Estimating a class of triangular simultaneous equations models without exclusion restrictions

Roger Klein a,*, Francis Vella b

a Rutgers University, United States
b Georgetown University, United States

A R T I C L E   I N F O

Article history:
Received 5 May 2009
Received in revised form 5 May 2009
Accepted 6 May 2009
Available online 4 August 2009

Keywords:
Triangular semiparametric models
Heteroscedasticity
Identification

A B S T R A C T

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.

1. Introduction

While there is general agreement that instrumental variables (IV) estimation is appropriate for a large class of models with endogeneity, there is frequently disagreement about the exclusion restrictions imposed in specific empirical applications. This difficulty frequently leads to choices which are sometimes seen to invalidate the final estimates while the use of uninformative instruments is responsible for the weak instruments literature (see, for example, Staiger and Stock, 1997).

It is well known that IV for a linear model with a single endogenous regressor is equivalent to an OLS regression that includes this variable’s reduced-form residual as an additional regressor to control for endogeneity. 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, and since the control is a linear combination of the endogenous regressor and exogenous variables, identification requires an exclusion restriction. However, when the error distribution depends on the exogenous variables, it is possible to develop a control whose impact is not constant. Here we consider the case of heteroscedastic errors and develop a “feasible” control whose impact is not constant. More explicitly, we show that even in the absence of exclusion restrictions which can be used to generate appropriate instruments it is possible to produce a control function style estimator for a model with endogenous regressors which identifies the parameters of interest. We show that this is true even when we treat the processes generating the heteroscedasticity in a general manner. In addition to providing the estimator, we describe a procedure for implementing it. We also derive the asymptotic properties of our proposed procedure.

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:

\[ Y_{1i} = X_1\theta o + Y_2\theta o + u_i \equiv Z\theta o + u_i \]

\[ Y_{2i} = X_1\pi o + u_i \]

* We would like to acknowledge Research Council support from Rutgers. We are grateful to participants at numerous seminars over the past six years for various comments which have resulted in improvements to the paper. We would also like to thank Ethel Fonseca and Chan Shen 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. Finally, we would like to thank the referees and the editor for suggestions and comments. Any remaining errors are the sole responsibility of the authors.

E-mail address: klein@econ.rutgers.edu (R. Klein).

1 This control function approach is equivalent to two-stage-least-squares.
where \( Y_{i1} \) and \( Y_{i2} \) are continuous endogenous variables; \( X_i \) is a vector of variables that are mean-independent of the error components \( u_i \) and \( v_i \).

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 \( \theta_0 \). A key feature of the model is that it allows the same \( X \)'s in both equations without imposing any restrictions on the parameter values.

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

\[
a_0 = \arg \min_a E \left[ u - a u \right]^2 = cov(u, v) / Var(v).
\]

By construction, \( \epsilon \equiv u - a_0 v \) is uncorrelated with \( v \) and therefore uncorrelated with \( Z \), which provides the basis for the controlled regression:

\[
Y_{i1} = Z \theta_0 + a_0 v_i + \epsilon_i.
\]

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 does depend on \( X \), obtain the (linear) conditional relation between the errors by the following conditional population regression:

\[
A_o (X_i) \equiv \arg \min_A E \left[ u_i - A v_i \right] = cov (u_i, v_i) / Var(v_i | X_i).
\]

In this case, \( \epsilon_i \equiv u_i - A_0 (X_i) v_i \) is uncorrelated with \( v_i \) conditioned on \( X_i \), which provides the basis for the controlled regression:

\[
Y_{i1} = Z \theta_0 + A_0 (X_i) v_i + \epsilon_i.
\]

We refer to the dependence of \( A_0 (X_i) \) on \( X_i \) as the variable impact property (VIP). With \( R_i \) as the matrix with ith row: \( R_i \equiv [Z_i \ \ A_o (X_i) v_i] \), then \( R_i \) will have full column rank if VIP holds. Identification will then follow if it is possible to estimate \( A_o (X_i) \).

Below we characterize a class of error structures for which VIP holds and for which it is possible to consistently estimate \( A_o (X_i) \). To do so, denote conditional variances as: \( S_o^n \equiv Var (u_i | X_i) \) and \( S^2_{v_i} \equiv Var (v_i | X_i) \) and write the error structure as:

\[
\begin{align*}
& u_i \equiv S_o^n u_i^*; \quad v_i \equiv S^2_{v_i} v_i^*; \quad E \left( u_i^* | X_i \right) = E \left( v_i^* | X_i \right) = 0 \quad (3) \\
& \rho_o \equiv E \left( u_i^* v_i^* | X_i \right) = E \left( u_i^* v_i^* \right).
\end{align*}
\]

With the correlation \( \rho_o \) constant and the ratio \((S_o^n/S^2_{v_i})\) depending on \( X_i \), VIP holds and the control is given as:

\[
A_o (X_i) v_i = \rho_o \left[ S_o^n / S^2_{v_i} \right] v_i.
\]

When \( A_o (X_i) \) has the above form, we will show that it can be consistently estimated.

To illustrate several error structures in which the control has the above form, define variances and scaled errors as above. As a first case let the unscaled errors be related according to the additive linear structure:

\[
\begin{align*}
& u^* = \rho_o v^* + \epsilon^*; \quad E \left( u^* | X_i \right) = E \left( v^* | X_i \right) = 0; \quad cov \left( \epsilon^*, \epsilon^* | X_i \right) = 0.
\end{align*}
\]

As a second case, with \( \epsilon_1 \) and \( \epsilon_2 \) being mean-zero error components that are independent of \( X \), consider the multiplicative error structure:

\[
\begin{align*}
& u = \alpha_1 (X) \omega^* \epsilon_1; \quad v = \alpha_2 (X) \omega^* \epsilon_2,
\end{align*}
\]

where \( \epsilon_1 \) and \( \epsilon_2 \) are independent of the common error component, \( \omega^* \). The conditional second moment for the common error component, \( \omega^* \), may or may not depend on \( X \). In this case, with \( \rho_o \) as the correlation between \( \epsilon_1 \) and \( \epsilon_2 \), \( A_o (X) \) has the required form.

Other papers exploit second moment information as a source of identification. 4 Vella and Verbeek (unpublished) 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. The underlying identifying source in the Vella and Verbeek approach is a covariance restriction. 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 Vella and Verbeek (unpublished) and Rigobon (2003), Lewbel (unpublished) 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:

\[
E \left( X_i u_i v_i \right) = E \left( X_i E(u_i v_i | X_i) \right) = 0; \quad E \left( X_i v_i^2 \right) \neq 0. \quad (4)
\]

The model considered here differs in several respects from those above. First, for the model we consider, the conditional covariance \( E(u_i v_i | X_i) \) depends on \( X_i \). Consequently, while the first restriction in (4) may hold in special cases, it will generally not hold for the model considered here. Second, with the conditional covariance and variance functions depending on \( X_i \), we model the conditional variance of each error as an unknown function of an index. This flexibility is important because economic theory seldom provides a detailed parametric specification for the error structure.

3. The estimator

3.1. Index structures and semiparametric least-squares

Since we impose index restrictions in estimating the model described below, we begin by briefly discussing these restrictions and the estimation method that makes use of them. Beginning with an illustrative probit model, assume that the value for the choice is given by \( Y = X \beta_0 \), and term this value as an index. In this case, with \( Y \) as the binary choice variable:

\[
Pr(Y = 1 | X) = \Phi \left( X \beta_0 \right),
\]

where \( \Phi \) is a standard normal distribution function. In a semiparametric formulation, we do not make a normality assumption. In this case:

\[
Pr(Y = 1 | X) = E(Y | X) = E(Y | X \beta_0) \equiv G(X \beta_0).
\]

---

2 At this stage we make no assumptions regarding the properties of the components of \( X \) as they differ depending on the treatment of the heteroscedasticity. However, below we explicitly state what properties \( X \) must possess in the models we consider. See (A4) for the restrictions on the \( K \geq 3 \) components of \( X \).

3 We would like to thank Whitney Newey for this interpretation of the control.

4 This paper examines a model with a continuous outcome and a continuous endogenous treatment. Klein and Vella (2009) consider a model where the endogenous treatment is binary and identification of the treatment effect is through the presence of heteroscedasticity. However while that paper focuses on issues related to those discussed here, the respective identification approaches and the estimation procedures are fundamentally different.
Unlike the parametric model, the function $G$ is unknown. However, as in the parametric case, the conditional choice probability, or equivalently the conditional expectation of the dependent variable, depends on a single index, $X\beta_0$. This index need not be linear, but we do require that it is a parametric function of explanatory variables and parameter values. Without such an index assumption, we would need to estimate $E(Y|X)$ nonparametrically, which is known to be problematic when the dimension of $X$ is not small. Under a single index assumption, the underlying dimension of the conditioning variables becomes one, and we can nonparametrically and reliably estimate $E(Y|X\beta)$ for any specified value of the parameter vector.

In the linear index case, identification requires that the index contains a continuous variable, $X_1$, with non-zero and finite coefficient $\beta_{10}$. Write:

$$X\beta_0 = \beta_{10} [X_1 + X_2\theta_0] + c_0,$$

where $c_0$ is a finite constant, $\theta_0 \equiv \beta_{20}/\beta_{10}$, and $X_2$ contains a vector of random variables. In the linear index case, parameter values are at most identified up to location and scale as:

$$E(Y|X) = E(Y|X_1 + X_2\theta_0).$$

Therefore, in this case, we seek an estimator for $\theta_0$.

In general, with $V_1 = V_1(X; \beta_0)$ as a parametric function (not necessarily linear) of exogenous variable and parameter values, we term $V_1$ as a single index and say that a single index restriction holds if:

$$E(Y|X) = E(Y|V_1).$$

Generalizing to double indices, let $V_2$ be a second (scalar) index, assume that:

$$E(Y|X) = E(Y|V_1(\theta_{10}), V_2(\theta_{20})) \equiv E(Y|V(\theta_0)).$$

In this case, a double index restriction holds. Again, as compared with the nonparametric case, by reducing the dimension of the conditioning variables to two, it is possible to estimate expectations quite well at moderate sample sizes.

To estimate index models, we will employ Ichimura’s (1993) semiparametric least-squares (SLS), a method whose estimator minimizes a sum of squared residuals. Namely, let $\hat{E}_i(\theta) \equiv \hat{E}(Y_i|V_i(\theta))$ be a nonparametric estimate of the indicated conditional expectation and define a residual as $\hat{u}_i(\theta) \equiv Y_i - \hat{E}_i(\theta).$ Then the SLS estimate of $\theta_0$ is defined by minimizing a sum of squared residuals:

$$\hat{\theta} = \arg \min \sum \tau_i \hat{u}_i^2(\theta)/N,$$

where $\tau_i$ is an indicator (trimming) function that deletes problematic observations. Namely, it can be shown that an estimated expectation is a ratio of functions, with the estimated density for an index (or indices) in the denominator. For technical reasons, we trim out those observations from this objective function where the true density is assumed to be zero (e.g. in the tails).

3.2. The secondary equation

With $u_{1i}$ as the error term in the secondary equation, to estimate its conditional variance function, $S^2_{\tau_i}$, we impose a single index structure:

$$S^2_{\tau_i} = E \left[ u_{1i}^2 \mid X_i \right] = E \left[ u_{1i}^2 \mid I_{1i}(\delta_0) \right],$$

where $I_{1i}(\delta_0) \equiv X_{1i} + X_{2i}\delta_0$. Next, as detailed in (D3-4) below, estimate $\delta_0$ using semiparametric least-squares with the squared residual, $\hat{u}_{1i}^2$, as the dependent variable. The estimator for the conditional variance function is then given as:

$$\hat{S}^2_{\tau_i} = \hat{E} \left( \hat{u}_{1i}^2 \mid I_{1i}(\hat{\delta}) \right),$$

where $\hat{E}$ is a nonparametric estimator for the indicated conditional expectation. Employing the above initial estimator $\hat{S}_{\tau_i}$, we then obtain better estimates by repeating the above process in a GLS step.

3.3. The primary equation and identification

As consistent residuals are not available for the primary equation, due to the presence of an endogenous explanatory variable, the conditional variance function for this equation and the parameters of interest are estimated simultaneously. We distinguish between two cases according to whether or not the conditional variance function for the primary equation has an index or a nonparametric structure. For technical reasons discussed above, we focus on the index case as it can be expected to perform better in practice. For the nonparametric case, it can be shown that the control adjustment, $A(X)$, is determined up to a multiplicative constant by sample moments. Identification follows, with Theorem 2 providing the argument.

For the index case, we employ a different identification strategy under which both indices are identified without exclusion restrictions. To outline the argument, recall that the standard consistency–identification argument proceeds as follows. Let $\hat{S}(\alpha)$ be an objective function (e.g. SLS) whose minimum characterizes the estimator $\hat{\theta}$. Here, $\hat{S}(\alpha)$ is estimated in the sense that it depends on an estimated semiparametric expectation. Define $S(\alpha)$ by replacing all estimated semiparametric expectations in $\hat{S}(\alpha)$ with their probability limits. A standard argument then first shows that $\hat{S}(\alpha) - S(\alpha)$ is uniformly close to 0. Second, a relatively straightforward argument shows that $S(\alpha)$ is uniformly close to its expectation, $ES(\alpha)$. Identification requires that this expectation or population objective function is uniquely minimized at the true parameter values, $\alpha_0$. In what follows, we focus on describing a population objective function for which identification holds. The formal estimator will then be described in the next section as one that minimizes a sample counterpart.

With $Z_i \equiv [X_i, Y_{2i}]$, we begin by denoting the error, away from the truth, as:

$$u_{1i}(\theta) \equiv (Y_{1i} - Z_i \theta).$$

The index restriction for the primary equation is then given as:

$$E \left[ u_{1i}^2(\theta_o) \mid X_i = E \left[ u_{1i}^2 \mid I_{1i}(\theta_o) \right], \quad I_{1i}(\theta_o) \equiv X_{1i} + X_{2i}\theta_o. \right.$$  

Consider the population objective function:

$$ES(b) \equiv \frac{1}{N} E \sum \tau_i \left[ u_{1i}^2(\theta_o) - E \left( u_{1i}^2(\theta_o) \mid I_{1i}(b) \right) \right]^2,$$

which is essentially a regression of squared residuals on a flexibly modeled function of the index. With $b^*$ minimizing this objective

\footnote{As discussed below, identification also holds under a nonparametric formulation.}

\footnote{For the secondary equation, we impose an index structure to obtain better performance in finite samples. However, the identification argument does not depend on whether or not an index structure is imposed.}
function, as a necessary condition for a minimum, the following index restriction must be satisfied\footnote{7}: 

\[ E \left( \left( u_{i}^{T} (\theta_2) | u_{i2} (b^*) \right) E \left( u_{i}^{T} (\theta_2) | X \right) = E \left( u_{i}^{T} (\theta_2) | u_{i} (b_0) \right) \right). \]

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

With \( \theta_2 \) being unknown, identification becomes problematic as it is difficult to impose an index restriction away from the truth. To explain the problem define \( Z_i \) and \( u_i (\theta) \) as above and let:

\[ S_{wi} (\theta, b) = E \left( u_{i}^{T} (\theta_2) | u_{i} (b) \right). \]

With \( \alpha = (\theta, b, \rho) \), define a population objective function and its minimizer as:

\[ A_i (\alpha) = \rho [S_{wi} (\theta, b) / S_{vi}] ; \quad M_{wi} (\alpha) = Z_i \theta + A_i (\alpha) \hat{v}_i \]

\[ EQ_i (\alpha) = \frac{1}{N} E \sum_i \left( v_i Y_{ii} - M_{wi} (\alpha) \right)^2 \]

\[ = \frac{1}{N} E \sum_i \left( Y_{ii} - M_{wi} (\alpha) \right) - \left( M_{wi} (\alpha) - M_{wi} (\alpha) \right)^2 \]

\[ \alpha^* = \arg \min E \left[ Q_i (\alpha) \right]. \]

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

\[ M_{wi} (\alpha^*) - M_{wi} (\alpha) = 0 \Rightarrow Z_i (\alpha^* - \theta_2) + [\rho^* S_{wi} (\theta^*, b^*) - \rho S_{wi} (\theta_2, b_2)] v_i / S_{vi} = 0. \]

With additional information relating \( S_{wi} (\theta^*, b^*) \) to \( S_{wi} (\theta_2, b_2) \), the identification strategy would be greatly simplified. For example, consider the case where minimizing values satisfy an index restriction with: \( S_{wi} (\theta^*, b^*) = S_{wi} (\theta_2, b_2) \). Letting \( R \) be the matrix with \( i \)th row: \( R_{i} = [X_i Y_{i2i} \left( S_{wi} (\theta, b_2) / S_{vi} \right) v_i] \), from above:

\[ R \left[ \theta^* - \theta_2 = 0 \right]. \]

Identification would follow from a full (column) rank assumption on \( R \).

While it may not be possible to guarantee the strong index restriction \( S_{wi} (\theta^*, b^*) = S_{wi} (\theta_2, b_2) \) a priori 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. Recalling that \( \alpha = (\theta, b, \rho) \), let:

\[ S_{wi}^{2} (\theta, b) = E \left( u_{i}^{T} (\theta_2) | u_{i} (b) \right) / S_{vi} \]

\[ M_{bi} (\alpha) = Z_i \theta + \varphi \left( \left[ S_{wi}^{2} (\theta, b) / S_{vi} \right] v_i \right) \]

\[ Q_{2} (\alpha) = \| \hat{v}_i [Y_{i1} - M_{i2}] \|. \]

Then, with \( \alpha = (\theta, b, \rho) \) consider the “overall” population objective function:

\[ Q (\alpha) = Q_{1} (\alpha) + Q_{2} (\alpha) \]

Denote \( \alpha^* \) as a minimizer for \( Q (\alpha) \). In the Appendix it is shown that \( \alpha_\circ \), 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 \( \alpha^* \) must also minimize each of these component objective functions. As a result, \( \alpha^* \) must satisfy 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 \( \alpha_\circ \) is the unique minimizer.

To indicate the nature of these restrictions, in an argument similar to that above we show that \( \alpha^* \) must satisfy the restrictions:

\( M_{k} (\alpha^*) = M_{k} (\alpha_\circ) \). \( k = 1, 2 \). Therefore, with \( X^* \equiv [u_{i} (b^*)] \):

\( Ra : Z_i (\theta^* - \theta_2) + [\rho^* S_{wi} (\theta^*, b^*) - \rho S_{wi} (\theta_2, b_2)] v_i / S_{vi} = 0 \)

\( Rb : Z_i (\theta^* - \theta_2) + [\rho^* S_{wi} (\theta^*, b^*) - \rho S_{wi} (\theta_2, b_2)] v_i / S_{vi} = 0 \)

\( Rc : E \left( u_{i}^{2} (\theta^*) | u_{i} (b^*) \right) = E \left( u_{i}^{2} (\theta^*) | X^* \right) \)

where the index restriction in \( Rc \) follows by differenting the first two restrictions and employing the definitions of \( S_{wi} \) and \( S_{vi} \). Intuitively, \( Rc \) may be interpreted as providing an index restriction away from the truth. In Theorem 2, we show these restrictions in conjunction with a full rank condition are sufficient to provide identification. When \( X \) has full column rank and the ratio \( S_{wi} / S_{vi} \) depends on \( X \), then \( R \) will have full rank.

From the above discussion, we are motivated to select an estimator so as to minimize a sample counterpart to the population objective function \( Q \) above.\footnote{8} We define this function and the corresponding estimator in the next section.

Before proceeding to the details underlying the estimation procedure, we briefly reiterate our approach. Given the nature of the model under examination the appropriate control function to be included is \( A_i (X_i) v_i = \rho_o [S_{wi} / S_{vi}] v_i. \) Here, \( \rho_o \) is a parameter which is estimated along with the other slope parameters in the primary equation and thus the construction of the control requires the three remaining components. An estimate of \( v_i \) is obtained as the residual from the regression of \( Y_{i2} \) on \( X_i \). An estimate of its standard error, \( S_{vi} \), is obtained as the square root of the expected value from the semiparametric least-squares regression of the secondary equation’s residuals squared on \( X_i \). The component which is more difficult to estimate is \( S_{wi} \), as this corresponds to the standard error of the primary equation error. We obtain this component by iterating over the values for the primary equation slope parameters and the parameters which appear in the index underlying the primary equation’s heteroscedasticity. In each iteration the estimate of \( S_{wi} \) is based on a nonparametric estimate which is obtained in a similar manner to that for \( S_{vi} \). As defined earlier, iteration is based on the minimization of \( Q (\alpha) \), which represents the sum of the two least-squares criterion functions based on different conditioning sets. The choice of the conditioning sets ensures identification.

4. Assumptions, definitions, and results

In obtaining asymptotic results, we make the following assumptions:

A1. The vector \( (Y_{i1}, Y_{i2}, X_i, u_i, v_i) \) is i.i.d distributed over \( i \), with the variables \( X_i \) bounded and with data generating process in (1)–(3).\footnote{9} With \( S_{wi}^{2} \) as the conditional variance function for \( v_i, S_{wi}^{2} > 0 \).

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

A3. Let \( f \) be the density of either \( u^2 \) or \( v^2 \). Assume there exists \( c > 0 \) such that for \( t > c \), \( f \) satisfies the tail condition:

\[ f (t) \leq 1 / [1 + t^2]^{(d+1)/2}, \quad d \geq 5. \]
A4. Write $X_i \equiv [X_{i1}, X_{i2}, X_{i3}]$, where $X_{i1}$ and $X_{i2}$ are continuous variables that are not functionally related. For the case where conditional variances depend on linear indices, $I_{i1}$ and $I_{i2}$, assume that $I_{i1}$ depends on $X_{i1}$ and that $I_{i2}$ depends on $X_{i2}$. Write the normalized linear indices as:

\[ I_{i1}(b_a) = X_{i1} + [X_{i2}, X_{i3}] b_a; \quad I_{i2}(\delta_a) = X_{i2} + [X_{i3}] \delta_a. \]

A5. Referring to (A4), and letting $y$ be either $v^2$ or $u^2$, let the conditional density $g(x_1, x_2, x_3, y)$ be bounded and have bounded partials to order 6 in $x_1$ and $x_2$: \[ \left| \nabla^2 \nabla^2 g(x_1, x_2, x_3, y) \right| = O(1), \quad r + s \leq 6. \]

A6. Referring to (A4), write $X = [X, X_0]$, where $X$ and $X_0$ are vectors of continuous and discrete variables respectively. For $x_0$ a support point with positive probability, let the conditional density $g_{x_0} = g(x_0, X_0)$ be supported on $\mathcal{X}$, a product space that does not depend on $x_0$. Assume $g_{x_0} > 0$ on the interior of $\mathcal{X}$.

A7. Let $R$ be a matrix with $r$th row: $[X_i, Y_{2i}, (S_{ui}/S_{vi}) v_i]$ and assume that $R$ has full column rank.

Assumptions (A1–A2) define the data generating process and the parameter space. Assumption (A3) simplifies uniform convergence proofs for unbounded random variables. Assumption (A5) guarantees that the densities of the indices in (A4) are sufficiently smooth for bias control. Assumption (A6) identifies where index densities are zero for trimming purposes. Lastly, an identification condition is given in (A7).

To implement the estimation strategy, estimated expectations are defined in (D1). These are based on twinning kernels [see Newey et al., 2004] as we have found that they perform better than other higher order kernels. Other definitions provide various trimming strategies that we implement.

D1. Define an estimated conditional expectation:

\[ \hat{E}(W_i | I_i) = \sum_{j \neq i} W_j K_{ij} \sum_{j \neq i} K_{ji}. \]

To define the kernel, $K$, with $\psi(z)$ a standard normal density, let:

\[ K_1(w) = 2 \psi(w) - \int \psi(w-z) \psi(z) dz; \]

\[ K_2(w) = 2 K_1(w) - \int K_1(w-z) K_1(z) dz. \]

When $I$ is a single index with standard deviation $s_1$:

\[ K_{ij} \equiv K_1 \left[ (I_i - I_j) / h_1 \right]; \quad h_1 = s_1 N^{-\gamma}, \quad 1/8 < \gamma < 2/15. \]

When $I$ contains two indices with respective standard deviations $s_k$,

\[ K_{ij} \equiv K_2 \left[ (I_{i1} - I_{j1}) / h_1 \right] K_2 \left[ (I_{i2} - I_{j2}) / h_2 \right]. \]

\[ h_k = s_k N^{-\gamma}, \quad 1/2 < \gamma < 1/10. \]

D2. Indicator trimming. Let $q_{ak}$ and $q_{bk}$ be lower and upper population quantiles for the continuous variables $X_{ak}, k = 1, \ldots, K$. Let $q_a$ be a vector of these quantiles. With $x = 1 \times K$, define $P_a(q_a) = \left\{ x : q_a < x < q_{ak}, k = 1, \ldots, K \right\}$. With $X_i \equiv [X_{i1}, \ldots, X_{ik}]$, define the trimming indicator:

\[ \tau_{i \delta} = \tau_1(q_{0}) = \left[ X_i \in P_a(q_a) \right]. \]

With $q$ as a vector of sample quantiles: $\hat{\tau}_{i \delta} \equiv \tau_1(\hat{q})$.

D3. $Y_1$-Model. With the model given in (1)–(3), recall that the error component, $v_i$, has the conditional variance structure\footnote{The model can be generalized so that the conditional expectation of $Y_1$ is given by $G(X_{0})$ where the function $G$ is unknown.}:

\[ \hat{\delta} \equiv \arg \min_{\delta} \left( \hat{\sigma}_2^2 - E \left[ \hat{\sigma}_2^2 | I_{i1}(\delta) \right] \right)^2; \]

\[ \hat{\sigma}_2^2 = \hat{E} \left( \hat{\sigma}_2^2 | I_{i1}(\delta) \right). \]

Employing this conditional variance, the entire process above is repeated with residuals obtained using this estimated conditional variance in a GLS step.\footnote{Though asymptotically negligible, under higher order kernels it is possible that $\hat{\sigma}_2^2 < 0$. We smoothly trim out observations for which this problem occurs.}

D4. Smooth trimming. With $\hat{\sigma}_2^2 \equiv \hat{E} \left( \hat{\sigma}_2^2 | I_{i1}(\delta) \right)$, define:

\[ \hat{f}_{i \delta} = \tau \left( \hat{\sigma}_2^2 \right) = \left[ 1 + \exp \left( -\alpha \hat{\sigma}_2^2 \right) \right]^{-1}, \quad \alpha_n = \ln(N)^2. \]

The function above will tend to 0 as $\hat{\sigma}_2^2$ becomes negative and to 1 otherwise.\footnote{While very unlikely, negative estimates are possible under higher order kernels.} As this function approximates an indicator, its derivative must become high in a neighborhood of zero. To control for the magnitude of the derivative, $\alpha_n$ is selected above.

D5. $Y_1$-Model. Referring to (1)–(3), recall $Z_i \equiv [X_i, Y_{2i}]$ and $u_i(\theta_i) \equiv (Y_{1i} - Z_i \theta_i)$. The error has conditional variance structure:

\[ E \left[ u_i^2(\theta_i) | X_i \right] = E \left[ u_i^2(b_a) \right], \quad I_{i1}(b_a) = X_{i1} + X_{i2} b_a. \]

To define the estimator for this model, with $\theta = (\theta, b, \rho)$ and $k = 1, 2$ let:

\[ \hat{\delta}_{i1} \equiv \hat{E} \left( \hat{\sigma}_2^2 | I_{i1}(\delta) \right); \]

\[ \hat{\delta}_{i2} \equiv \hat{E} \left( \hat{\sigma}_2^2 | I_{i2}(\delta) \right); \]

\[ \hat{M}_{i2} = Z_i \theta + \hat{A}_{i2}(\alpha) \hat{v}_i; \]

\[ \hat{Q}_a = \frac{1}{N} \sum_i \hat{f}_{i \delta} \left[ Y_{1i} - \hat{M}_{i2}(\alpha) \right]^2. \]

Employing the above assumptions and definitions, the Appendix provides all proofs for asymptotic results. In the remainder of this section, we summarize the main results and provide a brief outline of the proof strategy. Beginning with the secondary equation (Y2-Model), Theorem 1 provides the large-sample results for the estimators of the nuisance parameters.

**Theorem 1** (The Y2-Model). From (1)–(3), (A1–A6), and (D1–D3), estimates of regression and index parameters satisfy the characterization:

\[
\sqrt{N}[\hat{\pi} - \pi_0] = \sqrt{N} \sum_{i=1}^{N} \varepsilon_{pi}/N + o_p(1) \tag{a}
\]

\[
\sqrt{N}[\hat{\delta} - \delta_0] = \sqrt{N} \sum_{i=1}^{N} \varepsilon_{si}/N + o_p(1), \tag{b}
\]

where $\varepsilon_{pi}$ and $\varepsilon_{si}$ 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 Y1-Model, Theorem 2 provides the consistency/identification result.
Theorem 2 (Consistency and Identification: The $Y_1$-Model). With
\[ \alpha_0 \equiv (\theta_{x_1}, \rho, b_0) \text{ and } \hat{\alpha} \equiv \left( \hat{\theta}, \hat{\rho}, \hat{b} \right), \text{ under (1)-(3), (A1-A7) and (D1-5):} \]
\[ \hat{\alpha} \xrightarrow{p} \alpha_0. \]

Section 3.2 provides the intuition underlying the consistency argument and identification. For both nonparametric and semi-parametric formulations, in the Appendix we establish identification when under (A7) the matrix $[X, Y_2, (S_{x_1}/S_0) v]$ has full column rank and a constant correlation assumption holds. This rank condition will hold if $X$ has full column rank and $S_{x_1}/S_0$ depends on $X$. Theorem 3, which is proved in the Appendix, provides the normality result.

Theorem 3 (Normality: The $Y_1$-Model). Employing notation in Lemmas GA–GC in the Appendix, let
\[ G \equiv G_1 + G_2 + G_3, \]
with $H_0 \equiv E[H(\alpha_0, \eta_0)]$ and $G_3$ as the $i$th row of $G$, (see Lemma G), under (A1-A7) and (D1-D4):
\[ \sqrt{N} \left[ \bar{\alpha} - \alpha_0 \right] \Rightarrow Z, \quad Z \sim N(0, H^{-1} \{G, G^t \} H_0^{-1}). \]

To outline the argument, which is formally given in the Appendix, note that under a standard Taylor series argument for the gradient to the objective function and a uniform convergence argument for the Hessian, normality will follow if the normalized gradient is asymptotically distributed as normal.\textsuperscript{13} For expositional purposes, neglect first-stage estimation, which matters, but poses no technical difficulties.\textsuperscript{14} With $\hat{\tilde{\alpha}}_k \equiv \tilde{\alpha}_k V_k \tilde{M}_k$ termed a weight function, define the gradient to objective function $k, k = 1, 2,$ as:
\[ \sqrt{N} \tilde{G}_k = -\sqrt{N} \sum \tilde{t}_i [Y_{t_i} - M_{1i}] \hat{w}_{ik} / N \]
\[ + \sqrt{N} \sum \tilde{t}_i [\tilde{M}_k - M_{1i}] \hat{w}_{ik} / N \]
\[ = -\sqrt{N} \tilde{G}_k(\alpha_0) + \sqrt{N} \tilde{G}_k(\hat{\alpha}_s). \]
The normalized gradient to the sum of the two objective functions is then given by $\sqrt{N} \left[ \tilde{G}_1 + \tilde{G}_2 \right]$. In what follows, we indicate the argument for showing that these components simplify to a form for which normality readily follows.

For the A-components, from results in Pakes and Pollard (1989) and mean-square convergence arguments, estimated trimming and weight functions may be taken as known:
\[ \sqrt{N} \tilde{G}_k(\alpha_0) = \sqrt{N} \sum \tilde{t}_i [Y_{t_i} - M_{1i}] w_{ik} + o_p(1). \]

For the B-components, employing Cauchy's inequality and rate of convergence results, it can be shown that trimming and weight functions may be taken as known. In the Appendix, we then show that the resulting gradient component is close to a U-statistic. Standard U-statistic results then enable us to show that this term has a form enabling a central limit theorem to apply.

5. Simulation evidence

To analyze the finite sample performance of the estimator we generate data from a model where the same exogenous variables appear in the conditional means and the conditional variance functions. The two indices underlying the heteroscedasticity are also highly correlated. With i.i.d. observations, the model has the form:
\[ Y_1 = 1 + x_{1i} + x_{2i} + X_{2i} + u_i, \quad x_{1i} \sim N(0, 1), \quad x_{2i} \sim \chi^2(1) \]
\[ Y_{2i} = 1 + x_{1i} + x_{2i} + v_i \]
\[ u_i = 1 + \exp(.2 \times x_{1i} + .6 \times x_{2i}) \times u^* \]
\[ v_i = 1 + \exp(.6 \times x_{1i} + .2 \times x_{2i}) \times v^* \]
\[ u^* = .33 \times u^*_i + z_i, \quad v^* = .33 \times v^*_i + z_i. \]

Table 1 Simulation results.

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>CF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>.858</td>
<td>1.003</td>
</tr>
<tr>
<td>$x_1$</td>
<td>.858</td>
<td>1.003</td>
</tr>
<tr>
<td>$x_2$</td>
<td>.866</td>
<td>1.011</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>1.137</td>
<td>.993</td>
</tr>
<tr>
<td>$\rho$</td>
<td>.120</td>
<td>.108</td>
</tr>
<tr>
<td>$\beta$</td>
<td>.121</td>
<td>.203</td>
</tr>
<tr>
<td>$b$</td>
<td>.245</td>
<td>.245</td>
</tr>
<tr>
<td>$b_{sls}$</td>
<td>.358</td>
<td>.358</td>
</tr>
<tr>
<td>$\delta$</td>
<td>.298</td>
<td>.298</td>
</tr>
</tbody>
</table>

For the control function (CF) method formulated here, Table 1 provides OLS and CF simulation results when $N = 1000$ observations and $R = 100$ replications. Entries in parentheses provide the standard deviations over replications, while other entries provide the means over replications.

Column 1 of this table indicates that the bias in the OLS estimates for the primary model is on the order of 14%. 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 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. With $\rho = .3134$ as the true value of the correlation parameter, the average estimate of .298 is reasonable. The parameter $\delta$ 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, 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 $\delta$. 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

\textsuperscript{13} With estimated expectation functions converging uniformly to positive functions, this expansion is valid on a set with probability approaching one.

\textsuperscript{14} See Newey and McFadden (1994).
stepproducesworthwhilegains.Additionalgainsmaybegpssibleiftheentireerrordistri-butiondependsonasonalindex(asinthesimulations). Further gains are also possible under a “GLS” variant of the CF method presented here. The resulting improvement, while noticeable, was not sufficient to further complicate the exposition. Two important features of the design are the “amount” of heteroscedasticity in the model and the extent to which \((S_n/S_0)\) varies in the sample. While a detailed examination of these issues is beyond the scope of this paper we provide a preliminary examination of the former by setting \(S_n\) to a constant and varying the amount of heteroscedasticity as reflected by \(S_0\). We then estimated the model while treating \(S_n\) as known. To assess the degree of heteroscedasticity present we calibrated the model such that the Breusch–Pagan test, conducted at the 1% level, rejected the degree of heteroscedasticity present we calibrated the model such that the Breusch–Pagan test, conducted at the 1% level, rejected the hypothesis of heteroscedasticity at 25%, 50%, 75% and 95% of the one thousand replications when the model is estimated with \(n = 1000\). These results are reported in Table 2 and indicate that even with relatively little heteroscedasticity the estimator bias is eliminated. There is a large reduction in the variance of the estimator as the level of heteroscedasticity is increased. We note, however, that the \(\rho\) coefficient is estimated with some imprecision.

6. Conclusion

We have examined a triangular simultaneous model where there are no available exclusion restrictions to employ as instruments but where we allow for generalized forms of heteroscedasticity. 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.

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 straightforward 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 the conditional mean rather than the linear dependence of \(u\) on \(v\), it would also be possible to allow for nonlinearities in the endogenous variables.

\footnote{When the entire distribution of the errors depends on a single index, the squared residual and any function of it will satisfy a single index assumption. Presumably, such additional information could be exploited.}

\footnote{Given the normalizations employed the OLS estimates are largely unaffected by the degree of heteroscedasticity. For each of the designs the mean estimates for the constant and the coefficients of the \(x_i\)s were approximately .69 with a standard error of .05. The mean coefficient for \(Y_2\) was 1.31 with a standard error of .04.}

Appendix

The main section of the Appendix provides the proofs for identification, consistency, and asymptotic normality. All intermediate lemmas, which are required in these proofs, are stated and proved in another section of the Appendix.

A.1. Main results

The proof of Theorem 1 outlines a standard argument for the first-stage estimator.

\textbf{Proof of Theorem 1.} Define: \(\psi^2(\pi) \equiv (Y_2i - X_i\pi)^2\), \(S_{11i}(\pi, \delta) \equiv E \left[ \psi^2(\pi) | l_{ii}(\delta) \right]\), and \(X_i^* \equiv X_i/S_{ii}(\pi_o, \delta_o)\). Then, with \(\Omega \equiv p \lim (X_i^*X_i^*/N)\), part (a) follows as the GLS estimator of \(\pi, \hat{\pi}\), satisfies:

\[
\sqrt{N} \left( \hat{\pi} - \pi_o \right) = \Omega^{-1} \sqrt{N} \sum X_i^* \psi_i^* / N + o_p(1). \]

For part (b) of Theorem 1, let:

\[
R(\delta; \hat{\pi}) = N \sum_{i=1}^N \tau_i \tilde{r}_i(\delta; \pi). \quad \tilde{r}_i(\delta; \pi) = \psi^2(\pi) - S_{11i}(\pi, \delta)\]

\[
w_i = \tau_i \frac{\partial}{\partial \delta} \tilde{r}_i(\delta_o; \pi_o), \quad w_i^o = w_i - E \left[ w_i | l_{ii}(\delta_o) \right]. \]

Then, from Ichimura (1993) or from the intermediate lemmas below:\footnote{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.}

\[
\sqrt{N} \left[ \tilde{\delta} - \delta_o \right] = -R^{-1}_{11} \sqrt{N} \sum \tilde{r}_i(\delta_o; \pi_o) w_i^*/N + R_{21} \left[ \tilde{\pi} - \pi_o \right] + o_p(1). \]

\textbf{Proof of Theorem 2.} For the nonparametric case, we first show that \(\text{Var}(u(X)) = S_u^2\) is determined up to a multiplicative constant by sample moments.\footnote{We are grateful to one of the referees for this argument.} Let \(S_u^2 \equiv \text{Var}(V(X))\) and \(\rho \equiv \text{cov}(u, v)/S_uS_v\). With the correlation \(\rho\) constant and \(\rho^2 < 1\):

\[
\text{Var} \left( Y_1 | X \right) = \text{Var} \left( \nu \theta_2 + u | X \right) = \theta_2 \theta_2S_u^2 + 2\theta_2\rho S_uS_v + S_v^2 = \theta_2S_u^2 + \rho S_v^2 + (1 - \rho^2)S_u^2 \]

\[
\text{Var} \left( Y_2 | X \right) = \text{Var} \left( \nu \theta_2 + u | X \right) = \theta_2 \theta_2S_v^2 + 2\theta_2\rho S_uS_v + S_u^2 = \theta_2S_v^2 + \rho S_u^2 + (1 - \rho^2)S_v^2 \]

\[
\text{Var} \left( Y_1, Y_2 | X \right) = (1 - \rho^2)S_u^2 = (1 - \rho^2)S_v^2 \]

\[
\Rightarrow (1 - \rho^2)^{1/2} S_u \quad \Rightarrow \left[ \text{Var} \left( Y_1, Y_2 | X \right) - (1 - \rho^2) \text{cov}(Y_1, Y_2 | X) \right]^{1/2}. \]

Notice that:

\[
A^*(x) \equiv \left[ (1 - \rho^2)^{1/2} S_u/S_v \right]^2 \]

\[
A(x) v = \rho / (1 - \rho^2)^{1/2} A^*(x) v. \]
Therefore, if now we include $A^*(x)v$ as a variable in the model, under the full rank assumption of the theorem, we will directly identify all coefficients other than the correlation parameter. For the correlation parameter, we will identify $\rho/(1 - \rho^2)^{1/2}$. As this known ratio implies the sign of $\rho$, it can be shown that $\rho$ is also identified.

For the index case, let $\alpha \equiv (\theta, b, \rho)$ and recall that conditional variance, response, and objective functions are:

\[ \hat{S}_i(\theta, b)^2 \equiv \hat{E} \left[ u_i(\theta)^2 \mid I_{ui}(b^*) , I_i \right] \]

\[ \hat{S}_i(\theta, b)^2 \equiv \hat{E} \left[ u_i(\theta)^2 \mid I_{ui}(b) \right] \]

\[ \hat{M}_i \equiv \hat{Z}_i \theta + \rho \hat{S}_i(\theta, b) \hat{e}_i/\hat{S}_i; \quad \hat{M}_i \equiv \hat{Z}_i \theta + \rho \hat{S}_i(\theta, b) \hat{e}_i/\hat{S}_i \]

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

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

\[ \hat{\alpha} \equiv \arg \min_{\alpha} \hat{Q}^*(\alpha) \approx \arg \min_{\alpha} \hat{Q}^*(\alpha). \]

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 $k = 1, 2$:

\[ EQ^*(\alpha) = EQ(\alpha) - EQ(\alpha_o) = E\Delta_1 + E\Delta_2, \]

\[ E\Delta_k = E \sum_{i=1}^N t_j |M_i - M\|^2 /N, \]

\[ M_1(\alpha) \equiv Z \theta + \rho S_1(\alpha) v_i / S_i; \]

\[ M_2(\alpha) \equiv Z \theta + \rho S_2(\alpha) v_i / S_i. \]

Since, $M_1(\alpha) - M_2 = 0$, both $Q_1$ and $Q_2$ are separately minimized at the true parameter values. With $\alpha^*$ as a candidate for a minimum, $Q_1$ and $Q_2$ must also be separately minimized at $\alpha^*$. Therefore:

\[ M_1(\alpha^*) - M_2 = 0, \quad k = 1, 2. \]

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

\[ \rho S_{1i}(\alpha) v_i / S_i = Z_i (2B - \theta) + \rho S_2(\alpha) v_i / S_i. \]

Multiply (2B.1) by $v_i$, take an expectation conditioned on $X_i$, and divide by $S_i \neq 0$ to obtain:

\[ \rho \rho S_{1i}(\alpha) v_i = S_1(2B - \theta) + \rho S_2(\alpha). \]

Noting that the r.h.s. only depends on $X$ through $|I_{ui}(b^*) , I_i |$, for $\rho \neq 0$ it follows that:

\[ S_{1i}(\alpha) = E \left[ S_{1i}(\alpha) \mid I_{ui}(b^*) , I_i \right], \quad m = 1, 2. \]

Next, in (2B.1), solve for $\rho^2 S_{2i}(\alpha) v_i / S_i$, square both sides, and take an expectation conditioned on $I_{ui}(b^*)$ and $I_i$ to obtain:

\[ \rho^2 S_{2i}(\alpha)^2 = \rho^2 S_{1i}(\alpha) - B(X_i) + C(X_i). \]

\[ B(X_i) \equiv 2 \rho S_{1i}(\alpha)(\theta - \theta_o) \]

\[ C(X_i) \equiv (\theta - \theta_o) E \left[ Z_i |I_{ui}(b^*) , I_i \right] (\theta - \theta_o). \]

With $u_i(\theta) = Y_i - Z_i \theta, S_{2i}(\alpha)$ is defined as:

\[ S_{2i}(\alpha)^2 \equiv \left[ u_i(\theta)^2 \mid I_{ui}(b^*) , I_i \right] = E \left[ u_i(\theta) - Z_i (\theta - \theta_o)^2 \mid I_{ui}(b^*) , I_i \right] + C(X_i). \]

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

\[ E(u_i v_i | X_i) = \rho \rho S_{1i}(\theta_o) S_{2i} \]

\[ \Rightarrow E(u_i(\theta_o) v_i | I_{ui}(b^*), I_i) = \rho \rho S_{2i} [\theta_o (\theta_o) | I_{ui}(b^*), I_i] \]

\[ = \rho \rho S_{2i}(\theta_o) S_{2i}. \]

Substituting (2B.2) and (2B.5) into (2B.4):

\[ S_{2i}(\alpha)^2 = S_{1i}(\theta_o) - B(X_i) + C(X_i). \]

Differencing (2B.3) and (2B.6):

\[ (1 - \rho^2) S_{1i}(\alpha)^2 = (1 - \rho^2) S_{2i}(\theta_o). \]

Note that $\rho^2 < 1$, because $\rho^2 = 1$ implies $\rho^2 = 1$, violating a restriction. Let $r = \left( (1 - \rho^2) / (1 - \rho^2)^{1/2} \right)$ and substitute (2B.7) into (2B.1) to obtain:

\[ [Z_i, S_{1i}(\theta_o) / S_{2i} v_i] (\theta - \theta_o) / (\rho^2 - \rho_o) = r [\rho^2 - \rho_o] = 0. \]

Under (A7) and the assumption that relative heteroscedasticity depends on $X$: $\theta^* = \theta_o$ and $\rho^* = \rho_o$. Since $\theta^* = \theta_o$, from (2B.6) and (2B.7), $r = 1$. With $\rho^* = \rho_o$: $\rho^* = \rho_o$.

Since $M_{1i} = M_{2i}$,

\[ S_{2i}^2(\theta_o, b^*) = E \left[ u_i(\theta_o)^2 \mid I_{ui}(b^*) , I_i \right] = E \left[ u_i(\theta_o)^2 \mid I_{ui}(b^*) \right] = E \left[ u_i(\theta_o)^2 \mid I_{ui}(b_o) \right]. \]

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

**Proof of Theorem 3.** With $\alpha^c = (\hat{\alpha}, \alpha_o)$, on a set with probability tending to 1, from a standard Taylor series argument:

\[ \sqrt{N} \left[ \hat{\alpha} - \alpha_o \right] = \sqrt{N} \left[ \hat{H}(\alpha^c; \hat{n}) \right]^{-1} \left[ \sqrt{N} \left( \hat{G}_A + \hat{G}_B + \hat{C} \right) \right]. \]

From standard uniform convergence arguments: $\hat{H} \rightarrow H$. For the gradient, from Lemmas GA–GC, the result follows.

**A.2. Intermediate lemmas**

In the primary equation, recall that $u(\theta_o) = Y_i - Z_i \theta_o$ and that $\pi_o$ denotes the true regression coefficients for the secondary equation. Let

\[ r_2 = v_i \pi_o^2 v_i^2 = u_i(\theta)^2, \]

where the error is defined for an arbitrary value of the parameter vector.

For $m = 1, 2$ let $W_m \equiv X_1 + X_2 \gamma_m, \gamma = (\gamma_1, \gamma_2) = (\theta, \gamma)$, and $W = (W_1, W_2)$. With $w_{m} \equiv x_1 + x_2 \gamma_m$ as a conditioning value for $W_m$ and with $\lambda = (\theta, \gamma)$, define $w(\lambda)$ as the conditioning vector. Then, let:

\[ a(w(\lambda); \lambda, s) \equiv g(w) E \left( r_2^s \mid W = w \right), \quad s = 0, 1, \]

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

\[ \hat{a}(w(\lambda); \lambda, s) \equiv \sum_{j=1}^N r_2^s T_{2j} \left[ (w_m - W_m) / h \right]. \]

Referring to (D1), for $M = 1, K_1$ is the twicing kernel; while for $M = 2, K_2$ is the double twicing kernel. Note that

\[ \hat{a}(w(\lambda); \lambda, 1) / \hat{a}(w(\lambda); \lambda, 0) \equiv \hat{f} / \hat{g} \]
estimates a conditional expectation of the form shown in (D1). Define the derivative operator:
\[ \nabla^d_t (\hat{\alpha}) \equiv \frac{\partial}{\partial \alpha} \hat{\alpha}, \quad \text{with} \quad \nabla^d_t (\hat{\alpha}) \equiv \hat{\alpha}. \]

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

(a) \[ \sup_{w,\lambda,\alpha} E \left[ \nabla^d_t \hat{a} (w; \lambda, s) - \nabla^d_t a (w; \lambda, m) \right]^2 \]

\[ = \begin{cases} O (h^d) & : \ M = 1 \\ O (h^{11}) & : \ M = 2 \end{cases} \]

(b) \[ \sup_{w,\lambda,\alpha} \frac{1}{N h^{2(d+M) - 4}} = d = 0, 1, 2. \]

Proof of Lemma 1. The proof for the squared bias result in (a) follows from standard Taylor series expansions in \( h \) about 0. For (b), setting \( d = 0 \) and \( M = 1 \) gives the variance order:

\[ O \left( \frac{\var{1/h f_1 \left[ \left( w - w_i \right)/h \right]}{N} \right) = O \left( \frac{E \left[ (K_1^2 (w - w_i)/h) \right]}{N h^2} \right). \]

Letting \( z = \left( (w - w_i)/h \right) \), a factor of \( h \) disappears in the Jacobian, yielding the result. A similar argument holds for other cases. □

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

\[ \Delta \equiv \sup_{w,\lambda,\alpha} \left| \nabla^d_t \hat{a} (w; \lambda, s) - E \nabla^d_t a (w; \lambda, s) \right| = o_p (1). \]

Proof of Lemma 2. Here, we provide the result for \( d = 0, s = 1, X \equiv \dim(W) = 1 \), and \( r_{2i} = u^2_t (\theta) \). The proofs for other cases are similar. Write

\[ u^2_t (\theta) = u^2_t (\theta_0) - 2u_t (\theta_0) Z_t (\theta - \theta_0) + (\theta - \theta_0) Z_t Z_t (\theta - \theta_0). \]

We outline the argument for the first term as it is similar for the others. Following Ichimura (1993), define the \( t_j \) as an indicator on the indicated set: \( t_j = 1 \) if \( u^2_t (\theta_0) < N^{q_0} \). Then, for observations for which \( t_j = 1 \), the result follows from Hoefding’s inequality (see Lemma 1 of Klein and Spady, 1993). When \( t_j = 0 \) the result follows from an assumption on higher order moments or the tail conditions in (A3). □

To establish asymptotic normality, denote \( \hat{\eta} \) as the vector of estimated parameters from the Y-2-equation (including estimated parameters of the conditional error variance). Write \( \alpha_0 \equiv (\theta^0, \rho^0, b^0)' \) for the vector of true parameter values from the \( Y_1 \)-model. Refer to (D5) for the definitions of the estimated regression functions \( \hat{M}_{i\theta} \) and the objective function for estimating the principal equation (\( Y_1 \)-Model). Taken with respect to \( \alpha \) let \( \hat{G} (\alpha_0; \hat{\eta}) \) be the corresponding gradient. From (D4), recall that \( \hat{z}_{i\theta} \) is a smooth trimming function for observations where \( z_{i\theta} \leq 0 \). Denote \[ \hat{\eta}_{\theta} \equiv \hat{\eta}_{\theta; \theta_0} \]

\[ \left[ \nabla^d_t \hat{M}_{i\theta} (\alpha_0, \eta_0) \right], \quad \hat{w}_{i\theta} \equiv \left[ \hat{w}_{i\theta} + \hat{w}_{i\theta} \right], \quad \text{and} \quad \hat{G}_{i\theta} \equiv \nabla^d_t \hat{G} (\alpha_0, \eta_0) \]

\[ \equiv \hat{G}_{i\theta} + \hat{G}_{i\theta}, \quad \hat{G}_{i\theta} \equiv \sum_{i=1}^N \hat{t}_i \hat{M}_{i\theta} (\alpha_0, \eta_0) - M_{\theta i} \hat{w}_{i\theta}/N. \]

To analyze the above terms we need the following result on indicators:

Lemma 3 (Indicator Bounds). Let \( v_0 \equiv X_1 + X_2 \theta_0 \) and \( \hat{v}_1 \equiv X_1 + X_2 \partial \hat{\theta} \), define:

\[ \tau (X) \equiv 1 (c_1 < v < c_2): \quad \hat{\tau} (X) \equiv 1 (\hat{c}_1 < \hat{v} < \hat{c}_2) \]

\[ S (z) \equiv \left[ 1 + \exp[- (N^{(1-c)} + z) / N^{(1-c)/2}] \right]^{-1}, \quad 0 < c < s \]

\[ S_{a\theta} \equiv S (|\hat{v}_1 - v_0| + |\hat{c}_1 - c_1| - |v_0 - c_1|) + 1 - S (0) \]

\[ S_{a\theta} \equiv S (|v_1 - v_0| + |\hat{c}_2 - c_2| - |v_0 - c_2|) + 1 - S (0). \]

Then:

\[ |\tau_i - \hat{\tau}_i| \leq S_{a\theta} + S_{a\theta}. \quad (A) \]

Assume that \( x \equiv [x_1 x_2] \) is bounded and that for fixed \( q_0 \equiv (\theta_0, c_1, c_2) \):

\[ \hat{\theta} - \theta_0 = o_p (N^{-s}) \quad \text{and} \quad |\hat{c}_1 - c_1| = o_p (N^{-s}) \]

\[ |\hat{c}_2 - c_2| = o_p (N^{-s}). \]

Then, with \( q \equiv (\theta, c_1, c_2) \) restricted to an \( o_p (N^{-s}) \) neighborhood of \( q_0 \) and \( s > 0 \):

\[ \sup_{q \in \Theta} \frac{1}{N} \sum |\tau_i (q) - \tau_i (q_0)| = o_p (N^{-(s-c)}). \quad (B) \]

Proof of Lemma 3. The proof for (A) directly follows from Klein (1993, Lemma A.1) that uses a result due to Jim Powell. Part (B) will follow if \( S_{a\theta} \) and \( S_{\theta} \) are each uniformly \( o_p (N^{-(s-c)}) \). With the arguments for both terms being identical, consider \( S_{a\theta} \). For this term, since \( 1 - S (0) = o_p (N^{-s}) \) we require a uniform result on:

\[ S (\delta - w), \quad \delta \equiv v - v_0 + |c_1 - c_1| - |v_0 - c_1|; \]

\[ w \equiv |v_0 - c_1|. \]

Taylor expand \( S (\delta - w) \) in \( \delta \) about \( \delta = 0 \). The proof now follows similarly to that in Klein (1993, Proof of Lemma A.2).

Lemma GA (Gradient Component \( \hat{G}_{i\theta} \)). Refer to the definition of gradient components in (A21). With \( \hat{M}_{i\theta} (\alpha_0, \eta_0) \) defined in (D5) above and with \( z_{i\theta} \equiv (X_1, Y_2) \),

\[ \nabla^d_t \hat{M}_{i\theta} (\alpha_0, \eta_0) \equiv \left[ Z_{i\theta} + \rho_{i\theta} \left( \nabla^d_t \hat{G}_{i\theta} (\alpha_0, \eta_0) \right) \right], \quad \hat{G}_{i\theta} \equiv \left[ \hat{\eta}_{\theta} \right] \]

\[ \equiv \left[ \hat{\rho}_{i\theta} \left( \nabla^d_t \hat{G}_{i\theta} (\alpha_0, \eta_0) \right) \right]. \]

Recalling that \( \hat{w}_{i\theta} \equiv \hat{w}_{i\theta} + \hat{w}_{i\theta} \), define \( \hat{u}_{i\theta} \) by replacing all estimated functions with their probability limits. Then, with \( r_{\tau_0} \equiv \tau_i (q_0) \), for \( \hat{G}_{i\theta} \) in (A21):

\[ \sqrt{N} \hat{G}_{i\theta} = \sqrt{N} \hat{G}_{i\theta} + o_p (1). \]

\[ G_{i\theta} \equiv - \frac{1}{N} \sum r_{\tau_0} (Y_{i\theta} - M_{\theta i}) w_{i\theta}. \]

Proof of Lemma GA. From (D2), with \( \hat{\tau}_i \equiv \tau_i (\hat{q}) \), \( \sqrt{N} \left[ \hat{G}_{i\theta} - \hat{G}_{i\theta} \right] \) is the sum of the following three terms:

\[ A \equiv \frac{1}{N} \sum r_{\tau_0} (Y_{i\theta} - M_{\theta i}) \hat{w}_{i\theta} \]

\[ B \equiv \frac{1}{N} \sum r_{\tau_0} (Y_{i\theta} - M_{\theta i}) \left( \hat{w}_{i\theta} - w_{i\theta} \right) \]

\[ C \equiv \frac{1}{N} \sum r_{\tau_0} (Y_{i\theta} - M_{\theta i}) \tau_i (\hat{w}_{i\theta} - w_{i\theta}). \]
Employing a similar strategy to that in Klein (1993), denote \( q_0 \) as a vector of population quantiles (see (D2), Section 4) and let \( N_e := \{ q : |q - q_0| < \varepsilon \}, \varepsilon = o(1) \). Then, \( A = o_p(1) \) if
\[
A^* = \sup_{N_e} N^{-1/2} \sum_{i \in N_e} \left| Y_{ii} - M_{ii} \right| (\tau_1(q_i) - \tau_1(q_0)) |w_i| = o_p(1)
\]
for all \( \varepsilon = o(1) \).

This result follows from Pakes and Pollard (1989, Lemma 2.17, p. 1037). For \( B \), let \( \tau^*_1(q_1) \) be an indicator defined on the union of the sets over which \( \tau_1(q_i) \) and \( \tau_i(q_0) \) are defined. Letting \( N_\delta := \{ q : N^{(r-\delta)} |q - q_0| < \varepsilon \}, \delta > 0 \), it suffices to show that for all \( \varepsilon = o(1) \) (see previous footnote):
\[
B^* = \sup_{N_\delta} N^{-1/2} \sum_{i \in N_\delta} \left| Y_{ii} - M_{ii} \right| (\tau_1(q_i) - \tau_i(q_0)) \tau^*_1(q_i) |\hat{w}_i - u_i| \leq o_p(1).
\]

This result follows from Lemmas 1 and 3 and from Cauchy’s inequality.

For \( C \), the analysis is similar to that in Klein and Spady (1993), with the result obtained by showing that \( E(C^2) \) converges to 0.

Lemmas 4 and 5 simplify gradient component, \( \hat{G}_B \equiv G_{B1} + G_{B2} \), so as to enable a U-statistic argument in Lemma GB.

**Lemma 4.** With \( \hat{G}_{Bk} \) given in (A.21) and \( w_{ik} \) and \( \hat{w}_{ik} \) defined as in Lemma GA:
\[
N^{1/2} \hat{G}_{Bk} = \sqrt{N} \sum_{i} \left\{ \hat{M}_{ik} - M_{ik} \right\} w_{ik}/N + o_p(1).
\]

**Proof of Lemma 4.** Referring to \( \hat{r}_i \hat{w}_i \) as an estimated weight, the difference in terms with estimated and true weights is given as:
\[
\Delta_{ik} = N^{1/2} \sum_{i} \left\{ \hat{M}_{ik} - M_{ik} \right\} \left( \hat{r}_i \hat{w}_{ik} - r_{ik} w_{ik} \right)/N,
\]
and the result follows from Cauchy’s inequality and repeated application of Lemmas 1 and 3.

**Lemma 5.** Referring to Lemma 4 and the definition of \( \hat{M}_{ik} \) in (D5):
\[
N^{1/2} \hat{G}_{Bk} = N^{1/2} \sum_{i} \tau_i v_i \left( \sum_{i} \left( \hat{S}_{ii} - M_{ii} \right) \right) w_{ik}/N
\]
\[
= N^{1/2} T_{ik} = N^{1/2} T_{2ik}.
\]

Definition D1 provides the ratio forms for \( \hat{S}_{ii} \) (\( \alpha \); 1), \( \hat{S}_{ii} \) (\( \alpha \); 2), and \( \hat{S}_{ii} \) (\( \alpha \); 2). Accordingly, define:
\[
\hat{S}_{ii} (\alpha_1; k) = \hat{f}_{ii} (\alpha_1; k)/\hat{g}_{ii} (\alpha_1; k); \quad \hat{S}_{ii} (\alpha_2; k) = \hat{f}_{ii} (\alpha_2; k)/\hat{g}_{ii} (\alpha_2; k);
\]
\[
a_{ii} = \left[ \frac{\hat{S}_{ii} (\alpha_1; k) S_{ii} \hat{S}_{ii} (\alpha_2; k) S_{ii}}{2 \hat{g}_{ii} (\alpha_1; k) S_{ii} \hat{g}_{ii} (\alpha_2; k) S_{ii}} \right]^{-1}.
\]

Then, on a set with probability tendency to one:
\[
N^{1/2} \hat{G}_{Bk} = N^{1/2} \hat{T}_{ik} = N^{1/2} T_{2ik} + o_p(1),
\]
\[
T_{ik} = \sum_{i} \tau_i \left( \hat{f}_{ii} (\alpha; k) - \hat{g}_{ii} (\alpha; k) S_{ii} \right) a_{ii} w_{ik}/N
\]
\[
T_{2ik} = \sum_{i} \tau_i \left( \hat{f}_{ii} (\alpha; k) - \hat{g}_{ii} (\alpha; k) S_{ii} \right) a_{ii} w_{ik}/N.
\]

**Proof of Lemma 5.** For the term \( T_{ik} \), from Lemmas 1 and 3:
\[
N^{1/2} T_{ik} = N^{1/2} \sum_{i} \tau_i \left( \hat{S}_{ii} (\alpha; k) - S_{ii} \right) \left( \hat{S}_{ii} / S_{ii} \right) + o_p(1).
\]

On a set with probability approaching 1, Taylor expand \( \left( \hat{S}_{ii} (\alpha; k) \right) \) about \( S_{ii} \) and employ Lemmas 1 and 3 to obtain:
\[
N^{1/2} T_{ik} = N^{1/2} \sum_{i} \tau_i \left( \hat{S}_{ii} (\alpha; k) - S_{ii} \right) \left( \hat{S}_{ii} / S_{ii} \right) + o_p(1)
\]
\[
= N^{1/2} \sum_{i} \tau_i \left( \hat{S}_{ii} (\alpha; k) - S_{ii} \right) \left( \hat{S}_{ii} / S_{ii} \right) + o_p(1)
\]
\[
= N^{1/2} \sum_{i} \tau_i \left( \hat{S}_{ii} (\alpha; k) - S_{ii} \right) \left( \hat{S}_{ii} / S_{ii} \right) + o_p(1)
\]
\[
\times \left( \hat{S}_{ii} / S_{ii} \right) w_{ik}/N + o_p(1),
\]
which completes the argument. The proof for \( T_{2ik} \) is identical.

**Lemma GB (U-Statistic Projection).** Referring to Lemma 5, write the trimming indicator as \( \tau_i = \tau_i \tau_{ik} \), the product of index and X-trimming indicators. Then, with indices \( i_N \) and \( i_T \) evaluated at the true parameter values:
\[
N^{1/2} \hat{G}_B = N^{1/2} G_B + o_p(1),
\]
where \( G_B = G_{B1} + G_{B2} \)
\[
G_{Bk} = \sum_{i} \left( u_i^2 - \hat{S}_{ii} \right) E \left[ \tau_i a_{ii} w_{ik} | i_N \right] / N
\]
\[
- \sum \left( u_i^2 - \hat{S}_{ii} \right) E \left[ \tau_i a_{ii} w_{ik} | i_T \right] / N,
\]
where \( i_N \) (k) \( = i \) (l) \( = i_T \) and \( l_N \) (k) \( = i \) (l) \( = k = 2 \).

**Proof of Lemma GB.** Recalling the definition of the twining kernel \( \hat{k} \) in (D1) and the conditional variance index \( I_{ik} \) in (A4), define:
\[
k_{i} (i, j) \equiv K_i \left( \left( l_{ik} (j) - l_{ik} (j) / h \right) \right) / h.
\]

Then, for \( k = 1 \) (the argument for \( k = 2 \) is identical), refer to Lemma 5 and write:
\[
T_{ik} = U_N \equiv \sum_{i} \tau_i \left( \hat{f}_{ii} (\alpha; 1) - \hat{g}_{ii} (\alpha; 1) \right) \hat{S}_{ii} a_{ii} w_{ii}/N
\]
\[
= \frac{1}{N(N-1)} \sum_{j} \rho_{ij} \rho_{ij}^*\left( u_i^2 k_{i} (j) - k_{i} (j) \right) a_{ii} w_{ii}/N.
\]

As the above expression is a U-statistic with \( \sqrt{N} E (U_N) \to 0 \), from standard projection arguments:
\[
\sqrt{N} U_N = \sqrt{N} \frac{2}{N} \sum_{i} E \left( \left( \rho_{ij}^* + \rho_{ij}^* \right) / 2 \right) Y_{ii}, Z_i + o_p(1)
\]
\[
= \sum \left( u_i^2 - \hat{S}_{ii} \right) E \left[ \tau_i a_{ii} w_{ik} | i_N \right] / N.
\]
Lemma GC. Referring to the definition of gradient components in (A.21) and employing the notation in Theorem 1, define $D \equiv \lim (\nabla G_{\eta}(\theta_0, b_0; \eta_0))$. Then:

$$\sqrt{N}\hat{G}_C = \sqrt{N} (G_C) + o_p(1),$$

$G_C \equiv D \left( -R_1^{-1} \left[ r_1(\delta_0; \pi_0) w_1^* + R_2 \Omega^{-1} \varepsilon_{x_1} \right] \right).$

Proof of Lemma GC. The proof follows from (A.21), Theorem 1, and a standard uniform convergence result. □

Lemma G. With $\hat{G}$ as the gradient to the objective function (see (A.21)) and with $G_A$, $G_B$, and $G_C$ given in Lemmas GA–GC:

$$\sqrt{N}\hat{G} = \sqrt{N} G + o_p(1), \quad G = G_A + G_B + G_C$$

with $G$ being a $K \times 1$ vector of sample means of the form: $G = \sum G_i / N$. Let $D$ be the $N \times K$ matrix with ith row $D_i \equiv G_i$. Then, with $\Omega \equiv E [D' D / N] :$

$$\text{Var} \left[ \sqrt{N}\hat{G} \right] = \Omega; \quad \hat{\Omega} \equiv D' D / N \to \Omega.$$

References


