Quantile Estimation with Auxiliary Information: Simulation Results

Stima dei Quantili in presenza di Informazione Ausiliaria: Simulazioni

Lucio De Capitani and Leo Pasquazzi

Sommario In this work we compare finite population quantile estimators in a simulation study. We consider settings where complete auxiliary information is available, and quantile estimators that are obtained from inversion of several well-known estimators for the population cdf. The simulation results show that estimators based on separate estimates of the regression function and the error distributions are usually the most efficient ones.

Sommario *In questo lavoro confrontiamo stimatori per i quantili di una popolazione finita mediante simulazioni. Considereremo stimatori basati sull'inversione di stimatori per la funzione di ripartizione che sfruttano informazione ausiliaria completa. I risultati delle simulazioni mostrano che solitamente gli stimatori più efficienti sono quelli basati su stime separate della funzione di regressione e delle distribuzioni degli errori.*

Key words: superpopulation model, local-linear regression, model-based estimator, model-assisted estimator, model-calibrated estimator

1 Introduction

In the present work we compare the efficiency of quantile estimators based on inversion of several distribution function estimators in a simulation study. The distribution function estimators we consider are the same as those in Pasquazzi and De Capitani (2014). Hence we will use the same definitions and the same notation as

Lucio Decapitani
Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, e-mail: lucio.decapitani1@unimib.it

Leo Pasquazzi
Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, e-mail: leo.pasquazzi@unimib.it

in the latter work hereafter. Thus, the pair (y_i, x_i) denotes the values of a study variable Y and of an auxiliary variable X associated with unit i of a finite population $U = \{1, 2, \ldots, N\}$. While x_i is assumed to be known for all $i \in U$, y_i is known only for $i \in s$, where $s \subset U$ is a sample of units that has been selected according to some sampling design. The model-based estimators for $F_N(y) := \frac{1}{N}\sum_{i \in U} I(y_i \leq y)$ we considered for deriving quantile estimators will be denoted by $\widehat{F}_\bullet(y)$. They are all of the general form

$$\widehat{F}_\bullet(y) := \frac{1}{N}\left\{\sum_{i \in s} y_i + \sum_{i \notin s} \widehat{p}_{\bullet,i}\right\},$$

where the $\widehat{p}_{\bullet,i}$'s denote predictors for $P(y_i \leq y)$, $i \notin s$, under some superpopulation model. The predictors used in the simulations are those implicit in

- the Chambers and Dunstan (1986) estimator $\widehat{F}_{CD}(y)$:

$$\widehat{p}_{CD,i} := \frac{1}{n}\sum_{j \in s} I\left(y_j - x_j\widehat{\beta} \leq y - y_i x_i \widehat{\beta}\right)$$

  where n denotes the sample size, and $\widehat{\beta} := \sum_{k \in s} x_k y_k / \sum_{k \in s} x_k^2$;
- the Kuo (1988) estimator $\widehat{F}_K(y)$:

$$\widehat{p}_{K,i} := \sum_{j \in s} w_j(x_i) I(y_j \leq y),$$

  where $w_j(x)$ is defined as the jth component of $\mathbf{w} := \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}_x$. In the definition of $\mathbf{w}$, $\mathbf{e}_1^T := (1,0)$, $\mathbf{X}_x := [1, (x_i - x)]_{i \in s}$, and $\mathbf{W}_x := diag[K((x_i - x)/\lambda)]_{i \in s}$ with $\lambda > 0$ and with $K(u) := \frac{3}{4}(1-u^2) I(|u| \leq 1)$ for $u \in \mathbb{R}$;
- the Chambers, Dunstan and Wehrly (1993) estimator $\widehat{F}_{CDW}(y)$:

$$\widehat{p}_{CDW,i} := \widehat{p}_{CD,i} + \sum_{j \in s} w_j(x_i)\left[\widehat{p}_{K,j} - \widehat{p}_{CD,j}\right]$$

- the nonlinear version $\widehat{F}_{DH}(y)$ of the Chambers and Dunstan estimator proposed by Dorfman and Hall (1993):

$$\widehat{p}_{DH,i} := \frac{1}{n}\sum_{j \in s} I\left(y_j - \widehat{m}(x_j) \leq y - \widehat{m}(x_i)\right)$$

  where $\widehat{m}(x) := \sum_{k \in s} w_k(x) y_k$;
- the Double-Regression estimator $\widehat{F}_{DR}(y)$ proposed by Pasquazzi and De Capitani (2014):

$$\widehat{p}_{DR,i} := \sum_{j \in s} w_j(x_i) I\left(y_j - \widehat{m}(x_j) \leq y - \widehat{m}(x_i)\right).$$

In addition to the above model-based estimators, we also tested the performance of quantile estimators derived from model-assisted estimators $\widetilde{F}_\bullet(y)$ for $F_N(y)$. Except for the model-calibrated estimator $\widetilde{F}_{MC}(y)$ proposed by Rueda et al. (2010), they are

all generalized difference estimators of the form

$$\widetilde{F}_{\bullet}(y) := \frac{1}{N}\left\{\sum_{i\in s} d_i I(y_i \le y) + \sum_{i\in U} \widetilde{p}_{\bullet,i} - \sum_{i\in s} d_i \widetilde{p}_{\bullet,i}\right\},$$

where the d_i's denote the inverse sample inclusion probabilities, and the $\widetilde{p}_{\bullet,i}$'s are design-weighted versions of the predictors $\widehat{p}_{\bullet,i}$. As for the model-calibrated estimator $\widetilde{F}_{MC}(y)$ considered the simulations, it is defined by $\widetilde{F}_{MC}(y) := \sum_{i\in s} \omega_i I(y_i \le y)$, where the ω_i's minimize the chi-squared distance $\Phi_s := \sum_{i\in s}(\omega_i - d_i)^2/d_i$ under the constraints $\sum_{i\in s} \omega_i I(\widetilde{m}(x_i) \le y_r^*) = \sum_{i\in U} I(\widetilde{m}(x_i) \le y_r^*)$, $r = 1,2,3,4$. In the constraints, $\widetilde{m}(x)$ denotes the design-weighted version of $\widehat{m}(x)$, and y_r^*, $r = 1,2,3$, are the three quartiles of the $\widetilde{m}(x_i)$'s, $i \in U$, while $y_4^* := \max_{i\in U} \widetilde{m}(x_i)$.

Since not all estimators for $F_N(y)$ included in the simulation study do provide nondecreasing estimates $F(\cdot)$ with probability 1, we used the following rule for computing the inverses $F^{-1}(t)$ at the levels $t = 0.1, 0.25, 0.50, 0.75, 0.90$:

$$F^{-1}(t) := \frac{1}{2}\left(\inf\{y : F(y) \ge t\} + \sup\{y : F(y) \le t\}\right).$$

2 Simulation Study

To compare the efficiency of the quantile estimators we simulated $B = 1000$ samples of size $n = 100$ by simple random without replacement sampling. The samples were selected from eight finite populations of size $N = 1000$ generated from the following superpopulation models:

$$M1 \qquad y_i := x_i + \sigma\varepsilon_i$$
$$M2 \qquad y_i := \sqrt{x_i} + \sigma\varepsilon_i$$

The errors ε_i were generated independently from either the Student t distribution with $\nu = 5$ df (identically distributed error components), or from shifted noncentral Student t distributions with $\nu = 5$ dgf and with noncentrality parameter $\zeta = 15x_i$ (not identically distributed error components). The shifts applied to the error distributions in the latter case are aimed to make sure that their expectations equal zero. Note that in the case of not identically distributed error components the auxiliary variable influences not only the scale but also the shape of the error distributions. As for σ, we considered two values: $\sigma = 0.1$ (strong regression relationship), and $\sigma = 0.3$ (weak regression relationship). The x-values of the auxiliary variable were independently generated from the uniform distribution on $(0,1)$. The populations we considered are thus $2(\text{regression functions}) \times 2(\text{types of error components}) \times 2(\text{values of } \sigma) = 8$.

The simulation results are summarized in Tables 1 and 2, where the ratios between the simulated MSE referring to each estimator and the one referring to the

corresponding inverse of the Horvitz-Thompson estimator $\widehat{F}_{HT}(y)$ are reported. In the notation identifying estimators which involve local-linear regression, we added a subscript s or l according to whether the bandwidth λ was set to $\lambda_s := 0.1$ or to $\lambda_l := 0.3$. In the double-regression estimators we added two subscripts, because we also tested two different bandwidths for estimating the regression function (first subscript) and the error distributions (second subscript).

The simulation results confirm that the estimators based on auxiliary information do often provide a large gain in efficiency with respect to the inverse of $\widehat{F}_{HT}(y)$. In the case of identically distributed errors the gain in efficiency seems more pronounced when $\sigma = 0.1$ than when $\sigma = 0.3$, while in the case of not identically distributed error components evidence in favor of this statement is not as clear. Beyond this, it is worth noting that, unless the finite population does not follow the underlying super-population model (as may be the case with $\widehat{F}_{CD}(y)$, $\widehat{F}_{DH}(y)$ and their model-assisted counterparts), the estimators based on a separate estimate of the regression function appear to be more efficient than the quantile estimators derived from $\widehat{F}_{K}(y)$ and $\widehat{F}_{CDW}(y)$. In this respect, we emphasize the good overall performance of the quantile estimators derived from $\widehat{F}_{DR}(y)$. With the proper bandwidth combination they loose little efficiency with respect to those derived from $\widehat{F}_{CD}(y)$ and $\widehat{F}_{DH}(y)$ in the populations with identically distributed errors, and they are usually the most efficient estimators in the populations with not identically distributed errors.

Riferimenti bibliografici

1. Chambers, R.L., Dorfman, A.H., Wehrly, T.: Bias robust estimation in finite populations using nonparametric calibration. J. Amer. Statist. Assoc. 88, 268–277 (1993)
2. Chambers, R.L., Dunstan, R.: Estimating distribution functions from survey data. Biometrika 73, 597–604 (1986)
3. Dorfman, A.H., Hall, P.: Estimators of the finite population distribution function using nonparametric regression. The Annals of Statistics 21, 1452–1475 (1993)
4. Kuo, L.: Classical and prediction approaches to estimating distribution functions from survey data. In: Proceedings of the section on survey research methods, pp. 280-285, Amer. Statist. Assoc., Alexandria, VA (1988)
5. Pasquazzi, L., De Capitani, L.: A new estimator for a finite population distribution function in the presence of complete auxiliary information. Università degli Studi di Milano-Bicocca, Dip. di Statistica e Metodi Quantitativi, Working paper (2014)
http://boa.unimib.it/bitstream/10281/47528/1/Pasquazzi_Decapitani.pdf
6. Rueda, M., Sànchez-Borrego, I., Arcos, A., Martinez, S.: Model-calibration estimation of the distribution function using nonparametric regression. Metrika 71, 33–44 (2010)

Tabella 1 Simulated relative MSE for the populations generated from models with identically distributed error components

Model	M1										M2									
	$\sigma = 0.1$					$\sigma = 0.3$					$\sigma = 0.1$					$\sigma = 0.3$				
level	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
Estimator																				
$\widehat{F}_{HT}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\widehat{F}_{CD}$	0.624	0.371	0.248	0.333	0.740	0.833	0.551	0.999	0.920	0.797	0.476	0.427	0.565	0.294	0.967	0.641	0.444	0.687	0.785	0.837
$\widehat{F}_{DHs}$	0.763	0.407	0.337	0.384	0.802	0.849	0.590	1.029	0.930	0.739	0.219	0.246	0.252	0.316	0.783	0.668	0.469	0.770	0.784	0.823
$\widehat{F}_{DHl}$	0.761	0.385	0.305	0.358	0.795	0.874	0.576	1.007	0.942	0.770	0.224	0.200	0.220	0.273	0.690	0.665	0.456	0.729	0.778	0.829
$\widehat{F}_{Ks}$	0.984	0.775	0.651	0.703	0.873	1.076	0.991	0.997	1.138	1.016	0.688	0.458	0.419	0.652	0.832	1.138	0.901	0.939	1.020	0.990
$\widehat{F}_{Kl}$	1.202	0.861	0.653	0.672	0.781	1.146	1.067	1.015	1.081	1.017	1.362	0.722	0.433	0.683	0.749	1.389	1.050	0.937	0.962	0.927
$\widehat{F}_{CDWs}$	0.992	0.732	0.629	0.686	0.863	1.078	0.932	1.028	1.086	1.063	0.542	0.429	0.419	0.632	0.839	1.020	0.838	0.940	0.974	1.013
$\widehat{F}_{CDWl}$	1.032	0.698	0.603	0.671	0.803	1.105	0.968	1.080	1.049	1.103	0.687	0.409	0.391	0.613	0.888	1.117	0.900	0.936	0.943	0.957
$\widehat{F}_{DRs}$	0.825	0.587	0.479	0.558	0.777	0.953	0.794	1.055	1.024	0.931	0.319	0.329	0.313	0.506	0.749	0.800	0.743	0.911	0.896	0.956
$\widehat{F}_{DRl}$	0.802	0.471	0.424	0.450	0.701	0.940	0.790	1.074	0.958	0.930	0.264	0.255	0.267	0.389	0.675	0.803	0.669	0.916	0.814	0.867
$\widehat{F}_{DRsl}$	0.798	0.500	0.442	0.469	0.722	0.900	0.762	1.127	0.963	0.870	0.273	0.281	0.284	0.422	0.707	0.762	0.662	0.945	0.818	0.878
$\widehat{F}_{DRls}$	0.841	0.576	0.463	0.553	0.765	1.008	0.843	1.013	1.037	0.977	0.317	0.317	0.305	0.498	0.737	0.843	0.754	0.896	0.911	0.939
$\widetilde{F}_{CD}$	0.945	0.900	0.870	0.922	0.927	0.955	0.902	0.981	0.979	0.994	0.890	0.897	0.856	0.913	0.910	0.877	0.915	0.958	0.960	0.965
$\widetilde{F}_{DHs}$	0.886	0.726	0.616	0.685	0.877	0.998	0.923	1.034	0.994	1.059	0.480	0.441	0.407	0.625	0.828	0.898	0.853	0.901	0.926	1.006
$\widetilde{F}_{DHl}$	0.863	0.712	0.602	0.666	0.861	1.001	0.904	0.997	0.983	1.037	0.476	0.431	0.401	0.606	0.806	0.877	0.838	0.883	0.912	1.004
$\widetilde{F}_{Ks}$	0.944	0.732	0.641	0.734	0.886	1.056	0.997	0.973	1.092	1.010	0.502	0.435	0.420	0.651	0.858	1.017	0.881	0.908	1.011	0.980
$\widetilde{F}_{Kl}$	0.877	0.722	0.635	0.721	0.866	1.043	0.979	0.949	1.052	1.006	0.521	0.438	0.413	0.638	0.847	0.981	0.856	0.868	0.972	0.977
$\widetilde{F}_{CDWs}$	0.965	0.724	0.630	0.695	0.867	1.065	0.948	1.014	1.050	1.091	0.526	0.441	0.421	0.636	0.821	0.947	0.854	0.928	0.949	1.034
$\widetilde{F}_{CDWl}$	0.889	0.718	0.618	0.683	0.844	1.039	0.923	0.994	1.009	1.110	0.543	0.440	0.413	0.626	0.804	0.908	0.832	0.887	0.917	1.044
$\widetilde{F}_{DRs}$	0.929	0.723	0.609	0.685	0.871	1.034	0.941	1.029	1.039	1.052	0.469	0.445	0.414	0.630	0.828	0.905	0.858	0.915	0.955	1.047
$\widetilde{F}_{DRl}$	0.875	0.711	0.607	0.669	0.860	1.020	0.918	1.010	1.005	1.086	0.471	0.432	0.406	0.607	0.815	0.878	0.835	0.886	0.923	1.018
$\widetilde{F}_{DRsl}$	0.887	0.718	0.615	0.672	0.859	1.020	0.921	1.023	1.011	1.080	0.471	0.441	0.409	0.615	0.817	0.880	0.841	0.891	0.924	1.029
$\widetilde{F}_{DRls}$	0.918	0.720	0.600	0.673	0.871	1.054	0.931	1.015	1.038	1.069	0.473	0.439	0.407	0.619	0.830	0.898	0.849	0.910	0.951	1.026
$\widetilde{F}_{MCs}$	0.889	0.708	0.652	0.767	0.914	1.003	1.012	0.991	1.072	1.031	0.730	0.440	0.446	0.690	0.907	0.991	0.887	0.908	1.013	0.997
$\widetilde{F}_{MCl}$	0.887	0.698	0.642	0.750	0.908	1.004	0.976	0.960	1.065	0.994	0.731	0.438	0.443	0.677	0.900	0.953	0.871	0.880	1.017	0.984

Tabella 2 Simulated relative MSE for the populations generated from models with not identically distributed error components

Model	M1										M2									
	$\sigma = 0.1$					$\sigma = 0.3$					$\sigma = 0.1$					$\sigma = 0.3$				
level	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
Estimator																				
$\widehat{F}_{HT}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\widehat{F}_{CD}$	8.845	1.373	6.021	1.991	0.890	1.327	1.069	2.228	1.381	0.859	5.385	0.707	3.692	2.135	0.855	2.257	1.225	2.638	1.410	0.823
$\widehat{F}_{DHs}$	9.366	1.379	6.478	2.306	0.877	1.861	1.051	2.282	1.585	0.742	6.716	0.903	4.868	2.255	0.967	3.084	1.241	2.864	1.650	0.736
$\widehat{F}_{DHl}$	8.437	1.388	6.230	1.953	0.928	1.488	1.097	2.173	1.344	0.823	6.284	0.891	4.758	1.995	1.021	2.568	1.273	2.737	1.395	0.810
$\widehat{F}_{Ks}$	1.242	1.093	0.972	0.991	1.077	1.155	0.996	1.005	0.991	1.030	1.089	0.909	0.892	0.899	1.065	1.168	1.084	1.102	1.019	1.018
$\widehat{F}_{Kl}$	1.299	1.097	0.935	0.978	1.102	1.270	1.025	0.988	0.981	1.057	1.714	1.156	0.917	0.918	1.090	1.279	1.109	1.118	1.021	1.051
$\widehat{F}_{CDWs}$	1.090	1.094	1.046	1.000	1.022	1.005	0.951	1.003	1.049	1.035	0.852	0.860	0.870	0.915	1.024	1.040	1.038	1.027	1.051	1.023
$\widehat{F}_{CDWl}$	1.119	1.123	1.014	0.969	1.081	1.084	0.988	0.998	1.003	1.055	1.082	0.922	0.862	0.879	1.094	1.110	1.058	1.030	1.019	1.058
$\widehat{F}_{DRs}$	0.912	1.099	1.152	1.006	0.946	0.823	0.834	1.014	0.968	0.951	0.545	0.651	0.699	0.928	0.934	0.834	0.923	0.962	0.956	0.944
$\widehat{F}_{DRl}$	1.031	1.168	1.214	0.942	0.976	0.881	0.889	1.041	0.937	0.983	0.766	0.628	0.680	0.879	0.969	0.880	0.979	0.965	0.906	0.974
$\widehat{F}_{DRsl}$	1.085	1.222	1.403	1.003	0.904	0.853	0.886	1.084	0.948	0.900	0.773	0.656	0.776	0.949	0.894	0.863	1.008	0.969	0.939	0.889
$\widehat{F}_{DRls}$	0.908	1.078	1.059	0.978	0.984	0.891	0.870	1.005	1.000	1.002	0.575	0.618	0.682	0.905	0.971	0.890	0.954	0.996	0.972	0.994
$\widetilde{F}_{CD}$	0.934	1.009	1.060	0.980	0.988	0.880	0.946	1.009	0.995	0.999	0.839	0.937	0.930	0.952	0.981	0.924	0.982	0.947	1.001	0.997
$\widetilde{F}_{DHs}$	1.083	1.112	1.063	1.056	0.978	0.941	1.037	1.062	1.062	1.012	0.748	0.855	0.824	0.955	0.975	0.999	1.094	0.989	1.056	1.004
$\widetilde{F}_{DHl}$	1.094	1.115	1.040	1.007	0.991	0.926	1.012	1.034	1.029	1.016	0.759	0.841	0.823	0.922	0.992	0.984	1.081	0.971	1.029	1.004
$\widetilde{F}_{Ks}$	1.082	1.066	0.965	0.983	1.039	1.013	0.945	1.037	0.995	0.995	0.800	0.824	0.844	0.895	1.026	1.033	1.009	1.002	0.996	0.986
$\widetilde{F}_{Kl}$	1.076	1.027	0.946	0.945	1.034	0.994	0.932	1.025	0.979	0.991	0.791	0.814	0.821	0.865	1.023	1.022	0.989	0.995	0.984	0.988
$\widetilde{F}_{CDWs}$	1.006	1.094	1.053	1.035	1.003	0.912	0.903	1.042	1.061	1.020	0.730	0.836	0.842	0.940	1.005	0.952	0.993	0.955	1.060	1.022
$\widetilde{F}_{CDWl}$	1.002	1.048	1.039	0.987	1.000	0.909	0.886	1.030	1.062	1.019	0.718	0.820	0.812	0.902	1.001	0.949	0.973	0.948	1.052	1.023
$\widetilde{F}_{DRs}$	0.988	1.080	1.037	1.067	0.992	0.903	0.917	1.034	1.047	1.013	0.676	0.826	0.824	0.971	0.993	0.947	1.002	0.958	1.039	1.018
$\widetilde{F}_{DRl}$	0.972	1.051	1.014	1.001	0.994	0.892	0.883	1.003	1.042	1.013	0.676	0.814	0.809	0.911	0.993	0.914	0.978	0.948	1.026	1.018
$\widetilde{F}_{DRsl}$	1.051	1.021	0.948	0.988	1.064	1.033	1.037	1.068	1.018	1.039	0.938	0.881	0.836	0.915	1.044	1.007	1.016	1.035	1.016	1.012
$\widetilde{F}_{DRls}$	1.045	1.020	0.944	0.951	1.052	1.022	1.021	1.036	1.028	1.039	0.931	0.871	0.841	0.895	1.033	1.023	1.002	1.011	1.023	1.027
$\widetilde{F}_{MCs}$	0.987	1.072	1.042	1.013	0.991	0.882	0.916	1.025	1.040	1.009	0.669	0.824	0.811	0.930	0.996	0.915	0.999	0.951	1.036	1.013
$\widetilde{F}_{MCl}$	0.972	1.074	1.036	1.052	0.992	0.905	0.899	1.024	1.067	1.015	0.670	0.816	0.820	0.939	0.996	0.943	0.982	0.958	1.050	1.019