Draft: Quantile estimation with using auxiliary information

March 5, 2015

1 Motivation

Suppose we have a population, we can always observe the information X. But we can only observe the sample of Y from the population. Our target is to estimate the population quantile of Y.

2 Method

2.1 Direct Method

A very naive method is to use the sample of Y. Define

$$F_n(y) = \frac{\sum_{i \in A} d_i I\left(y_i \leq y\right)}{\sum_{i \in A} d_i}$$

. Then we can get the estimated quantile by

$$\hat{\theta}_d = \inf \left\{ y : F_n(y) \geq \tau \right\}$$

2.2 Regression estimator and difference estimate under pseudo empirical log-likelihood function

We put these two estimators together, because they are asymptotically same. The theorem 1 in the paper shows that.

$$\hat{\theta}_{df} = \hat{\theta}_d + N^{-1} \left\{ \sum_{i=1}^{N} q\left(\mathbf{x}_i; \hat{\beta}_{\tau_0}\right) - \sum_{i \in A} d_i q\left(\mathbf{x}_i; \hat{\beta}_{\tau_0}\right) \right\}$$

For this estimator, we can extend to multi-calibrations. We can use $\tau_0, \tau_1, ..., \tau_m$ to do the calibration together.

2.3 Using empirical distribution to do calibration

Define

$$F_N(x) = \frac{\sum_i^N I(x_i \le x)}{N}$$
$$F_w(x) = \sum_{i \in A} w_i I(x_i \le x)$$

We want to minimize

$$\arg\min_w Var\left(\hat{\theta}_w\right) \approxeq \left(\frac{1}{f(\hat{\theta}_\tau)}\right)^2 Var\left(\hat{F}_w(\theta)\right) \tag{1}$$

Given constraints:

$$F_N(x) = F_w(x)$$
$$\sum_{i \in A} w_i = 1$$

But unfortunately, this constraint is not always holds for any $x \in \mathbb{R}$. So I just use the moments to do the calibration. Because the theorem below:

Theorem 1 *(Frechet-Shohat Therom) Suppose $X_n, n \ge 0$ are random variables. If $\underline{\lim} EX_n^r = \beta_r$ for all r and if all β_r are the moments of a unique random variable X_0, then $X_n \longrightarrow X_0$ in distribution.*

So we can use the first four moments to do the calibration. But I don't get the minimized weights. I first to do the simulation to see if this works. Then we can minimize the w to see we can get the improvement.

2.4 Ratio estimator

For any quantile τ, we can get the true population quantile for X. So the ratio estimator is possible. Suppose the $\theta_{x,\tau}$ is the quantile of τ for population X.

The ratio estimator is

$$\hat{\theta}_{ratio} = \theta_{x,\tau} \frac{\hat{\theta}_{w,y}}{\hat{\theta}_{w,x}}$$

But in some cases if the $\hat{\theta}_{w,x}$ is 0 or closed to 0. That may cause problems. We may use the bias-corrected ratio estimator.

$$\hat{\theta}_{ratio} = \theta_{x,\tau} \frac{\hat{\theta}_{w,y}\hat{\theta}_{w,x} + \hat{C}\left(\hat{\theta}_{w,y}, \hat{\theta}_{w,x}\right)}{\hat{\theta}_{w,x}^2 + \hat{V}\left(\hat{\theta}_{w,x}\right)}$$

The problem is that can we use

$$\hat{C}\left(\hat{\theta}_{w,y}, \hat{\theta}_{w,x}\right) = ??$$

and

$$\hat{V}\left(\hat{\theta}_{w,x}\right) \approxeq \left(\frac{1}{f(\hat{\theta}_\tau)}\right)^2 \hat{V}ar\left(\hat{F}_w(\theta)\right)$$

3 A small simulation study

The simulation studies are conducted to compare the performance of different estimators. Two finite populations of size $N = 1000$ were generated from bivariate normal distribution respectively. The correlation between two variables in the first population is 0.9, then 0.6. For each simulation run , a simple random sample of size $n = 100$ was taken.

To compare with each other, we set the direct estimator as the base and define relative efficiency:

$$RE_* = \frac{MSE_*}{MSE_{Direct}}$$

.

Note: We use the same $\tau_0 = 0.5$ to get the difference estimator.