Statistical Papers (2026) 67:52
https://doi.org/10.1007/s00362-026-01802-5

REGULAR ARTICLE

Quantile Share Ratio Regression for the Study of Economic Inequality

Alessio Farcomeni[1]

Received: 24 July 2025 / Revised: 7 November 2025

Abstract
We introduce a distribution-free approach to quantile share ratio regression. Our proposal involves the specification of a generalised linear model for the ratio of tail areas above and below two pre-specified quantiles. The latter ratio is the quantile share ratio, a measure of primary interest in the study of income inequality. We derive inference through an efficient two-step approach for parameter estimation that entails estimation of the conditional cumulative distribution function at the first step. A scalable strategy is discussed for large sample sizes. We are motivated by the study of income inequality in the European Union. Using data from a sample of approximately 2.8 million households across twenty-three countries and fifteen years (2007-2021) we make formal claims on the significance of adjusted and unadjusted differences among countries, and time trends. Interestingly enough, we find independent negative associations of economic inequality with gender equality and control of corruption.

Keyword Conditional quantiles, Distribution-free methods, Income inequality, Palma ratio.

1 Introduction

There are several methods to measure income inequality (Davidson and Flachaire 2007; Cowell 2011). Alongside the classical ones, such as the Lorenz curve, quantile ratios and quantile share ratios have emerged for their ease of interpretation (Palma and Stiglitz 2016; Chancel and Piketty 2021; Chancel et al. 2022).

In this work, we focus on quantile share ratios (Beach and Davidson 1983). These involve computing the ratio of the total income above a τ_1-quantile and the total

The author acknowledges two anonymous referees for constructive and useful remarks. This work was partially supported by the PRIN 2022 research project LANNKC (CUP E53D23005810006), funded by NextGenerationEU programme of the European Union and by the Italian Ministry of University and Research.

✉ Alessio Farcomeni
alessio.farcomeni@uniroma2.it

[1] University of Rome "Tor Vergata", Rome, Italy

Published online: 21 April 2026

income below a τ_2-quantile, with $\tau_2 \leq \tau_1$. The most commonly used quantile share ratios are the quintile share ratio ($\tau_1 = 0.8$, $\tau_2 = 0.2$), and the Palma (2011) ratio ($\tau_1 = 0.9$, $\tau_2 = 0.4$). The combination $\tau_1 = 0.9$, $\tau_2 = 0.5$ is also quite common in the economic literature. Quantile share ratios answer the question: "By what factor does the total wealth of the top τ_1 percent of the population exceed that of the bottom τ_2 percent?". Although their main application field is certainly the study of economic inequality, quantile share ratios are also of interest in other areas, such as epidemiology (see, for example, Harper and Lynch (2007)), and location analysis (Eiselt and Marianov 2011). The (marginal) quantile share ratio has previously been discussed in the statistical literature. Langel and Tillé (2011) describe linearization techniques to approximate the quintile share ratio estimator and its variance. In Hulliger and Schoch (2014), a distribution-free robust estimator is developed. Kpanzou (2015) derive the influence function of a quintile share ratio estimator, and Kebe et al. (2023) a novel semiparametric kernel estimator. All these works pertain marginal quantile share ratios, but conditional quantile share ratios might be more of interest in empirical applications. Clearly, it is of interest to compare different populations, evaluate the effects of covariates, and even predict the quantile share ratio (that is, economic inequality) at the configuration of certain covariates; which are all tasks that would entail the definition of conditional quantile share ratios. Notably, conditional measures of income inequality are yet underexplored in the literature, with a few exceptions which we mention below.

Quantile share ratios are currently mostly used as descriptive statistics in the empirical economics literature. The common approach to evaluate association with covariates involves computing marginal quantile shares at an aggregate level (e.g., at county level) and then regress those on (possibly aggregated) covariates. This practice is clearly sub-optimal, and incurs strong risks of ecological bias. We will provide an empirical assessment in our simulation study.

In this work, we propose a distribution-free inferential method to make inference on *conditional* quantile share ratios, and therefore to associate them with predictors in a simple regressive framework. To the best of our knowledge, we will present the first formal inferential method apt at doing so; albeit our approach is related to the general frameworks proposed in Chernozhukov et al. (2013) and Firpo et al. (2018). Formal methods for Lorenz regression are proposed in Aaberge et al. (2005) and Heuchenne and Jacquemain (2022); and a formal method for quantile ratio regression is proposed in Farcomeni and Geraci (2024).

Our motivating example involves the study of income inequality in Europe in the very recent past. We base that on a data set of more than 2.8 million households, across twenty-three European countries, in the period 2007 to 2021.

The rest of the paper is as follows. In the next section, we present our Quantile Share Ratio Regression (QSRR) approach. A two-step inferential approach, together with a simple modification that can scale well to millions of observations, is presented in Section 2.1. In Section 2.2 we show a closed-form estimator for the standard errors of the parameters involved. In Section 2.3, we show the asymptotic normality of our estimates. In Sections 3 and 4 we show a brief simulation study, and the application to income inequality in Europe. Some concluding remarks are given in Section 5.

An R implementation of our proposed method can be found at https://github.com/afarcome/qsrr.

2 Quantile share ratio regression

Let $Y > 0$ denote an absolutely continuous random variable, of which an i.i.d. sample $y_1, \ldots, y_n$ is observed. Let also x denote a vector of p covariates. We let $F(y|x) = \Pr(Y \leq y|x)$ denote the conditional C.D.F. of Y, and $q(\tau, x) = \min_y\{y : F(y|x) \geq \tau\}$ the conditional quantile function. We define the conditional quantile share ratio $QSR(x)$ as

$$QSR(x) = \frac{\int_{q(\tau_1,x)}^{\infty} y\, dF(y|x)}{\int_0^{q(\tau_2,x)} y\, dF(y|x)} = \frac{\int_{\tau_1}^{1} q(\tau, x)\, d\tau}{\int_0^{\tau_2} q(\tau, x)\, d\tau} \tag{1}$$

for some $0 < \tau_2 \leq \tau_1 < 1$. It should be noted that, while in most applications $QSR(x) > 1$, it is also possible that $QSR(x) \leq 1$. This can happen for instance if both τ_2 and τ_1 are very large, so that the left tail might include a very large share of the population and the right tail only a tiny part of the rich population. In all cases, $QSR(x) > 0$. We wish to specify a linear model of the kind

$$QSR(x) = g^{-1}(x'\beta) \tag{2}$$

where $g(\cdot)$ is known link function, β is a vector of regression coefficients, and x possibly includes a leading constant to accommodate an intercept. Any link function mapping $\mathcal{R}^+$ to $\mathcal{R}$ is mathematically appropriate, as it will guarantee that the predictions are positive. The most natural choice is clearly $g(\cdot) = \log(\cdot)$, but we will compare a few alternatives in our real data analysis below. The logarithmic link function is also advantageous as it allows us to approximately interpret $\exp(\hat{\beta})$ as a fold change in the quantile share ratios. Indeed, consider a logarithmic link function, a single binary covariate, and a model of the kind

$$QSR(x) = \exp(\beta_0 + x\beta_1).$$

Suppose that $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 0.5$. Clearly, $\widehat{Q}SR(0) = e$, $\widehat{Q}SR(1) = \exp(1.5)$, and $QSR(1)/QSR(0) = \exp(0.5)$. When more than one covariate is involved, a similar reasoning approximately holds (see also Firpo et al. (2009) for a more detailed related discussion).

The population conditional quantile share ratio (1) is a functional of the conditional CDF, in connection to the results in Chernozhukov et al. (2013). It should be noted that we have made no parametric assumptions on $F(y|x)$, and therefore the model specification is distribution-free.

2.1 Inference

Let the covariates be stacked in an n by p design matrix X. Let also $I(\cdot)$ denote the indicator function, and $y_{(j)}$ the j-th ordered value of $y_1, \ldots, y_n$.

Given the definition of $QSR(x)$ in (1), it is difficult to obtain a direct estimator for the parameters involved in (2). We propose to proceed in two steps: first, we set up a direct non-parametric estimator for (1), based on a non-parametric estimator of the conditional CDF. Secondly, we minimise the residuals based on differences between the pseudo-outcomes estimated in the first step and the linear predictor $g^{-1}(x'\beta)$.

There are different methods that can be suggested for the estimation of $F(y|x)$ (Hall et al. 1999). One possibility is to implement kernel estimators (Li and Racine 2008; Li et al. 2013) of the kind:

$$\hat{F}(y|x) = \frac{\sum_{i=1}^{n} I(Y_i \leq y) K_\lambda(X_i, x)}{n \hat{f}_X(x)},$$

where $K_\lambda(\cdot, \cdot)$ denotes a (product) kernel estimator with bandwidth λ, and $\hat{f}_X(x)$ a kernel density estimator for the marginal density of X. Another possibility involves a method originally proposed by Peracchi (2002), and whose theoretical properties were established in Chernozhukov et al. (2013). This approach entails, for $j = 1, \ldots, n$, obtaining predictions from a binary regression model of the kind

$$F(y_{(j)}|x) = \Pr(Y \leq y_{(j)}|x) = h(x'\gamma_j), \tag{3}$$

where the link function $h(\cdot)$ (e.g., the inverse of the logit transformation) appropriately maps real numbers to the unit interval, and γ_j is a vector of parameters. Note that up to n separate generalised linear models might be fit, an operation which can be easily parallelised; resulting in a less computationally cumbersome approach than kernel-based methods. In our implementation we will use (3), which in our experience is more convenient than kernel approaches. In addition to being much less computationally cumbersome, it is more resistant to the curse of dimensionality as the number of covariates increases, and it does not require the selection of tuning parameters (like a bandwidth).

A generalised inverse of the non-parametric estimator of the conditional quantile function yields a non-parametric estimate of the conditional quantile function $q(\tau, x)$. One can finally estimate (1) by fixing

$$\widehat{QSR}(x_i) = \frac{\sum_{j=1}^{n} I(y_{(j)} \geq \hat{q}(\tau_1, x_i)) y_{(j)} (\hat{F}(y_{(j)}|x_i) - \hat{F}(y_{(j-1)}|x_i))}{\sum_{j=1}^{n} I(y_{(j)} \leq \hat{q}(\tau_2, x_i)) y_{(j)} (\hat{F}(y_{(j)}|x_i) - \hat{F}(y_{(j-1)}|x_i))} \tag{4}$$

for $i = 1, \ldots, n$; where $y_{(0)} = 0$ and $F(y_{(0)}|x_i) = 0$ by definition. The expression (4) provides an initial nonparametric estimate of the conditional quantile share ratio of interest.

At the second step, one shall simply treat $\widehat{QSR}(x_i)$ as observed pseudo-outcomes, and minimise the residuals

$$\min_{\beta} \sum_{i} \left(\widehat{QSR}(x_i) - g^{-1}(x_i'\beta) \right)^2. \tag{5}$$

The use of the squared norm is particularly convenient, as a closed form solution is available for (5): define $u = (u_1, \ldots, u_n)$, with $u_i = g(\widehat{QSR}(x_i))$, then $\hat{\beta} = (X'X)^{-1}X'u$.

The proposed method scales well up to about one hundred thousand units, depending on the number of predictors. For higher values of n, the estimation of $\hat{F}(y|x)$ might become too cumbersome. It should also be noted that storage might be an issue, as $\hat{F}(y|x)$ in general results in a n by n matrix (one row for each unique value of X, one column for each unique value of y). When the sample size is in the order of the millions, a scalable inferential strategy is simply obtained by estimating $\hat{F}(y|x)$ only at v landmark points $\tilde{y}_{(j)}$, $j = 1, \ldots, v$, with $v << n$ (e.g., $v = 1000$) values between $y_{(1)}$ and $y_{(n)}$; thus reducing the number of generalised linear models fit at the first step to v. Given the often encountered asymmetric distributions for Y, equally spaced landmark points are not recommended. In our experience, the lowest sensitivity is obtained with landmark points fixed at equally spaced order statistics, as suggested above.

It shall be remarked that the sample size n is not actually reduced, and the entire sample is always used to estimate the conditional CDF. We reduce only the number of points at which the conditional CDF is evaluated (from n to v). As soon as v is large enough (in the order of the hundreds), in our experience there is very little sensitivity to v. A sensitivity analysis is anyway recommended in specific empirical applications.

An additional approach, which reduces also the computational burden of fitting each separate model in the first step, is subsampling, that is, basing the first step only on a random subset of the data and estimating the conditional CDF everywhere. Note that subsampling might be used only in the first step, while the second step can almost always be based on the entire sample. We stress that we do not use any subsampling in this work.

2.2 Standard errors

Our main task in this section involves deriving an estimator for the standard errors for $\hat{\beta}$. To do so, we make use of the total variance law, which can be expressed as follows:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{E}_{\hat{\gamma}}(\mathrm{Var}_{\hat{\beta}}(\hat{\beta}|\hat{\gamma})) + \mathrm{Var}_{\hat{\gamma}}(\mathrm{E}_{\hat{\beta}}(\hat{\beta}|\hat{\gamma})). \tag{6}$$

For the estimation of the first addend, it is possible to set up a Huber-White type estimator of the kind

$$(X'X)^{-1}X'DX(X'X)^{-1},$$

where $D = \mathrm{diag}((u_i - x_i'\hat{\beta})^2)$.

For estimation of the second addend, on the other hand, one must use the Delta method as

$$\frac{\partial \hat{\beta}}{\partial F}' \mathrm{Var}(\hat{F}) \frac{\partial \hat{\beta}}{\partial F}. \tag{7}$$

In case (3) is used, the middle term becomes $\mathrm{Var}(\hat{\gamma})$ and can be well approximated by a block matrix made of p by p variance covariance matrices for the coefficients of

logistic regression models; which is a direct by-product of the estimator used at the first stage. In fact, the computation of standard errors is greatly simplified by the block structure of $\text{Var}(\hat{\gamma})$ and by the fact that the Jacobian involved in (7) is zero whenever $\tau_2 < \hat{F}(y_i|x_i) < \tau_1$ for all $i = 1, \ldots, n$. In case one estimates the conditional CDF on a grid, additionally, only entries such that $\hat{F}(\tilde{y}_{(j)}) \neq \hat{F}(\tilde{y}_{(j+1)})$ are non-zero, for all $j < v$. For simplicity, in our implementation the non-zero entries of the p-dimensional Jacobian are obtained through numerical differentiation. The resulting estimator for the standard errors will be evaluated in the simulation study in Section 3.

2.3 Theoretical results

In this section, we show consistency and asymptotic normality of our proposed quantile share ratio regression methodology.

Theorem 1 *Assume Y admits a density $f(y|x)$ that is uniformly bounded and continuous for any x. Fix $\tau_1 \geq \tau_2$. Assume $X'X/n$ converges to a positive definite matrix C and $y_1, \ldots, y_n$ are i.i.d. samples from $f(y|x_i)$. Assume that the estimator of the conditional quantile function used in the first step is consistent and asymptotically normal, with a rate of convergence $\rho_n \leq \sqrt{n}$. Then, there exists a positive definite matrix Ω such that*

$$\rho_n(\hat{\beta} - \beta) \xrightarrow{d} MVN(0, \Omega).$$

Proof First of all, note that $\widehat{QSR}(x)$ can also be expressed as the ratio of

$$1/n \sum_{j=1}^{n} I(y_{(j)} \geq \hat{q}(\tau_1, x_i)) y_{(j)} (\hat{F}(y_{(j)}|x_i) - \hat{F}(y_{(j-1)}|x_i)) \tag{8}$$

and

$$1/n - \frac{\sum_{j=1}^{n} I(y_{(j)} > \hat{q}(\tau_2, x_i)) y_{(j)} (\hat{F}(y_{(j)}|x_i) - \hat{F}(y_{(j-1)}|x_i))}{n}. \tag{9}$$

Under the assumptions, it is shown in Chernozhukov et al. (2013) (see Theorem 5.2 and Corollary 5.4) that the two quantities jointly converge in distribution to a bivariate normal. Now we can use the delta method to show that the ratio of (8) and (9) is asymptotically normal and converges to $QSR(x)$. Due to (2), $QSR(x) = g^{-1}(x'\beta)$, therefore, $\widehat{QSR}(x)$ converges almost surely to $g^{-1}(x'\beta)$. Since $\hat{\beta}$ is obtained as in (5), one can finally use the Slutzky theorem to prove the result.

The result is general and holds for any (consistent and asymptotically normal) estimator of the conditional CDF used in the first step. It should be remarked that kernel density estimators are slightly more flexible than (logistic) regression approaches, but their rate of convergence is invariably slightly slower than $\sqrt{n}$, and the bandwidths should be well calibrated (that is, some under-smoothing is required as the sample size grows). On this point, see also Geraci and Farcomeni (2022). On the other hand, (3) has the usual $\sqrt{n}$ rate of convergence, but it requires that the model is well specified (see also Chernozhukov et al. (2013)). As expected, it is straightforward to check that

if the conditional CDF is estimated on a grid of v values, Theorem 1 holds only as $v \to \infty$; and the convergence rate corresponds to ρ_v.

3 Simulations

In this section, we describe the results of a simulation study.

We focus on both the quintile ($\tau_1 = 0.8$, $\tau_2 = 0.2$) and the Palma ($\tau_1 = 0.9$, $\tau_2 = 0.4$) share ratios.

The data generating mechanism is based on two covariates. We sample X_1 from a standard Gaussian and X_2 from a Bernoulli with a success rate of 1/3; while $X_0 = 1$ to accommodate an intercept. We use a logarithmic link function and compute the linear predictor after fixing $\beta = \{(1.5, -0.5, 0.5), (1.5, 0.5, -0.5)\}$. These choices mirror true parameters that would be expected in real situations. More precisely, the intercept is very close to the intercepts estimated in the real data analysis below. The two slopes roughly correspond to medium and weak effect sizes in the classical Cohen's definition.

We generate the response based on three different sampling distributions: a Weibull, a log-normal, and a Pareto I. In the case of the Weibull distribution, we fix a unit scale and set individual-specific shape parameters to yield the quantile share ratios of interest at population level. We do so through (univariate) numerical optimisation. This is necessary since the population-level quantile share ratios are not fixed before hand, and they are not directly simulated. The data are generated from a parametric distribution, whose parameters must be fixed so that the population-level quantile share ratio corresponds to the one implied by (2). For the log-normal distribution we fix a zero mean on the logarithmic scale, and optimise individual-specific standard deviations to yield the quantile share ratios of interest at population level. For the Pareto distribution, we do not fix any parameter. We let $n = \{1000, 5000\}$, thus evaluating 24 simulation scenarios, combining two levels for (τ_1, τ_2), two sample sizes, two true β coefficients combinations, and three data generating distributions.

For each scenario we generate data $B = 1000$ times, and estimate our model and its standard errors both with our Huber-White approach and non-parametric bootstrap. We also estimate a naive competitor which strictly mimicks the current most common analysis in this setting. Presently, an analyst would proceed with a simple linear regression at an aggregated (e.g., region, discrict, country) level; with therefore a very small number of units (in the order of the tens). Hence, we obtain an aggregated data set by partitioning uniformly at random the generated data in thirty blocks. Within each block we compute the (marginal) quantile share ratio, and aggregate the covariates. A linear regression model is then estimated on the logarithms of the quantile share ratios thus obtained. Note that this approach should give optimistic results for the competitor, as no selection/ecological bias is driven by the independent aggregation strategy. This does not apply to real data analyses at the aggregated level, and hence in real data analysis the naive approach might have a worse performance than what is reported in this simulation study.

We evaluate the mean squared error (MSE) for parameter estimates and, in order to validate our proposed approach for standard error estimation, the coverage of the 95%

confidence intervals. MSE are reported in Table 1. It can be seen that, as expected, QSRR yields substantially smaller MSE than the naive approach, often by several orders of magnitude. In particular, the MSE decreases at the expected rate with n; and it is slightly smaller for the quintile share ratio. This could be expected, as the Palma ratio involves observations further in the right tail, which inevitably increases uncertainty.

Coverage of the 95% confidence intervals is reported in Table 2. We show coverage of confidence intervals for both the cases in which the proposed Huber-White approach is used, and for non-parametric bootstrap. The coverage of the 95% confidence intervals is good, although a bit variable. For the estimates based on the Huber-White approach, coverage can be expected to improve for larger sample sizes. The nonparametric bootstrap seems to perform slightly better, albeit its results are also a bit variable. All in all, we recommend using the nonparametric boostrap for small sample sizes, and the Huber-White approach for larger ones. The latter is clearly much faster.

4 Real data analysis

Our data analysis involves the cross-sectional component of the EU Statistics on Income and Living Conditions (EU-SILC) survey. We have data on $n = 2815104$ households, sampled by EUROSTAT in 23 countries in the period 2007 to 2021. The assumption that households are independent is tenable, as no household is repeatedly observed (i.e. we have a repeated cross-sectional data set, not a panel). Our response variable is the Equivalised Disposable Income (EDI), defined as the household income per equivalised adult, after tax and other deductions. Formally, EDI is computed by dividing the total income after tax and deductions by the number of household members converted into equivalised adults. The latter operation involves weighting each household member according to the modified OECD equivalence scale (unit weight to the first adult, 0.5 to the second adult and each subsequent person aged 14 and over, 0.3 to each child under 14). Our main interest resides in evaluating time trends and differences among countries in quantile share ratios. We also obtain data from the World Bank and Our World in Data (`https://ourworldindata.org/`) about country- and year-specific unemployment rates (as a proportion of the total labor force), a standardized index of corruption control, and proportion of seats held by women in national parliaments. Note that the second and third variables can be seen as proxies of justice and gender equalities. A second analysis involves study of the *adjusted* trends and differences among countries, after conditioning on the three additional covariates; and association between the three variables mentioned above and income inequality.

In Table 3 we report the country-specific and total number of households, median EDI, quantile ratios, and summary statistics for the three covariates. Substantial heterogeneity can be clearly seen, and it is also testified by the large quantile share ratios for the pooled data set.

As testified by the simulation study, a naive approach which would treat the measurements in Table 3 as the observed data would be biased. We now proceed with our

Table 1 Simulation study. MSE for parameter estimates for the proposed QSRR approach and a naive aggregation approach; for different values of the sample size n, true parameter configuration, and quantile share ratio of interest. Results are based on $B = 1000$ replicates.

n	β_2	τ_1	τ_2	QSRR	Naive
Weibull data					
1000	-0.5	0.9	0.4	0.064	2.299
1000	-0.5	0.8	0.2	0.022	0.697
1000	0.5	0.9	0.4	0.084	2.977
1000	0.5	0.8	0.2	0.032	1.058
5000	-0.5	0.9	0.4	0.015	2.561
5000	-0.5	0.8	0.2	0.006	0.772
5000	0.5	0.9	0.4	0.018	3.320
5000	0.5	0.8	0.2	0.007	1.151
Lognormal data					
1000	-1	0.9	0.4	0.077	3.111
1000	-1	0.8	0.2	0.018	0.640
1000	1	0.9	0.4	0.095	3.917
1000	1	0.8	0.2	0.080	0.946
5000	-1	0.9	0.4	0.026	3.645
5000	-1	0.8	0.2	0.008	0.611
5000	1	0.9	0.4	0.028	4.941
5000	1	0.8	0.2	0.008	0.963
Pareto data					
1000	-1	0.90	0.40	0.77	8.76
1000	-1	0.80	0.20	0.99	2.83
1000	1	0.90	0.40	0.77	10.33
1000	1	0.80	0.20	0.87	3.73
5000	-1	0.90	0.40	0.53	22.72
5000	-1	0.80	0.20	0.15	5.13
5000	1	0.90	0.40	0.58	27.86
5000	1	0.80	0.20	0.14	8.04

proposed method, which works at the household level. The EU-SILC survey involves complex sampling. We incorporate the sampling weights in our approach, which are computed by EUROSTAT, in order to obtain results that can be generalised to the entire popolation.

We compare different link functions: the logarithmic one, a scaled logit (defined as the logit of the outcome minus its minimum, divided by its range), a scaled probit (defined as the standard normal CDF computed on the outcome minus its minimum, divided by its range), and an identity link. The latter is inappropriate, but it is sometimes used in generalised linear models with positive responses for ease of interpretation. We make this comparison through Pregibon-type goodness-of-link statistics: after estimation of the pseudo-outcomes $\widehat{QSR(x)}$ at the first stage, we predict their transform $\widehat{g}QSR(x)$ through a linear model as in (2). For each link g_j we then estimate η

Table 2 Simulation study. Coverage of 95% confidence intervals for the proposed QSRR approach; for different values of the sample size n, true parameter configuration, and quantile share ratio of interest. The standard error can be either estimated in closed form as in (6), or via the non-parametric bootstrap. Results are based on $B = 1000$ replicates.

				Closed form SE			Bootstrap		
n	β_2	τ_1	τ_2	β_0	β_1	β_2	β_0	β_1	β_2
Weibull data									
1000	-1	0.9	0.4	0.954	0.966	0.972	0.949	0.953	0.964
1000	-1	0.8	0.2	0.998	0.991	0.992	0.988	0.977	0.943
1000	1	0.9	0.4	0.953	0.974	0.982	0.931	0.965	0.944
1000	1	0.8	0.2	0.994	1.000	0.978	0.957	0.991	0.933
5000	-1	0.9	0.4	0.978	0.973	0.978	0.987	0.984	0.953
5000	-1	0.8	0.2	0.947	0.976	0.949	0.966	0.961	0.969
5000	1	0.9	0.4	0.939	0.994	0.978	0.942	0.979	0.963
5000	1	0.8	0.2	0.958	0.981	0.973	0.991	0.967	0.942
Lognormal data									
1000	-1	0.9	0.4	0.931	0.950	0.958	0.986	0.990	1.000
1000	-1	0.8	0.2	1.000	0.983	1.000	0.995	0.979	0.997
1000	1	0.9	0.4	0.899	0.974	0.939	0.965	0.956	0.943
1000	1	0.8	0.2	0.993	0.996	1.000	0.957	0.952	0.954
5000	-1	0.9	0.4	0.992	0.996	0.993	0.955	0.981	0.977
5000	-1	0.8	0.2	0.987	0.991	0.996	0.969	0.958	0.959
5000	1	0.9	0.4	0.970	0.999	0.985	0.941	0.948	0.969
5000	1	0.8	0.2	0.952	0.986	0.990	0.944	0.984	0.966
Pareto data									
1000	-1	0.90	0.40	0.987	0.979	0.999	0.984	0.959	0.989
1000	-1	0.80	0.20	0.966	0.982	0.942	0.965	0.976	0.972
1000	1	0.90	0.40	0.995	0.984	0.962	0.944	0.964	0.970
1000	1	0.80	0.20	0.981	0.953	0.970	0.992	0.974	0.930
5000	-1	0.90	0.40	0.971	0.983	0.967	0.988	0.970	0.907
5000	-1	0.80	0.20	0.947	0.940	0.963	0.961	0.938	0.947
5000	1	0.90	0.40	0.980	0.972	0.984	0.952	0.915	0.971
5000	1	0.80	0.20	0.962	0.983	0.965	0.994	0.984	0.946

parameters in the model

$$E[\widehat{QSR(x)}] = g_j^{-1}(\eta_0 + \eta_1 \widehat{g}QSR(x) + \eta_2(\widehat{g}QSR(x))^2).$$

The classical Pregibon goodness-of-link approach postulates that a well-specified link will yield a small T statistic for $\hat{\eta}_2$. In Table 4 we report these figures for the four models of interest (two combinations of τ_1 and τ_2, and two sets of covariates, which will be better detailed below). Given the very large sample size, the fact that T statistics would be associated with significant tests should not be taken as an indication of bad

Table 3 Real data analysis. Number of households n, median equivalised disposable income in hundreds of Euro (EDI), quintile share ratio (Quintile SR), Palma ratio, unemployment rate (Unemp), index of control of corruption (Corr), and proportion of Women in parliament by country and across all countries (Total).

Country	n	EDI	Quintile SR	Palma	Unemp	Corr	Women
Belgium	86339	203	4.12	1.00	7.25	1.47	39.20
Cyprus	60836	152	4.99	1.31	9.48	0.87	15.00
Denmark	89042	295	3.66	0.93	6.03	2.30	38.55
Estonia	85046	68	6.05	1.42	8.05	1.22	24.11
France	169879	212	4.42	1.12	9.08	1.34	30.99
Germany	225073	208	4.87	1.15	5.25	1.79	32.63
Greece	178918	82	5.64	1.23	18.14	-0.02	19.00
Hungary	132395	46	3.87	0.96	7.19	0.22	10.86
Ireland	73365	211	4.86	1.23	9.52	1.62	16.98
Italy	295019	170	5.80	1.26	9.79	0.23	25.67
Latvia	89198	47	7.10	1.76	10.82	0.38	23.62
Lithuania	76606	49	7.03	1.65	9.66	0.47	23.05
Luxembourg	61753	349	4.72	1.13	5.47	2.03	26.19
Malta	59157	116	4.59	1.12	5.32	0.71	12.20
Netherlands	171975	231	3.58	0.92	5.04	1.99	37.78
Norway	89949	393	3.78	0.87	3.62	2.11	40.59
Poland	210883	53	5.18	1.23	7.05	0.59	24.57
Portugal	125319	87	5.72	1.41	10.22	0.91	32.73
Romania	112618	24	7.60	1.61	6.02	-0.23	14.85
Slovak Republic	81798	65	3.61	0.83	10.38	0.20	18.57
Slovenia	132056	122	3.59	0.86	6.80	0.84	24.81
Sweden	98394	242	3.94	0.91	7.50	2.16	45.72
Switzerland	109486	388	5.05	1.19	4.53	2.05	33.86
Total	2815104	139	13.20	2.56	7.95	1.10	26.01

fit or inappropriate link. In order to support this claim, in the table we also report the squared correlation between the pseudo-outcomes predicted at the first step (through the non-parametric estimator) and the ones predicted at the second step (after solving (2)). We do so only for the chosen link, which is always the logarithmic one. It can be seen that fit is always very good. It shall also be remarked that in all cases the largest pseudo-R^2 was associated with the chosen link, which is an additional minor evidence in favor of the choice made.

We therefore always specify logarithmic link functions. As main analysis, we include a linear and quadratic effect of time (in years) and dummies for all countries except Belgium, which is used as baseline. The conditional CDF is estimated at $v = 1000$ points. We estimated standard errors as proposed, and we compared the estimates with nonparametric bootstrap. The two estimates are very close to each other, as could be expected given the large sample size.

Table 4 Real data analysis. Pregibon-type goodness-of-link statistics, plus pseudo-R^2 for the chosen link. *Main* models include country indicators and quadratic time effects, *main+covs* model in addition estimate effects for unemployment, corruption, and women proportion in the parliament.

			Link function				
Model	τ_1	τ_2	log	logit	probit	identity	pseudo-R^2
Main	0.8	0.2	50.21	61.80	81.70	67.21	0.89
Main+covs	0.8	0.2	51.23	61.63	69.06	92.51	0.92
Main	0.9	0.4	54.05	96.60	88.74	65.99	0.92
Main+covs	0.9	0.4	28.95	46.2	43.30	38.23	0.95

Results are reported in Table 5 for the quintile share ratio, and in Table 6 for the Palma ratio.

Palma and quintile share ratios are on a slightly different scale (Palma ratios are smaller), but the covariate effects are quite similar. A brief interpretation follows, that mostly applies to both share ratios. Both marginally and conditionally on confounders, a significant decreasing time trend can be seen in both cases. Many countries are significantly more unequal than Belgium (prominently, Latvia, Lithuania, Portugal, Greece, Switzerland, Italy). Given our study design, we cannot make causal claims. However, unsurprisingly, better control of corruption and gender equality (measured by the proportion of women in the parliament) are both independently associated with lower income inequality. Interestingly, after adjusting for these predictors, some country differences are not significant anymore (e.g. countries like Cyprus, Italy, and Greece have a QSR very similar to that of Belgium after adjusting for control of corruption and women participation in the national parliament); and some effects are even reversed (e.g., the difference between Sweden and Belgium becomes positive after adjusting for covariates). Additionally, while positive differences with respect to Belgium often decrease after adjusting for covariates (an indication that more justice and gender equality might be associated with more economic equality for those countries); for some countries, like Switzerland, the differences even increase (an indication that economic inequality is not strongly linked to social justice in those countries, and might be driven by other factors).

5 Concluding remarks

We have proposed a distribution-free approach to modeling conditional quantile share ratios. The modeling specification is simple and in parallel with classical generalised linear models for conditional expectation, and quantile regression for conditional quantiles.

For model estimation, we have proposed a two-step approach, which seems to work well and is readily scalable. We have demonstrated the approach on a large-scale example on the study of income inequality in Europe in the recent past, where we have formally shown significant differences between countries and time trends. We have also proposed a Huber-White estimator for standard errors, which is computationally

Table 5 Real data analysis. Quintile share ratio regression ($\tau_2 = 0.2$, $\tau_1 = 0.8$).

Variable	$\hat{\beta}$	2.5% CI	97.5% CI	p-value	$\hat{\beta}$	2.5% CI	97.5% CI	p-value
Intercept	1.41	1.39	1.44	< 0.01	1.81	1.69	1.92	< 0.01
$(t - 2014)$	-0.01	-0.01	-0.00	< 0.01	-0.00	-0.00	0.00	0.10
$(t - 2014)^2/10$	-0.01	-0.02	-0.01	< 0.01	-0.01	-0.01	-0.00	< 0.01
Cyprus	0.22	0.18	0.25	< 0.01	0.04	-0.03	0.10	0.24
Denmark	-0.15	-0.18	-0.12	< 0.01	0.01	-0.05	0.07	0.72
Estonia	0.46	0.40	0.53	< 0.01	0.33	0.27	0.39	< 0.01
France	0.05	0.02	0.08	< 0.01	0.00	-0.03	0.04	0.80
Germany	0.17	0.13	0.21	< 0.01	0.20	0.15	0.25	< 0.01
Greece	0.35	0.29	0.41	< 0.01	-0.00	-0.13	0.13	0.96
Hungary	-0.03	-0.08	0.03	0.34	-0.33	-0.45	-0.22	< 0.01
Ireland	0.18	0.14	0.21	< 0.01	0.14	0.09	0.19	< 0.01
Italy	0.35	0.29	0.41	< 0.01	0.08	0.00	0.17	0.05
Latvia	0.66	0.60	0.72	< 0.01	0.37	0.27	0.47	< 0.01
Lithuania	0.60	0.54	0.65	< 0.01	0.31	0.21	0.40	< 0.01
Luxembourg	0.14	0.11	0.17	< 0.01	0.21	0.16	0.26	< 0.01
Malta	0.13	0.09	0.17	< 0.01	-0.10	-0.17	-0.03	0.01
Netherlands	-0.11	-0.14	-0.08	< 0.01	-0.02	-0.06	0.02	0.35
Norway	-0.09	-0.12	-0.06	< 0.01	0.04	-0.02	0.09	0.19
Poland	0.26	0.11	0.40	< 0.01	0.03	-0.05	0.10	0.45
Portugal	0.40	0.37	0.44	< 0.01	0.28	0.22	0.34	< 0.01
Romania	0.23	0.07	0.40	0.01	-0.04	-0.24	0.16	0.72
Slovak Republic	-0.09	-0.14	-0.03	< 0.01	-0.39	-0.49	-0.29	< 0.01
Slovenia	-0.11	-0.15	-0.07	< 0.01	-0.26	-0.32	-0.20	< 0.01
Sweden	-0.04	-0.08	-0.01	0.02	0.12	0.06	0.17	< 0.01
Switzerland	0.19	0.16	0.23	< 0.01	0.29	0.24	0.34	< 0.01
Unemp/10					-0.01	-0.04	0.02	0.64
Corr					-0.19	-0.26	-0.13	< 0.01
Women/10					-0.03	-0.05	-0.02	< 0.01

less intense than non-parametric bootstrap, and is recommended for large samples. We mention that we have also experimented with block-bootstrap, which in this setting does not seem to give any advantage.

There are several routes for further work that remain open. From an applied perspective, we believe that quantile share ratio regression and quantile ratio regression (Farcomeni and Geraci 2024) might be useful to investigate important open questions in the literature on economic inequality. These involve, for instance, the investigation of gender effects, and causal claims related to policies. Causal investigations nonetheless might not be straightforward, due to the necessity of identifying causal estimands in the highly non-linear setting of inequality indexes.

Table 6 Real data analysis. Palma ratio regression ($\tau_2 = 0.4$, $\tau_1 = 0.9$).

Variable	$\hat{\beta}$	2.5% CI	97.5% CI	p-value	$\hat{\beta}$	2.5% CI	97.5% CI	p-value
Intercept	-0.01	-0.03	0.01	0.38	0.61	0.48	0.74	< 0.01
$(t - 2014)$	-0.01	-0.01	-0.01	< 0.01	-0.00	-0.00	-0.00	0.02
$(t - 2014)^2/10$	-0.02	-0.02	-0.01	< 0.01	-0.01	-0.01	-0.00	0.04
Cyprus	0.29	0.25	0.33	< 0.01	-0.03	-0.11	0.05	0.40
Denmark	-0.12	-0.15	-0.10	< 0.01	0.11	0.06	0.16	< 0.01
Estonia	0.41	0.35	0.48	< 0.01	0.22	0.16	0.28	< 0.01
France	0.09	0.06	0.13	< 0.01	-0.01	-0.04	0.02	0.64
Germany	0.16	0.12	0.19	< 0.01	0.20	0.16	0.24	< 0.01
Greece	0.29	0.24	0.34	< 0.01	-0.31	-0.44	-0.19	< 0.01
Hungary	-0.01	-0.06	0.03	0.48	-0.53	-0.64	-0.42	< 0.01
Ireland	0.21	0.18	0.23	< 0.01	0.10	0.04	0.15	< 0.01
Italy	0.26	0.22	0.30	< 0.01	-0.16	-0.25	-0.07	< 0.01
Latvia	0.61	0.56	0.65	< 0.01	0.15	0.07	0.23	< 0.01
Lithuania	0.57	0.52	0.62	< 0.01	0.14	0.05	0.23	< 0.01
Luxembourg	0.12	0.09	0.14	< 0.01	0.19	0.15	0.24	< 0.01
Malta	0.15	0.12	0.18	< 0.01	-0.22	-0.30	-0.14	< 0.01
Netherlands	-0.06	-0.09	-0.04	< 0.01	0.08	0.04	0.12	< 0.01
Norway	-0.14	-0.16	-0.11	< 0.01	0.06	0.01	0.11	0.01
Poland	0.23	0.18	0.29	< 0.01	-0.12	-0.20	-0.05	< 0.01
Portugal	0.41	0.37	0.44	< 0.01	0.19	0.14	0.24	< 0.01
Romania	0.81	0.63	0.99	< 0.01	0.05	-0.18	0.28	0.68
Slovak Republic	-0.13	-0.17	-0.09	< 0.01	-0.63	-0.73	-0.53	< 0.01
Slovenia	-0.12	-0.16	-0.09	< 0.01	-0.37	-0.42	-0.32	< 0.01
Sweden	-0.08	-0.11	-0.05	< 0.01	0.16	0.11	0.20	< 0.01
Switzerland	0.16	0.13	0.18	< 0.01	0.29	0.25	0.33	< 0.01
Unemp/10					0.03	-0.00	0.07	0.07
Corr					-0.28	-0.34	-0.22	< 0.01
Women/10					-0.07	-0.09	-0.04	< 0.01

From the perspective of the statistical methodology, we believe that there are three issues that deserve attention. First, the proposed method requires independent observations. This assumption works well for the data at hand, which involve repeated cross-sectional surveys with no overlap of units, but it shall be relaxed in the presence of panel data. Secondly, we restricted our presentation to the case of absolutely continuous random variables. In real applications, it might happen that outcomes are rounded or coarsened, or maybe even discrete measures might be of interest. The proposed method might also be applied in the presence of ties and discrete data once the conditional CDF is appropriately estimated (see Geraci and Farcomeni (2022) for a discussion of this point). To do so, along the lines of Geraci and Farcomeni (2022), one might possibly only need to replace the conditional CDF estimator with an estimator

of the conditional mid-CDF, which is defined as

$$G(y|x) = F(y|x) - 0.5 \Pr(Y = y|x).$$

Clearly, a full exploration of this approach is left as further work. For instance, theoretical guarantees in the presence of ties are quite complicated; and it might be more appropriate to modify the definition of quantile share ratios to this end.

A further route for further work would be an extension of quantile ratio and quantile share ratio regression to the case of spatial data (e.g., Chen and Tokdar (2021); Castiglione et al. (2025); Xu et al. (2026)), which is often encountered in the study of economic inequality.

Finally, we are convinced that for both marginal and conditional quantile share ratios, the most pressing issue is robust inference. Quantile share ratios involve the estimation of (conditional) sums above certain thresholds at the nominator, which are intuitively very sensitive to outliers. This problem has already been considered for the case of marginal quantile share ratios (e.g. Cowell and Victoria-Feser (1996), Cowell and Flachaire (2007), Hulliger and Schoch (2014), and references therein and thereof). A possible route to robustification of our methodology would for instance involve the use of robust logistic regression at the first stage and least trimmed of squares at the second (Heritier et al. 2009; Farcomeni and Ventura 2012). A development of robust quantile ratio and quantile share ratio regression methodologies is also left for further work.

Funding Open access funding provided by Università degli Studi di Roma Tor Vergata within the CRUI-CARE Agreement.

References

Aaberge R, Bjerve S, Doksum K (2005) Decomposition of rank-dependent measures of inequality by subgroups. Metron LXIII:493–503

Beach CM, Davidson R (1983) Distribution-free statistical inference with Lorenz curves and income shares. Rev Econ Stud 50(4):723–735

Castiglione C, Arnone E, Bernardi M, Farcomeni A, Sangalli L (2025) PDE-regularised spatial quantile regression. J Multivar Anal 205:105381

Chancel L, Piketty T (2021) Global income inequality, 1820–2020: the persistence and mutation of extreme inequality. J Eur Econ Assoc 19:3025–3062

Chancel L, Piketty T, Saez E, Zucman G (2022) World inequality report 2022. Technical report, World Inequality Lab, Paris

Chen X, Tokdar ST (2021) Joint quantile regression for spatial data. Journal of the Royal Statistical Society (Series B) 83:826–852

Chernozhukov V, Fernandez-Val I, Melly B (2013) Inference on counterfactual distributions. Econometrica 81:2205–2268
Cowell F (2011) Measuring inequality. Oxford University Press, Oxford
Cowell FA, Flachaire E (2007) Income distribution and inequality measurement: The problem of extreme values. Journal of Econometrics 141:1044–1072
Cowell FA, Victoria-Feser M-P (1996) Robustness properties of inequality measures. Econometrica 64:77–101
Davidson R, Flachaire E (2007) Asymptotic and bootstrap inference for inequality and poverty measures. Journal of Econometrics 141:141–166
Eiselt HA, Marianov V (2011) Foundations of Location Analysis. Springer, New York
Farcomeni A, Geraci M (2024) Quantile ratio regression. Stat Comput 34:94
Farcomeni A, Ventura L (2012) An overview of robust methods in medical research. Stat Methods Med Res 21:111–133
Firpo S, Fortin NM, Lemieux T (2009) Unconditional quantile regressions. Econometrica 77:953–973
Firpo S, Fortin NM, Lemieux T (2018) Decomposing wage distributions using recentered influence function regression. Econometrics 6:28
Geraci M, Farcomeni A (2022) Mid-quantile regression for discrete responses. Stat Methods Med Res 31:821–838
Hall P, Wolff RCL, Yao Q (1999) Methods for estimating a conditional distribution function. J Am Stat Assoc 94(445):154–163
Harper S, Lynch J (2007) Using innovative inequality measures in epidemiology. Int J Epidemiol 36:926–928
Heritier S, Cantoni E, Copt S, Victoria-Feser MP (2009) Robust methods in biostatistics. Wiley, Chichester
Heuchenne C, Jacquemain A (2022) Inference for monotone single-index conditional means: A Lorenz regression approach. Comput Stat Data Anal 167:107347
Hulliger B, Schoch T (2014) Robust, distribution-free inference for income share ratios under complex sampling. AStA Advances in Statistical Analysis 98:63–85
Kebe M, Deme EH, Kpanzou TA, Manou-Abi SM, Sisawo E (2023) Kernel estimation of the quintile share ration index of inequality for heavy-tailed income distributions. European Journal of Pure and Applied Mathematics 16:2509–2543
Kpanzou TA (2015) On the influence function of the quintile share ratio. Communications in Statistics - Simulation and Computation 44:2492–2499
Langel M, Tillé Y (2011) Statistical inference for the quintile share ratio. Journal of Statistical Planning and Inference 141:2976–2985
Li Q, Lin J, Racine JS (2013) Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. Journal of Business & Economic Statistics 31(1):57–65
Li Q, Racine JS (2008) Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. Journal of Business & Economic Statistics 26(4):423–434
Palma JG (2011) Homogeneous middles vs. heterogeneous tails, and the end of the "inverted-U": It's all about the share of the rich. Dev Chang 42:87–153
Palma JG, Stiglitz JE (2016) Do nations just get the inequality they deserve? the "Palma ratio" re-examined. In: Basu K, Stiglitz JE (eds) 'Inequality and Growth: Patterns and Policy', International Economics Association. Palgrave Macmillan, London
Peracchi F (2002) On estimating conditional quantiles and distribution functions. Computational Statistics & Data Analysis 38(4):433–447
Xu X-Y, Wang J-F, Hu K, He S, Xia Y (2026) Spatial local linear quantile regression under association. Statistics & Probability Letters 228:110573