estimated percentiles of BMI as a function of age (shown 1st, 5th, 10th, 25th,50th, 75th, 90th, 95th and 99th percentileds)

Quantiles Definitions

- For $\alpha \in [0, 1]$, the αth quantile of Y, Y^{α}, can be define as the value such that

$$P(Y \leq Y^{\alpha}) = \alpha.$$

- The covariate specific quantile of Y is the quantile for $Y|X = x$, denote as $Y^{\alpha}(x)$ such that

$$P(Y \leq Y^{\alpha}(x)|X = x) = \alpha.$$

- Rigorous definition: $Y^{\alpha}(x)$ satisfies

$$\inf_{Y^{\alpha}(x)} \{P(Y \leq Y^{\alpha}(x)|X = x)\} \geq \alpha,$$

Previous Solutions: Bin and Smooth Estimation

- Bin and Smooth Quantiles: Empirical quantiles for each narrow interval of X, then smoothed it. (Hamill et al., 1977)
- Bin and Smooth Parameters: Model $f(Y|X)$ by $\theta(X)$, estimate $\theta(X)$ for each narrow inteval of X, 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.
- Limitation: (1) Require large sample size; (2) Curse of dimensionality.

Previous Solutions: Parametric Models

- Idea: Specify a fully parametric model for $Y|X$ index by parameter θ. Estimate θ via likelihood.
- LMS (Cole and Green 1992): Assume Y can be transformed to standard normal random variable Z as follow:

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

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

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

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

Previous Solutions: Nonparametric Models

- Idea: Directly estimate $Y^{\alpha}(X)$ without assume certain distribution for Y.
- Koenker and Bassett (1978) propose 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.
- Modification: He propose location-scale model for Y

$$Y = \mu(X) + \sigma(X)\varepsilon,$$

  $\mu(X)$ is median and $\sigma(X)$ is median absolute deviation.

New method: Semiparametric Models

- Allow the shape depend on X and do not specify specific distribution for Y. Model $\mu(X)$ and $\sigma(X)$ parametricly 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 location parameter, $\sigma(X_i; \boldsymbol{\theta})$ is 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.

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

Model details

- We can use splines to model $\mu(X)$ and $\sigma(X)$. Let $\boldsymbol{\theta} = \{\beta_1, \cdots, \beta_p, \gamma_1, \cdots, \gamma_q\}$.

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

$R(X) = \{R_1(X), \cdots, R_p(X)\}$ and $S(X) = \{S_1(X), \cdots, S_q(X)\}$ are pre-specified regression spline basis functions.

Estimation: General Idea

- Estimate parameters in location and scale part with quasi likelilyhood (or GEE2 in general).

$$\sum_i W_1(X_i, \theta)(Y_i - \mu(X_i, \theta)) = 0$$

$$\sum_i W_2(X_i, \theta)\{(Y_i - \mu(X_i, \theta))^2 - \sigma^2(X_i, \theta)\} = 0$$

- Obtain the residual

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

then estimate the baseline function $F_0(z, X)$ by locally weighted kernel density estimation (Or use other smooth technique).

Next step

- We will show details about the estimation procedure.
- Under the assumption of semiparametric model, we willl run simulation to show
  (1) Parametric model (LMS method) will be biased when distribution assumption fail.
  (2) Semiparametric model did not lose much efficiency when parametric model is correct.
  (3) Nonparametric model is less efficient than Semiparametric model.
- We will re-analyze the US children growth data to obtain standardized BMI for age and gender, standardized weight for height, age and gender.