Estimating a Class of Triangular Simultaneous Equations Models Without Exclusion Restrictions*

Roger Klein
Rutgers University

Francis Vella
Georgetown University

January 2006

Abstract

This paper provides a control function estimator to adjust for endogeneity in the triangular simultaneous equations model where there are no available exclusion restrictions to generate suitable instruments. Our approach is to exploit the dependence of the errors on exogenous variables (e.g. heteroscedasticity) to adjust the conventional control function estimator. The form of the error dependence on the exogenous variables is subject to restrictions, but is not parametrically specified. In addition to providing the estimator and deriving its large-sample properties, we present simulation evidence which indicates the estimator works well.

*We are grateful to participants at numerous seminars over the past four years for various comments which have resulted in improvements to the paper. We would also like to thank Ethel Fonseca for helpful comments and Bob Sherman for technical advice. We are particularly grateful to Whitney Newey for detailed comments which led to the current formulation of the problem. Any remaining errors are the sole responsibility of the authors.

1 Introduction

Instrumental variables (IV) is a method commonly employed in empirical applications for estimating models with endogenous regressors. However, while there is general agreement that IV is appropriate for a large class of models with endogeneity, there is frequently disagreement about the exclusion restrictions imposed in specific empirical applications. In fact, the difficulty in obtaining instruments has generated a rapidly growing and important literature related to inference in the presence of weak instruments (see, for example, Staiger and Stock 1999).

When the primary equation of interest contains an endogenous regressor, it is well known that IV is equivalent to an OLS regression that includes an additional regressor to control for endogeneity. Commonly, this additional variable or control is the reduced form residual for the endogenous regressor. In the linear case, as the control is a linear combination of the endogenous regressor and exogenous variables, the model is only identified in the presence of at least one exclusion restriction.[1]

In the above case the control's impact, as reflected by the residual's coefficient, is a constant that is estimated along with the parameters of interest. As a result, without further information, identification requires an exclusion restriction. However, when the error distribution depends on the exogenous variables, it is possible and in some sense natural to develop a control whose impact is not constant. In particular, as elaborated on below, we assume a generalized form of heteroscedasticity for both errors.[2] We then develop a "feasible" control whose impact is not constant and show that the model is identified without exclusion restrictions.

As discussed in section 3, other papers have explored identification via

[1] This control function approach is equivalent to two-stage-least-squares.

[2] Identification results are provided for both nonparametric and semiparametric specifications of the conditional variance functions. To obtain reasonable results at moderate sample sizes, most of this paper focuses on the semiparametric case.

second moments (e.g. Vella and Verbeek 1997, Rummery et al 1999, Sentana and Fiorentini 2001, Rigobon 2003 and Lewbel 2004). For the model that we consider, identification depends on there being heteroscedasticity in one or both equations of interest and that it "differs" across equations in a manner made precise below. The estimator is then based on estimating a semiparametric model of heteroscedasticity in each equation. For the structural equation of interest, such heteroscedasticity must be estimated simultaneously with the model's parameters as consistent residuals are unavailable. We do this in a setting where the conditional variance of each error is an unknown function of an index which needs to be estimated. While this semiparametric treatment of the unknown functions complicates the analysis, it avoids the reliance on parametric assumptions for identification.[3]

In the following section we outline the model. In section 3 we discuss the estimation method and how to implement it. Formal results are stated in section 4. This section also outlines the proof strategy for obtaining these results. Section 5 provides simulation evidence and section 6 concludes. The Appendix contains detailed proofs of all theorems and intermediate lemmas.

2 Model and Identification Sources

With θ_o and π_o as vectors of true parameter values, consider the following linear triangular model:

$$Y_{1i} = X_i\theta_{1o} + Y_{2i}\theta_{2o} + u_i \equiv Z_i\theta_o + u_i \quad (1)$$

$$Y_{2i} = X_i\pi_o + v_i, \quad (2)$$

where Y_{1i} and Y_{2i} are continuous endogenous variables; X_i is a vector of variables that are mean-independent of the error components u_i and v_i. We

[3]In the Appendix, we provide an identification result for both nonparametric and semiparametric models of heteroscedasticity.

further assume that these errors are correlated. We use the terms primary and secondary to refer to the first and second equations respectively. The main objective of estimation is to conduct inference on θ_o, the vector of true parameter values in the primary equation. Notice that the model allows the same $X's$ in both equations without imposing any restrictions on the parameter values.

When the errors do not depend on X, the (linear) relation between errors is captured by the following unconditional population regression:

$$a_o = \arg\min_a E\left[u - av\right]^2 \Rightarrow a_o = cov\left(u, v\right)/Var(v).$$

By construction, $\varepsilon \equiv u - a_o v$ is uncorrelated with v and therefore uncorrelated with Z, which provides the basis for the controlled regression:

$$Y_{1i} = Z_i\theta_o + a_o v_i + \varepsilon_i.$$

Provided that the matrix $[Z\ v\ ]$ has full column rank, the OLS estimator for this regression is consistent and would be implemented in practice by replacing v_i by the corresponding residual. However, in the absence of an exclusion restriction this full rank condition is not satisfied.

When the distribution of the errors depends on the X-variables, obtain the (linear) conditional relation between the errors by the following conditional population regression:[4]

$$\begin{aligned} A_o\left(X_i\right) &= \arg\min_A E\left[u_i - Av_i \mid X_i\right]^2 \Rightarrow \\ A_o(X_i) &= cov\left(u_i, v_i \mid X_i\right)/Var(v \mid X_i). \end{aligned}$$

In this case, $\varepsilon_i \equiv u_i - A_o\left(X_i\right) v_i$ is uncorrelated with v_i conditioned on X_i,

[4] We would like to thank Whitney Newey for this interpretation of the control.

which provides the basis for the controlled regression:

$$Y_{1i} = Z_i\theta_o + A_o(X_i)v_i + \varepsilon_i.$$

Let R be the matrix with ith row: $\mathrm{R}_i \equiv [Z_i \;\; A_o(X_i)v_i]$ and assume that A_o depends on X_i, which would be reasonable when the error distributions depend on X_i. In this case, the matrix R will have full column rank and identification will follow without exclusion restrictions.

As $A_o(X_i)$ is unknown, it must be estimated and restrictions must be imposed to obtain identification. Here, we explore the restrictions implied by a generalized form of heteroscedasticity. To this end, assume:

$$u_i \equiv S_{ui}u_i^*; \quad v_i \equiv S_{vi}v_i^*,$$

where

$$\begin{aligned} S_{ui}^2 &\equiv Var\left(u_i|X_i\right) \\ S_{vi}^2 &\equiv Var\left(v_i|X_i\right) \\ E\left(u_i|X_i\right) &= E(v_i|X_i) = 0. \end{aligned}$$

Further, there is a constant relation between unscaled error components:[5]

$$\rho_o \equiv E\left(u_i^*v_i^*|X_i\right) = E\left(u_i^*v_i^*\right).$$

Subject to the above restrictions, for observation i the error components can arbitrarily depend on X_i. With the correlation ρ_o constant, the control is given as:

$$A_o(X_i)v_i = \rho_o\left[S_{ui}/S_{vi}\right]v_i.$$

Before discussing how to implement the above control, note that if the

[5]Note that Bollerslev (1990) also employs a constant correlation assumption in a time-series context.

scaling functions are known or can be consistently estimated, then identification holds if these scaling functions "differ" in that for observation i, $S_{ui}/\ S_{vi}$ depends on X_i. As a specialized interpretation, view u_i^* and v_i^* as unobserved variables with non-constant impacts that depend on the endogenous variables. These impacts differ and are functions of X_i given by S_{ui} and S_{vi} respectively. For this type of error structure, parametric forms of conditional variance functions have been employed in a variety of applications, and we expect there will be a number of others where it will be relevant and reasonable.

Other papers exploit second moment information as a source of identification. Vella and Verbeek (1997) and Rummery et al (1999) develop an estimation procedure based on the rank order of an individual's position in the reduced form residual distribution for subsets of the data. The variable determining the selection of subsets is also assumed to be responsible for the heteroscedasticity. In the context of normal factor models, Sentana and Fiorentini (2001) examine heteroscedasticity as a source of identification. Rigobon (2003) formulates a model in which there are two known regimes. The parameters of interest and the covariance between the equations' errors do not depend on the regime indicator. However, the error variances do depend on the known regime indicator. Employing an error covariance restriction similar to that in Rigobon, Lewbel (2004) examines a model of heteroscedasticity with second moment information depending on a known vector of variables Z. As Z may coincide with X, for comparative purposes we focus on this case and without loss of generality take $E(X) = 0$. He then considers a model in which:

$$E\left(X_i u_i v_i\right) = E\left[X_i E\left(u_i v_i | X_i\right)\right] = 0;\ E\left(X_i v_i^2\right) \neq 0.$$

The model considered here differs in several respects from those above. First, for the model outlined earlier, $E\left(u_i v_i | X_i\right)$ depends on X_i . Consequently, while the first restriction above may hold in special cases, it will not

hold in general for the model considered here. Second, with the conditional covariance and variance functions depending on X_i, here the conditional variance of each error is modeled as an unknown function of an index.

In a different model, Klein and Vella (2004) we also exploit heteroscedasticity to estimate a triangular treatment model where the endogenous regressor is binary. To flexibly model both the shape and conditional variance for the error distribution in the binary model, a double index formulation is employed. In so doing, with the estimated binary response probability as an instrument, the model is "well-identified" without exclusion restrictions. While the treatment paper is related to this paper, the identification and estimation strategies are fundamentally different from those employed here.

Finally note that the use of instruments in the absence of exclusion restrictions is not limited to cases of heteroscedasticity. Dagenais and Dagenais (1997) and Lewbel (1999) also discuss estimation of models where there are endogenous regressors and no exclusion restrictions. They show that when there is measurement error of a specific form one is able to use instruments based on the higher powers of the included variables. The model and estimator presented below both differ from the approaches in these papers.

3 The Estimator: Implementing Strategies

3.1 The Secondary Equation

Before presenting the main results, this section outlines and motivates the estimation strategy. From the above discussion, we will require residuals and the conditional variance function for the secondary equation. Accordingly, first obtain consistent estimates of the secondary equation's conditional mean parameter values by regressing Y_2 on X to get $\widehat{\pi}$. We then estimate the

residuals as:[6]

$$\widehat{v} = Y_2 - X\widehat{\pi}.$$

To estimate S_v, we impose a single index structure:[7]

$$S^2_{v_i} \equiv E\left[v_i^2 \mid X_i\right] = E\left[v_i^2 \mid I_{vi}\left(\delta_o\right)\right],$$

where $I_{vi}\left(\delta_o\right) \equiv X_{1i} + X_{2i}\delta_o$. Next, estimate δ_o using semiparametric least squares with $\widehat{v}_i^2$ as the dependent variable (see Ichimura, 1993). Namely:

$$\hat{\delta} = \arg\min_{\delta} \sum \hat{\tau}_i \left[\hat{v}_i^2 - \hat{E}\left(\hat{v}_i^2 \mid I_{vi}\left(\delta\right)\right)\right]^2,$$

where $\hat{\tau}_i$ is a trimming function that restricts X_i to a compact set depending on sample quantiles. Employing the estimated index:

$$\hat{S}^2_{vi} = \widehat{E}\left(\widehat{v}_i{}^2 \mid I_{vi}\left(\hat{\delta}\right)\right)$$

where $\widehat{E}$ is a non-parametric estimator for the indicated conditional expectation. Employing the above initial estimator $\hat{S}_{vi}$, we then repeat the above process in a GLS step.[8] For notational convenience below, denote the vector of parameter estimates as: $\hat{\eta} \equiv \left(\hat{\pi}' \ \hat{\delta}'\right)'$. As our focus will be on the primary equation, we will refer to these parameters as nuisance parameters.

3.2 The Primary Equation

As consistent residuals are not available for the primary equation, the conditional variance function for this equation and the parameters of interest are

[6] These residuals could be obtained in a more general nonparametric or semiparametric regression.

[7] As discussed below, identification also holds under a nonparametric formulation.

[8] While it is possible to avoid a GLS step, we have found that the estimator for S_{vi} based on $\hat{\pi}_{GLS}$ is improved and that there is a corresponding improvement in the estimates of the primary equation of interest.

estimated simultaneously. In so doing, we distinguish two cases according to whether or not the conditional variance function has an index structure. Identification arguments for the primary equation do not depend on whether or not an index structure is imposed on the conditional variance function for the secondary equation. However, for the primary equation, identification arguments are quite different depending on whether or not an index structure is imposed on its conditional variance function. As will become clear below, while the index case can be expected to perform better in practice, it is more difficult to formulate and analyze an estimator for this case. Beginning with a nonparametric formulation for the conditional variance function, let $Z_i \equiv [X_i, Y_{2i}]$ and define:

$$\begin{aligned} u_i(\theta) &\equiv (Y_{1i} - Z_i\theta) \\ \hat{S}^*_{ui}(\theta)^2 &\equiv \hat{E}\left[u_i^2(\theta) \mid X_i\right]. \end{aligned}$$

With $\beta \equiv (\theta, \rho)$ and $i = 1, ..., N$ observations, let:

$$\begin{aligned} \hat{A}_i(\beta) &\equiv \rho\left[\hat{S}^*_{ui}(\theta)/\hat{S}_{vi}\right] \\ \hat{M}_i(\beta) &\equiv W_i\theta + \hat{A}_i(\beta)\hat{v}_i \\ \hat{Q}(\beta) &\equiv \frac{1}{N}\sum_i \hat{\tau}_i\left[Y_{1i} - \hat{M}_i(\beta)\right]^2. \end{aligned}$$

An estimator for the primary equation is now defined as:

$$\hat{\beta} \equiv \arg\min_{\beta} \hat{Q}(\beta).$$

Conditioning on X_i, assume that the conditional correlation between u_i^* and v_i^* is constant. Theorem 2a then provides consistency and identification results for this nonparametric case under a full rank condition. For this case, the conditional variance function from the secondary equation may be taken as known, and it is therefore irrelevant (for theoretical purposes) whether or

not it satisfies an index condition. However, the structure of the conditional variance function in the primarily equation does have a critical impact on the identification argument.

To obtain reliable parameter estimates at moderate sample sizes, the remainder of this paper imposes an index structure on both conditional variance functions. For reasons discussed below, identification becomes more difficult in an index formulation for the conditional variance function of the primary equation. In this case, the following index restriction holds at the true parameter values:

$$\begin{aligned} E\left[u_i^2\left(\theta_o\right) \mid X_i\right] &= E\left[u_i^2 \mid I_{ui}\left(b_o\right)\right], \\ I_{ui}\left(b_o\right) &\equiv X_{1i} + X_{2i} b_o. \end{aligned}$$

For whatever objective function that is employed, for purposes of identification it is important that the set of potential minimizers satisfy an index restriction. As an example of an objective function that implies this restriction, for illustrative purposes suppose that we knew θ_o. In this case, for the primary equation it would only remain to estimate the index parameter values of the conditional variance function. Employ SLS as was done for the secondary equation to obtain:

$$\hat{b} = \arg\min \hat{S},\ \hat{S} \equiv \frac{1}{N} \sum \hat{\tau}_i \left[u_i^2\left(\theta_o\right) - \hat{E}\left(\ u_i^2 | I_{ui}\left(b\right)\ \right)\right]^2.$$

It can be shown that $\hat{\mathrm{S}}$ is uniformly close to

$$\frac{1}{N} E \sum \tau_i \left[u_i^2 - E\left(\ u_i^2 | I_{ui}\left(b\right)\ \right)\right]^2.$$

Let b^* be a minimizing value of this objective function. As a necessary condition for a minimum, it can be shown that the following index restriction

must be satisfied:[9]

$$E\left(\ u_i^2\left(\theta_o\right)|I_{ui}\left(b^*\right)\ \right) = E\left(\ u_i^2\left(\theta_o\right)|X\ \right) = E\left(\ u_i^2\left(\theta_o\right)|I_{ui}\left(b_0\right)\ \right).$$

With this restriction holding away from the truth, the set of potential minimizers is sufficiently reduced so as to enable an identification argument.

With θ_o being unknown, identification becomes problematic as it is difficult to impose an index restriction away from the truth. To illustrate both the problem and a solution, return to the objective function employed to obtain nonparametric identification, with the appropriate modifications made to accommodate an index structure. In examining this case, for purposes of exposition, throughout we take N to be large and discuss the problem in terms of population objective functions. Let

$$S_{ui}^2\left(\theta, b\right) \equiv E\left(u_i^2\left(\theta\right)|I_{ui}\left(b\right)\right).$$

Then, similar to the strategy for the nonparametric case above, write $\|Q\| \equiv \Sigma Q_i^2/N$ and with $\alpha \equiv (\theta, b, \rho)$ define:

$$\begin{aligned} M_{1i}\left(\alpha\right) &\equiv Z_i\theta + \rho\left[S_{ui}\left(\theta, b\right)/S_{vi}\right]v_i \\ Q_1\left(\alpha\right) &\equiv Q\left(M_1\right) \equiv \|\hat{\tau}\left[Y_1 - M_1\right]\| \\ \alpha^* &\equiv \arg\min E\left[Q_1\left(\alpha\right)\right]. \end{aligned}$$

[9] Write:

$$\begin{aligned} &E\left(\left[u_i^2 - E\left(\ u_i^2|I_{ui}\left(b^*\right)\ \right)\right]^2|X\right) = \\ &E\left(\left[u_i^2 - E\ \left(u_i^2|X\right)\right]^2|X\right) + \\ &\left(\left[E\ \left(u_i^2|X\right) - E\left(\ u_i^2|I_{ui}\left(b^*\right)\ \right)\right]^2\right). \end{aligned}$$

The second term above attains a minimum of zero when $b^* = b_o$. Therefore, for any minimum, b^* :

$$E\left(u_i^2|X\right) = E\left(\ u_i^2|I_{ui}\left(b^*\right)\ \right)$$

on a set where $\tau \neq 0$.

With an orthogonality condition holding between $Y_1 - M_1(\alpha_o)$ and $[M_1(\alpha^*) - M_1(\alpha_o)]$, it can be shown that for any candidate for a minimum, α^*, must satisfy:

$$M_1(\alpha^*) - M_1(\alpha_o) = 0.$$

In other words:

$$Z_i(\theta^* - \theta_o) + [\rho^* S_{ui}(\theta^*, b^*) - \rho_o S_{ui}(\theta_o, b_o)]\, v/S_{vi} = 0.$$

With additional information relating $S_{ui}(\theta^*, b^*)$ to $S_{ui}(\theta_o, b_o)$, the identification strategy would be greatly simplified. For example, if minimizing values satisfied: $S_u(\theta^*, b^*) = S_u(\theta_o, b_o)$, then identification would readily follow. In this case, let R be the matrix with i^{th} row: $\mathrm{R}_i \equiv [X_i \quad Y_{2i} \quad (S_{ui}(\theta_o, b_o)/S_{vi})\, v_i\ ]$. From above:

$$R \begin{bmatrix} \theta^* - \theta_o \\ \rho^* - \rho_o \end{bmatrix} = 0.$$

Accordingly, identification would follow from a full (column) rank assumption on R.

While it does not appear possible to guarantee the strong index restriction: $S_{ui}(\theta^*, b^*) = S_{ui}(\theta_o, b_o)$ apriori as in the above example, it is possible to modify the objective function so as to ensure that the set of minimizers is sufficiently restricted to yield identification. To this end, let:

$$\begin{aligned} S^{*2}_{ui}(\theta, b) &\equiv E\left(u_i^2(\theta) \,|I_{ui}(b), I_{vi}\right) \\ M_{2i}(\theta, b) &\equiv Z_i\theta + \rho\,[S^*_{ui}(\theta, b)/S_{vi}]\, v_i \\ Q_2(\theta, b, \rho) &\equiv Q(M_2) \equiv \|\hat{\tau}\,[Y_1 - M_2]\|. \end{aligned}$$

Then, with $\alpha \equiv (\theta, b, \rho)$ consider the "overall" population objective function:

$$Q(\alpha) \equiv Q_1(\alpha) + Q_2(\alpha).$$

Denote α^* as a minimizer for $Q(\alpha)$. In the Appendix it is shown that α_o, the vector of true parameter values, is a minimizer not only for Q but also separately for Q_1 and Q_2. It follows that α^* must also minimize each of these component objective functions. As a result, α^* must satisfy minimizing conditions implied by minimizing each separate objective function. Taken together, we show below that these restrictions and a full rank condition suffice to establish that α_o is the unique minimizer.

To indicate the nature of these restrictions, in an argument similar to that above we show that α^* must satisfy the restrictions: $M_k(\alpha^*) = M_k(\alpha_o), k = 1, 2$. Therefore, with $X_i^* \equiv [I_{ui}(b^*), I_{vi}]$:

$$\begin{aligned}
(Ra) &: Z_i(\theta^* - \theta_o) + [\rho^* S_{ui}(\theta^*, b^*) - \rho_o S_{ui}(\theta_o, b_o)] v_i / S_{v_i} = 0 \\
(Rb) &: Z_i(\theta^* - \theta_o) + [\rho^* S_{ui}^*(\theta^*, b^*) - \rho_o S_{ui}(\theta_o, b_o)] v / S_{vi} = 0 \\
(Rc) &: E\left[u_i^2(\theta^*) | I_{ui}(b^*)\right] = E\left[u_i^2(\theta^*) | X_i^*\right],
\end{aligned}$$

where the index restriction in (Rc) follows by differencing the first two restrictions and employing the definitions of S_{ui}^* and S_{ui}. In Theorem 2b below, we show these restrictions in conjunction with a full rank assumption are sufficient to provide identification.

From the above discussion, we are motivated to formulate the following estimator for the primary equation under an index structure. Recall that

$$\begin{aligned}
S_{ui}^*(\theta, b)^2 &\equiv E\left[u_i^2(\theta) | I_{ui}(b), I_{vi}\right] \\
S_{ui}(\theta, b)^2 &\equiv E\left[u_i^2(\theta) | I_{ui}(b)\right] \\
S_{vi}^2 &\equiv E\left[v_i^2(\theta) | I_{vi}(\delta_o)\right].
\end{aligned}$$

As defined in the next section, let $\hat{S}_{ui}^*(\theta, b)$, $\hat{S}_{ui}(\theta, b)$, and $\hat{S}_{vi}$ be estimators for the above functions obtained from semiparametric regressions. Obtain $\hat{Q}_k$ from Q_k by replacing known functions with the above estimators, $k = 1, 2$. Then, with $\alpha \equiv [\theta, \rho, b]$, the estimator for the primary equation is now defined

as:[10]

$$\hat{\alpha} \equiv \arg\min_{\alpha} \hat{Q}(\alpha), \ \hat{Q}(\alpha) \equiv \sum_{k=1}^{2} \hat{Q}_k(\alpha).$$

4 Assumptions, Definitions, and Results

In obtaining asymptotic results, we make the following assumptions:

A1 The vector $(Y_{1i}, Y_{2i}, X_i, u_i, v_i)$ is i.i.d distributed over i, with the variables X_i being bounded.[11]

A2 The parameter vector: $\gamma \equiv (\pi, \theta, \delta, b, \rho)$ is in a compact parameter space, Θ, where γ_o is in the interior of Θ.

A3 Write the error components as:

$$\begin{aligned} u &\equiv S_u u^*, \ S_u^2 \equiv Var(u|X) \\ v &\equiv S_v v^*, \ S_v^2 \equiv Var(v|X); \end{aligned}$$

Assume:

$$E(u|X) = E(v|X) = 0$$

$$\rho_o \equiv E(u^* v^*|X) = E(u^* v^*), \ 0 < \rho_o < 1.[12]$$

A4 Let f be the density of either u^2 or v^2. Assume there exists $c > 0$ such

[10] It would be interesting to explore a variance minimizing weighting for these objective functions. However, this extension is beyond the scope of the present paper.

[11] While it is possible to handle the unbounded case, uniform convergence arguments are simplified under this assumption.

[12] If the errors are not sufficiently different ($\rho_o = 1$), we are unable to identify the parameters. If $\rho_o = 0$, the main parameters of interest are identified, but the index parameters of the primary equation are not identified by the estimator for the primary equation. Identification of index parameters from squared (primary-equation) residuals would hold when $\rho_o = 0$.

that for $t > c,: f$ satisfies the tail condition:[13]

$$f(t) \leq 1/\left[1+t^2\right]^{(r+1)/2},\ r\ \geqq 5.$$

A5 Write $X_i \equiv [X_{1i}\ X_{2i}\ X_{3i}]$, where X_{1i} and X_{2i} are continuous variables that are not functionally related. Assume that I_{ui} depends on X_{1i} and that I_{vi} depends on X_{2i}. Then, without imposing any exclusion restrictions on variables entering the indices, write the normalized linear indices as:

$$\begin{aligned} I_{ui}(b_o) &\equiv X_{1i} + [X_{2i}\ X_{3i}]\, b_o \\ I_{vi}(\delta_o) &\equiv X_{2i} + [X_{1i}\ X_{3i}]\, \delta_o. \end{aligned}$$

With $b_o \equiv [b_{1o}\ b_{2o}]$ and $\delta_o \equiv [\delta_{1o}\ \delta_{2o}]$ assume 1-$b_{2o}\delta_{1o} \neq 0$ and that :

$$\begin{aligned} S_u^2(\theta_o, b_o) &\equiv E\left(u(\theta_o)^2 | X\right) = E\left[u(\theta_o)^2 \mid I_{ui}(b_o)\right] > 0 \\ S_v^2(\delta_o) &\equiv E\left(v(\pi_o)^2 | X\right) = E\left[v(\pi_o)^2 \mid I_{vi}(\delta_o)\right] > 0. \end{aligned}$$

For X in a compact set, these functions and their first six derivatives are uniformly bounded. Further, each index depends on a continuous variable.

A6 Referring to (A5), assume that the joint conditional density $g(x_1, x_2|x_3)$ is bounded away from zero on the interior of its support and has bounded derivatives up to the sixth order.

A7 For estimating expectations and densities, assume that the kernel func-

[13] This assumption, which guarantees that both u^2 and v^2 have at least 4 moments, is stronger than is needed. We require uniform convergence results involving sample means whose elements u^2 and v^2, may be unbounded random variables. As in Ichimura (1993), we can reduce the required order of the kernel if we assume that at least the 3^{rd} moment of these variables is bounded. Under the stronger tail assumption employed here, we can simplify the proofs in addition to lowering the order of the required kernel.

tion, K, is a symmetric density with up to 4 bounded derivatives.

A8 Let R be a matrix with i^{th} row: $[X_i,\ Y_{2i},\ (S_{ui}/S_{vi})\, v_i]$ and assume that R has full column rank.

As noted above in a footnote, (A4) is stronger than is needed. This assumption simplifies uniform convergence proofs pertaining to unbounded random variables and makes it possible to reduce the required order of the kernels employed. If the standardized errors u^* and v^* are independent of X, then it is possible to relax this assumption significantly.[14] Most of the remaining assumptions are somewhat standard, with the last assumption required for identification (see Theorem 2 of this section). In addition to these assumptions, we need to define the estimators and a bias reduction device used to establish asymptotic normality. In defining the kernels used below, we are motivated to employ kernels that provide the required degree of bias reduction, perform reasonably well in finite samples, and for which tedious detail in the resulting proofs is minimized. We have found that twicing kernels (see Newey et. al. (2004)) satisfy these objectives under an appropriate trimming sequence.[15]. Assumptions (A4-5) are useful in obtaining bias expansions for the components of a nonparametric expectations estimator. For example, the denominator of an expectations estimator involves the joint density of an index. Derivatives of this density need to be bounded up to the sixth order. Assumption (A6) guarantees that this is the case.

[14] In this case, a variance function can be recovered up to a multiplicative constant (which is all that is required) with minimal assumptions on higher moments as follows. Let $W \equiv u^{2/m}$ or $v^{2/m}$ and estimate the variance function up to a multiplicative constant as:

$$\left[\hat{E}\,(W)\right]^m.$$

[15] It is possible to follow a mixed strategy with single index components being estimated under locally-smoothed kernels. Under this strategy, double index components would be estimated with a twicing kernel. While the finite sample performance of the estimator might be improved under this strategy, the resulting proofs would be significantly longer.

D1 Let $\{W_i, I_i\}$ be i.i.d., where I_i is a single index upon which W_i depends. Then, the estimator for the expectation of W_i conditioned on I_i is given by:

$$\hat{E}\left(W_i|I_i\right) = \sum_{j\neq i} \frac{W_j}{Nh} K_1\left[(I_{iu} - I_{ju})/h\right] / \sum_{j\neq i} \frac{1}{Nh} K_1\left[(I_{iu} - I_{ju})/h\right],$$

where, with $k(w)$ a standard normal:

$$K_1(w) \equiv 2k(w) - \int k(w-v)k(v)dv.$$

With s as the standard deviation of I, the window h is given by $h = sN^{-r}$, $1/8 < r < 2/15$.[16]

D2 Let $\{W_i, I_{1i}, I_{2_i}\}$ be i.i.d., where I_{1i} and I_{2i} are indices upon which W_i depends. Then, the estimator for the expectation of W_i conditioned on (I_{1i}, I_{2_i}) is given by

$$\hat{E}\left(W_i|I_{1i}, I_{2i}\right) = \frac{\sum_{j\neq i} \frac{W_j}{Nh_1h_2} K_2\left[(I_{1i} - I_{1j})/h_1\right] K_2\left[(I_{2i} - I_{2j})/h_2\right]}{\sum_{j\neq i} \frac{1}{Nh_1h_2} K_2\left[(I_{1i} - I_{1j})/h\right] K_2\left[(I_{2i} - I_{2j})/h\right]},$$

where with $K_1(w)$ given as in (D1):

$$K_2(w) \equiv 2K_1(w) - \int K_1(w-v)K_1(v)dv.$$

[16] The lower limit on r is required for bias control. Namely, for this kernel, numerator and denominator of the conditional expectation each has a bias uniformly of order h^4, and we require that the bias vanish faster than $N^{-1/2}$. The tight upper bound is required to establish uniform convergence of a second derivative when the conditional expectation is being applied to an unbounded random variable. This condition can be relaxed under either a kernel of higher order than K_1 or under a more restrictive tail condition than that in (A4).

With s_k as the standard deviation for I_k, set $h_k = s_k N^{-r}$, $1/12 < r \leq 1/10$.

To insure that various estimated denominators are bounded away from zero in large samples, we employ a trimming function that restricts the components of X to a compact set depending on estimated sample quantiles. As a result, the trimming function should be viewed as being estimated. [17] This trimming function is given in (D3).

D3 Indicator Trimming. Let $\underline{c}_k$ and $\bar{c}_k$ be lower and upper population quantiles for $X_{ik}, k = 1, ..., K$. Let q_o be the vector of these quantiles. With x: 1xK, define $\mathcal{P} \equiv \{x : \underline{c}_k < x_k < \bar{c}_k,\ k = 1, ..., K\}$. With $X_i \equiv [X_{i1}, ...X_{iK}]$, define the trimming indicator:

$$\tau_{ix} \equiv \ \tau_i(q_o) \equiv \ \{\ X_i \in \mathcal{P}\ \}.$$

With $\hat{q}$ as a vector of sample quantiles, the estimated trimming function is given as: $\hat{\tau}_{xi} \equiv \tau_i(\hat{q})$.

D4 Y_2-Model. Let $\hat{\pi}$ be the GLS estimator from the regression of Y_2 on X.[18] Define the residual:

$$\hat{v} \equiv Y_2 - X\hat{\pi}.$$

[17] For one of the gradient components (Lemma GA), estimated trimming may be taken as known under a result due to Pakes and Pollard (1989). In other gradient components (Lemma GB), under standard convergence arguments estimated trimming may be taken as known.

[18] First, obtain OLS residuals $\hat{v}_i$. Second, obtain $I_{vi}\left(\hat{\delta}\right)$, from the SLS estimator of δ_o. Next, define estimate $\hat{S}^2_{vi}$:

$$\hat{S}^2_{vi} = \hat{E}\left(\hat{v}_i^2 \mid I_{vi}\left(\hat{\delta}\right)\right).$$

Reweighting observations in the Y_2 model provides the GLS estimator of π_o. All of the results in this paper hold using the OLS estimator of π_o. The finite sample properties of the estimator for the Y_1 model are improved by employing the GLS estimator for the secondary equation.

The estimated index parameters of the conditional error variance are then given by:

$$\hat{\delta} = \arg\min_{\delta} \hat{R}(\delta), \ \hat{R}(\delta) \equiv \sum_{i=1}^{N} \hat{\tau}_i \left[\hat{v}_i^2 - \hat{E}\left(\hat{v}_i^2 \mid I_{vi}(\delta) \right) \right]^2 / N,$$

where $\hat{E}$ is a nonparametric estimated expectation defined above. With $\hat{\eta} \equiv \left(\hat{\pi}, \hat{\delta} \right)$ estimating $\eta_o \equiv (\pi_o, \delta_o)$, we refer to $\hat{\eta}$ as the (nuisance) vector of estimates for the secondary equation.

In estimating the model above, the monte-carlo results were improved under a two-stage trimming strategy similar to but less complicated than that in Klein and Spady (1993). Namely, in the first stage obtain estimates under X-trimming as described above with $\hat{\tau}_i \equiv \hat{\tau}_{xi}$. In a second-stage, re-estimate the model with index trimming and minimal (for technical reasons) X-trimming. By targeting the problem at its source, such index trimming provides a better control for small values of the index density.[19] In either case, as argued below, the trimming function can be taken as known.

Employing the above definitions, it is now possible to define the estimated conditional variance from the Y_2-Model.

D5 Estimated Conditional Variance. With $\hat{\delta}$ given in (D4) and with expectations estimated under the kernel in (D1):

$$\hat{S}_{vi}^2 \equiv \left| \hat{E}\left(\hat{v}_i^2 | I_{vi}\left(\hat{\delta} \right) \right) \right|.$$

[19] In the first stage, we trimmed on the basis of the .975 upper and .025 lower sample quantiles of the continuous X-variables. In the second stage, let $\hat{\tau}_{iv}$ denote trimming on the basis of the .975 and .025 index sample quantiles. Denote $\hat{\tau}_{ix}$ as a trimming function under "minimal" X-trimming, where the .995 upper and .005 lower quantiles formed the basis for such trimming. The trimming function employed in the second stage is then given by the product: $\hat{\tau}_{iv}\ \hat{\tau}_{ix}$. As an alternative to minimal X-trimming, it is possible to modify the expectations estimator as in Klein and Spady (1993).

Notice that absolute values are employed above. While the estimated expectation is positive under "regular" kernels, it can be negative under higher order kernels. While this is not a problem asymptotically, in any finite sample there can be a small fraction of observations for which the estimated expectation is negative. In estimating the primary, Y_1-model, we will smoothly trim out observations for which this problem occurs.[20] For this purpose, we employ the following smooth trimming function.

D6 Smooth Trimming. With $\hat{\sigma}_i^2 \equiv \hat{E}\left(\hat{v}_i^2 | I_{vi}\left(\hat{\delta}\right)\right)$, define:

$$\hat{\tau}_{si} \equiv \tau\left(\hat{\sigma}_i^2\right) = \left[1 + \exp\left(\ - a_n \hat{\sigma}_i^2\right)\right]^{-1}, \ a_n = Ln(N)^2.$$

The function above will tend to 0 as $\hat{\sigma}_i^2$ becomes negative and to 1 otherwise. As this function approximates an indicator, its derivative must become high in a neighborhood of zero. To control for the magnitude of the derivative, the slowly increasing sequence a_n is selected above. Note that this trimming function is based on the estimated index obtained from estimating the secondary equation and does not depend on any of the unknown parameter values for the primary equation.

D7 Y_1-Model. With the Y_1- model given as:

$$Y_1 = [X \ \ Y_2]\,\theta_o + u \equiv Z\theta_o + u,$$

[20] When trimming is based on an estimated linear index, the lemma in Pakes and Pollard applies to indicator trimming. However, when the trimming involves nonparametric expectations, it does not appear that the required Euclidean property of this lemma holds for indicator trimming. Moreover, we have found trimming to be important for those few but influential observations where estimated variance functions are negative. Accordingly, a smooth trimming function is employed to control for this problem.

define $u(\theta) \equiv Y_1 - Z\theta$ and let:

$$\begin{aligned} \hat{S}^2_{ui}(\alpha) &\equiv \left| \hat{E}\left(u_i^2(\theta) \,|I_{ui}(b)\right)\right| \\ \hat{S}^{*2}_{ui}(\alpha) &\equiv \left| \hat{E}\left(u^2(\theta) \,|I_u(b), \hat{I}_v\right)\right| \end{aligned}$$

where the first, single-index component is obtained under the kernel in (D1) and the second, double-index component is obtained under the kernel in (D2). Let:

$$\hat{S}_{ui}(\alpha; k) \equiv \hat{S}_{ui}(\alpha), k = 1;\ \hat{S}_{ui}(\alpha; k) \equiv \hat{S}^*_{ui}(\alpha), k = 2.$$

Then, for $k = 1, 2$:

$$\begin{aligned} \hat{M}_{ik}(\theta, b, \rho) &\equiv Z\theta + \rho\left[\hat{S}_i/\hat{S}_{vi}\right] v_i, \\ \hat{Q}_k(\theta, b, \rho) &\equiv \sum \hat{\tau}_{si}\hat{\tau}_i \left[Y_1 - \hat{M}_{ik}\right]^2 /N. \end{aligned}$$

Then, with $\alpha \equiv (\theta, b, \rho)$ and $\hat{Q} \equiv \hat{Q}_1 + \hat{Q}_2$:

$$\hat{\alpha} = \arg\min_{\alpha} \hat{Q}(\alpha).$$

As with the secondary equation, trimming is based on a two-stage process. Namely, $\hat{\tau}_i$ is obtained under X-trimming in the first stage at the same levels as for the secondary equation (see the discussion following (D4) above). In the second stage, $\hat{\tau}_i$ is a product of index and minimal X-trimming trimming similar to that for the secondary equation discussed above.

Employing the above assumptions and definitions, the Appendix provides all proofs for asymptotic results. In the remainder of this section, we summarize these results and provide a brief outline of the proof strategy.

Beginning with the secondary equation (Y_2-Model), Theorem 1 provides the large sample results for the estimators of the nuisance parameters.

Theorem 1 (The $\mathbf{Y_2}$-Model). Under the above assumptions and definitions, estimates of regression and index parameters satisfy the characterizations:

$$\sqrt{N}[\hat{\pi} - \pi_o] = \sqrt{N} \sum_{i=1}^{N} \varepsilon_{\pi i}/N \tag{a}$$

$$\sqrt{N}\left[\hat{\delta} - \delta_o\right] = \sqrt{N} \sum_{i=1}^{N} \varepsilon_{\delta i}/N + o_p(1), \tag{b}$$

where $\varepsilon_{\pi i}$ and $\varepsilon_{\delta i}$ each are $i.i.d.$ with 0 expectation and finite variance.

The first result above is immediate and the second follows from a standard Taylor series argument and Ichimura (1993). This second result also follows from the same type of U-statistic arguments used to establish asymptotic normality for estimator of the primary equation .

For the Y_1-Model, Theorem 2 below provides the consistency/identification result.

Theorem 2 (Consistency: the $\mathbf{Y_1}$-Model). With $\alpha_o \equiv (\theta_o, \rho_o, b_o)$ and $\hat{\alpha} \equiv \left(\hat{\theta}, \hat{\rho}, \hat{b}\right)$, under the above assumptions and definitions:

$$\hat{\alpha} \xrightarrow{p} \alpha_o.$$

To outline the consistency argument, which is provided in detail in the Appendix, recall from (D7) that:

$$\hat{\alpha} = \arg\min_{\alpha} \hat{Q}(\alpha).$$

Referring to (D7), obtain $M_{ik}(\alpha)$ from $\hat{M}_{ik}(\alpha)$ by replacing all estimated functions with their uniform probability limits. Then, define $Q(\alpha)$ by re-

placing $\hat{M}_{ik}(\alpha)$ in $\hat{Q}(\alpha)$ with $M_{ik}(\alpha)$. It can be shown that

$$\left|\hat{Q}(\alpha) - Q(\alpha)\right| \text{ and } |Q(\alpha) - E[Q(\alpha)]|$$

each converge in probability, uniformly in α to zero. Consistency follows if $E[Q(\alpha)]$ is uniquely minimized at α_o. From an orthogonality condition between $Y_{1i} - M_{ik}(\alpha_o)$ and $[M_{ik}(\alpha_o) - M_{ik}(\alpha)]$, any minimizing value of EQ must satisfy:

$$M_i(\alpha_o) - M_{ik}(\alpha) = 0, \ k = 1, 2.$$

Clearly, $\alpha = \alpha_o$ is a minimizer. Under a constant correlation assumption, in the appendix we establish identification (uniqueness) when the matrix $[X, Y_2, (S_u/S_v)v]$ has full column rank. The theorem in the appendix provides this result for both nonparametric and semiparametric specifications of conditional variance functions.

Theorem 3 below, which is proved in the appendix, provides the normality result.

Theorem 3 (Normality: the $\mathbf{Y_1}-$Model). Under the assumptions and definitions above:

$$\sqrt{N}[\hat{\alpha} - \alpha_o] \xrightarrow{d} Z, \ Z \tilde{} N(0, \Sigma).$$

To outline the argument, note that under a standard Taylor series argument for the gradient[21] to the objective function and a uniform convergence argument for the Hessian, normality will follow if the normalized gradient is asymptotically distributed as normal. To establish this result, in the Appendix it is established that the gradient has an i.i.d. sample mean characterization to which a standard central limit applies. To outline the argument, for

[21] With all estimating expectations functions converging uniformly to positive functions, this expansion is valid on a set with probability approaching one.

expositional purposes, neglect first-stage estimation, which matters,[22] but poses no technical difficulties. With $\hat{w}_{ik} \equiv \hat{\tau}_{is}\nabla_\alpha \hat{M}_{ik}$ termed a weight function, define:

$$\begin{aligned}
\sqrt{N}\hat{G}_k &\equiv -\sqrt{N}\sum \hat{\tau}_i \left[Y_{1i} - M_i\right] \hat{w}_{ik}/N + \sqrt{N}\sum \hat{\tau}_i \left[\hat{M}_{ik} - M_i\right] \hat{w}_{ik}/N \\
&\equiv \sqrt{N}\hat{G}_{Ak}\left(\alpha_o\right) + \sqrt{N}\hat{G}_{Bk}\left(\alpha_o\right), \ k = 1, 2. \\
\sqrt{N}\hat{G}_A &\equiv \sqrt{N}\left[\hat{G}_{A1} + \hat{G}_{A2}\right]; \ \sqrt{N}\hat{G}_B \equiv \sqrt{N}\left[\hat{G}_{B1} + \hat{G}_{B2}\right].
\end{aligned}$$

The normalized gradient to the objective function is then given by:

$$\sqrt{N}\hat{G} = \sqrt{N}\hat{G}_A + \sqrt{N}\hat{G}_B$$

For the A-component, from results in Pakes and Pollard (1989) and mean-square convergence arguments, Lemma GA shows that the estimated trimming ($\hat{\tau}_i$) and weight ($\hat{w}_{ik}$) functions may be taken as known. Accordingly:

$$\sqrt{N}\hat{G}_{Ak}\left(\alpha_o\right) = \sqrt{N}\sum \tau_i \left[Y_{1i} - M_{oi}\right] w_{ik} + o_p(1).$$

With

$$\varepsilon_{Ai} \equiv \tau_i \left[Y_{1i} - M_{oi}\right] \left[w_{i1} + w_{i2}\right]$$

it then follows that:

$$\sqrt{N}\hat{G}_A = \sqrt{N}\bar{\varepsilon}_A + o_p\left(1\right), \ \bar{\varepsilon}_A \equiv \sum \varepsilon_{Ai}/N.$$

For the B-component, Lemma 5 shows that the estimated trimming and weight functions may be taken as known:

$$\sqrt{N}\hat{G}_{Bk} = \sqrt{N}\sum \tau_i \left[\hat{M}_{ik} - M_{ik}\right] w_{ik}/N + o_p(1).$$

[22] See Newey and McFadden (1994).

Lemmas 5 and GB of the Appendix establishes that $\sqrt{N}\hat{G}_{Bk}$ is close in probability to a linear combination of U-statistics. From a standard projection argument, it is then possible to characterize this gradient component in the same form as the A-component.[23] Namely in the Appendix we define a vector ε_{Bi} which is i.i.d. with expectation zero and finite variance components. Then, with $\bar{\varepsilon}_B$ as the corresponding sample mean, we show that:

$$\sqrt{N}\hat{G}_B = \sqrt{N}\bar{\varepsilon}_B + o_p(1).$$

With first-stage estimation uncertainty having a similar i.i.d. characterization, asymptotic normality follows.

5 Simulation Evidence

To analyze the finite sample performance of the estimator we examine the following setting. We simulated the following model where the same exogenous variables appear in the conditional means and the conditional variances of both endogenous variables. The two indices underlying the heteroscedasticity are also highly correlated. Moreover, we use the same functional form for the heteroscedasticity in each equation. The model has the form:

$$\begin{aligned}
Y_{1i} &= 1 + x_{1i} + x_{2i} + Y_{2i} + u_i \\
Y_{2i} &= 1 + x_{1i} + x_{2i} + v_i \\
u_i &= 1 + \exp(.2 * x_{1i} + .6 * x_{2i}) * u_i^* \\
v_i &= 1 + \exp(.6 * x_{1i} + .2 * x_{2i}) * v_i^* \\
u^* &= .33 * v_i^* + N(0,1) \text{ and } v_i^* \sim N(0,1).
\end{aligned}$$

We generate x_{1i} and x_{2i} as standard normal random variables and then transform x_{2i} into a chi-squared variable with 1 degree of freedom. We then esti-

[23] See Serfling (1980) and Powell, Stock, and Stoker (1989).

mated the model by OLS and the control function procedure developed here, which we denote CF in the tables. The simulation results for $n = 1000$ and 100 replications are reported in Table 1.

An examination of Table 1 reveals a number of interesting features of the simulations. First, consider the OLS estimates noting that the entries in the Table represent the mean value from the 100 replications with the standard deviation of the replications reported in parentheses under the estimate. The OLS estimates for the main equation's parameters in this specification are severely biased with respect to their true values of 1 indicating that there is a large degree of endogeneity in this model. The estimates for each of the $x's$ are approximately .86, indicating a bias of around 14 percent, while the bias on the coefficient for the endogenous regressor is approximately 14 percent.

Column 2 shows results for the control function procedure. First consider the estimates of the parameters for the conditional mean of the main equation as these are our major focus. The average values of the coefficients for the $x's$ and Y_2 are all close to 1 indicating that the inclusion of the control function is accounting for the endogeneity bias. Moreover, while there is more variability in the estimates, in comparison to the OLS estimates, the estimates, as indicated by their standard deviations, are generally quite precise. Second, consider the auxiliary parameter estimates which are obtained in the estimation process. The parameter ρ corresponds to the coefficient on the control function. The true value is .3134 and thus the average estimate of .298 is reasonable. The parameter δ is the parameter in the index generating the heteroscedasticity in the secondary equation. The average estimate is .377 which is reasonably close to the true value of .33, noting that it has a relatively large standard deviation. The parameter b is the coefficient in the index generating heteroscedasticity in the main equation. The average point estimate of .245 is reasonable relative to the true value of .33, but again there is a very large standard deviation associated with this estimate. Recall that this estimate is obtained simultaneously with the slope coefficients and its

imprecision reflects that it is difficult to estimate this parameter accurately while simply minimizing the squared residuals for this model.

If conditional variance parameters in the primary equation are of direct interest, then as described below, it is possible to exploit other sources of information to increase their precision. One approach which we employed was to employ the residual from the primary equation using the final estimates. Using this squared residual as the dependent variable, we obtained the SLS estimator as was done for δ. The average estimate for b from this approach is reported as b_{sls} in Table 1. We see that there is a notable improvement with an average estimate of .358 and a large increase in the precision of the estimate. This would suggest that this additional step produces worthwhile gains.

Though not reported here, we also considered several variants on the estimator presented. Here, we have explored the case in which the conditional variance of the errors is characterized by a single index. However, suppose that the entire distribution of each error is characterized by a single index (as is the case in the simulations). Under this more restrictive index assumption, it may be possible to develop a modified version of the estimator presented here with better finite sample performance (especially for index parameters).[24] We have also examined several "GLS" variants of the CF method presented here. While these resulted in a noticeable improvement in the estimates, we judged the improvement not sufficient to warrant any further (albeit minor) lengthening of the Appendix.[25]

[24] When the entire distribution of the errors depends on a single index, any function of the squared residual will satisfy a single index assumption. Accordingly, in an SLS regression, there will be many ways of estimating the index parameters. For the multiplicative heteroscedasticity employed in the monte-carlo, the (trimmed) log transform of the squared residual would appear to a natural transformation to employ.

[25] We examined estimators based on consistent residuals from the primary equation to estimate the conditional variance parameter in a manner similar to that employed for the secondary equation. Not surprisingly, as the information on the conditional variance is largely contained in the residuals, the resulting estimator for the index parameter had a much smaller variance that obtained from the control method reported here. Indeed,

Table 1: Simulation Results

	OLS	**CF**
constant	.858	1.003
	(.122)	(.201)
x_1	.858	1.003
	(.120)	(.210)
x_2	.866	1.011
	(.121)	(.203)
Y_2	1.137	.993
	(.108)	(.119)
ρ		.298
		(.110)
b		.245
		(.435)
b_{sls}		.358
		(.193)
δ		.377
		(.245)

it would be interesting to combine this moment information with that in the first-order conditions to the minimization problem defined here. We have not explored this possibility in the present paper.

6 Conclusion

We have examined a triangular simulatenous model where there are no available exclusion restrictions to employ as instruments but where we allow for generalized forms of heteroscedasticity with the model's errors depending on the explanatory variables. We have shown that the model is identified and have formulated a method for estimating it. We have also established that the estimator is consistent and asymptotically distributed as normal. In a monte-carlo study the estimator for the parameters of interest in the primary equation performed quite well in finite samples. As indicated previously, there is scope for further improvements in the estimator for the index parameters. Such improvements would come from fully exploring index structure in the monte-carlo (see footnote 24) or from making use of all available moment information (see footnote 25).

We have focused on the linear structure in part because it is most often used in practice. More importantly, in the absence of other information, it is this structure for which identification fails without exclusion restrictions. Nevertheless, it would seem relatively straight-forward to extend the model to allow nonlinear functions of the exogenous variables to enter both primary and secondary equations. With a control modified to reflect conditional mean rather than linear dependence of u on v, it may also be possible to allow for nonlinearities in the endogenous variables.

7 Appendix

7.1 Intermediate Lemmas

The Appendix is organized into a section on intermediate lemmas and a main section providing consistency and asymptotic normality of the proposed estimator for both primary and secondary equations. We begin with convergence rates for the components of the expectations estimator. In the primary equation, recall that $u(\theta_o) \equiv Y_{1i} - Z_i\theta_o$ and that π_o denotes the true regression coefficients for the secondary equation. Let

$$r_{2i} \equiv v_i(\pi_o)^2 \equiv {v_i}^2 \text{ or } u_i(\theta)^2,$$

where the error in the primary equation, $u_i(\theta)$, is defined for an arbitrary value of the parameter vector.

To provide convergence results for estimated conditional expectations of the above squared errors and for estimated index densities, for $m = 1, 2$ let $W_m \equiv X_1 + X_2\gamma_m$, $\gamma \equiv (\gamma_1, \gamma_2) \equiv (\theta, \gamma)$, and $W \equiv (W_1, W_2)$. With $w_m \equiv x_1 + x_2\gamma_m$ as a conditioning value for W_m and with $\lambda \equiv (\theta, \gamma)$, define $w(\lambda)$ as the conditioning vector. Then, define:

$$a(w(\lambda); \lambda, s) \equiv g(w) E(r_2^s | W = w), \ s = 0, 1,$$

where $g(w)$ is the density for W and $\gamma \equiv (\gamma_1, \gamma_2)$. For $s = 0, 1$ and $m = 1, .., M \equiv \dim(W)$, write the estimator for a as:

$$\hat{a}(w(\lambda); \lambda, s) \equiv \sum_{j=1}^{N} r_{2i}^s \Pi_m \left[\frac{1}{hN} K_M \left[(w_m - W_{jm}) / h \right] \right].$$

For $M = 1$, K_1 is the twicing kernel in (D1) while for $M = 2$, K_2 is the

(double) twicing kernel in (D2). Note that

$$\hat{a}\left(w\left(\lambda\right);\lambda,1\right)/\hat{a}\left(w\left(\lambda\right);\lambda,0\right)\equiv\hat{f}/\hat{g}$$

estimates a conditional expectation of the form shown in (D1-2). Define the derivative operator:

$$\nabla_{\lambda}^{d}\left(\hat{a}\right)\equiv\frac{\partial}{\partial\lambda}\hat{a},\text{ with }\nabla_{\lambda}^{0}\left(\hat{a}\right)\equiv\hat{a}.$$

Employing the above notation, the following lemma provides convergence results useful for analyzing the gradient.

Lemma 1 Assume that $a(w)$ has bounded derivatives to order 4 for $M \equiv dim(W) = 1$ and to order 6 for $M = 2$. Then, for $d = 1, 2$, $s = 0, 1$ and for X in a compact set:

$$a) \quad : \quad \sup_{w}\left(\ E\left[\hat{a}\left(w;\lambda,s\right)\right]-a\left(w;\lambda,m\right)\ \right)^{2}=\begin{cases} O\left(h^{8}\right): & M=1 \\ O\left(h^{12}\right): & M=2 \end{cases}$$

$$b) \quad : \quad \sup_{x}\left|E\left[\nabla_{\gamma}^{d}\hat{a}\left(w;\lambda,s\right)\right]-\nabla_{\gamma}^{d}a\left(w;\lambda,s\right)\right|^{2}=O\left(h^{2}\right),\ d=1,2$$

$$c) \quad : \quad \sup_{x}E\left(\left[\ \nabla_{\gamma}^{d}\hat{a}\left(w;\lambda,s\right)-E\left(\nabla_{\gamma}^{d}\hat{a}\left(w;\lambda,s\right)\right)\ \right]^{2}\right)=O\left(\frac{1}{Nh^{2d-1+M}}\right).$$

Proof of Lemma 1 . Noting that the kernels in (D1-2) are higher order kernels, the proof for the squared bias readily follows from standard Taylor series arguments; the variance result is standard.

To establish consistency and to analyze the Hessian matrix, we require convergence uniform in the parameters. The following lemma provides these results.

Lemma 2 (Uniform Convergence). Under the assumptions and definitions in section 4, for $s = 0, 1$ and $d = 0, 1, 2$:

$$\Delta \equiv \sup_{z,\gamma} \left| \nabla_\gamma^d \hat{a}\left(w; \gamma, s\right) - E\nabla_\gamma^d \hat{a}\left(w; \gamma, s\right) \right| = o_p\left(1\right).$$

Proof of Lemma 2. Here, we provide the result for $d = 0$, $s = 1$, $M \equiv dim(W) = 1$, and $r_{2i} = u_i^2\left(\theta\right)$.The proofs for other cases are similar and somewhat simpler when $r_{2j} = v_j^2\left(\pi_o\right)$. Write

$$u_j^2\left(\theta\right) = u_j^2\left(\theta_o\right) - 2u_j\left(\theta_o\right) Z_j\left(\theta - \theta_o\right) + \left(\theta - \theta_o\right)' Z_j' Z_j\left(\theta - \theta_o\right).$$

With θ in a compact set, the argument for all three terms is quite similar. Below, we provide the argument for the first term. Following Ichimura (1993), define:

$$t_j \equiv \begin{array}{ll} 1: & \left|u_i^2\left(\theta_o\right)\right| < N^\alpha \\ 0: & Otherwise \end{array}$$

and write:

$$\begin{aligned} \hat{a}_1 &\equiv \sum_{j=1}^{N} t_j u_j^2\left(\theta_o\right) \left[\frac{1}{hN} K\left[\left(w - W_j\right)/h\right]\right] \\ a_{1n} &\equiv E\left[t_j\ u_j^2\left(\theta_o\right) \mid W_j = w\right] g\left(w\right). \end{aligned}$$

Similarly, define $\hat{a}_0$ and a_{0n} by replacing t_j with $\left(1 - t_j\right)$. Then, with Δ defined as above, $\Delta \leq \Delta_1 + \Delta_2 + \Delta_3$, where:

$$\Delta_1 \equiv \sup_{z,\gamma} \left|\hat{a}_1 - E\left(\hat{a}_1\right)\right|; \quad \Delta_2 \equiv \sup_{z,\gamma} \left|\hat{a}_0\right|; \quad \Delta_3 \equiv \sup_{z,\gamma} \left|E\left(\hat{a}_0\right)\right|.$$

The proof then follows by showing $\Delta_k = o_p\left(1\right)$, $k = 1, 2, 3$. Since $N^{-\alpha} h \Delta_1$ is bounded, from Hoeffding's inequality (see Lemma 1 of Klein and Spady

(1993)):

$$\Delta_1 = O_p\left(hN^{-[1/2-\alpha]}\right) = o_p\left(1\right)$$

for the selected window and α sufficiently small. For the second term:

$$\Delta_2 \leq \left(c\sum_{j=1}^{N}\left(1-t_j\right)u_j^2\left(\theta_o\right)/\left(hN\right)\right),$$

where c is a positive and finite constant. The above expression converges to its expectation. From the tail-assumption on r_{2i}, this expectation tends to zero as:

$$\frac{1}{h}\int_{N^\alpha}^{\infty}\left(\frac{w}{\left[1+w^2\right]^{(s+1)/2}}\right)dw < \frac{1}{h}\int_{N^\alpha}^{\infty} w^{-s}dw = o\left(1\right).$$

A similar argument shows that $\Delta_3 = o_p\left(1\right)$.

Lemma 3. Assume:

$$S_a \equiv \sum \hat{a}_i^2/N = O_p\left(N^{-s}\right),\ S_b \equiv \sum \hat{b}_i^2/N = O_p\left(N^{-t}\right),$$

where $s+t>1$. Then,

$$\sqrt{N}\sum \hat{a}_i\hat{b}_i/N = o_p(1).$$

Proof of Lemma 3. The result follows from Cauchy's inequality:

$$\left|\sqrt{N}\sum \hat{a}_i\hat{b}_i/N\right| \leq \sqrt{N}S_a^{1/2}S_b^{1/2}.$$

In the remainder of this intermediate section, we use the above lemmas to characterize the gradient to the objective function for the primary equation. Asymptotic normality for the estimator of the primary equation will then follow from this characterization. To preview the argument, we

require the following notation. Denote $\hat{\eta}$ as the vector of estimated parameters from the Y_2-equation (including estimated parameters of the conditional error variance). Write $\alpha_o \equiv (\theta_o', \rho_o, b_o')'$ for the vector of true parameter values from the Y_1-model (including correlation and conditional variance parameter values). Refer to the averaged objective function for the primary equation (Y_1-model) shown in (D7). Taken with respect to α let $\hat{G}(\alpha_o; \hat{\eta})$ and $\hat{H}(\alpha_o; \hat{\eta})$ be the corresponding gradient and Hessian when estimated variance functions are positive.

With estimated conditional variance functions converging uniformly to positive functions, the following Taylor series expansion is valid on a set with probability approaching one:

$$\sqrt{N}[\hat{\alpha} - \alpha_o] = -\hat{H}\left(\alpha^+; \hat{\eta}\right)^{-1} \sqrt{N}\hat{G}(\alpha_o; \hat{\eta}), \tag{1}$$

$\alpha^+ \epsilon [\alpha_o . \hat{\alpha}]$. Obtain $H(\alpha; \eta)$ from $\hat{H}$ by replacing all estimated nonparametric expectations by their probability limits in $\hat{H}$. Uniformly in the parameters, it can be shown that $\left|\hat{H}(\alpha; \eta) - H(\alpha; \eta)\right|$ and $|H(\alpha; \eta) - EH(\alpha; \eta)|$ each converge in probability to zero. Therefore, once consistency is established, $\hat{H}(\alpha^+; \hat{\eta})$ will converge in probability to $H_o \equiv EH(\alpha_o; \eta_o)$. Asymptotic normality will then follow if the gradient component is asymptotically normal.

To analyze the gradient, recall from (D7) that:

$$\begin{aligned} \hat{M}_{ik}(\alpha_o, \eta_o) &\equiv [X_i \; Y_{2i}]\theta_o + \rho_o \frac{\hat{S}_{ui}(\alpha_o; k)}{\hat{S}_{vi}(\eta_o)} \hat{v}_i(\eta_o) \\ M_i(\alpha_o, \eta_o) &\equiv [X_i \; Y_{2i}]\theta_o + \frac{S_{ui}(\alpha_o)}{S_{vi}(\eta_o)} v_i \equiv M_{oi}. \end{aligned} \tag{2}$$

From (D7) , recall that $\hat{\tau}_{si}$ is a smooth trimming function for observations where $S_{vi} < 0$. Denote $\hat{w}_{ik} \equiv \hat{\tau}_{si}\left[\nabla_\alpha \hat{M}_{ik}(\alpha_o, \eta_o)\right]$, $\hat{w}_i \equiv [\hat{w}_{i1} + \hat{w}_{i2}]$, and $\hat{G}_C \equiv \left[\nabla \hat{G}_\eta(\theta_o, b_o; \eta^+)\right][\hat{\eta} - \eta_o]$. Then, the gradient with respect to α at

α_o is given as:

$$\begin{aligned}
\hat{G}\left(\alpha_o;\hat{\eta}\right) &= \hat{G}\left(\alpha_o;\eta_o\right)+\hat{G}_C=\hat{G}_A+\hat{G}_B+\hat{G}_C, \qquad (3)\\
\hat{G}_A &\equiv -\sum\hat{\tau}_i\left[Y_{1i}-M_{oi}\right]\hat{w}_i/N,\ \hat{w}_i\equiv\left[\hat{w}_{i1}+\hat{w}_{i2}\right]\\
\hat{G}_B &\equiv \hat{G}_{B1}+\hat{G}_{B2},\ \hat{G}_{Bk}\equiv\sum_{i=1}^{N}\hat{\tau}_i\left[\hat{M}_{ik}\left(\alpha_o,\eta_o\right)-M_{oi}\right]\hat{w}_{ik}/N.
\end{aligned}$$

Lemmas GA and GB below provide appropriate characterizations for $\sqrt{N}\hat{G}_A$ and $\sqrt{N}\hat{G}_B$. The characterization of the remaining component will immediately follow from the characterization of the first-stage estimator, $\hat{\eta}$, in Theorem 1 of the next section. All asymptotic results hold if trimming is based on X throughout or on X and estimated indices.[26] Below we will show that known trimming functions may replace estimated trimming functions in a number of terms. In one critical term (see Lemma GA below), this result will follow from Pakes and Pollard (1989, Lemma 2.17, p. 1037). In other cases, we will employ results on convergence rates for indicators in Klein (1993), which are based on inequalities due to Jim Powell.

Lemma GA (First Gradient Component). With $\hat{M}_{ik}\left(\alpha,\eta_o\right)$ defined in (2) above:

$$\nabla_\alpha\hat{M}_{ik}\left(\alpha_o,\eta_o\right)=\begin{bmatrix} W_i+\rho_o\left(\nabla_\theta\hat{S}_{ui}\left(\alpha_o;k\right)\right)v_i/\hat{S}_{vi}\\ \left(\hat{S}_{ui}\left(\alpha_o;k\right)/\hat{S}_{vi}\right)v_i\\ \rho_o\left(\nabla_b\hat{S}_{ui}\left(\alpha_o;k\right)\right)v_i/\hat{S}_{vi}. \end{bmatrix}$$

Recalling that $\hat{w}_i\equiv\hat{\tau}_{si}\left[\hat{w}_{i1}+\hat{w}_{i2}\right]$,define w_i by replacing all estimated func-

[26] As stated in the assumptions section, we have found that the finite sample performance is improved when X-trimming is followed by trimming based on estimated indices (with minimal X-trimming maintained for technical reasons).

tions with their probability limits. With $\tau_{io} \equiv \tau_i(q_o)$, let:

$$\varepsilon_{1i} \equiv -\tau_{io}[Y_{1i} - M_{oi}]\, w_i;\ \bar{\varepsilon}_1 \equiv \sum_{i=1}^{N} \varepsilon_{1i}/N.$$

Then, for $\hat{G}_A$ in (3):

$$\sqrt{N}\hat{G}_A \equiv -N^{-1/2}\sum \tau_{io}[Y_{1i} - M_{oi}]\, w_i = \sqrt{N}\bar{\varepsilon}_A + o_p(1).$$

Proof of Lemma GA. With $\hat{\tau}_i \equiv \tau_i(\hat{q})$, $\sqrt{N}\hat{G}_A$ is the sum of the following three terms:

$$\begin{aligned}
\mathbf{A} &\equiv N^{-1/2}\sum [Y_{1i} - M_{oi}]\,[\hat{\tau}_i - \tau_{io}]\, w_i \\
\mathbf{B} &\equiv N^{-1/2}\sum [Y_{1i} - M_{oi}]\,[\hat{\tau}_i - \tau_{io}]\,[\hat{w}_i - w_i] \\
\mathbf{C} &\equiv N^{-1/2}\sum [Y_{1i} - M_{oi}]\,\tau_{io}\,[\hat{w}_i - w_i].
\end{aligned}$$

The proof will follow if each of these terms is $o_p(1)$. Employing a similar strategy to that in Klein (1993), denote q_o as a vector of population quantiles (see (D1), Section 4) and let $N_\varepsilon \equiv \langle q : |q - q_o| < \varepsilon \rangle$, $\varepsilon = o(1)$. Then, $\mathbf{A} = o_p(1)$ if

$$\mathbf{A}^* \equiv \sup_{N_\varepsilon} N^{-1/2}\sum [Y_{1i} - M_{oi}]\,[\tau_i(q) - \tau_i(q_o)]\, w_i = o_p(1)$$

for all $\varepsilon = o(1)$.[27] From Pakes and Pollard (1989, Lemma 2.17, p. 1037), $\mathbf{A}^* = o_p(1)$.

For the term **B**, note that $\hat{\tau}_i \equiv \tau_i(\hat{q})$ and $\tau_{io} \equiv \tau_i(q_o)$, where $|\hat{q} - q_o| \equiv o_p(N^{-s})$. Letting $N_\delta \equiv \langle q : |q - q_o| < \delta \rangle$, $\delta = o(N^{-s})$ it suffices to show

[27] If uniformity holds for $\alpha \in \mathcal{N}_\varepsilon$ for all $\varepsilon = o(1)$, then uniformity holds over $o_p(1)$ neighborhoods of q_o.

that for all $\delta = o(N^{-s})$:

$$\mathbf{B}^* \equiv \sup_{N_\delta} N^{-1/2} \left| \sum \left[Y_{1i} - M_{oi} \right] \left[\tau_i \left(q \right) - \tau_i \left(q_o \right) \right] \left[\hat{w}_i - w_i \right] \right| = o_p(1).$$

Let $\tau_i^* \left(q \right)$ be an indicator defined on the union of the sets on which the indicators $\tau_i \left(q \right)$ and $\tau_i \left(q_o \right)$ are defined. Then, it suffices to show that:

$$\mathbf{B}^* \equiv \sup_{N_\delta} N^{-1/2} \left| \sum \left[Y_{1i} - M_{oi} \right] \left[\tau_i \left(q \right) - \tau_i \left(q_o \right) \right] \tau_i^* \left(q \right) \left[\hat{w}_i - w_i \right] \right| = o_p(1).$$

From Cauchy's inequality (see Lemma 3):

$$\begin{aligned} \mathbf{B}^* &\leq N^{1/2} \mathbf{B}_1^* \mathbf{B}_2^*, \\ \mathbf{B}_1^* &= \sup_{N_\delta} \left[\sum \left[Y_{1i} - M_{oi} \right]^2 \left[\tau_i \left(q \right) - \tau_i \left(q_o \right) \right]^2 / N \right]^{1/2} \\ \mathbf{B}_2^* &= \sup_{N_\delta} \left[\sum \tau_i^* \left(q \right) \left[\hat{w}_i - w_i \right]^2 / N \right]^{1/2}. \end{aligned}$$

From Klein (1993), with indicators approximated by smooth functions, it can be shown that for any fixed ε arbitrarily small : $\mathbf{B}_1^* = o_p \left(N^{-s+\varepsilon} \right)$. It also can be shown that $\mathbf{B}_2^* = o_p \left(N^{-1/2+s-\varepsilon} \right)$, which completes the argument for $\mathbf{B}$.[28]

Turning to $\mathbf{C}$, the analysis is similar to that in Klein and Spady (1993), with the result following from a mean-square convergence argument. To illustrate the argument, with $r_i \equiv \left[Y_{1i} - M_{oi} \right] v_i$, the second component of

[28] When $\hat{q}$ is a X sample quantile, $s = 1/2 - \varepsilon$.

the weight vector generates the following component of $\mathbf{C}$:

$$\begin{aligned}
\mathbf{C}_2 &\equiv N^{-1/2}\sum \tau_{io} r_i \left[\hat{\tau}_{is}\frac{\hat{S}_{ui}}{\hat{S}_{vi}} - \tau_{is}\frac{S_{ui}}{S_{vi}}\right] \\
&= N^{-1/2}\sum \tau_{io} r_i \left[\frac{\left(\hat{\tau}_{is}\hat{S}_{ui} - \tau_{is}S_{ui}\right)}{\hat{S}_{vi}} - \tau_{is}\frac{S_{ui}}{S_{vi}}\frac{\left(\hat{\tau}_{is}\hat{S}_{vi} - \tau_{is}S_{vi}\right)}{\hat{S}_{vi}}\right] \\
&\equiv D_1 + D_2.
\end{aligned}$$

With the analysis for both of these terms being similar, focus on D_1. From Lemmas 1 and 3:

$$\begin{aligned}
D_1 &= D_{11} - D_{12} + o_p(1), \\
D_{11} &\equiv N^{-1/2}\sum \tau_{io} r_i \left[\frac{\hat{\tau}_{is}\left(\hat{S}_{ui} - S_{ui}\right)}{\hat{S}_{vi}}\right]\frac{\hat{S}_{vi}}{S_{vi}}, \\
D_{12} &\equiv N^{-1/2}\sum \tau_{io} r_i \left[\frac{S_{ui}\left(\hat{\tau}_{is} - \tau_{is}\right)}{\hat{S}_{vi}}\right]\frac{\hat{S}_{vi}}{S_{vi}}.
\end{aligned}$$

For D_{11}, from a Taylor series on $\hat{\tau}_{is}$ and Lemmas 1 and 3:

$$D_{11} = N^{-1/2}\sum \tau_{io} r_i \left[\frac{\tau_{is}\left(\hat{S}_{ui} - S_{ui}\right)}{S_{vi}}\right] + o_p(1)$$

On a set with probability approaching 1, from a Taylor series expansion of $\left(\hat{S}_{ui}^2\right)^{1/2}$ about S_{ui}^2 and Lemmas 1 and 3:

$$D_{11} = N^{-1/2}\sum \tau_{io} r_i \tau_{is}\left(\hat{S}_{ui}^2 - S_{ui}^2\right)/\left(2S_{ui}S_{vi}\right) + o_p(1)$$

Employing the ratio form of $\hat{S}_{ui}^2$ and Lemmas 1 and 3:

$$\begin{aligned} D_{11} &= N^{-1/2} \sum \tau_{io} r_i \left[\hat{f}_i / \hat{g}_i - S_{ui}^2 \right] \frac{\hat{g}_i}{g_i} \frac{\tau_{is}}{2 S_{ui} S_{vi}} + o_p(1) \\ &= D_{11}^* + o_p(1), \\ D_{11}^* &\equiv N^{-1/2} \sum \tau_{io} r_i \left[\hat{f}_{ii} - \hat{g}_i S_{ui}^2 \right] \left[\tau_{is} / \left(2 g_i S_{ui} S_{vi} \right) \right]. \end{aligned}$$

With the above term being linear in estimated functions and with r_i having expectation conditioned on X of 0, it can be shown that

$$E\left[(D_{11}^*)^2 \right] \to 0.$$

With $\hat{\tau}_{is}$ being a smoothed indicator with derivative controlled by a_N, the analysis for D_{12} is similar.

To simplify $\hat{G}_B \equiv G_{B1} + G_{B2}$, the second gradient component in (3), Lemma 4 below shows that the estimated trimming and weight functions may be taken as known.

Lemma 4. With w_{ik} and $\hat{w}_{ik}$ defined as in Lemma GA and with all terms evaluated at true parameter values:

$$N^{1/2} \hat{G}_{Bk} \equiv \sqrt{N} \sum \tau_i \left[\hat{M}_{ik} - M_{oi} \right] w_{ik} / N + o_p(1),$$

where $\hat{G}_{Bk}$ is a gradient given in (3).

Proof of Lemma 4. Referring to $\hat{\tau}_i \hat{w}_i$ as an estimated weight, the difference in terms with estimated and true weights is given as:

$$\Delta_k = N^{1/2} \sum \left[\hat{M}_{ik} - M_{oi} \right] \left[\hat{\tau}_i \hat{w}_{ik} - \tau_{io} w_{ik} \right] / N, k = 1, 2.$$

The result follows from repeated application of Lemmas 1 and 3, a Taylor series arugment for the smooth trimming component of $\hat{w}_{ik}$, and a convergence

rate for indicators in Klein (1993).

To further simplify $\hat{G}_B$, Lemma 5 below shows that the components of $\hat{G}_B$ can be written as a linear combination of estimated functions. As this form will be a U-statistic, standard projection arguments will complete the characterization of this gradient term.

Lemma 5. Referring to Lemma 4 and the definition of $\hat{M}_{ik}$ in (2):

$$\begin{aligned} N^{1/2}\hat{G}_{Bk} &= N^{1/2}\sum \tau_i v_i \left[\frac{\left(\hat{S}_{ui}\left(\alpha_o;k\right)-S_{ui}\right)}{\hat{S}_{vi}} - \frac{S_{ui}}{S_{vi}}\frac{\left(\hat{S}_{vi}-S_{vi}\right)}{\hat{S}_{vi}}\right] w_{ik}/N \\ &\equiv N^{1/2}T_{1k} - N^{1/2}T_{2k}. \end{aligned}$$

For the single index case, (D1) provides the ratio form for $\hat{S}^2_{ui}\left(\alpha_o;1\right)$ and $\hat{S}^2_{vi}$. In the double index case, (D2) provides the ratio form for $\hat{S}^2_{ui}\left(\alpha_o;2\right)$. Accordingly, write:

$$\begin{aligned} \hat{S}^2_{ui}\left(\alpha_o;k\right) &\equiv \hat{f}_{1i}\left(\alpha_o;k\right)/\hat{g}_{1i}\left(\left(\alpha_o;k\right)\right) \\ \hat{S}^2_{vi} &\equiv \hat{f}_{2i}/\hat{g}_{2i}. \end{aligned}$$

Define:

$$a_{1i} \equiv \left[\frac{1}{2g_{1i}\left(\alpha_o;k\right)S_{vi}S_{ui}}\right];\ a_{2i} \equiv \left[\frac{S_{ui}}{2g_{2i}S_{vi}S_{vi}}\right].$$

Then, on a set with probability tending to one:

$$\begin{aligned} N^{1/2}\hat{G}_{Bk} &= N^{1/2}T^*_{1k} - N^{1/2}T^*_{2k} + o_p\left(1\right), \\ T^*_{1k} &\equiv \sum \tau_i \left[\hat{f}_{1i}\left(\alpha_o;k\right) - \hat{g}_{1i}\left(\alpha_o;k\right)S^2_{ui}\right] a_{1i}w_{ik}/N \\ T^*_{2k} &\equiv \sum \tau_i \left[\hat{f}_{2i} - \hat{g}_{2i}S^2_{vi}\right] a_{2i}w_{ik}/N. \end{aligned}$$

Proof of Lemma 5: For the term T_{1k}, from Lemmas 1 and 3:

$$N^{1/2}T_{1k} = N^{1/2}\sum \tau_i\tau_{is}\frac{\left(\hat{S}_{ui}\left(\alpha_o;k\right)-S_{ui}\right)}{\hat{S}_{vi}}\frac{\hat{S}_{vi}}{S_{vi}}+o_p\left(1\right).$$

On a set with probability approaching 1, Taylor expand $\left[\hat{S}_{ui}^2\left(\alpha_o;k\right)\right]^{1/2}$ about S_{ui}^2 and employ Lemmas 1,3 to obtain:

$$\begin{aligned}
N^{1/2}T_{1k} &= N^{1/2}\sum \tau_i\left(\hat{S}_{ui}^2\left(\alpha_o;k\right)-S_{ui}^2\right)\frac{\tau_{is}}{2S_{ui}S_{vi}}w_{ik}/N+o_p\left(1\right)\\
&= N^{1/2}\sum \tau_i\left(\frac{\hat{f}_{1i}\left(\alpha_o;k\right)}{\hat{g}_{1i}\left(\alpha_o;k\right)}-S_{ui}^2\right)\frac{\tau_{is}}{2S_{ui}S_{vi}}w_{ik}/N+o_p\left(1\right)\\
&= N^{1/2}\sum \tau_i\left(\frac{\hat{f}_{1i}\left(\alpha_o;k\right)}{\hat{g}_{1i}\left(\alpha_o;k\right)}-S_{ui}^2\right)\left[\frac{\hat{g}_{1i}\left(\left(\alpha_o;k\right)\right)}{g_{1i}\left(\left(\alpha_o;k\right)\right)}\right]\frac{\tau_{is}}{2S_{ui}S_{vi}}w_{ik}/N+o_p\left(1\right),
\end{aligned}$$

which completes the argument. The proof for T_{2k} is identical.

From Lemma 5, $\hat{G}_{Bk}$ is a U-statistic to which standard projection arguments apply to complete the required characterization for this term. Lemma GB below provides this result.

Lemma GB (U-Statistic Projection): Referring to Lemma 5, write the trimming indicator as $\tau_i \equiv \tau_{Ii}\tau_{xi}$, the product of index and X-trimming indicators. Then, with indices I_{ui} and I_{vi} evaluated at the true parameter values:[29]

$$\begin{aligned}
N^{1/2}\hat{G}_{Bk} &= N^{1/2}G_{Bk}+o_p\left(1\right), \qquad \text{(a)}\\
G_{Bk} &\equiv \sum\left[u_i^2-S_{ui}^2\right]E\left[\tau_i a_{1i}w_{ik}|I_i\left(k\right)\right]/N-\\
&\quad \sum\left[v_i^2-S_{vi}^2\right]E\left[\tau_i a_{2i}w_{ik}|I_{vi}\right]/N,
\end{aligned}$$

[29] Within the expectations, the indices are naturally evaluated at true parameter values. Though not theoretically necessary, for reasons argued earlier, we have adopted a strategy of re-estimating the model and trimming on the basis of estimated indices.

where $I_i(k) = I_{ui}$ for k = 1 and (I_{ui}, I_{vi}) for k = 2. With ε_{Bi} as the i^{th} term of $G_{B1} + G_{B2}$ it then follows that:

$$N^{1/2}\hat{G}_B = N^{1/2}\left[G_{B1} + G_{B2}\right] \equiv N^{1/2}\sum \varepsilon_{Bi}/N \equiv N^{1/2}\bar{\varepsilon}_B. \qquad \text{(b)}$$

Proof of Lemma GB. For $k = 1$ (the argument for $k = 2$ is identical), refer to Lemma 5 and write:

$$\begin{aligned} T_{11}^* &\equiv \sum \tau_i \left[\hat{f}_{1i}(\alpha_o; 1) - \hat{g}_{1i}(\alpha_o; 1) S_{ui}^2\right] \tau_i a_{1i} w_{i1}/N \\ &= \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \rho_{ij}^*, \ \rho_{ij}^* \equiv \left[u_j^2 k_1[i,j] - k_1[i,j] S_{ui}^2\right] \tau_i a_{1i} w_{i1} \\ &= \begin{pmatrix} N \\ 2 \end{pmatrix}^{-1} \sum_i \sum_{j>i} \left[\rho_{ij}^* + \rho_{ji}^*\right]/2 \equiv U_N. \end{aligned}$$

As the above expression is a U-statistic with expectation 0, from standard projection arguments:

$$\sqrt{N}\left[U_N = \hat{U}_N\right] = o_p(1),$$

$$\begin{aligned} \hat{U}_N &= \frac{2}{N} \sum_i E\left(\left[\rho_{ij}^* + \rho_{ji}^*\right]/2 \mid Y_{1i}, Z_i\right) \\ &= \sum_i \left[u_i^2 - S_{ui}^2\right] E\left[\tau_i a_{1i} w_{ik} | I_{ui}\right]/N, \end{aligned}$$

which follows because ρ_{ij}^* has conditional expectation of $\mathrm{o}\left(N^{-1/2}\right)$ from the higher order kernel and ρ_{ji}^* has the conditional expectation shown above. Employing the same argument, T_{21}^* has a similar form. The characterization for $\hat{G}_{B1}$ follows. The analysis of $\hat{G}_{B2}$ is similar to that for $\hat{G}_{B1}$, which completes the argument for (a). The required form in (b) now directly follows from (a).

7.2 Main Results

Recall that the third gradient component for the second stage estimator depends on $\hat{\eta}$, the estimator for the nuisance parameter vector from the Y_1-model. To analyze such first-stage estimation uncertainty, Theorem 1 below characterizes the components of $\hat{\eta}$.

Theorem 1: First Stage Consistency and Characterization. Define:

$$v_i^2(\pi) \equiv (Y_{2i} - X_i\pi)^2,$$

where

$$E\left[v_i^2(\pi_o) \mid I_{vi}(\delta_o)\right] = E\left[Y_i(\pi) \mid X_i\right].$$

Define

$$\begin{aligned}
\hat{R}(\delta;\pi) &\equiv \frac{1}{N}\sum_{i=1}^{N} \hat{\tau}_i \hat{r}_i^2(\delta;\pi),\ \hat{r}_i(\delta;\pi) \equiv v_i^2(\pi) - \hat{E}\left[v_i^2(\pi) \mid I_{vi}(\delta)\right] \\
\hat{\delta}(\pi) &= \arg\min_{\delta} \hat{R}(\delta;\pi), \\
\hat{S}_{v_i}(\pi) &\equiv \left[\left|\hat{E}\left((Y_{2i} - X_i\pi)^2 | I_{vi}\left(\hat{\delta}(\pi)\right)\right)\right|\right]^{1/2}.
\end{aligned}$$

With $\hat{\pi}_{ols}$ as the OLS estimator for π_o, let:

$$\hat{X}_i^* \equiv X_i/\hat{S}_{v_i}(\hat{\pi}_{ols});\ X_i^* \equiv X_i/S_{v_i}(\pi_o).$$

Then, with $\Omega \equiv p\lim\left(X^{*'}X^*/N\right)$ and with $\varepsilon_{\pi i} \equiv X_i^{*\prime} v_i^*$, the GLS estimator of π, $\hat{\pi}$, satisfies:

$$\sqrt{N}\left[\hat{\pi} - \pi_o\right] = \Omega^{-1}\sqrt{N}\sum \varepsilon_{\pi i}/N + o_p(1). \qquad \text{(a)}$$

Define

$$\begin{aligned} R(\delta;\hat{\pi}) &\equiv \frac{1}{2N}\sum_{i=1}^{N}\tau_{io}r_i^2(\delta;\hat{\pi}),\ r_i(\delta;\pi)\equiv v_i^2(\hat{\pi})-E\left[v_i^2(\hat{\pi})\mid I_{vi}(\delta)\right] \\ w_i &\equiv \tau_{io}\frac{\partial}{\partial\delta}r_i(\delta_o;\pi_o),\ w_i^*\equiv w_i-E\left[w_i|I_{vi}(\delta_o)\right] \\ R_{11} &\equiv p\lim\left[\frac{\partial^2}{\partial\delta\partial\delta'}R(\delta_o;\pi_o)\right];\ R_{21}\equiv p\lim\left[\frac{\partial^2}{\partial\pi\partial\delta'}R(\delta_o;\pi_o)\right]. \end{aligned}$$

The estimator for the index parameters, $\hat{\delta}(\hat{\pi})$, satisfies:[30]

$$\sqrt{N}\left[\hat{\delta}-\delta_o\right]=-R_{11}^{-1}\sqrt{N}\left[\sum r_i(\delta_o;\pi_o)\,w_i^*/N+R_{21}\left[\hat{\pi}-\pi_o\right]\right]+o_p(1). \quad \text{(b)}$$

Proof of Theorem 1. The proof for (a) is immediate. For (b), accounting for estimation uncertainty in $\hat{\pi}$, the proof follows from Ichimura (1993) or from the intermediate lemmas above.[31]

Recall from (1-3) at the beginning of the Appendix that the second stage estimator has three gradient components, with the first two being characterized in Lemmas GA and GB above. Lemma GC characterizes the third gradient component.

Lemma GC. Referring to (1-3) and employing the notation in Theorem 1, define:

$$\varepsilon_{\mathrm{Ci}}\equiv p\lim\left(\nabla G_\eta(\theta_o,b_o;\eta_o)\right)\begin{bmatrix}\Omega^{-1}\varepsilon_{\pi i}\\ \left(-R_{11}^{-1}\left[r_i(\delta_o;\pi_o)\,w_i^*+R_{21}\Omega^{-1}\varepsilon_{\pi i}\right]\ \right)\end{bmatrix}.$$

[30] This characterization holds under a more general semiparametric formulation of the Y_2-model. Here, to emphasize identification issues, we have focused on the case where the Y_2 model is linear with an unknown conditional variance function.

[31] With the weight redefined for the second-stage estimator, first and second-stage gradients have a similar structure. Consequently, the intermediate lemmas used to prove Theorem 3 could also be employed to prove Theorem 1.

Then:

$$\sqrt{N}\hat{G}_C = \sqrt{N}\bar{\varepsilon}_C + o_p(1),\ \bar{\varepsilon}_C \equiv \sum_{i=1}^{N} \varepsilon_{\mathrm{Ci}}/N.$$

Proof of Lemma GC. The proof follows from (1-2), Theorem 1, and a standard uniform convergence result.

Theorem 2: Second Stage Consistency and Identification. With $Z_i \equiv [X_i\ Y_{2i}]$, let R be the matrix with i^{th} row:

$$[Z_i\ \ (\ S_{ui}(\theta_o, b_o)/S_{vi})\, v_i].$$

Then, the model is identified under the constant correlation assumption if R has full column rank and the correlation parameter satisfies:

$$0 < |\rho_o| < 1.$$

The proof for the above theorem is given separately for the case where $S_{ui}(\theta_o, b_o)$ is obtained nonparametrically (case A) and the case in which a single index structure is imposed (case B).

Proof of Theorem 2A (Nonparametric Case). Let:

$$\begin{aligned}
u_i(\theta) &\equiv Y_{1i} - Z_i\theta,\ Z \equiv [X\ \ Y_2];\ \hat{S}_{ui}(\theta)^2 \equiv \hat{E}\left[u_i(\theta)^2 | X_i\right] \\
\hat{M}_i &\equiv Z_i\theta + \rho\left[\frac{\hat{S}_{ui}(\theta)}{\hat{S}_{vi}}\right]\hat{v}_i \\
\hat{Q}(\alpha) &\equiv \frac{1}{2}\sum \hat{\tau}_i\hat{\tau}_{is}\left[Y_{1i} - \hat{M}_i\right]^2/N;\ \hat{Q}^*(\alpha) \equiv \left[\hat{Q}(\alpha) - \hat{Q}(\alpha_o)\right] \\
\hat{\alpha} &\equiv \arg\min \hat{Q}(\alpha) = \arg\min \hat{Q}^*(\alpha).
\end{aligned}$$

Replace estimated functions in $\hat{Q}^*(\alpha)$ with their uniform probability lim-

its to obtain $Q^*(\alpha) \equiv [Q(\alpha) - Q(\alpha_o)]$.[32] It can be shown that $\sup\left|\hat{Q}^*(\alpha) - Q^*(\alpha)\right|$ is, $o_p(1)$, uniformly in α. Further, the function $Q^*(\alpha)$ converges uniformly in the parameters to its expectation:

$$Q^*(\alpha) \xrightarrow{p} E[Q^*(\alpha)] \equiv E[Q(\alpha) - Q(\alpha_o)], \text{ uniformly in } \alpha.$$

With

$$M_i(\alpha) \equiv Z_i\theta + \rho\left[\frac{S_{ui}(\theta, b)}{S_{vi}}\right] v_i \quad \& \quad M_{oi} \equiv M_i(\alpha_o),$$

write $Y_{1i} - M_i = [Y_{1i} - M_{oi}] - [M_i - M_{oi}]$. It can be shown that

$$EQ^*(\alpha) = E\sum_{i=1}^{N} \tau_i [M_i - M_{oi}]^2 / N.$$

With $M_i - M_{oi} = 0$ at the true parameter values, consistency follows if this minimum is unique. If the minimum is not unique, it must be the case that $M_i - M_{oi} = 0$ at all potential minimizing parameter values. Then, for any minimizer, (θ^*, ρ^*)

$$Z_i(\theta^* - \theta_o) + [\rho^* S_{ui}(\theta^*) - \rho_o S_{ui}(\theta_o)] v_i / S_{vi} = 0, \tag{2A.1}$$

from which it follows that:

$$\begin{aligned} \rho^{*2} S_{ui}^2(\theta^*)\left(v_i^2/S_{vi}^2\right) &= \rho_o^2 S_{ui}^2(\theta_o)\left(v_i^2/S_{vi}^2\right) \\ &\quad -2\rho_o S_{ui}(\theta_o)(v_i/S_{vi}) Z_i(\theta^* - \theta_o) \\ &\quad + (\theta^* - \theta_o)' Z_i' Z_i (\theta^* - \theta_o). \end{aligned}$$

[32] The function $\hat{Q}^*$ is introduced to avoid convergence arguments for:

$$\frac{1}{N}\sum [Y_i - M_{oi}]^2.$$

Taking an expectation conditioned on X_i :

$$\begin{aligned} \rho^{*2} S_{ui}^2(\theta^*) &= \rho_o^2 S_{ui}^2(\theta_o) - 2\rho_o S_{ui}(\theta_o) S_{vi}(\theta_2 - \theta_{2o}) \\ &\quad + (\theta^* - \theta_o)' E[\, Z_i' Z_i \mid X_i] (\theta^* - \theta_o). \end{aligned} \tag{2A.2}$$

From the definition of $S_{ui}^2(\theta^*)$:

$$\begin{aligned} S_{ui}^2(\theta^*) &= E\left[(Y_i - Z_i\theta^*)^2 \mid X_i\right] \\ &= E\left[(\, u_i(\theta_o) - Z_i(\theta^* - \theta_o))^2 \mid X_i\right] \\ &= S_{ui}^2(\theta_o) - 2E(u_i v_i | X_i)(\theta_2^* - \theta_{2o}) \\ &\quad + (\theta^* - \theta_o)' E[\, Z_i' Z_i \mid X_i] (\theta^* - \theta_o) \\ &= S_{ui}^2(\theta_o) - 2\rho_o S_{ui}(\theta_o) S_{vi}(\theta_2^* - \theta_{2o}) \\ &\quad + (\theta^* - \theta_o)' E[\, Z_i' Z_i \mid X_i] (\theta^* - \theta_o). \end{aligned} \tag{2A.3}$$

Differencing the expressions in (2A.2) and (2A.3):

$$\rho^{*2} S_{ui}^2(\theta^*) - S_{ui}^2(\theta^*) = \rho_o^2 S_{ui}^2(\theta_o) - S_{ui}^2(\theta_o). \tag{2A.4}$$

Note that $\rho^{*2} < 1$, because $\rho^{*2} = 1$ implies $\rho_o^2 = 1$, which violates an identifying assumption. Let $r \equiv \left[(1-\rho_o^2)/(1-\rho^{*2})\right]^{1/2}$ and substitute (2A.4) into (2A.1) to obtain:

$$[Z_i \ , \ (\, S_{ui}(\theta_o)/S_{vi})\, v_i] \begin{bmatrix} \theta - \theta_o \\ \rho^* r - \rho_o \end{bmatrix} = 0.$$

Under a full rank assumption, $\theta = \theta_o$ and $\rho^* r = \rho_o$. Since $\theta = \theta_o$, from (2A.3), $S_{ui}(\theta_o) = S_{ui}(\theta^*)$. Consequently, from (2A.4), $r = 1$. With $\rho^* r = \rho_o$, it follows that $\rho^* = \rho_o$.

Proof of Theorem 2B (The Index Case). In addition to the notation introduced above, recall that estimated conditional variance functions are

given as:

$$\begin{aligned}\hat{S}_{ui}^{*}(\theta,b)^{2} &\equiv \hat{E}\left[u_i(\theta)^{2}\left|I_{ii}(b), I_{vi}\right.\right] \\ \hat{S}_{ui}(\theta,b)^{2} &\equiv \hat{E}\left[u_i(\theta)^{2}\left|I_{ii}(b)\right.\right].\end{aligned}$$

Write estimated response functions as:

$$\begin{aligned}\hat{M}_{1i} &\equiv Z_i\theta + \rho\hat{S}_{ui}(\theta,b)\,\hat{v}_i/\hat{S}_{vi} \\ \hat{M}_{2i} &\equiv Z_i\theta + \rho\hat{S}_{ui}^{*}(\theta,b)\,\hat{v}_i/\hat{S}_{vi}.\end{aligned}$$

Let $\alpha \equiv (\theta, b, \rho)$ and

$$\hat{Q}(\alpha) \equiv \sum_k \hat{Q}_k, \quad \hat{Q}_k \equiv \frac{1}{2N}\sum_k \hat{\tau}_i\hat{\tau}_s\left[Y_{1i} - \hat{M}_{ki}\right]^2.$$

Then, with $\hat{Q}^{*}(\alpha) \equiv \hat{Q}(\alpha) - \hat{Q}(\alpha_o)$, the estimator is given as:

$$\hat{\alpha} \equiv \arg\min_{\alpha}\hat{Q}(\alpha) = \arg\min_{\alpha}\hat{Q}^{*}(\alpha).$$

Similar to the argument above, with $Q^{*}(\alpha) \equiv E\left[Q(\alpha) - Q(\alpha_o)\right]$

$$\sup_{\alpha}\left|\hat{Q}^{*}(\alpha) - Q^{*}(\alpha)\right| = o_p(1).$$

It can be shown that for $k = 1, 2$:

$$\begin{aligned}EQ^{*}(\alpha) &= EQ(\alpha) - EQ(\alpha_o) = E\Delta_1 + E\Delta_2, \\ E\Delta_k &= E\sum_{i=1}^{N}\tau_i\left[M_{ki} - M_{oi}\right]^2/N,\end{aligned}$$

where

$$\begin{aligned} M_{1i}(\alpha) &\equiv Z_i\theta + \rho S_{ui}(\alpha)\, v_i/S_{vi} \\ M_{2i}(\alpha) &\equiv Z_i\theta + \rho S^*_{ui}(\alpha)\, v_i/S_{vi}. \end{aligned}$$

At the true parameter values, $M_{ki}(\alpha_o) - M_{oi} = 0$, $k = 1, 2$. Therefore, both Q_1 and Q_2 are separately minimized at the true parameter values. With α^* as a candidate for a minimum, Q_1 and Q_2 must also be separately be minimized at α^*. It then follows that:

$$M_{ki}(\alpha^*) - M_{oi} = 0,\ k = 1, 2.$$

For $k = 2$, from the above restriction:

$$Z_i(\theta^* - \theta_o) + \left[\rho^* S^*_{ui}(\alpha^*) - \rho_o S_{ui}(\alpha_o)\right] v_i/S_{vi} = 0. \tag{2B.1}$$

Multiply (2B.1) by v_i, take an expectation conditioned on X_i, divide by $S_{vi} \neq 0$, and solve for $\rho_o S_{ui}(\alpha_o)\, v_i/S_{vi}$ to obtain:

$$\rho_o S_{ui}(\alpha_o) = S_{vi}(\theta^*_2 - \theta_{2o}) + \rho^* S^*_{ui}(\alpha^*).$$

Noting that the r.h.s. only depends on X through $[I_v\ I_i(b^*)]$, for $\rho_o \neq 0$, it follows that:

$$S^m_{ui}(\alpha_o) = E\left[S^m_{ui}(\alpha_o)\,|I_v\ I_i(b^*)\right],\ m = 1, 2.$$

Returning to (2B.1), solve for $\rho^* S^*_{ui}(\alpha^*)\left(v_i^2/S_{vi}^2\right)$ to obtain:

$$\begin{aligned} \rho^{*2} S^*_{ui}(\alpha^*)^2\left(v_i^2/S_{vi}^2\right) &= \rho_o^2 S^2_{ui}(\alpha_o)\left(v_i^2/S_{vi}^2\right) - \\ &\quad 2\rho_o S_{ui}(\alpha_o)\left(v_i/S_{vi}\right) Z_i(\theta - \theta_o) + \\ &\quad (\theta^* - \theta_o)' Z_i' Z_i(\theta^* - \theta_o). \end{aligned} \tag{2B.3}$$

Letting:

$$\begin{aligned} B\left(X_i\right) &\equiv 2\rho_o S_{ui}\left(\alpha_o\right) S_{vi}\left(\theta_2-\theta_{2o}\right) \\ C\left(X_i\right) &\equiv \left(\theta^*-\theta_o\right)' E\left[Z_i'Z_i|I_u\left(b^*\right), I_v\right]\left(\theta^*-\theta_o\right), \end{aligned}$$

employ (2B.2) and take an expectation in (2B.3) conditioned on $I_u\left(b^*\right)$ and I_v to obtain:

$$\rho^{*2} S_{ui}^*\left(\theta^*, b^*\right)^2 = \rho_o^2 S_{ui}^2\left(\theta_o\right) - B\left(X_i\right) + C\left(X_i\right). \tag{2B.4}$$

Proceeding with a strategy similar to the nonparametric case above, from the definition of $S_{ui}^2\left(\alpha^*\right)$:

$$\begin{aligned} S_{ui}^*\left(\alpha^*\right)^2 &\equiv E\left[\left(Y_i - Z_i\theta\right)^2 \mid I_u(b^*), I_v\right] \\ &= E\left[\left(\, u_i\left(\theta_o\right) - Z_i\left(\theta-\theta_o\right)\right)^2 \mid I_u(b^*), I_{vi}\right] \\ &= E\left[S_{ui}^2\left(\alpha_o\right)|I_u(b^*), I_{vi}\right] - 2E(u_i v_i \mid I_u(b^*), I_v)\left(\theta_2^* - \theta_{2o}\right) \\ &\quad + C\left(X_i\right). \end{aligned} \tag{2B.5}$$

From the constant correlation assumption and (2B.2):

$$\begin{aligned} E(u_i v_i \mid X_i) &= \rho_o S_{ui}\left(\theta_o\right) S_{vi} \Rightarrow \\ E(u_i\left(\theta_o\right) v_i \mid I_u(b^*), I_v) &= \rho_o E\left[S_{ui}\left(\theta_o\right)|I_u(b^*), I_v\right] S_{vi} \\ &= \rho_o S_{ui}\left(\theta_o\right)\ S_{vi}. \end{aligned} \tag{2B.6}$$

Substituting (2B.2) and (2B.6) into (2B.5):

$$S_{ui}^*\left(\theta^*, b^*\right)^2 = S_{ui}^2\left(\theta_o\right) - B\left(X_i\right) + C\left(X_i\right). \tag{2B.7}$$

Differencing (2B.4) and (2B.7) :

$$\rho^{*2} S_{ui}^*\left(\theta^*, b^*\right)^2 - S_{ui}^*\left(\theta^*, b^*\right)^2 = \rho_o^2 S_{ui}^2\left(\theta_o\right) - S_{ui}^2\left(\theta_o\right). \tag{2B.8}$$

Note that $\rho^{*2} = 1 \Rightarrow \rho_o^2 = 1$, which contradicts an identification assumption. With $\rho^{*2} \neq 1$, let $r \equiv \left[\left(1-\rho_o^2\right)/\left(1-\rho^{*2}\right)\right]^{1/2}$ and from (2B.4) write:

$$S_{ui}^{*}\left(\theta^{*}, b^{*}\right) = rS\left(\theta_o\right). \tag{2B.9}$$

Employing an argument identical to that in the nonparametric case, it now follows from (2B1), a full rank condition, and the above results that $\theta^* = \theta_o$ and $\rho^* = \rho_o$.

Since $\theta^* = \theta_o$, $\rho^* = \rho_o$, and $M_{1i} = M_{2i}$

$$\begin{aligned} S_{ui}^{2*}\left(\theta_o, b^{*}\right) &= E\left[u_i\left(\theta_o\right)^2 \mid I_u\left(b^{*}\right), Iv\right] = E\left[u_i\left(\theta_o\right)^2 \mid I_u\left(b^{*}\right)\right] \\ &= S_{ui}^{2}\left(\theta_o\right) = E\left[u_i\left(\theta_o\right)^2 \mid I_u\left(b_o\right)\right]. \end{aligned}$$

It can now follows that $b^* = b_o$ (Ichimura 1993).

Theorem 3 : Asymptotic Normality of the Second Stage Estimator. Employing notation in Lemmas GA-C, let:

$$\varepsilon_i \equiv \varepsilon_{Ai} + \varepsilon_{Bi} + \varepsilon_{\mathrm{Ci}}.$$

From (1-3) of the previous section and with $H_o \equiv E\left[H\left(\alpha_o; \eta_o\right)\right]$:

$$\sqrt{N}\left[\hat{\alpha} - \alpha_o\right] \xrightarrow{d} Z, \; Z \tilde{} \; N\left(0, H_o^{-1} E\left(\varepsilon_i \varepsilon_i'\right) H_o^{-1}\right).$$

Proof of Theorem 3. With $\alpha^{+} \in (\hat{\alpha}, \alpha_o)$, in a set with probability tending to 1, from (1-3):

$$\sqrt{N}\left[\hat{\alpha} - \alpha_o\right] = -\left[\hat{H}\left(\alpha^{+}; \hat{\eta}\right)\right]^{-1}\left[\sqrt{N}\left(\hat{G}_1 + \hat{G}_2 + \hat{G}_3\right)\right],$$

For the Hessian term, from standard uniform convergence arguments: $\hat{H} \xrightarrow{p}$

H_o. For the gradient, from Lemmas GA-C:

$$\sqrt{N}\left(\hat{G}_A + \hat{G}_B + \hat{G}_C\right) = \sqrt{N}\bar{\varepsilon},\ \bar{\varepsilon} = \sum_{i=1}^{N} \varepsilon_i/N.$$

The result now follows.

References

[1] Bollerslev, T. (1990): " Modelling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized GARCH Approach," *Review of Economics and Statistics, 72,* 498-505.

[2] Dagenais, M., and D.Dagenais (1997): "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics, 76 (1-2),* 193-222.

[3] Ichimura, H, (1993):"Semiparametric least squares (SLS) and weighted SLS estimation of single index models" *Journal of Econometrics*, 58, 71-120.

[4] Klein, R. (1993): "Specification Tests for Binary Choice Models Based on Index Quantiles, *Journal of Econometrics,* 59, 343-375.

[5] Klein, R. and F.Vella (2004): "A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity," unpublished manuscript.

[6] Lewbel, A. (1997): "Constructing Instruments for Regressions with Measurement Error when No Additional Data are Available, With an Application to Patents and R&D," *Econometrica*, 65, 1201-1213.

[7] Lewbel, A. (2004): "Identification of Heteroskedastic Endogenous Models or Mismeasured Regressor Models," unpublished manuscript.

[8] Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, v. 4, Chapter 36, Amsterdam, North Holland.

[9] Newey, W., F. Hsieh, and J. Robins (2004): "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947-962.

[10] Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1058.

[11] Powell, J.L.,J.H. Stock, and T.M. Stoker (1989):"Semiparametric Estimation of Weighted Average Derivatives," *Econometrica*, 57, 1403-1430.

[12] Rigobon, R, (1999): "Identification through heteroscedasticity," *Review of Economics and Statistics,* 85, 777-792.

[13] Rummery, S., F.Vella and M.Verbeek (1999):"Estimating the Returns to Education for Australian Youth via Rank-Order Instrumental Variables," *Labour Economics,* 6, 491-507.

[14] Serfling, R.S. (1980) : *Approximation Theorems of Mathematical Statistics.* New York; Wiley.

[15] Sentana, E. and G.Fiorentini (2001), "Identification, Estimation and Testing of Conditional Heteroskedastic Factor Models,"*Journal of Econometrics,* 102, 143-164.

[16] Silverman, P. (1986): *Density Estimation.* New York; Chapman and Hall.

[17] Staiger, R. and J.Stock (1999): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 68, 1055-1096.

[18] Vella, F. and M.Verbeek (1997): "Rank Order as an Instrumental Variable" unpublished manuscript