

# Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data

Qi Li<sup>1</sup>    Jeff Racine<sup>2</sup>

<sup>1</sup>Department of Economics  
Texas A&M University  
College Station, TX USA 77843-4228

<sup>2</sup>Department of Economics  
McMaster University  
Hamilton, ON Canada L8S 4M4

CESG 2005

## Synopsis

- ▶ We propose a nonparametric conditional CDF and quantile estimator
- ▶ The approach admits a mix of discrete and categorical data
- ▶ We adapt a data-driven (i.e., automatic) conditional PDF based bandwidth selector proposed by Hall, Racine, & Li (2004, JASA) that can automatically remove irrelevant variables and has impressive performance in this setting
- ▶ Theoretical underpinnings including rates of convergence and limiting distributions are provided
- ▶ Simulations demonstrate that this approach performs quite well relative to its peers
- ▶ Two illustrative examples serve to underscore its value in applied settings

## Motivation

- ▶ The estimation of regression functions is a popular activity
- ▶ Sometimes, however, the regression function is not representative of the impact of the covariates on the dependent variable
- ▶ For example, when the dependent variable is left (or right) censored, the relationship given by the regression function is distorted
- ▶ In such cases, regression quantiles above (or below) the censoring point are robust to the presence of censoring
- ▶ Furthermore, the quantile function always provides a more comprehensive picture of the conditional distribution of a dependent variable than the conditional mean function

## Background

- ▶ A natural way to model a conditional quantile function is to invert a conditional cumulative distribution function (CDF) at the desired quantile
- ▶ Nonparametric estimation of conditional CDFs has received much recent attention
- ▶ See, for example, Hall, Wolff, & Yao (1999, JASA) and Cai (2002, ET)
- ▶ This work considers the case of continuous conditioning variables only
- ▶ In applied settings, however, we frequently encounter a mix of discrete and continuous data
- ▶ One could adopt a frequency approach in these settings by splitting the data into subsets
- ▶ This approach would suffer from finite-sample efficiency losses due to the reduction in the sub-sample size

## Kernel Estimation of Mixed Data Conditional CDFs

- ▶ Let  $Y$  be a continuous random variable
- ▶ We propose estimating  $F(y|x)$  by

$$\hat{F}(y|x) = \frac{n^{-1} \sum_{i=1}^n G\left(\frac{y-Y_i}{h_0}\right) K_\gamma(X_i, x)}{\hat{\mu}(x)}$$

- ▶  $\hat{\mu}(x) = n^{-1} \sum_{i=1}^n K_\gamma(X_i, x)$  is the kernel estimator of  $\mu(x)$
- ▶  $K_\gamma(X_i, x) = W_h(X_i^c, x^c) L_\lambda(X_i^d, x^d)$  is a generalized product kernel, where  $W_h(X_i^c, x^c) = \prod_{s=1}^q h_s^{-1} w\left(\frac{X_{is} - x_s}{h_s}\right)$ , and  $L_\lambda(X_i^d, x^d) = \prod_{s=1}^r l(X_{is}^d, x_s^d, \lambda_s)$
- ▶  $G(v) = \int_{-\infty}^v w(u) du$  is the distribution function derived from the density function  $w(\cdot)$
- ▶  $h_0$  is the bandwidth associated with  $Y_i$

## Properties of $\hat{F}(y|x)$

- ▶ We demonstrate that

$$MSE[\hat{F}(y|x)] = \left[ \sum_{s=0}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x) \right]^2 + \frac{V(y|x) - h_0 \Omega(y|x)}{nh_1 \dots h_q} + o(\eta_{4n}) + o(\eta_{5n}),$$

- ▶ Also,

$$(nh_1 \dots h_q)^{1/2} \left[ \hat{F}(y|x) - F(y|x) - \sum_{s=1}^q h_s^2 B_{1s}(y|x) - \sum_{s=1}^r \lambda_s B_{2s}(y|x) \right] \rightarrow N(0, V(y|x))$$

## Quantile Function Estimation with Mixed Data

- ▶ A conditional  $\alpha$ th quantile of a conditional distribution function  $F(\cdot|x)$  is defined by ( $\alpha \in (0, 1)$ )

$$q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\} = F^{-1}(\alpha|x).$$

Or equivalently,  $F(q_\alpha(x)|x) = \alpha$ .

- ▶ We can estimate the conditional quantile function  $q_\alpha(x)$  by inverting the estimated conditional CDF function, i.e.,

$$\hat{q}_\alpha(x) = \inf\{y : \hat{F}(y|x) \geq \alpha\} \equiv \hat{F}^{-1}(\alpha|x).$$

## Properties of $\hat{q}_\alpha(x)$

- ▶ Define  $B_{n,\alpha}(x) = B_n(q_\alpha(x)|x)/f(q_\alpha(x)|x)$ , where

$$B_n(y|x) = \left[ \sum_{s=0}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x) \right]$$

is the leading bias term of  $\hat{F}(y|x)$  ( $y = q_\alpha(x)$ )

- ▶ We demonstrate that  $\hat{q}_\alpha(x) \rightarrow q_\alpha(x)$  in probability and that

$$(nh_1 \dots h_q)^{1/2} [\hat{q}_\alpha(x) - q_\alpha(x) - B_{n,\alpha}(x)] \rightarrow N(0, V_\alpha(x))$$

in distribution

## Empirical Application: Conditional Value at Risk

Table: Conditional Value at Risk for a long position in IBM stock

Model	5% CVaR	1% CVaR
Inhomogeneous Poisson, GARCH(1,1)	\$303,756	\$497,425
Conditional Normal, IGARCH(1,1)	\$302,500	\$426,500
AR(2)-GARCH(1,1)	\$287,700	\$409,738
Student- $t_5$ AR(2)-GARCH(1,1)	\$283,520	\$475,943
Extreme Value	\$166,641	\$304,969
LSCV CDF	\$258,727	\$417,192

Observe that, depending on one's choice of parametric model, one can obtain estimates that differ by as much 82% for 5% CVaR and 63% for 1% for this example