Spady, Richard H.; Stouli, Sami

Working Paper

Dual regression

cemmap working paper, No. CWP04/16

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Spady, Richard H.; Stouli, Sami (2015) : Dual regression, cemmap working paper, No. CWP04/16, Centre for Microdata Methods and Practice (cemmap), London

This Version is available at:
http://hdl.handle.net/10419/130091

cemmap
centre for microdata methods and practice

Dual regression

Richard H. Spady
Sami Stouli

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP04/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

DUAL REGRESSION

RICHARD H. SPADY AND SAMI STOULI

Abstract. We propose an alternative ('dual regression') to the quantile regression process for the global estimation of conditional distribution functions under minimal assumptions. Dual regression provides all the interpretational power of the quantile regression process while largely avoiding the need for 'rearrangement' to repair the intersecting conditional quantile surfaces that quantile regression often produces in practice. Our approach relies on a mathematical programming characterization of conditional distribution functions which, in its simplest form, provides a simultaneous estimator of location and scale parameters in a linear heteroscedastic model. The statistical properties of this estimator are derived.

1. Introduction

Let Y be a random variable with continuous support and X a random vector. Then the conditional distribution function of Y given X, written $U = F_{Y|X}(Y \mid X)$, has three properties: (1) U is standard uniform, (2) U is independent of X, and (3) $F_{Y|X}(Y = y \mid X = x)$ is strictly increasing in y for any value x of X. We will refer to these three properties as "uniformity", "independence" and "monotonicity".

Supposing that we have a sample of n points $\{(x_i, y_i)\}_{i=1}^n$ drawn from the joint distribution $F_{YX}(Y, X)$, how might we estimate the n values $u_i = F_{Y|X}(Y = y_i \mid X = x_i)$ using only the requirement that the estimate displays uniformity, independence, and monotonicity? We explore this question by formulating a sequence of mathematical programming problems that embodies these requirements and that generalizes the dual formulation of the quantile regression problem.

Date: This version: November 5, 2015. First Arxiv version October 25, 2012. Acknowledgements: We are indebted to Andrew Chesher for many fruitful discussions on the topic of this paper, and to Roger Koenker for encouragement at a key early stage. We also thank Jelmer Ypma for his help with Ipoptr, and Dennis Kristensen, David Pacini, Yanos Zylberberg and seminar participants at Oxford, the Workshop on Semiparametric Econometrics in Essex and the Bristol Econometric Study Group for helpful comments. Sami Stouli gratefully acknowledges the financial support of the UK Economic and Social Research Council and of the Royal Economic Society. Authors' affiliations: Richard Spady: Department of Economics, Johns Hopkins University and CeMMAP. Sami Stouli: Department of Economics, University of Bristol.

The use of 'dual' is thus motivated by the general observation that the estimation problem for a conditional distribution function $F_{Y|X}$ indexed by a parameter θ is usually formulated in terms of a procedure that obtains θ directly and $F_{Y|X}$ as a byproduct that follows from a calculation from the representation evaluated at a specific value of θ. Two leading examples are the linear location shift model $F_{Y|X}(Y = y_i \mid X = x_i) = F\{(y_i - \beta \cdot x_i)/\sigma\}$, for some distribution function F, and the linear quantile regression model $F_{Y|X}(Y = y_i \mid X = x_i) = \int_0^1 1\{\beta(u) \cdot x_i \leq y_i\}du$, for which the parameters $\theta = (\beta, \sigma)^{\mathrm{T}}$ and $\theta(u) = \beta(u)$, $u \in (0,1)$, respectively, need to be estimated in order to obtain the n values u_i. Here we turn that process around, obtaining u_i, or a monotone transformation of u_i, first (from e.g. a mathematical programming problem) and 'backing out' θ afterwards, if at all.

Although this 'generalized dual' formulation seeks only to find the n values $u_i = F_{Y|X}(Y = y_i \mid X = x_i)$, *its dual* - the primal, so to speak - shows that the assignment of these n values admits a sequence of location-scale representations, the simplest element of which is a linear heteroscedastic model. For the class of location-scale distributions, this simplest representation, like the quantile regression process, provides a complete estimate of $F_{Y|X}(Y \mid X)$. Moreover, it is largely free of 'quantile-crossing' problems that the quantile regression process sometimes encounters in practice.

In its simplest form, dual regression augments the conventional median regression dual programming problem (Koenker & Bassett (1978)) with global second moment orthogonality constraints, and delivers a family of globally monotone location-scale representations, thereby providing a simultaneous estimator of the location and scale parameters of a linear heteroscedastic model. Adding further global orthogonality constraints gives rise to more flexible, generalized dual regression representations, which we introduce after having established the computational and statistical properties of our basic method.

2. **Basics**

2.1. **The dual regression problem.** The dual problem of the (linear) $0 \cdot 5$ quantile regression of y on X is (cf. Koenker (2005) p. 87, equation 3.12):

$$\max_u \{y^{\mathrm{T}}u \mid X^{\mathrm{T}}(u - \frac{1_n}{2}) = 0,\ u \in [0,1]^n\}, \tag{2.1}$$

where y is an $(n \times 1)$ vector of dependent variable values, X is an $(n \times k)$ matrix of explanatory variable values that includes an intercept, i.e an $(n \times 1)$ vector of ones denoted 1_n.

The solution to problem (2.1) produces values of u that are largely 0 and 1, with k sample points being assigned u values that are neither 0 nor 1. The points that are assigned 1 fall above the median quantile regression; the points receiving 0's fall below; and the remaining points fall on the median quantile regression plane. One direction of extension of equation (1) is to replace the "1/2" with values α that fall between 0 and 1 to obtain the α quantile regression.

Another extension is to augment problem (2.1) by adding k more constraints:

$$\max_u \{y^{\mathrm{T}}u \mid \begin{cases} X^{\mathrm{T}}(u - \frac{1_n}{2}) & = 0 \\ X^{\mathrm{T}}(u^2 - \frac{1_n}{3}) & = 0 \end{cases}, \ u \in [0,1]^n\}. \tag{2.2}$$

Apparently the solution to (2.1) does not satisfy (2.2): the variance of u (around 0) in the solution to (2.1) is approximately 1/2, not 1/3. To satisfy program (2.2), the u's have to be moved off $\{0\}$,$\{1\}$. Since X contains an intercept, the sample moments of u and u^2 will be 1/2 and 1/3; u and u^2 will be orthogonal to the components of X, relations that are necessary but not sufficient for uniformity and independence.

Both systems (2.1) and (2.2) impose monotonicity by maximally correlating y and u. It is worth noting that a violation of monotonicity requires there to be two observations that share the same X values but have different y values, with the lower of the two y values having the (weakly) higher value of u. But a 'solution' characterized by such a violation could be improved upon by exchanging the u assignments. In program (2.1), however, the set of admissible exchanges in u assignments is overly restricted by the fact that program (2.1) is dual to a linear program well-known to have solutions at which k observations are interpolated when k parameters are being estimated, i.e. the hyperplanes obtained by regression quantiles must interpolate k observations.

Some simplification (particularly in computation) is obtained by reformulating problem (2.2) into a constrained optimization problem over $\mathbb{R}^n$. This is feasible since the goal of assigning a value u_i to each observation such that uniformity, independence, and monotonicity is achieved could equally well be achieved by assigning a value $e_i \in \mathbb{R}$ to each observation, where $e = (e_1, \ldots, e_n)^{\mathrm{T}}$ obeys the independence and monotonicity requirements, but where e_i is given by $F^{-1}(u_i)$ for some distribution function F. Such a e solution is transformed into a corresponding u solution by taking $u_i = F(e_i)$; without loss of generality we can take F to correspond to a distribution with zero mean and unit variance. Doing this, the

problem corresponding to (2.2) becomes:

$$\max_{e\in\mathbb{R}^n}\{y^{\mathrm{T}}e \mid \begin{cases} X^{\mathrm{T}}e & = 0 \\ \frac{1}{2}X^{\mathrm{T}}(e^2 - 1_n) & = 0 \end{cases}, \tag{2.3}$$

where e can take on any real value (whereas u is restricted to $[0,1]^n$). It is then natural to take $u_i = F_n(e_i)$, the empirical cumulative distribution function of e, thereby imposing uniformity to high precision even at small n.

2.2. **Solving the dual problem.** The solution to the problem in equation (2.3) is easily found from the Lagrangian

$$\mathscr{L} = \sum_{i=1}^{n} y_i e_i - \lambda_1 \sum_{i=1}^{n} x_i e_i - \frac{1}{2}\lambda_2 \sum_{i=1}^{n} x_i(e_i^2 - 1).$$

Differentiating with respect to e_i, we obtain n first-order conditions:

$$\frac{\partial \mathscr{L}}{\partial e_i} = y_i - \lambda_1 \cdot x_i - (\lambda_2 \cdot x_i)e_i = 0.$$

Keeping in mind that x_i is a k component vector (and thus so are the Lagrange multipliers λ_1 and λ_2) we obtain for each e_i:

$$e_i = \frac{y_i - \lambda_1 \cdot x_i}{\lambda_2 \cdot x_i}, \tag{2.4}$$

which is of the familiar location-scale form $e_i = \{y_i - \mu(x_i)\}/\sigma(x_i)$ with the functions $\mu(x_i)$ and $\sigma(x_i)$ being linear in x_i.

Another view is obtained by writing the first-order conditions as

$$y_i = \lambda_1 \cdot x_i + (\lambda_2 \cdot x_i)e_i, \tag{2.5}$$

a linear location-scale representation for Y given X, with corresponding quantile regression representation

$$\begin{aligned} y_i &= (\lambda_1 + \lambda_2 e_i) \cdot x_i \\ &= \{\lambda_1 + \lambda_2 F_n^{-1}(u_i)\} \cdot x_i \\ &\equiv \beta(u_i) \cdot x_i. \end{aligned} \tag{2.6}$$

Thus, the dual regression program (2.3) provides a complete characterization of linear representations (2.5) and (2.6), as they arise from its first-order conditions.

The quantile regression representation of the first-order conditions of (2.3) sheds additional light on the monotonicity property of dual regression solutions, when there are no repeated X values. For $u, u' \in (0,1)$, $u' > u$, the no crossing property of conditional quantiles requires that

$$\beta(u') \cdot x_i - \beta(u) \cdot x_i > 0 \quad (i = 1, \ldots, n).$$

Replacing $\beta(u)$ by its expression in (2.6), the condition is

$$\lambda_2 \{F_n^{-1}(u') - F_n^{-1}(u)\} \cdot x_i > 0 \quad (i = 1, \ldots, n),$$

which holds under the simple condition that $\lambda_2 \cdot x_i$ be strictly positive for each i, and coincides with the n second-order conditions of system (2.3):

$$\frac{\partial^2 \mathscr{L}}{\partial e_i \partial e_i} = -(\lambda_2 \cdot x_i) < 0 \quad (i = 1, \ldots, n).$$

Therefore, an optimal e solution that violates the monotonicity property is ruled out by the requirement that for an observation with X value x_i, the ordering of the counterfactual Y values $\beta(u') \cdot x_i$ and $\beta(u) \cdot x_i$ must correspond to the ordering of the u values. Hence the correlation criterion of system (2.3) suffices to impose monotonicity.

2.3. **Formal duality.** By Lagrangian duality arguments (e.g. Boyd & Vandenberghe (2004), Chapter 5), the objective function of the dual of problem (2.3), the primal dual regression objective $Q_n(\lambda)$, is the convex conjugate, or Legendre transform, of the function $\sum_{i=1}^n C(x_i, e_i, \lambda)$:

$$Q_n(\lambda) = \sup_{e \in \mathbb{R}^n} \left\{ y^{\mathrm{T}} e - \sum_{i=1}^n C(x_i, e_i, \lambda) \right\},$$

where

$$C(x_i, e_i, \lambda) = (\lambda_1 \cdot x_i) e_i + \frac{1}{2} (\lambda_2 \cdot x_i)(e_i^2 - 1).$$

When $\lambda_2 \cdot x_i > 0$, $i = 1, \ldots, n$, $C(x_i, e_i, \lambda)$ is itself a convex function that represents $F_{Y|X}(y_i \mid x_i)$ once a distribution for e_i is given. Its Legendre transform therefore contains the same information. We further define the domain of $Q_n(\lambda)$ as $\Lambda_0 = \Lambda_1 \times \Lambda_2$, with $\Lambda_1 = \mathbb{R}^k$ and $\Lambda_2 = \{\lambda_2 \in \mathbb{R}^k : \ \lambda_2 \cdot x_i > 0, \ i = 1, \ldots, n\}$. Under Conditions 1 and 2 below, $Q_n(\lambda)$ is strictly convex over Λ_0 (cf Lemma 5 in the Appendix), and minimizing $Q_n(\lambda)$ over Λ_0 is equivalent to solving (2.3).

For $\lambda_2 \in \Lambda_2$, let $\Omega_n = \mathrm{diag}(\lambda_2 \cdot x_i)$, an $n \times n$ diagonal matrix with diagonal elements $\lambda_2 \cdot x_i$, $i = 1, \ldots, n$.

Condition 1. Y is continuously distributed conditional on X, with conditional density $f_{Y|X}(y \mid X)$ bounded away from 0.

Condition 2. For all $\lambda_2 \in \Lambda_2$, $X^{\mathrm{T}}\Omega_n^{-1}X = M_n$, a finite positive definite matrix of rank k.

Let $\lambda_n = (\lambda_{1n}, \lambda_{2n})^{\mathrm{T}}$ be a minimizer of $Q_n(\lambda)$ over Λ_0, and e^* a feasible solution to program (2.3). For clarity we also denote by λ^* the value of the Lagrange multiplier vector of program (2.3), corresponding to a solution e^*. Theorem 1 summarizes our results on formal duality; its proof is given in the Appendix.

Theorem 1. *Suppose that Conditions 1 and 2 hold. Then, for the dual regression problem*

$$(D) \qquad \max_{e\in\mathbb{R}^n} \{y^T e \mid \begin{cases} X^T e & = 0 \\ \frac{1}{2}X^T(e^2 - 1_n) & = 0 \end{cases},$$

the following holds:

(i) (Primal problem) The dual of Problem (D) is

$$(P) \qquad \min_{\lambda\in\Lambda_0} \sum_{i=1}^{n} \frac{1}{2}\left\{\left(\frac{y_i - \lambda_1 \cdot x_i}{\lambda_2 \cdot x_i}\right)^2 + 1\right\}(\lambda_2 \cdot x_i),$$

the primal dual regression problem.

(ii) (First-order conditions) Problem (D) admits the Method-of-Moments representation

$$\begin{aligned} X^T e &= 0 \\ \frac{1}{2}X^T(e^2 - 1_n) &= 0 \\ e_i &= \frac{y_i - \lambda_1 \cdot x_i}{\lambda_2 \cdot x_i} \quad (i = 1, \ldots, n), \end{aligned} \tag{2.7}$$

the first-order conditions of (P).

(iii) (a) (Uniqueness) The pair (λ_n, e^*) uniquely solves the primal and dual problems (P) and (D), and $\lambda_n = \lambda^*$; (b) (Strong duality) the value of (D) equals the value of (P).

Theorem 1 establishes formal duality of our initial assignment problem under orthogonality constraints and the global M-estimation problem (P). To a unique assignment of e values corresponds a unique linear representation of the form (2.5). The primal problem (P) of Theorem 1 is a locally heteroscedastic generalization of a simultaneous location-scale estimator proposed by Huber (1981) and further analyzed in Owen (2001). The linear heteroscedastic

model of equation (2.5) has been previously encountered in the quantile regression literature: see Koenker & Zhao (1994); He (1997). The former considers the efficient estimation of (2.5) via L-estimation while the latter develops a restricted quantile regression method that prevents quantile crossing. Compared to these quantile-based methods, dual regression trades local estimation and the convenient linear programming formulation of quantile regression for global estimation of location and scale parameters.

2.4. **Existence, structural interpretation and statistical properties.** If the data generating process (DGP) is of the linear heteroscedastic form

$$\begin{aligned} y_i &= \beta_1 \cdot x_i + (\beta_2 \cdot x_i)\varepsilon_i, \quad \beta_2 \cdot x_i > 0 \quad (i = 1, \ldots, n) \\ E(\varepsilon_i \mid x_i) = 0 \quad &\text{and} \quad E(\varepsilon_i^2 \mid x_i) = 1, \end{aligned}$$

then there exists a solution to the dual regression problem for n sufficiently large, it is unique, and dual regression consistently estimates the parameter vector $\beta = (\beta_1, \beta_2)^{\mathrm{T}}$. Formally, with $\mathcal{X}$ denoting the support of X, we impose the following conditions:

Condition 3. (i) $\{(y_i, x_i)\}_{i=1}^n$ are i.i.d.; (ii) $E(Y^2)$, $E\|X\|^4$ and $E(Y^2\|X\|^2)$ are finite; (iii) there exists a constant C such that $\inf_{x\in\mathcal{X}} \mathrm{var}(Y \mid X = x)^{1/2} \geq 1/C > 0$; (iv) for $\beta \in \Lambda_0$, $E(Y \mid X) = \beta_1 \cdot X$ and $\mathrm{var}(Y \mid X)^{1/2} = \beta_2 \cdot X$.

Condition 4. For all $\lambda_2 \in \Lambda_2$, $\lim n^{-1} M_n = M$, a finite positive definite matrix of rank k.

Condition 5. $E(Y^4)$, $E\|X\|^6$ and $E(Y^4\|X\|^2)$ are finite.

Together Conditions 1-4 are sufficient conditions for existence and consistency, whereas the additional Condition 5 is needed for asymptotic normality of dual regression estimates of β. In view of part (iii) of Theorem 1, these properties are shared by λ_n and λ^*, which we denote by $\hat{\lambda}$ for notational simplicity. Similarly, given the functional form of a solution e^*, we denote both e^* and the vector of indirect estimates $(y_i - \lambda_{1n} \cdot x_i)/(\lambda_{2n} \cdot x_i)$, $i = 1, \ldots, n$, constructed after solving (P), by $\hat{e}$, with empirical distribution function $F_n(e) = n^{-1}\sum_{i=1}^n 1(\hat{e}_i \leq e)$, $e \in \mathbb{R}$. Furthermore, part (ii) of Theorem 1 shows that while the solution e^* can be obtained directly by solving the mathematical program (D), knowledge that the solution obeys equation (2.4) can be exploited to write estimating equations for $\hat{\lambda}$ in the form of system (2.7). The computation of the asymptotic distribution of $\hat{\lambda}$ follows from this characterization.

For $e(y_i, x_i, \lambda) = (y_i - \lambda_1 \cdot x_i)/(\lambda_2 \cdot x_i)$, define $m_1(y_i, x_i, \lambda) = x_i e(y_i, x_i, \lambda)$, $m_2(y_i, x_i, \lambda) = x_i\{e(y_i, x_i, \lambda)^2 - 1\}/2$, and $m(y_i, x_i, \lambda) = (m_1(y_i, x_i, \lambda), m_2(y_i, x_i, \lambda))^{\mathrm{T}}$, and let $G = E\{\partial m(y_i, x_i, \lambda)/\partial\lambda\}|_{\lambda=\beta}$ and $S = E\{m(y_i, x_i, \beta)m(y_i, x_i, \beta)^{\mathrm{T}}\}$. The proof of Theorem 2 is given in the Supplementary Material.

Theorem 2. *If Conditions 1-5 hold, then: (i) there exists $\hat{\lambda}$ in Λ_0 with probability approaching one, (ii) $\hat{\lambda} \to_p \beta$, and (iii) $n^{1/2}(\hat{\lambda} - \beta) \to_d N(0, G^{-1}SG^{-1})$.*

Knowledge of the statistical properties of $\hat{\lambda}$ can be used to establish the limiting behaviour of the empirical distribution of $\hat{e}$. For F_ε the cumulative distribution function of the random variable $\varepsilon_i = e(y_i, x_i, \beta)$, we define the empirical dual regression process $\mathbb{U}_n(\cdot)$:

$$\mathbb{U}_n(e) = n^{1/2}\{F_n(e) - F_\varepsilon(e)\}, \quad e \in \mathbb{R}.$$

With $g(e) = E[f_{Y|X}\{(\beta_1 + \beta_2 e) \cdot x_i \mid x_i\}(x_i, x_i e)^{\mathrm{T}}]$, Theorem 3 establishes weak convergence of the empirical distribution of $\hat{e}$ and the limiting behaviour of $\mathbb{U}_n(\cdot)$, accounting for its dependence on the distribution of $n^{1/2}(\hat{\lambda} - \beta)$; the proof is given in the Supplementary Material.

Theorem 3. *Suppose that Conditions 1-5 hold, and further assume that uniformly in x over $\mathcal{X}$, $f_{Y|X}(y|x)$ is uniformly continuous in y, bounded and satisifies $\sup_{y\in\mathbb{R}} |y| f_{Y|X}(y|x) < \infty$. Then: the empirical dual regression process $\mathbb{U}_n(\cdot)$ converges weakly to a zero-mean Gaussian process $\mathbb{U}(\cdot)$ with covariance function $E\{\varphi_e(y_i, x_i, \beta)\varphi_{e'}(y_i, x_i, \beta)\}$, where*

$$\varphi_e(y_i, x_i, \beta) = 1\{e(y_i, x_i, \beta) \le e\} - F_\varepsilon(e) - g(e)^\top G^{-1} m(y_i, x_i, \beta).$$

Theorem 3 establishes that the empirical distribution of dual regression estimates is a consistent estimator of the distribution of ε_i. If in addition ε_i is independent of x_i, then $F_n(\hat{e}_i)$ consistently estimates $F_{Y|X}(y_i \mid x_i)$. Similarly, an estimate of the quantile regression coefficient vector $\beta(u)$ can be constructed as $\hat{\lambda}_1 + \hat{\lambda}_2 F_n^{-1}(u)$, exploiting the location-scale structure of the conditional u-quantile function of y_i given x_i, and independence of ε_i and x_i.

The third term in the expression for φ_e reflects the influence of imposing sample orthogonality constraints in (D) on the empirical distribution of e^*, or equivalently, of sample variability of parameter estimates λ_n on the empirical distribution of $e(y_i, x_i, \lambda_n)$, as expected from the classical result of Durbin (1973). Methods for inference on parametric empirical processes (see e.g. Koenker & Xiao (2002), Parker (2013)) provide a natural direction for future study of inference on the empirical dual regression process.

3. **Generalization**

3.1. **Framework.** Let $\mathcal{E} = [0,1]$ or $\mathbb{R}$, and for $j = 1, \ldots, J$, let $h_j : \mathcal{E} \to \mathcal{E}$ be a continuously differentiable function and define, for all $e \in \mathcal{E}$, $\widetilde{h}_j(e) = \int_{-\infty}^{e} h_j(s)ds$ and $m_j(e) = \widetilde{h}_j(e) - c_j$, an antiderivative of h_j, for some $c_j \in \mathbb{R}$.

The two approaches in systems (2.2) and (2.3) share the common structure

$$\max_{e \in \mathcal{E}^n} y^{\mathrm{T}} e \quad s.t. \ \sum_{i=1}^{n} x_{ij} m_j(e_i) = 0 \qquad (j = 1, \ldots, J), \tag{3.1}$$

which gives rise to the first-order conditions

$$\begin{aligned} y_i &= \sum_{j=1}^{J} (\lambda_j \cdot x_{ij}) h_j(e_i) \qquad (i = 1, \ldots, n) \\ \sum_{i=1}^{n} x_{ij} m_j(e_i) &= 0 \qquad (j = 1, \ldots, J). \end{aligned} \tag{3.2}$$

For $J = 2$ and $\widetilde{h}_1(e_i) = e_i$, $\widetilde{h}_2(e_i) = e_i^2/2$, systems (2.2) and (2.3) have the above structure with (c_1, c_2) set to $(1/2, 1/6)$ and $(0, 1/2)$, and $\mathcal{E}^n = [0,1]^n$ and $\mathbb{R}^n$, respectively.

The dual regression characterization of conditional distribution functions via the monotonicity element (the objective) and the independence element (the constraints) can be exploited to generate flexible representations for Y given X: for $J > 2$, Equation (3.2) already suggests a representation of Y conditional on X which is more flexible than a location-scale specification. The object of this section is to further analyze and specify conditions under which (3.1) can serve the purpose of characterizing flexible representations for Y given X.

Given a random sample $\{(y_i, x_i)\}_{i=1}^n$, suppose that the stochastic structure of Y given X can be represented as

$$y_i = H(x_i, e_{oi}) \equiv H_{x_i}(e_{oi}) \qquad e_{oi} | x_i \sim F, \tag{3.3}$$

for some cumulative distribution function F with support the real line, and where, for each x_i, $H_{x_i}(e_{oi})$ is strictly increasing in e_{oi}. To the monotone function H_{x_i} also corresponds a convex function $\widetilde{H}_{x_i}$ defined as

$$\widetilde{H}_{x_i}(e_{oi}) \equiv \int_{-\infty}^{e_{oi}} H_{x_i}(s)ds, \quad e_{oi} \in \mathbb{R}. \tag{3.4}$$

The monotonicity of $H_{x_i}(e_{oi})$ guarantees the convexity of $\widetilde{H}_{x_i}(e_{oi})$. At each value x_i, $\widetilde{H}_{x_i}(e_{oi})$ is a convex function of e_{oi}; the slope of this function gives the value of Y corresponding to the

value e_{oi} at $X = x_i$. Thus $F_{Y|X}(Y \mid X)$ corresponds to a collection of convex functions, with one element of this collection for each value of X, together with a single random variable whose distribution is common to all the convex functions: given one random variable e_{oi} with a particular distribution F, we can always monotonically transform it to another random variable and similarly transform the functions $\widetilde{H}_{x_i}$ so as to leave $F_{Y|X}(Y \mid X)$ unchanged.

3.2. **Infeasible generalized dual regression.** Equipped with $\widetilde{H}_{x_i}$, suppose we are tasked with assigning a value e_i to each observation in our sample $\{(y_i, x_i)\}_{i=1}^n$. Then, for $S_n = \sum_{i=1}^n \widetilde{H}_{x_i}(e_{oi})$, solving the (infeasible) optimization problem:

$$\max_{e \in \mathbb{R}^n} \; y^{\mathrm{T}} e \quad s.t. \; \sum_{i=1}^n \widetilde{H}_{x_i}(e_i) = S_n, \tag{3.5}$$

generates the correct '$y - e$' assignment, as established in Theorem 4 below; the constraint in (3.5) imposes that e_i be independent of x_i and specifies $e_i \sim F$, whereas the objective imposes monotonicity. Writing the Lagrangian

$$\mathscr{L} = y^{\mathrm{T}} e - \Lambda \left\{ \sum_{i=1}^n \widetilde{H}_{x_i}(e_i) - S_n \right\},$$

the n associated first-order conditions are:

$$\frac{\partial \mathscr{L}}{\partial e_i} = y_i - \Lambda H_{x_i}(e_i) = 0 \qquad (i = 1, \ldots, n). \tag{3.6}$$

Strict convexity of $\widetilde{H}_{x_i}$ then guarantees that $(\Lambda, e) = (1, e_o)$ satisfies these conditions. This demonstrates that maximizing $y^{\mathrm{T}} e$ generally suffices to match e's to y's, regardless of the form of $\widetilde{H}_{x_i}$.

Theorem 4. *Given a random sample $\{(y_i, x_i)\}_{i=1}^n$, suppose that representation (3.3) holds with $H_{x_i} : \mathbb{R} \to \mathbb{R}$ a continuously differentiable, strictly increasing function for each $x_i \in \mathcal{X}$. Then: for $\widetilde{H}_{x_i}(e_{oi}) = \int_{-\infty}^{e_{oi}} H_{x_i}(s)ds$, $(e_{oi}, x_i) \in \mathbb{R} \times \mathcal{X}$, solving the (infeasible) optimization problem (3.5) with $S_n = \sum_{i=1}^n \widetilde{H}_{x_i}(e_{oi})$ generates the correct 'y−e' assignment, i.e. $(\Lambda, e) = (1, e_o)$ uniquely solves the first-order conditions (3.6).*

Theorem 4 shows that problem (3.5) fully characterizes the '$y - e$' assignment problem: given $\widetilde{H}_{x_i}$, solving (3.5) assigns a value e_i to each sample point (y_i, x_i) and this value is the corresponding value $F_{Y|X}(y_i \mid x_i)$ up to a specified transformation F. Knowledge of $\widetilde{H}_{x_i}$ and S_n can thus be incorporated into a mathematical programming problem which delivers the values of $F_{Y|X}$ at the n sample points.

3.3. **Generalized dual regression representations.** Problem (3.5) is infeasible because neither $\widetilde{H}_{x_i}$ nor S_n is known. However, Theorem 4 motivates a feasible approach once H_{x_i} and F are specified. We adopt the following notation: define $x_i = (1, \tilde{x}_i)^T$, and let $x_{ij} = x_i$ if $j = 1, 2$ and $x_{ij} = \tilde{x}_i$ if $j > 2$. If $\tilde{x}_i$ is centered, we write x_i^c, i.e. we define the $(k-1)$-vectors $\bar{x} = n^{-1} \sum_{i=1}^n \tilde{x}_i$ and $x_i^c = \tilde{x}_i - \bar{x}$. Define x_{ij}^c analogously to x_{ij}, substituting x_i^c to $\tilde{x}_i$.

Without loss of generality, let $\tilde{x}_i$ be centered. We specify each of the strictly monotone functions H_{x_i} by a linear combination of J basis functions $\{h_j : j = 1, 2, \ldots, J\}$, such as splines or orthogonal polynomials (DeVore (1977a,b)), the coefficients of which depend on x_i:

$$H_{x_i}(e_{oi}) = \sum_{j=1}^{J} \beta_j(x_i) h_j(e_{oi}), \quad e_{oi} \in \mathbb{R}, \tag{3.7}$$

and we assume that H_{x_i} is linear in x_i and set:

$$\beta_j(x_i) = \gamma_j + \lambda_j \cdot x_i^c, \quad x_i \in \mathcal{X}, \quad (j = 1, \ldots, J). \tag{3.8}$$

Finally, we specify F by letting $h_1(e_{oi}) = 1, h_2(e_{oi}) = e_{oi}$ in (3.7), imposing that $\sum_{i=1}^n e_{oi} = 0$ and $\sum_{i=1}^n (e_{oi}^2 - 1)/2 = 0$, and setting $\gamma_j = 0$ for $j > 2$ in (3.8). This is not the only normalization possible; for instance, F can be fully specified to a known distribution by specifying all basis functions and sample moments of e_{oi} instead (cf problem (3.13) below).

Our normalization and (3.7)-(3.8) together yield the generalized dual regression representation

$$y_i = \gamma_1 + \gamma_2 e_{oi} + (\lambda_1 \cdot x_i^c) + (\lambda_2 \cdot x_i^c) e_{oi} + \sum_{j=3}^{J} (\lambda_j \cdot x_i^c) h_j(e_{oi}). \tag{3.9}$$

Equation (3.9) admits of the following interpretation. When $x_i^c = 0$, $y_i = \gamma_1 + \gamma_2 e_{oi}$ and $e_{oi} = (y_i - \gamma_1)/\gamma_2$, so that e_{oi} is just a re-scaled version of the distribution of y_i at $x_i^c = 0$. Since e_{oi} is independent of x_i, transformations of this 'shape' of e_{oi} must suffice to produce y_i at other values of x_i. The first two transformations - $(\lambda_1 \cdot x_i^c)$ and $(\lambda_2 \cdot x_i^c) e_{oi}$ - are translations of location and scale which do not essentially affect the 'shape' of y_i's response to changes in e_{oi} at all. The additional terms $(\lambda_j \cdot x_i^c) h_j(e_{oi})$ achieve that end.

For $\widetilde{h}_1(e_{oi}) = e_{oi}$, $\widetilde{h}_2(e_{oi}) = e_{oi}^2/2$, applying definition (3.4) to H_{x_i}, the corresponding convex function $\widetilde{H}_{x_i}(e_{oi})$ is

$$\widetilde{H}_{x_i}(e_{oi}) = \int_{-\infty}^{e_{oi}} H_{x_i}(s) ds = \sum_{j=1}^{J} (\theta_j \cdot x_{ij}^c) \widetilde{h}_j(e_{oi}), \quad e_{oi} \in \mathbb{R}, \tag{3.10}$$

where $\theta_j = (\gamma_j, \lambda_j)$ for $j = 1, 2$ and $\theta_j = \lambda_j$ for $j > 2$. Thus, for $\theta = (\theta_1, \ldots, \theta_J)$ and given the form of $\widetilde{H}_{x_i}$, the infeasible generalized dual regression problem becomes

$$\max_{e \in \mathbb{R}^n} \; y^{\mathrm{T}} e \quad s.t. \; \sum_{i=1}^{n} \widetilde{H}_{x_i}(e_i) = S_n.$$

with

$$S_n = \sum_{i=1}^{n} \widetilde{H}_{x_i}(e_{oi}) = \sum_{j=1}^{J} \left\{ \sum_{i=1}^{n} (\theta_j \cdot x_{ij}^c) \cdot \sum_{i=1}^{n} \widetilde{h}_j(e_{oi}) \right\},$$

where the specification of S_n follows from its definition in Theorem 4 and expansion (3.10), and imposes independence of e_{oi} and x_i. Substituting in the expressions for $\widetilde{H}_{x_i}$ and S_n, the Lagrangian for the '$y - e$' assignment problem is

$$\begin{aligned} \mathscr{L} &= y^{\mathrm{T}} e - \Lambda \left\{ \sum_{i=1}^{n} \widetilde{H}_{x_i}(e_i) - S_n \right\} \\ &= y^{\mathrm{T}} e - \Lambda \sum_{i=1}^{n} \left[\sum_{j=1}^{J} (\theta_j \cdot x_{ij}^c) \left\{ \widetilde{h}_j(e_i) - \sum_{i=1}^{n} \widetilde{h}_j(e_{oi}) \right\} \right] \\ &= y^{\mathrm{T}} e - \Lambda \sum_{j=1}^{J} \sum_{i=1}^{n} (\theta_j \cdot x_{ij}^c) m_j(e_i), \end{aligned}$$

with $m_1(e_i) = e_i$, $m_2(e_i) = (e_i^2 - 1)/2$, and $m_j(e_i) = \tilde{h}_j(e_i)$, for $j > 2$, since for $c_j = \sum_{i=1}^{n} \widetilde{h}_j(e_{oi})$ the centering of x_i^c implies $\sum_{i=1}^{n} \{(\theta_j \cdot x_{ij}^c) c_j\} = 0$ for $j > 2$.

Recalling from Theorem 4 that $\Lambda = 1$, if we add θ to the choice variables of the optimization problem, we obtain the $\dim(\theta)$ additional constraints

$$\frac{\partial \mathscr{L}}{\partial \theta_j} = -\sum_{i=1}^{n} x_{ij}^c m_j(e_i) \quad = \quad 0 \quad (j = 1, \ldots, J). \tag{3.11}$$

Equation (3.11) can be directly appended to the objective $\max_e \; y^{\mathrm{T}} e$ to obtain an optimization problem in which the Lagrange multiplier is θ:

$$\max_{e \in \mathbb{R}^n} \; y^{\mathrm{T}} e \quad s.t. \; \sum_{i=1}^{n} x_{ij}^c m_j(e_i) = 0 \quad (j = 1, \ldots, J). \tag{3.12}$$

Problem (3.12) gives a feasible formulation of the generalized '$y - e$' assignment problem. When the conditional quantile function of Y given X is linear in X, representation (3.9) determines the form of $\widetilde{H}_{x_i}$, a linear combination of basis functions generating an increasing

sequence of orthogonality conditions which, as $J \to \infty$, impose full independence between e and X while specifying a distribution F with zero mean and unit variance for e.

If an alternative specification for F is chosen, then approximation (3.9) ought to be altered. For instance if F is specified to be the standard uniform distribution, then the corresponding feasible generalized dual regression problem is

$$\max_{u\in[0,1]^n} \; y^{\mathrm{T}}u \quad s.t. \; \frac{1}{j}\sum_{i=1}^{n} x_i \left(u_i^j - \frac{1}{j+1} \right) = 0 \quad (j = 1, \ldots, J). \tag{3.13}$$

On the other hand, the special case of 'dual regression' corresponds to $J = 2$ and $m_1(e_i) = e_i$, $m_2(e_i) = (e_i^2 - 1)/2$, where imposing $\sum_{i=1}^n m_j(e_i) = 0$, for $j = 1, 2$, is a normalization. The simple basis $\{e_i, (e_i^2 - 1)/2\}$ is obviously 'impoverished' for the space of all convex functions, although quite practical for many applications once the flexibility in the distribution of e_i is taken into account.

A further generalization is obtained by regarding X as elementary regressors and defining $W = W(X)$ as a vector formed by transformations of X, cf. Belloni et al. (2011) for a detailed treatment of this series formulation in the context of quantile regression. Except in the notation $F_{Y|X}(Y \mid X)$, this type of series or sieve analysis in the foregoing is achieved by simply substituting W and w for X and x throughout. The remainder of the discussion is unaffected.

4. **Engel's Data Revisited**

4.1. **Empirical illustration.** The classical dataset collected by Engel consists of food expenditure and income measurements for 235 households, and has been studied in depth by Koenker (2005) by means of quantile regression methods. Koenker (2005) shows that the dispersion of food expenditure increases with household income, so that a location-scale model is particularly well-suited to the study of this data. We apply dual regression to the estimation of the statistical relationship between food expenditure and income, with household income as a single regressor and food expenditure as the outcome of interest.

All computational procedures are implemented in the software R (R Development Core Team (2014)). For dual regression we use Ipopt (Interior Point Optimizer), an open source software package for large-scale nonlinear optimization (Waechter & Biegler (2006)), and its

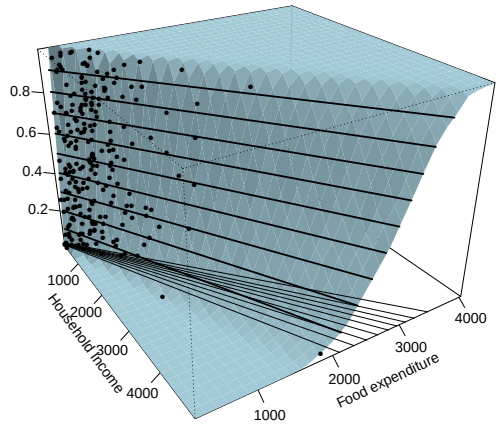
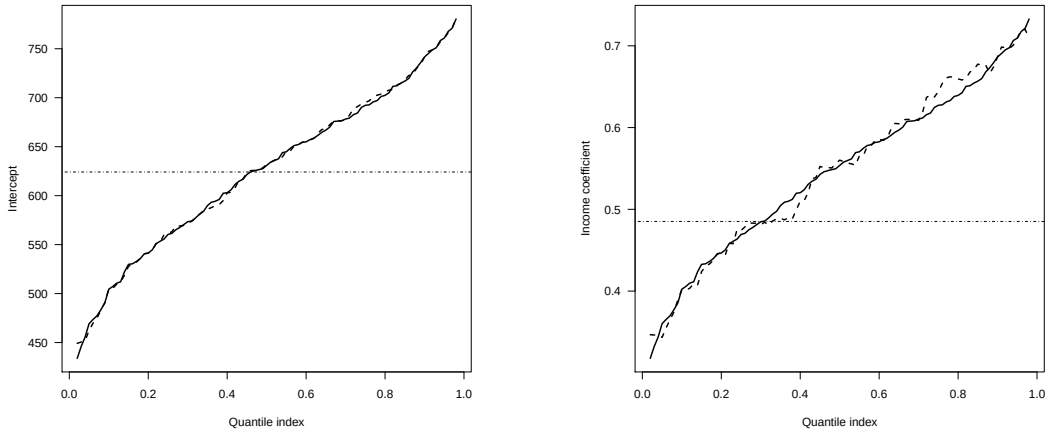


FIGURE 4.1. Dual regression estimate of the distribution of food expenditure conditional on income. Level sets (solid lines) are plotted for a grid of values ranging from 0.1 to 0.9. The projected 'shadow' level sets yield the respective conditional quantile functions appearing on the xy-plane.

R interface Ipoptr developed by Jelmer Ypma. Ipopt has proven to be an effective and easy-to-use solver for the dual regression constrained optimization problem (2.3), and quantile regression procedures in the package quantreg have been used to carry our comparisons.

Figure 4.1 illustrates our results and plots the estimated conditional distribution of food expenditure given household income. The sequence of estimates $\{u_i^*\}_{i=1}^n$, where $u_i^* = F_n(e_i^*)$, is used in order to plot each observation in the xyu-space with predicted coordinates (x_i, y_i, u_i^*), and the solid lines give the u-level sets for a grid of values $\{0.1, \ldots, 0.9\}$. Although non-standard, this representation can be related to standard quantile regression plots since the levels of the distribution function give the conditional quantiles of food expenditure for each value of income. These are the plotted 'shadow' solid lines corresponding for each u to the dual regression estimates of the conditional quantile functions of food expenditure given household income.

It is apparent from Fig. 4.1 that the predicted conditional distribution function obtained by dual regression is indeed endowed with all desired properties. Of particular interest is the fact that the estimated function satisfies the requirement of being monotone in food expenditure. Also, our estimates satisfy some basic smoothness requirements across probability levels,



(A) (B)

FIGURE 4.2. Engel coefficient plots revisited. Dual (solid) and quantile (dashes) regression estimates of the intercept (a) and income (b) coefficients as a function of the quantile index. Least squares estimates are also shown (dot-dash).

in the food expenditure values. This feature does not typically characterize estimates of the conditional quantile process by quantile regression methods, as conditional quantile functions are then estimated sequentially and independently of each other. The decreasing slope of the distribution function across values of income provides evidence that the data indeed follow a heteroscedastic generating process. This is the distributional counterpart of quantile functions having increasing slope across probability levels, a feature characterizing the conditional quantile functions on the xy plane and signalling increasing dispersion in food expenditure across household income values.

Figure 4.2 compares our estimates of the functional intercept and covariate quantile regression coefficients, with estimates obtained by quantile regression. Estimates of quantile regression coefficients for Engel's data are given in Koenker (2005). For interpretational purposes, we follow Koenker (2005) and estimate the functional coefficients after having recentered household income. This avoids having to interpret the intercept as food expenditure for households with zero income. After centering, the intercept coefficient can be interpreted as the u-th quantile of food expenditure for households with mean income: this is Tukey's 'centercept'. Fig. 4.2 shows the estimated quantile regression coefficients as a

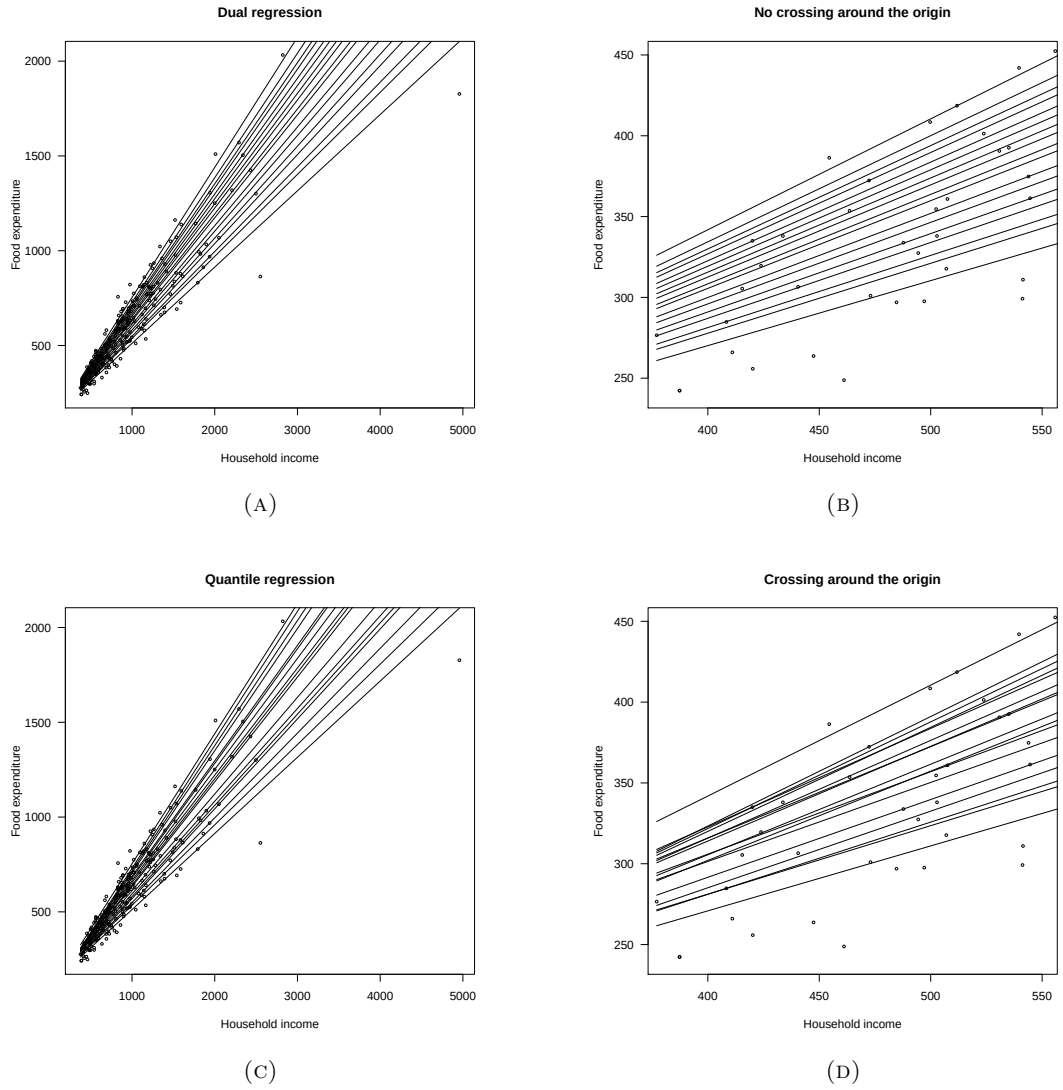


FIGURE 4.3. Scatterplots and dual (a) and quantile (c) regression estimates of the conditional 0.1 to 0.9 quantile functions (solid lines) for Engel's data, and their rescaled counterparts ((b),(d)).

function of u. It illustrates the fact that the location-scale structure imposed by dual regression yields estimates that are indeed smoother than their quantile regression counterpart, the latter having a somewhat erratic behaviour around the dual regression estimates.

Figure 4.3 gives the more familiar quantile regression plots. The plots presented show scatterplots of Engel's data as well as conditional quantile functions obtained by dual and

quantile regression methods. The rescaled plots in the right panels of Fig. 4.3 highlight some features of the two procedures. The fitted lines obtained from dual regression are not subject to crossing in this example, whereas several of the fitted quantile regression lines actually cross for small values of household income. Last, the more evenly spread dual regression conditional quantile functions illustrate the effect of imposing a functional form on the quantile regression coefficients.

4.2. **Simulations.** In this section we give results of several Monte Carlo simulations in order to assess dual regression finite-sample properties. The data generating process is:

$$y_i = \beta_{11} + \beta_{12}x_i + (\beta_{21} + \beta_{22}x_i)\varepsilon_i, \quad \varepsilon_i \sim N(0,1), \tag{4.1}$$

with parameter values calibrated to the Engel data empirical application. We first compare dual regression estimates of the conditional distribution function values $F_{Y|X}(y_i|x_i)$ to those obtained applying the rearrangement procedure of (Chernozhukov et al. (2010)) as a benchmark for dual regression estimates. The performance of dual regression in estimating conditional quantile functions is also studied and compared to linear quantile regression estimation of the functional coefficients $\beta_1(u) = \beta_{11} + \beta_{21}\Phi^{-1}(u)$ and $\beta_2(u) = \beta_{12} + \beta_{22}\Phi^{-1}(u)$. Implementation details and a description of the experiment are given in Appendix **??**. In Appendix E.2 we compare the empirical performance of dual regression to the noncrossing quantile regression method of Bondell et al. (2010).

Table 1 reports a first set of results of our Monte Carlo simulations regarding the accuracy of conditional distribution function estimates across simulations. It reports average estimation errors of dual regression and rearranged quantile regression, respectively, and their ratio in percentage terms. Average estimation errors are measured in L^p norms $\|\cdot\|_p$, $p = 1, 2$, and ∞, where for $f : \mathbb{R} \mapsto [0,1]$, $\|f\|_p = \left\{\int_{\mathbb{R}} |f(s)|^p \, ds\right\}^{1/p}$. For each simulation, the estimation errors $\|u^* - \Phi(\varepsilon)\|_p$ and $\left\|\hat{u}^{QR} - \Phi(\varepsilon)\right\|_p$ are computed, where $\hat{u}^{QR}$ are the rearranged quantile regression estimates, and the errors are averaged across simulations for each sample size. The results show that for this setup dual regression estimates systematically outperform rearranged quantile regression estimates, and that the spread in performance increases with sample size. Whereas the reduction in average estimation error is between 7 and 17%, depending on the norm, for $n = 235$, estimation error is reduced up to 30% when $n = 1000$.

Tables 2 and 3 summarize the results for three different sample sizes regarding the accuracy of the functional intercept and covariate coefficients estimates across simulations. For each coefficient, we compute the root mean absolute error (RMAE) of our estimates, obtained by

TABLE 1. **L^p estimation errors ($\times 100$) and ratios of L^p estimation errors of dual and rearranged quantile regression estimates of the conditional distribution function, for $p = 1, 2$ and ∞.**

Sample size	L^1_{DR}	L^1_{DR}/L^1_{QR}	L^2_{DR}	L^2_{DR}/L^2_{QR}	L^∞_{DR}	$L^\infty_{DR}/L^\infty_{QR}$
$n = 235$	2.67	92.82	3.68	90.48	16.99	82.51
$n = 500$	1.84	91.87	2.53	88.43	13.06	74.82
$n = 1000$	1.31	91.59	1.81	87.63	10.16	69.64

L^p_{DR} and L^p_{QR} are the average L^perrors of the dual and rearranged quantile regression estimates.

TABLE 2. **Summary results of the simulation study for the intercept coefficient: RMAE across quantile indices and sample sizes.**

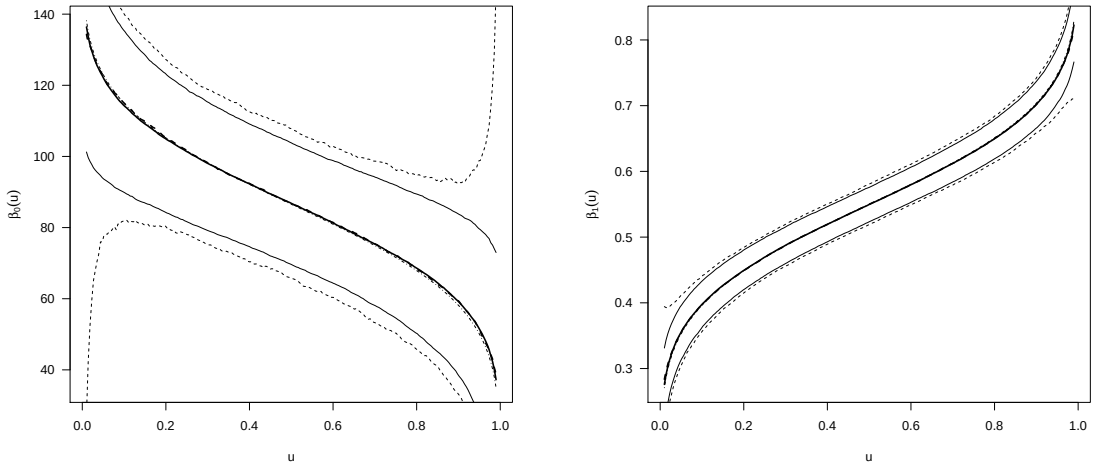
Sample size	Method	RMAE				
		Quantile index				
		$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
$n = 235$	DR	3.31	2.98	2.85	3.02	3.34
	QR	3.73	3.32	3.19	3.34	3.78
$n = 500$	DR	2.74	2.48	2.34	2.46	2.71
	QR	3.09	2.73	2.64	2.73	3.06
$n = 1000$	DR	2.30	2.08	1.98	2.09	2.31
	QR	2.57	2.32	2.20	2.30	2.55

RMAE, square root of mean absolute error across simulations; DR, dual regression; QR, quantile regression.

TABLE 3. **Summary results of the simulation study for the income coefficient: RMAE across quantile indices and sample sizes.**

Sample size	Method	RMAE				
		Quantile index				
		$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
$n = 235$	DR	0.13	0.12	0.11	0.12	0.13
	QR	0.14	0.13	0.12	0.13	0.14
$n = 500$	DR	0.11	0.10	0.09	0.10	0.11
	QR	0.12	0.11	0.10	0.10	0.12
$n = 1000$	DR	0.09	0.08	0.08	0.08	0.09
	QR	0.10	0.09	0.09	0.09	0.10

RMAE, square root of mean absolute error across simulations; DR, dual regression; QR, quantile regression.



(A) (B)

FIGURE 4.4. Simulation results for intercept (a) and covariate (b) coefficients: median estimates across simulations and 90% confidence intervals. Solid lines are dual regression estimates and dashed lines are quantile regression estimates. The truth (dot-dash) is covered by median estimates.

either method, by computing errors for quantile indices in $\{0.1, 0.25, 0.5, 0.75, 0.9\}$ for each replication and then computing the summary statistic. In all cases dual regression estimates have lower RMAE, which corroborates results shown in Fig. 4.4.

Figure 4.4 illustrates the results of our simulations with $n = 235$, the number of observations in Engel's data. For both dual and quantile regression, the solid line is the median estimate of intercept and covariate coefficients $\beta_1(u)$ and $\beta_2(u)$ across simulations. The 90% confidence bands are constructed pointwise by taking the 0.05 and 0.95 quantile estimates across simulations. For both coefficients, a striking feature is that dual regression bands follow the median estimates uniformly over the entire quantile process, whereas quantile regression confidence bands tend to get wider at extreme values of the probability index u. This is expected since dual regression is a global estimation method and is able to exploit the location-scale structure of the model, thus delivering well-behaved and more precise estimates than quantile regression in this example.

5. **Discussion**

If we designate problems such as (2.3) and (3.1) as (already) 'dual', then their solutions reveal a corresponding 'primal'. Typically, the Lagrange multipliers of the dual appear as parameters in the primal, and the primal has an interpretation as a DGP. So perhaps not surprisingly the constraints on the construction of the stochastic elements have 'shadow values' that are parameters of a data generating representation. In this way the relation between identification and estimation is made perspicuous: a *parameter* of the DGP is the Lagrange multiplier of a specific constraint on the construction of the stochastic element, so to specify that some parameters are non-zero and others are zero is to say that some constraints are (in the large-sample limit) binding and others are not.

Another way of expressing this is to say that when a primal corresponds to the DGP, additional moment conditions are superfluous: they will (in the limit) attract Lagrange multiplier values of zero and consequently not affect the value of the program (the objective function) nor the solution. In a sense, this is obvious: the parameters of the primal can typically be identified and estimated through an M–estimation problem that will generate k equations to be solved for the k unknown parameters. Nonetheless, the recognition that the only moment conditions that contribute to enforcing the independence requirement are those whose imposition simultaneously reduces the objective function while providing multipliers that are coefficients in the stochastic representation of Y suggests the futility of portmanteau approaches (e.g. those based on characteristic functions) to imposing independence. The dual formulation reveals that to specify the binding moment conditions *is* to specify a (approximating) DGP representation, which *then* can be extrapolated to provide estimates of objects of interest beyond the n explicitly estimated values of $u_i = F_{Y|X}(Y = y_i \mid X = x_i)$ that characterize the sample and the definition of the mathematical program.

As is well understood in mathematical programming, dual solutions provide lower bounds on the values obtained by primal problems. In the generic form of the problems we have considered here there is no gap between the primal and dual values; hence in econometrics these problems are said to display 'point identification'. We conjecture that the problems without point identification do have gaps between their dual and primal values, and that this characterization will enhance our understanding.

Appendix A. **Proof of Theorem 1**

A.1. **Convexity Lemma.**

Lemma 5. *Suppose that Conditions 1 and 2 hold. Then, the Hessian matrix of $Q_n(\lambda)$, the objective of the primal dual regression problem (P), is positive definite for all $\lambda \in \Lambda_0$.*

Proof. For $e(y_i, x_i, \lambda) = (y_i - \lambda_1 \cdot x_i)/(\lambda_2 \cdot x_i)$, using that

$$\frac{\partial Q_n}{\partial \lambda_1} = -\sum_{i=1}^{n}[x_i e(y_i, x_i, \lambda)] \tag{A.1}$$

$$\frac{\partial Q_n}{\partial \lambda_2} = -\sum_{i=1}^{n}\frac{1}{2}[x_i\{e(y_i, x_i, \lambda)^2 - 1\}], \tag{A.2}$$

the Hessian matrix $H_n(\lambda)$ is

$$H_n(\lambda) = \sum_{i=1}^{n}\begin{bmatrix} \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} & \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e(y_i, x_i, \lambda) \\ \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e(y_i, x_i, \lambda) & \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e(y_i, x_i, \lambda)^2 \end{bmatrix}.$$

Positive definiteness of M_n for all $\lambda_2 \in \Lambda_2$ under Condition 2 implies that $H_n(\lambda)$ is positive definite for all $\lambda \in \Lambda_0$ if and only if the Schur complement of M_n in $H_n(\lambda)$ is positive definite (Boyd & Vandenberghe (2004), Appendix A.5.5) for all $\lambda \in \Lambda_0$, i.e. if and only if

$$S(\lambda) = \left(\sum_{i=1}^{n}\frac{x_i x_i^\top}{\lambda_2 \cdot x_i} {e_i}^2\right) - \left(\sum_{i=1}^{n}\frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e_i\right)\left(\sum_{i=1}^{n}\frac{x_i x_i^\top}{\lambda_2 \cdot x_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e_i\right),$$

with $e_i = e(y_i, x_i, \lambda)$, satisfies $\det\{S(\lambda)\} > 0$, for all $\lambda \in \Lambda_0$. Letting

$$\Xi(\lambda) = \left(\sum_{i=1}^{n}\frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e_i\right)\left(\sum_{i=1}^{n}\frac{x_i x_i^\top}{\lambda_2 \cdot x_i}\right)^{-1},$$

for all $\lambda \in \Lambda_0$, $S(\lambda)$ is equal to

$$\sum_{i=1}^{n}\left[\left\{\frac{x_i e_i}{(\lambda_2 \cdot x_i)^{1/2}} - \Xi(\lambda)\frac{x_i}{(\lambda_2 \cdot x_i)^{1/2}}\right\}\left\{\frac{x_i e_i}{(\lambda_2 \cdot x_i)^{1/2}} - \Xi(\lambda)\frac{x_i}{(\lambda_2 \cdot x_i)^{1/2}}\right\}^\top\right], \tag{A.3}$$

a finite positive $k \times k$ semidefinite matrix, and equal to zero if and only if

$$x_i e_i = \Xi(\lambda) x_i \quad (i = 1, \ldots, n); \tag{A.4}$$

this is an application of the Cauchy-Schwarz inequality for matrices stated in Tripathi (1999).

If system (A.4) holds, then, with $\Xi_j(\lambda)$ denoting the jth row of $\Xi(\lambda)$,

$$x_{ij}y_i = (\lambda_1 \cdot x_i)x_{ij} + \Xi_j(\lambda)x_i, \quad j = 1, \ldots, k,$$

which implies that $x_{ij}^2 \text{var}(y_i|x_i) = 0$, $j = 1, \ldots, k$. These equalities cannot hold if $\text{var}(y_i|x_i) > 0$, which holds w.p.1. under Condition 1. □

A.2. **Proof of Theorem 1.** *Proof of part (i).* Define the Lagrange dual function (Boyd & Vandenberghe (2004), Chapter 5) $Q_n(\lambda) \equiv \sup_{e \in \mathbb{R}^n} \mathscr{L}(e, \lambda)$. In order to derive $Q_n(\lambda)$, we first show that for $\lambda \in \Lambda_0$ the maximum of the mapping $e \mapsto \mathscr{L}(e, \lambda)$ is attained and is unique, and then evaluate $e \mapsto \mathscr{L}(e, \lambda)$ at this value. We then show that $e \mapsto \mathscr{L}(e, \lambda)$ is unbounded above for all $\lambda \notin \Lambda_0$.

Step 1. For $\lambda \in \Lambda_0$ and $c \in \mathbb{R}$, consider the level sets $\mathcal{B}_c(\lambda) = \{e \in \mathbb{R}^n : -\mathscr{L}(e, \lambda) \leq c\}$ of $-\mathscr{L}$. These sets are compact. Consider a sequence $(e_{(m)})$ in $\mathbb{R}^n$ such that $||e_{(m)}|| \to \infty$ as $m \to \infty$. Let $z_{(m)} = \frac{e_{(m)}}{||e_{(m)}||}$, a bounded sequence with unit norm. By the Bolzano-Weierstrass theorem there exists a convergent subsequence $z_{(m_l)}$, $m_l \to \infty$ as $l \to \infty$, with limit z_o, say. Then, using that $\lambda_2 \cdot x_i > 0$, $i = 1, \ldots, n$, for $\lambda \in \Lambda_0$,

$$-\mathscr{L}(e_{(m_l)}, \lambda) = -||e_{(m_l)}|| \sum_{i=1}^n [y_i - (\lambda_1 \cdot x_i)] z_{i,(m_l)} + ||e_{(m_l)}||^2 \sum_{i=1}^n [\frac{1}{2}(\lambda_2 \cdot x_i) z_{i,(m_l)}^2] + \frac{1}{2} \sum_{i=1}^n (\lambda_2 \cdot x_i) \to \infty$$

as $l \to \infty$, since for $\lambda \in \Lambda_0$

$$\lim_{l \to \infty} -\mathscr{L}(e_{(m_l)}, \lambda) = \left(\lim_{l \to \infty} ||e_{(m_l)}||^2 \right) \frac{1}{2} \sum_{i=1}^n (\lambda_2 \cdot x_i) z_{i,o}^2 = \infty.$$

Therefore $-\mathscr{L}(e, \lambda)$ grows unboundedly as $||e|| \to \infty$, and $\mathcal{B}_c(\lambda)$ is bounded. Since $e \mapsto -\mathscr{L}(e, \lambda)$ is continuous over $\mathbb{R}^n$, $\mathcal{B}_c(\lambda)$ is also closed. It then follows from the Weierstrass theorem that there exists $e(\lambda) \in \arg\min_{e \in \mathbb{R}^n}(-\mathscr{L}(e, \lambda)) = \arg\max_{e \in \mathbb{R}^n} \mathscr{L}(e, \lambda)$.

Step 2. The Hessian matrix of the map $e \mapsto \mathscr{L}(e, \lambda)$ is diagonal with diagonal elements $\partial^2 \mathscr{L} / \partial e_i \partial e_i = -(\lambda_2 \cdot x_i)$, $i = 1, \ldots, n$, and is thus negative definite for all $\lambda \in \Lambda_0$. Therefore, $e \mapsto \mathscr{L}(e, \lambda)$ is strictly concave with unique maximum $e(\lambda)$, for all $\lambda \in \Lambda_0$.

Step 3. For $e(y_i, x_i, \lambda) = (y_i - \lambda_1 \cdot x_i)/(\lambda_2 \cdot x_i)$, it follows from the n first-order conditions of (D) that the maximum $e_i(\lambda) = e(y_i, x_i, \lambda)$, $i = 1, \ldots, n$, for all $\lambda \in \Lambda_0$. Define the function $L : \mathcal{X} \times \mathbb{R} \times \mathbb{R}^{2 \times k} \to \mathbb{R}$ as

$$L(x_i, y_i, \lambda) = \frac{1}{2} \left\{ \left(\frac{y_i - \lambda_1 \cdot x_i}{\lambda_2 \cdot x_i} \right)^2 + 1 \right\} (\lambda_2 \cdot x_i). \tag{A.5}$$

Then, evaluating $\mathscr{L}(e,\lambda)$ at $e = e(\lambda)$ yields after some algebra: $\mathscr{L}(e(\lambda),\lambda) = \sum_{i=1}^{n} L(x_i, y_i, \lambda)$, the maximum of the map $e \mapsto \mathscr{L}(e,\lambda)$, for all $\lambda \in \Lambda_0$.

Step 4. We next show that the domain of $Q_n(\lambda)$ is Λ_0, i.e. we show that if $\lambda_2 \cdot x_i \leq 0$ for some i, then $y^T e - \sum_{i=1}^{n} C(x_i, e_i, \lambda)$ is unbounded above. Define the sets $\mathcal{I}_- = \{i \in \{1,\dots n\} : \lambda_2 \cdot x_i < 0\}$, $\mathcal{I}_0 = \{i \in \{1,\dots n\} : \lambda_2 \cdot x_i = 0\}$ and $\mathcal{I} = \mathcal{I}_- \cup \mathcal{I}_0$. Choose $e_k = t$, for each $k \in \mathcal{I}$, and $e_i = 0$, $i \neq k$, and write the value of the Lagrangian evaluated at the chosen values of e:

$$\overline{\mathscr{L}}(\lambda) = [t \sum_{i \in \mathcal{I}_-} (y_i - \lambda_1 \cdot x_i) - t^2 \sum_{i \in \mathcal{I}_-} \frac{1}{2}(\lambda_2 \cdot x_i) - \frac{1}{2} \sum_{i \in \mathcal{I}_-} (\lambda_2 \cdot x_i)] + t \sum_{i \in \mathcal{I}_0} (y_i - \lambda_1 \cdot x_i).$$

The term in brackets is unbounded above since $-t^2 \sum_{i \in \mathcal{I}_-} \frac{1}{2}(\lambda_2 \cdot x_i) \to \infty$ as $t \to \infty$, and also as $t \to -\infty$. Finally, if $\sum_{i \in \mathcal{I}_0} (y_i - \lambda_1 \cdot x_i) > 0$, then $t \sum_{i \in \mathcal{I}_0} (y_i - \lambda_1 \cdot x_i) \to \infty$ as $t \to \infty$, and $\overline{\mathscr{L}}(\lambda)$ grows unboundedly as $t \to \infty$. If $\sum_{i \in \mathcal{I}_0} (y_i - \lambda_1 \cdot x_i) < 0$, then $t \sum_{i \in \mathcal{I}_0} (y_i - \lambda_1 \cdot x_i) \to \infty$ as $t \to -\infty$, and $\overline{\mathscr{L}}(\lambda)$ grows unboundedly as $t \to -\infty$. Therefore, $\mathscr{L}(e,\lambda)$ is unbounded above for all $\lambda \notin \Lambda_0$.

Step 5. Summarizing the above, the primal dual regression problem is

$$\min_{\lambda \in \mathbb{R}^{2\times k}} Q_n(\lambda) = \begin{cases} \sum_{i=1}^{n} L(x_i, y_i, \lambda) & \lambda_2 \cdot x_i > 0,\ i = 1,\dots,n \\ \infty & \text{otherwise.} \end{cases}$$

This yields the equivalent problem $\min_{\lambda \in \Lambda_0} \sum_{i=1}^{n} L(x_i, y_i, \lambda)$. Therefore, (P) is the dual of (D).

Proof of part (ii). The first-order conditions of (P) implied by (A.1) and (A.2) coincide with system (2.7). From the n first-order conditions of (D), a feasible solution is of the form $e_i = e(y_i, x_i, \lambda)$, $i = 1,\dots n$, and satisfies the constraints of (D). Substituting $e(y_i, x_i, \lambda)$, $i = 1,\dots n$, into the constraints yields the Method-of-Moments representation of (D).

Proof of part (iii). (a) Under Conditions 1 and 2, it follows from Lemma 5 that $Q_n(\lambda)$ is strictly convex over Λ_0. Therefore, λ_n is the unique minimum of $Q_n(\lambda)$ and uniquely solves system (2.7). By part (ii), the Lagrange multiplier vector λ^* associated with a solution to problem (D) satisfies system (2.7). It follows that $\lambda^* = \lambda_n$. Moreover, Step 2 in part (i) implies that, for all $\lambda \in \Lambda_0$, the map $e \mapsto \mathscr{L}(e,\lambda)$ admits a unique maximizer $e(\lambda)$. Thus $e(\lambda^*)$ is the unique maximizer of $\mathscr{L}(e,\lambda^*)$, and $e^* = e(\lambda^*)$, the unique feasible solution to (D). Therefore, the pair (λ_n, e^*) uniquely solves (P) and (D).

(b) By direct substitution, using that $\sum_{i=1}^n(\lambda_1^* \cdot x_i)e_i^* = 0$ and $\sum_{i=1}^n(\lambda_2^* \cdot x_i)(e_i^{*2} - 1) = 0$, at a solution the value of (D) is $\sum_{i=1}^n y_i e_i^* = \sum_{i=1}^n(\lambda_2^* \cdot x_i)$. Using that $\sum_{i=1}^n(\lambda_{2n} \cdot x_i)\{e(y_i, x_i, \lambda_n)^2 - 1\} = 0$, the value of (P) is $1/2\sum_{i=1}^n\{e(y_i, x_i, \lambda_n)^2 + 1\}(\lambda_{2n} \cdot x_i) = \sum_{i=1}^n(\lambda_{2n} \cdot x_i)$. Strong duality then follows from $\lambda_n = \lambda^*$ established in part (a).

Appendix B. **Proof of Theorem 2**

Define $Q_0(\lambda)$, the population objective function of the primal dual regression problem as $Q_0(\lambda) = E[L(x_i, y_i, \lambda)]$ where the function $L(x_i, y_i, \lambda)$ is defined in (A.5). We use the notation $a \vee b = \max(a, b)$.

Lemma 6. *Suppose that Conditions 1-4 hold. Then, for $\lambda \in \Lambda_0$, $Q_0(\lambda)$ is continuously differentiable and $\nabla_\lambda E[L(x_i, y_i, \lambda)] = E[\nabla_\lambda L(x_i, y_i, \lambda)]$.*

Proof. We first show that $E[|L(x_i, y_i, \lambda)|] < \infty$ for all $\lambda \in \Lambda_0$. For some positive constant C such that $\frac{1}{\lambda_2 \cdot x_i} \leq C < \infty$ for all $\lambda_2 \in \Lambda_2$,

$$\begin{aligned}\left(\frac{y_i - \lambda_1 \cdot x_i}{\lambda_2 \cdot x_i}\right)^2 \lambda_2 \cdot x_i &\leq C\{2y_i^2 + 2(\lambda_1 \cdot x_i)^2\} \\ &\leq C(2y_i^2 + 2\sup_{\lambda_1 \in \Lambda_1} ||\lambda_1||^2||x_i||^2), \qquad \text{(B.1)}\end{aligned}$$

and therefore $E[|L(x_i, y_i, \lambda)|] < \infty$ for all $\lambda \in \Lambda_0$ under Condition 3(ii). In addition, there exists an integrable function $\kappa(x_i, y_i)$ such that $||\nabla_\lambda L(x_i, y_i, \lambda)|| \leq \kappa(x_i, y_i)$. There is

$$\begin{aligned}\nabla_{\lambda_1} L(x_i, y_i, \lambda) &= -x_i e(y_i, x_i, \lambda) \\ \nabla_{\lambda_2} L(x_i, y_i, \lambda) &= -\frac{1}{2}x_i(e(y_i, x_i, \lambda)^2 - 1).\end{aligned}$$

Since $-1 \leq e(y_i, x_i, \lambda)^2 - 1$, steps similar to those leading to (B.1) yield

$$\begin{aligned}||x_i(e(y_i, x_i, \lambda)^2 - 1)|| &\leq ||x_i||\,|e(y_i, x_i, \lambda)^2 - 1| \\ &\leq ||x_i||\{1 \vee C^2(2y_i^2 + 2\sup_{\lambda_1 \in \Lambda_1} ||\lambda_1||^2||x_i||^2)\}.\end{aligned}$$

so that $E[||x_i(e(y_i, x_i, \lambda)^2 - 1)||] < \infty$ under Condition 3(ii). This and Holder's inequality then imply that $E[||x_i e(y_i, x_i, \lambda)||] < \infty$, therefore $E[\sup_{\lambda \in \Lambda_0} ||\nabla_\lambda L(x_i, y_i, \lambda)||] < \infty$ under Condition 3. Lemma 3.6 in Newey & Mc Fadden (1994) then implies that $Q_0(\lambda)$ is continuously differentiable in λ, and that the order of differentiation and integration can be interchanged. □

First, both existence and consistency of $\hat{\lambda}$ result from strict convexity of $Q_0(\lambda)$ over Λ_0, and pointwise convergence of $Q_n(\lambda)$ to $Q_0(\lambda)$. Strict convexity and pointwise convergence then together imply uniform convergence, as in, for instance, Theorem 2.7 in Newey & Mc Fadden (1994). The asymptotic distribution of $\hat{\lambda}$ then follows from the Method-of-Moments characterization of the estimates given in part (ii) of Theorem 1, verifying conditions of Theorem 3.4 in Newey & Mc Fadden (1994).

Proof of parts (i) and (ii). We verify the conditions of Theorem 2.7 in Newey & Mc Fadden (1994).

We first show identification. Under Conditions 1-4, $Q_0(\lambda)$ is continuously differentiable and the order of differentiation and integration can be interchanged by Lemma 6. We then show that $\nabla_\lambda Q_0(\lambda)$ is differentiable for $\lambda \in \Lambda$. There is:

$$\nabla_{\lambda\lambda} L(x_i, y_i, \lambda) = \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} \begin{bmatrix} 1 & e(y_i, x_i, \lambda) \\ e(y_i, x_i, \lambda) & e(y_i, x_i, \lambda)^2 \end{bmatrix}.$$

Applying steps similar to those leading to (B.1) in the proof of Lemma 6 yields the bound

$$\left\| \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e(y_i, x_i, \lambda)^2 \right\| \leq C^3 ||x_i||^2 (2y_i^2 + 2 \sup_{\lambda_1 \in \Lambda_1} ||\lambda_1||^2 ||x_i||^2),$$

which has finite expectation under Condition 3(ii). This and Holder's inequality then imply that $E[\sup_{\lambda \in \Lambda} ||\nabla_{\lambda\lambda} L(x_i, y_i, \lambda)||] < \infty$. Lemma 3.6 in Newey & Mc Fadden (1994) then implies that $\nabla_\lambda Q_0(\lambda)$ is continuously differentiable, and that the Hessian matrix of $Q_0(\lambda)$ is

$$H(\lambda) = E \begin{bmatrix} \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} & \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e(y_i, x_i, \lambda) \\ \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e(y_i, x_i, \lambda) & \frac{x_i x_i^\top}{\lambda_2 \cdot x_i} e(y_i, x_i, \lambda)^2 \end{bmatrix},$$

which is a finite positive definite matrix under Conditions 1-4 and steps similar to the proof of Lemma 5. Therefore, $\beta \in \Lambda_0$ is the unique minimizer of $Q_0(\lambda)$, and the identification condition (i) in Theorem 2.7 in Newey & Mc Fadden (1994) is thus verified.

Their condition (ii) follows by convexity of Λ_0 and Condition 3(iii)-(iv), as well as strict convexity of $Q_n(\lambda)$ established in Lemma 5. Finally, since the sample is i.i.d. under Condition 3, pointwise convergence of $Q_n(\lambda)$ to $Q_0(\lambda)$ follows from boundedness of $Q_0(\lambda)$ (established in the proof of Lemma 6) and application of Khinchine's law of large numbers. Hence, all conditions of Newey and McFadden's Theorem 2.7 are satisfied, and there exists $\hat{\lambda} \in \Lambda_0$ with probability approaching one and $\hat{\lambda} \to^p \beta$.

Proof of part (iii). By part (ii) of Theorem 1, the Lagrange multiplier vector $\hat{\lambda}$ solves the $2 \times k$ equations system

$$\frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, \lambda) = 0, \tag{B.2}$$

with $m(y_i, x_i, \lambda)$ defined in Section 2.4. System (B.2) can be equivalently viewed as minimizing

$$\mathcal{Q}_n^{MM}(\lambda) \;=\; \left[\begin{array}{c} \frac{1}{n}\sum_{i=1}^n x_i e(y_i, x_i, \lambda) \\ \frac{1}{n}\sum_{i=1}^n x_i\{e(y_i, x_i, \lambda)^2 - 1\} \end{array}\right]^{\mathrm{T}} \left[\begin{array}{c} \frac{1}{n}\sum_{i=1}^n x_i e(y_i, x_i, \lambda) \\ \frac{1}{n}\sum_{i=1}^n x_i\{e(y_i, x_i, \lambda)^2 - 1\} \end{array}\right].$$

Asymptotic normality of the Method-of-Moments estimator then follows after verifying conditions of Theorem 3.4 in Newey & Mc Fadden (1994).

Under Condition 3, $\beta \in \Lambda_0$ so that their condition (i) is satisfied. The mapping $\lambda \mapsto m(y_i, x_i, \lambda)$ is continuously differentiable in $\lambda \in \Lambda_0$ by inspection, so that their condition (ii) is satisfied. Since the properties of ε_i imply that $E\{m(y_i, x_i, \beta)\} = 0$, the first part of their condition (iii) is satified. In addition, steps similar to the proof of Lemma 6 show that $E\{||m(y_i, x_i, \beta)||^2\}$ and $E\{\sup_{\lambda\in\Lambda_o} ||\nabla_\lambda m(y_i, x_i, \lambda)||\}$ are finite under Conditions 3(ii) and 5, and their conditions (iii)-(iv) are verified. Finally, their full rank condition on $G = E\{\nabla_\lambda m(y_i, x_i, \lambda)\}|_{\lambda=\beta}$ is satisfied under Condition 4: the matrix G can be simplified by noting that the off-diagonal elements

$$E\left\{\frac{x_i x_i^{\mathrm{T}}}{\beta_2 \cdot x_i}\left(\frac{y_i - \beta_1 \cdot x_i}{\beta_2 \cdot x_i}\right)\right\} = 0_{k\times k}$$

and

$$E\left\{\frac{x_i x_i^{\mathrm{T}}}{\beta_2 \cdot x_i}\left(\frac{y_i - \beta_1 \cdot x_i}{\beta_2 \cdot x_i}\right)^2\right\} = E\left(\frac{x_i x_i^{\mathrm{T}}}{\beta_2 \cdot x_i}\right),$$

using that $E(\varepsilon_i \mid x_i) = 0$ and $E(\varepsilon_i^2 \mid x_i) = 1$. Thus, G is a block diagonal matrix with positive definite diagonal elements under Condition 4. Therefore, $n^{1/2}(\hat{\lambda} - \beta) \xrightarrow{d} N(0, G^{-1}S(G^{-1})^{\mathrm{T}})$. Exploiting the block diagonal structure of G, the variance-covariance matrix is $G^{-1}SG^{-1}$ and can be characterized explicitly. Partitioning G and S,

$$G = \left[\begin{array}{cc} G_{11} & 0_{k\times k} \\ 0_{k\times k} & G_{22} \end{array}\right], \quad S = \left[\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & S_{22} \end{array}\right],$$

a bit of algebra yields

$$G^{-1}SG^{-1} = \left[\begin{array}{cc} G_{11}^{-1}S_{11}G_{11}^{-1} & G_{11}^{-1}S_{12}G_{22}^{-1} \\ G_{22}^{-1}S_{21}G_{11}^{-1} & G_{22}^{-1}S_{22}G_{22}^{-1} \end{array}\right].$$

Appendix C. **Proof of Theorem 3**

We let, for $d_i = (y_i, x_i)$, $i = 1\ldots, n$, $\mathbb{E}_n f = \mathbb{E}_n f(d_i) = n^{-1}\sum_{i=1}^n f(d_i)$ and $\mathbb{G}_n f = \mathbb{G}_n\{f(d_i)\} = n^{-1/2}\sum_{i=1}^n [f(d_i) - E\{f(d_i)\}]$.

Define the class of functions

$$\mathcal{F} = \{1\{e(y_i, x_i, \lambda) \leq e\}, \lambda \in \Lambda_0, e \in \mathbb{R}\}.$$

Following van der Vaart & Wellner (2007), the empirical dual regression process $\mathbb{U}_n(e) = n^{1/2}(\mathbb{E}_n f_{e,\hat{\lambda}} - E f_{e,\beta})$ admits the following decomposition:

$$n^{1/2}(\mathbb{E}_n f_{e,\hat{\lambda}} - E f_{e,\beta}) = \mathbb{G}_n(f_{e,\hat{\lambda}} - f_{e,\beta}) + \mathbb{G}_n f_{e,\beta} + \sqrt{n}E(f_{e,\hat{\lambda}} - f_{e,\beta}). \tag{C.1}$$

The proof thus proceeds by (i) establishing that the first term on the right in (C.1) converges in probability to zero, (ii) using the fact that the second term converges in distribution to a mean zero Gaussian process, and (iii) expanding the last term uniformly in $e \in \mathbb{R}$.

Step 1. (Stochastic equicontinuity) By Theorem 2.1 in van der Vaart & Wellner (2007), since $\Pr(\hat{\lambda} \in \Lambda_0) \to 1$ by part (i) of Theorem 1, $\sup_{e\in\mathbb{R}} ||\mathbb{G}_n(f_{e,\hat{\lambda}} - f_{e,\beta})|| \to_p 0$ holds if the class of functions $\mathcal{F}$ is Donsker and if the pseudometric $\rho\{(e', \lambda'), (e'', \lambda'')\}^2 \equiv E[\{f_{e',\lambda'}(d_i) - f_{e'',\lambda''}(d_i)\}^2]$ satisfies $\delta_n \equiv \sup_{e\in\mathbb{R}} \rho\{(e, \hat{\lambda}), (e, \beta)\}^2 \to_p 0$.

We first show that the class of functions $\mathcal{F}$ is Donsker. Define the parametric class of functions $\tilde{\mathcal{F}} = \{e(y_i, x_i, \lambda), \lambda \in \Lambda_0\}$. For all $\lambda', \lambda'' \in \Lambda_0$, a mean-value expansion and Cauchy-Schwarz inequality yield

$$|e(y_i, x_i, \lambda') - e(y_i, x_i, \lambda'')| \leq ||\nabla_\lambda e(y_i, x_i, \lambda)|_{\lambda=\bar{\lambda}}|| \, ||\lambda' - \lambda''||,$$

where $\bar{\lambda}$ is on the line joining λ' and λ''. Steps similar to those in the proof of Theorem 2 show that $E[||\nabla_\lambda e(y_i, x_i, \lambda)|_{\lambda=\bar{\lambda}}||^2]$ is bounded under Condition 3, so that $\tilde{\mathcal{F}}$ is Donsker by Example 19.7 in van der Vaart (1998). Therefore, $\mathcal{F}$ is Donsker, by monotonicity of the indicator function, with unit envelope.

We now show that $\delta_n \to_p 0$. Let $\bar{f}$ denote the upper bound for $|y| f_{Y|X}(y|x)$, and set $\lambda(e) = \lambda_1 + \lambda_2 e$ and $\beta(e) = \beta_1 + \beta_2 e$, $e \in \mathbb{R}$. Upon using that $1\{e(y_i, x_i, \lambda) \leq e\} = 1\{y_i \leq \lambda(e) \cdot x_i\}$ for all $\lambda \in \Lambda_0$, the law of iterated expectations, a mean-value expansion and Cauchy-Schwarz

inequality yield:

$$\begin{aligned}\sup_{e\in\mathbb{R}} \rho((e,\hat{\lambda}),(e,\beta))^2 &= \sup_{e\in\mathbb{R}} E[|1\{y_i \leq \hat{\lambda}(e)\cdot x_i\} - 1\{y_i \leq \beta(e)\cdot x_i\}|] \\ &= \sup_{e\in\mathbb{R}} E(|(\hat{\lambda}-\beta)^\top [f_{Y|X}\{\bar{\lambda}(e)\cdot x_i|x_i\}(x_i, x_i e)^\top]|) \\ &\leq ||\hat{\lambda}-\beta||\bar{f}E||x_i||,\end{aligned}$$

where $\bar{\lambda}$ is on the line joining $\hat{\lambda}$ and β. Thus $\delta_n = o_p(1)$ by Condition 3(ii) and consistency of $\hat{\lambda}$.

Step 2. (Expansion) We show that the following expansion is valid uniformly in $e \in \mathbb{R}$:

$$E\{f_{e,\hat{\lambda}}(d_i) - f_{e,\beta}(d_i)\} = (\hat{\lambda}-\beta)^\top\{g(e) + o_P(1)\}. \tag{C.2}$$

Upon using that $1\{e(y_i, x_i, \lambda) \leq e\} = 1\{y_i \leq \lambda(e)\cdot x_i\}$ for all $\lambda \in \Lambda_0$, the law of iterated expectations and a mean-value expansion yield:

$$E\{f_{e,\hat{\lambda}}(d_i) - f_{e,\beta}(d_i)\} = (\hat{\lambda}-\beta)^\top E[\nabla_\lambda F_{Y|X}\{\lambda(e)\cdot x_i|x_i\}|_{\lambda=\bar{\lambda}}],$$

where $\bar{\lambda}$ is on the line joining $\hat{\lambda}$ and β.

Using that $\nabla_\lambda F_{Y|X}\{\lambda(e)\cdot x_i|x_i\} = f_{Y|X}\{\lambda(e)\cdot x_i|x_i\}(x_i, x_i e)^\top$ for all $\lambda \in \Lambda_0$, we obtain

$$E[f_{Y|X}\{\bar{\lambda}(e)\cdot x_i|x_i\}(x_i, x_i e)^\top] = E[f_{Y|X}\{\beta(e)\cdot x_i|x_i\}(x_i, x_i e)^\top] + o_p(1),$$

uniformly in $e \in \mathbb{R}$, by uniform continuity of the mapping $y \mapsto f_{Y|X}(y|x)$, uniformly in x over $\mathcal{X}$, uniform consistency of $\hat{\lambda}(e)$ implied by consistency of $\hat{\lambda}$ and linearity of $\hat{\lambda}(e)$ in e, and since $\sup_{e\in\mathbb{R}} |e| f_{Y|X}(\beta(e)\cdot X|X) \leq \bar{f}$ and $E[||x_i||] < \infty$ by Condition 3(ii). Hence (C.2) holds by definition of $g(e)$, uniformly in $e \in \mathbb{R}$.

Finally, the Method-of-Moments representation of dual regression implies that the dual regression estimator $\hat{\lambda}$ is asymptotically linear with influence function

$$\psi(y_i, x_i, \beta) = -G^{-1} m(y_i, x_i, \beta). \tag{C.3}$$

Thus (C.1)-(C.3) together imply that uniformly in $e \in \bar{\mathcal{E}}$

$$\begin{aligned}\mathbb{U}_n(e) &= \mathbb{G}_n(f_{e,\hat{\lambda}} - f_{e,\beta}) + \mathbb{G}_n f_{e,\beta} + n^{1/2}(\hat{\lambda}-\beta)^\top\{g(e) + o_P(1)\} \\ &= o_P(1) + \mathbb{G}_n f_{e,\beta} + g(e)^\top n^{-1/2}\sum_{i=1}^n \psi(y_i, x_i, \beta) + o_P(1) \\ &= n^{-1/2}\sum_{i=1}^n \varphi_e(y_i, x_i, \beta) + o_P(1).\end{aligned}$$

Hence the empirical dual regression process $\mathbb{U}_n(\cdot)$ weakly converges to the zero-mean Gaussian process $\mathbb{U}(\cdot)$ in $\ell^\infty(\mathbb{R})$, the set of uniformly bounded real functions on $\mathbb{R}$, and where $\mathbb{U}(\cdot)$ has covariance function $E\{\varphi_e(y_i, x_i, \beta)\varphi_{e'}(y_i, x_i, \beta)\}$.

Appendix D. **Proof of Theorem 4**

Lemma 7. *Suppose that $H_{x_i} : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable, strictly increasing function for each $x_i \in \mathcal{X}$. Then, $\Lambda = 1$ is the unique solution to the equation*

$$\sum_{i=1}^{n} \widetilde{H}_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\} - S_n = 0 \tag{D.1}$$

such that $\Lambda > 0$.

Proof. Equation (D.1) is the first-order condition of the minimization problem

$$\min_{\Lambda>0} q_n(\Lambda) \equiv \sum_{i=1}^{n} y_i H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right) - \Lambda\left[\sum_{i=1}^{n} \widetilde{H}_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\} - S_n\right],$$

where, for $\mathscr{L}^{IGDR}(e, \Lambda) = y^\top e - \Lambda\{\sum_{i=1}^n \widetilde{H}_{x_i}(e_i) - S_n\}$, $q_n(\Lambda) = \sup_{e\in\mathbb{R}^n} \mathscr{L}^{IGDR}(e, \Lambda)$ for all $\Lambda > 0$ such that $\mathscr{L}^{IGDR}(e, \Lambda) < \infty$. The function $q_n(\Lambda)$ is strictly convex over $(0, \infty)$: since $H_{x_i}(e_{oi})$ is continuously differentiable in e_{oi} for all $x_i \in \mathcal{X}$ by assumption, by the inverse function theorem $H_{x_i}^{-1}(y_i)$ is continuously differentiable in y_i for all $x_i \in \mathcal{X}$ and there are the following derivatives:

$$\frac{\partial H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)}{\partial \Lambda} = -\frac{1}{H'_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\}}\frac{y_i}{\Lambda^2} \tag{D.2}$$

$$\frac{\partial \widetilde{H}_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\}}{\partial \Lambda} = -\frac{y_i}{\Lambda}\frac{1}{H'_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\}}\frac{y_i}{\Lambda^2}, \tag{D.3}$$

for all $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Upon using (D.2) and (D.3), $q_n(\Lambda)$ has first derivative

$$\frac{\partial q_n}{\partial \Lambda} = -\left[\sum_{i=1}^{n} \widetilde{H}_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\} - S_n\right]$$

and second derivative

$$\frac{\partial^2 q_n}{\partial \Lambda \partial \Lambda} = \frac{1}{\Lambda}\sum_{i=1}^{n} \frac{1}{H'_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\}}\left(\frac{y_i}{\Lambda}\right)^2 > 0,$$

since strict monotonicity and continuous differentiability of H_{x_i} imply that $H'_{x_i} > 0$ and is bounded over its entire domain, for each $x_i \in \mathcal{X}$. Therefore, $q_n(\Lambda)$ is strictly convex

over $(0,\infty)$ and admits at most one minimum. Since $H_{x_i}^{-1}(y_i/\Lambda) = e_{oi}$ for $\Lambda = 1$ and $S_n = \sum_{i=1}^n \widetilde{H}_{x_i}(e_{oi})$ by definition, $\Lambda = 1$ solves the equation $\partial q_n/\partial\Lambda = 0$. The result follows. □

Proof of Theorem 4. In order to show that the pair $(\Lambda, e) = (1, e_o)$ uniquely solves the first-order conditions (3.6), write the Lagrangian: $\mathscr{L}^{IGDR}(e,\Lambda) = y^\top e - \Lambda\{\sum_{i=1}^n \widetilde{H}_{x_i}(e_i) - S_n\}$. By assumption, H_{x_i} is continuously differentiable and strictly increasing, and by application of the fundamental theorem of calculus, $\widetilde{H}_{x_i}$ is differentiable with derivative H'_{x_i} such that $H'_{x_i} > 0$ for all $x_i \in \mathcal{X}$, and the inverse function of H_{x_i}, denoted $H_{x_i}^{-1}$, is well-defined. The n first-order conditions of problem (3.5) are

$$\frac{\partial \mathscr{L}^{IGDR}}{\partial e_i} = y_i - \Lambda H_{x_i}(e_i) = 0 \qquad (i = 1,\dots,n), \tag{D.4}$$

so that $e_i = H_{x_i}^{-1}(y_i/\Lambda)$. The n second-order conditions are

$$\frac{\partial^2 \mathscr{L}^{IGDR}}{\partial e_i \partial e_i} = -\Lambda H'_{x_i}(e_i) < 0 \qquad (i = 1,\dots,n), \tag{D.5}$$

which are satisfied if and only if $\Lambda > 0$, since H'_{x_i} is strictly positive over its entire domain. For $\Lambda \leq 0$, (D.5) implies that $\mathscr{L}^{IGDR}$ is unbounded above since $e \mapsto \mathscr{L}^{IGDR}(e,\Lambda)$ is then convex. Therefore we need only consider a pair (Λ, e) solving the first-order conditions (D.4) and satisfying the constraint of problem (3.5) with $\Lambda > 0$. For $\widetilde{\Lambda} \neq 1$ and $\tilde{e}_i = H_{x_i}^{-1}(y_i/\widetilde{\Lambda})$, $i = 1,\dots,n$, let $(\widetilde{\Lambda}, \tilde{e})$ be such a pair.

Substituting $e_i = \tilde{e}_i$ into the constraint of problem (3.5) yields

$$\sum_{i=1}^n \widetilde{H}_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\widetilde{\Lambda}}\right)\right\} - S_n = 0. \tag{D.6}$$

By Lemma 7, $\widetilde{\Lambda} = 1$ is the only solution to (D.6) such that $\widetilde{\Lambda} > 0$, a contradiction. Therefore, $(1, e_o)$ is the unique pair that solves the first-order conditions (D.4).

Appendix E. **Numerical Simulations**

E.1. **Design and implementation of the numerical simulations.** Data is generated according to the location-scale model (4.1) calibrated to Engel's data. The value of β is set to the value of estimates obtained by the method suggested in Koenker & Xiao (2002): for a grid of $R = 235$ quantile indices $\{u_1,\dots,u_R\}$, $(\hat{\beta}_1^{QR}(u_r), \hat{\beta}_2^{QR}(u_r))$ are estimated by quantile regression, and β_1 and β_2 are set equal to the estimates obtained from linear regression of $(\hat{\beta}_1^{QR}(u_r), \hat{\beta}_2^{QR}(u_r))$ on $\{(1, \Phi^{-1}(u_r)) : 1,\dots,R\}$, where Φ^{-1} is the inverse standard normal

distribution. We set $\beta_1 = (86 \cdot 56, -22 \cdot 17)$ and $\beta_2 = (0 \cdot 55, 0 \cdot 12)$. Therefore, the quantile regression parameters are $\beta_1(u) = \beta_{11} + \beta_{12}\Phi^{-1}(u)$ and $\beta_2(u) = \beta_{21} + \beta_{22}\Phi^{-1}(u)$, and the conditional distribution function is $F_{Y|X}(y \mid x) = \Phi\{(y - \beta_1 \cdot x)/(\beta_2 \cdot x)\}$. $\tilde{x}_i$ is a scalar random variable drawn from a left-truncated normal distribution with truncation point equal to min(income) $-$ 100 = 277. Two alternative designs were considered with values of $\tilde{x}_i$ fixed to sample values of income across simulations for $n = 235$ or sampling from values of income for $n = (100, 500, 1000)$, and with $\tilde{x}_i \sim U(\min(\text{income}), \max(\text{income}))$; results are similar to the truncated normal design and are omitted, but are available upon request. The number of replications is 4999. The dual regression Lagrange multipliers yield estimated functional coefficients $\hat{\beta}_j(u) = \lambda^*_{1j} + \lambda^*_{2j} F_n^{-1}(u)$, $j = 1, 2$, where F_n^{-1} is the inverse empirical distribution function of e^*. As a benchmark, the conditional distribution function is also estimated by rearranged quantile regression (Chernozhukov et al. (2010)), as $\hat{u}_i^{QR} = \epsilon + \int_\epsilon^{1-\epsilon} 1\{\hat{\beta}_1^{QR}(u) + \hat{\beta}_2^{QR}(u)\tilde{x}_i \leq y_i\} du$, with $\epsilon = 0 \cdot 001$.

E.2. **Additional Simulations.** We provide additional simulations comparing dual regression to the noncrossing quantile regression method introduced by Bondell et al. (2010), replicating the experiments they propose. In their simulation study they consider three examples which are special cases of the linear heteroscedastic model

$$y_i = \gamma_1 + \beta_1 \cdot \tilde{x}_i + (\gamma_2 + \beta_2 \cdot \tilde{x}_i)\varepsilon_i, \quad \tilde{x}_{ij} \sim U(0, 1), \quad \varepsilon_i \sim N(0, 1),$$

with $\gamma_1 = \gamma_2 = 1$. Their method imposes noncrossing constraints on the quantile regressions estimated, and they show that it outperforms both linear quantile regression and the method of He (1997). The three examples are:

Example 1. $\dim(\tilde{x}_i) = 4$, $\beta_1 = (1, 1, 1, 1)^{\mathrm{T}}$, and $\beta_2 = (0 \cdot 1, 0 \cdot 1, 0 \cdot 1, 0 \cdot 1)^{\mathrm{T}}$.

Example 2. $\dim(\tilde{x}_i) = 10$, $\beta_1 = (1, 1, 1, 1, 0^{\mathrm{T}})^{\mathrm{T}}$, and $\beta_2 = (0 \cdot 1, 0 \cdot 1, 0 \cdot 1, 0 \cdot 1, 0^{\mathrm{T}})^{\mathrm{T}}$.

Example 3. $\dim(\tilde{x}_i) = 7$, $\beta_1 = (1, 1, 1, 1, 1, 1, 1)^{\mathrm{T}}$, and $\beta_2 = (1, 1, 1, 0, 0, 0, 0)^{\mathrm{T}}$.

For each example, 500 datasets of size 100, 200 and 500 are simulated. For the method of Bondell et al. (2010), six quantile curves are fitted to the data for each example, $u = \{0 \cdot 1, 0 \cdot 3, 0 \cdot 5, 0 \cdot 7, 0 \cdot 9, 0 \cdot 99\}$. We also implemented the noncrossing quantile regression method by fitting eleven quantile curves for the larger sequence $u = \{0 \cdot 01, 0 \cdot 1, 0.2, \ldots, 0 \cdot 9, 0 \cdot 99\}$, the results are similar and are thus omitted.

Table 4 shows the average root mean integrated squared errors over the 500 datasets along with their estimated standard errors, for each sample size, and for each of $u = \{0 \cdot 5, 0 \cdot 9, 0 \cdot$

99}. For each simulation, the empirical root mean integrated squared error is calculated as RMISE $= [n^{-1} \sum_{i=1}^{n} \{\hat{\beta}(u) \cdot x_i - \beta(u) \cdot x_i\}^2]^{1/2}$, where $\hat{\beta}(u)$ and $\beta(u)$ are the estimated and true vector of quantile regression coefficients, respectively. The results for the other quantiles are similar, and are thus omitted.

In all three examples dual regression significantly outperforms the noncrossing quantiles method for all quantiles and all sample sizes, except for $n = 100$ and $\tau = 0 \cdot 9$ in Example 2. The good relative performance of dual regression results from the imposed location-scale structure, which adds further smoothness and stability across quantile curves, beyond the noncrossing constraints imposed by noncrossing quantile regression. Since the DGP is a linear heteroscedastic model, it is expected that dual regression would perform better. This improvement is greater in the tails, as the location and scale parameters are estimated globally whereas the local nature of quantile regression affects estimation of extreme quantiles.

Example 1

	$\tau = 0 \cdot 5$	$\tau = 0 \cdot 9$	$\tau = 0 \cdot 99$
		$n = 100$	
DR	$26 \cdot 55\,(0 \cdot 41)$	$37 \cdot 43\,(0 \cdot 54)$	$58 \cdot 38\,(0 \cdot 88)$
NCRQ	$30 \cdot 74\,(0 \cdot 44)$	$41 \cdot 39\,(0 \cdot 57)$	$71 \cdot 11\,(0 \cdot 90)$
Ratio$\times 100$	$86 \cdot 36$	$90 \cdot 44$	$82 \cdot 10$
		$n = 200$	
DR	$18 \cdot 93\,(0 \cdot 27)$	$25 \cdot 72\,(0 \cdot 37)$	$41 \cdot 14\,(0 \cdot 65)$
NCRQ	$21 \cdot 93\,(0 \cdot 33)$	$30 \cdot 02\,(0 \cdot 45)$	$56 \cdot 39\,(0 \cdot 75)$
Ratio$\times 100$	$86 \cdot 29$	$85 \cdot 66$	$72 \cdot 97$
		$n = 500$	
DR	$12 \cdot 01\,(0 \cdot 17)$	$16 \cdot 36\,(0 \cdot 23)$	$26 \cdot 22\,(0 \cdot 43)$
NCRQ	$14 \cdot 01\,(0 \cdot 21)$	$19 \cdot 30\,(0 \cdot 27)$	$40 \cdot 41\,(0 \cdot 57)$
Ratio$\times 100$	$85 \cdot 74$	$84 \cdot 75$	$64 \cdot 89$

Example 2

	$\tau = 0 \cdot 5$	$\tau = 0 \cdot 9$	$\tau = 0 \cdot 99$
		$n = 100$	
DR	$40 \cdot 89\,(0 \cdot 40)$	$57 \cdot 50\,(0 \cdot 58)$	$89 \cdot 09\,(0 \cdot 87)$
NCRQ	$43 \cdot 76\,(0 \cdot 43)$	$53 \cdot 94\,(0 \cdot 51)$	$91 \cdot 14\,(0 \cdot 90)$
Ratio$\times 100$	$93 \cdot 44$	$106 \cdot 61$	$97 \cdot 74$
		$n = 200$	
DR	$28 \cdot 74\,(0 \cdot 28)$	$39 \cdot 05\,(0 \cdot 37)$	$59 \cdot 85\,(0 \cdot 60)$
NCRQ	$32 \cdot 03\,(0 \cdot 30)$	$39 \cdot 65\,(0 \cdot 38)$	$65 \cdot 35\,(0 \cdot 63)$
Ratio$\times 100$	$89 \cdot 75$	$98 \cdot 48$	$91 \cdot 58$
		$n = 500$	
DR	$17 \cdot 93\,(0 \cdot 17)$	$24 \cdot 19\,(0 \cdot 24)$	$37 \cdot 33\,(0 \cdot 42)$
NCRQ	$21 \cdot 04\,(0 \cdot 19)$	$27 \cdot 52\,(0 \cdot 26)$	$47 \cdot 56\,(0 \cdot 45)$
Ratio$\times 100$	$85 \cdot 24$	$87 \cdot 90$	$78 \cdot 48$

Example 3

	$\tau = 0 \cdot 5$	$\tau = 0 \cdot 9$	$\tau = 0 \cdot 99$
		$n = 100$	
DR	$70 \cdot 33\,(0 \cdot 83)$	$97 \cdot 96\,(1 \cdot 18)$	$153 \cdot 28\,(1 \cdot 75)$
NCRQ	$76 \cdot 78\,(0 \cdot 91)$	$100 \cdot 86\,(1 \cdot 21)$	$176 \cdot 85\,(1 \cdot 96)$
Ratio$\times 100$	$91 \cdot 60$	$97 \cdot 13$	$86 \cdot 67$
		$n = 200$	
DR	$50 \cdot 00\,(0 \cdot 58)$	$68 \cdot 22\,(0 \cdot 85)$	$106 \cdot 47\,(1 \cdot 36)$
NCRQ	$56 \cdot 49\,(0 \cdot 65)$	$74 \cdot 25\,(0 \cdot 91)$	$134 \cdot 98\,(1 \cdot 54)$
Ratio$\times 100$	$88 \cdot 51$	$91 \cdot 87$	$78 \cdot 88$
		$n = 500$	
DR	$30 \cdot 51\,(0 \cdot 36)$	$41 \cdot 64\,(0 \cdot 50)$	$66 \cdot 72\,(0 \cdot 90)$
NCRQ	$35 \cdot 64\,(0 \cdot 42)$	$47 \cdot 84\,(0 \cdot 57)$	$94 \cdot 45\,(1 \cdot 09)$
Ratio$\times 100$	$85 \cdot 59$	$87 \cdot 04$	$70 \cdot 64$

DR, dual regression; NCRQ, noncrossing quantile regression method of Bondell et al. (2010).

TABLE 4. *Replication of Bondell et al. (2010) experiments 1-3: average root mean integrated squared error (*$\times 100$*) over 500 simulations, with standard error in parentheses.*

References

Belloni, A., Chernozhukov, V. & Fernandez-Val, I. (2011). Conditional quantile processes based on series or many regressors. *Arxiv 1105.6154*.

Bondell, H., Reich, B. and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97**, 825–838.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Chernozhukov, V., Fernandez-Val, I. & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **81**, 2205–2268.

De Vore, R. (1977). Monotone approximation by splines. *SIAM Journal on Mathematical Analysis* **8**, 891–905.

De Vore, R. (1977). Monotone approximation by polynomials. *SIAM Journal on Mathematical Analysis* **8**, 906–921.

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

He, X. (1997). Quantile Curves without Crossing. *The American Statistician* **51**, 186–192.

Huber, P. (1981). *Robust Statistics*. Wiley, New York.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monograph Series, Vol. 38. Cambridge University Press.

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

Koenker, R. & Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica* **70**, 1583–1612.

Koenker, R. & Zhao, Q. (1994). L-estimation for linear heteroscedastic models. *Nonparametric Statistics* **3**, 223–235.

Newey, W. & Mc Fadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, ch. 36, 1st ed., pp. 2111–2245. Amsterdam: Elsevier.

Owen, A. (2001). *Empirical Likelihood*. Chapman&Hall/CRC, Boca Raton, USA.

Parker, T. (2013). A comparison of alternative approaches to supremum-norm goodness-of-fit tests with estimated parameters. *Econometric Theory* **29**, 968–1008.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Tripathi, G. (2006). A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters* **63**, 1–3.

Waechter, A. & Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* **106**, 25–57.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A.W. and Wellner, J. (2007). *Empirical processes indexed by estimated functions*. Lecture Notes-Monograph Series, 234–252.