

Superquantile Regression with Applications to Buffered Reliability, Uncertainty Quantification, and Conditional Value-at-Risk

R. T. Rockafellar

University of Washington, Seattle, WA

J. O. Royset

Naval Postgraduate School, Monterey, CA

S. I. Miranda

Portuguese Navy, Lisbon, Portugal

February 6, 2013

Abstract

The paper presents a generalized regression technique centered on a superquantile (also called conditional value-at-risk) that is consistent with that coherent measure of risk and yields more conservatively fitted curves than classical least-squares and quantile regressions. In contrast to other generalized regression techniques that approximate conditional superquantiles by various combinations of conditional quantiles, we directly and in perfect analog to classical regression obtain superquantile regression functions as optimal solutions of certain error minimization problems. We show the existence and possible uniqueness of regression functions, discuss the stability of regression functions under perturbations and approximation of the underlying data, and propose an extension of the coefficient of determination R-squared for assessing the goodness of fit. The paper presents two numerical methods for solving the error minimization problems and illustrates the methodology in several numerical examples in the areas of uncertainty quantification, reliability engineering, and financial risk management.

1 Introduction

Analysts and decision makers are often concerned with a random variable describing possible ‘cost,’ ‘loss,’ or ‘damage.’ The interest may be focused on a single ‘system’ or could involve study and comparison across a multitude of systems and designs. In either case, it may be beneficial to attempt to approximate such a *loss random variable* Y in terms of an n -dimensional *explanatory random vector* X that is more accessible in some sense. This situation naturally leads to least-squares regression and related models that estimate conditional expectations. While such models are adequate in many situations, they fall short in contexts where a decision maker is risk averse, i.e., is more concerned about upper-tail realizations of Y than average loss, and views errors asymmetrically with underestimating losses being more detrimental than overestimating. We focus on such contexts and therefore maintain an orientation of Y that implies that high realizations are unfortunate and low realizations are favorable. Of course, a parallel development with an opposite orientation of the random variable Y , focused on profits and gains, and concerns about overestimating instead of underestimating is also possible but not pursued here.

Quantile regression (see [16, 9] and references therein) accommodates risk-averseness and an asymmetric view of errors by estimating conditional quantiles at a certain probability level such as those in the tail of the conditional distribution of Y . A quantile corresponds to ‘value-at-risk’ (VaR)

Report Documentation Page			Form Approved OMB No. 0704-0188		
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 06 FEB 2013		2. REPORT TYPE		3. DATES COVERED 00-00-2013 to 00-00-2013	
4. TITLE AND SUBTITLE Superquantile Regression with Applications to Buffered Reliability, Uncertainty Quantification, and Conditional Value-at-Risk				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School, Monterey, CA, 93943				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES in review					
14. ABSTRACT The paper presents a generalized regression technique centered on a superquantile (also called conditional value-at-risk) that is consistent with that coherent measure of risk and yields more conservatively fitted curves than classical least-squares and quantile regressions. In contrast to other generalized regression techniques that approximate conditional superquantiles by various combinations of conditional quantiles, we directly and in perfect analog to classical regression obtain superquantile regression functions as optimal solutions of certain error minimization problems. We show the existence and possible uniqueness of regression functions, discuss the stability of regression functions under perturbations and approximation of the underlying data and propose an extension of the coefficient of determination R-squared for assessing the goodness of fit. The paper presents two numerical methods for solving the error minimization problems and illustrates the methodology in several numerical examples in the areas of uncertainty quantification reliability engineering, and financial risk management.					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 31	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

in financial terminology and relates to ‘failure probability’ in engineering terms. Quantile regression informs the decision maker about these quantities conditional on values of the explanatory random vector X . However, a quantile is not a *coherent* measure of risk in the sense of Artzner et al. [2] (see also [7]); it fails to be subadditive. Consequently, a quantile of the sum of two random variables may exceed the sum of the quantiles of each random variable at the same probability level, which runs counter to our understanding of what ‘risk’ should express. Moreover, quantiles cause computational challenges when incorporated into decision optimization problems as objective function, failure probability constraint, or chance constraint. The use of quantiles and the closely related failure probabilities is therefore problematic in risk-averse decision making; see [2, 25, 20, 21] for a detailed discussion.

A *superquantile* of a random variable, also called conditional value-at-risk, average value-at-risk, and expected shortfall¹, is an ‘average’ of certain quantiles as described further below. It’s a coherent measure of risk well suited for risk-averse decision making and optimization; see [28] for its application in financial engineering, [13] for military applications, and [20] for use in reliability engineering. While this risk measure has reached prominence in risk-averse optimization, there has been much less work on regression techniques that are consistent in some sense with it. In this paper, we derive such a *superquantile regression* methodology, study its properties, and propose means to assess the goodness-of-fit. The importance of such a regression methodology becomes apparent by considering the following two situations.

Suppose that a loss is given by a random variable Y , but our primary concern is with the conditional loss given that an explanatory random vector X takes on specific values. We aim to select these values judiciously in an effort to minimize the conditional loss. We denote by $Y(x)$ the conditional random variable Y given that $X = x \in \mathbb{R}^n$. Of course, ‘minimizing’ $Y(x)$ is not well-defined and a standard approach is to minimize a risk measure of $Y(x)$; see for example [21]. An attractive choice is to use a superquantile measure of risk, which as mentioned above is coherent and also computationally approachable. While in some contexts a superquantile of $Y(x)$ can be evaluated easily for any $x \in \mathbb{R}^n$, there are numerous situations, especially beyond the financial domain, where only a data base of realizations of $Y(x)$ is available for various x . In the latter situation, there is a need for building an approximating model, based on the data, for the relevant superquantile of $Y(x)$ as a function of x . We refer to this as *superquantile tracking*. In comparison, if the goal were to minimize the expectation of $Y(x)$, then least-squares regression would yield a model that approximates that conditional expectation. Likewise, if the goal were to minimize a quantile of $Y(x)$, quantile regression would provide a model of the conditional quantile. While these models are valuable for analysts and decision makers focused on the expectation and quantile risk measures, they don’t provide estimates of conditional superquantiles. In essence, the same need for estimating conditional superquantiles arises in reliability engineering when the goal is to determine a ‘design’ x with buffered failure probability of $Y(x)$ being no larger than a given probability level, which corresponds to a constraint on a superquantile of $Y(x)$ [20].

Another situation arises when the explanatory random vector X is beyond our direct control, but the dependence between the loss random variable Y and X makes us hopeful that, for a carefully selected *regression function* $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the random variable $f(X)$ may serve as a surrogate for Y . When the distribution of X is known, at least approximately, and f has been determined, then the distribution of $f(X)$ is usually easily accessible. That distribution may then serve as input to further analysis, simulation, and optimization in place of the unknown distribution of Y . Such *surrogate estimation* may arise in numerous contexts. ‘Factor models’ in financial investment applications (see for example [6, 15]), where Y may be the loss associated with a particular asset and X a vector describing a small number of macroeconomic ‘factors,’ is a result of surrogate estimation.

¹We prefer the application-neutral name ‘superquantile’ when deriving methods applicable broadly.

‘Uncertainty quantification’ (see for example [17, 8]) considers the output of a system described by a random variable Y , for example measuring damage, and estimates its moments and distribution from observed realizations as well as knowledge about the distribution of the input to the system characterized by a random vector X . A main approach here centers on surrogate estimation with $f(X)$ serving as an estimate of Y . In this situation, an essential question is what criterion should be used for selecting f . Clearly, one would like the *error random variable* $Z_f := Y - f(X)$ to be small in some sense. However, minimizing the mean-squared error of Z_f would not reflect a greater concern about underestimating Y , i.e., underestimating losses, than overestimating. We may want to assess the error of Z_f in a manner that is ‘consistent’ with our use of a superquantile as risk measure.

In this paper, we develop a ‘generalized’ regression technique that addresses the issue of superquantile tracking and surrogate estimation. The technique is an extension of least-squares and quantile regression, which center on expectations and quantiles, respectively, to one that focuses on superquantiles.

The foundation of least-squares and quantile regression is the fact that mean and quantiles minimize the expectation of certain convex random functions. A natural extension to superquantile regression could then possibly involve determining a random function that when minimizing its expectation, we obtain a superquantile. However, such a random function doesn’t exist [10, 5], which has lead to studies of indirect approaches to superquantile regression grounded in quantile regression.

For a random variable with a continuous cumulative distribution function, a superquantile equals a conditional expectation of the random variable given realizations no lower than the corresponding quantile. Utilizing this fact, studies have developed kernel-based estimators for the conditional probability density functions, which are then integrated and inverted to obtain estimators of conditional quantiles. An estimator of the conditional superquantile is then finally constructed by integrating the density estimator over the interval above the quantile [26, 4] or forming a sample average [14]. These studies also include asymptotic analysis of the resulting estimators under a series of assumptions, including that the data originates from certain time series.

A superquantile of a random variable is defined in terms of an integral of corresponding quantiles with respect to the probability level. Since the integral is approximated by a weighted sum of quantiles across different probability levels, an estimator of a conditional superquantile emerges as the sum of conditional quantiles obtained by quantile regression; see [18, 19], which also show asymptotic results under a set of assumptions including the continuous differentiability of the cumulative distribution function of the conditional random variables. Similarly, [5] utilizes the integral expression for a superquantile, but observes that a weighted sum of quantiles is an optimal solution of a certain minimization problem; see [21]. Analogously to the situation in least-squares and quantile regression, an optimization problem therefore yields an estimator of a conditional superquantile. Though, in contrast to the case of least-squares and quantile regression, the estimator is ‘biased’ due to the error induced by replacing an integral by a finite sum. Under a linear model assumption, [5] also constructs a conditional superquantile estimator using an appropriately shifted least-squares regression curve based on quantile estimates of residuals. In both cases, asymptotic results are obtained for a homoscedastic linear regression model. Under the same model, [27] studies ‘constrained’ regression, where the error random variable $Z_f = Y - f(X)$ is minimized in some sense, for example in terms of least square or absolute deviation, subject to a constraint that limits a superquantile of Z_f . While this approach doesn’t lead to superquantile regression in the sense we derive below, it highlights the need for alternative techniques for regression that incorporate superquantiles in some manner.

The need for moving beyond classical regression centered on conditional expectations is therefor

now well recognized and has driven even further research towards estimating conditional distribution function, i.e., $Prob(Y(x) \leq y)$ for all $y \in \mathbb{R}$, using nonparametric kernel estimators (see for example [11]) and transformation models (see for example [12]). Of course, conditional distribution functions provide the ‘full’ information about $Y(x)$ including its quantiles and superquantiles, and therefor also provide a means to inform a risk-averse decision maker. In this paper, however, we directly focus on superquantiles, which we believe deserve special attention due to their prominence in risk analysis.

A framework for ‘generalized’ regression is laid out in [22, 21] and regression functions are obtained as optimal solutions of optimization problems of the form $\min_f \mathcal{E}(Z_f)$, where $\mathcal{E}$ is a *measure of error* and f is restricted to a certain class of functions such as the affine functions. Least-squares regression is obtained by $\mathcal{E}(Z_f) = E[Z_f^2]$, quantile regression with the Koenker-Bassett measure of error, but many other possibilities exist. While it is not possible to determine a measure of error that is of the expectation type and yields a superquantile, in this paper we show that when allowing for a broader class of functionals, a measure of error that generates a superquantile is indeed available. Such a measure of error is also hinted at in our recent paper [24], but the present paper gives the first comprehensive treatment. In contrast to previous studies towards superquantile regression, which utilize indirect approaches and quantile regression, we here offer a natural extension of least-squares and quantile regression. We replace the mean-squares and Koenker-Bassett error measures by a new error measure, and then simply minimize that error of Z_f to obtain a regression function. Under few assumptions, we establish the existence of a regression function, discuss its uniqueness, and examine stability under perturbations of the distribution of (X, Y) for example caused by sampling. We omit a discussion of simple linear models with independent and identically distributed (iid) noise as we believe that there is little need for quantile and superquantile regression in such contexts as least-squares regression with an appropriate shift suffices. In fact, we don’t separate models into (additive) deterministic and stochastic terms. In many applications, especially in the area of uncertainty quantification, heteroscedasticity and dependence are prevalent making linear iid and additive models of little value.

Section 2 describes measures of regret and error, first in the context of quantile regression and then for the extension to superquantile regression. Section 3 defines superquantile regression as the minimization of a measure of error, discusses existence and uniqueness of the regression function, and provides asymptotic results. Section 4 proposes an approach for assessing the goodness-of-fit of regression function obtained by superquantile regression. Section 5 deals with computational methods for superquantile regression and Section 6 gives illustrative examples.

2 Quantiles, Superquantiles, and Errors

While our development centers on superquantiles, it is beneficial to maintain a parallel description of quantiles. As we see below, quantile regression, which is achieved by minimizing a Koenker-Bassett error of the random variable Z_f , provides a road map for the construction of superquantile regression, which is simply achieved by minimizing another measure of error. We start, however, with definitions of quantiles, superquantiles, and corresponding measures of regret and error.

2.1 Definitions

For $\alpha \in [0, 1]$, the α -quantile of a random variable Y with cumulative distribution function F_Y is defined as

$$q_\alpha(Y) := \min\{y \in \mathbb{R} \mid F_Y(y) \geq \alpha\}.$$

Its quantiles are as fundamental to Y as the distribution function, but are problematic to incorporate in risk analysis and optimization due to their lack of coherency as well as computational challenges. Superquantiles have more favorable properties. For $\alpha \in [0, 1)$, the α -superquantile of a random variable Y is defined as

$$\bar{q}_\alpha(Y) := \frac{1}{1-\alpha} \int_{\alpha}^1 q_\beta(Y) d\beta. \quad (1)$$

Since a superquantile is a coherent measure of risk and by the virtue of being an ‘average’ of quantiles is also more stable than a quantile in some sense, it’s well suited for applications. For $\alpha = 1$, we define $\bar{q}_\alpha(Y) := \sup Y$ (the essential supremum). Since $\bar{q}_0(Y) = E[Y]$, we therefore focus on $\alpha \in (0, 1)$ throughout the paper to avoid distractions by these special cases.

In reliability terminology, quantiles and superquantiles correspond to failure and buffered failure probabilities. The *failure probability* of a loss random variable Y is

$$p(Y) := \text{Prob}(Y > 0) = 1 - F_Y(0),$$

which corresponds to

$$p(Y) = 1 - \alpha \text{ with } \alpha \text{ such that } q_\alpha(Y) = 0$$

if there is no probability atom at zero. Analogously to the latter expression, the *buffered failure probability* (see [20]) of a loss random variable Y is defined as

$$\bar{p}(Y) := 1 - \alpha \text{ with } \alpha \text{ such that } \bar{q}_\alpha(Y) = 0. \quad (2)$$

A requirement that $\bar{p}(Y) \leq 1 - \alpha$ is therefore equivalent to the constraint that $\bar{q}_\alpha(Y) \leq 0$. Consequently, in applications with a buffered failure probability constraint on a (conditional) loss random variable $Y(x)$ as well as when the goal is to minimize a superquantile of $Y(x)$ directly, there are needs for estimating $\bar{q}_\alpha(Y(x))$ as a function of $x \in \mathbb{R}^n$. Quantiles and superquantiles are connected through a trade-off formula that leads to quantile regression as discussed next.

2.2 Measures of Regret and Error in Quantile Regression

Both α -quantiles and α -superquantiles, $\alpha \in [0, 1)$, of a loss random variable Y are expressed in terms of an optimization problem involving the quantity

$$\mathcal{V}_\alpha(Y) := \frac{1}{1-\alpha} E[\max\{Y, 0\}], \quad (3)$$

which is a *measure of regret* that quantifies the displeasure with realizations of Y above zero; see [21]. Quantiles and superquantiles then follow as

$$q_\alpha(Y) = \underset{C_0 \in \mathbb{R}}{\operatorname{argmin}} \{C_0 + \mathcal{V}_\alpha(Y - C_0)\} \quad (4)$$

$$\bar{q}_\alpha(Y) = \min_{C_0 \in \mathbb{R}} \{C_0 + \mathcal{V}_\alpha(Y - C_0)\}, \quad (5)$$

where we for simplicity assume that an optimal solution is unique. In general, this may not be the case and, traditionally, the lowest optimal solution has been defined as the quantile.

The expression for $q_\alpha(Y)$ is the essential building block for quantile regression, but since we ultimately would like to go beyond the class of constant functions as candidates for a regression function we need to pass to a *measure of error* $\mathcal{E}_\alpha$ constructed from $\mathcal{V}_\alpha$ by setting

$$\mathcal{E}_\alpha(Y) := \mathcal{V}_\alpha(Y) - E[Y]$$

for any loss random variable Y (with $E[|Y|] < \infty$). A measure of error quantifies the degree of ‘nonzeroness’ in a random variable; see [21]. Direct application of this definition and a recognition that a constant term in an objective function is immaterial with respect to the optimal solution gives that

$$q_\alpha(Y) = \operatorname{argmin}_{C_0 \in \mathbb{R}} \mathcal{E}_\alpha(Y - C_0), \quad (6)$$

where again the set of optimal solutions may not be a singleton and

$$\mathcal{E}_\alpha(Y) = \frac{1}{1-\alpha} E[\max\{Y, 0\}] - E[Y] = E \left[\frac{\alpha}{1-\alpha} \max\{Y, 0\} + \max\{-Y, 0\} \right]$$

is a (scaled) Koenker-Bassett error [16]. Quantile regression centers on computing this argmin with “minimizing the error of $Y - C_0$ over $C_0 \in \mathbb{R}$ ” replaced by “minimizing the error of $Y - f(X)$ over a class of functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ”, often taken to be the affine functions. We view $q_\alpha(Y)$ as the ‘closest’ scalar to the random variable Y under a Koenker-Bassett error.

If our goal simply were to estimate $\bar{q}_\alpha(Y)$ of a loss random variable Y for a given $\alpha \in (0, 1)$, the above expressions would have sufficed, possibly passing to an empirical distribution given by a sample if F_Y is unknown. In the present context, however, connections with the underlying explanatory random vector X and the focus on the ‘approximation’ of Y warrants a parallel development to that of quantile regression centered on a superquantile. In view of the above review of quantile regression, it’s clear that superquantile regression will involve the minimization of some measure of error that returns the superquantile as argmin². The next subsection develops such a measure by first constructing a corresponding measure of regret.

2.3 Measures of Regret and Error in Superquantile Regression

We start this subsection by establishing the finiteness of a superquantile under the assumption that the loss random variable Y has a finite second moment and write $Y \in \mathcal{L}^2(\Omega) := \{Y : \Omega \rightarrow \mathbb{R} \mid E[Y^2] < \infty\}$.

We know from [21] that $\bar{q}_\alpha$ is a convex, positively homogenous, monotonic, and averse³ functional on $\mathcal{L}^2(\Omega)$ for $\alpha \in (0, 1)$. A superquantile is also bounded by [24, Theorem 3], which we repeat here with a different proof. We adopt the notation $\mu(Y) = E[Y]$ and $\sigma^2(Y) = E[(Y - \mu(Y))^2]$.

Proposition 1 *For $Y \in \mathcal{L}^2(\Omega)$ and $\alpha \in (0, 1)$ one has that*

$$\bar{q}_\alpha(Y) \leq \mu(Y) + \frac{1}{\sqrt{1-\alpha}} \sigma(Y). \quad (7)$$

Proof: Suppose that the quantile $q_\alpha(Y)$, viewed as a function of the probability level, is continuous at α . Let I_α be the indicator function of the interval $[q_\alpha(Y), \infty)$ with probability $1 - \alpha$. We then have by the Schwartz inequality that

$$(1 - \alpha) \bar{q}_\alpha(Y - \mu(Y)) = E[(Y - \mu(Y)) I_\alpha] \leq \sqrt{E[(Y - \mu(Y))^2]} \sqrt{E[I_\alpha^2]} = \sigma(Y) \sqrt{1 - \alpha}.$$

Then, since $\bar{q}_\alpha(Y - \mu(Y)) = \bar{q}_\alpha(Y) - \mu(Y)$, the result follows from dividing by $1 - \alpha$. Thus, (7) is valid under the continuity assumption about the quantile, which is true for all but at most countably many α . By continuity of both sides of (7) with respect to α , it must then hold for all

²Classical least-squares regression can be viewed similarly as returning a (conditional) expectation as argmin when minimizing mean-square measure of error, i.e., $E[Y] = \operatorname{argmin}_{C_0 \in \mathbb{R}} E[(Y - C_0)^2]$.

³We recall that a functional $\mathcal{F} : \mathcal{L}^2(\Omega) \rightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ is averse if $\mathcal{F}(X) > E[X]$ for all nonconstant $X \in \mathcal{L}^2(\Omega)$.

$\alpha \in (0, 1)$. □

The measure of regret that serves in the context of superquantile regression is defined for any loss random variable Y and $\alpha \in (0, 1)$ as

$$\bar{\mathcal{V}}_\alpha(Y) := \frac{1}{1-\alpha} \bar{\mathcal{V}}_0(Y), \quad (8)$$

where

$$\bar{\mathcal{V}}_0(Y) := \int_0^1 \max\{0, \bar{q}_\beta(Y)\} d\beta. \quad (9)$$

These expressions appear in [24], which also explains their discovery. Here, we provide an alternative, direct proof of how they lead to a superquantile. We start, however, with two preliminary results and the definition of a corresponding error measure.

Lemma 1 For $Y \in \mathcal{L}^2(\Omega)$,

$$\bar{\mathcal{V}}_0(Y) \leq \sigma(Y) + \max\{0, \mu(Y) + \sigma(Y)\}. \quad (10)$$

Proof: From (7) and (9) we have

$$\bar{\mathcal{V}}_0(Y) \leq \int_0^1 \max\{0, \theta_Y(\beta)\} d\beta \quad \text{for} \quad \theta_Y(\beta) = \mu(Y) + \frac{1}{\sqrt{1-\beta}} \sigma(Y). \quad (11)$$

We consider three cases. In Case 1, we suppose that $\theta_Y(\beta) \geq 0$ for all $\beta \in [0, 1]$. Then the right hand side of (11) is given by

$$\int_0^1 \theta_Y(\beta) d\beta = \mu(Y) + \sigma(Y) \int_0^1 (1-\beta)^{-1/2} d\beta \quad \text{with} \quad \int_0^1 (1-\beta)^{-1/2} d\beta = 2. \quad (12)$$

Therefore, $\bar{\mathcal{V}}_0(Y) \leq \mu(Y) + 2\sigma(Y)$ in Case 1. In Case 2a, we suppose that $\theta_Y(\beta) \leq 0$ for all $\beta \in (0, 1)$. Then obviously $\bar{\mathcal{V}}_0(Y) \leq 0$. Finally, in Case 2b, let $\theta_Y(\beta) < 0$ for some $\beta \in (0, 1)$, but not all. Then necessarily $\sigma(Y) > 0$ and $\mu(Y) \leq -\sigma(Y)$, and $\theta_Y(\beta)$ strictly increases with respect to β . Let $\bar{\alpha}$ be the unique $\beta \in (0, 1)$ with $\theta_Y(\bar{\alpha}) = 0$, namely when

$$\sqrt{1-\bar{\alpha}} = \frac{\sigma(Y)}{-\mu(Y)}. \quad (13)$$

Then we have that

$$\begin{aligned} \int_0^1 \max\{0, \theta_Y(\beta)\} d\beta &= \int_{\bar{\alpha}}^1 \theta_Y(\beta) d\beta = (1-\bar{\alpha})\mu(Y) + \sigma(Y) \int_{\bar{\alpha}}^1 (1-\beta)^{-1/2} d\beta \\ &= (1-\bar{\alpha})\mu(Y) + 2\sigma(Y)\sqrt{1-\bar{\alpha}} \\ &= \frac{\sigma(Y)^2}{\mu(Y)^2} \mu(Y) + 2\sigma(Y) \frac{\sigma(Y)}{-\mu(Y)} \\ &= \frac{\sigma(Y)^2}{-\mu(Y)} \leq \sigma(Y). \end{aligned}$$

Thus, in Case 2b we get $\bar{\mathcal{V}}_0(Y) \leq \sigma(Y)$. The conclusion then follows by putting together the three cases. $\square$

We observe that for $\alpha \in (0, 1)$, $\bar{\mathcal{V}}_\alpha$ is a convex, positively homogeneous, monotonic, and averse functional on $\mathcal{L}^2(\Omega)$, which follows from the properties of the superquantile [21], and by the above result it is also finite, and consequently continuous. A corresponding measure of error is defined for $Y \in \mathcal{L}^2(\Omega)$ by

$$\bar{\mathcal{E}}_\alpha(Y) := \bar{\mathcal{V}}_\alpha(Y) - E[Y] \quad (14)$$

and referred to as a *superquantile error*. Obviously, $\bar{\mathcal{E}}_\alpha$ is also convex and positively homogeneous. It also satisfies the following properties.

Proposition 2 *For any $\alpha \in (0, 1)$ and $Y \in \mathcal{L}^2(\Omega)$, a superquantile error satisfies*

- (a) $\bar{\mathcal{E}}_\alpha(Y) = 0$ when $Y \equiv 0$,
- (b) $\bar{\mathcal{E}}_\alpha(Y) > 0$ when $Y \not\equiv 0$, and
- (c) $\bar{\mathcal{E}}_\alpha(Y) \geq \min\{1, \alpha/(1 - \alpha)\} |E[Y]|$.

Proof: Since $\bar{q}_\beta(0) = 0$ for all $\beta \in [0, 1]$, (a) follows trivially.

Since $\bar{\mathcal{V}}_\alpha$ is averse, we have that for $Y \in \mathcal{L}^2(\Omega)$, $\bar{\mathcal{E}}_\alpha(Y) = \bar{\mathcal{V}}_\alpha(Y) - E[Y] > E[Y] - E[Y] = 0$ when Y is not a constant. To complete part (b), we therefore only need to consider nonzero constants. If Y is a positive constant K , then

$$\frac{1}{1 - \alpha} \int_0^1 \max\{0, \bar{q}_\beta(Y)\} d\beta - E[Y] > \int_0^1 \max\{0, \bar{q}_\beta(Y)\} d\beta - E[Y] = K - E[Y] = 0.$$

If Y is a negative constant K , then

$$\frac{1}{1 - \alpha} \int_0^1 \max\{0, \bar{q}_\beta(Y)\} d\beta - E[Y] = \frac{1}{1 - \alpha} \int_0^1 \max\{0, K\} d\beta - E[Y] = 0 - E[Y] > 0,$$

which completes part (b).

Since $\bar{q}_\beta(Y) \geq E[Y]$ for all $\beta \in [0, 1]$, we have whenever $E[Y] \geq 0$ the bound

$$\frac{1}{1 - \alpha} \int_0^1 \max\{0, \bar{q}_\beta(Y)\} d\beta - E[Y] \geq \frac{1}{1 - \alpha} \int_0^1 \max\{0, E[Y]\} d\beta - E[Y] = \frac{\alpha}{1 - \alpha} E[Y].$$

When $E[Y] < 0$,

$$\frac{1}{1 - \alpha} \int_0^1 \max\{0, \bar{q}_\beta(Y)\} d\beta - E[Y] \geq \frac{1}{1 - \alpha} \int_0^1 \max\{0, E[Y]\} d\beta - E[Y] = -E[Y].$$

Part (c) then follows by combining the two results. $\square$

In view of Proposition 2 and the above discussion, $\bar{\mathcal{E}}_\alpha$ is a *regular* measure of error in the sense of [21].

We are now ready to show that a superquantile is a unique optimal solution of optimization problems involving $\bar{\mathcal{V}}_\alpha$ and $\bar{\mathcal{E}}_\alpha$. As mentioned, the connection between a superquantile and $\bar{\mathcal{V}}_\alpha$ is also reached in Theorem 7 of [24] through different means. The direct proof in the present paper and the connection with a superquantile error are new.

Theorem 1 (*Superquantile as optimal solution*) *For $Y \in \mathcal{L}^2(\Omega)$ and $\alpha \in (0, 1)$,*

$$\bar{q}_\alpha(Y) = \operatorname{argmin}_{C_0 \in \mathcal{R}} \{C_0 + \bar{\mathcal{V}}_\alpha(Y - C_0)\} = \operatorname{argmin}_{C_0 \in \mathcal{R}} \bar{\mathcal{E}}_\alpha(Y - C_0). \quad (15)$$

Proof: Let $\varphi(C) = C + \bar{\mathcal{V}}_\alpha(Y - C)$ and $\psi_\beta(C) = \max\{0, \bar{q}_\beta(Y) - C\}$. These are both convex functions of C , and ψ_β is nonincreasing. We can use the criterion that

$$\bar{C} \in \operatorname{argmin}_C \varphi(C) \iff \varphi'_+(\bar{C}) \geq 0, \varphi'_-(\bar{C}) \leq 0,$$

where, because of the monotonicity of ψ_β ,

$$\begin{aligned} \varphi'_+(C) &= 1 + \frac{1}{1-\alpha} \int_0^1 (\psi_\beta)'_-(C) d\beta, & \varphi'_-(C) &= 1 + \frac{1}{1-\alpha} \int_0^1 (\psi_\beta)'_+(C) d\beta, \\ (\psi_\beta)'_+(C) &= \begin{cases} -1 & \text{if } \bar{q}_\beta(Y) > C, \\ 0 & \text{if } \bar{q}_\beta(Y) \leq C, \end{cases} & (\psi_\beta)'_-(C) &= \begin{cases} -1 & \text{if } \bar{q}_\beta(Y) \geq C, \\ 0 & \text{if } \bar{q}_\beta(Y) < C. \end{cases} \end{aligned}$$

Therefore

$$\int_0^1 (\psi_\beta)'_+(C) d\beta = \int_0^1 (\psi_\beta)'_-(C) d\beta = -(1-\gamma) \text{ for } C = \bar{q}_\gamma(Y),$$

in which case $(\psi_\beta)'(C) = (\psi_\beta)'_+(C) = (\psi_\beta)'_-(C) = 1 - (1-\gamma)/(1-\alpha)$. Thus, $(\psi_\beta)'(C) = 0$ corresponds to $C = \bar{q}_\gamma(Y)$ for $\gamma = \alpha$. Consequently, the first equality of the theorem holds. The second follows directly from (14) and the fact that a constant in an objective function is immaterial with regard to the argmin. $\square$

Being analogous to (4) and (6), the foundations for quantile regression, the expressions (15) provide the path to superquantile regression as developed in the remainder of the paper. In fact, Theorem 1 shows that $\bar{q}_\alpha(Y)$ is the uniquely ‘closest’ scalar to Y in the sense of the superquantile error.

While not the focus here, the optimal objective function value in (15) defines a *measure of risk* (see [24])

$$\bar{\mathcal{R}}_\alpha(Y) := \min_{C_0 \in \mathbf{R}} \{C_0 + \bar{\mathcal{V}}_\alpha(Y - C_0)\} = \bar{q}_\alpha(Y) + \bar{\mathcal{V}}_\alpha(Y - \bar{q}_\alpha(Y))$$

for $Y \in \mathcal{L}^2(\Omega)$ analogously to $\bar{q}_\alpha(Y)$ in (5). A corresponding *measure of deviation*, which quantifies the nonconstancy in a random variable, is given by

$$\bar{\mathcal{D}}_\alpha(Y) := \min_{C_0 \in \mathbf{R}} \bar{\mathcal{E}}_\alpha(Y - C_0) = \bar{\mathcal{R}}_\alpha(Y) - E[Y].$$

We note that parallel to (1) (see [24]), $\bar{\mathcal{R}}_\alpha(Y) = 1/(1-\alpha) \int_\alpha^1 \bar{q}_\beta(Y) d\beta$ and, consequently,

$$\bar{\mathcal{D}}_\alpha(Y) = \frac{1}{1-\alpha} \int_\alpha^1 \bar{q}_\beta(Y) d\beta - E[Y].$$

The measures of regret, error, risk, and deviation $\bar{\mathcal{V}}_\alpha$, $\bar{\mathcal{E}}_\alpha$, $\bar{\mathcal{R}}_\alpha$, and $\bar{\mathcal{D}}_\alpha$, $\alpha \in (0, 1)$, form a family of *risk quadrangles* in the sense of [21] that corresponds to the *statistic* $\bar{q}_\alpha$. The measure of deviation $\bar{\mathcal{D}}_\alpha$ plays a central role in the remainder of the paper as it facilitates simplifications, goodness-of-fit tests, and computational methods.

3 Superquantile Regression

Theorem 1 and the development leading to quantile regression direct us to a new regression methodology that is centered on a superquantile error. The next subsection poses the regression problem, provides its properties, and discusses stability under perturbations. The section ends with a discussion of superquantile tracking.

3.1 Regression Problem

While Theorem 1 shows that the ‘best’ scalar approximation of a random variable Y in the sense of a superquantile error is the corresponding superquantile, we now go beyond the class of constant functions to utilize the connection with an underlying explanatory random vector X . We focus on regression functions of the form

$$f(x) = C_0 + \langle C, h(x) \rangle, \quad C_0 \in \mathbb{R}, C \in \mathbb{R}^m,$$

for a given ‘basis’ function $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$. This class satisfies most practical needs including that of linear regression where $m = n$ and $h(x) = x$. Extensions beyond this class are also possible but not dealt with here.

For any $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $\alpha \in (0, 1)$, we define the *superquantile regression problem*

$$P : \min_{C_0 \in \mathbb{R}, C \in \mathbb{R}^m} \bar{\mathcal{E}}_\alpha(Z(C_0, C)),$$

where

$$Z(C_0, C) := Y - (C_0 + \langle C, h(X) \rangle)$$

is the *error random variable*, whose distribution depends on C_0, C, h , and the joint distribution of (X, Y) . We denote by $\bar{\mathcal{C}} \subset \mathbb{R}^{m+1}$ the set of optimal solutions of P and refer to $(\bar{C}_0, \bar{C}) \in \bar{\mathcal{C}}$ as a *regression vector*.

The objective function $\bar{\mathcal{E}}_\alpha(Z(\cdot, \cdot))$ is well-defined and finite when the distribution of (X, Y) and h is such that $Z(C_0, C) \in \mathcal{L}^2(\Omega)$ for all $C_0 \in \mathbb{R}, C \in \mathbb{R}^m$. A sufficient condition that ensures this property is that $Y, h_1(X), \dots, h_m(X) \in \mathcal{L}^2(\Omega)$ as shown next, where we adopt the notation

$$H = h(X), \quad H_i = h_i(X), \quad i = 1, 2, \dots, m.$$

Lemma 2 *If $Y, H_1, \dots, H_m \in \mathcal{L}^2(\Omega)$, then $Z(C_0, C) \in \mathcal{L}^2(\Omega)$ for all $C_0 \in \mathbb{R}, C \in \mathbb{R}^m$.*

Proof: Let $M < \infty$ be such that $E[Y^2] \leq M$ and $E[H_i^2] \leq M, i = 1, 2, \dots, m$. Since $|\langle C, H \rangle| \leq \|C\| \sum_{i=1}^m |H_i|$ and $\langle C, H \rangle^2 \leq \|C\|^2 \sum_{i=1}^m H_i^2$, we find that $E[|\langle C, H \rangle|] \leq \|C\|mM$ and $E[\langle C, H \rangle^2] \leq \|C\|^2mM$. Consequently,

$$\begin{aligned} E[(Y - C_0 - \langle C, H \rangle)^2] &\leq E[(Y - C_0)^2] + 2|E[(Y - C_0)\langle C, H \rangle]| + E[\langle C, H \rangle^2] \\ &\leq M + 2(\|C\|m^{1/2}M + (M + |C_0|)\|C\|mM) + \|C\|^2mM. \end{aligned} \quad (16)$$

□

In surrogate estimation, $\bar{C}_0 + \langle \bar{C}, h(X) \rangle$, with $(\bar{C}_0, \bar{C}) \in \bar{\mathcal{C}}$, provides the best approximation of Y in the sense of a superquantile error. For example, after having computed $(\bar{C}_0, \bar{C})$, the analysis could proceed with examining the moments, quantiles, and superquantiles of $\bar{C}_0 + \langle \bar{C}, h(X) \rangle$ as surrogates for the corresponding quantities of Y . If X is Gaussian and h is affine, then $\bar{C}_0 + \langle \bar{C}, h(X) \rangle$ is a Gaussian approximation of Y easily examined and utilized in further studies. It may also be of interest to examine $\bar{C}_0 + \langle \bar{C}, h(X) \rangle$ under hypothetical distributions of X .

A direct consequence of the Regression Theorem in [21] (see also Theorem 3.1 in [22]) we obtain that a regression vector can equivalently be determined from a measure of deviation $\bar{\mathcal{D}}_\alpha$.

Proposition 3 *Suppose that $Y, H_1, \dots, H_m \in \mathcal{L}^2(\Omega)$. Then, the set of regression vectors $\bar{\mathcal{C}}$ of P is equivalently obtained as*

$$\bar{\mathcal{C}} = \left\{ (\bar{C}_0, \bar{C}) \in \mathbb{R}^{m+1} \mid \bar{C} \in \underset{C \in \mathbb{R}^m}{\operatorname{argmin}} \bar{\mathcal{D}}_\alpha(Z_0(C)), \quad \bar{C}_0 = \bar{q}_\alpha(Z_0(\bar{C})) \right\},$$

where $Z_0(C) := Y - \langle C, h(X) \rangle$.

Proposition 3 implies computational advantages as the $(m + 1)$ -dimensional optimization problem P is replaced by a problem in m dimensions with a simpler objective function, which we fully utilize in Sections 5 and 6. Moreover, the result also proves beneficial in analysis of regression vectors.

The existence of a regression vector is ensured by the next result, which also provides conditions for uniqueness.

Theorem 2 (*Existence and uniqueness of regression vector*) *If $Y, H_1, \dots, H_m \in \mathcal{L}^2(\Omega)$, then P is a convex problem with a set of optimal solutions $\bar{\mathcal{C}}$ that is nonempty, closed, and convex.*

(a) *$\bar{\mathcal{C}}$ is bounded if and only if the random vector X and the basis function h satisfy the condition that $\langle C, h(X) \rangle$ is not constant unless $C = 0$.*

(b) *If in addition, for every $(C_0, C), (C'_0, C') \in \mathbb{R}^{m+1}$, with $C \neq C'$, there exists a $\beta_0 \in [0, 1)$ such that*

$$0 \leq \bar{q}_\beta(Z(C_0, C) + Z(C'_0, C')) < \bar{q}_\beta(Z(C_0, C)) + \bar{q}_\beta(Z(C'_0, C')) \quad (17)$$

for all $\beta \in [\beta_0, 1)$, then $\bar{\mathcal{C}}$ is a singleton.

Proof. Since $Y \in \mathcal{L}^2(\Omega)$ implies that $\bar{\mathcal{E}}_\alpha(Y) < \infty$ by Lemma 1, we deduce the two first conclusions from Theorem 3.1 in [22]. Hence, we only need to show that $\bar{\mathcal{C}}$ is a singleton.

Suppose for the sake of a contradiction that $(C_0, C), (C'_0, C') \in \bar{\mathcal{C}}$ and $(C_0, C) \neq (C'_0, C')$, with corresponding optimal value $\xi \geq 0$, i.e., $\xi = \bar{\mathcal{E}}_\alpha(Z(C_0, C)) = \bar{\mathcal{E}}_\alpha(Z(C'_0, C'))$. We consider two cases.

First, suppose that $\xi = 0$. By Proposition 2, $Z(C_0, C) = Z(C'_0, C') = 0$ and consequently

$$C_0 + \langle C, H \rangle = C'_0 + \langle C', H \rangle,$$

which implies that $\langle C - C', H \rangle = C'_0 - C_0$. Under the assumption that $\langle C, h(X) \rangle$ is only constant when $C = 0$, we must have that $C - C' = 0$. Then, also $C'_0 - C_0 = 0$ follows, which contradicts the hypothesis that $(C_0, C) \neq (C'_0, C')$.

Second, suppose that $\xi > 0$. If $C = C'$, then a direct consequence of Proposition 3 and the fact that every random variable has a unique superquantile at each probability level, is that also $C_0 = C'_0$, which again contradicts our hypothesis. Consequently, we focus on the case with $C \neq C'$, for which there exists a β_0 such that (17) holds for all $\beta \in [\beta_0, 1)$. Trivially, then

$$\max\{0, \bar{q}_\beta(Z(C_0, C) + Z(C'_0, C'))\} < \max\{0, \bar{q}_\beta(Z(C_0, C))\} + \max\{0, \bar{q}_\beta(Z(C'_0, C'))\}$$

for $\beta \in [\beta_0, 1)$. If $\beta \in (0, 1)$ is such that $\bar{q}_\beta(Z(C_0, C) + Z(C'_0, C')) < 0$, then

$$\max\{0, \bar{q}_\beta(Z(C_0, C) + Z(C'_0, C'))\} \leq \max\{0, \bar{q}_\beta(Z(C_0, C))\} + \max\{0, \bar{q}_\beta(Z(C'_0, C'))\}$$

as the left-hand side vanishes and the right-hand side is nonnegative. Hence,

$$\int_0^1 \max\{0, \bar{q}_\beta(Z(C_0, C) + Z(C'_0, C'))\} d\beta < \int_0^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta + \int_0^1 \max\{0, \bar{q}_\beta(Z(C'_0, C'))\} d\beta$$

and also

$$\bar{\mathcal{E}}_\alpha(Z(C_0, C) + Z(C'_0, C')) < \bar{\mathcal{E}}_\alpha(Z(C_0, C)) + \bar{\mathcal{E}}_\alpha(Z(C'_0, C')). \quad (18)$$

Let

$$(C''_0, C'') = (1/2)(C_0, C) + (1/2)(C'_0, C')$$

and therefore

$$2Z(C''_0, C'') = Z(C_0, C) + Z(C'_0, C').$$

By the optimality of ξ , the positive homogeneity of $\bar{\mathcal{E}}_\alpha$, and (18), we find that

$$2\xi \leq 2\bar{\mathcal{E}}_\alpha(Z(C''_0, C'')) = \bar{\mathcal{E}}_\alpha(2Z(C''_0, C'')) < \bar{\mathcal{E}}_\alpha(Z(C_0, C)) + \bar{\mathcal{E}}_\alpha(Z(C'_0, C')) = 2\xi,$$

which cannot hold. In view of this contradiction, the conclusion follows. $\square$

While Theorem 2 gives a sufficient condition for uniqueness of the regression vector, in general uniqueness cannot be expected. For example, suppose that the random vector (X, Y) , with X scalar valued, has the possible and equally likely realizations $(1, 1)$, $(2, 2)$, and $(3, 1)$. Then, $\bar{q}_\beta(Z_0(C)) = \max\{1 - C, 2 - 2C, 1 - 3C\}$ for $\beta > 2/3$ and $E[Z_0(C)] = 4/3 - 2C$. It's straightforward to show that for $\alpha > 2/3$, any $C \in [-1, 1]$ minimizes $\bar{D}_\alpha(Z_0(\cdot))$. Consequently, in view of Proposition 3, any $C \in [-1, 1]$, with a corresponding $C_0 = \max\{1 - C, 2 - 2C, 1 - 3C\}$, minimizes $\bar{\mathcal{E}}_\alpha(Z(\cdot, \cdot))$ for $\alpha > 2/3$. The minimum error is $2/3$.

A unique regression vector is indeed achieved in the normal case as stated next.

Proposition 4 *Suppose that (H, Y) is normally distributed with positive definite variance-covariance matrix. Then, $\bar{\mathcal{C}}$ is a singleton.*

Proof: Let Σ be the variance-covariance matrix of (H, Y) , with Cholesky decomposition $\Sigma = LL^\top$. For any $\beta \in (0, 1)$ and $C \in \mathbb{R}^m$, $Z_0(C)$ is also normal with mean $\mu(Z_0(C)) = \langle \tilde{C}, E[(H, Y)] \rangle$ and variance $\sigma^2(Z_0(C)) = \langle \tilde{C}, \Sigma \tilde{C} \rangle$, where $\tilde{C} = (-C, 1)$. Thus,

$$\bar{q}_\beta(Z_0(C)) = \mu(Z_0(C)) + k_\beta \sigma(Z_0(C)) = \mu(Z_0(C)) + k_\beta \|L^\top \tilde{C}\|,$$

where $k_\beta = \phi(\Phi^{-1}(\beta))/(1 - \beta)$, with ϕ and Φ being the standard normal probability density and cumulative distribution functions, respectively.

For $C, C' \in \mathbb{R}^m$, with $C \neq C'$, there is no constant $k > 0$ such that $(-C, 1) = k(-C', 1)$. Let $\tilde{C} = (-C, 1)$ and $\tilde{C}' = (-C', 1)$. Since Σ is positive definite, the upper-triangular matrix $L^\top$ is unique and full rank. Consequently, the null space of $L^\top$ contains only the zero vector and $L^\top(\tilde{C} - k\tilde{C}') \neq 0$ for all scalars $k > 0$. Since the triangle inequality for two vectors holds strictly whenever the two vectors cannot be expressed as a positive multiple of each other, we therefore find that

$$\|L^\top \tilde{C} + L^\top \tilde{C}'\| < \|L^\top \tilde{C}\| + \|L^\top \tilde{C}'\|.$$

Now suppose for the sake of a contradiction that $C, C' \in \mathbb{R}^m$ both minimize $\bar{D}_\alpha(Z_0(\cdot))$ and attain the minimum value $\xi \in \mathbb{R}$, but $C \neq C'$. Let

$$C'' = (1/2)C + (1/2)C',$$

$\tilde{C}'' = (-C'', 1)$, and $\gamma_\alpha = \int_\alpha^1 k_\beta d\beta / (1 - \alpha) > 0$. Then,

$$\begin{aligned} \bar{D}_\alpha(Z_0(C'')) &= \frac{1}{1 - \alpha} \int_\alpha^1 \bar{q}_\beta(Z_0(C'')) d\beta - E[Z_0(C'')] \\ &= \mu(Z_0(C'')) + \gamma_\alpha \|L^\top \tilde{C}''\| - \mu(Z_0(C'')) \\ &= \frac{\gamma_\alpha}{2} \|L^\top \tilde{C} + L^\top \tilde{C}'\| \\ &< \frac{\gamma_\alpha}{2} (\|L^\top \tilde{C}\| + \|L^\top \tilde{C}'\|) \\ &= \frac{1}{2} \left(\mu(Z_0(C)) + \gamma_\alpha \|L^\top \tilde{C}\| - \mu(Z_0(C)) \right) + \frac{1}{2} \left(\mu(Z_0(C')) + \gamma_\alpha \|L^\top \tilde{C}'\| - \mu(Z_0(C')) \right) \\ &= \frac{1}{2} (\bar{D}_\alpha(Z_0(C))) + \frac{1}{2} (\bar{D}_\alpha(Z_0(C'))) \\ &= \frac{1}{2} (\xi + \xi) = \xi. \end{aligned}$$

However, this contradicts the optimality of C, C' and we reach the conclusion. $\square$

We next turn to consistency and stability of the regression vector. Of course, the joint distribution of (X, Y) is rarely available in practice and one may need to pass to an approximating empirical distribution generated by a sample. Moreover, perturbations of the ‘true’ distribution of (X, Y) may occur due to measurement errors in the data and other factors. We consider these possibilities and let (X^ν, Y^ν) be a random vector whose joint distribution approximates that of (X, Y) in some sense. For example, (X^ν, Y^ν) may be governed by the empirical distribution generated by an independent and identically distributed sample of size ν from (X, Y) . Presumably, as $\nu \rightarrow \infty$, the approximation of (X, Y) by (X^ν, Y^ν) improves as stated formally below. Regardless of the nature of (X^ν, Y^ν) , we define the *approximate error random variable*

$$Z^\nu(C_0, C) := Y^\nu - C_0 - \langle C, h(X^\nu) \rangle,$$

and the corresponding *approximate superquantile regression problem*

$$P^\nu : \min_{C_0 \in \mathbb{R}, C \in \mathbb{R}^m} \bar{\mathcal{E}}_\alpha(Z^\nu(C_0, C)).$$

The next result shows that as (X^ν, Y^ν) approximates (X, Y) , a regression vector obtained from P^ν approximates one from P , which provides the justification for basing a regression analysis on P^ν . Below, we let $\rightarrow^d$ denote convergence in distribution and

$$H^\nu = h(X^\nu) \text{ and } H_i^\nu = h_i(X^\nu), \quad i = 1, 2, \dots, m.$$

Theorem 3 (*Stability of regression vector*) Suppose that (X^ν, Y^ν) , $\nu = 1, 2, \dots$, and (X, Y) are $n + 1$ -dimensional random vectors such that $(X^\nu, Y^\nu) \rightarrow^d (X, Y)$ and that the basis function h is continuous except possibly on a subset $S \subset \mathbb{R}^n$ with $\text{Prob}(X \in S) = 0$. Moreover, let $H_i, Y \in \mathcal{L}^2(\Omega)$, $\sup_\nu E[(H_i^\nu)^2] < \infty$, $i = 1, 2, \dots, m$, and $\sup_\nu E[(Y^\nu)^2] < \infty$.

If $\{(\bar{C}_0^\nu, \bar{C}^\nu)\}_{\nu=1}^\infty$ is a sequence of optimal solutions of P^ν , with $\alpha \in (0, 1)$, then every accumulation point of that sequence is a regression vector of P .

Proof: Let $(C_0, C) \in \mathbb{R}^{m+1}$ be arbitrary. By the continuous mapping theorem (see for example Theorem 29.2 [3]),

$$Z^\nu(C_0, C) = Y^\nu - C_0 - \langle C, h(X^\nu) \rangle \rightarrow^d Z(C_0, C) = Y - C_0 - \langle C, h(X) \rangle.$$

By the assumed moment conditions, there exists a constant $M < \infty$ that bounds from above the terms

$$\max_i E[|H_i|], \max_i E[(H_i)^2], \sup_{\nu, i} E[|H_i^\nu|], \sup_{\nu, i} E[(H_i^\nu)^2], E[|Y|], E[Y^2], \sup_\nu E[|Y^\nu|], \sup_\nu E[(Y^\nu)^2].$$

In view of Lemma 2 and its proof, we deduce that

$$E[(Y^\nu - C_0 - \langle C, H^\nu \rangle)^2] \leq M + 2(\|C\| m^{1/2} M + (M + |C_0|) \|C\| m M) + \|C\|^2 m M \quad (19)$$

for all ν . Hence, $Z^\nu(C_0, C)$ is uniformly integrable (for fixed C_0, C) and

$$E[Z^\nu(C_0, C)] \rightarrow E[Z(C_0, C)] < \infty; \quad (20)$$

see [3], Theorem 25.12 and its corollary.

By [24, Theorem 4], a sequence of random variables converges in distribution to a random variable if and only if the corresponding α -superquantiles, viewed as functions of the probability level α , converge uniformly on every closed subset of $(0, 1)$. Consequently, $\bar{q}_\beta(Z^\nu(C_0, C)) \rightarrow \bar{q}_\beta(Z(C_0, C))$ uniformly in β on closed subsets of $(0, 1)$. Moreover, since the 0-superquantile coincides with the expectation, (20) implies that $\bar{q}_0(Z^\nu(C_0, C)) \rightarrow \bar{q}_0(Z(C_0, C))$ also holds. These facts and the observation that the superquantile of any random variable is continuous and nondecreasing as a function of the probability level, ensure that for any $\epsilon > 0$ and $\delta \in (0, 1)$, there exists an integer $\nu(\epsilon, \delta)$ such that for all $\nu \geq \nu(\epsilon, \delta)$,

$$\sup_{\beta \in [0, 1-\delta]} |\bar{q}_\beta(Z^\nu(C_0, C)) - \bar{q}_\beta(Z(C_0, C))| \leq \frac{\epsilon}{2(1-\delta)}. \quad (21)$$

Then,

$$\left| \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta - \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \right| \quad (22)$$

$$\leq \int_0^{1-\delta} |\bar{q}_\beta(Z^\nu(C_0, C)) - \bar{q}_\beta(Z(C_0, C))| d\beta \quad (23)$$

$$\leq \int_0^{1-\delta} \frac{\epsilon}{2(1-\delta)} d\beta = \frac{\epsilon}{2} \quad (24)$$

for all $\nu \geq \nu(\epsilon, \delta)$. Following an argument similar to that in Lemma 1, we find that

$$\int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \leq \delta^{1/2} \sigma(Z(C_0, C)) + \max\{0, \delta \mu(Z(C_0, C)) + \delta^{1/2} \sigma(Z(C_0, C))\}. \quad (25)$$

Moreover, the reasoning that lead to (19) also gives

$$|\mu(Z(C_0, C))| \leq M + |C_0| + \|C\|mM. \quad (26)$$

These facts show that there exists a positive constant $\tilde{M} < \infty$ (which depends on C_0 and C) such that $|\mu(Z(C_0, C))|, \sigma(Z(C_0, C)) \leq \tilde{M}$. Hence, from (25), we find that

$$\int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \leq 3\tilde{M}\delta^{1/2}. \quad (27)$$

Let $\epsilon < 12\tilde{M}$ and $\delta_\epsilon = (\epsilon/(12\tilde{M}))^2$. Then, $3\tilde{M}\delta_\epsilon^{1/2} = \epsilon/4$ and

$$\int_{1-\delta_\epsilon}^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \leq \frac{\epsilon}{4}. \quad (28)$$

An identical result holds for $Z^\nu(C_0, C)$. Consequently, for all $\nu \geq \nu(\epsilon, \delta_\epsilon)$,

$$\begin{aligned} & \left| \int_0^1 \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta - \int_0^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \right| \\ & \leq \left| \int_0^{1-\delta_\epsilon} \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta - \int_0^{1-\delta_\epsilon} \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \right| \\ & + \int_{1-\delta_\epsilon}^1 \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta + \int_{1-\delta_\epsilon}^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \\ & \leq \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon. \end{aligned}$$

This fact, (20), and the assumption that (C_0, C) is arbitrary, imply that $\bar{\mathcal{E}}_\alpha(Z^\nu(\cdot, \cdot)) \rightarrow \bar{\mathcal{E}}_\alpha(Z(\cdot, \cdot))$ pointwise on $\mathbb{R}^{m+1}$. Lemma 1 and the above moment assumptions imply that $\bar{\mathcal{E}}_\alpha(Z^\nu(\cdot, \cdot))$ and $\bar{\mathcal{E}}_\alpha(Z(\cdot, \cdot))$ are finite-valued functions. They are also convex, which follows directly from the convexity of $\bar{\mathcal{E}}_\alpha$ on $\mathcal{L}^2(\Omega)$ and the affine form of Z^ν and Z as functions of C_0 and C . Consequently, by Theorem 7.17 in [23], $\bar{\mathcal{E}}_\alpha(Z^\nu(\cdot, \cdot))$ epiconverges to $\bar{\mathcal{E}}_\alpha(Z(\cdot, \cdot))$. The result then follows from Theorem 7.31 in [23]. $\square$

When the approximating problem P^ν is constructed using an independent identically distributed sample of size ν from the distribution of (X, Y) , we obtain the following corollary which follows from the properties of the empirical distribution.

Corollary 1 *Suppose that the basis function h is continuous except possibly on a subset $S \subset \mathbb{R}^n$ with $\text{Prob}(X \in S) = 0$ and that $H_i, Y \in \mathcal{L}^2(\Omega)$, $i = 1, 2, \dots, m$. Moreover, let (X^ν, Y^ν) be distributed according to the empirical distribution generated by an independent and identically distributed sample of size ν from the distribution of (X, Y) . Then, the conclusion of Theorem 3 holds.*

We next examine the rate of convergence of regression vectors obtained from the approximate problem P^ν to those of P corresponding to the ‘true’ distribution. Quantification of the stability of the set of optimal solutions of an optimization problem under perturbations depends on a ‘growth condition’ of the problem, which is difficult to quantify for P ; see [23, Section 7J]. Consequently, we focus on the better behaved ϵ -regression vectors of P defined for $\epsilon > 0$ as

$$\bar{\mathcal{C}}_\epsilon := \left\{ (C_{0,\epsilon}, C_\epsilon) \in \mathbb{R}^{m+1} \mid \bar{\mathcal{E}}_\alpha(Z(C_{0,\epsilon}, C_\epsilon)) \leq \min_{C_0 \in \mathbb{R}, C \in \mathbb{R}^m} \bar{\mathcal{E}}_\alpha(Z(C_0, C)) + \epsilon \right\},$$

with an analogous definition of the ϵ -regression vectors of P^ν denoted by $\bar{\mathcal{C}}_\epsilon^\nu$. The rate with which $\bar{\mathcal{C}}_\epsilon^\nu$ tends to $\bar{\mathcal{C}}_\epsilon$ depends, naturally, on the rate with which (X^ν, Y^ν) , underlying P^ν , tends to (X, Y) of P in some sense. Before we make a precise statement, we introduce a convenient notion of distances between any two nonempty sets $A, B \subset \mathbb{R}^{m+1}$. For $\rho \geq 0$, let

$$\hat{d}_\rho(A, B) := \inf\{\eta \geq 0 \mid A \cap \rho \mathcal{B} \subset B + \eta \mathcal{B}, B \cap \rho \mathcal{B} \subset A + \eta \mathcal{B}\},$$

where $\mathcal{B}$ is the Euclidean ball in $\mathbb{R}^{m+1}$ with unit radius and center at the origin. Roughly, $\hat{d}_\rho(A, B)$ is the smallest amount the sets need to be ‘enlarged’ to ensure they contain the other one, with an exclusive focus on points no further from the origin than ρ . This restriction facilitates the treatment of unbounded sets.

As we see next, the rate of convergence is directly related to the rate with which the random vector

$$\Delta^\nu := (H^\nu - H, Y^\nu - Y),$$

describing the approximation error, tends to zero.

Theorem 4 *(Rate of convergence of regression vector) Suppose that (X^ν, Y^ν) , $\nu = 1, 2, \dots$, and (X, Y) are $n + 1$ -dimensional random vectors generating P^ν and P , respectively. Moreover, let $H_i, Y \in \mathcal{L}^2(\Omega)$, $\sup_\nu E[(H_i^\nu)^2] < \infty$, $i = 1, 2, \dots, m$, and $\sup_\nu E[(Y^\nu)^2] < \infty$. Let $\rho_0 > 0$ be such that $\rho_0 \mathcal{B} \cap \bar{\mathcal{C}} \neq \emptyset$ and $\rho_0 \mathcal{B} \cap \bar{\mathcal{C}}^\nu \neq \emptyset$.*

Then, for $\rho > \rho_0$, there exist positive constants k_1, k_2 , and k_3 (dependent on ρ) such that for any $\epsilon > 0$ and $\nu = 1, 2, \dots$,

$$\hat{d}_\rho(\bar{\mathcal{C}}_\epsilon^\nu, \bar{\mathcal{C}}_\epsilon) \leq \left(1 + \frac{4\rho}{\epsilon}\right) \left[E[\|\Delta^\nu\|] \left(k_1 \max \left\{ 0, \log \left(\frac{1}{E[\|\Delta^\nu\|]} \right) \right\} + k_2 \right) + k_3 E[\|\Delta^\nu\|] \right]$$

whenever $E[\|\Delta^\nu\|] > 0$ and $\hat{d}_\rho(\bar{\mathcal{C}}_\epsilon^\nu, \bar{\mathcal{C}}_\epsilon) = 0$ otherwise.

Proof: By Theorem 3(a) of [24], for $\beta \in [0, 1)$,

$$\begin{aligned} |\bar{q}_\beta(Z^\nu(C_0, C)) - \bar{q}_\beta(Z(C_0, C))| &\leq \frac{1}{1-\beta} E[|Z^\nu(C_0, C) - Z(C_0, C)|] \\ &= \frac{1}{1-\beta} E[|\langle \tilde{C}, \Delta^\nu \rangle|] \\ &\leq \frac{1}{1-\beta} \|\tilde{C}\| E[\|\Delta^\nu\|], \end{aligned} \quad (29)$$

where $\tilde{C} = (-C, 1)$. Then, for $\delta \in (0, 1)$,

$$\begin{aligned} &\left| \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta - \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \right| \\ &\leq \int_0^{1-\delta} |\bar{q}_\beta(Z^\nu(C_0, C)) - \bar{q}_\beta(Z(C_0, C))| d\beta \\ &\leq \|\tilde{C}\| E[\|\Delta^\nu\|] \int_0^{1-\delta} \frac{1}{1-\beta} d\beta = -\|\tilde{C}\| E[\|\Delta^\nu\|] \log \delta. \end{aligned} \quad (30)$$

Let $\rho > \rho_0$ and M be an upper bound on first and second moments of $|H_i|$, $|H_i^\nu|$, $|Y|$, and $|Y^\nu|$ as in the proof of Theorem 3. Then, for $\|(C_0, C)\| \leq \rho$, it follows by (26) that

$$|\mu(Z(C_0, C))| \leq M + \rho + \rho m M$$

and by (16) that

$$\sigma(Z(C_0, C)) \leq (M + 2(\rho m^{1/2} M + (M + \rho) \rho m M) + \rho^2 m M)^{1/2},$$

with identical bounds for $|\mu(Z^\nu(C_0, C))|$ and $\sigma(Z^\nu(C_0, C))$. Let M_ρ be the larger of the two previous right-hand sides.

By (25), analogously to (27), we have that for $\|(C_0, C)\| \leq \rho$,

$$\int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \leq 3M_\rho \delta^{1/2} \quad (31)$$

and similarly with $Z(C_0, C)$ replaced by $Z^\nu(C_0, C)$.

We also find that for $\|(C_0, C)\| \leq \rho$,

$$|E[Z^\nu(C_0, C)] - E[Z(C_0, C)]| = |\langle \tilde{C}, E[\Delta^\nu] \rangle| \leq \|\tilde{C}\| \|E[\Delta^\nu]\| \leq (1 + \rho) \|E[\Delta^\nu]\|. \quad (32)$$

Then, collecting the results of (30), (31), and (32), we obtain that for $\|(C_0, C)\| \leq \rho$,

$$\begin{aligned} &|\bar{\mathcal{E}}_\alpha(Z^\nu(C_0, C)) - \bar{\mathcal{E}}_\alpha(Z(C_0, C))| \\ &\leq \left| \int_0^1 \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta - \int_0^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \right| + |E[Z^\nu(C_0, C)] - E[Z(C_0, C)]| \\ &\leq \left| \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta - \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \right| \\ &\quad + \int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z^\nu(C_0, C))\} d\beta + \int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z(C_0, C))\} d\beta \\ &\quad + |E[Z^\nu(C_0, C)] - E[Z(C_0, C)]| \\ &\leq -(1 + \rho) E[\|\Delta^\nu\|] \log \delta + 6M_\rho \delta^{1/2} + (1 + \rho) \|E[\Delta^\nu]\|. \end{aligned} \quad (33)$$

We next determine the choice of $\delta \in (0, 1)$ that minimizes the previous bound and consider two cases. First, if

$$0 < k_\rho(E[\|\Delta^\nu\|])^2 < 1,$$

with

$$k_\rho := (2(1 + \rho)/(6M_\rho))^2,$$

then differentiation gives that the bound is minimized with $\delta = k_\rho(E[\|\Delta^\nu\|])^2$. Second, if

$$k_\rho(E[\|\Delta^\nu\|])^2 \geq 1,$$

then

$$M_\rho \leq 4(1 + \rho)E[\|\Delta^\nu\|]/6$$

and the bound

$$\begin{aligned} & -(1 + \rho)E[\|\Delta^\nu\|] \log \delta + 6M_\rho \delta^{1/2} + (1 + \rho)\|E[\Delta^\nu]\| \\ \leq & -(1 + \rho)E[\|\Delta^\nu\|] \log \delta + 4(1 + \rho)E[\|\Delta^\nu\|]\delta^{1/2} + (1 + \rho)\|E[\Delta^\nu]\| \end{aligned}$$

for any $\delta \in (0, 1)$. Consequently, combining the two cases, there exist constants k_1 , k_2 , and k_3 (which depend on ρ), such that for $\|(C_0, C)\| \leq \rho$,

$$\begin{aligned} & |\bar{\mathcal{E}}_\alpha(Z^\nu(C_0, C)) - \bar{\mathcal{E}}_\alpha(Z(C_0, C))| \\ \leq & k_1 E[\|\Delta^\nu\|] \max \left\{ 0, \log \left(\frac{1}{E[\|\Delta^\nu\|]} \right) \right\} + k_2 E[\|\Delta^\nu\|] + k_3 \|E[\Delta^\nu]\| \\ \leq & E[\|\Delta^\nu\|] \left(k_1 \max \left\{ 0, \log \left(\frac{1}{E[\|\Delta^\nu\|]} \right) \right\} + k_2 \right) + k_3 \|E[\Delta^\nu]\|. \end{aligned}$$

Direct application of Example 7.62 and Theorem 7.69 of [23] then yields the conclusion for $E[\|\Delta^\nu\|] > 0$, where the additional coefficient $(1 + 4\rho/\epsilon)$ originates in that theorem. Finally, if $E[\|\Delta^\nu\|] = 0$, then, in view of (29) and the fact that this implies that $\|E[\Delta^\nu]\| = 0$, we find that for $\|(C_0, C)\| \leq \rho$,

$$|\bar{\mathcal{E}}_\alpha(Z^\nu(C_0, C)) - \bar{\mathcal{E}}_\alpha(Z(C_0, C))| = 0.$$

The final conclusion then follows by again invoking Example 7.62 and Theorem 7.69 of [23]. $\square$

Theorem 4 shows that the distance between $\bar{\mathcal{C}}_\epsilon^\nu$ and $\bar{\mathcal{C}}_\epsilon$ is almost proportional to $E[\|\Delta^\nu\|]$, but with a minor correction by a logarithmic term. If the approximation (X^ν, Y^ν) is caused by measurement errors of magnitude $1/\nu$, i.e., the absolute value of each component of $(X^\nu - X, Y^\nu - Y)$ is no greater than $1/\nu$ almost surely, then $E[\|\Delta^\nu\|] \leq \sqrt{m+1}/\nu$ and the expressions can be simplified. For $\xi > 0$, $\log x \leq x^\xi$ for sufficiently large $x \in \mathbb{R}$. Consequently, for any $\xi \in (0, 1)$ and sufficiently large ν ,

$$\hat{d}_\rho(\bar{\mathcal{C}}_\epsilon^\nu, \bar{\mathcal{C}}_\epsilon) \leq \left(1 + \frac{4\rho}{\epsilon}\right) \frac{k}{\nu^{1-\xi}},$$

where $k > 0$ can be determined from k_1, k_2, k_3 , and m . That is, the Euclidean distance between an ϵ -regression vector of P^ν to one of P is $O(\nu^{\xi-1})$ for $\xi \in (0, 1)$ arbitrarily close to zero.

3.2 Superquantile Tracking

We next turn to the situation where we seek to estimate $\bar{q}_\alpha(Y(x))$ for $x \in \mathbb{R}^n$, or a subset thereof, with the goal of eventually minimizing, at least approximately, $\bar{q}_\alpha(Y(x))$ by a judicious choice of x . Of course, with incomplete knowledge about the distributions of $Y(x)$ this is a difficult task that can be achieved only approximately. For example, there is no guarantee that a regression function $f = \bar{C}_0 + \langle \bar{C}, h(\cdot) \rangle$, with $(\bar{C}_0, \bar{C}) \in \bar{\mathcal{C}}$ obtained by solving P using $\alpha \in (0, 1)$, tracks $\bar{q}_\alpha(Y(x))$, i.e., $f(x) = \bar{q}_\alpha(Y(x))$ for all $x \in \mathbb{R}^n$. The hope of such ‘exact’ tracking becomes even less realistic when P must be replaced by an approximation P^ν as typically required in practice. However, ‘local’ tracking is possible, at least approximately, with an appropriate weighing of the data available as we discuss next.

We consider the situation where there is a sample of $Y(x)$ for a set of x , but the sample is not large enough to allow pointwise estimation of $\bar{q}_\alpha(Y(x))$ for every x of interest. There may even be no x for which there are multiple samples of $Y(x)$. Concentrating on a particular $\hat{x} \in \mathbb{R}^n$, we hope to estimate $\bar{q}_\alpha(Y(\hat{x}))$ by using samples from $Y(x)$ for x near $\hat{x}$, weighted appropriately. The weights should be nonnegative, sum to one, and can be thought of as an artificially constructed probability distribution associated with the sample. Specifically, suppose that $x_i, i = 1, \dots, \nu$, are the points where the sample is observed and $y_i, i = 1, \dots, \nu$, are the corresponding realizations of $Y(x_i)$. When estimating a superquantile at $\hat{x}$, we put more ‘trust’ on sample points taken near $\hat{x}$ and consequently the weight of (x_i, y_i) may be inversely proportional to $\|x_i - \hat{x}\|$, with an appropriate adjustment if $\hat{x}$ coincides with an x_i .

A justification for the approach follows directly from Theorem 3 through the next proposition.

Proposition 5 *Suppose that the assumptions of Theorem 3 hold and that the probability distribution of (X, Y) is degenerate at $\hat{x} \in \mathbb{R}^{n+1}$ in the sense that $\text{Prob}((X, Y) \leq (x, y)) = \varphi(y)$, for all $y \in \mathbb{R}$ and $x \geq \hat{x}$, where $\varphi(y) = \text{Prob}(Y(\hat{x}) \leq y)$, and $\text{Prob}((X, Y) \leq (x, y)) = 0$ otherwise. If $\{(\bar{C}_0^\nu, \bar{C}^\nu)\}_{\nu=1}^\infty$ is a sequence of optimal solutions of P^ν , with $\alpha \in (0, 1)$, then along every convergent subsequence we have that $\bar{C}_0^\nu + \langle \bar{C}^\nu, h(\hat{x}) \rangle$ tends to $\bar{q}_\alpha(Y(\hat{x}))$.*

Proof. For the given degenerate distribution of (X, Y) , $C_0 + \langle C, h(X) \rangle = C_0 + \langle C, h(\hat{x}) \rangle$ almost surely. Consequently, P reduces to the error minimization problem of Theorem 1 and $\bar{C}_0 + \langle \bar{C}, h(\hat{x}) \rangle = \bar{q}_\alpha(Y(\hat{x}))$ for every $(\bar{C}_0, \bar{C}) \in \bar{\mathcal{C}}$. The conclusion then follows from Theorem 3. $\square$

Suppose that the weights of (x_i, y_i) , $i = 1, 2, \dots, \nu$, in the above construction are chosen to approximate the degenerate distribution of Proposition 5, for example by setting them inversely proportional to $\|x_i - \hat{x}\|$. Then, in view of Proposition 5, a solution of P^ν , constructed using those weights as an artificial probability distribution for (X^ν, Y^ν) , leads to an approximation of the considered superquantile at $\hat{x}$. Of course, this procedure can be repeated for different points $\hat{x}$ to generate a ‘global’ assessment of $\bar{q}_\alpha(Y(x))$ as a function of x and eventually facilitate optimization over x . Moreover, the process can be repeated with new or augmented sample points in a straightforward manner. In a situation where a sample is not fully randomly generated but x -points are determined by an analyst, the approach may even motivate scattering those points near a point of interest $\hat{x}$ instead of concentrating them all at $\hat{x}$ exactly. The former approach certainly results in a better ‘global’ understanding of a superquantile as a function of x , but may prove to be a more economical route to estimate a superquantile at $\hat{x}$ too. We examine this situation numerically in Section 6.

4 Validation Analysis

Regression modeling must be associated with means of assessing the goodness-of-fit of a computed regression vector. In least-squares regression, the *coefficient of determination*

$$R^2 = 1 - \frac{SS_{\text{Res}}}{SS_{\text{T}}},$$

where SS_{Res} denotes the residual sum of squares and SS_{T} the total sum of squares, provides a means for such an assessment. While R^2 can't be relied on exclusively, it provides an indication of the goodness of fit that is easily extended to the present context of superquantile regression. In our notation,

$$R^2 = 1 - \frac{E[Z(C_0, C)^2]}{\sigma^2(Y)}, \quad (34)$$

and similarly when passing to an approximate random vector (X^ν, Y^ν) . From Example 1' in [21], we know that the numerator in (34) is an error measure applied to $Z(C_0, C)$ and that it corresponds to the deviation measure $\sigma^2(\cdot)$. Moreover, the minimization of that error of $Z(C_0, C)$ results in the least-squares regression vector. According to [21], these error and deviation measures are in correspondence and belong to a 'risk quadrangle' that yields the expectation as its statistic. This observation motivates the following definition of a coefficient of determination for superquantile regression model.

Definition 1 *In superquantile regression, the coefficient of determination of a regression vector $(C_0, C) \in \mathbb{R}^{m+1}$ is given by*

$$\bar{R}_\alpha^2(C_0, C) := 1 - \frac{\bar{\mathcal{E}}_\alpha(Z(C_0, C))}{\bar{\mathcal{D}}_\alpha(Y)}. \quad (35)$$

In fact, a similar definition can be formulated for any generalized regression consisting of minimizing an error of Z_f , with then another measure of error in the numerator and a corresponding deviation measure, in the sense of [21], in the denominator. As in the classical case, higher values of $\bar{R}_\alpha^2$ are better, at least in some sense. However, $\bar{R}_\alpha^2 \leq 1$, which is apparent from the nonnegativity of the error and deviation measures. Indeed, P aims to minimize the error of $Z(C_0, C)$ by wisely selecting the regression vector (C_0, C) and thereby also maximizes $\bar{R}_\alpha^2$. The error is 'normalized' with the overall 'nonconstancy' in Y as measured by its deviation measure to more easily allow for comparison of coefficients of determination across data sets.

It's possible to obtain large coefficients of determination by adding explanatory terms to a regression model, i.e., increasing m , but without necessarily achieving a more useful model. Hence, it's usual in least-squares regression to also evaluate an adjusted coefficient of determination that penalizes any term added to the model that doesn't reduce variability substantially. This quantity only increases if a new term reduces $SS_{\text{Res}}/(\nu - m)$ as seen by the definition

$$R_{\text{Adj}}^2 = 1 - \frac{SS_{\text{Res}}/(\nu - m)}{SS_{\text{T}}/(\nu - 1)}, \quad (36)$$

where ν is the number of observations. Naturally, then, we define an adjusted coefficient of determination for superquantile regression similarly in the case where the distribution of (X, Y) has a finite support of cardinality ν .

Definition 2 *In superquantile regression, the adjusted coefficient of determination of a regression vector $(C_0, C) \in \mathbb{R}^{m+1}$ is given by*

$$\bar{R}_{\alpha, \text{Adj}}^2(C_0, C) := 1 - \frac{\bar{\mathcal{E}}_\alpha(Z(C_0, C))/(\nu - m)}{\bar{\mathcal{D}}_\alpha(Y)/(\nu - 1)}. \quad (37)$$

Again, similar expressions are available for other generalized regression techniques.

5 Computational Methods

The computational task of carrying out superquantile regression consists of solving the convex optimization problem P , or in practice the approximate problem P^ν due to incomplete distributional information and other sources of approximations. In this section, we describe convenient means for solving P^ν when (X^ν, Y^ν) has a discrete joint distribution with ν possible realizations. Regardless of the distribution of (X^ν, Y^ν) , a reformulation of P^ν in terms of the deviation measure $\bar{D}_\alpha$ is beneficial. In view of Proposition 3, the task of determining a regression vector $(\bar{C}_0^\nu, \bar{C}^\nu)$ reduces to that of minimizing $\bar{D}_\alpha(Z_0^\nu(\cdot))$, setting $\bar{C}^\nu$ equal to an optimal solution, and then setting $\bar{C}_0^\nu = \bar{q}_\alpha(Z_0^\nu(\bar{C}^\nu))$. Since it's straightforward to compute every superquantile of a random variable with a discrete probability distribution, we focus on the minimization problem, which takes the following form after writing out the expression for the deviation measure in this case

$$D^\nu : \min_{C \in \mathbb{R}^m} \frac{1}{1-\alpha} \int_\alpha^1 \bar{q}_\beta(Z_0^\nu(C)) d\beta - E[Z_0^\nu(C)].$$

The next subsections describe two computational methods for solving D^ν when the distribution of (X^ν, Y^ν) is discrete.

5.1 Analytical Integration

While one might at first get the impression that numerical integration is required in solving D^ν , this may not actually be needed as shown next. Suppose that (X^ν, Y^ν) has a discrete distribution with support (x^j, y^j) , $j = 1, 2, \dots, \nu$, and $\text{Prob}((X^\nu, Y^\nu) = (x^j, y^j)) = 1/\nu$ for $j = 1, 2, \dots, \nu$. This is the case typically encountered in applications, where (x^j, y^j) , $j = 1, 2, \dots, \nu$, is the data assumed to be equally likely to occur. We then obtain significant simplifications in D^ν .

For any fixed $C \in \mathbb{R}^m$, the cumulative distribution function of $Z_0^\nu(C)$ is a piecewise constant function with at most ν steps. The range of the distribution function is $\{0, 1/\nu, 2/\nu, \dots, 1\}$ or a subset thereof. By partitioning the integral over β in D^ν according to this range, accounting for the fact that the integral starts at α , the problem can in this case be written as

$$\min_{C \in \mathbb{R}^m} \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu} \int_{\beta_{i-1}}^{\beta_i} \bar{q}_\beta(Z_0^\nu(C)) d\beta - E[Z_0^\nu(C)], \quad (38)$$

where $\nu_\alpha := \lceil \nu\alpha \rceil$, with $\lceil a \rceil$ being the smallest integer no smaller than $a \in \mathbb{R}$, $\beta_{\nu_\alpha-1} = \alpha$, and $\beta_i = i/\nu$, for $i = \nu_\alpha, \nu_\alpha + 1, \dots, \nu$. In view of (4) and (5),

$$\begin{aligned} \bar{q}_\beta(Z_0^\nu(C)) &= \min_{U_\beta \in \mathbb{R}} U_\beta + \frac{1}{1-\beta} E[\max\{Z_0^\nu(C) - U_\beta, 0\}] \\ &= q_\beta(Z_0^\nu(C)) + \frac{1}{1-\beta} E[\max\{Z_0^\nu(C) - q_\beta(Z_0^\nu(C)), 0\}] \end{aligned} \quad (39)$$

for each $\beta \in [0, 1)$. However, the special piecewise-constant structure of the cumulative distribution function of $Z_0^\nu(C)$ implies that $q_\beta(Z_0^\nu(C))$ is constant as a function of β on (β_{i-1}, β_i) for every $i = \nu_\alpha, \nu_\alpha + 1, \dots, \nu$. Consequently, U_β , $\beta \in (\alpha, 1)$ in (39) can be replaced by a finite number of variables so that (38) takes the form

$$\min_{C \in \mathbb{R}^m} \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu} \int_{\beta_{i-1}}^{\beta_i} \min_{U_i \in \mathbb{R}} \left(U_i + \frac{1}{1-\beta} E[\max\{Z_0^\nu(C) - U_i, 0\}] \right) d\beta - E[Z_0^\nu(C)].$$

The last integral simplifies further since for $\beta \in (\beta_{\nu-1}, \beta_\nu) = (1 - 1/\nu, 1)$,

$$\bar{q}_\beta(Z_0^\nu(C)) = M(C) := \max_{j=1,2,\dots,\nu} y^j - \langle C, x^j \rangle.$$

Consequently, (38) takes the form

$$\min_{C \in \mathbb{R}^m} \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu-1} \int_{\beta_{i-1}}^{\beta_i} \min_{U_i \in \mathbb{R}} \left(U_i + \frac{1}{1-\beta} E[\max\{Z_0^\nu(C) - U_i, 0\}] \right) d\beta + \frac{M(C)}{\nu(1-\alpha)} - E[Z_0^\nu(C)].$$

The order of minimization is immaterial and we can equivalently consider

$$\min_{C \in \mathbb{R}^m, U \in \mathbb{R}^{\nu-\nu_\alpha}} \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu-1} \int_{\beta_{i-1}}^{\beta_i} \left(U_i + \frac{1}{1-\beta} E[\max\{Z_0^\nu(C) - U_i, 0\}] \right) d\beta + \frac{M(C)}{\nu(1-\alpha)} - E[Z_0^\nu(C)],$$

where we let $U = (U_{\nu_\alpha}, U_{\nu_\alpha+1}, \dots, U_{\nu-1})$. For $i = \nu_\alpha, \nu_\alpha + 1, \dots, \nu - 1$, we define

$$a_i := \int_{\beta_{i-1}}^{\beta_i} \frac{1}{1-\beta} d\beta = \log(1 - \beta_{i-1}) - \log(1 - \beta_i).$$

Using this notation, (38) simplifies further to

$$\begin{aligned} \min_{C \in \mathbb{R}^m, U \in \mathbb{R}^{\nu-\nu_\alpha}} \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu-1} (\beta_i - \beta_{i-1}) U_i &+ \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu-1} E[\max\{Z_0^\nu(C) - U_i, 0\}] a_i \\ &+ \frac{M(C)}{\nu(1-\alpha)} - E[Z_0^\nu(C)]. \end{aligned}$$

By introducing another set of auxiliary variables and using the standard transcription technique for handling max-functions, we reach the linear program

$$\begin{aligned} D_{LP}^\nu : \quad & \min_{C, U, V, W} \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu-1} (\beta_i - \beta_{i-1}) U_i &+ \frac{1}{\nu(1-\alpha)} \sum_{i=\nu_\alpha}^{\nu-1} \sum_{j=1}^{\nu} a_i V_{ij} \\ & + \frac{1}{\nu(1-\alpha)} W - \frac{1}{\nu} \sum_{j=1}^{\nu} (y^j - \langle C, h(x^j) \rangle) \\ \text{s.t.} \quad & y^j - \langle C, h(x^j) \rangle - U_i \leq V_{ij}, \quad i = \nu_\alpha, \dots, \nu-1, j = 1, \dots, \nu \\ & 0 \leq V_{ij}, \quad i = \nu_\alpha, \dots, \nu-1, j = 1, \dots, \nu \\ & y^j - \langle C, h(x^j) \rangle \leq W, \quad j = 1, \dots, \nu \\ & C \in \mathbb{R}^m \\ & U = (U_{\nu_\alpha}, \dots, U_{\nu-1}) \in \mathbb{R}^{\nu-\nu_\alpha} \\ & V = (V_{\nu_\alpha,1}, \dots, V_{\nu-1,\nu}) \in \mathbb{R}^{(\nu-\nu_\alpha)\nu} \\ & W \in \mathbb{R}. \end{aligned}$$

This equivalent reformulation of D^ν involves $m + (\nu - \nu_\alpha)(\nu + 1) + 1$ variables and $2(\nu - \nu_\alpha)\nu + \nu$ inequality constraints. While $\nu_\alpha = \lceil \nu\alpha \rceil$ may be relatively close to ν in practice, the linear program could become large-scaled when ν is large and decomposition algorithms may be needed.

Alternatively, we consider next a numerical integration-based scheme that avoids some auxiliary variables and constraints, and also handles the situation when the distribution of (X^ν, Y^ν) is not uniformly discrete.

5.2 Numerical Integration

The integral in D^ν is easily approximated by standard numerical integration schemes. Suppose that the interval $[\alpha, 1]$ is divided into μ subintervals, where $\alpha \leq \beta_0 < \beta_1 < \dots < \beta_{\mu-1} < \beta_\mu \leq 1$ and $w_i \geq 0, i = 0, 1, \dots, \mu$, are factors specific to the integration scheme. An approximation of D^ν then takes the form

$$D^{\nu, \mu} : \min_{C \in \mathbb{R}^m} \frac{1}{1 - \alpha} \sum_{i=0}^{\mu} w_i \bar{q}_{\beta_i}(Z_0^\nu(C)) - E[Z_0^\nu(C)].$$

For large μ , an optimal solution of $D^{\nu, \mu}$ is close to that of D^ν , as seen next, under conditions that are satisfied by essentially all commonly used numerical integration schemes.

Proposition 6 *Suppose that for any continuous function $g : [\alpha, 1] \rightarrow \mathbb{R}$, a numerical integration scheme with discretization points $\alpha \leq \beta_0 < \beta_1 < \dots < \beta_{\mu-1} < \beta_\mu \leq 1$ and factors $w_i \geq 0, i = 0, 1, \dots, \mu$, satisfies*

$$\left| \sum_{i=0}^{\mu} w_i g(\beta_i) - \int_{\alpha}^1 g(\beta) d\beta \right| \rightarrow 0$$

as $\mu \rightarrow \infty$. Let $\{\bar{C}^{\nu, \mu}\}_{\mu=1}^{\infty}$ be a sequence of optimal solutions of $D^{\nu, \mu}$ under this numerical integration scheme. Then, every accumulation point of $\{\bar{C}^{\nu, \mu}\}_{\mu=1}^{\infty}$ is an optimal solution of D^ν .

Proof: For any $C \in \mathbb{R}^m$, $\bar{q}_\beta(Z_0^\nu(C))$ is finite and continuous as a function of β . Consequently, the assumption on the numerical integration scheme applies and the objective function of $D^{\nu, \mu}$ converges pointwise to that of D^ν , as $\mu \rightarrow \infty$. The objective functions are also finite and convex in C , which follows directly from the convexity of $\bar{q}_\alpha$ on $\mathcal{L}^2(\Omega)$ and the affine form of Z_0^ν as a function of C . Consequently, by Theorem 7.17 in [23], the objective function of $D^{\nu, \mu}$ epiconverges to that of D^ν and the conclusion follows from Theorem 7.31 in [23]. $\square$

While specialized solvers such as Portfolio Safeguard [1] handle $D^{\nu, \mu}$ directly with little difficulty under many circumstances, the problem is typically nonsmooth and standard nonlinear programming solvers may fail. However, following a simple reformulation of $D^{\nu, \mu}$, utilizing (5), yields the following equivalent linear program, where we assume for convenience that $\beta_\mu < 1$:

$$\begin{aligned} \min_{C, U, V} \quad & \frac{1}{1 - \alpha} \sum_{i=0}^{\mu} w_i \left(U_i + \frac{1}{1 - \beta_i} \sum_{j=1}^{\nu} p^j V_{ij} \right) - \sum_{j=1}^{\nu} p^j (y^j - \langle C, h(x^j) \rangle) \\ \text{s.t.} \quad & y^j - \langle C, h(x^j) \rangle - U_i \leq V_{ij}, \quad i = 0, 1, \dots, \mu, j = 1, \dots, \nu \\ & 0 \leq V_{ij}, \quad i = 0, 1, \dots, \mu, j = 1, \dots, \nu \\ & C \in \mathbb{R}^m \\ & U = (U_0, U_1, \dots, U_\mu) \in \mathbb{R}^{\mu+1} \\ & V = (V_{0,1}, \dots, V_{\mu,\nu}) \in \mathbb{R}^{(\mu+1)\nu}. \end{aligned}$$

If $\beta_\mu = 1$, then a straightforward modification is required based on the fact that $\bar{q}_1(Z_0^\nu(C)) = \max_{j=1,2,\dots,\nu} y^j - \langle C, x^j \rangle$. The linear program consists of $m + \mu + 1 + \nu(\mu + 1)$ variables and $2\nu(\mu + 1)$ constraints, which may be substantially less than what follows from the analytical integration approach for large ν . In practice, we find that a moderately large μ suffices as shown in the next section.

6 Numerical Examples

In this section, we illustrate superquantile regression in three numerical examples. The first example is artificially constructed, with known conditional superquantiles. The second example is an instance from the uncertainty quantification literature. The third example arises in investment analysis. Computations are mostly carried out in Matlab version 7.14 on a 2.26 GHz laptop with 8.0 GB of RAM using Portfolio Safeguard [1] with VAN as the optimization solver for $D^{\nu,\mu}$. When solving D_{LP}^{ν} we employ GAMS version 23.7 with the CPLEX 12.3 solver on a 4.0 GB, 2.50 GHz laptop.

6.1 Example 1: Solutions Methods and Tracking

We start by considering a loss random variable

$$Y = X_1 + X_2\epsilon, \text{ almost surely,}$$

where ϵ is a standard normal random variable and $X = (X_1, X_2)$ is uniformly distributed on $[-1, 1] \times [0, 1]$, with ϵ, X_1 , and X_2 independent. We consider a regression function of the form $f(x) = C_0 + C_1x_1 + C_2x_2$ and set $\alpha = 0.90$.

We first examine the computational effort required to obtain an approximate regression vector. Table 1 shows computing times for solving D_{LP}^{ν} for increasingly larger sample sizes ν obtained by independent draws from (ϵ, X_1, X_2) . While the results correspond to single instances of D_{LP}^{ν} , the times vary little between two samples of the same size and the computing times are therefore representative. As expected from the discussion at the end of Section 5.1, the computing time grows quickly as the sample size ν increases. In addition to the inconvenience of long computing times, memory requirements become problematic. D_{LP}^{ν} has a special structure and we anticipate significant reduction in computing times and memory needs resulting from tailored algorithms. However, the development of such algorithms is beyond the scope of the paper.

ν	100	200	300	400	500	600	700	800	900	1000	1500	2000
Time	0	0	2	6	17	32	45	65	163	174	996	2972

Table 1: Computing times (sec.) to solve D_{LP}^{ν} for increasing sample size in Example 1.

Second, we consider the alternative approach based on solving $D^{\nu,\mu}$. While this approach introduces a numerical integration error, Proposition 6 indicates that the error is negligible for large μ . In fact, as we see next empirically, moderately large μ suffices. Moreover, the substantial reduction in problem size, as compared to that of D_{LP}^{ν} , reduces computing times dramatically.

Since $\bar{q}_{\beta}(Z_0^{\nu}(C))$ may be nonsmooth as a function of β , standard numerical integration error bounds may not apply. However, since $\bar{q}_{\beta}(Z_0^{\nu}(C))$ is continuous and nondecreasing as a function of β , the use of left-endpoint and right-endpoint numerical integration rules in $D^{\nu,\mu}$ provide lower and upper bounds on the optimal value of D^{ν} , respectively. Table 2 shows solution vectors (C_0, C_1, C_2) for $\mu = 100$, $\mu = 1000$, left-endpoint, right-endpoint, and Simpson's numerical integration rules, and sample sizes of $\nu = 100$ and $\nu = 10000$. Each solution of $D^{\nu,\mu}$ is obtained quickly, in about 0.5 and 5 seconds for $\nu = 100$ and $\nu = 10000$, respectively; see the last column of Table 2. We also show the corresponding coefficient of determination $\bar{R}_{\alpha}^2$ for each instance. For $\nu = 100$, the solutions and $\bar{R}_{\alpha}^2$ are insensitive to the numerical integration rule as well as μ . The obtained solutions are essentially identical to the regression vector obtained from D_{LP}^{ν} ; see Row 8 of Table 2. For $\mu = 10000$, we note some differences but magnitudes are small. In this case, we are unable to solve D_{LP}^{ν} due to its size. We observe that as indicated by the coefficients of determination, the

Rule	ν	μ	C_0	C_1	C_2	$\bar{R}_{0.90}^2$	Time
Left Endpoint	100	100	0.0630	1.0951	1.5841	0.568	0.07
Left Endpoint	100	1000	0.0630	1.0951	1.5841	0.568	0.79
Right Endpoint	100	100	0.0630	1.0951	1.5841	0.568	0.08
Right Endpoint	100	1000	0.0630	1.0951	1.5841	0.568	0.83
Simpson's	100	100	0.0630	1.0951	1.5841	0.568	0.09
Simpson's	100	1000	0.0630	1.0951	1.5841	0.568	0.77
Analytic	100	NA	0.0630	1.0951	1.5841	0.568	0.05
Left Endpoint	10000	100	0.0835	1.0049	1.6374	0.392	0.58
Left Endpoint	10000	1000	0.0820	1.0048	1.6423	0.392	5.91
Right Endpoint	10000	100	0.0799	1.0050	1.6492	0.392	0.56
Right Endpoint	10000	1000	0.0816	1.0048	1.6435	0.392	5.00
Simpson's	10000	100	0.0818	1.0048	1.6429	0.392	0.56
Simpson's	10000	1000	0.0818	1.0048	1.6430	0.392	5.27

Table 2: Solution vectors, coefficients of determination, and computing times (sec.) for Example 1 with varying integration rule as well as number of intervals μ and observations ν .

linear model $f(x) = C_0 + C_1x_1 + C_2x_2$ doesn't fully capture the variability of the data and a study of other models may be warranted. However, we omit such an investigation and instead turn to superquantile tracking.

Third, we examine conditional values of Y given realizations of $X = (X_1, X_2)$, i.e., superquantile tracking. For $x = (x_1, x_2)$, $Y(x) = Y|X = x$ is normally distributed with mean x_1 and variance x_2^2 . Consequently, it is straightforward to compute that

$$\bar{q}_{0.9}(Y(x)) = x_1 + 1.7550x_2.$$

Table 2 shows vectors that only track $\bar{q}_{0.9}(Y(\cdot))$ approximately, as C_0 , C_1 , and C_2 deviate from 0, 1, and 1.755, respectively. In fact, there is in general no guarantee that every regression function f will satisfy $f(x) = \bar{q}_\alpha(Y(x))$ for all x , even for large sample sizes. As indicated by Proposition 5, however, a superquantile of $Y(x)$ can be estimated by approximating a degenerate distribution of (X, Y) at x . Table 3 shows such 'local' estimates of $\bar{q}_{0.9}(Y(x))$ near $x = (0.5, 0.5)$. Specifically, using $\nu = 500$ we compute C_0 , C_1 , and C_2 by solving D_{LP}^ν as above, with X sampled uniformly from $[-1, 1] \times [0, 1]$. We repeat these calculations 10 times with independent samples and obtain the aggregated statistics of Column 2 of Table 3. The second row gives an approximate 95% confidence interval for the mean value of $C_0 + 0.5C_1 + 0.5C_2$ across the 10 meta-replications. The interval contains $\bar{q}_{0.9}(Y((0.5, 0.5))) = 1.3775$, but is somewhat wide. Proposition 5 indicates that sampling from a smaller set $[0.45, 0.55] \times [0.45, 0.55]$ will tend to improve the estimate of $\bar{q}_{0.9}(Y((0.5, 0.5)))$. Column 3 of Table 3 illustrates this effect, by showing results comparable to those of Column 2 and Row 2, but for the smaller interval. As expected, the confidence interval for $C_0 + 0.5C_1 + 0.5C_2$ narrows around the correct value. The last column shows similar results, but now for sampling of X uniformly on $[0.495, 0.505] \times [0.495, 0.505]$. The estimate of $\bar{q}_{0.9}(Y((0.5, 0.5)))$ improves only marginally, with the residual uncertainty being due to the inherent variability in the (relatively small) samples. The narrow sampling interval causes the last estimate to be similar to that obtained by the standard empirical estimate from 500 realization of $Y((0.5, 0.5))$, which yields the confidence interval (1.312, 1.462).

While sampling on smaller sets gives better local estimates of $\bar{q}_{0.9}(Y(x))$, the global picture deteriorates. The last three rows of Table 3 show corresponding approximate 95% confidence intervals for C_0 , C_1 , and C_2 , respectively. While $C_0 + C_1x_1 + C_2x_2$ generated by the set $[-1, 1] \times [0, 1]$

provides a reasonably good global picture of $\bar{q}_{0.9}(Y(x))$, the smaller sets lose that quality as seen from the wide confidence intervals. In view of the above results, we see that an analyst that can choose “design points,” i.e., points x at which to sample $Y(x)$, should balance the need for accurate local estimates with that of global estimates. In fact, even if the primary focus is on estimating $\bar{q}_\alpha(Y(x))$ for a given x , as we see in this example, it may be equally effective to spread the samples of X near x instead of exactly at x , and then obtain some global information about $\bar{q}_\alpha(Y(\cdot))$ too. Our methodology provides a flexible framework for estimating $\bar{q}_\alpha(Y(x))$ even if there is only a small number of realization of $Y(x)$, or even none, available. The estimates are based on realization of $Y(x')$ for x' near x . None of the numerical examples in this paper include data with more than one realization of $Y(x)$ for any x .

X range:	$[-1, 1] \times [0, 1]$	$[0.45, 0.55]^2$	$[0.495, 0.505]^2$
$C_0 + 0.5C_1 + 0.5C_2$	(1.349, 1.575)	(1.329, 1.475)	(1.330, 1.473)
C_0	(0.029, 0.123)	(-2.414, 1.784)	(-23.715, 18.329)
C_1	(0.971, 1.075)	(-0.229, 3.597)	(-11.063, 25.656)
C_2	(1.523, 1.975)	(-1.686, 5.186)	(-33.916, 35.701)

Table 3: Approximate 95% confidence intervals when tracking $\bar{q}_{0.9}(Y(\cdot))$ in Example 1 near $x = (0.5, 0.5)$ using shrinking sampling ranges for X . The correct value $\bar{q}_{0.9}(Y((0.5, 0.5))) = 1.378$.

6.2 Example 2: Uncertainty Quantification

The next example arises in uncertainty quantification of a rectangular cross section of a structural column under uncertain material properties and uncertain loads; see [8] for details. The performance of the column is described by the random variable

$$Y = -1 + \frac{4X_1}{wd^2X_3} + \frac{X_2^2}{w^2d^2X_3^2}, \text{ almost surely,} \quad (40)$$

where the moment load X_1 and the axial load X_2 are normally distributed with mean 2000 and standard deviation 400, and mean 500 and standard deviation 100, respectively, and the material strength X_3 is lognormally distributed with parameters 5 and 0.5, with X_1 , X_2 , and X_3 independent. (We note that the orientation of the performance random variable is switched compared to that of [8] for consistency with our focus on ‘losses’ instead of ‘gains.’) We set the width $w = 3$, and the depth $d = 12$.

We seek to quantify the ‘uncertainty’ in Y by surrogate estimation. Of course, in this case, this is hardly necessary; direct use of (40) suffices. However, in practice, an analytic expression for a system performance, as in (40), is rarely available. One then proceeds with determining a regression function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$, based on a sample of input-output realizations, such that $f(X)$, with $X = (X_1, X_2, X_3)$, approximates Y in some sense. To mimic this situation, we consider a sample of size 50000 drawn independently from X , the corresponding realizations of Y according to (40), and two forms of the regression function. The first model is linear and takes the form

$$f_1(x) = C_0 + C_1x_1 + C_2x_2 + C_3x_3$$

and the second one utilizes basis functions $h_1(x) = x_1/x_3$ and $h_2(x) = (x_2/x_3)^2$ and is of the form

$$f_2(x) = C_0 + C_1x_1/x_3 + C_2x_2^2/x_3^2.$$

In view of (40), we expect f_1 to be unable to capture interaction effects between variables and its explanatory power may be limited. In contrast, f_2 uses the correct basis functions, but even then

Model	α	C_0	$10^2 C_1$	$10^4 C_2$	$10^4 C_3$	$\bar{R}_\alpha^2$
f_1	0.999	-0.6797	0.0156	7.9000	-9.1100	0.154
f_1	0.99	-0.8084	0.0150	3.8000	-8.2700	0.190
f_1	0.9	-0.8579	0.0107	1.5900	-7.7000	0.260
f_1	0.75	-0.8705	0.0090	1.0800	-7.5900	0.301
f_1	LS	-0.8827	0.0070	0.5921	-7.7180	0.571*
f_2	0.999	-1.0457	1.8640	0.0300	NA	0.902
f_2	0.99	-1.0450	1.6182	0.0400	NA	0.891
f_2	0.9	-1.0308	1.3393	0.0200	NA	0.894
f_2	0.75	-1.0261	1.2595	0.0200	NA	0.893
f_2	LS	-1.0179	1.1315	0.0056	NA	0.979*

Table 4: Approximate regression vectors and coefficients of determination in Example 2 for varying α and least-squares (LS) regression. An asterisk indicates that the coefficient of determination is determined by (34).

$f_2(X)$ may deviate from Y due to the finite sample size used to determine the regression vector. Table 4 confirms this intuition by showing approximate regression vectors for both models over a range of probability levels α as well as for the least-squares (LS) regression. The vectors are obtained in less than 15 seconds by solving $D^{\nu,\mu}$, with $\nu = 50000$, $\mu = 1000$, and Simpson's rule. The last column of Table 4 shows $\bar{R}_\alpha^2$ (classical coefficient of determination according to (34) in the case of least-squares regression), which is low for f_1 and high for f_2 as expected.

In uncertainty quantification and elsewhere, surrogate estimates such as $f_1(X)$ and $f_2(X)$ are important input to further analysis and simulation. Table 5 illustrates the quality of these surrogate estimates in this regard by showing various statistics of $f_1(X)$ and $f_2(X)$ as compared to those of Y . Row 2, Columns 3-10 show estimated mean, standard deviation, superquantiles at 0.75, 0.9, 0.99, 0.999, probability of failure, and buffered probability of failure (see (2)) of Y , respectively, using a sample size of 10^7 and standard estimators. Coefficients of variation for these estimators are ranging, approximately, from 10^{-5} for the mean to 0.02 for the probability of failure. Rows 3-6 of Table 5 show similar results, using the same sample, for $f_1(X)$, with $\alpha = 0.999, 0.99, 0.9$, and 0.75 , respectively. We notice that as α increases, $f_1(X)$ becomes increasingly conservative. In fact, for $\alpha = 0.999$, $f_1(X)$ is conservative in all statistics. Superquantile regression with smaller α fails to be conservative for some 'upper-tail' statistics. Interestingly, $f_1(X)$ based on α is conservative for all superquantiles up to and including $\bar{q}_\alpha$ in these tests. These observations indicate that in surrogate estimation the probability level α should be selected in accordance with the superquantile statistic of interest. We can then expect to obtain conservative estimates even for relatively poor surrogates. Row 7 of Table 5 gives corresponding results for $f_1(X)$ under the least-squares regression fit. While this fit provides an accurate estimate of the mean (see Column 3), the upper-tail behavior is represented in a nonconservative manner.

Rows 8-12 of Table 5 show comparable results to those above, but for the $f_2(X)$ models. As also indicated in Table 4, $f_2(X)$ is a much better surrogate of Y than $f_1(X)$ and essentially all quantities improve in accuracy. For example, $f_2(X)$ based on superquantile regression overestimates the buffered failure probability only moderately with $\alpha = 0.999, 0.99$, and 0.9 , and slightly underestimates with $\alpha = 0.75$; see the last column of Table 5. In contrast, least-squares regression underestimates the buffered failure probability substantially even for this supposedly 'accurate' model. Of course, least-squares regression centers on conditional expectations and as basis for estimating tail behavior may hide potentially dangerous risks.

Model	α	μ	σ	$\bar{q}_{0.75}$	$\bar{q}_{0.9}$	$\bar{q}_{0.99}$	$\bar{q}_{0.999}$	$10^3 p$	$10^3 \bar{p}$
Y	NA	-0.8436	0.0996	-0.7113	-0.6211	-0.3501	0.0091	0.3575	1.052
$f_1(X)$	0.999	-0.1259	0.1297	0.0305	0.0856	0.1868	0.2635	158.1838	376.995
$f_1(X)$	0.99	-0.4575	0.1027	-0.3370	-0.2963	-0.2225	-0.1669	0	0
$f_1(X)$	0.9	-0.6940	0.0828	-0.6016	-0.5728	-0.5219	-0.4843	0	0
$f_1(X)$	0.75	-0.7641	0.0777	-0.6795	-0.6544	-0.6106	-0.5786	0	0
$f_1(X)$	LS	-0.8439	0.0748	-0.7653	-0.7439	-0.7077	-0.6819	0	0
$f_2(X)$	0.999	-0.7611	0.1647	-0.5381	-0.3961	-0.0053	0.44953	3.4410	9.713
$f_2(X)$	0.99	-0.7979	0.1431	-0.6042	-0.4808	-0.1413	0.25383	1.4909	4.206
$f_2(X)$	0.9	-0.8263	0.1184	-0.6660	-0.5640	-0.2831	0.04375	0.4702	1.332
$f_2(X)$	0.75	-0.8337	0.1113	-0.6830	-0.5870	-0.3229	-0.0155	0.3194	0.899
$f_2(X)$	LS	-0.8451	0.1000	-0.7097	-0.6235	-0.3864	-0.1104	0.1539	0.440

Table 5: Statistics of $f_1(X)$ and $f_2(X)$ in Example 2 as compared to those of Y . Columns 3-10 show mean, standard deviation, superquantiles at 0.75, 0.9, 0.99, 0.999, probability of failure, and buffered probability of failure, respectively.

6.3 Example 3: Investment Analysis

The last example is a case study taken from the “Style Classification with Quantile Regression” documentation in Portfolio Safeguard [1] and deals with the negative return of the Fidelity Magellan Fund as predicted by the explanatory variables Russell 1000 Growth Index (X_1 , RLG), Russell 1000 Value Index (X_2 , RLV), Russell Value Index (X_3 , RUJ), and Russell 2000 Growth Index (X_4 , RUO). (We change the orientation from ‘return’ to ‘negative return’ to be consistent with the orientation of a loss random variable in the present paper.) The indices classify the style of the fund; see [1] for details. There are $\nu = 1264$ total observations available.

Regression	α	C_0	C_1 (RLG)	C_2 (RLV)	C_3 (RUJ)	C_4 (RUO)	$\bar{R}_\alpha^2$
Least-squares	NA	0.0010	-0.5089	-0.5180	0.0484	0.0061	0.9824*
Quantile	0.75	0.0045	-0.5438	-0.4518	0.0159	0.0173	—
Superquantile	0.75	0.0095	-0.5036	-0.4723	0.0192	0.0009	0.8735
Quantile	0.90	0.0089	-0.5177	-0.4602	0.0156	-0.0001	—
Superquantile	0.90	0.0138	-0.4837	-0.4912	0.0223	-0.0019	0.8722

Table 6: Approximate regression vectors and coefficients of determination in Example 3 for model f_1 . An asterisk indicates that coefficient of determination is determined by (34).

We start by considering a linear model $f_1(x) = C_0 + C_1x_1 + C_2x_2 + C_3x_3 + C_4x_4$ and compare the obtained approximate regression vectors for least-squares, quantile, and superquantile regression under $\alpha = 0.75$ and 0.90, as shown in Table 6. D^ν is solved through $D^{\nu,\mu}$ with Simpson’s rule and $\mu = 1000$, while quantile regression is carried out directly in Portfolio Safeguard’s Shell Environment [1]. Table 6 also shows the coefficients of determination, where for least-squares regression we use (34). The fits are good and a majority of the variability in the data is captured. However, the small values of C_4 and also the corresponding p -value from the least-squares regression point to the possible merit of dropping X_4 (RUO) as explanatory variable. We from now on focus on superquantile regression. A new model $f_2(x) = C_0 + C_1x_1 + C_2x_2 + C_3x_3$ yields the approximate regression vectors of Table 7, which also shows the obtained adjusted coefficients of determination $\bar{R}_{\alpha,Adj}^2$. The switch from $\bar{R}_\alpha^2$ to $\bar{R}_{\alpha,Adj}^2$ enable us to better compare fits across models with different number of explanatory variables. In comparison, adjusted coefficients of determination for f_1 , with $\alpha = 0.75$ and 0.90, are 0.8732 and 0.8719, respectively. Consequently, the fit improves slightly by

dropping X_4 (RUO).

Regression	α	C_0	C_1 (RLG)	C_2 (RLV)	C_3 (RUJ)	$\bar{R}_{\alpha,Adj}^2$
Superquantile	0.75	0.0095	-0.5028	-0.4728	0.0200	0.8733
Superquantile	0.90	0.0138	-0.4855	-0.4906	0.0210	0.8720

Table 7: Approximate regression vectors and adjusted coefficients of determination in Example 3 for model f_2 .

We further reduce the model to a single explanatory variable and examine the four possibilities in Table 8. We find that $\bar{R}_{\alpha,Adj}^2$ deteriorates, but only moderately for the model $C_0 + C_1 X_1$. This simple model captures much of the variability in the data set. A somewhat poorer fit is achieved by X_2 (RLV), which is illustrated in Figure 1 for $\alpha = 0.90$. That figure also depicts the corresponding quantile and least-squares regression lines. It's apparent that superquantile regression provides a distinct perspective from the other regression techniques of potential significant value to a decision maker.

Model	α	C_0	C_1 (RLG)	C_2 (RLV)	C_3 (RUJ)	C_4 (RUO)	$\bar{R}_{\alpha,Adj}^2$
$C_0 + C_1 X_1$	0.75	0.0137	-0.8228	—	—	—	0.7380
$C_0 + C_1 X_1$	0.90	0.0218	-0.8189	—	—	—	0.7248
$C_0 + C_2 X_2$	0.75	0.0321	—	-1.0668	—	—	0.5940
$C_0 + C_2 X_2$	0.90	0.0475	—	-1.0727	—	—	0.5702
$C_0 + C_3 X_3$	0.75	0.0515	—	—	-0.7745	—	0.4103
$C_0 + C_3 X_3$	0.90	0.0714	—	—	-0.6949	—	0.4162
$C_0 + C_4 X_4$	0.75	0.0344	—	—	—	-0.5498	0.3962
$C_0 + C_4 X_4$	0.90	0.0512	—	—	—	-0.5145	0.2593

Table 8: Approximate regression vectors and adjusted coefficients of determination in Example 3 for superquantile regression with single-variable models.

7 Conclusions

The paper presents a superquantile regression methodology centered on the minimization of a measure of error analogous to classical least-squares and quantile regression. We establish the existence of a regression function, discuss its possible uniqueness, and its stability under perturbation, for example caused by sample approximations of a true distribution. A new coefficient of determination allows us to quantify the goodness of fit. We show that superquantile regression requires the solution of a linear program, as in the case of quantile regression, or alternatively of an optimization problem with superquantile (conditional value-at-risk) constraints. Our computational tests demonstrate that superquantile regression is computationally tractable, provides new insight about tail-behavior for quantities of interest, and offers a complementary tool for the risk-averse decision maker.

Acknowledgement

This work was supported by the Air Force Office of Scientific Research under Grants FA9550-11-1-0206 and F1ATAO1194GOO1. The authors thank Prof. S. Uryasev, University of Florida, for enabling and supporting numerical tests in Portfolio Safeguard.

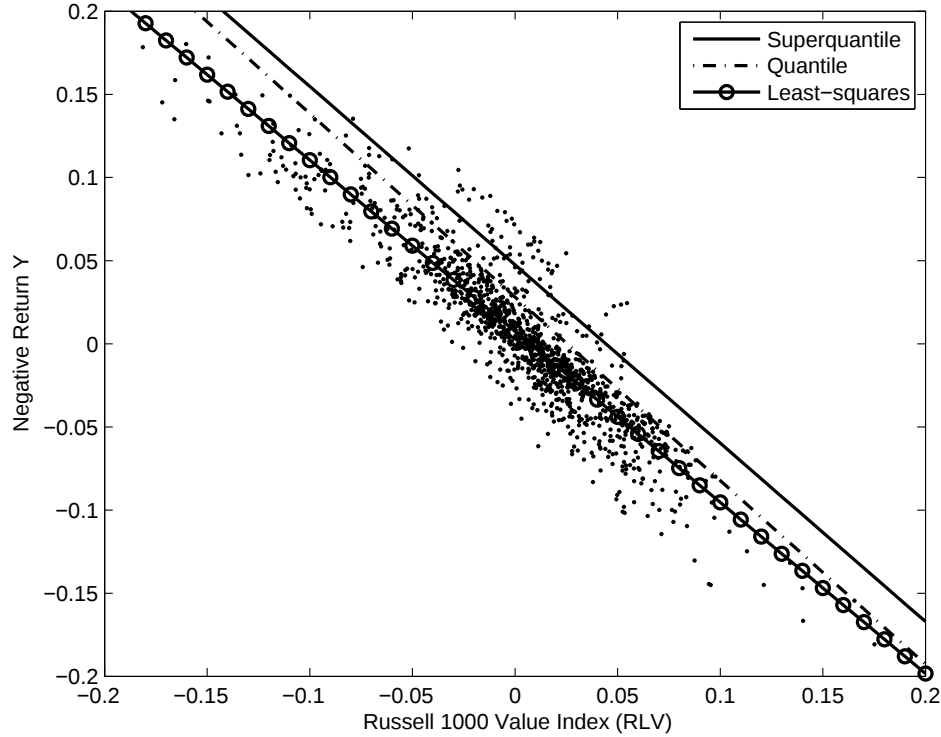


Figure 1: Regression lines in Example 3 for model $C_0 + C_2 X_2$.

References

- [1] American Optimal Decisions, Inc. *Portfolio Safeguard (PSG) in Windows Shell Environment: Basic Principles*. AORDA, Gainesville, FL, 2011.
- [2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–227, 1999.
- [3] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., New York, NY, 3. edition, 1995.
- [4] Z. Cai and X. Wang. Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics*, 147(1):120–130, 2008.
- [5] S. Y. Chun, A. Shapiro, and S. Uryasev. Conditional value-at-risk and average value-at-risk: Estimation and asymptotics. *Operations Research*, 60(4):739–756, 2012.
- [6] G. Conner. The three types of factor models: A comparison of their explanatory power. *Financial Analysts Journal*, 15:42–46, 1995.
- [7] F. Delbaen. Coherent measures of risk on general probability spaces. In P.J. Schonbucher K. Sandmann, editor, *Advances in Finance and Stochastics, Essays in Honor of Dieter Sondermann*, pages 1–37. Springer, Berlin, Germany, 2002.
- [8] M.S. Eldred and L.P. Swiler. Efficient algorithms for mixed aleatory-epistemic uncertainty quantification with application to radiation-hardened electronics. part 1: Algorithms and

- benchmark results. Technical Report SAND2009-5805, Sandia National Laboratories, Albuquerque, New Mexico, 2009.
- [9] W. Gilchrist. Regression revisited. *International Statistical Review*, 76(3):401–418, 2008.
 - [10] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011.
 - [11] P. Hall and H.G. Muller. Order-preserving nonparametric regression, with applications to conditional distribution and quantile function estimation. *Journal of the American Statistical Association*, 98(463):598–608, 2003.
 - [12] T. Hothorn, T. Kneib, and P. Buhlmann. Conditional transformation models. *Journal of the Royal Statistical Society*, to appear, 2013.
 - [13] K. Kalinchenko, A. Veremyev, V. Boginski, D.E. Jeffcoat, and S. Uryasev. Robust connectivity issues in dynamic sensor networks for area surveillance under uncertainty. *Structural and Multidisciplinary Optimization*, 7(2):235–248, 2011.
 - [14] K. Kato. Weighted Nadaraya-Watson estimation of conditional expected shortfall. *Journal of Financial Econometrics*, 10(2):265–291, 2012.
 - [15] J. Knight and S. Satchell (Eds.). *Linear Factor Models in Finance*. Butterworth-Heinemann, Oxford, UK, 2005.
 - [16] R. Koenker. *Quantile regression*. Cambridge University Press, Cambridge, UK, 2005.
 - [17] H. Lee and W. Chen. A comparative study of uncertainty propagation methods for black-box-type problems. *Structural and Multidisciplinary Optimization*, 37(3):239–253, 2009.
 - [18] S. Leorato, F. Peracchi, and A.V. Tanase. On estimating the conditional expected shortfall. *Applied Stochastic Models in Business and Industry*, 24:471–493, 2008.
 - [19] S. Leorato, F. Peracchi, and A.V. Tanase. Asymptotically efficient estimation of the conditional expected shortfall. *Computational Statistics & Data Analysis*, 56(4):768–784, 2012.
 - [20] R. T. Rockafellar and J. O. Royset. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety*, 95:499–510, 2010.
 - [21] R. T. Rockafellar and S. Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, to appear, 2013.
 - [22] R. T. Rockafellar, S. Uryasev, and M. Zabarankin. Risk tuning with generalized linear regression. *Mathematics of Operations Research*, 33(3):712–729, 2008.
 - [23] R. T. Rockafellar and R. J-B Wets. *Variational Analysis*. Springer, Berlin, Germany, 1998.
 - [24] R.T. Rockafellar and J.O. Royset. Random variables, monotone relations and convex analysis. *Mathematical Programming B*, in review, 2013.
 - [25] R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

- [26] O. Scaillet. Nonparametric estimation of conditional expected shortfall. *Insurance and Risk Management Journal*, 72:639–660, 2005.
- [27] A. Trindade, S. Uryasev, A. Shapiro, and G. Zrazhevsky. Financial prediction with constrained tail risk. *Journal of Banking and Finance*, 31(11):3524–3538, 2007.
- [28] C.-J. Wang and S. Uryasev. Efficient execution in the secondary mortgage market: a stochastic optimization model using CVaR constraints. *Journal of Risk*, 10(1):41–66, 2007.