

ON USING LINEAR QUANTILE REGRESSIONS FOR CAUSAL INFERENCE

RYUTAH KATO AND YUYA SASAKI
Johns Hopkins

We show that the slope parameter of the linear quantile regression measures a weighted average of the local slopes of the conditional quantile function. Extending this result, we also show that the slope parameter measures a weighted average of the partial effects for a general structural function. Our results support the use of linear quantile regressions for causal inference in the presence of nonlinearity and multivariate unobserved heterogeneity. The same conclusion applies to linear regressions.

1. INTRODUCTION

Is it appropriate to use the linear quantile regression (QR) for causal inference? To answer this question, we study relationships among the linear QR, the conditional quantile function (CQF), and general structural functions. First, we show that the slope parameter of the linear QR represents a weighted average of the local slopes of the CQF. Second, we cite an existing result demonstrating that the slope of the CQF represents a weighted average of the structural partial effects. Third, by chaining these two arguments, we show that the slope parameter of the linear QR represents a weighted average of the structural partial effects. Our results imply that the linear QR can be used for structural causal inference in the presence of nonlinearity and multiple unobserved heterogeneity.

Endogeneity and misspecification could be two major obstacles in using the linear QR for causal inference. Since an extensive body of the econometrics literature has solved the endogeneity issue in various contexts,¹ we assume exogeneity and focus on the misspecification problem throughout this paper. The linear QR generally misspecifies the true causal structure in two ways. First, the linearity of the linear QR fails to capture the nonlinearity of general structural functions.² Second, the monotonicity of the linear QR with respect to a scalar latent variable fails to capture the nonmonotonicity of general structural functions with respect to generally multivariate unobservables. Despite these two sources of potential misspecification, what we develop in this paper reconciles the linear QR with the true causal structure.

We were benefited from useful suggestions from V. Chernozhukov, I. Fernández-Val, and K. Kato. All the remaining errors are ours. Address correspondence to Ryutah Kato and Yuya Sasaki, Johns Hopkins University, Department of Economics, Wyman Park Building 544E, 3400 N. Charles St., Baltimore, MD 21218, USA; e-mail: rkato3@jhu.edu and sasaki@jhu.edu.

It is a well-known feature of the ordinary least squares (OLS) method that it minimizes the average squared distance between the linear regression function and the conditional expectation function (CEF). In fact, the OLS does not only minimize the fit. Yitzhaki (1996) and Angrist and Krueger (1999) show that the OLS slope coefficient under misspecification can be explicitly written as weighted average derivatives of the CEF. This result is viewed by empirical researchers to be useful for an interpretation of regression coefficients.³ One contribution of this paper is to provide a generalization of this weighted-average-derivatives interpretation for arbitrary linear functions that minimize weighted mean squared distances. Our result applies to the linear QR in particular, since Angrist, Chernozhukov, and Fernández-Val (2006) show that the linear QR parameters minimize a weighted average of squared distances between the linear function and the CQF.⁴

In Section 2, we show that the slope parameter of the linear QR equals a weighted average of the local slopes of the CQF. In Section 4, we use the existing result that the slope of the CQF equals a weighted average of the structural partial effects. We provide a framework for claiming the two auxiliary results together to show that the slope parameter of the linear QR equals a weighted average of the structural partial effects, even if the underlying structural function is nonlinear in the explanatory variable and/or is nonmonotone with respect to possibly multivariate latent variable(s). We therefore conclude in Section 6 that linear QR (Koenker and Bassett, 1978)⁵ is a useful tool for causal inference in the presence of unobserved heterogeneity, even if it misspecifies the true structure both in terms of nonlinearity and nonmonotonicity. A couple of extensions to the baseline result are presented in Section 3, and we also argue that the same conclusion applies to the linear regression in Section 5 as well as the linear QR.

2. RELATIONSHIP BETWEEN LINEAR AND NONLINEAR FUNCTIONS

2.1. General Setting

We first introduce basic definitions and fix relevant notations. Let Y and X denote dependent and independent variables supported on $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}$, respectively. Let $\mathcal{B}(\mathcal{X})$ denote the Borel sigma algebra on \mathcal{X} . First, we present an auxiliary lemma that a nonlinear function and a linear function related by a minimum weighted mean squared distance have at least two intersection points. Let $M : \mathcal{X} \rightarrow \mathbb{R}$ be a general function and $L : \mathcal{X} \rightarrow \mathbb{R}$ be an affine function specified by the parametric form $L(x) = \alpha + \beta x$. We state the following assumption for these two functions and the distribution of X .

Assumption 1. (i) $E[M(X)^2]$ and $E[X^2]$ are finite. (ii) (α, β) uniquely solves $\min_{(a,b)} E[\omega(X) \cdot (M(X) - a - bX)^2]$ for some weight function $\omega : \mathcal{X} \rightarrow \mathbb{R}$, where $\omega(x) \geq 0$ a.e. (iii) $\omega(X)$ is bounded a.s. (iv) M is continuous. (v) \mathcal{X} is convex. (vi) X is continuously distributed.

In this assumption, the linear function L is characterized as the unique weighted least squares approximation to the nonlinear function M . This applies to the

relation between the linear regression function L and the CEF M , for example. In addition, this also applies to the relation between the linear QR function L and the CQF M by Angrist, Chernozhukov, and Fernández-Val (2006). The current section proceeds with a general setting, and we will branch into these two concrete examples in Sections 2.3 and 2.4.

We first obtain the following auxiliary lemma with Assumption 1.

LEMMA 1 (Existence of Two Intersection Points). *If Assumption 1 is true, then M and L intersect at least at two points in \mathcal{X} .*

A proof of this lemma is provided in Section A.1 in the appendix. We also present an intuitive illustration of this lemma in Section 2.2, describing how the minimum weighted mean-squares characterization leads to the existence of two intersection points. Applying Lemma 1, we next claim that the slope β of L is a weighted average of the local slopes M' of M . To this end, we invoke an additional assumption that ensures that this derivative M' of interest exists.

Assumption 2. M is continuously differentiable.

An application of Lemma 1 and the Fundamental Theorem of Calculus yields the following result.

THEOREM 1. *If Assumptions 1 and 2 are true, then there exist $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$ such that*

$$\beta = E\left[w_{x_1, x_2}(X) \cdot M'(X) \mid x_1 \leq X \leq x_2 \right]$$

holds, where the weight function w_{x_1, x_2} is given by

$$w_{x_1, x_2}(x) := \frac{\frac{1}{f_X(x)}}{\int_{x_1}^{x_2} \frac{1}{f_X(\xi)} dF_X(\xi)} \quad \text{for all } x \in \mathcal{X}.$$

A proof of this theorem is provided in Section A.2 in the appendix, but it is straightforward from Lemma 1. This result characterizes β as a weighted average of M' . Furthermore, note that the weight function w_{x_1, x_2} is strictly positive on $[x_1, x_2]$. In other words, the slope β of L represents a strict convex combination of the local slopes M' of M within an interval $[x_1, x_2] \subset \mathcal{X}$, which is guaranteed to be nonempty.

To mathematically formalize the weighted-average interpretation, we define the function $\lambda_{x_1, x_2} : \mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R}$ by

$$\lambda_{x_1, x_2}(T) = \int_{T \cap [x_1, x_2]} w_{x_1, x_2}(x) dF_X(x) \quad \text{for each } T \in \mathcal{B}(\mathcal{X}). \tag{2.1}$$

Note that this λ_{x_1, x_2} can be shown to be a probability measure. With this notation, the conclusion of Theorem 1 can be restated as

$$\beta = E_{\lambda_{x_1, x_2}} [M'(X)]. \tag{2.2}$$

In other words, the slope β of L is the mean of the local slopes M' of M with respect to the probability measure λ_{x_1, x_2} defined in (2.1).

A remaining question is whether there can exist more than two intersection points. In general, there may exist more than two. For example, in the extreme case where M is also an affine function, the minimum weighted mean squared distance restriction necessarily leads to $L = M$, and there exists a continuum of intersection points. Note that, in this affine case, one can take any pair of points, x_1 and x_2 , from \mathcal{X} to form the weight function w_{x_1, x_2} defined in Theorem 1, since $\beta = L' = M'$ holds globally. In general, when there exist more than two intersection points between L and M , we can choose any pair of those intersection points, x_1 and x_2 , to form the weight function w_{x_1, x_2} . In this sense, the weight function need not be unique.

2.2. Intuition

We provide an intuitive illustration for Lemma 1, where we claim that there exist at least two intersection points of M and L in \mathcal{X} . Since L minimizes the weighted mean squared distance from M , no other affine function can have a strictly smaller w -weighted mean squared distance from the function M than the affine function L . This restriction plays the major role in our theoretical and intuitive arguments.

By way of contradiction, assume that L and M have less than two intersection points in \mathcal{X} . We can then branch into three representative cases, illustrated by (a)–(c) in Figure 1. Panel (a) shows a case where there is no intersection point. In this case, either $L - M > 0$ or $L - M < 0$ is true by the continuity of M . If $L - M < 0$ like the graph, then shifting L upward to produce \tilde{L} would achieve a smaller weighted mean squared distance between M and \tilde{L} than between M and L , contradicting the premise that L minimizes the weighted mean squared distance from M . Thus, we rule out case (a) of zero intersection point.

Panels (b) and (c) show cases where there is only one intersection point. In these cases, rotating and/or shifting L to produce \tilde{L} would achieve a smaller weighted mean squared distance between M and \tilde{L} than between M and L , contradicting the characterization of L as the minimizer of the weighted mean squared distance from M . Thus, we rule out cases (b) and (c) of one intersection point.

While this illustration is intuitive, establishing the existence of such a rotation and/or shift to achieve a strictly smaller weighted mean squared distance is not necessarily a trivial problem, except in case (a). We refer readers to Section A.1 in the appendix for mathematical details of this illustration.

2.3. Linear QR and CQF

We demonstrate an application of Theorem 1 to the case of quantile regressions. For a given $\tau \in (0, 1)$, the τ -th CQF $Q_\tau : \mathcal{X} \rightarrow \mathcal{Y}$ is defined by

$$Q_\tau(x) = \inf\{y \mid F_{Y|X}(y \mid x) \geq \tau\},$$

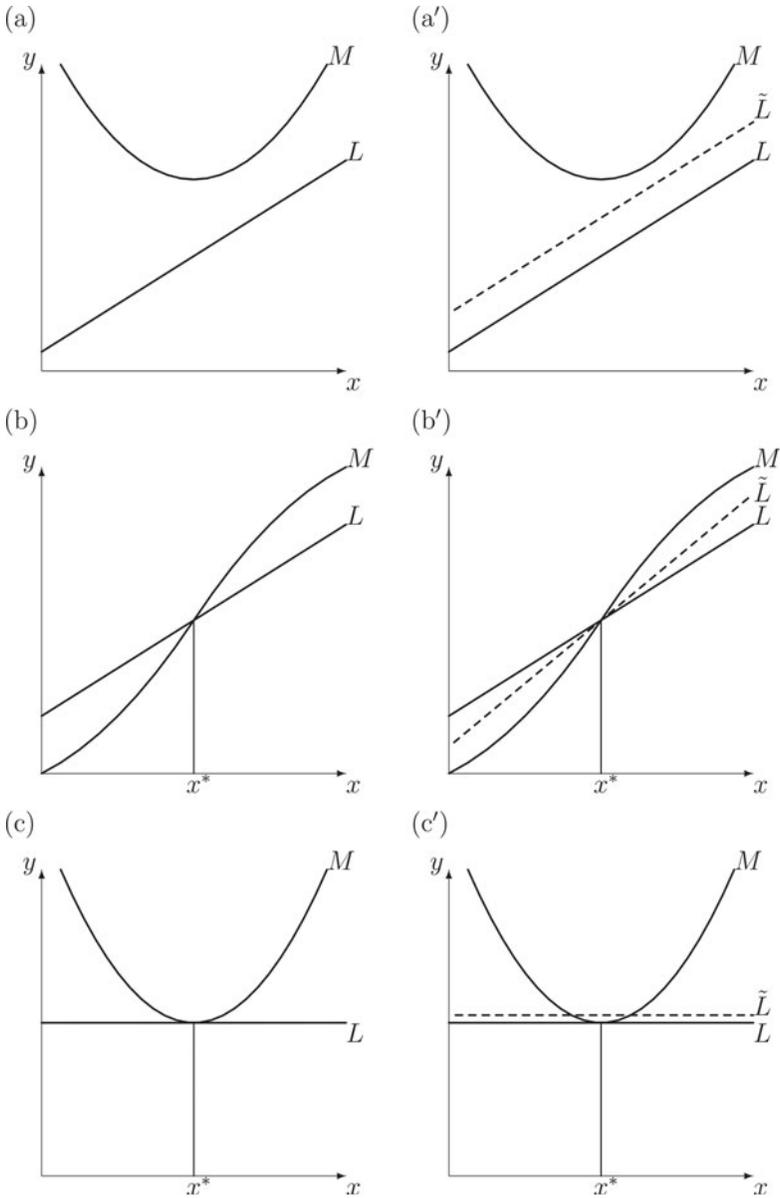


FIGURE 1. Illustrations of why M and L must have at least two intersection points. Panel (a) shows a case where there is no intersection point. Panels (b) and (c) show cases where there is only one intersection point. In each of these cases, moving L to \tilde{L} may reduce the weighted mean squared distance from M , and hence L cannot minimize the weighted mean squared distance.

where $F_{Y|X}(y | x)$ denotes the conditional cumulative distribution function of Y given X , which is assumed to be regularly defined throughout the paper. The τ -th linear QR L_τ provides the best linear prediction of Y given X under the loss function $\rho_\tau(u) : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\rho_\tau(u) = (\tau - \mathbb{1}(u \leq 0)) \cdot u$, called the check function. In other words, $L_\tau(x) = \alpha(\tau) + \beta(\tau)x$ where $(\alpha(\tau), \beta(\tau))$ satisfies

$$(\alpha(\tau), \beta(\tau)) \in \arg \min_{(\alpha, \beta) \in \Theta} E[\rho_\tau(Y - \alpha - \beta X)]$$

for some parameter set $\Theta \subset \mathbb{R}^2$. Note that we have $Q_\tau \neq L_\tau$ in general.

Angrist, Chernozhukov, and Fernández-Val (ACF, 2006; Theorem 2) show that, under the following assumption, the linear QR minimizes a weighted mean squared distance from the CQF. We state their results below.

Assumption 3 (ACF, 2006). (i) The conditional density $f_{Y|X}(\cdot | X)$ exists and is bounded a.s. (ii) $E[Y]$, $E[Q_\tau(X)^2]$, and EX^2 are finite. (iii) $(\alpha(\tau), \beta(\tau))$ uniquely solves $\min_{(\alpha, \beta)} E[\rho_\tau(Y - \alpha - \beta X)]$.

THEOREM 2 (ACF, 2006; Theorem 2). *If Assumption 3 is true, then $(\alpha(\tau), \beta(\tau))$ uniquely solves the weighted least squares problem*

$$\min_{\alpha, \beta} E[\bar{\omega}_\tau(X, \alpha(\tau), \beta(\tau)) \cdot (L_\tau - Q_\tau)^2(X, \alpha, \beta)], \tag{2.3}$$

where $\bar{\omega}_\tau(x, \alpha(\tau), \beta(\tau))$ is defined by

$$\bar{\omega}_\tau(x, \alpha(\tau), \beta(\tau)) := \frac{1}{2} \int_0^1 f_{Y|X}(uL_\tau(x) + (1-u)Q_\tau(x) | x) du.$$

To claim that the slope $\beta(\tau)$ of the τ -th linear QR L_τ is a weighted average of the slopes Q'_τ of the CQF, we now invoke an additional assumption that the support of X is convex and that the derivative Q'_τ of interest exists.

Assumption 4. (i) \mathcal{X} is convex. (ii) Q_τ is continuously differentiable. (iii) \mathcal{X} is continuously distributed.

Note that Assumptions 3 and 4 imply Assumption 1 by Theorem 2. Note also that Assumption 4(ii) implies Assumption 2. Combining Theorems 1 and 2 together, therefore, we get the following corollary that the slope $\beta(\tau)$ of the τ -th linear QR L_τ is a weighted average of the slopes Q'_τ of the CQF.

COROLLARY 1 (Weighted Average: CQF). *If Assumptions 3 and 4 are true, then there exist $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$ such that*

$$\beta(\tau) = E[w_{x_1, x_2}(X) \cdot Q'_\tau(X) | x_1 \leq X \leq x_2]$$

holds, where the weight function w_{x_1, x_2} is defined in Theorem 1.

As in the general case, this result can be formally interpreted as a weighted average based on the probability measure defined in (2.1). Specifically, the

conclusion of Corollary 1 can be restated as

$$\beta(\tau) = E_{\lambda_{x_1, x_2}} [Q'_\tau(X)]. \quad (2.4)$$

2.4. Linear Regressions and CEF

While the main focus of this paper is on quantile regressions, we demonstrate in the current subsection an application of Theorem 1 to the case of mean regressions. The linear regression L_m is a linear function which gives the minimum mean-squared-errors prediction of Y given X , i.e., $L_m(x) = \alpha + \beta x$ where

$$(\alpha, \beta) \in \arg \min_{(a, b)} E[(Y - a - bX)^2].$$

In addition, the linear regression also provides the minimum mean-squared-distance approximation to the CEF, i.e., L_m also satisfies

$$(\alpha, \beta) \in \arg \min_{(a, b)} E[(m(X) - a - bX)^2],$$

where $m : \mathcal{X} \rightarrow \mathbb{R}$ denotes the CEF. The approximation properties of the linear regression have been emphasized by White (1980), Chamberlain (1984), Goldberger (1991), and Angrist and Krueger (1999), among others. To apply Theorem 1 to the mean regressions, we state the following assumptions.

Assumption 5. (i) $E[Y^2]$ and $E[X^2]$ are finite. (ii) (α, β) uniquely solves $\min_{(a, b)} E[(Y - a - bX)^2]$. (iii) \mathcal{X} is convex. (iv) X is continuously distributed. (v) m is continuously differentiable.

Note that, if (α, β) uniquely solves $\min_{(a, b)} E[(Y - a - bX)^2]$, then (α, β) also uniquely solves $\min_{(a, b)} E[(m(X) - a - bX)^2]$. Therefore, Assumption 5 implies Assumptions 1 and 2, and we consequently obtain from Theorem 1 the following corollary that the slope β of the linear regression L_m is a weighted average of the slopes m' of the CEF.

COROLLARY 2 (Weighted Average: CEF). *If Assumption 5 is true, then there exist $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$ such that*

$$\beta = E[w_{x_1, x_2}(X) \cdot m'(X) | x_1 \leq X \leq x_2]$$

holds, where the weight function w_{x_1, x_2} is defined in Theorem 1.

As in the general case, this result can be formally interpreted as a weighted average based on the probability measure defined in (2.1). The conclusion of Corollary 2 can be restated as

$$\beta = E_{\lambda_{x_1, x_2}} [m'(X)].$$

We conclude this section with a remark about existing results that derive weighted average interpretations for the OLS slope parameter. Yitzhaki (1996;

Proposition 2) shows that the OLS slope parameter can be written as a weighted average of the local slopes of the CEF with the weight function w_Y defined by

$$w_Y(x) = \frac{1}{\text{Var}(X)} [E[X] \cdot F_X(x) - \Theta_X(x)] = \frac{F_X(x)}{\text{Var}(X)} (E[X] - E[X|X \leq x]),$$

where $\Theta_X(x) = \int_{-\infty}^x tf(t)dt = F_X(x)E[X|X \leq x]$. Likewise, Angrist and Krueger (1999; page 1311) show that the OLS slope parameter can be written as

$$\beta = E \left[\frac{X^2}{E[X^2]} m'(\zeta(X)) \right],$$

where $\zeta(x)$ derives from the mean value expansion around 0, and is thus between 0 and x . If ζ is invertible and continuously differentiable, then the resultant weight function w_{AK} can be defined by

$$w_{AK}(x) = \frac{\zeta^{-1}(x)^2 \cdot f_X(\zeta^{-1}(x))}{\zeta'(\zeta^{-1}(x)) \cdot f_X(x) \cdot E[X^2]}.$$

The weight function w_Y of Yitzhaki, the weight function w_{AK} of Angrist and Krueger, and our weight function w_{x_1, x_2} defined in Theorem 1 are different from each other. The linear regression slope parameter can thus admit multiple weights for the weighted-average interpretation. Each of the different weights can provide a unique way of interpretation. For example, w_Y has an advantage in that it is strictly positive on the support \mathcal{X} , while w_{x_1, x_2} has a truncated support $[x_1, x_2] \subset \mathcal{X}$. Thus, Yitzhaki’s result is useful for interpreting the slope parameter as a weighted average over the entire population. A similar remark applies to the weight of Angrist and Krueger, except for the zero weight at the point $x = 0$. On the other hand, w_{x_1, x_2} allows for a more intuitive interpretation as the inverse probability weighting. Thus, our result is useful for interpreting the slope parameter as a standardized average over a subpopulation. Because all the weights are true anyway, empirical practitioners can take advantage of all the convenient interpretations.

3. EXTENSIONS

This section presents a couple of extensions to the baseline result. First, we demonstrate that our basic idea also applies to the case of discrete regressors. Second, we demonstrate that our result also extends to the case of multivariate regressors.

3.1. The Result for the Case of Discrete Regressors

The main result (Theorem 1) was obtained assuming that the regressor X is continuously distributed. In the current section, we argue that a similar idea applies to the case of a discrete regressor, which is very common in economic applications.⁶ As we move from a continuous distribution to a discrete distribution, we also

change the object of interest from derivatives to differences. The following short-hand notations are introduced.

$$\beta = \Delta L(x) := L(x') - L(x)$$

$$\Delta M(x) := M(x') - M(x),$$

where $x' = \min\{\tilde{x} \in \mathcal{X} | \tilde{x} > x\}$. As in the continuous case, we first prove an auxiliary lemma under the following assumption.

Assumption 6. (i) $E[M(X)^2]$ and $E[X^2]$ are finite. (ii) (α, β) uniquely solves $\min_{(a,b)} E[\omega(X) \cdot (M(X) - a - bX)^2]$. (iii) $\omega(X)$ is bounded for all $x \in \mathcal{X}$. (iv) $\mathcal{X} \subset \mathbb{Z}$ and $|\mathcal{X}| \geq 2$.

This assumption is analogous to Assumption 1, which we made for the case of continuous X . Compared to that assumption, the current assumption drops the continuity of M and the continuous distribution of X . Instead, it adds the restriction that \mathcal{X} is integer-valued, although it can be relaxed to arbitrary discrete supports with isolated points. Under this assumption, we obtain the following auxiliary lemma.

LEMMA 2. *If Assumption 6 is true, then there exist two distinct points $x_1, x_2 \in \mathcal{X}$ such that $(\Delta L(x_1) - \Delta M(x_1)) \cdot (\Delta L(x_2) - \Delta M(x_2)) \leq 0$ holds.*

A proof is found in Section A.3 in the appendix. It is proved analogously to the logic used in the proof of Lemma 1, which we developed for the case of continuous X . This lemma claims that there is some point $x_1 \in \mathcal{X}$ at which $\beta = \Delta L(x_1) \geq \Delta M(x_1)$ is true and another point $x_2 \in \mathcal{X}$ at which $\beta = \Delta L(x_2) \leq \Delta M(x_2)$ is true. In other words, the slope parameter β of the linear function is bounded from both above and below by the differences ΔM . This property immediately implies that β can be written as a convex combination of $\{\Delta M(x) | x \in \mathcal{X}\}$, as formally stated in the following theorem.

THEOREM 3. *If Assumption 6 is true, then*

$$\beta(\tau) = E[\tilde{w}(X) \cdot \Delta M(X)]$$

holds for a non-negative weight function \tilde{w} such that $E[\tilde{w}(X)] = 1$.

A proof is provided in Section A.4 in the appendix. While our proof constructs a particular weight function w to establish the equality, such a weight function need not be unique.

Like the case of continuous regressors, we can apply Theorem 3 to quantile regressions in particular. Let $\Delta Q_\tau(x) := Q_\tau(x') - Q_\tau(x)$. The following corollary follows from Theorem 3.

COROLLARY 3. *If Assumption 3 and Assumption 6(iv) are true, then*

$$\beta(\tau) = E[\tilde{w}(X) \cdot \Delta Q_\tau(X)]$$

holds for a non-negative weight function \tilde{w} such that $E[\tilde{w}(X)] = 1$.

Likewise, we can apply Theorem 3 to mean regressions as well. Let $\Delta m(x) := m(x') - m(x)$. The following corollary follows from Theorem 3.

COROLLARY 4. *If Assumption 5(i)–(ii) and Assumption 6(iv) are true, then*

$$\beta(\tau) = E[\tilde{w}(X) \cdot \Delta m(X)]$$

holds for a non-negative weight function \tilde{w} such that $E[\tilde{w}(X)] = 1$.

3.2. The Result for the Case of Multivariate Regressors

The main result (Theorem 1) was obtained assuming that the regressor X is univariate. In the current section, we argue that a similar idea applies to the case of a multivariate regressor. To this goal, we make the following assumption.

Assumption 7. (i) $E[M(X)^2]$ and $E\|X\|^2$ are finite. (ii) $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^k$ uniquely solve $\min_{(a,b)} E[\omega(X) \cdot (M(X) - a - X'b)^2]$ for some weight function $\omega : \mathcal{X} \rightarrow \mathbb{R}$, where $\omega(x) \geq 0$ a.e. (iii) $\omega(X)$ is bounded a.s. (iv) M is twice continuously differentiable. (v) $\mathcal{X} = \mathbb{R}^k$. (vi) X is continuously distributed. (vii) There exists $C > 0$ such that $\frac{\partial^2 M(x)}{\partial x_i^2} > C$ for all $1 \leq i \leq k$ for all $x \in \mathcal{X}$.

Parts (i)–(vi) of this assumption are analogous to Assumption 1. In addition, part (vii) requires the nonlinear function M be convex.⁷ In econometrics, it is not unusual to impose such shape restrictions, in light of the fact that many functions in economics must be convex or concave to satisfy important economic properties. We can substitute the assumption that M is concave, and arguments below then can be straightforwardly modified by reversing the inequalities. With x_{-1} denoting the $(k - 1)$ -dimensional subvector $(x_2, x_3, \dots, x_k)'$ of x , we state the following two auxiliary lemmas that follow from the preceding assumption.

LEMMA 3. *If Assumption 7 is true, then there exists $x \in \mathcal{X}$ such that $L(x) > M(x)$.*

LEMMA 4. *Under Assumption 7, if there exists $x^* \in \mathcal{X}$ such that $L(x^*) > M(x^*)$, then $L(\cdot, x_{-1}^*)$ and $M(\cdot, x_{-1}^*)$ intersect at two distinct points.*

Proofs of Lemmas 3 and 4 are provided in Sections A.5 and A.6 in the appendix, respectively. Like the main result, an application of the Fundamental Theorem of Calculus together with these auxiliary lemmas yields the following weighted-average interpretation result.

THEOREM 4. *If Assumption 7 is true, then*

$$\beta_1 = E \left[w_{\mathcal{D}}(X) \cdot \frac{\partial M(X)}{\partial X_1} \middle| X \in \mathcal{D} \right]$$

holds for the weight function $w_{\mathcal{D}}$ defined by

$$w_{\mathcal{D}}(x) := \frac{\frac{1}{x_1''(x_{-1}) - x_1'(x_{-1})} \cdot \frac{1}{f_X(x)}}{\int_{\xi \in \mathcal{D}_{-1}} \frac{1}{f_{X_{-1}}(\xi)} dF_{X_{-1}}(\xi)} \quad \text{for all } x \in \mathcal{X},$$

where

$$\begin{aligned} \mathcal{D} &:= \{x \in \mathcal{X} \mid L(x) > M(x)\}, \\ \mathcal{D}_{-1} &:= \{x_{-1} \in \mathbb{R}^{k-1} \mid \exists x_1 \in \mathbb{R} \text{ s.t. } L(x) > M(x)\}, \\ x_1''(x_{-1}) &= \max \{x_1 \in \mathbb{R} \mid L(x_1, x_{-1}) = M(x_1, x_{-1})\}, \quad \text{and} \\ x_1'(x_{-1}) &= \min \{x_1 \in \mathbb{R} \mid L(x_1, x_{-1}) = M(x_1, x_{-1})\}. \end{aligned}$$

A proof is provided in Section A.7 in the appendix. We remark that this proof in particular shows that the weight function $w_{\mathcal{D}}$ is well-defined, as its denominator is strictly positive.

4. LINEAR QR AND STRUCTURAL FUNCTIONS

A general class of structural functions can be expressed by a nonseparable function $g : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ where $\mathcal{U} \subset \mathbb{R}^M$. The cumulative distribution function of U is denoted by F_U . Letting U denote an M -dimensional random vector of unobserved variables supported on \mathcal{U} , we can use g to write the relation among (Y, X, U) by $Y = g(X, U)$.

In the subsequent subsections, we explore relationships between the slopes of the linear QR and the structural partial effects $\partial g / \partial x$, where the latter object measures the *ceteris paribus* causal effects of X on Y .

4.1. Monotone Structural Functions

It is known that, if X is exogenous and g is monotone with respect to a scalar U , i.e., $M = 1$, then the CQF Q_{τ} can be used to represent the structural function g . Specifically, if $M = 1$ and $g(x, \cdot)$ is increasing for each $x \in \mathcal{X}$, then $Q_{\tau}(x) = g(x, u)$ holds for all $x \in \mathcal{X}$ where $\tau = F_U(u)$. In this case, Corollary 1 implies that the slope $\beta(\tau)$ of the τ -th linear QR L_{τ} identifies a weighted average of the structural partial effects $\partial g / \partial x$. We formally present this implication as the following corollary.

COROLLARY 5 (Weighted Average: Monotone Structural Function). *Suppose that $M = 1$, $g(x, \cdot)$ is monotone for each $x \in \mathcal{X}$, g is continuously differentiable, and X is exogenous. If Assumptions 3 and 4 are true for $\tau := F_U(u)$ and $u \in \mathcal{U}$, then there exist $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$ such that*

$$\beta(\tau) = E \left[w_{x_1, x_2}(X) \cdot \frac{\partial g(X, u)}{\partial x} \mid x_1 \leq X \leq x_2 \right]$$

holds, where the weight function w_{x_1, x_2} is defined in Theorem 1.

This result shows that the linear QR parameter $\beta(\tau)$ is useful for us to learn about the structural partial effects $\partial g/\partial x$ when the structural function g is monotone with respect to a scalar latent variable U . In the next subsection, we explore an interpretation of $\beta(\tau)$ while relaxing this monotonicity assumption.

4.2. General Structural Functions

We claim that the linear QR parameter can be still expressed as a weighted average of the structural partial effects even if the monotonicity assumption used in Corollary 5 is dropped. To this end, we combine our Corollary 1 with a result that exists in the literature, which connects the slope Q'_τ of the CQF to the structural partial effects $\partial g/\partial x$ under an arbitrary dimension M of the latent variables U .

We define the lower contour set $V(y, x) = \{u \in \mathbb{R}^M \mid g(x, u) \leq y\}$. Its boundary is denoted by $\partial V(y, x)$. Next, we define the algebra $\mathcal{B}(y, x) := \{S \cap \partial V(y, x) \mid S \in \mathcal{B}(\mathbb{R}^M)\}$ on $\partial V(y, x)$, where $\mathcal{B}(\mathbb{R}^M)$ is the Borel σ -algebra. Note that every element $S \in \mathcal{B}(y, x)$ is also a Borel set. Let m^M denote the Lebesgue measure on \mathbb{R}^M , and let H^{M-1} denote the $(M - 1)$ -dimensional Hausdorff measure restricted to $(\partial V(y, x), \mathcal{B}(y, x))$. The velocities of the boundary $\partial V(y, x)$ at u with respect to a change in y and a change in x are denoted by $\partial v(y, x; u)/\partial y$ and $\partial v(y, x; u)/\partial x$, respectively. With Σ denoting an $(M - 1)$ -dimensional rectangle, the boundary $\partial V(\cdot, \cdot)$ can be represented by a map $\Sigma \times \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}^M$, i.e., the $(M - 1)$ -dimensional manifold $\partial V(y, x)$ can be parameterized through a map $\pi_{(y,x)} : \Sigma \rightarrow \partial V(y, x)$ for each $(y, x) \in \mathcal{Y} \times \mathcal{X}$. The velocity $\partial v(y, \cdot; \cdot)/\partial x$ of the boundary with respect to a change in x can be represented by a map $\mathcal{X} \rightarrow L^1(\Sigma)$ for each $y \in \mathcal{Y}$, provided that the function $\partial v(y, x; \cdot)/\partial x : \partial V(y, x) \rightarrow \mathbb{R}^M$ is absolutely integrable with respect to the parameterization over Σ for each $(y, x) \in \mathcal{Y} \times \mathcal{X}$. Similarly, $\partial v(\cdot, x; \cdot)/\partial y$ can be represented by a map $\mathcal{Y} \rightarrow L^1(\Sigma)$ for each $x \in \mathcal{X}$, provided absolute integrability. We cite the following assumption and auxiliary lemma, which follows from an application of a conservation law for fluid mechanics—details and discussions can be found in Sasaki (2015).

Assumption 8. (i) g is continuously differentiable. (ii) $\|\nabla_u g(x, \cdot)\| \neq 0$ on $\partial V(y, x)$. (iii) The distribution of U is absolutely continuous with respect to m^M . (iv) $\int_{\partial V(y,x)} f_U(u) dH^{M-1}(u) > 0$. (v) $\partial V(y, \cdot) \in C^1(\Sigma \times \mathcal{X}; \mathbb{R}^M)$ holds for each $y \in \mathcal{Y}$, and $\partial V(\cdot, x) \in C^1(\Sigma \times \mathcal{Y}; \mathbb{R}^M)$ holds for each $x \in \mathcal{X}$. (vi) $\partial v(y, \cdot; \cdot)/\partial x \in C^1(\mathcal{X}; L^1(\Sigma))$ holds for each $y \in \mathcal{Y}$, and $\partial v(\cdot, x; \cdot)/\partial y \in C^1(\mathcal{Y}; L^1(\Sigma))$ holds for each $x \in \mathcal{X}$. (vii) X is exogenous. (viii) There exist values $p \geq 1$ and $q \geq 1$ satisfying $p^{-1} + q^{-1} = 1$ such that $\|\gamma(x, \cdot)\|_{L^p(\partial V(y,x), H^{M-1})} < \infty$ and $\|f_U\|_{L^q(\partial V(y,x), H^{M-1})} < \infty$ hold, where $\gamma(x, u) := \|\nabla_u g(x, u)\|^{-1}$.

LEMMA 5 (Sasaki, 2015). *If Assumption 8 is true, then the function $\mu_{y,x} : \mathcal{B}(y, x) \rightarrow \mathbb{R}$ defined by*

$$\mu_{y,x}(S) := \frac{\int_S \frac{f_U(u)}{\|\nabla_u g(x,u)\|} dH^{M-1}(u)}{\int_{\partial V(y,x)} \frac{f_U(u)}{\|\nabla_u g(x,u)\|} dH^{M-1}(u)} \quad \text{for all } S \in \mathcal{B}(y, x)$$

is a probability measure on $(\partial V(y, x), \mathcal{B}(y, x))$, and the slope Q'_τ of the CQF at $\tau = F_{Y|X}(y|x)$ can be written as the weighted mean structural partial effect with respect to $\mu_{y,x}$:

$$Q'_\tau(x) = E_{\mu_{y,x}} \left[\frac{\partial g(x, U)}{\partial x} \right].$$

To make this auxiliary result applicable to our framework, we take an additional step of translating the constructed probability measure $\mu_{y,x}$ into a regular conditional probability measure on $\mathcal{B}(\mathcal{U})$ given the events in $\mathcal{B}(\mathcal{X})$. We define the function $\mu_\tau : \mathcal{B}(\mathcal{U}) \times \mathcal{X} \rightarrow [0, 1]$ by

$$\mu_\tau(S, x) = \mu_{Q_\tau(x), x}(S \cap \partial V(Q_\tau(x), x)). \tag{4.1}$$

By the following lemma, claiming that this μ_τ is a regular conditional probability measure, we can rewrite the conclusion of Lemma 5 formally in terms of the conditional expectation

$$Q'_\tau(x) = E_{\mu_\tau} \left[\frac{\partial g(X, U)}{\partial x} \middle| X = x \right]. \tag{4.2}$$

LEMMA 6. *If Assumption 8 is satisfied, then μ_τ is a regular conditional probability measure on $\mathcal{B}(\mathcal{U})$ given the events in $\mathcal{B}(\mathcal{X})$.*

A proof of this lemma is found in Section A.8 in the appendix. We remark that this weighted measure μ_τ need not be the unique conditional probability measure that relates Q'_τ and $\partial g/\partial x$. For instance, Chernozhukov, Fernández-Val, Hoderlein, Holzmann, and Newey (2015) show that Q'_τ can be written as an unweighted average of $\partial g/\partial x$ under a partial monotonicity assumption in the presence of multiple unobservables. Since the measure μ_τ does not necessarily simplify to the uniform measure under the partial monotonicity assumption, weights connecting Q'_τ and $\partial g/\partial x$ need not be unique.

Thus far, we have developed a marginal probability measure λ_{x_1, x_2} on $\mathcal{B}(\mathcal{X})$ defined in (2.1), and a regular conditional probability measure μ_τ on $\mathcal{B}(\mathcal{U})$ given the events in $\mathcal{B}(\mathcal{X})$ defined in (4.1). Combining the respective mean (2.4) and the conditional mean (4.2) using the law of iterated expectations, we obtain

$$\beta(\tau) = E_{\lambda_{x_1, x_2}} \left[E_{\mu_\tau} \left[\frac{\partial g(X, U)}{\partial x} \middle| X \right] \right] = E_{\mu_\tau \times \lambda_{x_1, x_2}} \left[\frac{\partial g(X, U)}{\partial x} \right].$$

We formally state this result as the following theorem.

THEOREM 5 (Weighted Average: General Structural Function). *If Assumptions 3, 4, and 8 are true, then there exist a marginal probability measure λ_{x_1, x_2} on $\mathcal{B}(\mathcal{X})$ and a regular conditional probability measure μ_τ on $\mathcal{B}(\mathcal{U})$ given the events in $\mathcal{B}(\mathcal{X})$ such that*

$$\beta(\tau) = E_{\mu_\tau \times \lambda_{x_1, x_2}} \left[\frac{\partial g(X, U)}{\partial x} \right]$$

holds.

This result formally characterizes the slope parameter $\beta(\tau)$ of the τ -th linear QR L_τ as a weighted average of the structural partial effects $\partial g/\partial x$ with respect to the measure $\mu_\tau \times \lambda_{x_1, x_2}$. Because the weights are strictly positive on the interval $[x_1, x_2]$, the slope parameter of the linear QR represents a strict convex combination of the structural partial effects within the interval. This implies that, if $\beta(\tau)$ is positive, for example, there exists a nontrivial subpopulation in $[x_1, x_2]$ such that the structural partial effects are positive. In other words, even if the linear QR is quite different from the structural function, the linear QR can be useful for inference of the structural causal effects.

5. LINEAR REGRESSIONS AND STRUCTURAL FUNCTIONS

While the main contribution of this paper is about quantile regressions, we show in this section that the linear regression parameter can be similarly expressed as a weighted average of the structural partial effects. Let λ_U be the probability measure of U supported on \mathcal{U} . We state the following assumption.

Assumption 9. (i) g is continuously differentiable. (ii) $g(x, \cdot) \in \mathcal{L}^1(\lambda_U)$ for every $x \in \mathcal{X}$. (iii) There exists some function $h \in \mathcal{L}^1(\lambda_U)$ such that $\left| \frac{\partial g(x, u)}{\partial x} \right| \leq h(u)$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$.

Under this assumption, the slope of the CEF equals the conditional mean of the structural partial effects, i.e.,

$$m'(x) = E_{\lambda_U} \left[\frac{\partial g(X, U)}{\partial x} \mid X = x \right].$$

Thus, from Corollary 2, we obtain

$$\beta = E_{\lambda_{x_1, x_2}} \left[E_{\lambda_U} \left[\frac{\partial g(X, U)}{\partial x} \mid X \right] \right] = E_{\lambda_U \times \lambda_{x_1, x_2}} \left[\frac{\partial g(X, U)}{\partial x} \right].$$

We formally state this result in the following theorem.

THEOREM 6 (Weighted Average: Linear Regression and General Structural Function). *If Assumptions 5 and 9 are true, then there exists a probability measure λ_{x_1, x_2} on $\mathcal{B}(\mathcal{X})$ such that*

$$\beta = E_{\lambda_U \times \lambda_{x_1, x_2}} \left[\frac{\partial g(X, U)}{\partial x} \right]$$

holds.

This result formally characterizes the slope parameter β of the linear regression L_m as a weighted average of the structural partial effects $\partial g/\partial x$ with respect to the measure $\lambda_U \times \lambda_{x_1, x_2}$. Because the weights are strictly positive on the interval $[x_1, x_2]$, the slope parameter of the linear regression represents a strict convex combination of the structural partial effects within the interval. Therefore, even if

the linear regression is quite different from the structural function, it can be useful for inference of the structural causal effects.

6. CONCLUSION

The slope parameter of the linear QR can be written as weighted averages of the local slopes of the CQF among a nontrivial subpopulation of individuals. Likewise, the slope parameter of the linear regression can be written as weighted averages of the local slopes of the CEF among a nontrivial subpopulation of individuals. These results follow from the property that the linear and nonlinear functions related by weighted least squares must have at least two intersection points. We present an intuitive illustration as well as a formal derivation of this result.

When the structural function g is monotone with respect to a scalar latent variable U , the structural function can be represented by the CQF Q_τ . Therefore, our first result directly implies that the slope parameter $\beta(\tau)$ of the τ -th linear QR measures a weighted average of the structural partial effects $\partial g/\partial x$. Furthermore, even if the structural function g fails to exhibit monotonicity, we establish that the slope parameter $\beta(\tau)$ still measures a weighted average of the structural partial effects $\partial g/\partial x$. We also obtain a similar result for the linear regression.

From these findings, we conclude that the linear regression and the linear QR can be used for causal inference even if they misspecify the true structure both in terms of nonlinearity and nonmonotonicity.

NOTES

1. See Chernozhukov and Hansen (2005) and many subsequent papers.
2. In addition to causal interpretation, the linear misspecification of the linear QR can cause the problem of quantile crossing. Chernozhukov, Fernández-Val, and Galichon (2010) propose rearrangement as a solution.
3. Furthermore, even under endogeneity, the two-stage least squares have the weighted-average interpretation for discrete treatment effects (Angrist and Imbens, 1995).
4. Lee (2014) provides an alternative measure of fit for the linear QR parameters.
5. See also Koenker and Hallock (2001), Koenker (2005), and Chernozhukov and Hansen (2013) for surveys.
6. We thank K. Kato for suggesting this extension.
7. We impose this additional assumption as we move from univariate X to multivariate X , because multivariate extension is not straightforward otherwise. Similar difficulties arose also for the weighted-average interpretations of the linear regressions by Yitzhaki (1996) and Angrist and Krueger (1999).

REFERENCES

- Angrist, J., V. Chernozhukov, & I. Fernández-Val (2006) Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* 74(2), 539–563.
- Angrist, J.D. & G.W. Imbens (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 90(430), 431–442.
- Angrist, J.D. & A.B. Krueger (1999) Empirical strategies in labor economics. In O.C. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. 3, Part A, 1277–1366.

Chamberlain, G. (1984) Panel data. In Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, 1247–1318.

Chernozhukov, V., I. Fernández-Val, & A. Galichon (2010) Quantile and probability curves without crossing. *Econometrica* 74(2), 539–563.

Chernozhukov, V., I. Fernández-Val, S. Hoderlein, H. Holzmann, & W. Newey (2015) Nonparametric identification in panels using quantiles. *Journal of Econometrics* 188(2), 378–392.

Chernozhukov, V. & C. Hansen (2005) An IV model of quantile treatment effects. *Econometrica* 73(1), 245–261.

Chernozhukov, V. & C. Hansen (2013) Quantile models with endogeneity. *Annual Review of Economics* 5(1), 57–81.

Goldberger, A.S. (1991) *A Course in Econometrics*. Harvard University Press.

Imbens, G.W. & W.K. Newey (2009) Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77(5), 1481–1512.

Koenker, R. (2005) *Quantile Regression*. Cambridge University Press.

Koenker, R. & G. Bassett, Jr. (1978) Regression quantiles. *Econometrica* 46(1), 33–50.

Koenker, R. & K.F. Hallock (2001) Quantile regression. *Journal of Economic Perspectives* 15(4), 143–156.

Lee, Y.-Y. (2013) Interpretation and semiparametric efficiency in quantile regression under misspecification. Working Paper, University of Wisconsin-Madison.

Sasaki, Y. (2015) What do quantile regressions identify for general structural functions? *Econometric Theory* 31(5), 1102–1116.

White, H. (1980) Using least squares to approximate unknown regression functions. *International Economic Review* 21(1), 149–170.

Yitzhaki, S. (1996) On using linear regressions in welfare economics. *Journal of Business and Economic Statistics* 14 (4), 478–486.

A Mathematical Appendix

A.1. Proof of Lemma 1

Proof. By way of contradiction, suppose that M and L intersect at most at one point of x . Define the short-hand notation

$$\Delta(x, \alpha, \beta) := \alpha + \beta x - M(x) = L(x) - M(x).$$

Under the current assumption, we have at most one zero for $\Delta(\cdot, \alpha, \beta)$ in \mathcal{X} . By Assumption 1(iv)–(v), it follows that either one of the following three cases is true.

- (I) There exists a point $x^* \in \mathcal{X}$ such that $\Delta(x, \alpha, \beta) \cdot \Delta(x', \alpha, \beta) < 0$ for all $x \in \mathcal{X} \cap (x^*, \infty)$ and for all $x' \in \mathcal{X} \cap (-\infty, x^*)$.
- (II) There exists a point $x^* \in \mathcal{X}$ such that $\Delta(x^*, \alpha, \beta) = 0$, but $\Delta(x, \alpha, \beta) \cdot \Delta(x', \alpha, \beta) > 0$ for all $x, x' \in \mathcal{X}$ such that $x \neq x^*$ and $x' \neq x^*$.
- (III) $\Delta(x, \alpha, \beta) \cdot \Delta(x', \alpha, \beta) > 0$ for all $x, x' \in \mathcal{X}$.

We claim below that each of the cases (I)–(III) contradicts Assumption 1(ii) that (α, β) uniquely solves $\min_{(a,b)} E[\omega(X) \cdot (M(X) - a - bX)^2]$.

First, consider case (I). Without loss of generality, we normalize the location to $x^* = 0$ and assume $\Delta(x, \alpha, \beta) < 0$ for all $x > 0$. For each $\epsilon > 0$, define the set

$$A(\epsilon) = \{x \in \mathbb{R}_{++} \cap \mathcal{X} \mid \Delta(x, \alpha, \beta + \epsilon) \leq 0\} \cup \{x \in \mathbb{R}_{--} \cap \mathcal{X} \mid \Delta(x, \alpha, \beta + \epsilon) \geq 0\}.$$

We let B^1 and B^2 denote arbitrary compact intervals contained in $\mathcal{X} \cap \mathbb{R}_{--}$ and $\mathcal{X} \cap \mathbb{R}_{++}$, respectively. Note that $\min_{x \in B^1 \cup B^2} \left\{ \frac{\Delta(x, \alpha, \beta)}{-x} \right\}$ exists due to the compactness of $B^1 \cup B^2$ and the continuity of $\frac{\Delta(x, \alpha, \beta)}{-x}$ with respect to x on $B^1 \cup B^2$. Thus, if we choose $\epsilon := \min_{x \in B^1 \cup B^2} \left\{ \frac{\Delta(x, \alpha, \beta)}{-x} \right\}$, then $\Delta(x, \alpha, \beta + \epsilon) \geq 0$ for all $x \in B^1$ and $\Delta(x, \alpha, \beta + \epsilon) \leq 0$ for all $x \in B^2$. Hence, for any compact intervals $B^1 \subset \mathcal{X} \cap \mathbb{R}_{--}$ and $B^2 \subset \mathcal{X} \cap \mathbb{R}_{++}$, there exists $\epsilon > 0$ such that $(B^1 \cup B^2) \subset A(\epsilon)$. Furthermore, note that $\epsilon > 0$ is true. Now, observe from the definition of $A(\epsilon)$ that the inequality

$$\begin{aligned} & E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] - E[(\Delta^2(X, \alpha, \beta + \epsilon) \cdot \omega(x))] \\ &= \int_{\mathcal{X}} (-2\Delta(x, \alpha, \beta)\epsilon x - (\epsilon x)^2) \cdot \omega(x) dF_X(x) \\ &\geq \int_{A(\epsilon)} (\epsilon x)^2 \cdot \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A(\epsilon)} (\epsilon x)^2 \cdot \omega(x) dF_X(x) \end{aligned} \tag{A.1}$$

holds for any $\epsilon > 0$. Let $B_n := \mathcal{X} \cap \left(\left[-n, -\frac{1}{n} \right] \cup \left[\frac{1}{n}, n \right] \right)$ for each integer $n > 1$. Then for each $n > 1$, the above argument implies that there exists $\epsilon_n > 0$ such that $B_n \subset A(\epsilon_n)$. But then, for each $n > 1$, there exists $\epsilon_n > 0$ such that

$$\begin{aligned} & \int_{B_n} x^2 \cdot \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus B_n} x^2 \cdot \omega(x) dF_X(x) \\ &\leq \int_{A(\epsilon_n)} x^2 \cdot \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A(\epsilon_n)} x^2 \cdot \omega(x) dF_X(x). \end{aligned} \tag{A.2}$$

By Assumption 1 (i) and (iii) and the Monotone Convergence Theorem,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \int_{\mathcal{X} \setminus B_n} x^2 \cdot \omega(x) dF_X(x) \\ &= \lim_{n \rightarrow \infty} \int_{\mathcal{X}} x^2 \cdot \omega(x) \cdot f_X(x) \cdot \mathbb{1}\{x \in \mathcal{X} \setminus B_n\} dx \\ &= \int_{\mathcal{X}} \lim_{n \rightarrow \infty} x^2 \cdot \omega(x) \cdot f_X(x) \cdot \mathbb{1}\{x \in \mathcal{X} \setminus B_n\} dx = 0 \end{aligned} \tag{A.3}$$

is true. Likewise, Assumption 1 (i) and (iii) and the Monotone Convergence Theorem yield

$$\lim_{n \rightarrow \infty} \int_{B_n} x^2 \cdot \omega(x) dF_X(x) = \int_{\mathcal{X}} x^2 \cdot \omega(x) \cdot f_X(x) dx =: c. \tag{A.4}$$

Note that $c \geq 0$, and it holds with equality only if $\omega(x) \cdot f_X(x) = 0$ almost everywhere on \mathcal{X} . But it is not true that $\omega(x) \cdot f_X(x) = 0$ almost everywhere on \mathcal{X} , from Assumption 1(ii) that (α, β) uniquely minimizes the weighted mean squared distance. Thus, it follows from (A.1)–(A.4) that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] - E[(\Delta^2(X, \alpha, \beta + \epsilon_n) \cdot \omega(x))]}{\epsilon_n^2} \\ & \qquad \qquad \qquad \geq c > 0 \end{aligned}$$

is true. But then, there exists $n^* > 1$ such that

$$\frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] - E[(\Delta^2(X, \alpha, \beta + \epsilon_{n^*}) \cdot \omega(x))]}{\epsilon_{n^*}^2} > 0.$$

This inequality implies that

$$E \left[\Delta^2(X, \alpha, \beta) \cdot \omega(x) \right] > E \left[(\Delta^2(X, \alpha, \beta + \epsilon_{n^*}) \cdot \omega(x)) \right],$$

and it contradicts Assumption 1(ii), that (α, β) uniquely minimizes the weighted mean squared distance.

Next, consider case (II). Without loss of generality, we normalize the location to $x^* = 0$ and assume $\Delta(x, \alpha, \beta) > 0$ for all $x \neq 0$. For each $\epsilon > 0$, define the set

$$A'(\epsilon) = \{x \in \mathcal{X} \mid \Delta(x, \alpha - \epsilon, \beta) \geq 0\}.$$

Let B^1 and B^2 denote arbitrary compact intervals contained in $\mathcal{X} \cap \mathbb{R}_{--}$ and $\mathcal{X} \cap \mathbb{R}_{++}$, respectively. Note that $\min_{x \in B^1 \cup B^2} \{\Delta(x, \alpha, \beta)\}$ exists due to the compactness of $B^1 \cup B^2$ and the continuity of $\Delta(x, \alpha, \beta)$ with respect to x on $B^1 \cup B^2$. If we choose $\epsilon := \min_{x \in B^1 \cup B^2} \{\Delta(x, \alpha, \beta)\}$, then $\Delta(x, \alpha - \epsilon, \beta) \geq 0$ for all $x \in B^1 \cup B^2$. Hence, for any compact intervals $B^1 \subset \mathcal{X} \cap \mathbb{R}_{--}$ and $B^2 \subset \mathcal{X} \cap \mathbb{R}_{++}$, there exists $\epsilon > 0$ such that $(B^1 \cup B^2) \subset A'(\epsilon)$. Furthermore, note that $\epsilon > 0$ is true. It follows that the inequality

$$\begin{aligned} E \left[\Delta^2(X, \alpha, \beta) \cdot \omega(x) \right] &- E \left[\Delta^2(X, \alpha - \epsilon, \beta) \cdot \omega(x) \right] \\ &= \int_{\mathcal{X}} \left(2\Delta(x, \alpha, \beta)\epsilon - \epsilon^2 \right) \cdot \omega(x) dF_X(x) \\ &\geq \int_{A'(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A'(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) \end{aligned} \tag{A.5}$$

holds for any $\epsilon > 0$. By the same argument as the one used in case (I), for each integer $n > 1$, there exists $\epsilon_n > 0$ such that

$$\begin{aligned} \int_{B_n} \omega(x) dF_X(x) &- \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) \\ &\leq \int_{A'(\epsilon_n)} \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A'(\epsilon_n)} \omega(x) dF_X(x), \end{aligned} \tag{A.6}$$

where

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) = 0 \quad \text{and} \tag{A.7}$$

$$\lim_{n \rightarrow \infty} \int_{B_n} \omega(x) dF_X(x) = \int_{\mathcal{X}} \omega(x) \cdot f_X(x) dx =: c' > 0. \tag{A.8}$$

Thus, it follows from (A.5)–(A.8) that

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{E \left[\Delta^2(X, \alpha, \beta) \cdot \omega(x) \right] - E \left[(\Delta^2(X, \alpha - \epsilon_n, \beta) \cdot \omega(x)) \right]}{\epsilon_n^2} \\ \geq c' > 0 \end{aligned}$$

is true. But then, there exists $n^{**} > 1$ such that

$$\frac{E \left[\Delta^2(X, \alpha, \beta) \cdot \omega(x) \right] - E \left[(\Delta^2(X, \alpha - \epsilon_{n^{**}}, \beta) \cdot \omega(x)) \right]}{\epsilon_{n^{**}}^2} > 0.$$

This inequality implies that

$$E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] > E[(\Delta^2(X, \alpha - \epsilon_{n^{**}}, \beta) \cdot \omega(x))],$$

and it contradicts Assumption 1(ii), that (α, β) uniquely minimizes the weighted mean squared distance.

Lastly, consider case (III). Without loss of generality, assume that $\Delta(x, \alpha, \beta) > 0$ for all $x \in \mathcal{X}$. As in case (II), if B^1 and B^2 are arbitrary compact intervals contained in $\mathcal{X} \cap \mathbb{R}_{--}$ and $\mathcal{X} \cap \mathbb{R}_{++}$, respectively, then there exists $\epsilon > 0$ such that $(B^1 \cup B^2) \subset A'(\epsilon)$. Thus,

$$E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] - E[(\Delta(X, \alpha - \epsilon, \beta))^2 \cdot \omega(x)] \geq \int_{A'(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A'(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) \tag{A.9}$$

holds for any $\epsilon > 0$ similarly to case (II). By the same argument as the one used in case (I), for each integer $n > 1$, there exists $\epsilon_n > 0$ such that

$$\int_{B_n} \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) \leq \int_{A'(\epsilon_n)} \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A'(\epsilon_n)} \omega(x) dF_X(x), \tag{A.10}$$

where

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) = 0 \quad \text{and} \tag{A.11}$$

$$\lim_{n \rightarrow \infty} \int_{B_n} \omega(x) dF_X(x) = \int_{\mathcal{X}} \omega(x) \cdot f_X(x) dx =: c' > 0. \tag{A.12}$$

Thus, it follows from (A.9)–(A.12) that

$$\lim_{n \rightarrow \infty} \frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] - E[(\Delta^2(X, \alpha - \epsilon_n, \beta) \cdot \omega(x))]}{\epsilon_n^2} \geq c' > 0$$

is true, similarly to case (II). But then, there exists $n^{**} > 1$ such that

$$\frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] - E[(\Delta^2(X, \alpha - \epsilon_{n^{**}}, \beta) \cdot \omega(x))]}{\epsilon_{n^{**}}^2} > 0.$$

This inequality implies that

$$E[\Delta^2(X, \alpha, \beta) \cdot \omega(x)] > E[(\Delta^2(X, \alpha - \epsilon_{n^{**}}, \beta) \cdot \omega(x))],$$

and it contradicts Assumption 1(ii), that (α, β) uniquely minimizes the weighted mean squared distance. ■

A.2. Proof of Theorem 1

Proof. From Lemma 1, Assumption 1 guarantees the existence of two points $x_1, x_2 \in \mathcal{X}$ such that $L(x_1) = M(x_1)$ and $L(x_2) = M(x_2)$ are both true. Without loss of generality, let $x_1 < x_2$. By Assumption 2, we have $M(x_2) - M(x_1) = \int_{x_1}^{x_2} M'(\xi) d\xi$ by the Fundamental Theorem of Calculus. Similarly, we have $L(x_2) - L(x_1) = \int_{x_1}^{x_2} L'_\tau(\xi) d\xi = \beta \cdot (x_2 - x_1)$. Combining all these equalities together yields

$$\int_{x_1}^{x_2} M'(\xi) d\xi = \beta \cdot (x_2 - x_1).$$

By Assumption 1(v), $f_X > 0$ almost everywhere on $[x_1, x_2]$. Thus, we can write

$$\beta = \frac{\int_{x_1}^{x_2} M'(\xi) d\xi}{\int_{x_1}^{x_2} d\xi} = \frac{\int_{x_1}^{x_2} \frac{M'(\xi)}{f_X(\xi)} dF_X(\xi)}{\int_{x_1}^{x_2} \frac{1}{f_X(\xi)} dF_X(\xi)} = E[w_{x_1, x_2}(X) \cdot M'(X) | x_1 \leq X \leq x_2].$$



A.3. Proof of Lemma 2

Proof. By way of contradiction, suppose that $\beta - \Delta M(x) < 0$ is true for all $x \in \mathcal{X}$. We consider the following two cases.

- (I) There exist x, x' such that $L(x) - M(x) < 0$ and $L(x') - M(x') > 0$.
- (II) $L(x) - M(x) \geq 0$ for all $x \in \mathcal{X}$.

We derive a contradiction under each of these two cases to complete a proof. Specifically, we show below that each of the cases (I) and (II) contradicts Assumption 6(ii) that (α, β) uniquely solves $\min_{(a,b)} E[\omega(X) \cdot (M(X) - a - bX)^2]$.

First, we consider case (I). Since $\beta - \Delta M(x) < 0$ for all $x \in \mathcal{X}$, if $\Delta(x, \alpha, \beta) < 0$ for some $x \in \mathcal{X}$, then $\Delta(\tilde{x}, \alpha, \beta) < 0$ for all $\tilde{x} > x$. Also, if $\Delta(x', \alpha, \beta) > 0$ for some $x' \in \mathcal{X}$, then $\Delta(\tilde{x}, \alpha, \beta) > 0$ for all $\tilde{x} < x'$. Hence, there exists a unique $x^* \in \mathcal{X}$ such that $\Delta(x, \alpha, \beta) \geq 0$ for all $x \leq x^*$ and $\Delta(x, \alpha, \beta) < 0$ for all $x > x^*$. Without loss of generality, we normalize the location to $x^* = 0$. For each $\epsilon > 0$, define the set

$$A(\epsilon) = \{x \in \mathcal{X} \cap \mathbb{Z}_+ \mid \Delta(x, \alpha, \beta + \epsilon) \leq 0\} \cup \{x \in \mathcal{X} \cap \mathbb{Z}_- \mid \Delta(x, \alpha, \beta + \epsilon) \geq 0\}.$$

We let B denote an arbitrary nonempty finite subset of \mathcal{X} .

If we choose $\epsilon := \min_{x \in B} \left\{ \frac{\Delta(x, \alpha, \beta)}{-x} \right\}$, then $\Delta(x, \alpha, \beta + \epsilon) \geq 0$ for all $x \in B \cap \mathcal{X}_+$ and $\Delta(x, \alpha, \beta + \epsilon) \leq 0$ for all $x \in B \cap \mathcal{X}_-$. Hence, for any finite subset B , there exists $\epsilon > 0$ such that $B \subset A(\epsilon)$. Furthermore, note that $\epsilon > 0$ is true. Now, observe from the definition of $A(\epsilon)$ that the inequality

$$\begin{aligned} & E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[(\Delta^2(X, \alpha, \beta + \epsilon) \cdot \omega(x))] \\ &= \sum_{x \in \mathcal{X}} (-2\Delta(x, \alpha, \beta)\epsilon x - (\epsilon x)^2) \cdot \omega(x) \cdot p(x) \\ &\geq \sum_{x \in A(\epsilon)} (\epsilon x)^2 \cdot \omega(x) \cdot p(x) - \sum_{x \in \mathcal{X} \setminus A(\epsilon)} (\epsilon x)^2 \cdot \omega(x) \cdot p(x) \end{aligned} \tag{A.13}$$

holds for any $\epsilon > 0$. Let $B_n := \{x \in \mathcal{X} \mid -n \leq x \leq n\}$ for each integer $n > 1$. Then for each, $n > 1$, the preceding argument implies that there exists $\epsilon_n > 0$ such that $B_n \subset A(\epsilon_n)$. But then, for each $n > 1$, there exists $\epsilon_n > 0$ such that

$$\begin{aligned} & \sum_{x \in B_n} x^2 \cdot \omega(x) \cdot p(x) - \sum_{x \in \mathcal{X} \setminus B_n} x^2 \cdot \omega(x) \cdot p(x) \\ & \leq \sum_{x \in A(\epsilon_n)} x^2 \cdot \omega(x) \cdot p(x) - \sum_{x \in \mathcal{X} \setminus A(\epsilon_n)} x^2 \cdot \omega(x) \cdot p(x). \end{aligned} \tag{A.14}$$

By Assumption 6 (i), (iii),

$$\lim_{n \rightarrow \infty} \sum_{x \in \mathcal{X} \setminus B_n} x^2 \cdot \omega(x) \cdot p(x) = 0 \tag{A.15}$$

is true. Likewise, Assumption 6 (i), (iii) yield

$$\lim_{n \rightarrow \infty} \sum_{x \in B_n} x^2 \cdot \omega(x) \cdot p(x) = \sum_{x \in \mathcal{X}} x^2 \cdot \omega(x) \cdot p(x) =: c(\tau). \tag{A.16}$$

Note that $c(\tau) \geq 0$, and it holds with equality only if $\omega(x) \cdot p(x) = 0$ for all $x \in \mathcal{X}$. But it is not true that $\omega(x) \cdot p(x) = 0$ for all $x \in \mathcal{X}$, from the property of (α, β) that it uniquely solves $\min_{(a,b)} E[\omega(X)(M(X) - a - bX)]$. Thus, it follows from (A.13)–(A.16) that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[(\Delta^2(X, \alpha, \beta + \epsilon_n) \cdot \omega(X))]}{\epsilon_n^2} \\ & \geq c(\tau) > 0 \end{aligned}$$

is true. But then, there exists $n^* > 1$ such that

$$\frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[(\Delta^2(X, \alpha, \beta + \epsilon_{n^*}) \cdot \omega(X))]}{\epsilon_{n^*}^2} > 0.$$

This inequality implies that

$$E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] > E[(\Delta^2(X, \alpha, \beta + \epsilon_{n^*}) \cdot \omega(X))],$$

and it contradicts Assumption 6(ii), that (α, β) uniquely solves $\min_{(a,b)} E[\omega(X) \cdot (M(X) - a - bX)^2]$.

Next, we consider case (II). Suppose there exists x^{**} such that $\Delta(x^{**}, \alpha, \beta) = 0$. If $x^* = \max \mathcal{X}$, we can derive a contradiction from an argument similar to case (I). If $x^* \neq \max \mathcal{X}$, then since $\beta - \Delta M(x) < 0$ for all $x \in \mathcal{X}$, $\Delta(x, \alpha, \beta) < 0$ for all $x > x^{**}$. This contradicts $\Delta(x, \alpha, \beta) \geq 0$ for all $x \in \mathcal{X}$. Hence, $\Delta(x, \alpha, \beta) > 0$ for all $x \in \mathcal{X}$. For each $\epsilon > 0$, define the set

$$A'(\epsilon) = \{x \in \mathcal{X} \mid \Delta(x, \alpha - \epsilon, \beta) \geq 0\}.$$

Let B denote an arbitrary nonempty finite subset of \mathcal{X} . If we choose $\epsilon := \min_{x \in B} \{\Delta(x, \alpha, \beta)\}$, then $\Delta(x, \alpha - \epsilon, \beta) \geq 0$ for all $x \in B$. Hence, for any finite subset

$B \subset \mathcal{X}$, there exists $\epsilon > 0$ such that $B \subset A'(\epsilon)$. Furthermore, note that $\epsilon > 0$ is true. It follows that the inequality

$$\begin{aligned} E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[\Delta^2(X, \alpha - \epsilon, \beta) \cdot \omega(X)] \\ = \sum_{x \in \mathcal{X}} (2\Delta(x, \alpha, \beta)\epsilon - \epsilon^2) \cdot \omega(x) \cdot p(x) \\ \geq \sum_{x \in A'(\epsilon)} \epsilon^2 \cdot \omega(x) \cdot p(x) - \sum_{x \in \mathcal{X} \setminus A'(\epsilon)} \epsilon^2 \cdot \omega(x) \cdot p(x) \end{aligned} \tag{A.17}$$

holds for any $\epsilon > 0$. Let $B_n = \{x \in \mathcal{X} \mid -n \leq x \leq n\}$ for each integer $n > 1$. Then, for each $n > 1$, there exists ϵ_n such that $B_n \subset A'(\epsilon_n)$. But then, for each $n > 1$, there exists $\epsilon_n > 0$ such that,

$$\begin{aligned} \sum_{x \in B_n} \omega(x) \cdot p(x) - \sum_{x \in \mathcal{X} \setminus B_n} \omega(x) \cdot p(x) \\ \leq \sum_{x \in A'(\epsilon_n)} \omega(x) \cdot p(x) - \sum_{x \in \mathcal{X} \setminus A'(\epsilon_n)} \omega(x) \cdot p(x). \end{aligned} \tag{A.18}$$

Since $\sum_{x \in \mathcal{X}} \omega(x) \cdot p(x) < \infty$,

$$\lim_{n \rightarrow \infty} \sum_{x \in \mathcal{X} \setminus B_n} \omega(x) \cdot p(x) = 0 \quad \text{and} \tag{A.19}$$

$$\lim_{n \rightarrow \infty} \sum_{x \in B_n} \omega(x) \cdot p(x) = \sum_{x \in \mathcal{X}} \omega(x) \cdot p(x) = c'(\tau) > 0. \tag{A.20}$$

Thus, it follows from (A.17)–(A.20) that

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[(\Delta^2(X, \alpha - \epsilon_n, \beta) \cdot \omega(X))]}{\epsilon_n^2} \\ \geq c'(\tau) > 0 \end{aligned}$$

is true. But then, there exists $n^{**} > 1$ such that

$$\frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[(\Delta^2(X, \alpha - \epsilon_{n^{**}}, \beta) \cdot \omega(X))]}{\epsilon_{n^{**}}^2} > 0.$$

This inequality implies that

$$E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] > E[(\Delta^2(X, \alpha - \epsilon_{n^{**}}, \beta) \cdot \omega(X))],$$

and it contradicts Assumption 6(ii), that (α, β) uniquely solves $\min_{(a,b)} E[\omega(X) \cdot (M(X) - a - bX)^2]$. ■

A.4. Proof of Theorem 3

Proof. First, consider the case where there exists $\tilde{x} \in \mathcal{X}$ such that $\Delta L(\tilde{x}) - \Delta M(\tilde{x}) = 0$. In this case,

$$\beta = \Delta L(\tilde{x}) = \Delta M(\tilde{x}) = E[\tilde{w}(X) \cdot \Delta M(X)] \tag{A.21}$$

holds, where $\tilde{w}(x) = 1/p(x)$ if $x = \tilde{x}$ and $\tilde{w}(x) = 0$ otherwise.

Next, consider the case where $\Delta L(x) - \Delta M(x) \neq 0$ for all $x \in \mathcal{X}$. By Lemma 2, there exist $x_1, x_2 \in \mathcal{X}$ such that $(\Delta L(x_1) - \Delta M(x_1)) \cdot (\Delta L(x_2) - \Delta M(x_2)) < 0$. Thus,

$$\begin{aligned} \beta &= \frac{\beta - \Delta M(x_2)}{\Delta M(x_1) - \Delta M(x_2)} \cdot \Delta M(x_1) + \frac{\Delta M(x_1) - \beta}{\Delta M(x_1) - \Delta M(x_2)} \cdot \Delta M(x_2) \\ &= E[\lambda(X) \cdot \Delta M(X)] \end{aligned}$$

holds, where $\lambda(x_1) = \frac{\beta - \Delta M(x_2)}{\Delta M(x_1) - \Delta M(x_2)} \cdot \frac{1}{p(x_1)}$, $\lambda(x_2) = \frac{\Delta M(x_1) - \beta}{\Delta M(x_1) - \Delta M(x_2)} \cdot \frac{1}{p(x_2)}$, and $\lambda(x) = 0$ for all $x \in \mathcal{X} \setminus \{x_1, x_2\}$. Note that this weight function λ is non-negative because of $(\Delta L(x_1) - \Delta M(x_1)) \cdot (\Delta L(x_2) - \Delta M(x_2)) \leq 0$. ■

A.5. Proof of Lemma 3

Proof. Define the short-hand notation

$$\Delta(x, \alpha, \beta) := \alpha + x'\beta - M(x) = L(x) - M(x).$$

By way of contradiction, suppose $M \geq L$ for all $x \in \mathcal{X}$. Then, either one of the following two cases is true.

- (I) There exists a point $x^* \in \mathcal{X}$ such that $\Delta(x^*, \alpha, \beta) = 0$.
- (II) $\Delta(x, \alpha, \beta) < 0$ for all $x \in \mathcal{X}$.

We claim below that each of the cases (I)–(II) contradicts Assumption 7 that (α, β) uniquely solves $\min_{(a,b)} E[\omega(X) \cdot \Delta^2(X, a, \beta)]$.

First, consider case (I). We show that x^* is unique. Suppose there exists $x^{**} \neq x^*$ such that $\Delta(x^{**}, \alpha, \beta) = 0$. Since M is strictly convex from Assumption 7(vii), for any $0 < \lambda < 1$, $M(\lambda x^* + (1 - \lambda)x^{**}) < \lambda \cdot M(x^*) + (1 - \lambda)M(x^{**}) = \alpha + (\lambda x^* + (1 - \lambda)x^{**})'\beta$. This contradicts $M \geq L$ for all $x \in \mathcal{X}$. Hence, x^* is unique. Without loss of generality, we normalize the location to $x^* = 0$. For each $\epsilon > 0$, define the set

$$A(\epsilon) = \{x \in \mathcal{X} \mid -\epsilon - \Delta(x, \alpha, \beta) \geq 0\}.$$

Let B^1 denote an arbitrary compact rectangle in \mathcal{X} and B^2 denote an open rectangle in \mathcal{X} such that $B^2 \subset B^1$ and $0 \in B^2$. Let $B := B^1 \setminus B^2$. Note that B is compact in \mathcal{X} and $\min_{x \in B} \{-\Delta(x, \alpha, \beta)\}$ exists due to the compactness of B and the continuity of $\Delta(x, \alpha, \beta)$ with respect to x on B .

If we choose $\epsilon := \min_{x \in B} \{-\Delta(x, \alpha, \beta)\}$, then $-\epsilon - \Delta(x, \alpha, \beta) \geq 0$ for all $x \in B$. Hence, for any B^1, B^2 which satisfies the aforementioned property, there exists $\epsilon > 0$ such that $B \subset A(\epsilon)$. Furthermore, note that $\epsilon > 0$ is true. Now observe from the definition of $A(\epsilon)$ that the inequality

$$\begin{aligned} E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] &- E[(\Delta^2(X, \alpha + \epsilon, \beta) \cdot \omega(X))] \\ &= \int_{\mathcal{X}} (-2\Delta(x, \alpha, \beta)\epsilon - \epsilon^2) \cdot \omega(x) dF_X(x) \\ &\geq \int_{A(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) \end{aligned} \tag{A.22}$$

holds for any $\epsilon > 0$. Let $B_n^1 := [-n, n]^k$, $B_n^2 := (-\frac{1}{n}, \frac{1}{n})^k$, and $B_n = B_n^1 \setminus B_n^2$ for each integer $n > 1$. Then, for each integer $n > 1$, there exists $\epsilon_n > 0$ such that $B_n \subset A(\epsilon_n)$. But then, for each $n > 1$, there exists $\epsilon_n > 0$ such that

$$\begin{aligned} & \int_{B_n} \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) \\ & \leq \int_{A(\epsilon_n)} \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A(\epsilon_n)} \omega(x) dF_X(x). \end{aligned} \tag{A.23}$$

By Assumption 7(iii) and the Monotone Convergence Theorem,

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) = 0 \quad \text{and} \tag{A.24}$$

$$\lim_{n \rightarrow \infty} \int_{B_n} \omega(x) dF_X(x) = \int_{\mathcal{X}} \omega(x) \cdot f_X(x) dx =: c' > 0. \tag{A.25}$$

Thus, it follows from (A.22)–(A.25) that

$$\lim_{n \rightarrow \infty} \frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[\Delta^2(X, \alpha + \epsilon_n, \beta) \cdot \omega(X)]}{\epsilon_n^2} \geq c' > 0$$

is true. But then, there exists $n^{**} > 1$ such that

$$\frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[\Delta^2(X, \alpha + \epsilon_{n^{**}}, \beta) \cdot \omega(X)]}{\epsilon_{n^{**}}^2} > 0.$$

This inequality implies that

$$E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] > E[\Delta^2(X, \alpha + \epsilon_{n^{**}}, \beta) \cdot \omega(X)],$$

and it contradicts Assumption 7(ii), that β uniquely minimizes the weighted mean squared distance.

Next, we consider case (II). As in case (I), if B^1 and B^2 satisfy the property mentioned in case (I), then there exists $\epsilon > 0$ such that $(B^1 \setminus B^2) \subset A(\epsilon)$. Thus,

$$\begin{aligned} & E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[(\Delta^2(X, \alpha + \epsilon, \beta) \cdot \omega(X))] \\ & \geq \int_{A(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A(\epsilon)} \epsilon^2 \cdot \omega(x) dF_X(x) \end{aligned} \tag{A.26}$$

holds for any $\epsilon > 0$ similarly to case (I). By the same argument as the one used in case (I), for each integer $n > 1$, there exists $\epsilon_n > 0$ such that

$$\begin{aligned} & \int_{B_n} \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) \\ & \leq \int_{A(\epsilon_n)} \omega(x) dF_X(x) - \int_{\mathcal{X} \setminus A(\epsilon_n)} \omega(x) dF_X(x). \end{aligned} \tag{A.27}$$

Assumption 7(iii) and Monotone Convergence Theorem yield

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X} \setminus B_n} \omega(x) dF_X(x) = 0 \quad \text{and} \tag{A.28}$$

$$\lim_{n \rightarrow \infty} \int_{B_n} \omega(x) dF_X(x) = \int_{\mathcal{X}} \omega(x) \cdot f_X(x) dx =: c'. \tag{A.29}$$

Note that $c \geq 0$, and it holds with equality only if $\omega(x, \alpha, \beta\tau) \cdot f_X(x) = 0$ almost everywhere on \mathcal{X} . But it is not true that $\omega(X) \cdot f_X(x) = 0$ almost everywhere on \mathcal{X} , from Assumption 7(ii) that (α, β) uniquely minimizes the weighted mean squared distance.

Thus, it follows from (A.26)–(A.29) that

$$\lim_{n \rightarrow \infty} \frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[\Delta^2(X, \alpha + \epsilon_n, \beta) \cdot \omega(X)]}{\epsilon_n^2} \geq c' > 0$$

is true, similarly to case (II). But then, there exists $n^{**} > 1$ such that

$$\frac{E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] - E[\Delta^2(X, \alpha + \epsilon_{n^{**}}, \beta) \cdot \omega(X)]}{\epsilon_{n^{**}}^2} > 0.$$

This inequality implies that

$$E[\Delta^2(X, \alpha, \beta) \cdot \omega(X)] > E[\Delta^2(X, \alpha + \epsilon_{n^{**}}, \beta) \cdot \omega(X)],$$

and it contradicts Assumption 7(ii), that β uniquely minimizes the weighted mean squared distance. ■

A.6. Proof of Lemma 4

Proof. From Assumption 7(vii), there exist $x'_1 > x_1^*$ and $x''_1 < x_1^*$ such that $\frac{\partial M(x'_1, x_{-1}^*)}{\partial x_1} > \beta_1$ and $\frac{\partial M(x''_1, x_{-1}^*)}{\partial x_1} < \beta_1$.

If $M(x'_1, x_{-1}^*) > L(x'_1, x_{-1}^*)$, then by the Intermediate Value Theorem there exists $\tilde{x}_1 \in [x_1^*, x'_1]$ such that $M(\tilde{x}_1, x_{-1}^*) = L(\tilde{x}_1, x_{-1}^*)$. Now consider the case where $M(x'_1, x_{-1}^*) < L(x'_1, x_{-1}^*)$. Let $\bar{\beta}_1 := \frac{\partial M(x'_1, x_{-1}^*)}{\partial x_1}$. Recall that $\bar{\beta} > \beta$. From Assumption 7(vii), for all $x_1 > x'_1$,

$$M(x_1, x_{-1}^*) > \bar{\beta}_1(x_1 - x'_1) + x_{-1}^* \beta_{-1} + M(x'_1, x_{-1}^*). \tag{A.30}$$

Note that $\bar{\beta}_1(x_1 - x'_1) + x_{-1}^* \beta_{-1} + M(x'_1, x_{-1}^*)$, as a function of x_1 , is the tangent line to the curve $M(\cdot, x_{-1}^*)$ at the point $x_1 = x'_1$. Since $\bar{\beta}_1 > \beta_1$, $\bar{\beta}_1(x_1 - x'_1) + x_{-1}^* \beta_{-1} + M(x'_1, x_{-1}^*)$ and $L(x_1, x_{-1}^*)$ cross at the point $x_1 = \frac{M(x'_1, x_{-1}^*)}{\bar{\beta} - \beta}$. Further note that $\frac{M(x'_1, x_{-1}^*)}{\bar{\beta} - \beta} > x'_1$. Hence, from equation (A.30), we have

$$\begin{aligned} M\left(\frac{M(x'_1, x_{-1}^*)}{\bar{\beta} - \beta}, x_{-1}^*\right) &> \bar{\beta}_1 \left(\frac{M(x'_1, x_{-1}^*)}{\bar{\beta} - \beta} - x'_1\right) + x_{-1}^* \beta_{-1} + M(x'_1, x_{-1}^*) \\ &= L\left(\frac{M(x'_1, x_{-1}^*)}{\bar{\beta} - \beta}, x_{-1}^*\right). \end{aligned}$$

From the Intermediate Value Theorem, there exists $\bar{x}_1 \in \left[x'_1, \frac{M(x'_1, x_{-1}^*)}{\bar{\beta} - \beta} \right]$ such that

$$M(\bar{x}_1, x_{-1}^*) = L(\bar{x}_1, x_{-1}^*).$$

Similarly we can show that there exists $\underline{x}_1 < x_{-1}^*$ such that $M(\underline{x}_1, x_{-1}^*) = L(\underline{x}_1, x_{-1}^*)$. ■

A.7. Proof of Theorem 4

Proof. From Lemma 3, there exists $x^* \in \mathcal{X}$ such that $L(x^*) > M(x^*)$. Since M is continuous, there exists $\epsilon > 0$, such that for all $x \in B_\epsilon(x^*)$, $L(x) > M(x)$ where B_ϵ denotes the open ϵ -ball of $x^* \in \mathcal{X}$. Hence, $\int_{x \in \mathcal{D}} dx \geq \int_{x \in B_\epsilon} dx > 0$. Further note that

$$\int_{x_{-1} \in \mathcal{D}_{-1}} dx_{-1} > 0.$$

From Lemma 3 and 4, for each $x_{-1} \in \mathcal{D}_{-1}$, there exist $x'_1(x_{-1}) \in \mathbb{R}$ and $x''_1(x_{-1}) \in \mathbb{R}$ such that $x'_1(x_{-1}) < x''_1(x_{-1})$, $L(x'_1(x_{-1}), x_{-1}) = M(x'_1(x_{-1}), x_{-1})$ and $L(x''_1(x_{-1}), x_{-1}) = M(x''_1(x_{-1}), x_{-1})$. Hence, from the Fundamental Theorem of Calculus, for all $x_{-1} \in \mathcal{D}_{-1}$,

$$\int_{x'_1(x_{-1})}^{x''_1(x_{-1})} \frac{\partial M(x_1(x_{-1}), x_{-1})}{\partial x_1} dx_1 = \beta \cdot (x''_1(x_{-1}) - x'_1(x_{-1})).$$

By Assumption 7(v), $f_X > 0$ almost everywhere on \mathcal{D} . Also recall that $\int_{x_{-1} \in \mathcal{D}_{-1}} dx_{-1} > 0$.

Thus, we can write

$$\begin{aligned} \beta_1 &= \frac{\int_{x_{-1} \in \mathcal{D}_{-1}} \int_{x'_1(x_{-1})}^{x''_1(x_{-1})} \frac{1}{(x''_1(x_{-1}) - x'_1(x_{-1}))} \cdot \frac{\partial M(x_1, x_{-1})}{\partial x_1} dx_1 dx_{-1}}{\int_{x_{-1} \in \mathcal{D}_{-1}} dx_{-1}} \\ &= \frac{\int_{x \in \mathcal{D}} \frac{1}{(x''_1(x_{-1}) - x'_1(x_{-1}))} \cdot \frac{\partial M(x)}{\partial x_1} \cdot \frac{1}{f_X(x)} dF_X(x)}{\int_{x_{-1} \in \mathcal{D}_{-1}} \frac{1}{f_{X_{-1}}} dF_{X_{-1}}(x_{-1})} \\ &= E \left[w_{\mathcal{D}}(X) \cdot \frac{\partial M(X)}{\partial X_1} \middle| X \in \mathcal{D} \right]. \end{aligned}$$

■

A.8. Proof of Lemma 6

Proof. Define the function $P_\tau : \mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R}$ by

$$P_\tau(\Gamma) = \int \mu_\tau(\varpi(\Gamma, x), x) dF_X(x)$$

for each $\Gamma \in \mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{X})$, where $\varpi(\Gamma, x)$ is defined by

$$\varpi(\Gamma, x) = \{u \in \mathcal{U} \mid (u, x) \in \Gamma\}.$$

Note that this set $\varpi(\Gamma, x)$ belongs to $\mathcal{B}(\mathcal{U})$, as Γ belongs to the tensor product sigma algebra $\mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{X})$. We first show that this function P_τ is a probability measure.

That $P_\tau(\Gamma) \in [0, 1]$ for all $\Gamma \in \mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{X})$ follows from the fact that $\mu_{y,x}$ is a probability measure (Lemma 5) and F_X is a probability measure. Likewise, by Lemma 5, we have

$$P_\tau(\emptyset) = \int \mu_{Q_\tau(x),x}(\emptyset) dF_X(x) = 0 \quad \text{and}$$

$$P_\tau(\mathcal{U} \times \mathcal{X}) = \int \mu_{Q_\tau(x),x}(\partial V(Q_\tau(x), x)) dF_X(x) = \int_{\mathcal{X}} dF_X(x) = 1.$$

Let $\{\Gamma_i\}_i \subset \mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{X})$ be a countable collection of disjoint sets. Note that, for each $x \in \mathcal{X}$, $\varpi(\Gamma_i, x) \cap \varpi(\Gamma_j, x) = \emptyset$ whenever $i \neq j$, and $\varpi(\cup_i \Gamma_i, x) = \cup_i \varpi(\Gamma_i, x)$. Therefore, from the fact that $\mu_{y,x}$ is a probability measure (Lemma 5), we have the sigma additivity

$$P_\tau(\cup_i \Gamma_i) = \int \mu_\tau(\varpi(\cup_i \Gamma_i, x), x) dF_X(x) = \int \mu_\tau(\cup_i \varpi(\Gamma_i, x), x) dF_X(x)$$

$$= \int \sum_i \mu_\tau(\varpi(\Gamma_i, x), x) dF_X(x) = \sum_i P_\tau(\Gamma_i),$$

where the last step uses the Fubini–Tonelli Theorem.

To show that μ_τ is a regular conditional probability measure given the events in $\mathcal{B}(\mathcal{X})$, it remains to show that

$$P_\tau(S \times T) = \int_T \mu_\tau(S, x) dF_X(x)$$

holds for all $(S, T) \in \mathcal{B}(\mathcal{U}) \times \mathcal{B}(\mathcal{X})$. This follows straightforwardly, as

$$P_\tau(S \times T) = \int \mu_\tau(\varpi(S \times T, x), x) dF_X(x)$$

$$= \int \mu_\tau(\{u \in \mathcal{U} \mid (u, x) \in S \times T\}, x) dF_X(x) = \int_T \mu_\tau(S, x) dF_X(x). \quad \blacksquare$$