Expectile Regression Process under Misspecification

Zixin Yang
Shanghai University of Finance and Economics

Xiaojun Song*
Peking University

January 17, 2026

Abstract

Allowing for misspecification in the linear conditional expectile function, this paper provides an interpretation and asymptotic distribution theory for the expectile regression (ER) parameter in Newey and Powell (1987). The first result on interpretation shows that ER minimizes a weighted mean-squared error loss function. The weighting function is related to an average of the conditional distribution function of the dependent variable near the true conditional expectile. The weighted least squares interpretation of ER is further used to derive a formula for omitted variables bias. We also establish the asymptotic distribution of the ER process under the misspecification of the conditional expectile function. The result provides a foundation for simultaneous confidence intervals and a basis for global tests of hypotheses about distributions. The approximation properties of ER are illustrated using wage data from the U.S. Census.

Keywords: *Conditional expectile function, misspecification, asymmetric least squares, best linear approximation.*
JEL Classification: C14, C21

*Corresponding author: Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University. Email: sxj@gsm.pku.edu.cn. This work was supported by the National Natural Science Foundation of China [Grant Numbers 72373007 and 72333001]. The author also gratefully acknowledges the research support from the Center for Statistical Science of Peking University, China, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University) of the Ministry of Education, China.

1 Introduction

Ordinary least squares (OLS) has long been the standard method for modeling conditional expectations in econometrics. It provides a parsimonious approach to quantifying the impact of covariates on the mean of a response variable. However, many applications—particularly in labor economics, finance, and risk management—require insight into the entire conditional distribution of a response variable, not just its mean. When covariate effects vary across the distribution or the error terms are heteroskedastic, skewed, or heavy-tailed, focusing solely on the conditional mean is insufficient.

To address such heterogeneity, methods such as quantile regression (Koenker and Bassett, 1978) and expectile regression (ER, Newey and Powell, 1987) have been widely used. Quantile regression estimates conditional quantiles by minimizing an asymmetric absolute loss, allowing covariate effects to vary across the outcome distribution. ER, in contrast, minimizes an asymmetric squared loss to estimate conditional expectiles. An expectile can be viewed as a generalized mean that places more weight on observations above or below a given threshold, depending on the level (Philipps, 2022). It is analogous to a quantile but responds more strongly to extreme values. Like quantiles, expectiles provide a complete characterization of the distribution. Expectile regression, in turn, allows researchers to examine how covariates influence not just the center but also the tails of the conditional outcome distribution.

Compared to quantile regression, ER exhibits complementary strengths and weaknesses. As Newey and Powell (1987) and Efron (1991) observe, quantile regression is robust to outliers and heavy-tailed errors, but it involves a non-differentiable objective function. In contrast, ER uses a smooth squared-error loss and is fully efficient under the assumption of Gaussianity. This smoothness makes ER numerically expedient: scalable algorithms such as gradient methods can compute expectile estimates quickly. However, the squared-loss focus also means that ER can be sensitive to extreme values, unlike the check loss of quantiles. Conceptually, the key trade-off is that quantiles have a direct probabilistic interpretation (they invert the cumulative distribution function), while expectiles are less intuitive but yield the least-squares analogue of those distributional points. Finally, quantile regression requires nonparametric estimation of the conditional density for standard inference, whereas ER does not. The ER coefficients can be obtained using modified least squares routines. These features make ER a flexible and powerful alternative for studying conditional distributions, particularly in settings with skewed or heteroskedastic data.

As a result, ER has been applied in a range of areas including labor economics (Dawber et al., 2022; Bonaccolto-Töpfer and Bonaccolto, 2023), risk assessment (Kuan et al., 2009; Kim and Lee, 2016; Xu et al., 2022), and forecasting evaluation (Guler et al., 2017). In finance, expectiles play a foundational role in the class of elicitable and coherent risk mea-

sures, underpinning the expectile-based value-at-risk (EVaR). Bellini and Bignozzi (2015) show that expectiles are the only risk measures satisfying both coherence and elicitability, further cementing their relevance in financial decision-making.

Despite the advantages mentioned above, nearly all existing studies on ER assume that a linear conditional expectile model is correctly specified. In practice, however, model misspecification is the rule rather than the exception. When a researcher posits a linear expectile model, that model may be only an approximation to the true data-generating process. It is well known that under misspecification, an estimator typically converges not to a "true" parameter but to a pseudo-true value that minimizes a population criterion. For example, OLS converges to the best linear predictor (in MSE) even if the conditional mean is nonlinear, and similarly, a misspecified quantile regression converges to the best linear quantile approximation. Analogously, the pseudo-true ER coefficient is the minimizer of an asymmetric population least-squares loss. Characterizing this pseudo-true parameter and the large-sample distribution of the ER estimator is essential for interpretation and inference when the linear ER model is an approximation. To our knowledge, a general theory of ER under misspecification has not been fully developed in the econometric literature.

This paper fills that gap by developing a robust inferential theory for ER under misspecification. Our contribution is threefold. First, we demonstrate that ER provides the best linear approximation to the conditional expectile function, utilizing a weighted mean squared error loss function. We also illustrate how this approximation property can be used to interpret multivariate ER coefficients as partial regression coefficients and to develop an omitted variables bias formula for ER. A second contribution is to investigate the limiting distribution of the ER process, which accounts for the possible misspecification of a linear conditional expectile function. The results can be used to test the global hypothesis of the conditional distribution and to construct simultaneous confidence intervals. We finally propose a simple test for the correct specification of a linear conditional expectile function over a continuum of expectile levels. The test can be applied to assess the validity of inferences about the effect of covariates on the distribution of outcomes.

The rest of the article is organized as follows. Section 2 presents two interpretations of the ER vector under misspecification. The approximation properties are used to interpret multivariate ER coefficients as partial regression coefficients and to develop an omitted variables bias formula in Section 3. We present in Section 4 the asymptotic theory of the ER process under possible misspecification of the conditional expectile function. We present in Section 5 a simple test for the correct specification of the linear conditional expectile model. Section 6 presents the results of a set of Monte Carlo experiments. Section 7 illustrates ER approximation properties with U.S. census data. The last section concludes. The Appendix provides the proof of the theorems.

2 Interpreting ER under misspecification

Given a continuous response variable Y and a $d \times 1$ regressor vector X, we are interested in the conditional expectile function (CEF) of Y given X. Assuming integrability, the CEF is defined as

$$\mu_\tau(Y|X) = \arg\min_{m \in \mathbb{R}} \mathbb{E}[\rho_\tau(Y - m) - \rho_\tau(Y) \,|\, X],$$

where $\rho_\tau(\lambda) = |\tau - 1(\lambda < 0)| \cdot \lambda^2$ for $\tau \in (0, 1)$ and $1(A)$ denotes the indicator function for the event A. Clearly, the CEF can be viewed as an asymmetric generalization of the conditional mean function, which is $\mu_{1/2}(Y|X)$. It offers insights into the entire conditional distribution of Y given X in much the same way that the conditional quantile function does, while enjoying both computational convenience and statistical efficiency under normality conditions (Efron, 1991).

It may be possible to capture important features of the CEF using a linear model. This motivates linear ER. The linear ER vector solves the population minimization problem

$$\beta(\tau) \equiv \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}[\rho_\tau(Y - X^\intercal \beta) - \rho_\tau(Y)], \tag{1}$$

see Newey and Powell (1987) for the conditions that ensure the existence and uniqueness of the solution. The first-order conditions for this minimization problem imply that for each τ, $\beta(\tau)$ is a solution of the equation

$$\beta(\tau) = \left\{\mathbb{E}\big[|\tau - 1(Y < X^\intercal \beta(\tau))| X X^\intercal\big]\right\}^{-1} \mathbb{E}\big[|\tau - 1(Y < X^\intercal \beta(\tau))| X Y\big].$$

If $\mu_\tau(Y|X)$ is, in fact, linear, the ER minimand will find it. Otherwise, ER provides the best linear predictor of Y given X under the asymmetric least squares loss function, ρ_τ. When $\tau = 0.5$, ER (OLS regression) provides a minimum mean-squared error fit to the conditional expectation function. When $\tau \neq 0.5$, ER may have an approximation property similar to that of OLS, but the exact nature of the linear approximation has remained unknown.

Our first result fills this gap, demonstrating that ER is the best linear approximation to the conditional expectile function using a weighted mean squared error loss. To introduce the result, for any expectile index $\tau \in (0, 1)$, define the ER specification error as

$$D_\tau(X, \beta) \equiv X^\intercal \beta - \mu_\tau(Y|X).$$

Similarly, let ϵ_τ be an expectile-specific residual, defined as the deviation of the response variable from the conditional expectile of interest,

$$\epsilon_\tau \equiv Y - \mu_\tau(Y|X).$$

Finally, let $F_{\epsilon_\tau}(e|X)$ and $f_{\epsilon_\tau}(e|X)$ be the conditional distribution function and density of ϵ_τ given X, evaluated at $\epsilon_\tau = e$, respectively.

Theorem 1. *(Approximation Property) Suppose that (i) the conditional density $f_Y(y|X)$ exists a.s., (ii) $\mathbb{E}|Y|^2$, $\mathbb{E}[\mu_\tau(Y|X)]^2$, and $\mathbb{E}\|X\|^2$ are finite, and (iii) $\mathbb{E}[XX^\mathsf{T}]$ is nonsingular. Then, for each $\tau \in (0,1)$, a unique solution $\beta(\tau)$ to (1) exists, and*

$$\beta(\tau) = \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}[w_\tau(X,\beta) \cdot D_\tau^2(X,\beta)], \tag{2}$$

where

$$\begin{aligned} w_\tau(X,\beta) &= \tau + 2(1-2\tau)\int_0^1 (1-u) \cdot F_{\epsilon_\tau}(uD_\tau(X,\beta)|X)du \\ &= \tau + 2(1-2\tau)\int_0^1 (1-u) \cdot F_Y(u \cdot X^\mathsf{T}\beta + (1-u) \cdot \mu_\tau(Y|X)|X)du \geq 0. \end{aligned}$$

Theorem 1 states that the population ER vector minimizes the expected weighted mean-squared approximation error, i.e., the square of the difference between the true CEF and a linear approximation, with the weighting function $w_\tau(X,\beta)$. The weights depend on the index τ as well as the average distribution function of the response variable over a line from the point of approximation, $X^\mathsf{T}\beta$, to the true conditional expectile, $\mu_\tau(Y|X)$. When $\tau = 0.5$, the weights are equal to 0.5 for all values in the support of X. The result (2) then reduces to the well-known fact that OLS estimates provide the minimum mean-squared error linear approximation to a conditional expectation function $\mathbb{E}(Y|X)$.

As in Angrist et al. (2006), we refer to the function $w_\tau(X,\beta)$ as *importance weights* since this function determines the importance the ER minimand gives to points in the support of X for a given distribution of X. To understand what determines the shape of the importance weights, we consider the following approximation. Suppose that Y has a conditional density $f_Y(y|X)$, then for β in the neighborhood of $\beta(\tau)$,

$$\begin{aligned} w_\tau(X,\beta) &= \tau + (1-2\tau)F_Y(\mu_\tau(Y|X)|X) + r_\tau(X), \\ \big|r_\tau(X)\big| &\leq |1-2\tau|/3 \cdot |D_\tau(X,\beta)| \cdot \bar{f}(X). \end{aligned}$$

Here, $r_\tau(X)$ is a remainder term and the density $f_Y(y|X)$ is assumed to be bounded in y in absolute value by $\bar{f}(X)$ a.s.[1] It is interesting to note that the primary determinants of the importance weights, i.e., $\tau + (1-2\tau)F_Y(\mu_\tau(Y|X)|X)$, is constant across X in the location-scale models.[2]

[1] The remainder term $r_\tau(X) = w_\tau(X,\beta) - \tau - (1-2\tau)F_{\epsilon_\tau}(0\,|\,X)$ is bounded as $|r_\tau(X)| = 2|(1-2\tau)\int_0^1 (1-u)(F_{\epsilon_\tau}(uD_\tau(X,\beta)\,|\,X) - F_{\epsilon_\tau}(0\,|\,X))du| \leq 2|1-2\tau||D_\tau(X,\beta)|\bar{f}(X)\int_0^1 u(1-u)du = |1-2\tau|/3 \cdot |D_\tau(X,\beta)| \cdot \bar{f}(X)$.

[2] A location-scale model is any model of the form $Y = \mu(X) + \sigma(X) \cdot e$, where e is independent of X. In

In addition to the importance weights, the probability distribution of X also plays a role in the ultimate weight given to different values of X in the least squares problem (2). To see this, note that the ER minimand can also be expressed as $\beta(\tau) = \arg\min_{\beta\in\mathbb{R}^d} \int D_\tau^2(x,\beta) w_\tau(x,\beta)\pi(x)dx$, where $\pi(x)$ denotes the density function of X. Thus, the overall weight varies in the distribution of X according to $w_\tau(x,\beta)\cdot\pi(x)$.

The following theorem gives a second approximation property for expectile regression. This property is particularly useful for the development of a partial regression decomposition and the derivation of an omitted variables bias formula for ER.

Theorem 2. *(Iterative Approximation Property) Suppose that the conditions in Theorem 1 hold. For each $\tau\in(0,1)$, $\bar\beta(\tau)=\beta(\tau)$ uniquely solves the equation*

$$\bar\beta(\tau) = \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}\big[\bar w_\tau(X,\bar\beta(\tau))\cdot D_\tau^2(X,\beta)\big],$$

where

$$\begin{aligned}\bar w_\tau(X,\bar\beta(\tau)) &= \tau + (1-2\tau)\int_0^1 F_{\epsilon_\tau}\big(u\cdot D_\tau(X,\bar\beta(\tau))|X\big)du \\ &= \tau + (1-2\tau)\int_0^1 F_Y\big(u\cdot X^\mathsf{T}\bar\beta(\tau) + (1-u)\cdot\mu_\tau(Y|X)|X\big)du.\end{aligned}$$

The above theorem states that ER solves a weighted least squares approximation problem $\beta(\tau)=\arg\min_{\beta\in\mathbb{R}^d}\mathbb{E}[\bar w_\tau(X)\cdot D_\tau^2(X,\beta)]$, where the weight $\bar w_\tau(X)=\bar w_\tau(X,\bar\beta(\tau))$ is a function of X only. It also shows that the ER coefficient is the unique fixed point of an iterated minimum distance approximation. As in Theorem 1, when $\tau\neq 0.5$, the weighting function $\bar w_\tau(X)$ is related to the conditional distribution function of the dependent variable. In particular, when Y has a conditional density, we have, by a Taylor approximation,

$$\begin{aligned}\bar w_\tau(X,\beta(\tau)) &= \tau + (1-2\tau)F_Y\big(\mu_\tau(Y|X)|X\big) + \bar r_\tau(X), \\ \big|\bar r_\tau(X)\big| &\le |1-2\tau|/2\cdot|D_\tau(X,\beta(\tau))|\cdot\bar f(X),\end{aligned}$$

where $\bar r_\tau(X)$ is a remainder term, and the density $f_Y(y|X)$ is assumed to be bounded in y in absolute value by $\bar f(X)$ a.s.. When either $D_\tau(X,\beta(\tau))$ or $\bar f(X)$ is small, we have

$$\bar w_\tau(X,\beta(\tau)) \approx w_\tau(X,\beta(\tau)) \approx \tau + (1-2\tau)F_Y(\mu_\tau(Y|X)|X),$$

i.e., the approximate weighting function is the same as derived in Theorem 1.

this model, we have $\mu_\tau(Y|X)=\mu(X)+\sigma(X)\xi_\tau$, where ξ_τ is the τth expectile of e. Suppose that $\sigma(X)>0$. It is easily seen that $F_Y(\mu_\tau(Y|X)|X)=F_e(\xi_\tau|X)$, which is independent of X.

3 Partial ER and omitted variable bias

In this section, we derive an omitted variables bias formula and introduce the concept of partial expectile regression, utilizing the approximation properties established above. The partial expectile regression is defined with regard to a partition of the regressor vector X into variables X_1 and the remaining variables X_2, along with the corresponding partition of ER coefficients $\beta(\tau)$ into $\beta_1(\tau)$ and $\beta_2(\tau)$. We decompose $\mu_\tau(Y|X)$ and X_1 using orthogonal projections onto X_2 weighted by $\bar{w}_\tau(X) = \bar{w}_\tau(X, \beta(\tau))$ defined in Theorem 2:

$$\mu_\tau(Y|X) = \pi_\mu^\mathsf{T} X_2 + q_\tau(Y|X), \text{ where } \mathbb{E}\big[\bar{w}_\tau(X) \cdot X_2 \cdot q_\tau(Y|X)\big] = 0,$$

and

$$X_1 = \pi_1^\mathsf{T} X_2 + V_1, \text{ where } \mathbb{E}\big[\bar{w}_\tau(X) \cdot X_2 \cdot V_1^\mathsf{T}\big] = 0.$$

In the decomposition, $q_\tau(Y|X)$ and V_1 are residuals created by a weighted linear projection of $\mu_\tau(Y|X)$ and X_1 on X_2, respectively, using $\bar{w}_\tau(X)$ as the weight.[3] Using Theorem 2 and standard least squares algebra, we have

$$\beta_1(\tau) = \arg\min_{\beta_1} \mathbb{E}\Big[\bar{w}_\tau(X)\big(q_\tau(Y|X) - V_1^\mathsf{T}\beta_1\big)^2\Big],$$

and also $\beta_1(\tau) = \arg\min_{\beta_1} \mathbb{E}\big[\bar{w}_\tau(X)\big(\mu_\tau(Y|X) - V_1^\mathsf{T}\beta_1\big)^2\big]$. This implies that $\beta_1(\tau)$ is a partial ER coefficient in the sense that it can be obtained from a weighted least squares regression of $\mu_\tau(Y|X)$ on X_1, once we have partialled out the effect of X_2. Both the first-step and second-step regressions are weighted by $\bar{w}_\tau(X)$.

An omitted variable bias formula for ER can be similarly derived. Suppose that we are interested in an expectile regression with explanatory variables $X = (X_1^\mathsf{T}, X_2^\mathsf{T})^\mathsf{T}$, but X_2 is not available. We run ER on X_1 only, where the population coefficient vectors are given by $\gamma_1(\tau) = \arg\min_{\gamma_1} \mathbb{E}[\rho_\tau(Y - X_1^\mathsf{T}\gamma_1)]$. The long regression coefficient vectors are $(\beta_1^\mathsf{T}(\tau), \beta_2^\mathsf{T}(\tau))^\mathsf{T} = \arg\min_{\beta_1,\beta_2} \mathbb{E}[\rho_\tau(Y - X_1^\mathsf{T}\beta_1 - X_2^\mathsf{T}\beta_2)]$. Using Theorem 2, we can find that

$$\gamma_1(\tau) = \beta_1(\tau) + \big(\mathbb{E}[\tilde{w}_\tau(X) \cdot X_1 X_1^\mathsf{T}]\big)^{-1} \mathbb{E}[\tilde{w}_\tau(X) \cdot X_1 R_\tau(X)], \tag{3}$$

where $R_\tau(X) \equiv \mu_\tau(Y|X) - X_1^\mathsf{T}\beta_1(\tau)$, $\tilde{w}_\tau(X) \equiv \tau + (1-2\tau)\int_0^1 F_{\epsilon_\tau}(u \cdot D_\tau(X, \gamma_1(\tau)) \mid X)du$, $D_\tau(X, \gamma_1) \equiv X_1^\mathsf{T}\gamma_1 - \mu_\tau(Y|X)$, and $\epsilon_\tau \equiv Y - \mu_\tau(Y|X_1)$. Note that $R_\tau(X)$ is the part of the CEF not explained by the linear function of X_1 in the long ER. If the CEF is linear, then $R_\tau(X) = X_2^\top \beta_2(\tau)$.

For the OLS regression ($\tau = 0.5$), we have $\tilde{w}_\tau(X) = 0.5$ and Equation (3) reduces to $\gamma_1(\tau) = \beta_1(\tau) + (\mathbb{E}[X_1 X_1^\mathsf{T}])^{-1}\mathbb{E}[X_1(Y - X_1^\mathsf{T}\beta_1(\tau))]$. When $\tau \neq 0.5$, the short ER coefficients

[3] Thus, $\pi_\mu = \mathbb{E}[\bar{w}_\tau(X)X_2X_2^\mathsf{T}]^{-1}\mathbb{E}[\bar{w}_\tau(X)X_2\mu_\tau(Y|X)]$ and $\pi_1 = \mathbb{E}[\bar{w}_\tau(X)X_2X_2^\mathsf{T}]^{-1}\mathbb{E}[\bar{w}_\tau(X)X_2X_1^\mathsf{T}]$.

are equal to the corresponding long ER coefficients plus the coefficients in weighted projections of omitted effects on the included variables. The effect of omitted variables appears through the remainder term $R_\tau(X)$. Equation (3) also parallels the omitted variables formula for quantile regression. The main difference is that in the quantile case, the weighting function is related to an integral of the conditional density of Y given X, as seen in Section 2.3 of Angrist et al. (2006).

4 ER process under misspecification

In analogy to the literature on robust inference for quantile regression, it is of interest to examine how specification errors affect inference in ER. This section develops the large-sample distribution theory for ER without assuming a correctly specified CEF. Although ER does not consistently estimate the true nonlinear CEF under misspecification, it consistently estimates a best linear approximation as characterized in Theorems 1 and 2. To quantify the sampling variability in these estimates, we derive the asymptotic distribution of the sample ER process $\widehat{\beta}_n(\cdot)$, which is defined as

$$\widehat{\beta}_n(\tau) = \arg\min_{\beta\in\mathbb{R}^d} n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - X_i^\mathsf{T}\beta), \tag{4}$$

where $\tau \in \mathcal{T}$, a closed subset of $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$. We study the entire ER process $\widehat{\beta}_n(\cdot)$ because it facilitates global hypothesis testing and the construction of simultaneous confidence regions for the conditional distribution. While Newey and Powell (1987) examined pointwise inference for expectile coefficients under potential misspecification, the behavior of the entire ER process has not, to our knowledge, been addressed in the existing literature. The following theorem fills this gap by establishing uniform consistency and joint asymptotic normality for $\widehat{\beta}_n(\cdot)$.

Assumption 1. *(a)* $\{W_i = (Y_i, X_i^\mathsf{T})^\mathsf{T}\}_{i=1}^n$ *are independent and identically distributed (i.i.d.) on the probability space* $(\Omega, \mathcal{F}, P)$ *for each* n*. (b) The conditional density* $f_Y(y|X=x)$ *exists, and is uniformly continuous in* y *for almost all* x *over the support of* X*. (c)* $\mathbb{E}\|W\|^{4+\eta} < \infty$ *for some* $\eta > 0$*. (d)* $\mathbb{E}[XX^\mathsf{T}]$ *is nonsingular.*

Theorem 3. *Suppose Assumption 1 holds. Then, the ER process is uniformly consistent, i.e.,* $\sup_{\tau\in\mathcal{T}}\|\widehat{\beta}_n(\tau) - \beta(\tau)\| = o_p(1)$*, and* $J(\cdot)\sqrt{n}(\widehat{\beta}_n(\cdot) - \beta(\cdot))$ *converges in distribution to a zero mean Gaussian process* $z(\cdot)$*, where* $J(\tau) \equiv \mathbb{E}[|\tau - 1(Y < X^\mathsf{T}\beta(\tau))|XX^\mathsf{T}]$*, and* $z(\cdot)$ *is defined by its covariance function* $\Sigma(\tau,\tau') \equiv \mathbb{E}[z(\tau)z(\tau')^\mathsf{T}]$ *with*

$$\Sigma(\tau,\tau') = \mathbb{E}\big[\big|\big(\tau - 1(Y < X^\mathsf{T}\beta(\tau))\big)\big(\tau' - 1(Y < X^\mathsf{T}\beta(\tau'))\big)\big|\big(Y - X^\mathsf{T}\beta(\tau)\big)\big(Y - X^\mathsf{T}\beta(\tau')\big)XX^\mathsf{T}\big].$$

This result generalizes the joint asymptotic normality of finitely many expectile coefficients established by Newey and Powell (1987, Theorem 3). A key implication is that for any finite collection $\{\tau_k\}_{k=1}^K \subset \mathcal{T}$, the vector $\sqrt{n}(\widehat{\beta}_n(\tau_k) - \beta(\tau_k))$, $k = 1, 2, \ldots$ is jointly asymptotically normal with covariance $J(\tau_k)^{-1}\Sigma(\tau_k, \tau_l)J(\tau_l)^{-1}$. It is worth emphasizing that even under correct specification, the covariance structure $\Sigma(\tau, \tau')$ does not simplify without stronger model assumptions. This contrasts with quantile regression, where the asymptotic covariance under misspecification is typically more complicated than under correct specification (Angrist et al., 2006). In this regard, a notable advantage of ER is that its asymptotic covariance retains a unified form across both well-specified and misspecified settings, simplifying inference and implementation.

Remark 1. *While Theorem 3 provides a general characterization of the asymptotic distribution of the ER process, the resulting covariance expressions can be difficult to evaluate or interpret. To provide further intuition, we now consider a location-scale model as a concrete example in which simpler expressions for the asymptotic variance can be derived. Specifically, consider the model*

$$Y_i = X_i^\mathsf{T}\beta + (X_i^\mathsf{T}\gamma)\epsilon_i, \tag{5}$$

where $\{X_i\}$ are i.i.d. vectors with first component equal to 1, β and γ are comfortable vectors of unknown parameters, and $\{\epsilon_i\}$ are i.i.d. errors independent of $\{X_i\}$. For this model, the conditional τth expectile satisfies $\mu_\tau(Y|X) = X^\mathsf{T}(\beta + \xi_\tau\gamma)$, where ξ_τ is the τth expectile of ϵ, so $\beta(\tau) = \beta + \xi_\tau\gamma$. Assuming $X^\mathsf{T}\gamma > 0$ almost surely, we have

$$\mathbb{E}[|\tau - 1(Y < X^\mathsf{T}\beta(\tau))| \,|\, X] = \tau + (1 - 2\tau)F_Y(X^\mathsf{T}\beta(\tau)|X) = \tau + (1 - 2\tau)F_\epsilon(\xi_\tau),$$

and

$$\mathbb{E}[|\tau - 1(Y < X^\mathsf{T}\beta(\tau))|^2(Y - X^\mathsf{T}\beta(\tau))^2 \,|\, X] = (X^\mathsf{T}\gamma)^2 \cdot \mathbb{E}[|\tau - 1(\epsilon < \xi_\tau)|^2(\epsilon - \xi_\tau)^2].$$

Without loss of generality, we normalize $Var(\epsilon) = 1$ to identify the scale of γ. It follows that the asymptotic variance of $\widehat{\beta}_n(\tau)$ is given by

$$AVar\big(\widehat{\beta}_n(\tau)\big) = c_\tau \Sigma_X^{-1}\Omega_X\Sigma_X^{-1},$$

where $c_\tau = \mathbb{E}[|\tau - 1(\epsilon < \xi_\tau)|^2(\epsilon - \xi_\tau)^2]/[\tau + (1 - 2\tau)F_\epsilon(\xi_\tau)]^2$, $\Sigma_X = \mathbb{E}[XX^\mathsf{T}]$, and $\Omega_X = \mathbb{E}[(X^\mathsf{T}\gamma)^2 XX^\mathsf{T}]$. Notably, when $\tau = 1/2$, ER reduces to OLS, $\xi_{1/2} = 0$, and $c_{1/2} = \mathbb{E}(\epsilon^2) = 1$. In this case, the expression above simplifies to the familiar sandwich variance for the OLS estimator $AVar\big(\widehat{\beta}_n(1/2)\big) = \Sigma_X^{-1}\Omega_X\Sigma_X^{-1}$.

Furthermore, the variance expression simplifies under the assumption of homoskedasticity. Partition the vector γ as $\gamma = (\gamma_1, \gamma_2^\mathsf{T})^\mathsf{T}$, where γ_1 is a scalar corresponding to

the intercept and $\gamma_2 \in \mathbb{R}^{d-1}$ *corresponds to the slope components. If the errors are homoskedastic, i.e.,* $\gamma_2 = 0$*, the asymptotic variance simplifies to* $AVar(\widehat{\beta}_n(\tau)) = c'_\tau \Sigma_X^{-1}$*, where* $c'_\tau = c_\tau \gamma_1^2$*. This result illustrates that under homoskedasticity, the variance of the ER estimator is proportional to the inverse of the regressor covariance matrix, as in classical OLS. Lastly, if instead* $X^\mathsf{T}\gamma < 0$ *almost surely, the same conclusion holds with* c_τ *replaced by* $\tilde{c}_\tau = \mathbb{E}[|\tau - 1(\epsilon > \xi_\tau)|^2(\epsilon - \xi_\tau)^2]/[\tau + (1-2\tau)(1 - F_\epsilon(\xi_\tau))]^2$.

Inference on the ER process is useful for testing basic hypotheses of the form

$$R(\tau)\beta(\tau) = r(\tau), \quad \text{for all } \tau \in \mathcal{T}, \tag{6}$$

where $R(\tau)$ denotes an $l \times d$ matrix with $l \leq d$, and $r(\tau)$ is an $l \times 1$ vector. For example, we may be interested in testing whether a subset of variables $j \in \{k+1, \ldots, d\}$ enters the model for all conditional expectiles with zero coefficients, i.e., whether $\beta_j(\tau) = 0$ for all $\tau \in \mathcal{T}$ and $j = k+1, \ldots, d$. This corresponds to $R(\tau) = [\mathbf{0}_{(d-k)\times k} \quad I_{d-k}]$ and $r(\tau) = \mathbf{0}_{d-k}$. Another two particular hypotheses of interest are testing the constancy and symmetry of $\beta(\cdot)$. In these examples, the component $r(\tau)$ is a function of the conditional distribution and thus needs to be estimated, see Remark 2 below for detailed discussion.

Let $\widehat{R}_n(\tau)$ and $\widehat{r}_n(\tau)$ denote the estimators of $R(\tau)$ and $r(\tau)$, respectively. To test (6), we consider the empirical process

$$v_n(\tau) \equiv \widehat{R}_n(\tau)\widehat{\beta}_n(\tau) - \widehat{r}_n(\tau),$$

and form the test statistics as a continuous functional of v_n, say, $\Gamma(v_n)$. Commonly used functionals include the Kolmogorov–Smirnov (KS)-type functional

$$K_n \equiv \sqrt{n} \sup_{\tau \in \mathcal{T}} \left\| v_n(\tau) \right\|_{\widehat{V}(\tau)}, \tag{7}$$

and the Cramér–von Mises (CvM)-type functional

$$C_n \equiv n \int_{\mathcal{T}} \left\| v_n(\tau) \right\|^2_{\widehat{V}(\tau)} d\Phi(\tau), \tag{8}$$

where $\|a\|_V$ denotes $\sqrt{a^\mathsf{T} V a}$, $\widehat{V}(\tau)$ is a symmetric weight matrix such that $\widehat{V}(\tau) = V(\tau) + o_p(1)$ for some positive definite and continuous symmetric matrix $V(\tau)$ uniformly in τ, and Φ is some integrating measure on $\mathcal{T}$. The null hypothesis could be rejected if the realized values of K_n and C_n appear in the right tail of their asymptotic null distributions.

Remark 2. *Tests of homoskedasticity and conditional symmetry can be based on the hypothesis given in* (6)*. Assume that the data are generated by the linear model* (5)*. As shown in Re-*

mark 1, for this model, the population ER vector is given by $\beta(\tau) = \beta + \xi_\tau \gamma$, *where* ξ_τ *is the* τ*th expectile of* ϵ_i. *Partition* $\beta(\tau)$ *as* $\beta(\tau) = (\beta_0(\tau), \beta_1(\tau)^\mathsf{T})^\mathsf{T}$, *where* $\beta_0(\tau)$ *is a scalar and* $\beta_1(\tau)$ *is a* $(d-1)\times 1$ *slope vector. Then, if the scale of* Y_i *varies with* X_i, *so that heteroskedasticity is present in the regression equation, the slope coefficients* $\beta_1(\tau)$ *also vary with* τ. *Hence, heteroskedasticity can be detected by checking whether* $\beta_1(\tau) = \beta_1$ *for some* $\beta_1 \in \mathbb{R}^{d-1}$ *and for all* τ. *This corresponds to hypothesis (6) with* $R(\tau) = [\mathbf{0}_{(d-1)\times 1}, I_{d-1}]$ *and* $r(\tau) = \beta_1$. *We can estimate* $r(\tau)$ *in this case by* $\widehat{r}_n(\tau) = R(\tau)\widehat{\beta}_n$, *where* $\widehat{\beta}_n = (\sum_{i=1}^n X_i X_i^\mathsf{T})^{-1} \sum_{i=1}^n X_i Y_i$. *The test for homoskedasticity is then based on the empirical process* $v_n(\tau) = R(\tau)\widehat{\beta}_n(\tau) - \widehat{r}_n(\tau)$.

Nonsymmetry can be detected by checking whether the symmetrically placed ER estimators average up to the least squares estimator. As shown in Newey and Powell (1987, Theorem 2), if the distribution of Y_i *conditional on* X_i *is symmetric around* $X_i^\mathsf{T}\beta$ *with probability one, then* $[\beta(\tau) + \beta(1-\tau)]/2 = \beta(1/2)$. *This implies that tests for symmetry can be based on the empirical process* $v_n(\tau) = [\widehat{\beta}_n(\tau) + \widehat{\beta}_n(1-\tau)]/2 - \widehat{\beta}_n$.

In the following, we study the weak convergence of $v_n(\cdot)$ as an element of $l^\infty(\mathcal{T})$, the space of real-valued functions that are uniformly bounded on $\mathcal{T}$. The space is furnished with the supremum metric, and let $\mathcal{B}_{d_\infty}$ denote the corresponding Borel σ-algebra. We denote by "$\Rightarrow$" the weak convergence on $(l^\infty(\mathcal{T}), \mathcal{B}_{d_\infty})$ in the sense of Hoffmann–Jørgensen; see, e.g., Van Der Vaart and Wellner (1996, Definition 1.3.3) and Dudley (2014, p.136).

The following assumptions specify the data-generating process and the requirement for nuisance parameter estimates to derive the asymptotic distribution of the test statistics.

Assumption 2. *For each* $n = 1, 2, 3, \ldots$, $\{W_{i,n} \equiv (Y_{i,n}, X_{i,n}^\mathsf{T})^\mathsf{T}\}_{i=1}^n$ *is a sequence of arrays of i.i.d. variables satisfying*

$$R(\tau)\beta_n(\tau) - r(\tau) = p(\tau)/\sqrt{n},$$

where $p(\cdot) : \mathcal{T} \to \mathbb{R}^l$ *is a fixed and continuous function.*[4]

Assumption 3. *Under the local alternatives in Assumption 2, the expectile estimators and nuisance parameter estimators satisfy that* $\sqrt{n}J(\cdot)(\widehat{\beta}_n(\cdot) - \beta_n(\cdot)) \Rightarrow z(\cdot)$, $\sqrt{n}(\widehat{R}_n(\cdot) - R(\cdot)) \Rightarrow \rho(\cdot)$, *and* $\sqrt{n}(\widehat{r}_n(\cdot) - r(\cdot)) \Rightarrow \zeta(\cdot)$ *jointly in* $l^\infty(\mathcal{T})$, *where* $J(\cdot)$ *and* $z(\cdot)$ *are defined in Theorem 3, and* $(z, vec(\rho)^\mathsf{T}, \zeta^\mathsf{T})^\mathsf{T}$ *is a zero mean continuous Gaussian process with a non-degenerate covariance kernel.*[5]

Corollary 1 describes the limiting distribution of the process $v_n(\cdot)$ and the test statistics.

Corollary 1. *Suppose that Assumptions 1-3 hold. Then, in* $l^\infty(\mathcal{T})$

$$\sqrt{n}v_n(\cdot) \Rightarrow v(\cdot) = v_0(\cdot) + p(\cdot), \quad v_0(\cdot) \equiv u(\cdot) + d(\cdot),$$

[4]The subscript n in $\beta_n(\tau)$ makes explicit the dependence of $\beta(\tau)$ on n under the local alternatives.

[5]For an $n \times m$ matrix $A = (a_1, \ldots, a_m)$, vec(A) denotes the vectorization of A, i.e., vec$(A) = (a_1^\mathsf{T}, \ldots, a_m^\mathsf{T})^\mathsf{T}$.

where $u(\tau) \equiv R(\tau)J(\tau)^{-1}z(\tau)$, $d(\tau) \equiv \rho(\tau)\beta(\tau) - \zeta(\tau)$, *and* $\beta(\tau) \equiv \lim_{n\to\infty} \beta_n(\tau)$. *Under the null hypothesis* (6), $p = 0$, *the test statistics* $K_n \xrightarrow{d} \mathcal{K} \equiv \sup_{\tau\in\mathcal{T}} \|v_0(\tau)\|_{V(\tau)}$, *and* $C_n \xrightarrow{d} \mathcal{C} \equiv \int_{\mathcal{T}} \|v_0(\tau)\|^2_{V(\tau)} d\Phi(\tau)$.

Corollary 1 shows that the limit process $v(\cdot)$ is the sum of three components, $u(\cdot)$, $d(\cdot)$, and $p(\cdot)$. The usual component $u(\cdot)$ is a Gaussian process with a non-standard covariance kernel. The component $d(\cdot)$ is a Durbin term that is present due to the estimation effects of $R(\cdot)$ and $r(\cdot)$.[6] The component $p(\cdot)$ describes deviations from the null and determines the power of the tests.

Since the limiting distributions of K_n and C_n are nonstandard, the critical values for confidence regions and tests cannot be tabulated in general. To overcome this problem, we propose a multiplier bootstrap procedure, which is easy to implement and does not require the computation of new parameter estimates at each bootstrap replication. We first state a new assumption, which facilitates the simulation of the asymptotic distributions of $\widehat{R}_n(\cdot)$ and $\widehat{r}_n(\cdot)$.

Assumption 4. *Under the null hypothesis* (6), *uniformly in* $\tau \in \mathcal{T}$,

$$\sqrt{n}(\widehat{R}_n(\tau) - R(\tau)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_\tau(W_i, R(\tau)) + o_p(1),$$

where $L_\tau(W_i, R(\tau))$ *is an* $l \times d$ *matrix taking values in* $\Xi_R \subseteq \mathbb{R}^{l\times d}$ *such that* $\mathbb{E}[L_\tau(W_i, R(\tau))] = 0$ *and* $\mathbb{E}[vec(L_\tau(W_i, R(\tau)))vec^{\mathsf{T}}(L_\tau(W_i, R(\tau)))]$ *exists and is positive definite for all* τ; *and*

$$\sqrt{n}(\widehat{r}_n(\tau) - r(\tau)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\tau(W_i, r(\tau)) + o_p(1),$$

where $l_\tau(W_i, r(\tau))$ *is an* $l \times 1$ *vector taking values in* $\Xi_r \subseteq \mathbb{R}^l$ *such that* $\mathbb{E}[l_\tau(W_i, r(\tau))] = 0$ *and* $\mathbb{E}[l_\tau(W_i, r(\tau))l_\tau^{\mathsf{T}}(W_i, r(\tau))]$ *exists and is positive definite for all* τ. *The function classes* $\{W \to L_\tau(W, R) : \tau \in \mathcal{T}, R \in \Xi_R\}$ *and* $\{W \to l_\tau(W, r) : \tau \in \mathcal{T}, r \in \Xi_r\}$ *are P-Donsker.*[7]

Let $\widehat{J}_n(\tau)$ be a consistent estimate of $J(\tau)$ and

$$\psi(W_i, \beta(\tau), \tau) \equiv \big|\tau - 1\big(Y_i < X_i^{\mathsf{T}}\beta(\tau)\big)\big|\big(Y_i - X_i^{\mathsf{T}}\beta(\tau)\big). \tag{9}$$

We propose to approximate the asymptotic distribution of the test statistic $\Gamma(v_n)$ by that of

[6] Analogous problems arise in the theory of the one-sample Kolmogorov–Smirnov test when there are estimated parameters under the null, see, e.g., Durbin (1973).

[7] The definition of the P-Donsker class can be found in Van Der Vaart (1998, p. 269).

$\Gamma(v_n^*)$, where

$$v_n^*(\tau) \equiv \frac{1}{n}\sum_{i=1}^{n}\Big\{L_\tau(W_i, \widehat{R}_n(\tau))\widehat{\beta}_n(\tau) + \widehat{R}_n(\tau)\widehat{J}_n(\tau)^{-1}\psi(W_i, \widehat{\beta}_n(\tau), \tau)X_i - l_\tau\big(W_i, \widehat{r}_n(\tau)\big)\Big\}V_i,$$

and $\{V_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with zero mean, unit variance, bounded support, and independent with $\{W_i\}_{i=1}^n$. A popular example is i.i.d. Bernoulli variables $\{V_i\}$ with $\mathrm{P}(V = 1-\kappa) = \kappa/\sqrt{5}$ and $\mathrm{P}(V = \kappa) = 1 - \kappa/\sqrt{5}$, where $\kappa = (\sqrt{5}+1)/2$ (Mammen, 1993). The bootstrap empirical distribution of $\Gamma(v_n^*)$, i.e., $\widehat{F}_n^*(t|\{W_i\}_{i=1}^n) = \mathrm{P}[\Gamma(v_n^*) \leq t|\{W_i\}_{i=1}^n]$ is shown to be a consistent estimate of the asymptotic null distribution function of $\Gamma(v_n)$. Hence, the null hypothesis (6) will be rejected at the α-level of significance whenever $\Gamma(v_n) \geq c_{n,\alpha}^*$, where $c_{n,\alpha}^*$ is such that $\widehat{F}_n^*(c_{n,\alpha}^*|\{W_i\}_{i=1}^n) = 1-\alpha$. Alternatively, the null could be rejected when $p_n^* < \alpha$, where $p_n^* = \mathrm{P}[\Gamma(v_n) < \Gamma(v_n^*)|\{W_i\}_{i=1}^n]$.

Theorem 4 establishes the asymptotic validity of the proposed multiplier bootstrap procedure.

Theorem 4. *Suppose that Assumptions 1-4 hold. Then,* $\sqrt{n}v_n^*(\cdot) \underset{*}{\Rightarrow} v_0(\cdot)$ *in probability, where* $v_0(\cdot)$ *is the Gaussian process defined in Corollary 1, and* $\underset{*}{\Rightarrow}$ *denotes the weak convergence under the bootstrap law, i.e., conditional on the original sample* $\{W_i\}_{i=1}^n$*. Additionally,* $K_n^* \equiv \sqrt{n}\sup_{\tau\in\mathcal{T}}\|v_n^*(\tau)\|_{\widehat{V}(\tau)} \underset{*}{\xrightarrow{d}} \mathcal{K}$*, and* $C_n^* \equiv n\int_{\mathcal{T}}\|v_n^*(\tau)\|_{\widehat{V}(\tau)}^2 d\Phi(\tau) \underset{*}{\xrightarrow{d}} \mathcal{C}$ *in probability.*

The inference procedure above requires a consistent estimator of $\Sigma(\tau,\tau')$ and $J(\tau)$, which can be obtained by the following sample analog:

$$\begin{aligned}\widehat{\Sigma}_n(\tau,\tau') &= \frac{1}{n}\sum_{i=1}^{n}\psi\big(W_i, \widehat{\beta}_n(\tau), \tau\big)\psi(W_i, \widehat{\beta}_n(\tau'), \tau')X_iX_i^\intercal, \\ \widehat{J}_n(\tau) &= \frac{1}{n}\sum_{i=1}^{n}|\tau - 1(Y_i < X_i^\intercal\widehat{\beta}_n(\tau))|X_iX_i^\intercal.\end{aligned} \tag{10}$$

We show in the Appendix that these estimators are consistent uniformly in τ under the conditions of Theorem 3. Compared to tests based on the quantile regression process, the covariance matrix of the ER estimator does not involve the conditional density function of the dependent variable; therefore, nonparametric techniques with user-chosen parameters (such as kernel and bandwidth) are avoided to obtain a consistent estimate. It also completely circumvents the "curse of dimensionality" faced by nonparametric estimation.

5 A simple test for correct expectile specification

In this section, we consider testing the correct specification of a linear conditional expectile function over a continuum of expectile levels. The null hypothesis of interest is

$$H_0 : \mathbb{E}\left[\psi(W, \beta_0(\tau), \tau) \,|X\right] = 0 \text{ a.s. for some } \beta_0 \in \mathscr{B} \text{ and for all } \tau \in \mathcal{T}, \tag{11}$$

where $\psi(W, \beta(\tau), \tau)$ is given in (9), and $\mathscr{B}$ is a family of uniformly bounded functions from $\mathcal{T}$ to $\mathcal{B} \subseteq \mathbb{R}^d$. The alternative H_1 is the negation of H_0. The null implies the moment restriction $\mathbb{E}\big[\psi(W, \beta_0(\tau), \tau)m(X)\big] = 0$, for all measurable functions $m(\cdot)$ such that $\mathbb{E}|m(X)| < \infty$ and all $\tau \in \mathcal{T}$. Given a random sample $\{W_i = (Y_i, X_i^\mathsf{T})^\mathsf{T}\}$, it is natural to base the test for H_0 on the following residual marked empirical process

$$\widehat{R}_n(\tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \widehat{\beta}_n(\tau), \tau) M(X_i), \tag{12}$$

where $M(\cdot)$ is a measurable $s \times 1$ vector function. Note that to avoid degeneracies, $M(X_i)$ cannot have elements equal to any element of X_i, as the first order condition implies that $n^{-1/2} \sum_{i=1}^{n} \psi(W_i, \widehat{\beta}_n(\tau), \tau) X_i = 0$ no matter whether H_0 holds.

The test statistics are defined as continuous functionals of $\widehat{R}_n$, say, $\Gamma(\widehat{R}_n)$, for example,

$$\widetilde{K}_n = \sup_{\tau \in \mathcal{T}} \left\| \widehat{R}_n(\tau) \right\|, \quad \text{and} \quad \widetilde{C}_n = \int_{\mathcal{T}} \left\| \widehat{R}_n(\tau) \right\|^2 d\Phi(\tau). \tag{13}$$

To derive the asymptotic distribution of the test statistics, we make the following assumption.

Assumption 5. *The function $M(\cdot)$ satisfies $\mathbb{E}\|M(X)\|^2 < \infty$ and $\mathbb{E}[\|W\|\|M(X)\|]^{2+\delta} < \infty$ for some $\delta > 0$.*

As shown in the Appendix, if the linear conditional expectile model is correctly specified and Assumptions 1 and 5 hold, then

$$\sup_{\tau \in \mathcal{T}} \left| \widehat{R}_n(\tau) - R_n(\tau) + G(\beta(\tau), \tau) n^{-1/2} \sum_{i=1}^{n} l(W_i, \beta(\tau), \tau) \right| = o_p(1), \tag{14}$$

where $R_n(\tau) \equiv n^{-1/2} \sum_{i=1}^{n} \psi(W_i, \beta(\tau), \tau) M(X_i)$, $G(\beta(\tau), \tau) \equiv \mathbb{E}\big[(\tau + (1 - 2\tau) F_Y(X^\mathsf{T}\beta(\tau) \mid X)) M(X) X^\mathsf{T}\big]$, and

$$l(W_i, \beta(\tau), \tau) = J(\tau)^{-1} \psi(W_i, \beta(\tau), \tau) X_i. \tag{15}$$

Theorem 5 presents the asymptotic distribution of $\widehat{R}_n$ under the null hypothesis H_0.

Theorem 5. *Suppose that Assumptions 1 and 5 hold. Under the null hypothesis* H_0, $\widehat{R}_n \Rightarrow R^1_\infty$ *in* $l^\infty(\mathcal{T})$, *where* R^1_∞ *is a zero-mean Gaussian process with covariance function*

$$\begin{aligned}\mathbb{K}^1(\tau,\tau') =& \mathbb{E}\big[\psi(W,\beta(\tau),\tau)\psi(W,\beta(\tau'),\tau')M(X)M(X)^\mathsf{T}\big] \\ &+ G(\beta(\tau),\tau)J(\tau)^{-1}\Sigma(\tau,\tau')J(\tau')^{-1}G(\beta(\tau'),\tau')^\mathsf{T} \\ &- \mathbb{E}\big[\psi(W,\beta(\tau),\tau)M(X)l(W,\beta(\tau'),\tau')^\mathsf{T}\big]G(\beta(\tau'),\tau')^\mathsf{T} \\ &- G(\beta(\tau),\tau)\mathbb{E}\big[\psi(W,\beta(\tau'),\tau')l(W,\beta(\tau),\tau)M(X)^\mathsf{T}\big].\end{aligned}$$

We see from Theorem 5 that the structure of $\mathbb{K}^1$ does not allow for a simple representation of R^1_∞ in terms of a well-known distribution-free Gaussian process. To obtain the critical values, we propose a multiplier bootstrap procedure similar to the one described in Section 4. Specifically, we propose to approximate the asymptotic behavior of $\widehat{R}_n(\tau)$ by that of

$$\widehat{R}^*_n(\tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big[\psi(W_i,\widehat{\beta}_n(\tau),\tau)M(X_i) - \widehat{G}_n(\widehat{\beta}_n(\tau),\tau)\widehat{l}(W_i,\widehat{\beta}_n(\tau),\tau)\Big]V_i, \tag{16}$$

where $\{V_i\}_{i=1}^n$ is the sequence of i.i.d. random variables with zero mean, unit variance, bounded support and independent of the sample $\{W_i\}_{i=1}^n$, the vector $\widehat{l}(\cdot)$ is defined in the same way as $l(\cdot)$ in (15) except that $J(\cdot)$ is substituted by $\widehat{J}_n(\cdot)$ in (10), and

$$\widehat{G}_n(\widehat{\beta}_n(\tau),\tau) = \frac{1}{n}\sum_{i=1}^n \big|1\big(Y_i < X_i^\mathsf{T}\widehat{\beta}_n(\tau)\big) - \tau\big|M(X_i)X_i^\mathsf{T}$$

is a consistent estimate of the function $G(\beta(\tau),\tau)$ defined below (14). With $\widehat{R}^*_n(\tau)$ at hands, the bootstrapped version of our test statistics $\Gamma(\widehat{R}_n)$ is simply given by $\Gamma(\widehat{R}^*_n)$. For example, the bootstrap version of the KS and CvM test statistics in (13) are given by

$$\widetilde{K}^*_n = \sup_{\tau\in\mathcal{T}}\big\|\widehat{R}^*_n(\tau)\big\|, \text{ and } \widetilde{C}_n = \int_{\mathcal{T}}\big\|\widehat{R}^*_n(\tau)\big\|^2 d\Phi(\tau).$$

The bootstrap empirical distribution of $\Gamma(\widehat{R}^*_n)$, i.e., $\widehat{F}^*_n(t|\{W_i\}_{i=1}^n) = \mathrm{P}(\Gamma(\widehat{R}^*_n) \le t|\{W_i\}_{i=1}^n)$ is shown to be a consistent estimate of the asymptotic null distribution function of $\Gamma(\widehat{R}_n)$, i.e. $F_\infty(t) = \mathrm{P}(\Gamma(R^1_\infty) \le t)$. The null hypothesis will be rejected at the τ-level of significance when $\Gamma(\widehat{R}_n) \ge c^*_{n,\tau}$, where $c^*_{n,\tau}$ is such that $\widehat{F}^*_n(c^*_{n,\tau}|\{W_i\}_{i=1}^n) = 1-\tau$. We can also use bootstrap p-values in this context. In this case, the null could be rejected if $p^*_n < \tau$, where $p^*_n = \mathrm{P}\big(\Gamma(\widehat{R}^*_n) \ge \Gamma(\widehat{R}_n)|\{W_i\}_{i=1}^n\big)$.

The following theorem establishes the asymptotic validity of the multiplier bootstrap test presented above, which is a simple Monte Carlo resampling method to simulate the critical values of the test.

Theorem 6. *Suppose that Assumptions 1 and 5 hold. If the linear conditional expectile function is correctly specified, we have* $\widehat{R}_n^* \underset{*}{\Rightarrow} R_\infty^1$ *in probability, where* R_∞^1 *is the Gaussian process defined in Theorem 5. For any continuous functional* $\Gamma(\cdot)$ *from* $l^\infty(\mathcal{T})$ *to* $\mathbb{R}$*, we have* $\Gamma(\widehat{R}_n^*) \underset{*}{\xrightarrow{d}} \Gamma(R_\infty^1)$ *in probability.*

Theorem 6 implies that a test based on the multiplier bootstrap p-value would yield the correct asymptotic level for $\Gamma(\widehat{R}_n)$. Our test does not involve nonparametric smoothing and thus does not depend on any user-chosen tuning parameters, such as bandwidth. In addition, our test is computationally very fast. Nevertheless, it may not be consistent against all nonparametric alternatives. A consistent test can be constructed by first estimating the conditional expectile model nonparametrically and then comparing the nonparametric estimation with the parametric (linear or nonlinear) one, see, e.g., Zheng (1998) in the quantile regression setting. Alternatively, following He and Zhu (2003), we can consider a consistent test for parametric expectile regression based on a cusum process of the gradient vector. We leave these directions for future study.

6 Simulation

6.1 Estimation of ER process

We conduct simulation experiments to investigate the finite sample properties of the ER process. Samples are generated according to the following data-generating processes (DGPs):

$$\begin{aligned}
&\text{DGP 1: } Y_i = X_{1i} + X_{2i} + u_i,\\
&\text{DGP 2: } Y_i = X_{1i} + X_{2i} + X_{1i}^2 + X_{2i}^2 + 0.5X_{1i}X_{2i} + u_i,\\
&\text{DGP 3: } Y_i = X_{1i} + X_{2i} + \exp(0.5 * (X_{1i} + X_{2i})) + u_i,\\
&\text{DGP 4: } Y_i = X_{1i} + X_{2i} + \cos(0.5\pi * (X_{1i} + X_{2i})) + u_i,
\end{aligned}$$

where $\{(X_{1i}, X_{2i})\}_{i=1}^n$ are i.i.d. multivariate normal random vectors with $N(0,1)$ marginals and correlation coefficients 0.5. The error term u_i's are generated in two ways: (a) u_i's are i.i.d. $N(0,1)$, and (b) $u_i = \sqrt{0.5(1 + X_{1i}^2 + X_{2i}^2)}\epsilon_i$, where ϵ_i's are i.i.d. $N(0,1)$.

We consider sample sizes $n \in \{200, 500, 1000\}$ and a number of replications of 1,000. For each replication, we fit the following expectile regression model

$$Y_i = \widehat{\beta}_{0n}(\tau) + \widehat{\beta}_{1n}(\tau)X_{1i} + \widehat{\beta}_{2n}(\tau)X_{2i} + e_i(\tau),$$

where $\widehat{\beta}_n(\tau) = (\widehat{\beta}_{0n}(\tau), \widehat{\beta}_{1n}(\tau), \widehat{\beta}_{2n}(\tau))^\intercal$ is obtained by (4), over $m = 30$ equidistributed points $\{\tau_l\}_{l=1}^m$ on the interval $\mathcal{T} = [0.1, 0.9]$. The estimated coefficient $\widehat{\beta}_n(\tau)$ is compared

with the population ER vector $\beta(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau))^\mathsf{T}$ as defined in (1).[8] To evaluate the estimation performance, we record the root mean squared error (RMSE) and the coverage rate of the 95% simultaneous confidence interval (CR95) for $\beta_k(\cdot)$, $k = 0, 1, 2$. The RMSE for $\beta_k(\cdot)$ is calculated as $\sqrt{m^{-1}\sum_{l=1}^m (\widehat{\beta}_{kn}(\tau_l) - \beta_k(\tau_l))^2}$. The simultaneous confidence interval is constructed as $\left[\widehat{\beta}_{kn}(\tau_l) \pm \widehat{\kappa}(\alpha)\sqrt{V(\tau_l)/n}\right]$, where $V(\tau_l) = e_{k+1}^\mathsf{T} J(\tau_l)^{-1}\Sigma(\tau_l, \tau_l)J(\tau_l)^{-1}e_{k+1}$, e_{k+1} is a 3-dimensional vector with $(k+1)$th element being 1 and others being zero, and $\widehat{\kappa}(\alpha)$ is a consistent estimate of the 95% quantile of $\tilde{\mathcal{K}} = \sup_l |V(\tau_l)^{-1/2} e_{k+1}^\mathsf{T} J(\tau_l)^{-1} z(\tau_l)|$. In the simulation, the distribution of $\tilde{\mathcal{K}}$ is approximated by the multiplier bootstrap procedure proposed in Section 4 using $B = 200$ bootstrap replications.

Table 1 displays the RMSE and the coverage rate of the simultaneous confidence interval for β_0, β_1, and β_2. For all the DGPs, the RMSE is satisfactorily small with moderate sample sizes. The coverage rate of the 95% confidence interval is slightly below the nominal level, but it improves significantly as the sample size increases. Moreover, compared with the homoskedastic case, the estimation performance under heteroskedastic errors is mostly worse. This is expected because the signal-to-noise ratio is smaller when heteroscedasticity is present, resulting in a more variable estimate of the population ER vector. Finally, unreported simulation results also show that this pattern is robust to the distribution of errors.

6.2 Testing of homoskedasticity and symmetry

Next, we investigate the finite sample properties of test statistics K_n and C_n in (7) and (8) by simulation studies. We generate data from the following location-scale model:

$$Y_i = X_i^\mathsf{T}\beta + \sigma_i\epsilon_i, \quad i = 1, \ldots, n, \tag{17}$$

where $X_i = (X_{1i}, X_{2i})^\mathsf{T}$, and $\sigma_i = 1 + X_i^\mathsf{T}\eta_1 + 1(\epsilon_i > 0)X_i^\mathsf{T}\eta_2$. The random variables X_{1i} and X_{2i} are taken to be i.i.d. $N(0,1)$ and mutually independent. The errors ϵ_i's are i.i.d. $N(0,1)$ and independent of $\{X_i\}_{i=1}^n$. We aim to test the homoskedasticity and conditional symmetry, using the test procedures discussed in Remark 2. For the former test, the null hypothesis corresponds to model (17) with $\eta_1 = \eta_2 = 0$. For the latter test, the null hypothesis corresponds to model (17) with $\eta_2 = 0$.

We consider sample sizes $n \in \{100, 200\}$ and an interval of expectiles $\mathcal{T} = [0.1, 0.9]$. The test statistics are given by K_n and C_n in (7) and (8), respectively, with $\widehat{V}(\tau) = I_l$ and Φ a

[8] For DGP 1 with homogeneous errors, we have $\mu_\tau(Y|X) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2$ with $\beta_1(\tau) = \beta_2(\tau) = 1$ and $\beta_0(\tau)$ being the τth expectile of the standard normal distribution. For DGPs 2-4, the ER vector $\beta(\tau)$ in (1) is not obvious. In these cases, we simulate the ER vector $\beta(\tau)$ using the averaged sample ER process, which is calculated over 500 replications, each with 20,000 samples.

Table 1: Estimation of the ER vector $(\beta_0(\tau), \beta_1(\tau), \beta_2(\tau))$

			DGP 1		DGP 2		DGP 3		DGP 4	
Error	n	Coef	RMSE	CR95	RMSE	CR95	RMSE	CR95	RMSE	CR95
(a)	200	β_0	0.0638	0.919	0.2238	0.905	0.0988	0.872	0.0796	0.919
	200	β_1	0.0735	0.903	0.2876	0.901	0.1248	0.871	0.0881	0.908
	200	β_2	0.0721	0.912	0.2895	0.895	0.1237	0.883	0.0908	0.897
	500	β_0	0.0413	0.924	0.1365	0.925	0.0658	0.892	0.0501	0.925
	500	β_1	0.0460	0.927	0.1898	0.927	0.0817	0.886	0.0569	0.920
	500	β_2	0.0455	0.933	0.1844	0.921	0.0851	0.901	0.0568	0.923
	1000	β_0	0.0288	0.939	0.0958	0.948	0.0423	0.942	0.0353	0.923
	1000	β_1	0.0331	0.937	0.1381	0.923	0.0555	0.932	0.0410	0.939
	1000	β_2	0.0337	0.927	0.1365	0.929	0.0573	0.927	0.0396	0.931
(b)	200	β_0	0.0852	0.888	0.2290	0.923	0.1154	0.881	0.0935	0.906
	200	β_1	0.1115	0.899	0.3101	0.875	0.1550	0.878	0.1248	0.891
	200	β_2	0.1150	0.888	0.3138	0.906	0.1565	0.861	0.1281	0.877
	500	β_0	0.0538	0.929	0.1373	0.944	0.0729	0.924	0.0594	0.926
	500	β_1	0.0726	0.914	0.1997	0.918	0.1004	0.906	0.0836	0.896
	500	β_2	0.0727	0.915	0.1985	0.924	0.0991	0.920	0.0812	0.912
	1000	β_0	0.0378	0.939	0.1017	0.938	0.0535	0.923	0.0424	0.935
	1000	β_1	0.0510	0.937	0.1399	0.932	0.0699	0.928	0.0564	0.927
	1000	β_2	0.0517	0.921	0.1363	0.929	0.0708	0.924	0.0556	0.937

uniform discrete distribution over a grid of 60 equidistributed points from 0.1 to 0.9. The number of Monte Carlo replications is set to 1,000, and the number of sequences of bootstrap multipliers generated for each replication is set to 200.

Table 2 displays the rejection frequencies of K_n and C_n tests for testing of homoskedasticity. The nominal levels are given by 0.10, 0.05, and 0.01. The parameters η_1 and η_2 in model (17) are set to be $\eta_1 = (c_1, c_1)^\mathsf{T}$ and $\eta_2 = (c_2, c_2)^\mathsf{T}$, where c_1 and c_2 take values in $\{0, 0.25, 0.50\}$. Our findings are as follows. First, under the null hypothesis ($c_1 = c_2 = 0$), both tests are moderately oversized when $n = 100$, while the size inflation becomes milder as the sample size increases. Second, the empirical power is reasonably high for both tests. The CvM test appears to dominate the KS test for all the DGPs we consider when $c_1 \neq 0$ or $c_2 \neq 0$.

Table 3 displays the rejection frequencies of K_n and C_n tests for testing of conditional symmetry. The parameters η_1 and η_2 take the same values as in the test of homoskedasticity above. Under the null hypothesis ($c_2 = 0$), the empirical rejection probabilities of both K_n and C_n tests are close to the nominal levels. When $c_2 \neq 0$, our tests exhibit nontrivial power. As the sample size increases, the empirical power of both tests significantly improves. Note that larger values of $|c_2|$ imply higher power for the tests considered, whereas larger values of $|c_1|$ imply lower power of the tests. This is expected because in the presence of heteroskedasticity due to the term $X_i^\mathsf{T}\eta_1$ in σ_i, the signal/noise ratio is smaller than the case

Table 2: Rejection frequencies for the test of homoskedasticity

		$n = 100$						$n = 200$					
		K_n			C_n			K_n			C_n		
c_1	c_2	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
0	0	0.127	0.083	0.031	0.135	0.083	0.027	0.114	0.058	0.018	0.125	0.063	0.023
0.25	0	0.966	0.933	0.837	0.989	0.977	0.924	0.999	0.997	0.992	0.999	0.999	0.998
0.50	0	0.997	0.992	0.967	0.998	0.997	0.993	0.999	0.998	0.997	1	1	0.999
0	0.25	0.729	0.626	0.458	0.685	0.589	0.374	0.966	0.926	0.801	0.931	0.877	0.711
0.25	0.25	0.997	0.992	0.968	0.998	0.996	0.990	1	1	0.999	1	1	1
0.50	0.25	0.975	0.957	0.905	0.986	0.975	0.955	0.996	0.993	0.989	1	1	0.993
0	0.50	0.986	0.963	0.898	0.975	0.949	0.874	1	1	0.999	1	0.999	0.997
0.25	0.50	0.992	0.985	0.958	0.997	0.993	0.982	0.998	0.997	0.997	1	0.999	0.999
0.50	0.50	0.947	0.915	0.840	0.959	0.946	0.894	0.996	0.990	0.968	0.995	0.995	0.990

with pure nonsymmetry.

Table 3: Rejection frequencies for the test of symmetry

		$n = 100$						$n = 200$					
		K_n			C_n			K_n			C_n		
c_1	c_2	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
0	0	0.109	0.062	0.017	0.096	0.049	0.013	0.094	0.044	0.015	0.096	0.042	0.010
0.25	0	0.107	0.050	0.011	0.093	0.050	0.014	0.118	0.061	0.022	0.113	0.059	0.019
0.50	0	0.097	0.056	0.017	0.092	0.044	0.012	0.100	0.049	0.011	0.091	0.050	0.014
0	0.25	0.252	0.168	0.068	0.256	0.165	0.068	0.499	0.393	0.196	0.540	0.411	0.206
0.25	0.25	0.172	0.105	0.041	0.169	0.108	0.035	0.208	0.130	0.060	0.229	0.142	0.060
0.50	0.25	0.156	0.087	0.031	0.129	0.073	0.023	0.166	0.095	0.036	0.168	0.101	0.037
0	0.50	0.484	0.336	0.188	0.485	0.370	0.194	0.844	0.736	0.535	0.863	0.763	0.562
0.25	0.50	0.272	0.178	0.078	0.247	0.167	0.079	0.447	0.342	0.187	0.469	0.350	0.191
0.50	0.50	0.187	0.109	0.042	0.169	0.099	0.041	0.286	0.190	0.086	0.296	0.200	0.090

6.3 Testing the correct expectile specification

We finally examine the performance of the test for the correct expectile specification given in Section 5. The DGPs considered are the same as in Section 6.1, expect that the error terms are generated in two new ways: (a) u_i's are i.i.d. $N(0,1)$, and (c) $u_i = (1+0.5X_{1i}+0.5X_{2i})\epsilon_i$, where ϵ_i's are i.i.d. $N(0,1)$, and $\{\epsilon_i\}_{i=1}^n$ are independent of $\{(X_{1i}, X_{2i})\}_{i=1}^n$. Then, the null hypothesis (11) is true in DGP 1, but not in DGPs 2-4.

Table 4 reports the rejection frequencies of $\widetilde{K}_n$ and $\widetilde{C}_n$ in (13) over 1,000 Monte Carlo repetitions at the 10%, 5%, and 1% significance levels. The weighting function $M(\cdot)$ in (12) is chosen as $M(X_i) = (X_{1i}^2, X_{2i}^2, X_{1i} * X_{2i})^\intercal$. Our findings from Table 4 are as follows. First, the results of DGP 1 suggest that there is mild size inflation for both $\widetilde{K}_n$ and $\widetilde{C}_n$ tests at 10% and 5% levels. Second, when the true CEF is nonlinear and highly correlated with the $M(X)$

Table 4: Rejection frequencies for testing the correct specification of linear CEF

		$n = 100$						$n = 200$					
		K_n			C_n			K_n			C_n		
Error	DGP	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
(a)	1	0.109	0.040	0.009	0.118	0.046	0.010	0.115	0.060	0.010	0.126	0.065	0.012
	2	1	0.986	0.921	1	0.994	0.954	1	1	0.994	1	1	0.999
	3	0.743	0.569	0.249	0.774	0.623	0.296	0.957	0.900	0.677	0.961	0.926	0.745
	4	0.623	0.501	0.285	0.648	0.533	0.312	0.866	0.804	0.625	0.867	0.800	0.640
(c)	1	0.129	0.061	0.009	0.143	0.080	0.009	0.119	0.054	0.012	0.129	0.053	0.014
	2	0.999	0.982	0.901	1	0.989	0.937	1	0.998	0.993	1	1	0.997
	3	0.569	0.401	0.157	0.588	0.430	0.187	0.871	0.756	0.454	0.881	0.772	0.479
	4	0.468	0.340	0.166	0.475	0.356	0.193	0.630	0.536	0.354	0.610	0.518	0.343

function, as in DGP 2, our test exhibits good power properties for all sample sizes, regardless of whether heteroscedasticity is present. In contrast, if the CEF is nonlinear but not highly correlated with $M(X)$, as in DGPs 3 and 4, the power of the tests decreases; however, it can be improved by increasing the sample size. Lastly, the presence of heteroscedasticity has a negative impact on both the size and power performance of the tests.

7 Application to U.S. wage data

Returns to education have attracted significant attention in the field of labor economics since the 1970s. When analyzing returns to education across the wage distribution, most studies have employed the quantile regression approach. For example, Buchinsky (1994) and Martins and Pereira (2004) used linear quantile regression to show that education has a greater effect on the wages of individuals at the top of the wage distribution compared to those at the bottom. In this section, we demonstrate a similar conclusion, but using the ER approach. Our results suggest that linear ER provides a useful approximation to the conditional wage distribution and accurately captures changes in the wage distribution from 1980 to 2000. Furthermore, it offers critical advantages over the quantile regression method in terms of statistical efficiency and computational expediency.

The dataset investigated in this empirical study is the one used by Angrist et al. (2006). It comprises a sample of more than 65,023 U.S.-born black and white men aged 40-49 with a minimum of five years of education in 1980, 1990, and 2000. In our analysis, the response variable Y is the real log of weekly wage, calculated as the log of reported annual income from work divided by the number of weeks worked in the previous year, and the regressor consists of a years-of-schooling variable and other basic controls.

We are first interested in whether a linear specification of $\mu_\tau(Y|X)$ is correct over $\tau \in \mathcal{T}$, where Y is the real log weekly wage, X is the years of schooling, and $\mathcal{T}$ denotes one of

$[0.1, 0.9]$, $[0.1, 0.5]$ and $[0.5, 0.9]$. Implementing the bootstrap test in Section 5, we found that the null hypothesis of linearity is rejected at 1% significance level for both KS and CvM tests when either $\mathcal{T} = [0.1, 0.9]$, $[0.1, 0.5]$ or $[0.5, 0.9]$ is considered. This implies that $X^\intercal\beta(\tau)$ is just a linear approximation of the true CEF.

The nature of the ER approximation property is illustrated in Figure 1. Panels A-C plot a nonparametric estimate of the conditional expectile function, and the linear ER fit for the 0.10, 0.50, and 0.90 expectiles. We also compare the ER fit to an explicit minimum distance (MD) fit. The MD estimator for ER is the sample analog of the vector $\tilde{\beta}(\tau)$ that solves

$$\tilde{\beta}(\tau) = \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}[(\mu_\tau(Y|X) - X^\intercal\beta)^2] = \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}[D_\tau^2(X,\beta)].$$

That is, $\tilde{\beta}(\tau)$ is the slope of the linear regression of $\mu_\tau(Y|X)$ on X, weighted only by the probability mass function of X, $\pi(x)$. In contrast to the ER estimator, the MD estimator relies on the ability to estimate $\mu_\tau(Y|X)$ in a nonparametric first step. Figure 1 plots the MD fit with a dashed line. We see that the ER and MD regression lines are close, but they are not identical because the additional weighting by $w_\tau(X,\beta)$ in the ER fit accentuates the quality of the fit at values of X where $\mathbb{E}[|\tau - 1(Y \leq \mu_\tau(Y|X))| \,|\, X]$ is relatively large. To further investigate the ER weighting functions, panels D-F in Figure 1 present the estimates of the overall ER weights $w_\tau(X,\beta(\tau))\cdot\pi(X)$, importance weights $w_\tau(X,\beta(\tau))$, and their distribution function approximations $\tau + (1-2\tau)F_Y(\mu_\tau(Y|X)|X)$. It is seen that the importance weights and the distribution function weights are fairly close. Both are constant over X when $\tau = 0.5$. The overall weighting function assigns the highest weight to 12 years of schooling, implying that the linear ER fit should be the best in the middle of the design.

We next examine the ability of ER to capture the changes over time in expectile-based measures of conditional inequality. The column labeled "CE" in panel A of Table 5 shows nonparametric estimates of the average 90-10 expectile spread conditional on schooling, potential experience, and race. This spread increased from 0.98 to 1.03 from 1980 to 1990, and then to about 1.11 from 1990 to 2000. Expectile regression estimates match this almost perfectly, as expected by our theory. However, the fit is not as good when a specific schooling group is considered, as is shown in panels B and C of the table. Table 5 also shows that the expectile-based conditional inequality increases mainly in the upper half of the wage distribution from 1980 to 2000. Moreover, the increase in conditional inequality has been much larger for college graduates than for high school graduates.

Figure 2 provides a partial explanation for the patterns and changes in Table 5. It shows estimates of the schooling coefficient expectile process, along with robust simultaneous 95% confidence bands. These estimates are from expectile regressions of log earnings on years of schooling, race, and a quadratic function of experience, using the data from the 1980, 1990,

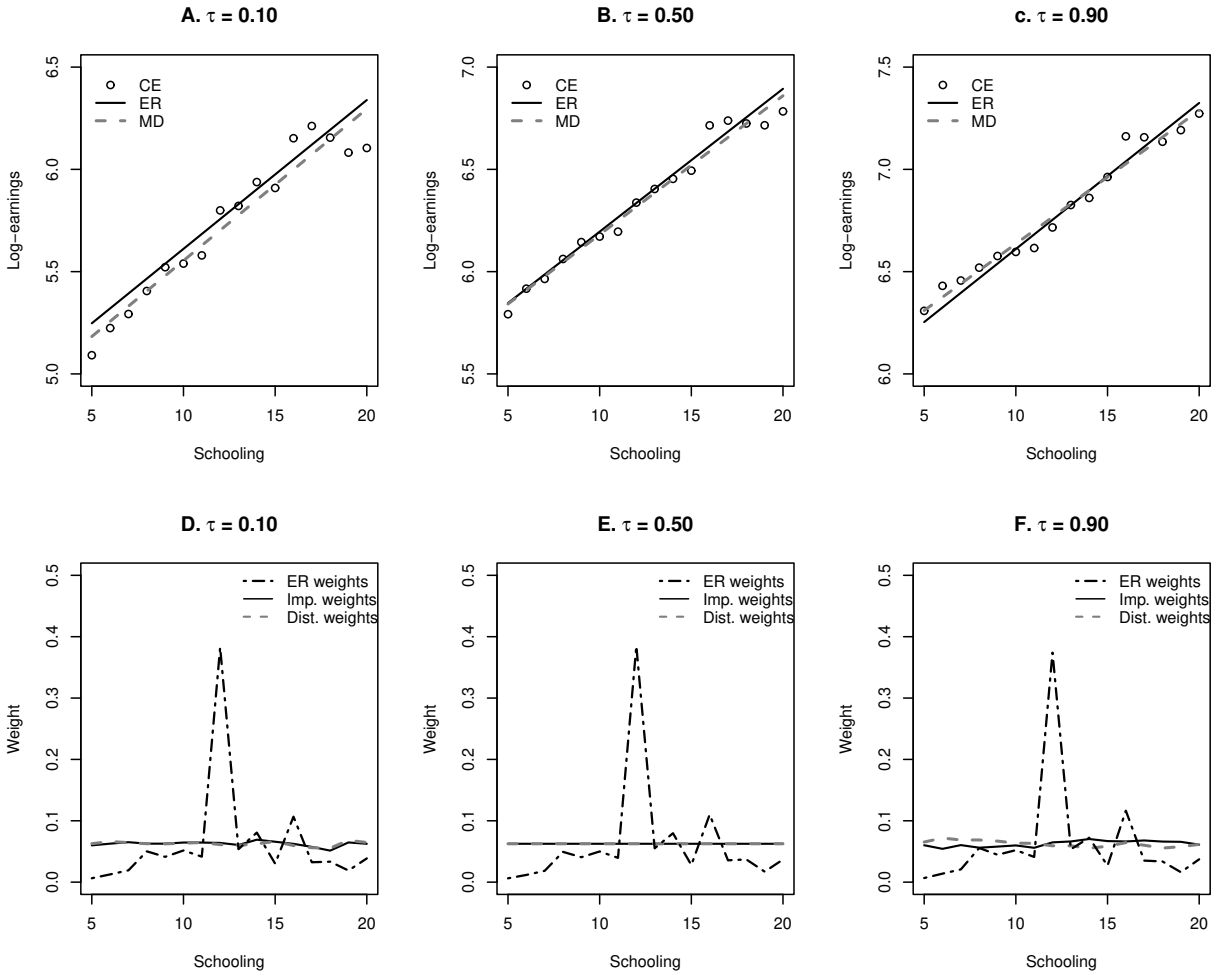

Figure 1: Conditional expectile function and weighting schemes in the 1980 Census.

and 2000 censuses.[9] The figure shows that the returns to schooling were low and essentially constant across expectiles in 1980, which is consistent with the finding by Buchinsky (1994). However, the return to schooling increased and became more heterogeneous in 1990 and especially in 2000. Since the simultaneous confidence bands do not contain a horizontal line, we reject the hypothesis of constant returns to schooling for 1990 and 2000. Further, the fact that there are expectile segments where the confidence bands do not overlap implies a statistically significant difference across years at those segments. For instance, the 1990 band does not overlap with the 1980 band, suggesting a statistically significant change in the relationship between schooling and the wage distribution in this period.

Compared with the quantile regression process, the ER process has narrower confidence bands. Indeed, using the subsampling method in Angrist et al. (2006), we find that

[9]The simultaneous bands were obtained by multiplier bootstrap using 200 bootstrap replicates with a grid of expectiles $\mathcal{T}_n = \{0.10, 0.11, \ldots, 0.90\}$.

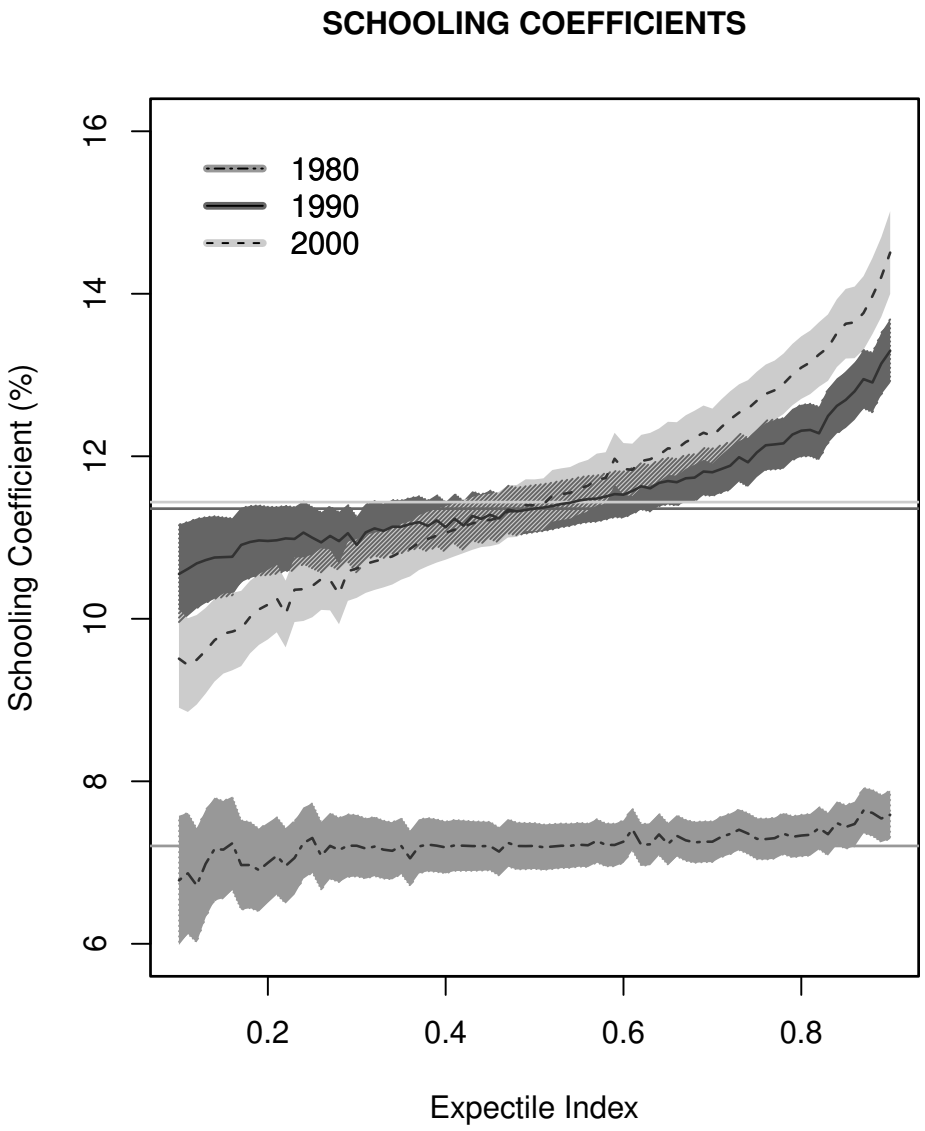

Figure 2: Schooling coefficients in the 1980, 1990, and 2000 censuses (for U.S.-born black and white men aged 40–49). The simultaneous 95% confidence bands are given by the shaded regions. The horizontal lines indicate OLS estimates of the schooling coefficients.

Table 5: Comparison of CEF and ER-based interexpectile spreads

		Interexpectile Spread					
		90-10		90-50		50-10	
Census	Obs.	CE	ER	CE	ER	CE	ER
		A. Overall					
1980	65,023	0.98	0.99	0.41	0.42	0.57	0.57
1990	86,785	1.03	1.04	0.47	0.48	0.56	0.56
2000	97,397	1.11	1.12	0.53	0.54	0.58	0.58
		B. High school graduates					
1980	25,020	0.91	0.98	0.38	0.41	0.53	0.57
1990	22,837	0.96	1.01	0.43	0.45	0.53	0.55
2000	27,697	1.00	1.04	0.47	0.49	0.53	0.55
		C. College graduates					
1980	7,158	1.00	0.99	0.44	0.42	0.56	0.56
1990	15,517	1.10	1.08	0.53	0.51	0.58	0.57
2000	18,329	1.24	1.22	0.61	0.61	0.62	0.62

the confidence interval lengths for the schooling coefficient quantile process averaged over $\{0.1,0.2,\ldots,0.9\}$ quantiles are 0.94, 1.11, and 1.56 for the 1980, 1990, and 2000 censuses, respectively. For the ER process plotted in Figure 2, the confidence interval lengths averaged over $\{0.1,0.2,\ldots,0.9\}$ expectiles are 0.76, 0.71, and 0.84 for 1980, 1990, and 2000 censuses, respectively. This suggests that ER regression would be a more efficient method for analyzing the entire conditional wage distribution in this data.

Finally, Figure 3 offers another view of the stylized facts in Table 5. This figure shows the changes in the approximate conditional expectiles based on an ER fit, where the covariates are evaluated at their mean values for each year. The 95% simultaneous confidence bands are also plotted. This figure provides a visual representation of the finding that the conditional wage inequality increases more in the upper half of the wage distribution than in the lower half. Changes in schooling coefficients across expectiles and years, sharper above the mean than below, contributed to the fact that recent inequality growth has been mostly confined to the upper half of the wage distribution.

8 Conclusion

This paper shows how linear ER provides a weighted least squares approximation to an unknown and possibly nonlinear conditional expectile function. The ER approximation property is further utilized to examine a partial ER relationship and the bias caused by omitted variables. All the results can be seen as an asymmetric generalization of OLS. We

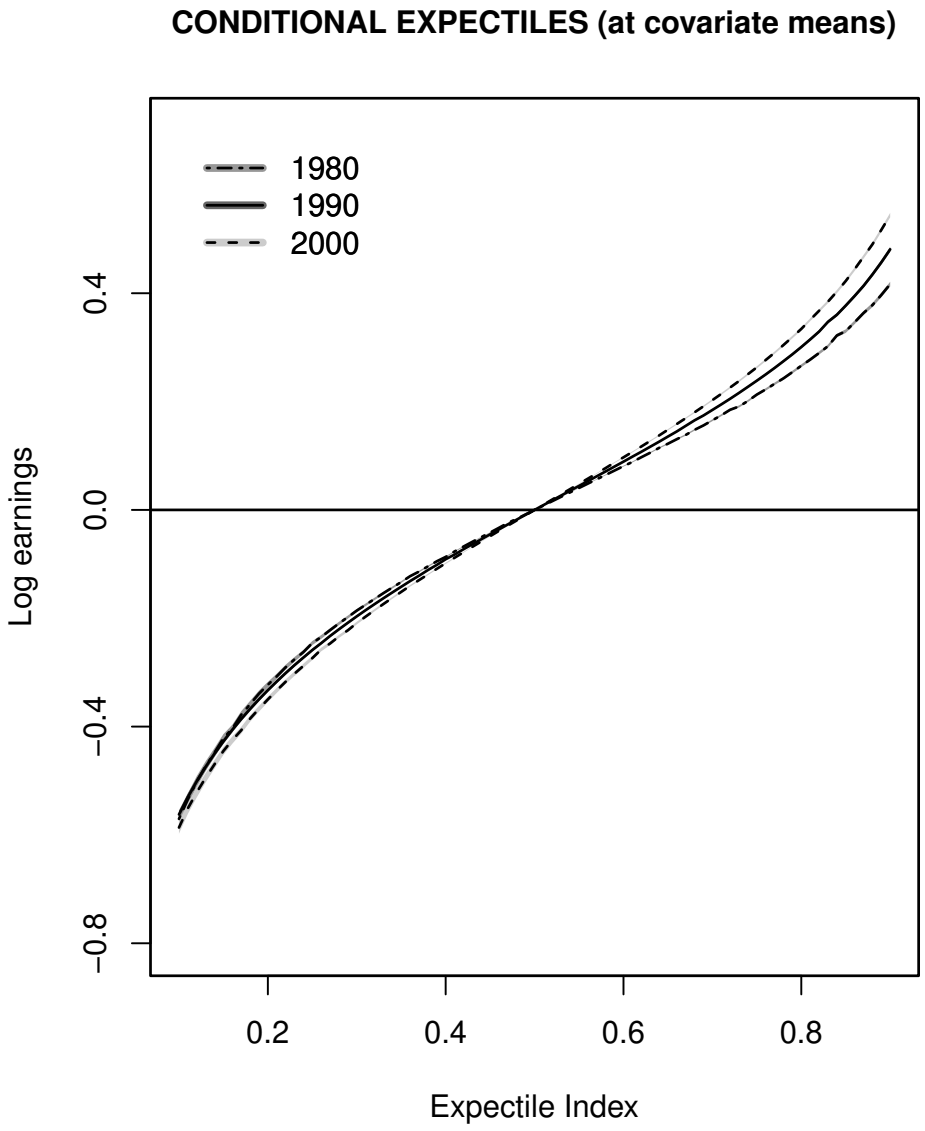

Figure 3: Conditional expectiles of log earnings in the 1980, 1990, and 2000 Census (for U.S.-born black and white men aged 40–49). The figure displays simultaneous 95% confidence bands for the ER approximation to the conditional expectile function, given schooling, race, and a quadratic function of experience. For each expectile τ and year, $\mathbb{E}[X]^{\mathsf{T}}(\widehat{\beta}_n(\tau)-\widehat{\beta}_n(0.5))$ is plotted.

also present a distribution theory for the ER process, allowing for possibly misspecified conditional expectile functions. This provides a foundation for simultaneous confidence intervals and a basis for global tests of hypotheses about distributions. An analysis of wage data from the U.S. Census illustrates the usefulness of the proposed inference procedures.

Appendix

Throughout the appendix, the symbol C is a generic constant that may change from one expression to another. For a random variable X, let $\|X\|_2 = [\mathbb{E}(X^2)]^{1/2}$. For a normed space of real functions $(\mathcal{G}, \rho)$, let $N_{[\cdot]}(\epsilon, \mathcal{G}, \rho)$ denote the covering number with bracketing, and $J_{[\cdot]}(\delta, \mathcal{G}, \rho) = \int_0^\delta \sqrt{\log N_{[\cdot]}(\epsilon, \mathcal{G}, \rho)} d\epsilon$ denote the entropy with bracketing, see Van Der Vaart (1998, p.270) for details.

A Proof of Theorems

Proof of Theorem 1.

The unique existence of $\beta(\tau)$ can be derived using the arguments of Theorem 3 in Newey and Powell (1987). The rest of the proof proceeds by proving the equivalence of the two objective functions. We have

$$\beta(\tau) = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \mathbb{E}[\rho_\tau(\epsilon_\tau - D_\tau(X, \beta))] - \mathbb{E}[\rho_\tau(\epsilon_\tau)] \right\}.$$

By the definition of ρ_τ, it follows further that

$$\beta(\tau) = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \mathbb{E}[\mathcal{T}(X, \beta)] - 2\mathbb{E}[\mathcal{B}(X, \beta)] + \mathbb{E}[\mathcal{C}(X, \beta)] \right\}, \tag{1}$$

where

$$\begin{aligned}
\mathcal{T}(X, \beta) &\equiv \mathbb{E}[|\tau - 1(\epsilon_\tau < D_\tau(X, \beta))| D_\tau^2(X, \beta) \mid X], \\
\mathcal{B}(X, \beta) &\equiv \mathbb{E}[|\tau - 1(\epsilon_\tau < D_\tau(X, \beta))| \epsilon_\tau D_\tau(X, \beta) \mid X], \\
\mathcal{C}(X, \beta) &\equiv (1 - 2\tau)\mathbb{E}[\epsilon_\tau^2 [1(\epsilon_\tau < D_\tau(X, \beta)) - 1(\epsilon_\tau < 0)] \mid X].
\end{aligned}$$

First, it is easy to see that

$$\mathcal{T}(X, \beta) = \left[\tau + (1 - 2\tau) F_{\epsilon_\tau}(D_\tau(X, \beta) \mid X)\right] \cdot D_\tau^2(X, \beta).$$

For $\mathcal{B}(X, \beta)$, suppose first that $D_\tau(X, \beta) > 0$. By the first order condition, we have $\mathbb{E}[|\tau -$

$1(\epsilon_\tau < 0)|\epsilon_\tau \mid X] = 0$. Then,

$$\begin{aligned}\mathcal{B}(X,\beta) &= \mathbb{E}\big\{\big[\big|\tau - 1(\epsilon_\tau < D_\tau(X,\beta))\big| - \big|\tau - 1(\epsilon_\tau < 0)\big|\big]\epsilon_\tau D_\tau(X,\beta) \mid X\big\} \\ &= (1-2\tau)\int_0^{D_\tau(X,\beta)} y f_{\epsilon_\tau}(y|X)dy \cdot D_\tau(X,\beta) \\ &= (1-2\tau)\int_0^1 u f_{\epsilon_\tau}(uD_\tau(X,\beta)|X)du \cdot D_\tau^3(X,\beta). \end{aligned} \tag{2}$$

Finally,

$$\begin{aligned}\mathcal{C}(X,\beta) &= (1-2\tau)\int_0^{D_\tau(X,\beta)} y^2 f_{\epsilon_\tau}(y|X)dy \\ &= (1-2\tau)\int_0^1 u^2 f_{\epsilon_\tau}(uD_\tau(X,\beta)|X)du \cdot D_\tau^3(X,\beta). \end{aligned} \tag{3}$$

A similar argument shows that (2) and (3) also hold when $D_\tau(X,\beta) < 0$. If $D_\tau(X,\beta) = 0$, then $\mathcal{B}(X,\beta) = \mathcal{C}(X,\beta) = 0$, so that (2) and (3) hold in this case too. Combining the above results yields $\beta(\tau) = \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}[w_\tau(X,\beta)\cdot D_\tau^2(X,\beta)]$, where

$$\begin{aligned} w_\tau(X,\beta) &= \tau + (1-2\tau)\bigg(F_{\epsilon_\tau}(D_\tau(X,\beta)\,|\,X) - \int_0^1 D_\tau(X,\beta)u(2-u)f_{\epsilon_\tau}(uD_\tau(X,\beta)\,|\,X)du\bigg) \\ &= \tau + (1-2\tau)\bigg(F_{\epsilon_\tau}(D_\tau(X,\beta)\,|\,X) - \int_0^1 u(2-u)dF_{\epsilon_\tau}(uD_\tau(X,\beta)\,|\,X)\bigg) \\ &= \tau + 2(1-2\tau)\int_0^1 (1-u)F_{\epsilon_\tau}(uD_\tau(X,\beta)\mid X)du. \end{aligned}$$

Here, the third equation follows from integration by parts. This proves Theorem 1. □

Proof of Theorem 2.

It suffices to show that

$$\beta(\tau) = \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}[\rho_\tau(Y - X^\mathsf{T}\beta) - \rho_\tau(Y)] \tag{4}$$

is equivalent to the fixed point $\bar\beta(\tau)$ that uniquely solves

$$\bar\beta(\tau) = \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}\big[\bar w_\tau(X,\bar\beta(\tau))\cdot D_\tau^2(X,\beta)\big]. \tag{5}$$

Both objective functions are finite by the stated conditions.

By convexity of (5) in β, any fixed point $\beta = \bar\beta(\tau)$ solves the first-order condition $\mathcal{F}(\beta) \equiv 2\mathbb{E}\big[\bar w_\tau(X,\beta)D_\tau(X,\beta)X\big] = 0$. Since (4) is convex on β, the ER vector solves the first-order

condition $\mathcal{H}(\beta) \equiv \mathbb{E}[\mathcal{H}(X,\beta)] = 0$, where

$$\mathcal{H}(X,\beta) = -2\mathbb{E}\big[\big|\tau - 1(\epsilon_\tau < D_\tau(X,\beta))\big|(\epsilon_\tau - D_\tau(X,\beta))X \mid X\big].$$

An argument similar to that used to establish equation (2) yields

$$\begin{aligned}
\mathcal{H}(X,\beta) =\ & 2\mathbb{E}\big[\big|\tau - 1(\epsilon_\tau < D_\tau(X,\beta))\big|\big|X\big]D_\tau(X,\beta)X \\
& - 2\mathbb{E}\big[\big(\big|\tau - 1(\epsilon_\tau < D_\tau(X,\beta))\big| - \big|\tau - 1(\epsilon_\tau < 0)\big|\big)\epsilon_\tau \mid X\big]X \\
=\ & 2\big[\tau + (1-2\tau)F_{\epsilon_\tau}(D_\tau(X,\beta) \mid X)\big] \cdot D_\tau(X,\beta)X \\
& - 2(1-2\tau)\int_0^1 u f_{\epsilon_\tau}\big(uD_\tau(X,\beta)|X\big)du \cdot D_\tau^2(X,\beta)X \\
=\ & 2\bigg(\tau + (1-2\tau)\int_0^1 F_{\epsilon_\tau}\big(u \cdot D_\tau(X,\beta)|X\big)du\bigg) \cdot D_\tau(X,\beta)X \\
=\ & 2\bar{w}_\tau(X,\beta)D_\tau(X,\beta)X.
\end{aligned}$$

The functions $\mathcal{F}(\beta)$ and $\mathcal{H}(\beta)$ are therefore identical. Since $\beta = \beta(\tau)$ uniquely satisfies $\mathcal{H}(\beta) = 0$, it also uniquely satisfies $\mathcal{F}(\beta) = 0$. Therefore, $\beta = \beta(\tau) = \bar{\beta}(\tau)$ is the unique solution to both (4) and (5). □

Proof of Theorem 3.

For $W = (Y, X^\mathsf{T})^\mathsf{T}$, denote $\mathbb{E}_n[f(W)] = n^{-1}\sum_{i=1}^n f(W_i)$ and $\mathbb{G}_n[f(W)] = n^{-1/2}\sum_{i=1}^n (f(W_i) - \mathbb{E}[f(W_i)])$. For a square matrix A, let $\lambda_{\min}(A)$ denote its minimum eigenvalue.

For each $\tau \in \mathcal{T}$, $\widehat{\beta}_n(\tau)$ minimizes $Q_n(\tau,\beta) = \mathbb{E}_n[\rho_\tau(Y - X^\mathsf{T}\beta) - \rho_\tau(Y - X^\mathsf{T}\beta(\tau))]$. Define $Q_\infty(\tau,\beta) = \mathbb{E}[\rho_\tau(Y - X^\mathsf{T}\beta) - \rho_\tau(Y - X^\mathsf{T}\beta(\tau))]$. It can be easily shown that $\mathbb{E}\|W\|^2 < \infty$ implies that $\mathbb{E}|\rho_\tau(Y - X^\mathsf{T}\beta)| < \infty$. Therefore, $Q_\infty(\tau,\beta)$ is finite and, under the stated assumptions, it is uniquely minimized at $\beta(\tau)$ for each $\tau \in \mathcal{T}$.

We first show the uniform convergence, namely for any compact set $\mathcal{B}$, $Q_n(\tau,\beta) = Q_\infty(\tau,\beta) + o_p(1)$ uniformly in $(\tau,\beta) \in \mathcal{T} \times \mathcal{B}$. Pointwise convergence follows Khinchin's law of large numbers. Further, the empirical process $(\tau,\beta) \mapsto Q_n(\tau,\beta)$ is stochastically equicontinuous because $|Q_n(\tau',\beta') - Q_n(\tau'',\beta'')| \le C_{1n} \cdot |\tau' - \tau''| + C_{2n} \cdot \|\beta' - \beta''\|$, where $C_{1n} = 2\mathbb{E}_n\|W\|^2(1 + \sup_{\beta\in\mathcal{B}}\|\beta\|^2) = O_p(1)$, and $C_{2n} = C\mathbb{E}_n\|W\|^2(1 + \sup_{\beta\in\mathcal{B}}\|\beta\|) = O_p(1)$. Hence, the convergence also holds uniformly. Using similar arguments as Angrist et al. (2006, p. 559), we can show that $\widehat{\beta}_n(\cdot)$ is uniformly consistent. The details are omitted for brevity.

Next, we establish the asymptotic Gaussianity of the sample ER process. With abuse of notation, denote $\psi_\tau(\lambda) = |\tau - 1(\lambda < 0)| \cdot \lambda$. Then we have $\partial\rho_\tau(Y_i - X_i^\mathsf{T}\beta)/\partial\beta = -2X_i\psi_\tau(Y_i - X_i^\mathsf{T}\beta)$. Recall that $\widehat{\beta}_n(\tau)$ minimizes $Q_n(\tau,\beta) = \mathbb{E}_n[\rho_\tau(Y - X^\mathsf{T}\beta) - \rho_\tau(Y - X^\mathsf{T}\beta(\tau))]$. By the

first order condition, we have for all $\tau \in \mathcal{T}$,

$$\sqrt{n}\mathbb{E}_n[\psi_\tau(Y - X^\mathsf{T}\widehat{\beta}_n(\tau))X] = 0. \tag{6}$$

Second, $(\tau, \beta) \mapsto \mathbb{G}_n[\psi_\tau(Y - X^\mathsf{T}\beta)X]$ is stochastically equicontinuous over $\mathcal{T} \times \mathcal{B}$, where $\mathcal{B}$ is any compact set, with respect to the $L_2(P)$ pseudometric $\rho((\tau', \beta'), (\tau'', \beta''))^2 \equiv \max_{j=1,\ldots,d} \mathbb{E}[(\psi_{\tau'}(Y - X^\mathsf{T}\beta')X_j - \psi_{\tau''}(Y - X^\mathsf{T}\beta'')X_j)^2]$, where $j = 1, \ldots, d$ indexes the components of X. To see this, note that $|\psi_{\tau'}(Y - X^\mathsf{T}\beta')X_j - \psi_{\tau''}(Y - X^\mathsf{T}\beta'')X_j| \leq |\tau' - \tau''|F_1(W) + \|\beta' - \beta''\|F_2(W)$, where $F_1(W) = \|W\|^2(1 + \sup_{\beta\in\mathcal{B}} \|\beta\|)$ and $F_2(W) = 2\|X\|^2$ are two square integrable functions. Applying Theorem 3 of Chen et al. (2003), we have $J_{[\cdot]}(\delta_n, H, \rho) \to 0$ for every $\delta_n \downarrow 0$, where H denotes the function class $\{W \to \psi_\tau(Y - X^\mathsf{T}\beta)X : \tau \in \mathcal{T}, \beta \in \mathcal{B}\}$. Besides, note that the class H has an envelope function with finite $(2+\delta)$th moment for some $\delta > 0$ under Assumption 1(c). Stochastic equicontinuity follows from the proof of Theorem 19.28 in Van Der Vaart (1998).

Third, by stochastic equicontinuity of $(\tau, \beta) \mapsto \mathbb{G}_n[\psi_\tau(Y - X^\mathsf{T}\beta)X]$ we have that

$$\mathbb{G}_n[\psi_\tau(Y - X^\mathsf{T}\widehat{\beta}_n(\tau))X] = \mathbb{G}_n[\psi_\tau(Y - X^\mathsf{T}\beta(\tau))X] + o_p(1) \quad \text{in } l^\infty(\mathcal{T}), \tag{7}$$

which follows from $\sup_{\tau\in\mathcal{T}} \|\widehat{\beta}_n(\tau) - \beta(\tau)\| = o_p(1)$ and resulting convergence with respect to the pseudometric $\sup_{\tau\in\mathcal{T}} \rho[(\tau, \widehat{\beta}_n(\tau)), (\tau, \beta(\tau))]^2 = o_p(1)$. The latter is immediate from $\sup_{\tau\in\mathcal{T}} \rho[(\tau, \beta(\tau)), (\tau, b(\tau))]^2 \leq C \sup_{\tau\in\mathcal{T}} \|\beta(\tau) - b(\tau)\|^2$, where $C = 2\mathbb{E}\|W\|^4 < \infty$.

Furthermore, the following expansion is valid uniformly in τ:

$$\mathbb{E}[\psi_\tau(Y - X^\mathsf{T}\beta)X]|_{\beta=\widehat{\beta}_n(\tau)} = -[J(\tau) + o_p(1)]\big(\widehat{\beta}_n(\tau) - \beta(\tau)\big). \tag{8}$$

Indeed, by Taylor expansion, $\mathbb{E}[\psi_\tau(Y - X^\mathsf{T}\beta)X]|_{\beta=\widehat{\beta}_n(\tau)} = -\mathbb{E}[|\tau - 1(Y < X^\mathsf{T}b(\tau))|XX^\mathsf{T}]|_{b(\tau)=\beta_n^*(\tau)} \times (\widehat{\beta}_n(\tau) - \beta(\tau))$, where $\beta_n^*(\tau)$ is on the line connecting $\widehat{\beta}_n(\tau)$ and $\beta(\tau)$ for each τ. Then (8) follows by the uniform consistency of $\widehat{\beta}_n(\tau)$, the continuity of $\mathbb{E}[|\tau - 1(Y < X^\mathsf{T}\beta)|XX^\mathsf{T}]$ with respect to β, and Assumption 1(d).

Fourth, we have that

$$o_p(1) = -[J(\cdot) + o_p(1)]\sqrt{n}\big(\widehat{\beta}_n(\cdot) - \beta(\cdot)\big) + \mathbb{G}_n[\psi_\cdot(Y - X^\mathsf{T}\beta(\cdot))X], \tag{9}$$

because the left-hand side of (6) is equal to the left-hand side of $n^{1/2}$ times (8) plus the left-hand side of (7). Since $\lambda_{\min}(J(\tau)) \geq \lambda > 0$, uniformly in $\tau \in \mathcal{T}$,

$$\sup_{\tau\in\mathcal{T}} \big\|\mathbb{G}_n[\psi_\cdot(Y - X^\mathsf{T}\beta(\cdot))X] + o_p(1)\big\| \geq (\sqrt{\lambda} + o_p(1)) \cdot \sup_{\tau\in\mathcal{T}} \sqrt{n}\|\widehat{\beta}_n(\tau) - \beta(\tau)\|. \tag{10}$$

Fifth, the mapping $\tau \mapsto \beta(\tau)$ is continuous by the implicit function theorem and

the stated assumptions. Because $\beta(\tau)$ solves $\mathbb{E}[|\tau - 1(Y < X^\mathsf{T}\beta)|(Y - X^\mathsf{T}\beta)X] = 0$, $d\beta(\tau)/d\tau = -J(\tau)^{-1}\mathbb{E}[|Y - X^\mathsf{T}\beta(\tau)|X]$. Hence $\tau \mapsto \mathbb{G}_n[\psi_\tau(Y - X^\mathsf{T}\beta(\tau))X]$ is stochastically equicontinuous over $\mathcal{T}$ for the pseudometric given by $\rho(\tau', \tau'') \equiv \rho((\tau', \beta(\tau')), (\tau'', \beta(\tau'')))$. Stochastic equicontinuity of $\tau \mapsto \mathbb{G}_n[\psi_\tau(Y - X^\mathsf{T}\beta(\tau))X]$ and a multivariate central limit theorem yield that

$$\mathbb{G}_n[\psi_\cdot(Y - X^\mathsf{T}\beta(\cdot))X] \Rightarrow z(\cdot) \quad \text{in } l^\infty(\mathcal{T}), \tag{11}$$

where $z(\cdot)$ is a Gaussian process with covariance function $\Sigma(\cdot, \cdot)$. Therefore, the left-hand side of (10) is $O_p(n^{-1/2})$, implying that $\sup_{\tau\in\mathcal{T}} \|\sqrt{n}(\widehat{\beta}_n(\tau) - \beta(\tau))\| = O_p(1)$. Finally, the latter fact and (9)-(11) yield that in $l^\infty(\mathcal{T})$,

$$J(\cdot)\sqrt{n}\big(\widehat{\beta}_n(\cdot) - \beta(\cdot)\big) = \mathbb{G}_n[\psi_\cdot(Y - X^\mathsf{T}\beta(\cdot))X] + o_p(1) \Rightarrow z(\cdot).$$

This completes the proof of the theorem. □

Proof of Corollary 1. Under the local alternative in Assumption 2, we have

$$\begin{aligned}\sqrt{n}v_n(\tau) &= \sqrt{n}\big(\widehat{R}_n(\tau)\widehat{\beta}_n(\tau) - \widehat{r}_n(\tau)\big) \\ &= R(\tau)\sqrt{n}\left[\widehat{\beta}_n(\tau) - \beta_n(\tau)\right] + \sqrt{n}\left[\widehat{R}_n(\tau) - R(\tau)\right]\widehat{\beta}_n(\tau) - \sqrt{n}[\widehat{r}_n(\tau) - r(\tau)] + p(\tau).\end{aligned}$$

Corollary 1 follows from Assumption 3, the consistency of $\widehat{\beta}_n(\cdot)$ and the continuous mapping theorem. □

Proof of Theorem 4. Using similar arguments as Theorem 3, we can show that the class of functions $\{(W, V) \to \psi(W, \beta, \tau)XV : \tau \in \mathcal{T}, \beta \in \mathcal{B}\}$ is Donsker. It follows from a stochastic equicontinuity argument and the consistency of $\widehat{\beta}_n(\cdot)$ that, uniformly in $\tau \in \mathcal{T}$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_\tau(W_i, \widehat{\beta}_n(\tau), \tau)X_iV_i = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_\tau(W_i, \beta(\tau), \tau)X_iV_i + o_p(1). \tag{12}$$

In a similar manner, we can obtain that uniformly in $\tau \in \mathcal{T}$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n L_\tau(W_i, \widehat{R}_n(\tau))V_i = \frac{1}{\sqrt{n}}\sum_{i=1}^n L_\tau(W_i, R(\tau))V_i + o_p(1), \tag{13}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n l_\tau(W_i, \widehat{r}_n(\tau))V_i = \frac{1}{\sqrt{n}}\sum_{i=1}^n l_\tau(W_i, r(\tau))V_i + o_p(1) \tag{14}$$

under Assumptions 3 and 4. Then, by the multiplier central limit theorem, see, e.g., Van Der Vaart and Wellner (1996, Theorem 2.9.6), we have $n^{-1/2}J(\cdot)^{-1}\sum_{i=1}^n \psi_\cdot(Y_i -$

$X_i^\mathsf{T}\beta(\cdot))X_iV_i \underset{*}{\Rightarrow} z(\cdot)$, $n^{-1/2}\sum_{i=1}^n L_\cdot(W_i, R(\cdot))V_i \underset{*}{\Rightarrow} \rho(\cdot)$, and $n^{-1/2}\sum_{i=1}^n l_\cdot(W_i, r(\cdot))V_i \underset{*}{\Rightarrow} \zeta(\cdot)$ in probability. The above results and the consistency of $\widehat{R}_n(\cdot)$ and $\widehat{\beta}_n(\cdot)$ yield that $\sqrt{n}v_n^*(\cdot) \underset{*}{\Rightarrow} v_0(\cdot)$ in probability. The rest of Theorem 4 follows from the continuous mapping theorem. □

Proof of uniform consistency of $\widehat{J}_n(\tau)$ and $\widehat{\Sigma}_n(\tau,\tau')$

Write $\widehat{\Sigma}_n(\tau,\tau') = n^{-1}\sum_{i=1}^n g(W_i, \widehat{\beta}(\tau), \widehat{\beta}(\tau'), \tau, \tau')X_iX_i^\mathsf{T}$, where $g(W_i, \beta, \beta', \tau, \tau') = \psi_\tau(Y_i - X_i^\mathsf{T}\beta)\psi_{\tau'}(Y_i - X_i^\mathsf{T}\beta')$. We aim to show that

$$\widehat{\Sigma}_n(\tau,\tau') - \Sigma(\tau,\tau') = o_p(1) \text{ uniformly in } (\tau,\tau') \in \mathcal{T}\times\mathcal{T}, \tag{15}$$

where $\Sigma(\tau,\tau')$ is defined in Theorem 3. It can be easily verified that $\{W \to g(W,\beta,\beta',\tau,\tau') : (\beta,\beta',\tau,\tau') \in \mathcal{B}\times\mathcal{B}\times\mathcal{T}\times\mathcal{T}\}$ is Donsker, and hence a Glivenko–Cantelli class, for any compact set $\mathcal{B}$, e.g., using Theorem 2.10.6 in Van Der Vaart and Wellner (1996). This implies that $\mathbb{E}_n[g(W,\beta,\beta',\tau,\tau')XX^\mathsf{T}] - \mathbb{E}[g(W,\beta,\beta',\tau,\tau')XX^\mathsf{T}] = o_p^*(1)$ uniformly in $(\beta,\beta',\tau,\tau') \in \mathcal{B}\times\mathcal{B}\times\mathcal{T}\times\mathcal{T}$. The latter, the continuity of $\mathbb{E}[g(W,\beta,\beta',\tau,\tau')XX^\mathsf{T}]$ in $(\beta,\beta',\tau,\tau')$, and the consistency of $\widehat{\beta}_n(\cdot)$ imply (15). In an analogous way, we can show that $\widehat{J}_n(\tau) - J(\tau) = o_p(1)$ uniformly in $\tau \in \mathcal{T}$. □

Proof of Theorem 5.

To show Theorem 5, we first state a weak convergence theorem. Given a sequence $\{W_i = (Y_i, X_i^\mathsf{T})^\mathsf{T}\}_{i=1}^n$ of i.i.d. variables, define the weighted empirical process

$$V_n(\gamma) \equiv n^{-1/2}\sum_{i=1}^n \big(\psi(W_i, \beta(\tau), \tau) - \mathbb{E}\big[\psi(W_i, \beta(\tau), \tau) \mid X_i\big]\big)\, M(X_i),$$

which is indexed by $\gamma = (\beta,\tau) \in \Gamma \equiv \mathscr{B}\times\mathcal{T}$, where $\mathscr{B}$ is a class of bounded, Lipschitz $\mathbb{R}^d$-valued functions on $\mathcal{T}$. Consider the following pseudo-metric

$$\rho(\gamma_1,\gamma_2) = \|\beta_1 - \beta_2\|_\mathcal{T} + |\tau_1 - \tau_2|, \tag{16}$$

where $\gamma_j = (\beta_j, \tau_j) \in \Gamma$, $j = 1, 2$, and $\|\beta_1 - \beta_2\|_\mathcal{T} = \sup_{\tau\in\mathcal{T}} \|\beta_1(\tau) - \beta_2(\tau)\|$.

We first show that under Assumptions 1 and 5, the process $V_n(\gamma)$ is ρ-stochastically equicontinuous. For $(\beta,\tau) \in \mathscr{B}\times\mathcal{T}$, define $\xi(W, \beta(\tau), \tau) = 1(Y < X^\mathsf{T}\beta(\tau)) - \tau$, $h(W,\beta,\tau) = \psi(W, \beta(\tau), \tau) - \mathbb{E}\big[\psi(W, \beta(\tau), \tau) \,|\, X\big]$, and $\mathcal{H} \equiv \{w \to h(w,\beta,\tau)M(x) : (\beta,\tau) \in \mathscr{B}\times\mathcal{T}\}$. Fix $(\beta_1,\tau_1) \in \mathscr{B}\times\mathcal{T}$. By the triangle inequality, we have

$$\mathbb{E}\left[\sup_{\beta:\|\beta-\beta_1\|_\mathcal{T}\le\delta}\sup_{\tau:|\tau-\tau_1|\le\delta} \big|h(W,\beta_1,\tau_1) - h(W,\beta,\tau)\big|^2\right]$$

$$
\begin{aligned}
&\leq C_1\mathbb{E}\bigg[\sup_{\beta:\|\beta-\beta_1\|_{\mathcal{T}}\leq\delta}\sup_{\tau:|\tau-\tau_1|\leq\delta}\big|\psi(W,\beta_1(\tau_1),\tau_1)-\psi(W,\beta(\tau),\tau)\big|^2\bigg]\\
&\quad+C_1\mathbb{E}\bigg[\sup_{\beta:\|\beta-\beta_1\|_{\mathcal{T}}\leq\delta}\sup_{\tau:|\tau_1-\tau|\leq\delta}\big|\mathbb{E}[\psi(W,\beta_1(\tau_1),\tau_1)\mid X]-\mathbb{E}[\psi(W,\beta(\tau),\tau)\mid X]\big|^2\bigg]\\
&\leq C_2\mathbb{E}\bigg[\sup_{\beta:\|\beta-\beta_1\|_{\mathcal{T}}\leq\delta}\sup_{\tau:|\tau-\tau_1|\leq\delta}\big|\xi(W,\beta_1(\tau_1),\tau_1)-\xi(W,\beta(\tau),\tau)\big|^2\big(Y-X^{\mathsf{T}}\beta_1(\tau_1)\big)^2\bigg]\\
&\quad+C_2\mathbb{E}\bigg[\sup_{\beta:\|\beta-\beta_1\|_{\mathcal{T}}\leq\delta}\sup_{\tau:|\tau-\tau_1|\leq\delta}\big|X^{\mathsf{T}}\beta(\tau)-X^{\mathsf{T}}\beta_1(\tau_1)\big|^2\bigg]\\
&\leq C_3\delta^2\mathbb{E}\big|Y-X^{\mathsf{T}}\beta_1(\tau_1)\big|^2+C_3\mathbb{E}\bigg[\|X\|^2\sup_{\beta:\|\beta-\beta_1\|_{\mathcal{T}}\leq\delta}\sup_{\tau:|\tau-\tau_1|\leq\delta}\big\|\beta(\tau)-\beta_1(\tau_1)\big\|^2\bigg]\\
&\leq C_4\delta^2,
\end{aligned}
$$

where the second inequality follows from conditional Jansen inequality, and the third inequality follows from the fact that $|(\xi(W,\beta_1(\tau_1),\tau_1)-\xi(W,\beta(\tau),\tau))\cdot(Y-X^{\mathsf{T}}\beta_1(\tau_1))|\leq\big[|1(Y<X^{\mathsf{T}}\beta(\tau))-1(Y<X^{\mathsf{T}}\beta_1(\tau_1))|+|\tau_1-\tau|\big]\cdot|Y-X^{\mathsf{T}}\beta_1(\tau_1)|\leq|X^{\mathsf{T}}(\beta_1(\tau_1)-\beta(\tau))|+|\tau_1-\tau|\cdot|Y-X^{\mathsf{T}}\beta_1(\tau_1)|$.

Apply Theorem 3 of Chen et al. (2003) and a slight modification of Van Der Vaart and Wellner (1996, Theorem 2.7.11) to obtain $J_{[\cdot]}(\delta_n,\mathcal{H},\|\cdot\|_2)\to 0$ for every $\delta_n\downarrow 0$. It follows from Theorem 19.28 of Van Der Vaart (1998) that the process $V_n(\gamma)$ is ρ-stochastically equicontinuous, which implies that

$$
\begin{aligned}
\sup_{\tau\in\mathcal{T}}\bigg|\widehat{R}_n(\tau)-\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(W_i,\beta(\tau),\tau)M(X_i)+\frac{1}{\sqrt{n}}\\
\times\sum_{i=1}^n\big(\mathbb{E}[\psi(W_i,\beta(\tau),\tau)|X_i]-\mathbb{E}[\psi(W_i,\widehat{\beta}_n(\tau),\tau)|X_i]\big)M(X_i)\bigg|=o_p(1).
\end{aligned}
$$

Note that $\int_{-\infty}^{\alpha}(y-\alpha)f_Y(y|x)dy$ is continuously differential in α, and the derivative is $-\int_{-\infty}^{\alpha}f_Y(y|x)dy$. Applying the mean-value theorem, we have

$$
\begin{aligned}
o_p(1)&=\sup_{\tau\in\mathcal{T}}\bigg|\frac{1}{\sqrt{n}}\sum_{i=1}^n\big(\mathbb{E}[\psi(W_i,\beta(\tau),\tau)|X_i]-\mathbb{E}[\psi(W_i,\widehat{\beta}_n(\tau),\tau)|X_i]\big)M(X_i)\\
&\qquad-\frac{1}{n}\sum_{i=1}^n\big[(\tau-(1-2\tau)F_Y(X_i^{\mathsf{T}}\tilde{\beta}_n(\tau)|X_i))M(X_i)X_i^{\mathsf{T}}\big]\times\sqrt{n}(\widehat{\beta}_n(\tau)-\beta(\tau))\bigg|\\
&=\sup_{\tau\in\mathcal{T}}\bigg|\frac{1}{\sqrt{n}}\sum_{i=1}^n\big(\mathbb{E}[\psi(W_i,\beta(\tau),\tau)|X_i]-\mathbb{E}[\psi(W_i,\widehat{\beta}_n(\tau),\tau)|X_i]\big)M(X_i)\\
&\qquad-\frac{1}{n}\sum_{i=1}^n\big[\tau-(1-2\tau)F_Y(X_i^{\mathsf{T}}\beta(\tau)|X_i)\big]M(X_i)X_i^{\mathsf{T}}\times\sqrt{n}(\widehat{\beta}_n(\tau)-\beta(\tau))\bigg|+o_p(1)
\end{aligned}
$$

$$= \sup_{\tau\in\mathcal{T}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\mathbb{E}[\psi(W_i, \beta(\tau), \tau)|X_i] - \mathbb{E}[\psi(W_i, \widehat{\beta}_n(\tau), \tau)|X_i] \big) M(X_i) \right.$$
$$\left. - G(\beta(\tau), \tau)\sqrt{n}(\widehat{\beta}_n(\tau) - \beta(\tau)) + o_p(1), \right.$$

where $\tilde{\beta}_n(\cdot)$ is such that $\left|\tilde{\beta}_n(\tau) - \beta(\tau)\right| \leq \left|\widehat{\beta}_n(\tau) - \beta(\tau)\right|$ *a.s.* for each $\tau \in \mathcal{T}$. The second equality is from the uniform convergence of $\widehat{\beta}_n$ and Assumption 1, and the last equality follows from the Glivenko–Cantelli Theorem, since $\tau \to [\tau - (1-2\tau)F_Y(X^\intercal\beta(\tau)|X)]M(X)X^\intercal$ for $\tau \in \mathcal{T}$ is Glivenko–Cantelli.

Combining the above results, we obtain

$$\sup_{\tau\in\mathcal{T}} \left| \widehat{R}_n(\tau) - R_n(\tau) + G(\beta(\tau), \tau)n^{1/2}(\widehat{\beta}_n(\tau) - \beta(\tau)) \right| = o_p(1).$$

Besides, it is shown in the proof of Theorem 3,

$$Q(\tau) \equiv \sqrt{n}(\widehat{\beta}_n(\tau) - \beta(\tau)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l(W_i, \beta(\tau), \tau) + o_p(1),$$

where $l(W_i, \beta(\tau), \tau) = J(\tau)^{-1}\psi(W_i, \beta(\tau), \tau)X_i$. A combination of the above results yields (14), and the theorem follows immediately. □

Proof of Theorem 6. We have the following decomposition

$$\widehat{R}^*_n(\tau) = \widehat{S}^*_n(\tau) - \widehat{G}_n(\widehat{\beta}_n(\tau), \tau)\widehat{J}_n(\tau)^{-1}\widehat{S}^*_{1n}(\tau),$$

where

$$\widehat{S}^*_n(\tau) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \widehat{\beta}_n(\tau), \tau)M(X_i)V_i,$$
$$\widehat{S}^*_{1n}(\tau) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \widehat{\beta}_n(\tau), \tau)X_iV_i.$$

The class of functions $\{(W, V) \to \psi(W, \beta(\tau), \tau)M(X)V : \beta \in \mathscr{B}, \tau \in \mathcal{T}\}$ and $\{(W, V) \to \psi(W, \beta(\tau), \tau)XV : \beta \in \mathscr{B}, \tau \in \mathcal{T}\}$ are Donsker, as can be shown by applying similar arguments as in Theorem 5. We obtain from a stochastic equicontinuity argument and the uniform consistency of $\widehat{\beta}_n$ that uniformly in $\tau \in \mathcal{T}$,

$$\widehat{S}^*_n(\tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \beta(\tau), \tau)M(X_i)V_i + o_p(1),$$

and

$$\widehat{S}^*_{1n}(\tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \beta(\tau), \tau) X_i V_i + o_p(1).$$

Also, we have $\sup_{\tau \in \mathcal{T}} \|\widehat{G}_n(\widehat{\beta}_n(\tau), \tau) - G(\beta(\tau), \tau)\| = o_p(1)$ and $\sup_{\tau \in \mathcal{T}} \|\widehat{J}_n^{-1}(\tau) - J^{-1}(\tau)\| = o_p(1)$. Thus

$$\widehat{R}^*_n(\tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left[\psi(W_i, \beta(\tau), \tau) M(X_i) - G(\beta(\tau), \tau) l(W_i, \beta(\tau), \tau)\right] + o_p(1).$$

The theorem follows from the multiplier central limit theorem, see, e.g., Theorem 2.9.6 of Van Der Vaart and Wellner (1996), and the continuous mapping theorem. □

References

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006), 'Quantile regression under misspecification, with an application to the U.S. wage structure', *Econometrica* **74**(2), 539–563.

Bellini, F. and Bignozzi, V. (2015), 'On elicitable risk measures', *Quantitative Finance* **15**(5), 725–733.

Bonaccolto-Töpfer, M. and Bonaccolto, G. (2023), 'Gender wage inequality: new evidence from penalized expectile regression', *Journal of Economic Inequality* **21**(3), 511–535.

Buchinsky, M. (1994), 'Changes in the U.S. wage structure 1963-1987: Application of quantile regression', *Econometrica* **62**(2), 405–458.

Chen, X., Linton, O. and Van Keilegom, I. (2003), 'Estimation of semiparametric models when the criterion function is not smooth', *Econometrica* **71**(5), 1591–1608.

Dawber, J., Salvati, N., Fabrizi, E. and Tzavidis, N. (2022), 'Expectile regression for multicategory outcomes with application to small area estimation of labour force participation', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **185**(S2), S590–S619.

Dudley, R. M. (2014), *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, 2 edn, Cambridge University Press.

Durbin, J. (1973), 'Weak convergence of the sample distribution function when parameters are estimated', *The Annals of Statistics* **1**(2), 279 – 290.

Efron, B. (1991), 'Regression percentiles using asymmetric squared error loss', *Statistica Sinica* **1**(1), 93–125.

Guler, K., Ng, P. T. and Xiao, Z. (2017), 'Mincer–Zarnowitz quantile and expectile regressions for forecast evaluations under aysmmetric loss functions', *Journal of Forecasting* **36**(6), 651–679.

He, X. and Zhu, L.-X. (2003), 'A lack-of-fit test for quantile regression', *Journal of the American Statistical Association* **98**(464), 1013–1022.

Kim, M. and Lee, S. (2016), 'Nonlinear expectile regression with application to value-at-risk and expected shortfall estimation', *Computational Statistics & Data Analysis* **94**, 1–19.

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

Kuan, C.-M., Yeh, J.-H. and Hsu, Y.-C. (2009), 'Assessing value at risk with CARE, the Conditional Autoregressive Expectile models', *Journal of Econometrics* **150**(2), 261–270. Recent Development in Financial Econometrics.

Mammen, E. (1993), 'Bootstrap and wild bootstrap for high dimensional linear models', *The Annals of Statistics* **21**(1), 255–285.

Martins, P. S. and Pereira, P. T. (2004), 'Does education reduce wage inequality? Quantile regression evidence from 16 countries', *Labour Economics* **11**(3), 355–371.

Newey, W. K. and Powell, J. L. (1987), 'Asymmetric least-squares estimation and testing', *Econometrica* **55**(4), 819–847.

Philipps, C. (2022), Interpreting Expectiles, Working Papers 2022-01.

Van Der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Van Der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Xu, W., Hou, Y. and Li, D. (2022), 'Prediction of extremal expectile based on regression models with heteroscedastic extremes', *Journal of Business & Economic Statistics* **40**(2), 522–536.

Zheng, J. X. (1998), 'A consistent nonparametric test of parametric regression models under conditional quantile restrictions', *Econometric Theory* **14**(1), 123–138.