Quantile Analysis When Covariates Are Functions

Kehui Chen and Hans-Georg Müller

Department of Statistics
University of California, Davis

August 5, 2010

Growth Charts

- Growth charts are tables containing a list of quantiles, widely used to screen growth status. Example: height growth of children.
- Traditional growth charts are marginal quantiles computed at each age.
- Conditional growth charts based on longitudinal growth measurements. (Cole 1994, Royston 1995, Thompson and Fatti 1997, Wei and He 2006.)
- Longitudinal measurements often have variable time spacing and their timing may be subject-specific.
- Functional data analysis perspective: viewing the entire growth history as a functional but latent covariate.

Two Approaches For Conditional Quantiles Estimation

- Definition: $g_\alpha(X)$ is the conditional α-th quantile of Y if

$$P(Y \leq g_\alpha(x) \mid X = x) = \alpha.$$

- Approaches:
  1. Estimating Equation: minimize $E[l_\alpha(Y - g_\alpha(X))]$, for $l_\alpha(u) = |u| + (2\alpha - 1)u$ (Koenker and Bassett 1978). Extension of quantile linear regression to functional covariates by expanding X using B-splines (Cardot et al. 2005).
  2. Using Conditional CDF, $g_\alpha(X) = \inf\{y : F(y \mid X) \geq \alpha\}$.

Extending To Functional Covariates

- We aim at estimating conditional distribution function first and then compute conditional quantiles.
- The covariates we consider are random trajectories X, for which one records an associated response Y, a one-dimensional random variable.
- Observe: $F(y \mid X) = P(Y \leq y \mid X) = E(I(Y \leq y) \mid X)$, so that $F(y \mid X)$ can be viewed as the regression of the indicator $I(Y \leq y)$ on the functional predictor X.

Functional Generalized Linear Regression Model

- Observe: for a fixed y, the indicator $I(Y \leq y)$, given the covariate function X, has a binomial distribution with mean parameter $F(y \mid X)$.
- Functional model for conditional CDF:

$$F(y \mid X) = E(I(Y \leq y) \mid X) = g^{-1}\{\beta_0(y) + \int X^c(t)\beta(y,t)dt\},$$

where $X^c(t) = X(t) - \mu(t)$, and g is a monotone link function, for example, the logit link, with $g^{-1}(z) = \exp(z)/(1 + \exp(z))$.

Using Functional Principal Components as Predictors

- X is L^2 random process, with continuous covariance function. $EX(t) = \mu(t)$ and $Cov(X(s), X(t)) = \sum_k \lambda_k \phi_k(s)\phi_k(t)$. ϕ_k are orthogonal eigenfunctions.

- $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$ (Karhunen-Loève).

- $\xi_1,\ \xi_2, \ldots$ are FPC scores (uncorrelated, zero means and variances$=\lambda_k$).

- Expanding $\beta(y,t)$ in the eigenbasis,
  $\beta(y,t) = \sum_{k=1}^{\infty} \beta_k(y)\phi_k(t)$,
  the model becomes $F(y \mid X) = g^{-1}\left(\beta_0(y) + \sum_{k=1}^{\infty} \beta_k(y)\xi_k\right)$

Estimating The Model Components

- Truncated version of the model: we can show that $I(Y \leq y)$, conditional on $\xi_1, \ldots, \xi_p$, has Bernoulli distribution with parameter $\pi = g_{p,y}^{-1}(\mu(t) + \sum_{k=1}^{p} \xi_k \beta_k(y))$.
- The discrepancy between $g_{p,y}$ and g vanishes asymptotically.
- This motivates the consideration of an increasing series of truncated binomial models, with fixed link function g.
- Recover the latent functions from longitudinal observations (potentially irregular, sparse and noisy), and estimate the principal components. Implemented in PACE package. http://anson.ucdavis.edu/∼mueller/data/pace.html. (see Yao et al. 2005 or Müller and Stadtmüller 2005)

Simulation Scenarios

- X: sparsely (S) or densely (D) observed. Sparse: number of observations is 4 to 14 with equal probability, and locations where to take measurements are uniformly distributed on [0,10]. Dense: 30 equally spaced observations.
- Y conditional on X has Gaussian Mixture distribution.
- Compare (1) the proposed conditional functional quantiles, (2) functional quantile linear regression as proposed in Cardot et al. (2005), and (3) unconditional empirical quantiles.

Simulation Results

	$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.25$		$\alpha = 0.50$	
	S	D	S	D	S	D	S	D
Proposed	14.06	10.75	11.29	9.01	10.29	8.47	6.71	4.72
Cardot	45.36	19.07	42.44	15.38	38.66	10.37	31.10	1.67
Unconditional	142.2	143.8	95.13	94.13	60.25	60.59	51.03	51.63

Table: Average MSE over 200 simulation runs. The upper quantiles were found to behave similarly and are not reported.

Data Description

- Berkeley Growth Data, 54 female growth curves, from age 0 to 18.
- X: growth path from 0 to 12; construct the conditional CDF for Y: height at age 18 (adult height).
- Fit the logistic regression for a grid of y, obtain coefficient function $\beta(y, t)$.
- For a new subject i, given growth history X_i, construct the smooth conditional CDF for the adult height. Construct conditional quantiles for $\alpha \in (0, 1)$

Illustrative Example

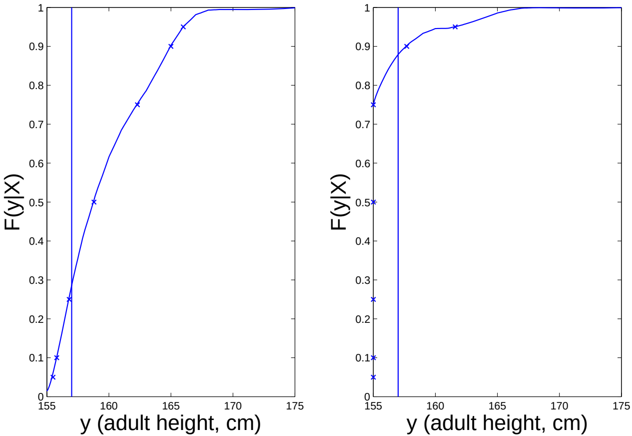

Figure: Estimated conditional CDFs for two girls. Their heights at age 12 are 142.1 cm (girl A) and 141 cm (girl B), respectively. Girl B has a large probability of about 0.9 to have adult height below the 10th quantile 157 cm, while this probability is predicted to be below 0.3 for girl A. The actual observed adult height is 164 cm for girl A and 154.5 cm for girl B.

Asymptotic Property

Theorem (Uniform Consistency)

Under regularity conditions, for a suitably chosen sequence $p(n) \to \infty$, as $n \to \infty$, on intervals I,

$$\sup_{y \in I} |\hat{F}(y \mid X) - F(y \mid X)| \xrightarrow{P} 0$$

where $\hat{F}(y \mid X) = g^{-1}(\sum_{k=0}^{p(n)} \hat{\beta}_k(y)\xi_k)$. Therefore, for any $0 < \alpha < 1$, the estimator $\hat{Q}(\alpha) = \inf\{y : \hat{F}(y \mid X) \geq \alpha\}$ of the αth conditional quantile of Y given X is a consistent estimator of $Q(\alpha) = \inf\{y : F(y \mid X) \geq \alpha\}$.

Thank You

- Thank you for your attention!