QUANTILE REGRESSION METHODS FOR RECURSIVE STRUCTURAL EQUATION MODELS

LINGJIE MA AND ROGER KOENKER

ABSTRACT. Two classes of quantile regression estimation methods for the recursive structural equation models of Chesher (2003) are investigated. A class of weighted average derivative estimators based directly on the identification strategy of Chesher is contrasted with a control variate estimation method. The latter imposes stronger restrictions achieving an asymptotic efficiency bound with respect to the former class. An application of the methods to the study of the effect of class size on the performance of Dutch primary school students shows that (i.) reductions in class size are beneficial for good students in language and for weaker students in mathematics, (ii) larger classes appear beneficial for weaker language students, and (iii.) the impact of class size on both mean and median performance is negligible.

1. INTRODUCTION

Classical two-stage least squares methods and the limited information maximum likelihood estimator provide attractive methods of estimation for Gaussian linear structural equation models with additive errors. However, these methods offer only a conditional mean view of the structural relationship, implicitly imposing quite restrictive location-shift assumptions on the way that covariates are allowed to influence the conditional distributions of the endogenous variables. Quantile regression methods seek to broaden this view, offering a more complete characterization of the stochastic relationship among variables and providing more robust, and consequently more efficient, estimates in some non-Gaussian settings.

Amemiya (1982) was the first to seriously consider quantile regression methods for the structural equation model showing the consistency and asymptotic normality of a class of two-stage median regression estimators. Subsequent work of Powell (1983) and Chen and Portnoy (1996) extended this approach, but maintained the focus primarily on the conditional median problem. Recent work has sought to broaden the perspective. Abadie, Angrist and Imbens (2002) considered quantile regression methods for estimating endogenous treatment effects focusing on the binary treatment case. Sakata (2000) has considered an ℓ_1 analogue of the LIML estimator. Chernozhukov and Hansen (2001) have proposed a novel instrumental variables approach.

Version: November 18, 2003. The authors would like to express their appreciation to Andrew Chesher for stimulating conversations regarding this work. They would also like to thank Annelie van der Wind for her extensive help with the interpretation of the PRIMA data. This work was partially supported by NSF Grant SES-0240781.

In an important series of recent papers Chesher (2001, 2002, 2003) has considerably expanded the scope of quantile regression methods for structural econometric models. He considers a general nonlinear specification whose crucial feature is its triangular stochastic structure. By recursively conditioning, a sequence of conditional quantile functions are available to characterize the model and identify the structural effects. The approach may be viewed as a natural generalization of the "causal chain" models advocated by Strotz and Wold (1960).

Chesher has elegantly laid out the structural interpretation of his proposed models and dealt with the ensuing identification issues. Our objective is to consider more pragmatic problems of estimation and inference. We will describe two general classes of the estimation methods. The first is a class of average derivative methods based directly on the Chesher identification strategy. The second is a "control variate" approach. In parametric settings with covariate effects assumed to satisfy certain location-scale shift restrictions we are able to compare the asymptotic behavior of the two approaches and show that the control variate methods attain an efficiency bound corresponding to an optimally weighted form of the average derivative estimator. In typical applications where the precise specification of the covariate effects are subject to dispute the two estimation strategies are useful complements, offering a valuable framework for inference.

The next section introduces the recursive structural model and describes our two classes of estimators. We will focus primarily on a simple two equation setting, with some brief remarks on the extension to larger models. Sections 3 and 4 are devoted to the asymptotic behavior of the estimators and their asymptotic relative efficiency. Section 5 discusses some issues related to the specification of linear conditional quantile models. Section 6 reports the results of a small simulation experiment designed to explore the finite sample performance of the two approaches. Section 7 describes an application of the models to the problem of estimating structural effects of changes in class size on student performance in Dutch primary schools.

2. Recursive Structural Models and Their Estimation

To motivate Chesher's approach it is worthwhile to briefly reconsider the simple, exactly identified, triangular model,

$$Y_{i1} = Y_{i2}\alpha_1 + x_i^\top \alpha_2 + \nu_{i1} + \lambda\nu_{i2} \tag{2.1}$$

$$Y_{i2} = z_i\beta_1 + x_i^\top \beta_2 + \nu_{i2}. \tag{2.2}$$

Suppose that the unobserved errors ν_{i1} and ν_{i2} are stochastically independent and identically distributed with $\nu_{i1} \sim F_1$ and $\nu_{i2} \sim F_2$. Assume further that the ν_{ij}'s are independent of $(z_i, x_i^\top)^\top$, and that for convenience Y_{i2} and z_i are scalars.

We will focus on the estimate of the scalar structural parameter α_1. The classical two stage least squares estimator of α_1 may be written as,

$$\hat{\alpha}_1 = (\hat{Y}_2^\top M_x \hat{Y}_2)^{-1} \hat{Y}_2^\top M_x Y_1$$

where $\hat{Y}_2 = z\hat{\beta}_1 + X\hat{\beta}_2$, $\hat{\beta}_1 = (z'M_X z)^{-1}z'M_X Y_2$, $\hat{\beta}_2 = (X'M_z X)^{-1}X'M_z Y_2$, $M_z = I - z(z'z)^{-1}z'$, and $M_X = I - X(X'X)^{-1}X'$. A somewhat less conventional interpretation of $\hat{\alpha}_1$ can be derived substituting for ν_{i2} in (2.1) to obtain

$$Y_1 = Y_2(\alpha_1 + \lambda) - z\beta_1\lambda + X(\alpha_2 - \lambda\beta_2) + \nu_1 \equiv W\delta + \nu_1, \tag{2.3}$$

where $W = [Y_2 \vdots z \vdots X]$ and $\delta = (\delta_1, \delta_2, \delta_3^\top)^\top = (\alpha_1 + \lambda, -\beta_1\lambda, \alpha_2^\top - \lambda\beta_2^\top)^\top$.

Now, suppose we estimate the hybrid structural equation (2.3) by ordinary least squares. We have the following result.

Proposition 1. *$\hat{\alpha}_1 = \hat{\delta}_1 + \hat{\beta}_1^{-1}\hat{\delta}_2$, where $\hat{\delta} = (W'W)^{-1}W'Y_1$.*

The proof of this result is somewhat involved and is, therefore, relegated to the Appendix, as are proofs of subsequent results, but its interpretation is simple and straightforward. The two stage least squares estimator may be viewed as a bias corrected form of the least squares estimator of the structural effect in the hybrid model (2.3).

The same strategy can be employed to estimate the conditional quantile effects in this model. We have the conditional quantile functions

$$\begin{aligned} Q_1(\tau_1|Y_2, z, x) &= Y_2(\alpha_1 + \lambda) - z\beta_1\lambda + x^\top(\alpha_2 - \lambda\beta_2) + F_1^{-1}(\tau_1) \\ Q_2(\tau_2|x, z) &= z\beta_1 + x^\top\beta_2 + F_2^{-1}(\tau_2) \end{aligned}$$

Provided that $\nabla_z Q_2(\tau_2|z, x) = \beta_1 \neq 0$ we may write *à la* Chesher (2003),

$$\begin{aligned} \alpha_1 &= \nabla_{Y_2} Q_1(\tau_1|Y_2, x, z) + \frac{\nabla_z Q_1(\tau_1|Y_2, x, z)}{\nabla_z Q_2(\tau_2|x, z)} \\ \alpha_2 &= \nabla_x Q_1(\tau_1|Y_2, x, z) - \frac{\nabla_z Q_1(\tau_1|Y_2, x, z)}{\nabla_z Q_2(\tau_2|x, z)}\nabla_x Q_2(\tau_2|x, z), \end{aligned}$$

adopting the convention that $Q_1(\tau_1|Y_2, x, z)$ is always evaluated at $Y_2 = Q_2(\tau_2|x, z)$. In this case, because the covariate effects take the simple location shift form, the structural parameters α_1 and α_2 are globally constant independent of τ_1 and τ_2 and of the exogenous variables x and z. As we will now see, this is highly unusual.

2.1. **Quantile Treatment Effects for Recursive Structural Models.** Now consider the nonlinear recursive model

$$Y_{i1} = \varphi_1(Y_{i2}, x_i, \nu_{i1}, \nu_{i2}) \tag{2.4}$$

$$Y_{i2} = \varphi_2(z_i, x_i, \nu_{i2}) \tag{2.5}$$

where as earlier we assume that ν_{i1} and ν_{i2} are independent, and identically distributed with $\nu_{ij} \sim F_j$. The pairs (ν_{i1}, ν_{i2}) are also maintained to be independent of $(z_i, x_i^\top)^\top$. The function φ_1, is assumed strictly monotonic in ν_1, and differentiable with respect to Y_2 and x, and φ_2 is assumed strictly monotonic in ν_2, and differentiable

with respect to both z and x. Under these conditions, we can write the conditional quantile functions,

$$\begin{aligned} Q_1(\tau_1|Q_2(\tau_2|x,z),x) &= \varphi_1(Q_2(\tau_2|x,z),x,F_1^{-1}(\tau_1),F_2^{-1}(\tau_2)) \\ Q_2(\tau_2|x,z) &= \varphi_2(z,x,F_2^{-1}(\tau_2)) \end{aligned}$$

How should we measure the effect of Y_2 on Y_1 in this model? Given the stochastic character of the "treatment", Y_2, we must evaluate the treatment effect at various quantiles of the treatment *distribution*. We may view this as corresponding to a thought experiment in which we exogenously alter not the value of Y_2 as we would with a treatment fully under our control, but instead alter the distribution of Y_2. Thus, for example, in our anticipated study of class-size effects on educational performance, we may imagine altering the prevailing distribution of class-sizes and exploring the consequences of this perturbation on various quantiles of the distribution of students' attainment. Of course, in the model Y_2 is determined according to (2.5), so to assume otherwise requires some sort of "willing suspension of disbelief" in the model. But this is inevitable in structural models and we are always entitled to interpret effects as long as they can be formulated in terms of well-posed *gedankan* experiments.

In their (infamous) tryptych on causal chain systems Strotz and Wold (1960) illustrate this point with a vivid fresh water example:

> Suppose z is a vector whose various elements are the amounts of various fish feeds (different insects, weeds, etc.) available in a given lake. The reduced form
>
> $$y' = B^{-1}\Gamma z' + B^{-1}u$$
>
> would tell us specifically how the number of fish of any species depends upon the availabilities of different feeds. The coefficient of any z is the partial derivative of a species population with respect to a food supply. It is to be noted, however, that the reduced form tells us nothing about the interactions among the various fish populations – it does not tell us the extent to which one species of fish feeds on another species. Those are the causal relations among the y's.
>
> Suppose, in another situation, we continuously restock the lake with species g, increasing y_g by any desired amount. How will this affect the values of the other y's? If the system were recursive and we had estimates of the elements of B, *we would simply strike the gth equation out of the model and regard y_g, the number of fish of species g, as exogenous* – as a food supply or, when appearing with a negative coefficient as a poison. (pp. 421-2, emphasis added)

Recursive conditioning enables us to contemplate similar kinds of policy experiments in the context of the triangular structural models considered by Chesher; related models have also been recently considered by Imbens and Newey (2001). In contrast to the linear structural models of the Cowles Commission era, whose causal effects were restricted to take the form of location shifts of the conditional distributions of the endogenous variables, recent work poses the identification of structural

effects in a general non-parametric framework so structural effects can take quite heterogeneous forms. We will focus on a more restricted finite dimensional parametric formulation, a formulation that is more conducive to our asymptotic analysis. Extensions to sequences of models with the parametric dimension tending to infinity could be considered in subsequent work.

To explore this further, consider the following model in which Y_2 exerts both a location and a scale shift effect on Y_1;

$$Y_{i1} = Y_{i2}\alpha_1 + x_i^\top \alpha_2 + \delta Y_{i2}(\nu_{i1} + \lambda\nu_{i2}) \tag{2.6}$$

$$Y_{i2} = z_i\beta_1 + x_i^\top \beta_2 + \gamma z_i \nu_{i2} \tag{2.7}$$

Maintaining our prior assumptions on (ν_{i1}, ν_{i2}), and assuming that $\delta \neq 0$ and $\gamma \neq 0$, we can again substitute for ν_{i2} in (2.6) to obtain,

$$Y_{i1} = Y_{i2}(\alpha_1 + \delta\nu_{i1} - \delta\beta_1\lambda/\gamma) + x_i^\top\alpha_2 + \frac{Y_{i2}^2}{z_i}\left(\frac{\delta\lambda}{\gamma}\right) - \frac{Y_{i2}x_i^\top}{z_i}\left(\frac{\delta\lambda\beta_2}{\gamma}\right)$$

$$Y_{i2} = z_i(\beta_1 + \gamma\nu_{i2}) + x_i^\top\beta_2$$

and again by recursive conditioning we have the conditional quantile functions,

$$\begin{aligned} Q_1(\tau_1|Q_2(\tau_2|x,z),x,z) &= Q_2(\tau_2|x,z)(\alpha_1 + \delta(F_1^{-1}(\tau_1) + \lambda F_2^{-1}(\tau_2))) + x^\top\alpha_2 \\ Q_2(\tau_2|x,z) &= z(\beta_1 + \gamma F_2^{-1}(\tau_2)) + x^\top\beta_2 \end{aligned}$$

The structural effect of interest is therefore

$$\pi_1(\tau_1,\tau_2) = \alpha_1 + \delta(F_1^{-1}(\tau_1) + \lambda F_2^{-1}(\tau_2)).$$

Given the separate contributions of $F_1^{-1}(\tau_1)$ and $F_2^{-1}(\tau_2)$, it is clear that $\pi(\tau_1,\tau_2)$ reflects not only the fact that the stochastic effect of Y_2 on Y_1 arises from two distinct sources, but also provides structural insight into how these sources are related. Suppose we fix τ_1 so ν_1 is fixed at its τ_1 quantile, changes in τ_2 in $\pi_1(\tau_1,\tau_2)$ reflect how the distribution of ν_2 affects the τ_1 quantile of the response Y_1. On the other hand, if we fix τ_2, and allow τ_1 to change, this sheds light on how the τ_2 quantile of Y_2 influences the whole distribution of the response Y_1. By considering variation in both τ_1 and τ_2 we obtain a panoramic view of the stochastic relationship between Y_2 and Y_1.

Recalling that integrating the quantile function $F_X^{-1}(\tau)$ of a random variable, X, over the domain $[0,1]$, yields its expectation, that is,

$$EX = \int_0^1 F_X^{-1}(t)dt,$$

we can define a mean quantile treatment effect by integrating out τ_2, and denoting $\mu_i = E\nu_i$,

$$\bar{\pi}_1(\tau_1) = \int_0^1 (\alpha_1 + \delta(F_1^{-1}(\tau_1) + \lambda F_2^{-1}(\tau_2)))d\tau_2 \equiv \alpha_1 + \delta F_1^{-1}(\tau_1) + \delta\lambda\mu_2$$

Averaging again, this time with respect to τ_1 yields the mean treatment effect

$$\bar{\pi}_1 = \int_0^1 (\alpha_1 + \delta F_1^{-1}(\tau_1) + \delta\lambda\mu_2)d\tau_1 \equiv \alpha_1 + \delta\mu_1 + \delta\lambda\mu_2.$$

This mean treatment effect would be what is estimated by the two stage least squares estimator in the pure location shift version of the model, but when the effects are more heterogeneous as in this location-scale shift model the structural quantile treatment effect $\pi_1(\tau_1, \tau_2)$ represents a deconstruction the mean effect into its elementary components. Figure 2.1 illustrates the three versions of the treatment effect $\pi_1(\tau_1, \tau_2), \bar{\pi}_1(\tau_1)$ and $\bar{\pi}_1$ for a particular parametric instance of the model (2.6-7).

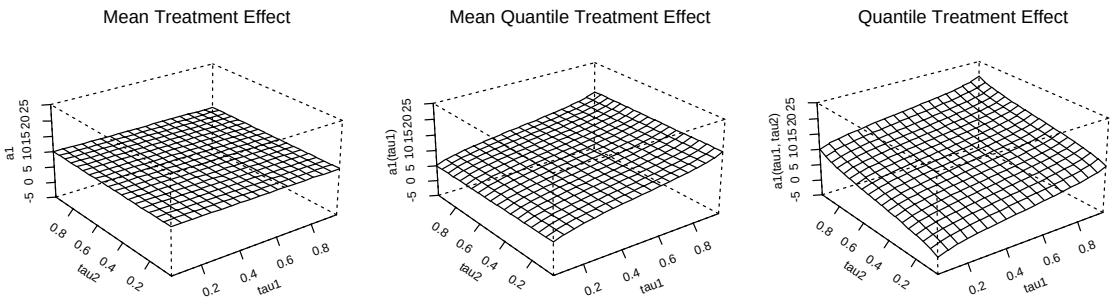

FIGURE 2.1. Quantile Treatment Effects for the Structural Model: The figure illustrate three different notions of the structural treatment effect for the linear location-scale structural equation model: (2.6-7) with $(\alpha_1, \alpha_2, \delta, \lambda) = (10, 4, 3, 2)$, $(\beta_1, \beta_2, \gamma) = (1, 2, 3)$, $\nu_1 \sim N(0, 1)$, $\nu_2 \sim N(0, 0.5)$. The left figure depicts $\bar{\pi}_1$ =10, the mean treatment effect; the middle figure shows $\bar{\pi}_1(\tau_1) = 10 + 3F_1^{-1}(\tau_1)$, the mean quantile treatment effect; the right figure shows $\pi_1(\tau_1, \tau_2) = 10 + 3(F_1^{-1}(\tau_1) + 2F_2^{-1}(\tau_2))$, the general quantile treatment effect.

2.2. **Estimation of Structural Quantile Treatment Effects.** In this section we will describe two general classes of estimators for the parametric recursive structural model,

$$\begin{aligned} Y_{i1} &= \varphi_1(Y_{i2}, x_i, \nu_{i1}, \nu_{i2}; \alpha) \\ Y_{i2} &= \varphi_2(z_i, x_i, \nu_{i2}; \beta) \end{aligned}$$

We will maintain our assumptions on the ν_{ij}'s and the functions φ_1 φ_2 and we will explicitly assume that the functions φ_1 and φ_2 are known up to the finite dimensional parameter vectors α and β. Under these conditions we have an inverse function for φ_2 with respect to ν_2, say $\tilde{\varphi}_2$, allowing us to write

$$\nu_{i2} = \tilde{\varphi}_2(Y_{i2}, z_i, x_i; \beta)$$

and thus we have,

$$Y_{i1} = \varphi_1(Y_{i2}, x_i, \nu_{i1}, \tilde{\varphi}_2(Y_{i2}, z_i, x_i; \beta); \alpha).$$

We will write the conditional quantile functions of Y_1 and Y_2 as,

$$\begin{aligned} Q_1(\tau_1|Y_{i2}, x_i, z_i) &= h_1(Y_{i2}, x_i, z_i; \theta) \\ Q_2(\tau_2|z_i, x_i) &= h_2(z_i, x_i; \beta). \end{aligned}$$

Fixing τ_1 and τ_2 we can estimate $\theta(\tau_1)$ and $\beta(\tau_2)$ by solving the possibly nonlinear weighted quantile regression problems,

$$\hat{\theta}(\tau_1) = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \sigma_{i1} \rho_{\tau_1}(Y_{i1} - h_1(Y_{i2}, x_i, z_i; \theta) \tag{2.8}$$

$$\hat{\beta}(\tau_2) = \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \sum_{i=1}^{n} \sigma_{i2} \rho_{\tau_2}(Y_{i2} - h_2(z_i, x_i, \beta)) \tag{2.9}$$

The weights σ_{ij} are assumed to be strictly positive and will play a important role in the efficiency comparisons made in Section 4. The function $\rho_\tau(u) = u(\tau - I(u < 0))$ is as in Koenker and Bassett (1978). Methods for computing quantile regression estimates for models that are nonlinear in parameters are described in Koenker and Park (1996). When h_1 and h_2 yield specifications that are nonlinear in parameters, then we may be required to specify compact domains Θ and $\mathcal{B}$.

Our primary objective will be to estimate the weighted average quantile treatment effect implied by the Chesher formula,

$$\pi_1(\tau_1, \tau_2) = \int \left\{ \nabla_y Q_1(\tau_1|y, x_i, z_i) + \frac{\nabla_z Q_1(\tau_1|y, x_i, z_i)}{\nabla_z Q_2(\tau_2|x_i, z_i)} \right\} w(x, z) dx dz$$

with y evaluated as before, at $Q_2(\tau_2|x_i, z_i)$. A secondary object will be to estimate the corresponding structural effect of the the exogonous variables x,

$$\pi_2(\tau_1, \tau_2) = \int \left\{ \nabla_x Q_1(\tau_1|Y_{i2}, x_i, z_i) - \frac{\nabla_z Q_1(\tau_1|y, x_i, z_i)}{\nabla_z Q_2(\tau_2|x_i, z_i)} \nabla_x Q_2(\tau_2|x_i, z_i) \right\} w(x, z) dx dz$$

Since, in general, the above integrands depend upon the point of evaluation in the space of the exogenous covariates we consider the class of weighted average derivative estimators,

$$\hat{\pi}_1(\tau_1, \tau_2) = \sum_{i=1}^{n} w_i \left\{ (\nabla_y \hat{h}_1(\tau_1|y, x_i, z_i, \hat{\theta}) + \frac{\nabla_z \hat{h}_1(\tau_1|y, x_i, z_i, \hat{\theta})}{\nabla_z \hat{h}_2(\tau_2|x_i, z_i.\hat{\beta})} \right\}$$

again evaluating at $y = \hat{h}_2(\tau_2|x_i, z_i, \hat{\beta})$. A weighted average derivative estimator for the structural effects of x is defined similarly as,

$$\hat{\pi}_2(\tau_1, \tau_2) = \sum_{i=1}^{n} w_i \left\{ \nabla_x \hat{h}_1(\tau_1|y, x_i, z_i, \hat{\theta}) - \frac{\nabla_z \hat{h}_1(\tau_1|y, x_i, z_i, \hat{\theta})}{\nabla_z \hat{h}_2(\tau_2|x_i, z_i, \hat{\beta})} \nabla_x \hat{h}_2(\tau_2|x_i, z_i, \hat{\beta}) \right\}$$

The weights are assumed to be positive and sum to one. A convenient choice would be $w_i \equiv n^{-1}$. In some cases, like the location shift model the dependence on the exogenous covariates vanishes so the weights are irrelevant. The foregoing considerations have presumed a situation of exact identification in which there is a single "instrumental variable," z, available. In over-identified settings we may have several

versions of $\hat{\pi}(\tau_1, \tau_2)$ corresponding to different choices of the variable z and we may wish to again consider weighted averages. This point will be addressed in more detail when we come to asymptotics.

The estimator $\hat{\pi}_n(\tau_1, \tau_2) = (\hat{\pi}_1(\tau_1, \tau_2), \hat{\pi}_2^\top(\tau_1, \tau_2))^\top$ is based squarely on Chesher's identification strategy. Its advantage is that it takes a rather skeptical attitude toward the original model and is thereby based on a rather loosely restricted form of the two conditional quantile functions. This complements nicely the more restrictive form of the estimators described in the next subsection and consequently may eventually prove to be advantageous from a specification diagnostics and testing viewpoint.

2.3. **A Control Variate Estimator.** To motivate the control variate approach to estimation of the structural quantile treatment effect, it is helpful to return briefly to the classical two stage least squares estimator of the location shift model (2.1-2) and recall its control variate interpretation. Suppose that rather than replacing Y_2 by $\hat{Y}_2$ in (2.1) and estimating the resulting model by least squares, we instead compute $\hat{\nu}_2 = Y_2 - \hat{Y}_2$, the residuals from the first stage of 2SLS. Now consider including $\hat{\nu}_2$ as an additional covariate in (2.1) and estimating by least squares. It is easy to show that the resulting estimates of α_1 and α_2 are the same as those produced by 2SLS. This result holds much more generally: Y_{i2} and z_i may be vector-valued and the model may be overidentified. A definitive original reference for this equivalence is however difficult to identify, see for example, Blundell and Powell (2003).

To apply the control variate approach to the estimation of the structural quantile treatment effect we must first estimate the conditional τ_2 quantile function of Y_2 to recover an estimate of $\nu_2(\tau_2) = \nu_2 - F_2^{-1}(\tau_2)$. Let

$$Q_1(\tau_1 | Y_{i2}, x_i, \nu_{i2}(\tau_2)) = g_1(Y_{i2}, x_i, \nu_{i2}(\tau_2); \alpha(\tau_1, \tau_2))$$
$$Q_2(\tau_2 | z_i, x_i) = g_2(z_i, x_i; \beta(\tau_2))$$

denote the conditional quantile functions of the response variables conditioning on the control variate, $\nu_{i2}(\tau_2)$. Solving

$$\hat{\beta}(\tau_2) = \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \sum \sigma_{i2} \rho_{\tau_2}(Y_{i2} - g_2(z_i, x_i; \beta))$$

our conditions on φ_2 insure that we can invert to obtain the function

$$\nu_2 = \tilde{\varphi}_2(Y_2, z, x, \beta)$$

so

$$F_2^{-1}(\tau_2) = \tilde{\varphi}_2(g_2(z, x; \beta), z, x; \beta)$$

and we have

$$\hat{\nu}_{i2}(\tau_2) = \tilde{\varphi}_2(Y_{i2}, z_i, x_i; \hat{\beta}) - \tilde{\varphi}_2(g_2(z_i, x_i; \hat{\beta}), z_i, x_i; \hat{\beta})$$

Note that the above procedure is valid regardless of the dimension of z_i, so as long as the model is identifiable $\hat{\nu}_{i2}(\tau_2)$ incorporates information on all of the available instruments. But it does so in a much more parsimonious fashion than by introducing z_i directly into what we have referred to as the hybrid form of the first structural equation.

Once $\hat{\nu}_{i2}(\tau_2)$ is available we can estimate the parameters of the first structural equation by reexpressing φ_1 as

$$g_1(Y_{i2}, x_i, \hat{\nu}_{i2}(\tau_2); a) = \varphi_1(Y_{i2}, x_i, F_1^{-1}(\tau_1), \hat{\nu}_{i2}(\tau_2); \alpha)$$

absorbing $F_1^{-1}(\tau_1)$ into the new parameter vector a, and solving,

$$\hat{\alpha}(\tau_1, \tau_2) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \sum_{i=1}^{n} \sigma_{i1} \rho_{\tau_1}(Y_{i1} - g_1(Y_{i2}, x_i, \hat{\nu}_{i2}(\tau_2); a)).$$

In the next section we will investigate the asymptotic behavior of this control variate estimator and compare its asymptotic performance with the weighted average derivative estimator. Before doing so we might remark that the restrictions imposed by the control variate procedure avoid the considerable complications of the weighted average derivative method apparent in the location-scale model (2.6-7).

2.4. **Extension to m Equations.** As shown by Chesher (2003) there are no real impediments to the extension of the recursive structural model to more than two equations, except some obvious notational ones. Maintaining the triangular structure we may consider the system of m structural equations,

$$\begin{aligned} Y_1 &= \varphi_1(Y_2, ..., Y_m, z, \nu_1, ..., \nu_m, \alpha_1) \\ Y_2 &= \varphi_2(Y_3, ..., Y_m, z, \nu_2, ..., \nu_m, \alpha_2) \\ &\vdots \\ Y_m &= \varphi_m(z, \nu_m, \alpha_m). \end{aligned}$$

The ν's are assumed stochastically independent and independent of the exogonous variables, z. Again, we can recursively condition to obtain the conditional quantile functions of the Y's and this leads to a natural generalization of the weighted average derivative estimators. Chesher (2003) describes the exclusion restrictions and other conditions required for identification in this case.

Similarly, we can adapt the control variate estimation method to the multiple equation setting. The estimation strategy is a quite straightforward extension of the two equation situation. Starting with the last equation we estimate the control variate $\hat{\nu}_m(\tau_m)$ and substitute it into the $(m-1)$th equation, thus obtaining the control variate $\hat{\nu}_{m-1}(\tau_{m-1})$, and so forth. The asymptotic representation also generalizes in a straightforward fashion so that for the first equation, for example, we obtain a sum of m independent terms in the Bahadur representation.

3. ASYMPTOPIA

The asymptotic behavior of the estimators described in the previous section can be developed with the aid of existing results on the asymptotics of nonlinear (in parameters) quantile regression estimation. We will employ the following regularity conditions; See, e.g., Oberhofer (1982) and Jurečková and Procházka(1994).

A.1: The conditional distribution functions $F_{Y_1}(y_1|Y_{i2}, x_i, z_i)$ and $F_{Y_2}(y_2|z_i, x_i)$ are absolutely continuous with continuous densities f_{i1} and f_{i2} that are uniformly bounded away from 0 and ∞ at the points $\xi_{i1} = Q_1(\tau_1|Q_2(\tau_2|z_i, x_i), x_i, z_i)$ and $\xi_{i2} = Q_2(\tau_2|z_i, x_i)$, for $i = 1, \ldots, n$. The weights σ_{ij} are positive and uniformly bounded away from 0 and ∞.

A.2: There exist positive definite matrices $J_1, \bar{J}_1, J_2, \bar{J}_2$ such that

$$\lim_{n\to\infty} n^{-1} \sum \sigma_{ij}^2 \dot{h}_{ij} \dot{h}_{ij}^\top = J_j, \quad \lim_{n\to\infty} n^{-1} \sum \sigma_{ij} f_{ij}(\xi_{ij}) \dot{h}_{ij} \dot{h}_{ij}^\top = \bar{J}_j,$$

where $\dot{h}_{i1} = \nabla_\theta h_{i1}$ and $\dot{h}_{i2} = \nabla_\beta h_{i2}$.

A.3: $\max_{i=1,\ldots,n} \| \dot{h}_{ij} \| / \sqrt{n} \to 0, \quad j = 1, 2.$

A.4: There exist constants l_1, l_2, u_1, u_2 and an integer $n_0 > 0$ such that for $(\theta_j, \theta_j') \subset \Theta$, $(\beta_j, \beta_j') \subset \mathcal{B}$, $j = 1, 2$ and $n > n_0$,

$$\begin{aligned} l_1 \| \theta - \theta' \| &\leq (n^{-1} \sum (h_1(Y_{i2}, x_i, z_i, \theta) - h_1(Y_{i2}, x_i, z_i, \theta')^2)^{1/2} \leq u_1 \| \theta - \theta' \| \\ l_2 \| \beta - \beta' \| &\leq (n^{-1} \sum (h_2(x_i, z_i, \beta) - h_2(x_i, z_i, \beta')^2)^{1/2} \leq u_2 \| \beta - \beta' \| \end{aligned}$$

Theorem 1. *For the parametric model (2.4-5) satisfying conditions A.1-4, the weighted average derivative estimator $\hat{\pi}_n(\tau_1, \tau_2)$ has the asymptotic linear (Bahadur) representation*

$$\begin{aligned} \sqrt{n}(\hat{\pi}_n(\tau_1, \tau_2) - \pi(\tau_1, \tau_2)) &= W_1 \bar{J}_1^{-1} n^{-1/2} \sum_{i=1}^n \sigma_{i1} \dot{h}_{i1} \psi_{\tau_1}(Y_{i1} - \xi_{i1}) \\ &+ W_2 \bar{J}_2^{-1} n^{-1/2} \sum_{i=1}^n \sigma_{i2} \dot{h}_{i2} \psi_{\tau_2}(Y_{i2} - \xi_{i2}) + o_p(1) \\ &\rightsquigarrow \mathcal{N}(0, \omega_{11} W_1 \bar{J}_1^{-1} J_1 \bar{J}_1^{-1} W_1^\top + \omega_{22} W_2 \bar{J}_2^{-1} J_2 \bar{J}_2^{-1} W_2^\top) \end{aligned}$$

where $\omega_{jj} = \tau_j(1 - \tau_j)$, $W_1 = \nabla_\theta \pi(\tau_1, \tau_2)$ and $W_2 = \nabla_\beta \pi(\tau_1, \tau_2)$.

Remark: It is immediately apparent that the optimal choice of the weights, σ_{ij} involves setting $\sigma_{ij} = f_{ij}(\xi_{ij})$. In this case the sandwich form of the limiting covariance matrix simplifies, and we have

$$\sqrt{n}(\hat{\pi}_n(\tau_1, \tau_2) - \pi(\tau_1, \tau_2)) \rightsquigarrow \mathcal{N}(0, \omega_{11} W_1 J_1^{-1} W_1^\top + \omega_{22} W_2 J_2^{-1} W_2^\top)$$

We will not address the somewhat delicate issues involved in estimating weights, but the interested reader could consult Koenker and Zhao (1994) and/or Zhao (2001). ■

Example: Recall that in the pure location shift version of the model (2.1-2) the structural effect $\pi_1(\tau_1, \tau_2)$ is a constant α_1. In this case we have model (2.1-2) and $\sqrt{n}(\hat{\alpha}_1(\tau_1, \tau_2) - \alpha_1)$ is asymptotically Gaussian with mean 0 and variance

$$v = \left(\frac{\tau_1(1 - \tau_1)}{f_1^2(\xi_1)} + \lambda^2 \frac{\tau_2(1 - \tau_2)}{f_2^2(\xi_2)} \right) v_0$$

where $v_0 = \lim_{n\to\infty} n^{-1} \beta_1' Z' M_X Z \beta_1$, and $M_X = I - X(X'X)^{-1}X'$. The parameter λ may be interpreted as a degree of endogeneity of the model, so the second term in v may be viewed as a performance penalty for this endogeneity effect. It may

be noted that under these special conditions the estimator $\hat{\alpha}_1(\tau_1, \tau_2)$ is equivalent to the so-called two-stage quantile regression estimator which replaces Y_2 in (2.1) by $\hat{Y}_2(\tau_2)$ the fitted values in the τ_2 quantile regression estimate of (2.2) and then estimates the τ_1 quantile regression of Y_1 on $\hat{Y}_2(\tau_2)$ and x. A special case of this procedure is Amemiya's two stage least absolute deviations estimator. To the best of our knowledge no general analysis of its asymptotic behavior has been undertaken although it has been employed in several empirical studies. ■

To study the asymptotic behavior of the control variate estimators we require a slightly modified version of our previous regularity conditions.

B.1: The conditional distribution functions $F_{Y_{i1}|Y_{i2},x_i,\nu_{i2}}$ and $F_{Y_{i2}|z_i,x_i}$ are absolutely continuous with continuous densities f_{i1} and f_{i2} uniformly bounded away from 0 and ∞ at the points $\xi_{i1} = Q_1(\tau_1|Y_{i2}, z_i, x_i), x_i, \nu(\tau_2))$ and $\xi_{i2} = Q_2(\tau_2|z_i, x_i)$, respectively for $i = 1, 2, \ldots, n$. The weights σ_{ij} are positive and uniformly bounded away from 0 and ∞.

B.2: There exist positive definite matrices $D_1, \bar{D}_1, D_2, \bar{D}_2$ such that

$$\lim_{n\to\infty} n^{-1} \sum \sigma_{ij}^2 \dot{g}_{ij} \dot{g}_{ij}^\top = D_j, \quad \lim_{n\to\infty} n^{-1} \sum \sigma_{ij} f_{ij}(\xi_{ij}) \dot{g}_{ij} \dot{g}_{ij}^\top = \bar{D}_j,$$

where $\dot{g}_{i1} = \nabla_\alpha g_{i1}$ and $\dot{g}_{i2} = \nabla_\beta g_{i2}$.

B.3: $\max_{i=1,\ldots,n} \| \dot{g}_{ij} \| / \sqrt{n} \to 0, \quad j = 1, 2.$

B.4: There exist constants l_1, l_2, u_1, u_2 and an integer $n_0 > 0$ such that such that, for α, $\alpha' \in \mathcal{A}$, β, $\beta' \in \mathcal{B}$ and $n > n_0$,

$$l_1||\alpha - \alpha'|| \leq (n^{-1} \sum_{i=1}^n (g_1(Y_{i2}, x_i, \nu_{i2}(\tau_2), \alpha) - g_1(Y_{i2}, x_i, \nu_{i2}(\tau_2), \alpha'))^2)^{1/2} \leq u_1||\alpha - \alpha'||$$

$$l_2||\beta - \beta'|| \leq (n^{-1} \sum_{i=1}^n (g_2(x_i, z_i, \beta) - g_2(x_i, z_i, \beta))^2)^{1/2} \leq u_2||\beta - \beta'||.$$

These conditions are the natural analogues of our previous conditions. It may be noted that in contrast to the prior conditions, however, the possibility of overidentification is now permitted by the modified conditions. We can now describe the asymptotic behavior of the control variate estimator.

Theorem 2. *For the parametric model (2.4-5) satisfying conditions B.1-4, the control variate estimator $\hat{\alpha}_n(\tau_1, \tau_2)$ has the asymptotic linear (Bahadur) representation,*

$$\begin{aligned}
\sqrt{n}(\hat{\alpha}_n(\tau_1, \tau_2) - \alpha(\tau_1, \tau_2)) &= \bar{D}_1^{-1} n^{-1/2} \sum_{i=1}^n \sigma_{i1} \dot{g}_{i1} \psi_{\tau_1}(Y_{i1} - \xi_{i1}) \\
&+ \bar{D}_1^{-1} \bar{D}_{12} \bar{D}_2^{-1} n^{-1/2} \sum_{i=1}^n \sigma_{i2} \dot{g}_{i2} \psi_{\tau_2}(Y_{i2} - \xi_{i2}) + o_p(1) \\
&\rightsquigarrow \mathcal{N}(0, \omega_{11} \bar{D}_1^{-1} D_1 \bar{D}_1^{-1} + \omega_{22} \bar{D}_1^{-1} \bar{D}_{12} \bar{D}_2^{-1} D_2 \bar{D}_2^{-1} \bar{D}_{12}^\top \bar{D}_1^{-1})
\end{aligned}$$

where $\bar{D}_{12} = \lim_{n\to\infty} n^{-1} \sum \sigma_{i1} f_{i1} \eta_i \dot{g}_{i1} \dot{g}_{i2}^\top$ and $\eta_i = (\partial g_{1i}/\partial \nu_{i2}(\tau_2))(\nabla_{\nu_{i2}} \varphi_{i2})^{-1}$.

Remark: Again, we see that the choice of the weights $\sigma_{ij} = f_{ij}(\xi_{ij})$ is optimal. It may appear that the use of symbols σ_{ij} for the weights for both classes of estimators is an abuse of notation, but careful examination of the conditioning reveals that the conditional densities are identical in conditions A.1 and B.1 so this economy is justified at least in the two cases of primary interest: weights identically equal to one, and optimally weighted estimation according to the conditional densities. ∎

For purposes of inference it is crucial that we have not only the marginal distribution of $\hat{\alpha}_n$ for fixed τ_1 and τ_2, but also the joint distribution of $\hat{\alpha}_n$ evaluated at several τ_1's and τ_2's. But this follows immediately from the Bahadur representation of the preceding theorem.

Corollary 1. *Let $\mathcal{T}_1 = \{\tau_{11}, ...\tau_{1q}\}$ and $\mathcal{T}_2 = \{\tau_{21}, ...\tau_{2r}\}$ with elements $\tau_{ij} \in (0,1)$, then under the conditions of Theorem 2, the joint asymptotic distribution of $\{\hat{\alpha}_n(\tau_1, \tau_2) : \tau_1 \in \mathcal{T}_1,\ \tau_2 \in \mathcal{T}_2\}$ is Gaussian with typical covariance block,*

$$Acov\big(\sqrt{n}\hat{\alpha}(\tau_1,\tau_2),\ \sqrt{n}\hat{\alpha}(\tau_3,\tau_4)\big) = \omega_{13}\bar{D}_1^{-1}D_{13}\bar{D}_3^{-1} + \omega_{24}\bar{D}_1^{-1}\bar{D}_{12}\bar{D}_2^{-1}D_{24}\bar{D}_4^{-1}\bar{D}_{34}^{\top}\bar{D}_3^{-1},$$

where $D_{rs} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \sigma_{ir}\sigma_{is}\dot{g}_{ir}\dot{g}_{is}^{\top}$, $\omega_{rs} = \min\{\tau_r, \tau_s\} - \tau_r\tau_s$, with $\{\tau_1, \tau_3\} \subset \mathcal{T}_1$ and $\{\tau_2, \tau_4\} \subset \mathcal{T}_2$.

4. Asymptotic Relative Efficiency of the Structural Estimators

Naturally, we would like to compare the performance of our two classes of estimators. The first and most obvious prerequisite for this is to ensure that they are really estimating the same quantity. For linear in parameters specifications the situation is quite straightforward so we will consider this case in some detail first, treating it as a rehearsal for the general result embodied in Theorem 4. To formalize what we mean by linear models, suppose that

$$\varphi_1(Y_{i2}, x_i, \nu_{i2}, \alpha, F_1^{-1}(\tau_1)) = \dot{g}_{i1}^{\top}\alpha(\tau_1,\tau_2) = \dot{h}_{i1}^{\top}\theta(\tau_1) \tag{4.1}$$

$$\varphi_2(z_i, x_i, F_2^{-1}(\tau_2), \beta) = \dot{g}_{i2}^{\top}\beta(\tau_2) = \dot{h}_{i2}^{\top}\beta(\tau_2) \tag{4.2}$$

where the vectors $\dot{g}_{ij}$ and $\dot{h}_{ij}$ are free of dependence on the parameters. The linearity of φ_1 implies that there is a linear mapping, $W_1 = \partial\pi/\partial\theta$, such that

$$W_1\theta = \pi.$$

Writing G_j for the matrix with typical row $n^{-1/2}(f_{ij}\dot{g}_{ij}^{\top})$ for $j = 1, 2$, and similarly let H_j denote the matrix with typical row $n^{-1/2}(f_{ij}\dot{h}_{ij}^{\top})$. Note that $G_2 = H_2$ and that there is a matrix A such that $G_1 = H_1A$ so $A\alpha = \theta$. Thus we have $W_1A\alpha = \pi$. Further, let $L = W_1A$, so $L\alpha = \pi$. The transformation L reduces the dimensionality of the α vector, eliminating the components that are required to describe the ν_2-effect and allowing us to focus attention on the performance of the control variate estimator of the π parameter.

We can now compare the performance of our two estimators of π: the weighted average derivative estimator $\hat{\pi}_n$ and the control variate estimator $\tilde{\pi}_n = L\hat{\alpha}_n$. To facilitate this comparison it is convenient to restrict attention to the optimally weighted

form of both estimators for which $\sigma_{ij} = f_{ij}$. In this case, the asymptotic covariance matrix of $\hat{\pi}_n$ specializes to

$$\text{Avar}(\sqrt{n}\hat{\pi}_n) = \omega_{11} W_1 J_1^{-1} W_1^\top + \omega_{22} W_2 J_2^{-1} W_2^\top$$

while that of $\hat{\alpha}_n$ specializes to

$$\text{Avar}(\sqrt{n}\hat{\alpha}_n) = \omega_{11} D_1^{-1} + \omega_{22} D_1^{-1} D_{12} D_2^{-1} D_{12} D_1^{-1}$$

where $D_i^{-1} = \lim_{n\to\infty} n^{-1} \sum f_{ij}^2 \dot{g}_{ij} \dot{g}_{ij}^\top$ and $D_{12} = \lim_{n\to\infty} n^{-1} \sum f_{i1}^2 \eta_i \dot{g}_{i1} \dot{g}_{i2}^\top$. Equivalently, we can write,

$$\text{Avar}(\sqrt{n}\hat{\alpha}_n) = \omega_{11} (G_1^\top G_1)^{-1} + \omega_{22} (G_1^\top G_1)^{-1} G_1 P_{G_2} G_1^\top (G_1^\top G_1)^{-1}$$

where P_G generically denotes the projection $G(G^\top G)^{-1} G^\top$ onto the column space of the matrix G. Thus, $\tilde{\pi} = L\hat{\alpha}$, we have,

$$\text{Avar}(\sqrt{n}\tilde{\pi}) = \omega_{11} L (G_1^\top G_1)^{-1} L^\top + \omega_{22} L (G_1^\top G_1)^{-1} G_1 P_{G_2} G_1^\top (G_1^\top G_1)^{-1} L^\top$$

Note that

$$\begin{aligned} L(G_1^\top G_1)^{-1} L^\top &= W_1 A (A^\top H_1^\top H_1 A)^{-1} A^\top W_1^\top \\ &= W_1 J_1^{-1} H_1^\top H_1 A (A^\top H_1^\top H_1 A)^{-1} A H_1^\top H_1 J_1^{-1} W_1^\top \\ &= W_1 J_1^{-1} H_1^\top P_{G_1} H_1 J_1^{-1} W_1^\top \\ &\leq W_1 J_1^{-1} W_1^\top, \end{aligned}$$

where $\leq$ signifies the conventional ordering of matrices in the sense of positive definite differences. Similarly, we have,

$$L(G_1^\top G_1)^{-1} G_1 P_{G_2} G_1^\top (G_1^\top G_1)^{-1} L^\top \leq W_2 J_2^{-1} W_2^\top,$$

so we have established that the control variate estimator, $\tilde{\pi}_n$, has smaller asymptotic variance than the weighted average derivative estimator $\hat{\pi}_n$.

The efficiency advantage of the control variate estimator clearly derives from the more restricted form of the estimator. While the restricted form of the $\tilde{\pi}_n$ estimator yields an efficiency gain when we are confident about the model specification, it clearly offers some disadvantages in situations in which we are not so confident. Indeed, tests of model specification based on the unrestricted form of the estimators $(\hat{\theta}_n, \hat{\beta}_n)$ might be viewed as a reasonable precaution in the early stages of model construction.

When the model is nonlinear in parameters the situation is much the same from an asymptotic viewpoint. Jacobians of the nonlinear transformations, W_1, A, and L evaluated at the true parameters now play the role of the matrices in the previous development, and the δ-method yields the following general result.

Theorem 3. *For the parametric model (2.4-5) with the optimal weighting, $\sigma_{ij} = f_{ij}$, let $\Lambda(\alpha) = \pi$ denote the mapping from the structural parameter α to the weighted average derivative parameter π. Suppose that the Jacobian, $L = \partial\Lambda/\partial\alpha$ is continuous in a neighborhood of the true parameters. Then the optimally-weighted average derivative estimator, $\hat{\pi}_n$, and the optimally-weighted control variate estimator, $\tilde{\pi}_n = \Lambda(\hat{\alpha}_n)$, have limiting Gaussian behavior with asymptotic covariance matrices:*

$$Avar(\sqrt{n}\hat{\pi}_n) = \omega_{11} W_1 J_1^{-1} W_1^\top + \omega_{22} W_2 J_2^{-1} W_2^\top$$

$$Avar(\sqrt{n}\tilde{\pi}) = \omega_{11} L(G_1^\top G_1)^{-1} L^\top + \omega_{22} L(G_1^\top G_1)^{-1} G_1 P_{G_2} G_1^\top (G_1^\top G_1)^{-1} L^\top$$

and $Avar(\sqrt{n}\tilde{\pi}_n) \leq Avar(\sqrt{n}\hat{\pi})$.

Remark: It is worth emphasizing at this point that the superior asymptotic performance of the control variate estimator asserted in Theorem 3 is particularly appealing when the model is overidentified. In such cases the weighted average derivative approach becomes somewhat cumbersome, while the control variate method remains entirely straightforward. ■

5. CAUTION: QUANTILES CROSSING

Monotonicity of the conditional distribution function, $F(y|x)$ in y implies that the associated conditional quantile functions must be monotone increasing in their argument τ. Some caution must be exercised as a consequence, particularly in simulation experiments, to insure that this condition is met. To illustrate the problem consider a simple location-scale shift model,

$$y_i = \beta_0 + x_i\beta_1 + (\gamma_0 + \gamma_1 x_i)\nu_i. \tag{5.1}$$

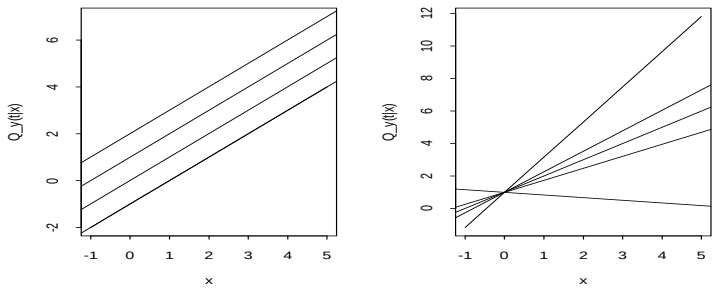

FIGURE 5.1. Non-Crossing vs Crossing: crossing occurs at $x = 0$ for model $Q_y(\tau|x) = 1 + x\beta(\tau)$, $x \in [-1, 5]$, in the right plot.

Suppose that the unobserved ν_i are iid with distribution function F_ν. When the scale parameter $\gamma_1 = 0$, then we have a pure location shift model and the family of conditional quantile functions are parallel as depicted in the left panel of Figure 5.1. When $\gamma_0 = 0$ and $\gamma_1 > 0$ we have a family of condition quantile functions that all pass through the point $(0, \beta_0)$ as illustrated in the right panel. This is fine as long as we contemplate using the model in region to the right of the point of intersection. If for example we know that the x_i's will be strictly positive we may have some compelling reason to believe that the response variable has a degenerate distribution at $x = 0$ and the model is perfectly appropriate. If, however, venturing into more dangerous territory, we impose the condition that $\gamma_0 = 0$ and then the x_i's take both positive and negative values then the quantile functions are no longer linear. If ν has quantile

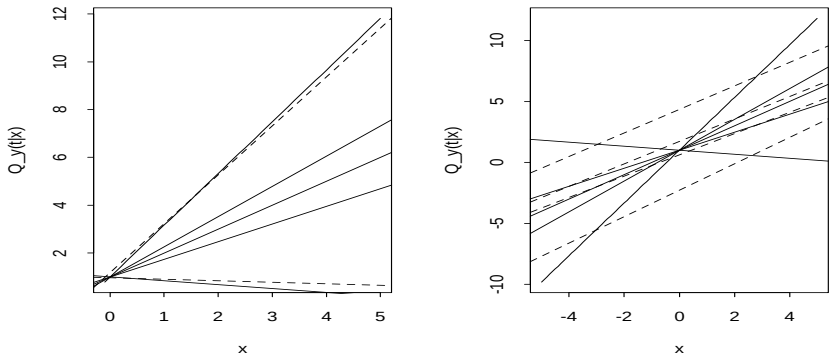

FIGURE 5.2. Degree of Distortion: $Q_y(\tau|x) = 1 + x(1 + F_\nu^{-1}(\tau))$, $\nu \sim N(0, 0.5^2)$, $\tau \in \{0.01, 0.3, 0.5, 0.7, 0.99\}$. The crossing occurs at $x = 0$. The dashed lines are the "distorted" fitted curves. In the left panel we generate x as uniform on the interval $[0.1, 5]$, and illustrate the fitted linear quantile functions for $\tau = .01$ and $\tau = .99$, based on a sample of $n = 100$ observations. We see that there is little distortion even at the extreme tails. In the right panel we repeat the exercise with $x \sim U[-5, 5]$, so crossing occurs at the middle of the domain. Now all the fitted lines are distorted except at the median, and the fitted quantile functions mimic an iid error model with parallel quantile functions even though the true conditional quantile functions are highly nonlinear.

function $F_\nu^{-1}(\tau)$ then $-\nu$ has quantile function, $-F_\nu^{-1}(1-\tau)$, so the quantile functions may be written in the piecewise linear form,

$$Q_{y_i}(\tau|x_i) = \begin{cases} \beta_0 + x_i\beta_1 + (\gamma_0 + \gamma_1 x_i)F_\nu^{-1}(\tau) & \text{if} \gamma_0 + \gamma_1 x_i \leq 0 \\ \beta_0 + x_i\beta_1 + (\gamma_0 + \gamma_1 x_i)F_\nu^{-1}(1-\tau) & \text{if} \gamma_0 + \gamma_1 x_i > 0 \end{cases} \tag{5.2}$$

If we persist in fitting linear models in the face of this piecewise linearity of the true model, we can be badly misled. In Figure 5.2 we illustrate two cases: in the left panel the quantile functions cross at $x = 0$ and we observe x's uniformly distributed on the interval $[-0.1, 5]$. The dashed lines representing the fitted linear approximation to the nonlinear conditional quantile functions are only slightly distorted. In the right panel, the x's are now uniform on $[-5, 5]$, and we see that the attempt to fit the piecewise linear truth with a strictly linear model completely misrepresents reality. The true V-shaped and Λ-shaped quantile functions appear, thanks to the symmetry of the situation to produce parallel quantile functions as if there were iid error. Some intermediate cases, with asymmetric support of the x's are illustrated in Figure 5.3.

In structural models since the response in one equation becomes a conditioning covariate in a subsequent equation, the problem of crossing becomes more acute in models that are specified to be linear in covariates. In the simulations reported in the next section, for example, parameters have been chosen to insure that the specified scale effects remain positive on the full support of the exogonous covariates. It should

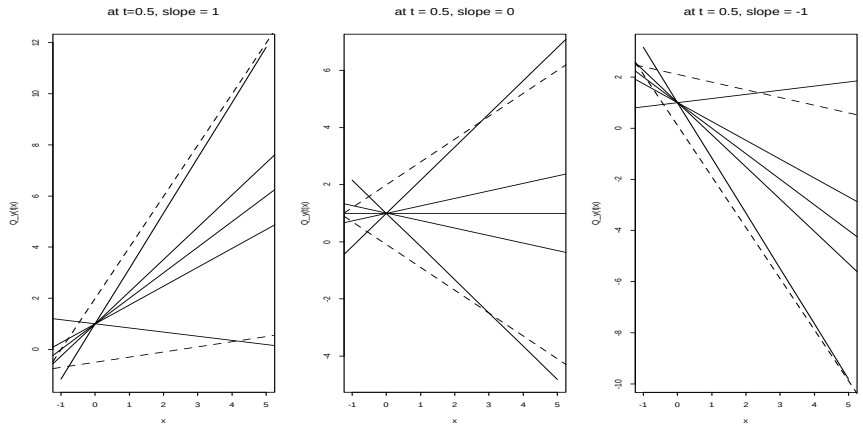

FIGURE 5.3. Direction of Distortion: $Q_y(\tau|x) = 1 + x(\beta + F_\nu^{-1}(\tau))$, $\nu \sim N(0, 0.5^2)$, $\tau \in \{0.01, 0.3, 0.5, 0.7, 0.99\}$, $x \in [-1, 5]$. The crossing occurs at $x = 0$. The dashed lines are the "distorted" fitted curves at $\tau = 0.01, 0.99$. The left figure is for $\beta = 1$, there is little distortion at the upper tail while there is large distortion at lower tail; the middle figure is for $\beta = 0$, the distortions happen in a symmetric way from the tails toward the median; the right figure is for $\beta = -1$, contrary to the case where $\beta = 1$, there is little distortion at the lower tail while large distortion occurs at upper tail.

be emphasized that in applications it is often advisable to consider specifications that are nonlinear in covariates, especially when estimated quantile functions are found to cross inside the convex hull of the exogonous covariates.

6. MONTE-CARLO

In this section we very briefly report on some simulation experiments designed to evaluate the performance of the estimation methods considered above. We restrict attention to sample size $n = 100$. The number of replications is set to $R = 1000$. The computational results reported in this and the following section were carried out in the R language, Ihaka and Gentleman (1996) using the quantile regression package of Koenker (1998).

We consider a simple location-scale shift model:

$$Y_1 = \alpha_1 + \alpha_2 x + (\alpha_3 + \delta(\lambda\nu_2 + \nu_1))Y_2 \tag{6.1}$$

$$Y_2 = \beta_1 + \beta_2 x + \beta_3 z + \nu_2 \tag{6.2}$$

where x, z, ν_1 and ν_2 are generated as the following: $x \sim t_3$, $z \sim N(15, 2^2)$, $\nu_1 \sim N(0, 1)$. and $\nu_2 \sim N(0, 0.5^2)$, We specify the parameter vectors as following, $(\alpha_1, \alpha_2, \alpha_3, \delta, \lambda) = (3, 4, 4, 5, 3)$, and $(\beta_1, \beta_2, \beta_3) = (1, 2, 3)$. For this model, both the weighted average derivative (WAD) and the control variate (CV) estimators for the structural quantile treatment effect of Y_2 on Y_1 will converge to the population value of $4+15F_{\nu_2}^{-1}(\tau_2)+5F_{\nu_1}^{-1}(\tau_1)$. For the sake of simplicity, we set $\tau_1 = \tau_2 = \tau$. The results are reported in Table 6.1 for the estimators at quantiles $\tau = (0.1, 0.3, 0.5, 0.7, 0.9)$.

	Coeffcient	Bias	Std. Error	RMSE
$\tau_1 = \tau_2 = \mathbf{0.1}$				
Theoretical Value	-12.019	0.000	0.000	0.000
CVQR	-10.799	1.221	11.715	11.778
WADQR	-10.748	1.271	12.057	12.124
2SQRQ	-7.191	4.829	11.505	12.478
2SQRA	-7.149	4.871	11.473	12.464
2SQRS	-7.152	4.867	11.473	12.463
QR	-2.788	9.231	11.820	14.997
$\tau_1 = \tau_2 = \mathbf{0.3}$				
Theoretical Value	-2.555	0.000	0.000	0.000
CVQR	-1.969	0.586	8.905	8.925
WADQR	-1.876	0.679	9.280	9.305
2SQRQ	-0.345	2.210	9.225	9.486
2SQRA	-0.337	2.218	9.229	9.492
2SQRS	-0.330	2.225	9.226	9.490
QR	4.031	6.586	9.086	11.221
$\tau_1 = \tau_2 = \mathbf{0.5}$				
Theoretical Value	4.000	0.000	0.000	0.000
CVQR	3.715	-0.285	8.656	8.661
WADQR	3.722	-0.278	8.934	8.939
2SQRQ	3.847	-0.153	8.488	8.490
2SQRA	3.847	-0.153	8.488	8.490
2SQRS	3.855	-0.145	8.490	8.492
QR	8.006	4.006	8.313	9.228
$\tau_1 = \tau_2 = \mathbf{0.7}$				
Theoretical Value	10.555	0.000	0.000	0.000
CVQR	9.945	-0.610	8.953	8.974
WADQR	9.968	-0.587	9.506	9.524
2SQRQ	8.417	-2.138	8.895	9.148
2SQRA	8.425	-2.130	8.896	9.148
2SQRS	8.425	-2.130	8.900	9.152
QR	12.626	2.071	8.694	8.937
$\tau_1 = \tau_2 = \mathbf{0.9}$				
Theoretical Value	20.019	0.000	0.000	0.000
CVQR	19.507	-0.513	11.166	11.177
WADQR	19.367	-0.653	12.390	12.407
2SQRQ	14.750	-5.270	11.617	12.756
2SQRA	14.796	-5.223	11.665	12.781
2SQRS	14.787	-5.232	11.656	12.776
QR	19.191	-0.828	11.385	11.415

TABLE 6.1. Simulation Results

We see first, that both estimators exhibit very modest bias. Secondly, in terms of the standard error and root mean square error, the control variate estimator outperforms the weighted derivative estimator at all considered quantiles.

For the sake of comparison we consider four other estimators:

QR: Naive quantile regression applied to (6.1) without any attempt to deal with the endogoneity of Y_2.

2SQRQ: Two stage quantile regression replacing Y_2 by the predicted $\hat{Y}_2$ from the $\tau = \tau_2$ quantile regression estimation of (6.2).

2SQRA: Two stage quantile regression replacing Y_2 by the predicted $\hat{Y}_2$ from the $\tau = 1/2$ median regression estimation of (6.2).

2SQRS: Two stage quantile regression replacing Y_2 by the predicted $\hat{Y}_2$ from the ordinary least squares (mean) regression estimation of (6.2).

The performance of the other estimators is quite unsatisfactory by comparison with the WADQR and CVQR proposals. At the median the two-stage methods all have good bias and variance performance, as one would expect from the results of Amemiya (1982). But at all other quantiles they exhibit serious bias problems. Naive quantile regression estimation of the structural equation, as expected, is also badly biased, except (amusingly) at $\tau = 0.9$, where countervailing bias effects seem to fortuitously cancel.

7. The Effect of Class Size on Student Performance in Dutch Primary Schools

In this section we reconsider an application of Levin (2001) investigating the effect of class size on student performance in Dutch primary schools. We will apply both weighted derivative and the control variate methods to a structural equation model of the impact of class size on student achievement. Our main objective is to demonstrate how these new approaches can be employed to reveal new aspects of the sample and thus yield more detailed and constructive policy analysis. We find that the two methods produce quite similar results, especially for language performance, a finding that somewhat reenforces our confidence in our model specification. Both estimators indicate that the class size effects vary significantly across quantiles of the class size distribution and student achievement distribution. For the lower attainment students, bigger classes improve language performance, while smaller classes improve math scores. For average students, class sizes have insignificant effects on both language and math performance. For high attainment students smaller classes are slightly better for language performance, but class size effects are not significant for math performance. These findings suggest that a general policy of class size reduction is unlikely to have large beneficial effects on overall student achievement and should be approached with some skepticism.

7.1. **A Brief Review of the Literature on Class Size Effect.**

Student academic performance is of paramount importance to parents, teachers and educational policy makers. Among policy tools available to school administrators reductions in class size appear among the most promising prescriptions for improving student achievement.

However, the statistical evidence on the linkages between class size and student performance is mixed.[1] Since the publication of the influential Coleman report (1966), there have been literally hundreds of studies examining the relationship between class size and student achievement. The results span the full range of possible conclusions: some find that there is a significant and positive relationship between class size and student achievement; some find that smaller classes are more effective; some find that there is no discernible relationship. Inevitably, some of the uncertainty in the literature derives from the fact that there is no uniformly agreed specification of the model or estimation method for the causal effect of class size. Most empirical studies have employed least squares methods to obtain estimates of the effect of class size on student achievement, and thus present a mean treatment view of class size effect. Recognizing the heterogeneity in the potential effects several authors have recently suggested that a more disaggregated estimation of the policy effects would be preferred, see e.g. Hanushek (1986), Krueger (1997), Card (2001) and Angrist and Krueger (2001). However, to the best of our knowledge, only two studies take up the challenge to investigate class size effects across quantiles of school achievement distribution.

Eide and Showalter (1998) using US data, apply quantile regression methods to a model of student achievement and find that the class size effect is insignificantly different from zero at all quantiles of students achievement distribution. It should be emphasized that this model does not include students' family background, or peer effects, and that they treat the class size variable as exogenous. Noting the endogeneity problem, Levin (2001) applies a variant of Amemiya's (1982) methods to a structural equation model, but also finds little empirical support for beneficial effects of smaller classes at most quantiles with or without peer effects added to the model. Note that both Eide and Showalter (1998) and Levin (2001) present what we have characterized as a mean quantile treatment effect view of class size effects: How does mean class size affect the distribution of academic outcomes? By revealing the variations of class size effects across quantiles of students achievement, the MQTE approach offers a more complete view than earlier work. However, the effect of variations across quantiles of the distribution of class sizes remains obscure. As a consequence, it is hard to evaluate the class size effect without acknowledging that various class sizes have different influences on students' academic performance. For broader view of class size effects, we consider the structural quantile treatment effect in the framework that we have set out in Section 2, in an effort to explore the

[1]For meta-analysis, see Glass and Smith (1979), Glass et al. (1982), Porwoll (1978), Robinson and Wittebols (1986) and Hanushek (1998). See also, Summers and Wolfe (1977), Hanushek (1986,1997), Angrist and Lavy (1998) and Krueger (2000). The Tennessee Student/Teacher Achievement Ratio experiment, known as project STAR, involved 11,600 students from 80 schools over four years Finn and Archilles (1990). Initiated in 1996, the California Class Size Reduction, namely the CSR program, cost over $1 billion per year and affected over 1.6 million students (Class Size Reduction in California: Early Evaluation Findings: 1996-1998, 1999). Dutch policy makers have recently dedicated more than $500 million to reduce class sizes in primary education (Levin, 2001).

	Minimum	Maximum	Mean	Std. Dev.
Language Score	841.80	1261.20	1073.26	51.56
Math Score	822.70	1361.30	1123.49	83.94
Pupil's Gender (Female=1)	0	1	0.50	0.50
IQ	4.00	37.00	25.53	4.95
Socio-Economic Status (SES)	0	1	0.53	0.50
Risk	1.00	5.00	2.20	0.87
Peer Effects (Language)	935.65	1179.10	1073.19	40.99
Peer Effects (Math)	852.67	1271.16	1123.44	69.70
Class Size*	5	39	23.81	6.46
Teacher's Experience (Years)*	1	40	19.05	8.06
School Denomination (Public = 1)**	0	1	0.72	0.44
Weighted School Enrollment (WSE)**	23	684	250.35	120.42

TABLE 7.1. Sample Summary Statistics: There are 5698, 5368 and 5608 observations for grade 4, 6, and 8, respectively, which after pooling and deleting cases with missing values for important variables yielded 12,203 observations.

potential heterogeneity in the class size effect over both the distribution of students achievement as well as the distribution of class sizes.

7.2. **Data Description.** The data we employ is the first wave of the PRIMA cohort study, which contains detailed information on Dutch primary school students in grades 2, 4, 6, and 8 as well as the associated teacher and school characteristics for the school year 1994/1995.[2] The PRIMA cohort study is a comprehensive survey of primary education in Holland, enabling researchers to explore relationships between pupil's achievements, their characteristics, those of their parents, as well as class level and school level characteristics. Pupils are tested with regard to intelligence, reading abilities, the Dutch language and mathematics. Background data are gathered through parents and teachers and detailed school level data are furnished by the directors of the participating schools. In total, there are about 57,000 pupils from 700 primary schools in the survey. Of these, 450 schools form the representative random sample that we use in this paper. Only grades 4, 6 and 8 are considered and the three grades are pooled together in our analysis.[3]

A brief statistical summary of the variables used in our modeling is reported in Table 7.1. The average class size is 24 and ranges from 5 to 39, but about 70% of classes are between 15–35. It may be noted that the variability of math scores is considerably higher than that of the language scores. About 72% of the schools in the sample are public, but it probably should be emphasized that the distinction between private and public schools in Holland is not nearly so great as one may be

[2]This data has been previously used by Dobblelsteen et al (1998) and Levin (2001).

[3]The ages of pupils in grade 4, 6 and 8 are around 7–8, 9–10 and 11–12, respectively.

led to expect from the vantage point of the US. Estimates of the interaction of school denomination and class size indicate that there is no significant difference in class size effects between public and private schools.

7.3. **Model Specification.** Before considering the formal model, there are two concerns about class size effects that should be addressed. The first one is the causal mechanism: class size *per se* should not contribute to students' academic achievement. Presumably, class size operates through various channels that exert influences on student performance. For example, smaller classes may induce changes in instructional methods and change the nature of peer effects. Both these factors are thought to play important roles in students' academic performance. Lazear (1999), for example, has focused on the public good aspect of classroom teaching and investigates the congestion effects of class size from a theoretical perspective. But there seems to be no generally accepted theory of the causal mechanism that links class size to performance.

A second major concern for the emprical study of class size effects is potential endogeneity. Parents may make location decisions based on the quality of local public schools attempting to ensure that their children attend small classes; school administrators may have a desire to put the lower attainment students in smaller classes or try to assign better teachers to bigger classes. Correspondingly, to treat the endogeneity problem of class size, there are two approaches in the literature: one is to sidestep endogeneity issues by focusing on "experimental" settings like the Tennessee STAR experiment, or related "natural experiments" as in Hoxby (2000); the other is to use instrumental variable methods to correct for the bias induced by endogenous covariates. e.g., Krueger (1997), Angrist and Lavy (1998), Hanushek (2001). While most studies adopt the IV approach, a good IV is notoriously hard to find. Empirically, researchers have taken the assigned class size Krueger (1997), school enrollment Akerheilm (1995), Iacovou (2001), Levin, 2001) and grade enrollment Angrist and Lavy (1998) as instrumental variables for actual class size in either continuous or non-continuous forms.

Given the observational, i.e. non-experimental, nature of our data, we may begin by considering a conventional approach based on a linear structural equation model of the form

$$y = \alpha_0 + X_i\alpha_1 + X_c\alpha_2 + X_s\alpha_3 + Y\delta + u \tag{7.1}$$

$$Y = \beta_0 + X_i\beta_1 + X_c\beta_2 + X_s\beta_3 + Z\gamma + U. \tag{7.2}$$

The precise specification of the random components u and U will be delayed momentarily while we consider the observable variables. Math or language test scores are denoted by y_i for student i in class c and school s; X_i are individual i's characteristic variables including pupil's gender, IQ, socioeconomic status (SES), peer effects and

risk level;[4] X_c are class c's characteristic variables including teacher's experience;[5] X_s are school s's characteristic variables, including the school denomination (public or nonpublic) only; Y is the covariate for class size and Z denotes the instrument for class size; u and U denote unobserved random components. As we have already noted, in the pure location shift form of the model the structural effect of class size is unambiguous: the parameter δ captures this effect and it may be interpreted as the shift in location of test scores induced by a change in class size that describes the effect at all quantiles of the academic performance distribution and at all quantiles of the class size distribution.

What *is* z, the instrumental variable for class size? The Dutch Ministry of Education imposed a new funding allocation rule during the time period of the first wave of the PRIMA survey. Each primary school reported weighted school enrollment (WSE) to the Ministry with weights determined by the socio-economic status of the enrolled students. Based on the value of this WSE, the Ministry allocated funding to each school and this funding determined how many teachers the school could hire. It is clear that this WSE variable is closely related to the actual class size but has no direct relation with student achievements conditional on characteristics. Following Levin (2001) we employ WSE as our instrumental variable for class size. This weighted school enrollment is calculated according to the following formula:

$$z_i = 1.03 \max\{\sum_{j=1}^{n_i} s_{ij} - .09 n_i, n_i\}, \tag{7.3}$$

where n_i is the total school enrollment of school i and s_{ij} is the weight determined by the socioeconomic status of each student j in school i. The variable s_{ij} takes values $\{1.0, 1.25, 1.4, 1.7, 1.9\}$ with 1 being the reference level and 1.9 being the worst family background. Based on this formula, we see that schools located in poorer neighborhoods will have more teachers.

Since z_i varies only between schools not within schools, a natural question may be, are we actually just using the school size as the IV? Preliminary tests indicate that although z_i and school size are closely related, z_i is quite distinct from school size. This is shown clearly by the top plot of Figure 7.1. where the upper conditional quantile functions of z_i given school enrollment have different slopes. The scatter plot also reveals that when the school size is smaller than 100 or bigger than 500, z_i is quite close to the school size, however, when the school size is between 100 and 500, z_i can be significantly different from the school size. This can be well explained by the fact that smaller schools, typically located in small towns or villages where most families

[4]Students are defined as "at risk" based on observed cognitive and/or behavioral problems. School must document students problems regularly. Based on information from the student profiles, each student is given a scaled score ranging from 1 to 5 in ascending order of riskiness. For detailed information on socio-economic status (SES), see Levin (2001), for the simplicity, we take recode SES as binary, with 1 indicating higher SES. The peer effect is measured by the classmates' average test score.

[5]Preliminary estimation indicated that teachers' age, sex and level of education were insignificant influences on students' achievement.

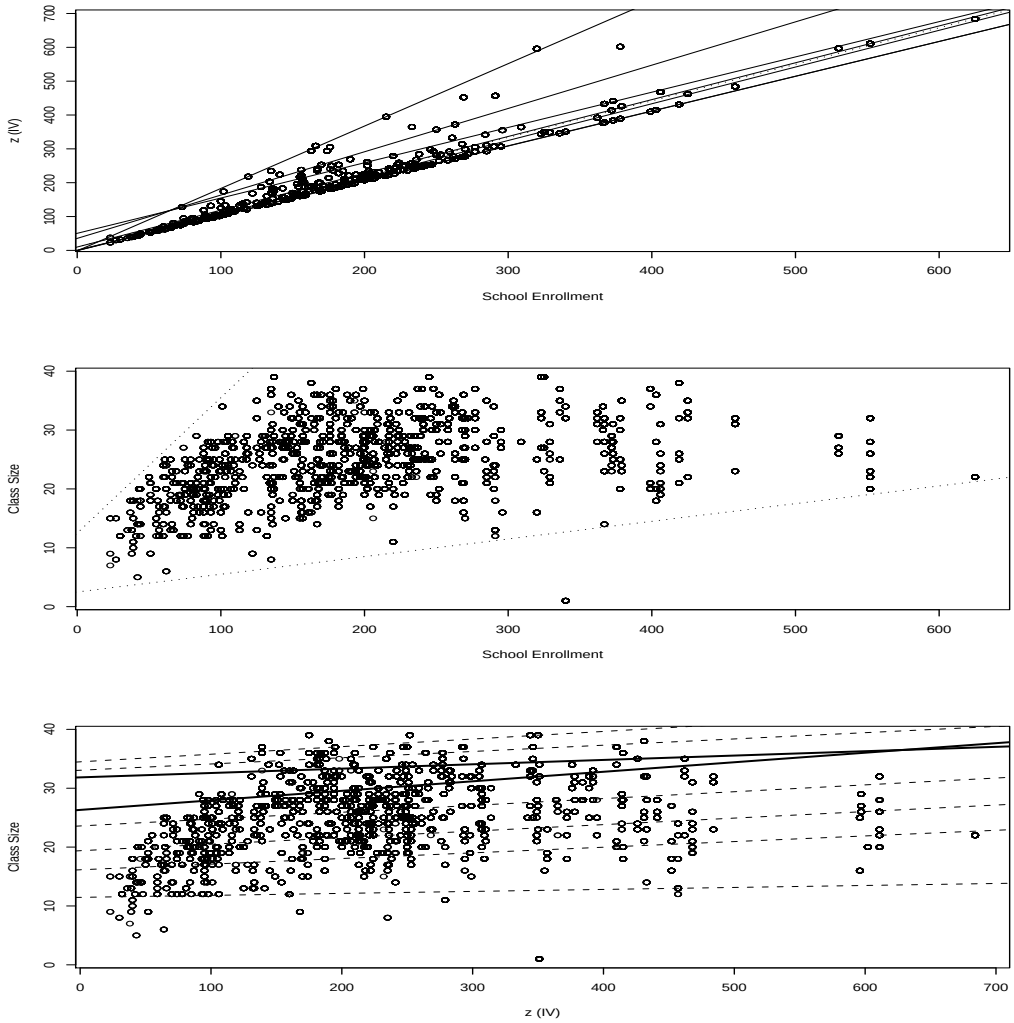

FIGURE 7.1. The top plot indicates that the weighted school enrollment variable, z, used as an instrument, is significantly different from the school size; the middle plot shows that class sizes are not strongly related to school sizes. The bottom plot shows that there is some heteroscedasticity in the relationship between class size and the WSE instrumental variable, the two solid lines represent the 0.75 and 0.90 quantiles.

are more homogeneous, have z_i that would be roughly similar to a scaled value of school size; for bigger schools, however, there are more varied family backgrounds. So z_i may diverge substantially from school size. Another concern is: how is class size is related to school size? Is it true that bigger schools imply bigger class sizes? The answer is no. Though the class size has more variability in bigger schools, it does not increase with the school size. This can be seen clearly from the bottom plot in Figure 7.1.

Regarding the performance of z_i, since our instrumental variable is at the school level, the more variation of class sizes is from "between schools", the better is the IV. We have estimated variance components for class size variable. The unconditional variance of class sizes is 41, the variance "between schools" is 28 and "within schools" is 13, so 70% of the variation of class sizes is "between schools". It should be emphasized that this does not imply that the variation comes from different school sizes! Furthermore, 83% of schools have only one class for each grade and the variation of class sizes within schools is due mainly to variation between grade levels. This is further supported by noting that in a decomposition of the "within school" variation the "between grades" variation in class size accounts for more than 92% of the within-school variation.

After some specification search we have selected a model in which class size is allowed to influence both the location and scale of the student performance distribution. Explicitly, we will assume that, $u_i = (\lambda\nu_{i2} + \nu_{1i})(Y_i\xi + 1)$ and $U_i = \nu_{2i}$, where ν_1 and ν_2 are independent of one another and iid over individuals. We will consider both weighted average derivative and control variate methods of estimation. As we have shown above, when the model is correctly specified both methods yield consistent estimators with the latter being more efficient. Substituting for ν_2 in the y_i equation yields a rather complicated form of what we have called the hyrid structural equation that is estimated in the weighted average derivative approach; it involves the location shift effects of the original specification plus a quadratic term in Y_i and interactions of Y_i with the other exogonous variables including z_i. In the case of the control variate estimator the situation is considerably simpler: the estimate $\hat{\nu}_2(\tau_2)$ is computed in the first stage, and then it is included along with its interaction with Y_i as additional covariates in the τ_1 quantile regression of the first y_i equation. In large samples like ours we would expect both estimators would produce similar results, provided that the model was correctly specified. When the model is misspecified, the weighted average derivative method is clearly preferable, the control variate method will be used primarly for checking the credibility of the specified structural model.

We will focus on the estimation of the structural class size effect. It should be emphasized that peer effects are also an very important influence on student performance. Moreover, since peer effects and class size effects are highly interconnected, their interaction should also be carefully explored. The endogeneity of peer effects makes this inquiry particularly challenging, but it is especially important from a policy standpoint to explore the distributional consequences of peer effects. We plan to address this issue in subsequent work.

7.4. **Empirical Analysis.** Before considering the structural estimation of the model we briefly describe some preliminary quantile regression results based on treating class size as exogonous. These results are illustrated in Figures 7.2 and 7.3 for language and math performance, respectively. Considering the class size effect first. The plots suggest that class size effects are roughly similar for math and language performance: both are significant, both are downward sloping, indicating that while class

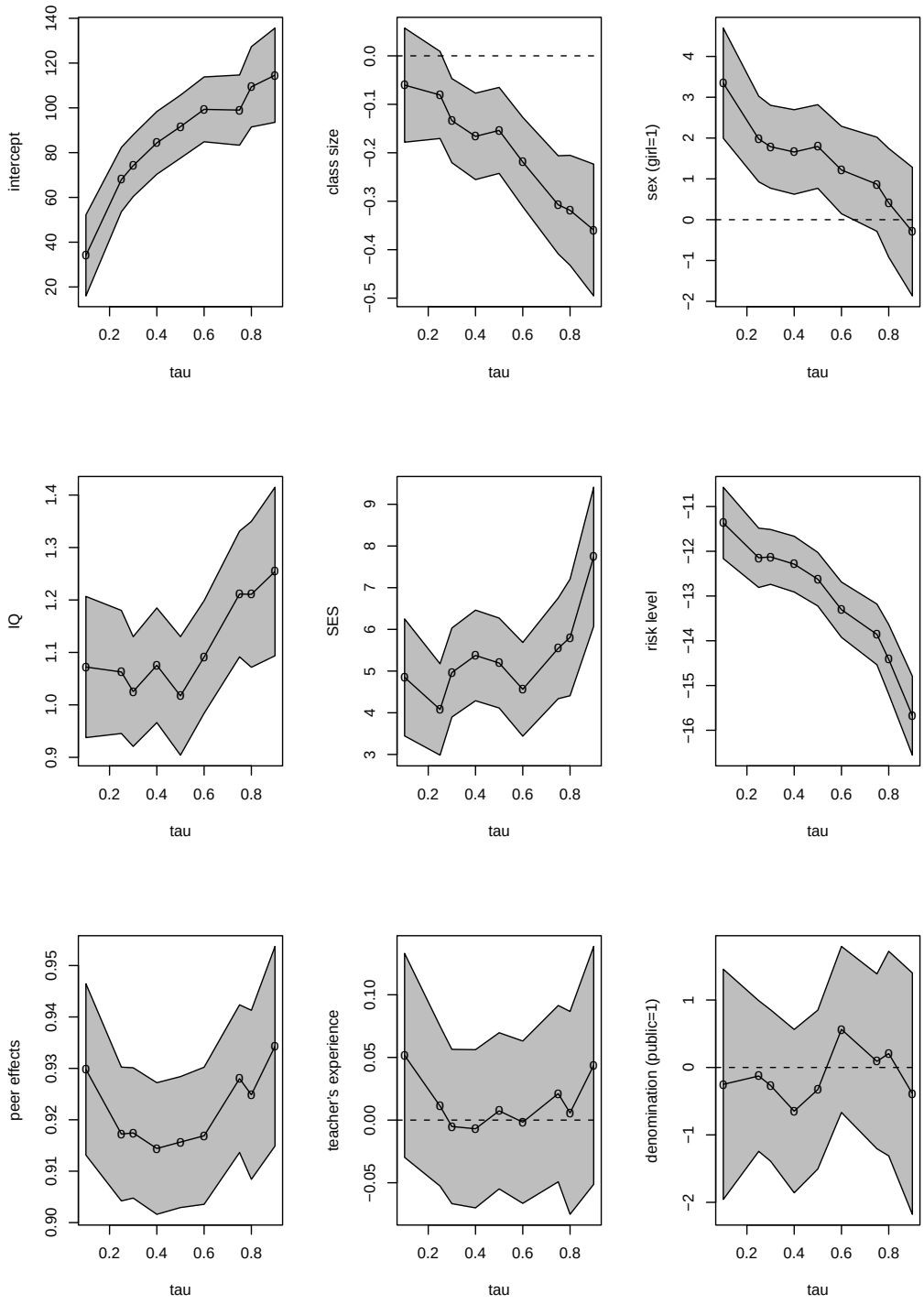

FIGURE 7.2. Quantile Regression Covariate Effects for Language Performance: Class Size Treated as Exogenous.

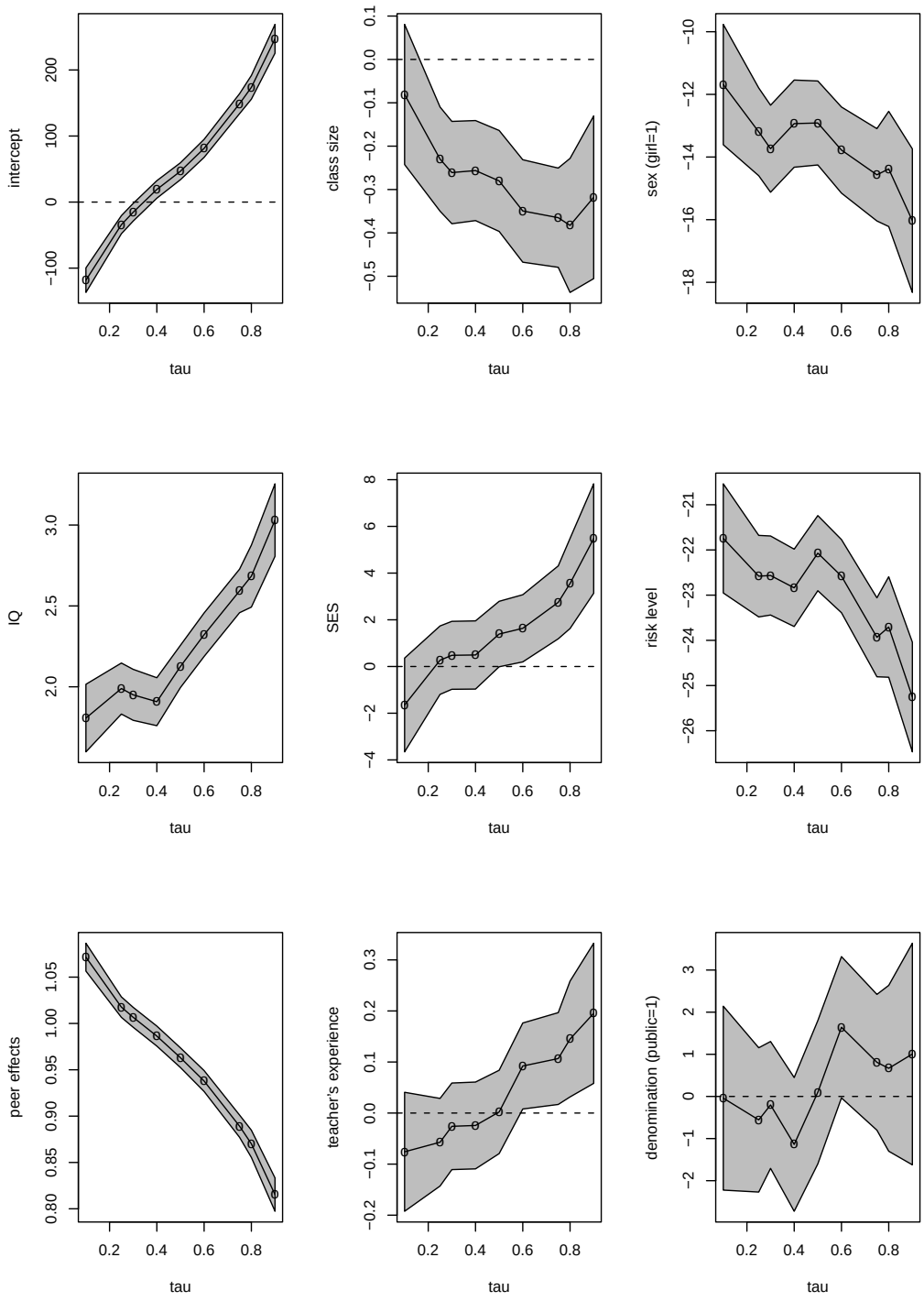

FIGURE 7.3. Quantile Regression Covariate Effects for Math Performance: Class Size Treated as Exogenous.

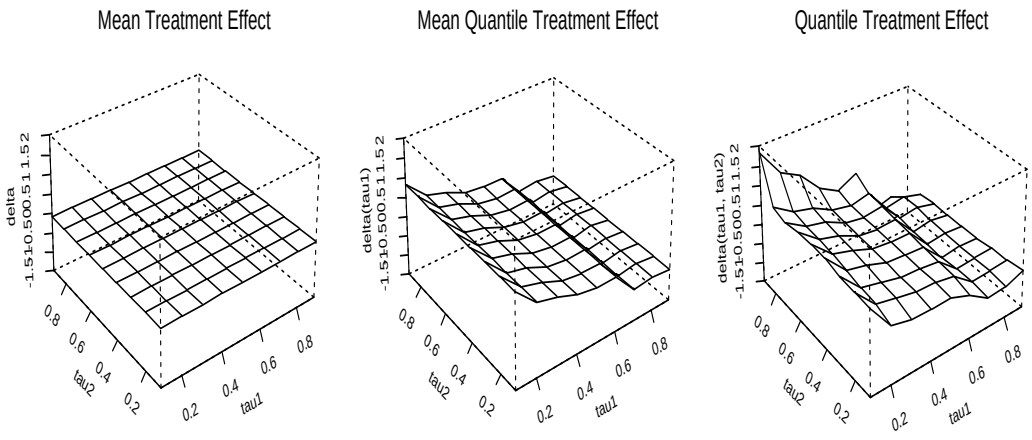

FIGURE 7.4. Structural Class Size Effects for Language: τ_1-students achievement, τ_2-class size.

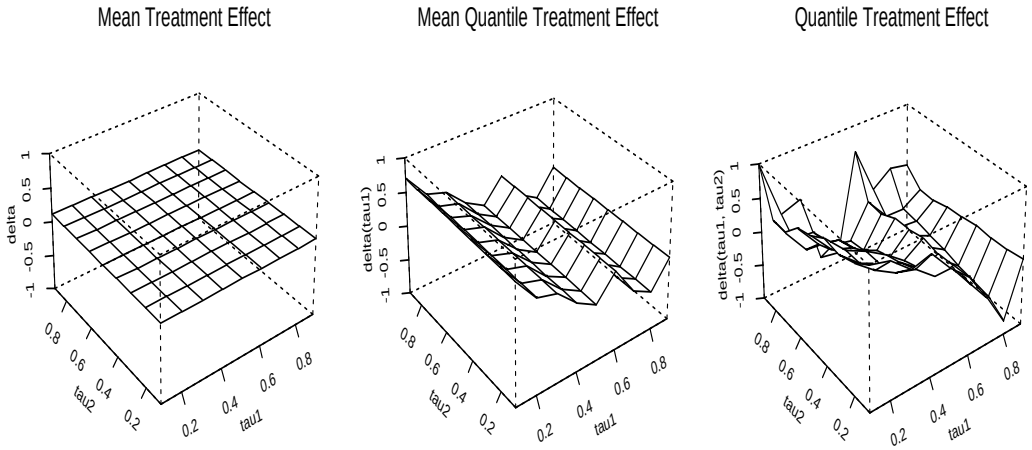

FIGURE 7.5. Structural Class Size Effects for Math: τ_1-students achievement, τ_2-class size.

size reductions are beneficial to all students they are more beneficial to better students conditional on the other covariates. The plots also suggest that peer effects are quite important especially for math, although considerable caution is required in the interpretation of these effects. Individual student characteristics are also quite interesting. Girls appear to be clearly disadvantaged in math, but exhibit a modest advantage in language. The "at risk" variable has a large impact, suggesting that students' attitude and behavior towards school work is crucial for their scholastic

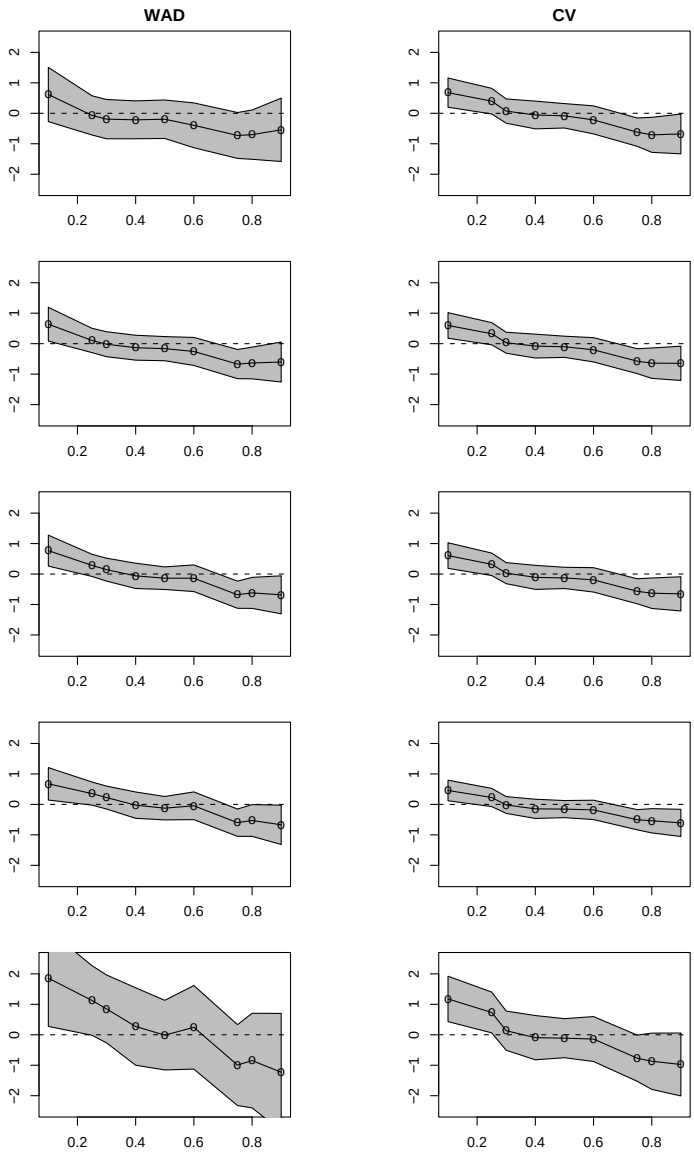

FIGURE 7.6. Structural Class Size Effect on Language Scores: The figure presents both the weighted average derivative (WAD) and control variate (CV) estimates of the structural class size effect on language performance. Five quantiles of the class size distribution are presented for each estimator in descending order from the top of the plot $\tau_2 \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$.

performance, although again, exogoneity may be controversial. As expected, family background plays a important role in students' academic performance, especially in language. Socio-economic status has a significantly positive effect across all quantiles of students achievement distribution and the effect increases as we move to higher

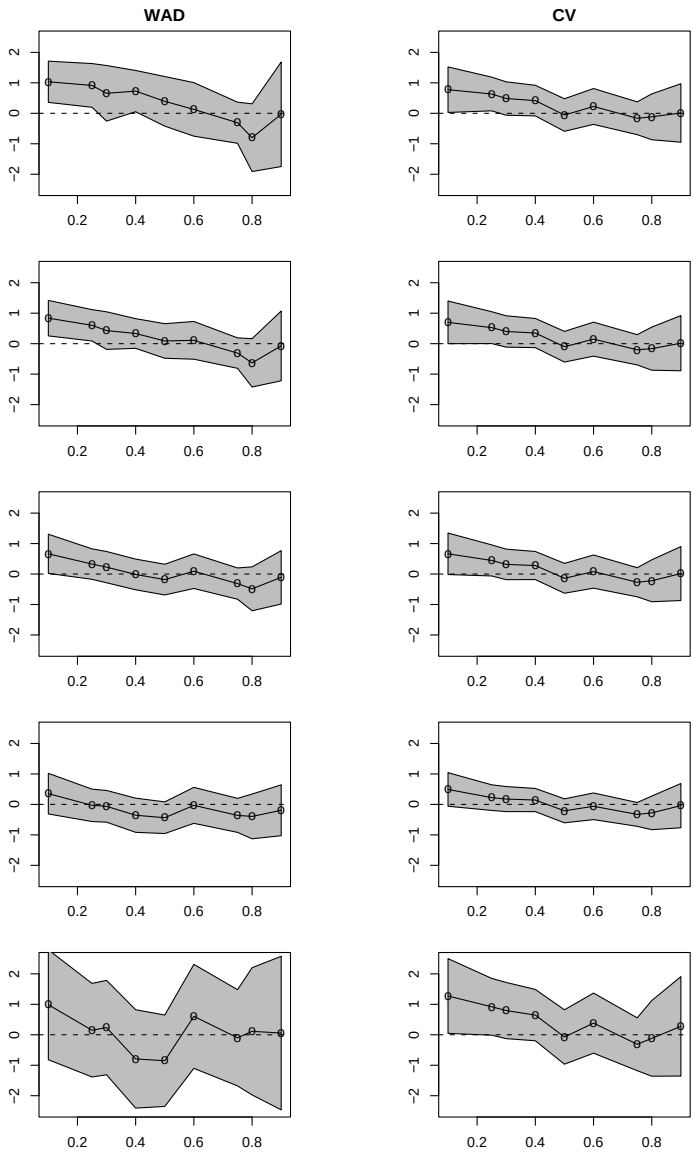

FIGURE 7.7. Structural Class Size Effect on Math Scores: The figure presents both the weighted average derivative (WAD) and control variate (CV) estimates of the structural class size effect on mathematics performance. Five quantiles of the class size distribution are presented for each estimator in descending order from the top of the plot $\tau_2 \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$.

quantiles of student achievement. IQ has the expected positive effect on students achievement with the magnitude of this effect larger on the math scores than on the language scores. Interestingly, more experienced teachers have no significant impact on language performance, but do seem to have a desirable effect on the upper quantiles

of math performance. A public versus parochial school effect on student attainment is not distinguishable across the quantiles considered.

We now turn to the estimation of the class size effect in our structural framework. A concise visual summary of the structural estimates of the class size effect on language and math scores is provided in Figures 7.4 and 7.5 respectively. In the left panel we depict the conventional two stage least squares estimate of the mean shift effect of class size viewed as a constant function of τ_1 and τ_2. In the middle panel we show what we have called the mean quantile treatment effect obtained by integrating out the τ_2 effect from the weighted average derivative estimate of the $\hat{\delta}(\tau_1, \tau_2)$ estimate of the structural class size effect. In the right panel we present $\hat{\delta}(\tau_1, \tau_2)$.

The two stage least squares estimate of the class size effect is -0.07 with a standard error of 0.20, a finding consistent with many other unsuccessful attempts to discern a significant effect of class size. However, our estimates of the mean quantile treatment effect of class size in the middle panel reveals a somewhat more nuanced view. Both math and language plots show a positive effect of around 0.7 at low quantiles and falling gradually to about -0.5 at the upper quantiles, suggesting that poorer students benefit from larger classes, while better students do better in smaller classes. Further disaggregating, the plots in the right panel indicate dispersion in the class size effect in both the τ_1 and τ_2 directions, but the picture is roughly similar: positive effects at the lower quantiles of test scores, and negative effects at the upper quantiles. In such circumstances it is not surprising that averaging over both quantile dimensions yields a result that is statistically negligible.

To examine the structural estimates more closely we plot in Figures 7.6 and 7.7 cross-sectional slices of the foregoing perspective plots. Superimposed on these plots is a .90 (pointwise) confidence band. To contrast the weighted average derivative approach and the control variate method we illustrate both estimates in Figure 7.6 for language performance and in Figure 7.7 for math. The similarity of the WAD and CV estimates provides some support for the model specification. We summarize our findings briefly as follows:

- The class size effect on language scores:
  - For weaker students the plots indicate that bigger classes are better.
  - For near median students class size effects are not significant.
  - For better students smaller classes appear marginally better.
- The class size effect on math scores:
  - For weaker students smaller classes are better
  - For the average and good students the class size effect is not significant.

Our finding that class size has an insignificant influence on median performance in language and math is quite consistent with previous literature indicating similarly insignificant conditional mean effects. However, especially in the case of language performance, we find that one should interpret findings of insignificant mean effects with considerable caution since it appears that they arise from averaging significant benefits from reductions in class size for good students and significant benefits from increases in class size for poorer students.

We would again stress the point that changes in class sizes *per se* cannot produce academic gains, but in combination with other instructional practices and institutional arrangements such changes may have benefits. By providing a more nuanced view of the apparently heterogeneous effects of class size, structural methods based on quantile regression may be able to constructively contribute to the policy debate on these important issues.

APPENDIX A. PROOFS

Lemma 1. *Let Y and Z be $N \times K$ matrices of rank K and X be a $N \times L$ matrix of rank L. If $\hat{\beta}_1 = (Z^\top M_X Z)^{-1} Z^\top M_X Y$, with $M_X = I - X(X^\top X)^{-1} X^\top$, then*

$$\begin{bmatrix} 1 & \hat{\beta}_1^{-1} \end{bmatrix} \begin{bmatrix} Y^\top M_X Y & Y^\top M_X Z \\ Z^\top M_X Y & Z^\top M_X Z \end{bmatrix}^{-1} = \begin{bmatrix} 0 & \hat{\beta}_1^{-1} (Z^\top M_X Z)^{-1} \end{bmatrix}. \tag{A.1}$$

Proof: Define $\tilde{Y} = M_X Y$ and $\tilde{Z} = M_X Z$, we have:

$$\begin{bmatrix} \tilde{Y}^\top \tilde{Y} & \tilde{Y}^\top \tilde{Z} \\ \tilde{Z}^\top \tilde{Y} & \tilde{Z}^\top \tilde{Z} \end{bmatrix}^{-1} = \begin{bmatrix} (\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} & F \\ F^\top & (\tilde{Z}^\top M_{\tilde{Y}} \tilde{Z})^{-1} \end{bmatrix}, \tag{A.2}$$

where F satisfies

$$(\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} \tilde{Y}^\top \tilde{Z} + F \tilde{Z}^\top \tilde{Z} = 0, \tag{A.3}$$

or,

$$(\tilde{Z}^\top M_{\tilde{Y}} \tilde{Z})^{-1} \tilde{Z}^\top \tilde{Z} + F^\top \tilde{Y}^\top \tilde{Z} = I. \tag{A.4}$$

Using (A.2), the left hand side of (A.1) can be written as,

$$\begin{bmatrix} 1 & \hat{\beta}_1^{-1} \end{bmatrix} \begin{bmatrix} (\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} & F \\ F^\top & (\tilde{Z}^\top M_{\tilde{Y}} \tilde{Z})^{-1} \end{bmatrix} = [(\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} + \hat{\beta}_1^{-1} F^\top, F + \hat{\beta}_1^{-1} (\tilde{Z}^\top M_{\tilde{Y}} \tilde{Z})^{-1}].$$

From (A.3) and (A.4), we have, respectively,

$$\begin{aligned} F &= -(\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} \tilde{Y}^\top \tilde{Z} (\tilde{Z}^\top \tilde{Z})^{-1} \\ &= -(\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} \hat{\beta}_1^\top, \\ (\tilde{Z}^\top M_{\tilde{Y}} \tilde{Z})^{-1} &= (I - F^\top \tilde{Y}^\top \tilde{Z})(\tilde{Z}^\top \tilde{Z})^{-1} \\ &= (\tilde{Z}^\top \tilde{Z})^{-1} - F^\top \hat{\beta}_1^\top. \end{aligned}$$

Consequently,

$$\begin{aligned} (\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} + \hat{\beta}_1^{-1} F^\top &= (\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} - \hat{\beta}_1^{-1} \hat{\beta}_1 (\tilde{Y}^\top M_{\tilde{Z}} \tilde{Y})^{-1} \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} F + \hat{\beta}_1^{-1} (\tilde{Z}^\top M_{\tilde{Y}} \tilde{Z})^{-1} &= F + \hat{\beta}_1^{-1} (\tilde{Z}^\top \tilde{Z})^{-1} - \hat{\beta}_1^{-1} F^\top \hat{\beta}_1^\top \\ &= F + \hat{\beta}_1^{-1} (\tilde{Z}^\top \tilde{Z})^{-1} - F \\ &= \hat{\beta}_1^{-1} (Z^\top M_X Z)^{-1}, \end{aligned}$$

■

Proof of Proposition 1. The 2SLS estimator of α_1 in model (2.1-2) is

$$\hat{\alpha}_1 = (\hat{Y}_2^\top M_X \hat{Y}_2)^{-1} \hat{Y}_2^\top M_X \, Y_1,$$

where $\hat{Y}_2 = z\hat{\beta}_1 + X\hat{\beta}_2$, $\hat{\beta}_1 = (z^\top M_X z)^{-1} z^\top M_X \, Y_2$, and $M_X = I - X(X^\top X)^{-1}X^\top$. Solving for ν_2 from (2.2) and substituting into (2.1), we have,

$$Y_1 = X(\alpha_2 - \beta_2\lambda) + V\tilde{\delta} + \nu_1, \tag{A.5}$$

where $V = (Y_2 \vdots z)$, and $\tilde{\delta} = (\delta_1, \ \delta_2) = (\alpha_1 + \lambda, \ -\beta_1\lambda)$. Our estimator for α_1 is $\hat{\delta}_1 + \hat{\delta}_2\hat{\beta}_1^{-1}$ where

$$\begin{aligned}
(\hat{\delta}_1 \quad \hat{\delta}_2)^\top &= (V^\top M_X V)^{-1} V^\top M_X \, Y_1 \\
&= \begin{bmatrix} Y_2^\top M_X Y_2 & Y_2^\top M_X z \\ z^\top M_X Y_2 & z^\top M_X z \end{bmatrix}^{-1} \begin{bmatrix} Y_2^\top \\ z^\top \end{bmatrix} M_X \, Y_1
\end{aligned}$$

By Lemma 1,

$$\begin{aligned}
\hat{\delta}_1 + \hat{\delta}_2\hat{\beta}_1^{-1} &= [0 \quad \hat{\beta}_1^{-1}(z^\top M_X z)^{-1}] V^\top M_X \, Y_1 \\
&= \hat{\beta}_1^{-1}(z^\top M_X z)^{-1} z^\top M_X \, Y_1 \\
&= [(z\hat{\beta}_1)^\top M_X (z\hat{\beta}_1)]^{-1} (z\hat{\beta}_1)^\top M_X \, Y_1 \\
&= [(z\hat{\beta}_1 + X\hat{\beta}_2)^\top M_X (z\hat{\beta}_1 + X\hat{\beta}_2)]^{-1} (z\hat{\beta}_1 + X\hat{\beta}_2)^\top M_X \, Y_1 \\
&= (\hat{Y}_2^\top M_X \hat{Y}_2)^{-1} \hat{Y}_2^\top M_X \, Y_1.
\end{aligned}$$

■

Proof of Theorem 1. Conventional asymptotic theory for quantile regression in the nonlinear in parameters model, e.g. Jurečková and Procházka (1994), implies that

$$\sqrt{n}(\hat{\theta}(\tau_1) - \theta(\tau_1)) = \bar{J}_1^{-1} n^{-1/2} \sum_{i=1}^{n} \sigma_{i1} \dot{g}_{i1} \psi_{\tau_1}(Y_{i1} - \xi_{i1}) + \mathbf{o}_p(1),$$

$$\sqrt{n}(\hat{\beta}(\tau_2) - \beta(\tau_2)) = \bar{J}_2^{-1} n^{-1/2} \sum_{i=1}^{n} \sigma_{i2} \dot{g}_{i2} \psi_{\tau_2}(Y_{i2} - \xi_{i2}) + \mathbf{o}_p(1).$$

Taylor expansion of $\hat{\pi}(\tau_1, \tau_2)$ at $(\theta(\tau_1), \ \beta(\tau_2))$ yields

$$\begin{aligned}
\sqrt{n}(\hat{\pi}_n(\tau_1, \tau_2) - \pi(\tau_1, \tau_2)) &= \begin{bmatrix} \nabla_{\theta(\tau_1)}\pi & \nabla_{\beta(\tau_2)}\pi \end{bmatrix} \begin{bmatrix} \sqrt{n}\hat{\theta}(\tau_1) - \theta(\tau_1) \\ \sqrt{n}\hat{\beta}(\tau_2) - \beta(\tau_2) \end{bmatrix} + \mathbf{o}_p(1) \\
&\equiv W_1 \sqrt{n}(\hat{\theta}(\tau_1) - \theta(\tau_1)) + W_2 \sqrt{n}(\hat{\beta}(\tau_2) - \beta(\tau_2)) + \mathbf{o}_p(1).
\end{aligned}$$

By hypothesis ν_{i1} is independent of ν_{i2}, so the result follows by the application of the δ-method. ■

The following Lemma will be used for the proof of Theorem 2.

Lemma 2. *Let $A(x)$ be a $n \times p$ matrix of functions defined on a set $S \in \mathbb{R}^m$. Suppose x_0 is an interior point of S at which A is continuously differentiable and $A(x)$ has rank $p < n$ in some neighborhood of x_0, then A has a G-inverse, $A^- = (A^\top A)^{-1} A^\top$ and at x_0,*

$$\frac{\partial A^-}{\partial x} A A^- = -A^- \frac{\partial A}{\partial x} A^-. \tag{A.6}$$

Proof: This is an immediate consequence of a more general result for G-inverses when A is allowed to have rank $q \leq p$. In that case we have, e.g. Harville (1997),

$$A\frac{\partial A^-}{\partial x}A = -AA^-\frac{\partial A}{\partial x}A^-A.$$

Multiplying from the left and right by A^-, and noting that $A^-A = I_p$ by the rank hypothesis, yields (A.6). ∎

Proof of Theorem 2 Note that $\hat{\alpha}(\tau_1,\tau_2) = \hat{\alpha}_{\hat{\nu}_2(\tau_2)}(\tau_1)$ and write

$$\sqrt{n}(\hat{\alpha}(\tau_1,\tau_2) - \alpha(\tau_1,\tau_2)) = \sqrt{n}\big(\hat{\alpha}_{\hat{\nu}_2(\tau_2)}(\tau_1) - \hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)\big) + \sqrt{n}(\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1) - \alpha_{\nu_2(\tau_2)}(\tau_1)).$$

Consider the second term, as in the proof of Theorem 1,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}_{\nu_2} - \alpha_{\nu_2}) &= D_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_{i1}\dot{g}_{i1}\psi_{\tau_1}(e_{i1}) + \mathbf{o}_p(1)\\ &\rightsquigarrow \mathcal{N}(0,\ \omega_{11}\bar{D}_1^{-1}D_1\bar{D}_1^{-1}),\end{aligned}$$

where $e_{i1} = Y_{i1} - g_{i1}$. Expanding the first term we have,

$$\sqrt{n}(\hat{\alpha}_{\hat{\nu}_2} - \hat{\alpha}_{\nu_2}) = \sqrt{n}\Big(\frac{\partial\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)}{\partial\nu_2(\tau_2)}\Big)^\top(\hat{\nu}_2 - \nu_2) + \mathbf{o}_p(1).$$

Considering first the $(\hat{\nu}_2 - \nu_2)$ term, by denoting $\tilde{\varphi}_2(Y,z,x,\beta)$ as a $n\times 1$ vector with the ith row $\tilde{\varphi}_2(Y_i,z_i,x_i,\beta)$, we have,

$$\nu_2(\tau_2) - \hat{\nu}_2(\tau_2) = \tilde{\varphi}_2(Y,z,x,\beta) - \tilde{\varphi}_2(g_2,z,x,\beta) - \tilde{\varphi}_2(Y,z,x,\hat{\beta}) + \tilde{\varphi}_2(\hat{g}_2,z,x,\hat{\beta}).$$

Thus, we have

$$\begin{aligned}&\sqrt{n}(\hat{\alpha}_{\hat{\nu}_2} - \hat{\alpha}_{\nu_2})\\ =\ & -\sqrt{n}\Big(\frac{\partial\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)}{\partial\nu_2(\tau_2)}\Big)^\top\big((\nabla_\beta(\tilde{\varphi}_2(Y,z,x,\beta) - \tilde{\varphi}_2(Y,z,x,\beta)))^\top(\hat{\beta}_n - \beta)\\ & +(\nabla_Y\tilde{\varphi}_2(Y,z,x,\beta))^\top(\hat{g}_2 - g_2)\big)|_{Y=g_2} + \mathbf{o}_p(1)\\ =\ & -\sqrt{n}\Big(\frac{\partial\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)}{\partial\nu_2(\tau_2)}\Big)^\top(\nabla_Y\tilde{\varphi}_2(Y,z,x,\beta))^\top(\hat{g}_2 - g_2)|_{Y=g_2} + \mathbf{o}_p(1)\\ =\ & -\sqrt{n}\Big(\frac{\partial\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)}{\partial\nu_2(\tau_2)}\Big)^\top(\nabla_{\nu_2}\varphi_2)^{-1\top}(\hat{g}_2 - g_2) + \mathbf{o}_p(1)\\ =\ & -\sqrt{n}\Big(\frac{\partial\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)}{\partial\nu_2(\tau_2)}\Big)^\top G(\hat{\beta}_n - \beta) + \mathbf{o}_p(1),\end{aligned}$$

where G denotes the matrix with the ith row $(\nabla_{\nu_{i2}}\varphi_{i2})^{-1}\dot{g}_{i2}^\top$.

The Bahadur representation for $\sqrt{n}\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)$ can be written as

$$\begin{aligned}\sqrt{n}\hat{\alpha}_{\nu_2(\tau_2)}(\tau_1) &= (n^{-1}\sum_{i=1}^n \sigma_{i1}f_{i1}\dot{g}_{i1}\dot{g}_{i1}^\top)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_{i1}\dot{g}_{i1}(f_{i1}\dot{g}_{i1}^\top\alpha_{\nu_2(\tau_2)}(\tau_1) + \psi_{\tau_1}(e_{i1})) + \mathbf{o}_p(1)\\ &= (n^{-1}\sum_{i=1}^n \sigma_{i1}f_{i1}\dot{g}_{i1}\dot{g}_{i1}^\top)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_{i1}\dot{g}_{i1}(f_{i1}g_{i1} + \psi_{\tau_1}(e_{i1})) + \mathbf{o}_p(1).\end{aligned}$$

Now differentiating, noting that the contribution of the $\psi_{\tau_1}(e_{i1})$ term is $\mathbf{o}_p(1)$, and using Lemma 2 gives

$$\begin{aligned}\Big(\frac{\partial \hat{\alpha}_{\nu_2(\tau_2)}(\tau_1)}{\partial \nu_2(\tau_2)}\Big)^\top G &= -\bar{D}_1^{-1}\big(\frac{1}{n}\sum_{i=1}^{n}\sigma_{i1}f_{i1}\eta_i\dot{g}_{i1}\dot{g}_{i2}^\top\big) + \mathbf{o}_p(1) \\ &\equiv -\bar{D}_1^{-1}\bar{D}_{12} + \mathbf{o}_p(1),\end{aligned}$$

where $\eta_i = \frac{\partial g_{i1}}{\partial \nu_{i2}(\tau_2)}(\nabla_{\nu_{i2}}\varphi_{i2})^{-1}$. Thus, we get immediately the limiting behavior of the first term,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}_{\hat{\nu}_2} - \hat{\alpha}_{\nu_2}) &= \bar{D}_1^{-1}\bar{D}_{12}\sqrt{n}(\hat{\beta}(\tau_2) - \beta(\tau_2)) + \mathbf{o}_p(1) \\ &\rightsquigarrow \mathcal{N}\big(0,\ \omega_{22}\bar{D}_1^{-1}\bar{D}_{12}\bar{D}_2^{-1}D_2\bar{D}_2^{-1}\bar{D}_{12}^\top\bar{D}_1^{-1}\big).\end{aligned}$$

Combining the results for the two terms completes the proof. ■

REFERENCES

[1] Abadie, A., Angrist, J. and Imbens, G.: "Instrumental Variables Estimates of the Effects of Subsidized Training on the Quantiles of Trainee Earnings", *Econometrica*, 70, 91-117, 2002.

[2] Akerhielm, K.: "Does Class Size Matter?", *Economics of Education Review*, 14, 229–241, 1995.

[3] Amemiya, T.: "Two Stage Least Absolute Deviations Estimators", *Econometrica*, 50, 689–711, 1982.

[4] Angrist J. and Krueger, A.: "Instrumental Variables and The Search for Identification: From Supply and Demand to Natural Experiments", *Journal of Economic Perspectives*, 114, 533–575, 1998.

[5] Angrist J. and Lavy, V.: "Using Maimonides' Rule to Estimate The Effects of Class Size on Scholastic Achievement", *Quarterly Journal of Economics*, 15, 69–85, 2001.

[6] Blundell, R and Powell, J.: "Endogeneity in Nonparametric and Semiparametric Regression Models", in *Advances in Economics and Econonometrics: Theory and Applications*, Eighth World Congress, Cambridge University Press, 2003.

[7] Card, D.: "Estimating The Return to Schooling: Progress on Some Persistent Econometric Problems", *Econometrica*, 69, 1127–1160, 2001.

[8] Chen, L. and Portnoy, S.: "Two-Staged Regression Quantiles and Two-Staged Trimmed Least-Squares Estimators for Structural Equation Model", *Communication in Statistics: Theory and Methods*, 25, 1005–1032, 1996.

[9] Chernozhukov, V. and Hansen, C.: "An IV Model of Quantile Treatment Effects", working paper, MIT, 2001.

[10] Chesher, A.: "Exogenous Impact and Conditional Quantile Functions", working paper, UCL, 2001.

[11] Chesher, A.: "Quantile Driven Identification of Structural Derivatives", working paper, UCL, 2002.

[12] Chesher, A.: "Identification in Nonseparable Models", *Econometrica*, 71, 1405–1441, 2003.

[13] Coleman, J.: "Equality of Educational Opportunity", Report for US Department of Health, Education and Welfare, Office of Education, Washington DC, 1966.

[14] Dobbelsteen, S., Levin, J. and Oosterbeek, H.: "The Causal Effect of Class Size on Scholastic Achievement: Distinguishing The Pure Class Size Effect from The Effect of Changes in Class Composition", Manuscript, University of Amsterdam, 1998.

[15] Eide, E. and Showalter, M.: "The Effect of School Quality on Student Performance: A Quantile Regression Approach", *Economics Letters*, 58, 345–350, 1998.

[16] Finn, J and Achilles, C.: "Answers and Questions about Class Size: A Statewide experiment", *American Educational Research Journal*, 24, 557–577, 1990.

[17] Glass, G. and Smith, M.: "Meta-Analysis of Research on Class Size and Achievement", *Educational Evaluation and Policy Analysis*, 1, 2–16, 1979.
[18] Glass, G. and Cahen, L., Smith, M. and Filby, N.: *School Class Size: Research and Policy*, Beverly Hills: Saga, 1982.
[19] Hanushek, E.: "The Economics of Schooling: Production and Efficiency in Public Schools", *Journal of Economic Literature*, 24, 1141–1177, 1986.
[20] Hanushek, E.: "School Resources and Student Performance", in *Does Money Matter? The effect of School Resources on Student Achievement and Adult Success*, edited by Burtless, Brookings Institution: Washington, DC, 1996.
[21] Hanushek, E.: "Assessing the Effects of School Resource on Student Performance: An Update", *Educational Evaluation and Policy Analysis*, vol. 19, issue 2, pp. 141–164, 1997.
[22] Harville, D. A. *Matrix Algebra From a Statistician's Perspective*. New York: Springer-Verlag, 1997.
[23] Hoxby, C.: "The Effects of Class Size on Student Achievement: New Evidence from Population Variation", *Quarterly Journal of Economics*, 115, 1239–1285, 2000.
[24] Iacovou, M.: "Class Size in The Early Years: Is Smaller Really Better?", working paper, 2001.
[25] Ihaka, R. and Gentleman, R.: "R, A Language for Data Analysis and Graphics", *Journal of Graphical and Computational Statistics*, 5, 299-314, 1996.
[26] Imbens, G. and Newey, W.: "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity", working paper, MIT, 2002.
[27] Jurečková, J. and Procházka, B.: "Regression Quantiles and Trimmed Least Squares in Nonlinear Regression Model", *Journal of Nonparametric Statistics*, 3, 201-222, 1994.
[28] Koenker, R. and Bassett, G.: "Regression Quantiles", *Econometrica*, 46, 33–50, 1978.
[29] Koenker, R. and Park, B.: "An Interior Point Algorithm for Nonlinear Quantile Regression", *Journal of Econometrics*, 71, 265-285, 1996.
[30] Koenker, R. and Zhao, Q.:"L-estimation for the Linear Heteroscedastic Models", *Journal of Nonparametric Statistics*. 3, 223-235, 1994.
[31] Koenker, R. "Quantreg: A Quantile Regression Package for R," `http://cran.r-project.org`, 1998.
[32] Krueger, A.: "Experimental Estimates of Education Production Functions", working paper, Princeton University and NBER, 1997.
[33] Krueger, A.: "Economic Considerations and Class Size", *Economic Journal*, 113, 34-63, 2003.
[34] Lazear, E.: "Educational Production", *Quarterly Journal of Economics*, 116, 777-803, 2001.
[35] Levin, J.: "For Whom the Reductions Count: A Quantile Regression Analysis of Class Size and Peer Effects on Scholastic Achievement", *Empirical Economics*, 26, 221–246, 2001.
[36] Oberhofer, W.: "The Consistency of Nonlinear Regression Minimizing The L_1 Norm", *The Annals of Statistics*, 10, 316-19, 1982.
[37] Porwoll, P.: *Class Size: A Summary of Research*, Arlington, VA: Educational Research Service, Inc., 1978.
[38] Powell, J.: "The Asymptotic Normality of Two Stage Least Absolute Deviations Estimators", *Econometrica*, 51, 1569–1575, 1983.
[39] Robinson, G. and Wittebols, J.: *Class Size Research: A Related Cluster Analysis for Decision Making*, Arlington, VA: Educational Research Service, Inc., 1986.
[40] Sakata, S.:"Instrumental Variable Estimator Based on Mean Absolute Deviation", working paper, U. Michigan, 2000.
[41] Strotz, R. and Wold, H.: "Recursive v.s. Nonrecursive Systems: an Attempt at Synthesis", *Econometrica*, 28, 417–427, 1960.
[42] Summers, A. and Wolfe, B.: "Do Schools Make a Difference", *American Economic Review*, 67, 1977.
[43] Zhao, Q.: "Asymptotically efficient median regression in the presence of heterocesdasticity of unknown form", *Econometric Theory*, 17, 765-84, 2001.

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN