G. S. Maddala, C. R. Rao and H. D. Vinod, eds., *Handbook of Statistics, Vol. 11*

# Efficient Estimation of Models with Conditional Moment Restrictions

*Whitney K. Newey*

## 1. Introduction

Often an economic data set is used to answer questions for which it was not designed, because it is too expensive to collect data to answer each new question. The lack of control of empirical workers over the form of the data suggests the need for models that impose few restrictions on the distribution of the data. Also, because economic data often embodies response of individual agents to market conditions and/or economic incentives, it is important to have models that control for these responses. The simultaneous equations models of econometrics, and corresponding instrumental variable methods, are designed to control for such phenomena: see Hausman (1984) for further motivation and discussion.

A useful type of model that imposes few restrictions and can allow for simultaneity is a conditional moment restriction model, where all that is specified is that a vector of *residuals*, consisting of known, prespecified functions of the data and parameters, has conditional mean zero given known variables. An example is regression with a disturbance that has conditional mean zero given the regressors, where the residual is the usual difference of the dependence variable and linear combination of the regressors. Other examples, to be discussed below, can allow for incorporation of information about the variance of the disturbance and for simultaneity. Estimators for the parameters of these models can be constructed by interacting functions of the residuals with functions of the conditioning variables and choosing the parameter estimates so that the sample moments of these interactions are zero. These estimators are conditional, implicit versions of the method of moments, that are typically referred to as *instrumental variables* (IV) estimators, where the *instruments* are the functions of conditioning variables that interacted with the residuals. These estimators have the usual advantage of method of moments over maximum likelihood, that their consistency only depends on correct specification of the residuals and conditioning variables, and not on the correctness of a likelihood function. Of course, maximum likelihood may be more efficient than IV if the distribution is correctly specified, so that the usual bias/efficiency tradeoff is present for IV and maximum likelihood. Further description and motivation for IV estimators is given in Sections 2 and 3.

The precision of estimators for conditional moment models is a concern, because of the restrictions imposed are so weak. The purpose of this chapter is to discuss (asymptotically) efficient estimation of the parameters of conditional moment restriction models. Two notions of efficiency are of interest here. One is efficiency within the class of IV estimators and the other is efficiency in the semiparametric model sense of Stein (1956). It turns out that the two notions result in the same estimator, because the optimal estimator in the class of IV estimators is efficient for the semiparametric model. Thus, as far as efficient estimation is concerned, it suffices to restrict attention to the class of IV estimators.

Several approaches to efficient estimation are considered. Each is based on constructing an estimator of the optimal instruments, i.e., of those functions that are interacted with the residual to obtain the IV estimator with smallest asymptotic variance. These approaches will work because suitably regular estimation of the instruments has no effect on the distribution of estimators of the parameters of interest (a kind of information matrix block diagonality for the parameters of interest and parameters of the instruments). One approach is to specify the optimal instruments to be functions of auxiliary parameters, and replace the auxiliary parameters with consistent estimators. The resulting estimator of the parameters of interest will be efficient if the optimal instruments take the specific parametric form, and will be consistent even if they do not. Another approach is to use a nonparametric estimator of the optimal instruments. The resulting estimator of the parameter of interest will be efficient over all possible forms for the optimal instruments, subject to regularity conditions. Two types of nonparametric estimators are considered, one based on nearest neighbor estimation of conditional expectations that form the optimal instruments and the other on estimation of the optimal instruments by linear combinations of prespecified functions. Asymptotic efficiency of the corresponding estimators of parameters of interest is shown in this chapter, and a small Monte Carlo study of their properties in the heteroskedastic linear regression model is given. Also, some data based methods for choosing the number of nearest neighbors or the number of approximating functions in the nonparametric estimator are given.

The new results in this chapter are those on nonparametric estimation of the optimal instrumental variables. Other results have previously been given elsewhere. The form of the optimal instruments was derived by Amemiya (1974) and Berndt, Hall, Hall and Hausman (1974) for the homoskedastic residual case and Hansen (1985) for the general case. Chamberlain (1987) showed that the semiparametric efficiency bound for conditional moment restrictions models was attained by optimal instrumental variables estimators. Also, results on estimation when the optimal instruments are replaced by parametric estimators are given in Carroll and Ruppert (1988), nonparametric estimation of the optimal instruments has been considered by Carroll (1982), Robinson (1987), and Newey (1988) for the linear regression model, and by Newey (1990) for the homoskedastic residual case.

Section 2 of the paper introduces the model, describes IV estimators, derives the form of the optimal instruments, and describes parametric estimation of the optimal instruments. Section 3 gives a number of examples, including estimation of regression models, models with second moment information, transformation models, and models with simultaneity. In each case a parametric estimator of the optimal instruments is discussed. Section 4 deals with nearest neighbor nonparametric estimation of the optimal instruments, Section 5 with estimation via linear combinations of functions, and Section 6 reports on a small Monte Carlo study.

## 2. Conditional moment restrictions and instrumental variables estimation

The general type of model that will be dealt with in this chapter can be described as follows. Let $z$ denote a single $p \times 1$ observation on all the variables, and denote the data by $z_1, \ldots, z_n$. Let $\theta$ be a $q \times 1$ vector of parameters, $\rho(z, \theta)$ an $s \times 1$ *residual* vector of functions of a data observation and the parameter, and $x$ a vector of conditioning variables. The conditional moment restriction model considered here is one where the true distribution of the data satisfies

$$\mathrm{E}[\rho(z, \theta_0)|x] = 0\,, \tag{2.1}$$

where $\theta_0$ denotes the true value of the parameters. An example of this type of model is the linear regression model

$$y = x'\beta_0 + \varepsilon\,, \qquad \mathrm{E}[\varepsilon\,|\,x] = 0\,, \tag{2.2}$$

where $z = (y, x')$, $\theta = \beta$, $s = 1$, and the residual is $\rho(z, \theta) = y - x'\theta$.

The conditional moment restriction of equation (2.1) implies that $\rho(z, \theta_0)$ is uncorrelated with functions of $x$, an unconditional moment restriction. This restriction can be used to estimate $\theta_0$ by setting the sample cross-product of $\rho(z, \theta)$ with functions of $x$ close to zero. To describe this estimator, let $A(x)$ denote an $r \times s$ matrix of functions of $x$. Then $\mathrm{E}[A(x)\rho(z, \theta_0)] = 0$ by equation (2.1) and iterated expectations, suggesting a method of moments estimator that sets the sample moment of $A(x)\rho(z, \theta)$ equal to its population value of zero. When $r > q$ it will generally not be possible to set the sample moments to be zero, but a similar method of moments estimator can be obtained by minimizing a quadratic form in the sample moments. Let $\hat{P}$ denote an $r \times r$ positive semi-definite matrix that may be random, and consider the estimator

$$\hat{\theta} = \mathrm{argmin}_{\theta \in \Theta} \hat{g}_n(\theta)'\hat{P}\hat{g}_n(\theta)\,, \qquad \hat{g}_n(\theta) = \sum_{i=1}^{n} A(x_i)\rho(z_i, \theta)/n\,, \tag{2.3}$$

where $\Theta$ is some set of feasible values for $\theta$.

This type of estimator was developed and analyzed in a number of econometrics papers, including Sargan (1959), Amemiya (1974), Jorgenson and Laffont (1974), Burguete, Gallant and Souza (1982), Hansen (1982), and Newey (1990). It is often referred to as a nonlinear IV estimator, and $A(x)$ as instruments. This type of estimator dates to Reiersol (1945), who suggested IV estimation as a method to treat measurement error in linear regression models. Also, IV estimators are very general, in that they include many familiar estimators as special cases. For example, in the linear model of equation (2.2), for a weight function $w(x)$ the weighted least squares estimator $\hat{\beta} = (\Sigma_{i=1}^{n} w(x_i)x_ix_i')^{-1} \Sigma_{i=1}^{n} w(x_i)x_iy_i$ solves equation (2.3) for instruments $A(x) = w(x)x$.

It is straightforward to derive the asymptotic variance of $\hat{\theta}$, as needed to discuss its (asymptotic) efficiency, by applying the usual mean-value expansion argument to the first-order conditions $\partial\hat{g}_n(\hat{\theta})/\partial\theta'\hat{P}\hat{g}_n(\hat{\theta}) = 0$ for $\hat{\theta}$. Suppose that there is a positive semi-definite matrix $P$ such that $\hat{P} \xrightarrow{p} P$, that a uniform law of large numbers gives $\partial\hat{g}_n(\bar{\theta})/\partial\theta \xrightarrow{p} \mathrm{E}[A(x)\,\partial\rho(z,\theta_0)/\partial\theta] = G$ for any $\bar{\theta} \xrightarrow{p} \theta_0$, and that $z_1, \ldots, z_n$ are iid so that a central limit theorem gives $\sqrt{n}\hat{g}_n(\theta_0) \xrightarrow{d} \mathrm{N}(0, V)$, $V = \mathrm{E}[A(x)\rho(z,\theta_0)\rho(z,\theta_0)'A(x)']$. The expanding $\hat{g}_n(\hat{\theta})$ around $\theta_0$, solving for $\sqrt{n}(\hat{\theta} - \theta_0)$, replacing estimated averages by their probability limits, and applying the Slutzky theorem gives

$$\begin{aligned} n(\hat{\theta} - \theta_0) &= -\left(\frac{\partial\hat{g}_n(\hat{\theta})}{\partial\theta'}\,\frac{\hat{P}\,\partial\hat{g}_n(\bar{\theta})}{\partial\theta}\right)^{-1} \frac{\partial\hat{g}_n(\hat{\theta})}{\partial\theta'}\,\hat{P}\sqrt{n}\hat{g}_n(\theta_0) \\ &= -(G'PG)^{-1}G'P\sqrt{n}\hat{g}_n(\theta_0) + \mathrm{o}_{\mathrm{p}}(1) \\ &\xrightarrow{d} \mathrm{N}(0, (G'PG)^{-1}G'PVPG(G'PG)^{-1})\,. \end{aligned} \tag{2.4}$$

The asymptotic variance of $\hat{\theta}$ depends on the assumptions of independent observations through the form of $V$. With dependent observations, $V$ might include covariances between $A(x)\rho(z,\theta_0)$ for different observations, rather than being just an expected outer product. Because of this complication the analysis of optimality is much more difficult with dependent observations, although Hansen (1985) and Hansen, Heaton and Ogaki (1988) have made progress on this problem. Here attention is restricted to the iid case to avoid further complications, and because the assumption of iid observations is sufficiently general to cover many cross-section and longitudinal data models of interest.

The asymptotic variance matrix $(G'PG)^{-1}PVPG(G'PG)^{-1}$ depends on both $P$ and $A(x)$. As shown by Hansen (1982), the optimal choice of $P$, that minimizes the asymptotic variance, is $P = V^{-1}$. This optimal $P$ can be implemented by using $\hat{P} = \hat{V}^{-1}$ for a consistent estimator $\hat{V}$ of $V$ (because the asymptotic variance depends only on the probability limit of $\hat{P}$).

The main focus here is the optimal, asymptotic variance minimizing choice of instruments $A(x)$. It will turn out that the optimal $A(x)$ has $q$ rows, so that the asymptotic variance of the corresponding estimator does not depend on $P$. To describe the optimal $A(x)$, let

$$D(x) \equiv \mathrm{E}[\partial\rho(z,\theta_0)/\partial\theta \,|\, x]\,, \qquad \Omega(x) \equiv \mathrm{E}[\rho(z,\theta_0)\rho(z,\theta_0)' \,|\, x]\,. \tag{2.5}$$

The optimal instruments are

$$B(x) = C \cdot D(x)'\Omega(x)^{-1}\,, \tag{2.6}$$

where $C$ is any nonsingular matrix. The asymptotic variance matrix for these instruments is

$$\Lambda = (\mathrm{E}[D(x)'\Omega(x)^{-1}D(x)])^{-1}\,. \tag{2.7}$$

For example, in the linear model of equation (2.2), $D(x) = -x'$ and $\Omega(x) = \mathrm{E}[\varepsilon^2 \,|\, x]$, so that for $C = -I$, $B(x) = x/\sigma^2(x)$. The corresponding IV estimator is just a weighted least squares estimator with weight $1/\sigma^2(x)$, i.e., heteroskedasticity corrected generalized least squares.

The form of the optimal instruments has an intuitive explanation by way of an analogy with the linear model. The term $\Omega(x)^{-1}$ is a correction for heteroskedasticity (and correlation between different components of $\rho(z,\theta_0)$) similar to that for the linear model. The derivatives $\partial\rho(z,\theta_0)/\partial\theta$ correspond to the regressors, because as usual the model can be treated as approximately linear when calculating the asymptotic variance. These derivatives are not allowed as instruments in general, because they may not depend just on $x$, but the matrix $D(x)$ is the function of $x$ that is most closely correlated with $\partial\rho(z,\theta_0)/\partial\theta$.

It is straightforward to show that $B(x)$ gives the optimal instruments. First, note that $\Lambda$ does not depend on $C$, so that it suffices to show the result with $C = I$ in equation (2.6). Let $m_A = G'PA(x)\rho(z,\theta_0)$ and $m_B = B(x)\rho(z,\theta_0)$. Then by iterated expectations, $\mathrm{E}[m_A m_B'] = G'PE[A(x)\Omega(x)B(x)'] = G'PE[A(x)\Omega(x)B(x)'] = G'PE[A(x)D(x)] = G'PG$, $G'PVPG = \mathrm{E}[m_A m_A']$, and $\Lambda = (\mathrm{E}[m_B m_B'])^{-1}$. Therefore,

$$\begin{aligned}
&(G'PG)^{-1}G'PVPG(G'PG)^{-1} - \Lambda \\
&\quad = (\mathrm{E}[m_A m_B'])^{-1}\mathrm{E}[m_A m_A'](\mathrm{E}[m_B m_A'])^{-1} - (\mathrm{E}[m_B m_B'])^{-1} \\
&\quad = (\mathrm{E}[m_A m_B'])^{-1}\{\mathrm{E}[m_A m_A'] \\
&\qquad - \mathrm{E}[m_A m_B'](\mathrm{E}[m_B m_B'])^{-1}\mathrm{E}[m_B m_A']\}\mathrm{E}[m_B m_A'] = \mathrm{E}[RR']\,, \\
&R = (\mathrm{E}[m_A m_B'])^{-1}\{m_A - \mathrm{E}[m_A m_B'](\mathrm{E}[m_B m_B'])^{-1}m_B\}\,,
\end{aligned} \tag{2.8}$$

and $\mathrm{E}[RR']$ is positive semi-definite, showing that $\Lambda$ is a lower bound for the asymptotic variance of all IV estimators.

Chamberlain (1987) showed that $\Lambda$ is a lower bound in an even stronger sense. In the semiparametric model where the only substantive restriction imposed on the distribution of the data is equation (2.1), $\Lambda_B$ is an asymptotic minimax bound. Consequently, $\Lambda_B$ is a lower bound for the asymptotic variance of any consistent, asymptotically normal (regular) estimator, and not just for IV estimators.

It is generally not feasible to use the optimal instruments $B(x)$ to form an efficient estimator, because they depend on unknown parameters and/or functions. Feasible approaches to efficient estimation can be based on an estimator of $\hat{B}(x)$ of $B(x)$. Such an estimated optimal instrument could be used in place of $A(x)$ in equation (2.3), with $\hat{P} = (\Sigma_{i=1}^n \hat{B}(x_i)\hat{B}(x_i)'/n)^{-1}$. The choice of $\hat{P}$ will not matter asymptotically, but this one has the virtue of making the objective function invariant to nonsingular linear transformations of the instruments and so may improve computation. The resulting estimator will be

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \sum_{i=1}^n \rho(z, \theta)' \hat{B}(x_i)' \left[ \sum_{i=1}^n \hat{B}(x_i)\hat{B}(x_i)' \right]^{-1} \sum_{i=1}^n \hat{B}(x_i)\rho(z, \theta) \right\} . \tag{2.9}$$

This approach, based on using estimators of the optimal instruments, should work because estimation of optimal instruments should not affect the asymptotic distribution of the estimator. Intuitively, equation (2.1) implies that variation of $A(x)$ around $B(x)$ that is asymptotically small in an appropriate sense will not affect the asymptotic distribution of $\hat{\theta}$.

Linearized versions of feasible efficient estimators are convenient because an initial estimator is often needed anyway to form an estimator of the optimal instruments and the linearized estimator does not require iteration. Let $\hat{\theta}$ be an initial estimator, e.g., obtained form equation (2.3) with $A(x)$ and $\hat{P}$ known. One Newton–Raphson step toward the solution of (2.3) starting at $\hat{\theta}$, with $A(x) = \hat{B}(x)$ (i.e., one Newton–Raphson step toward the solution of $\Sigma_{i=1}^n \hat{B}(x_i)\rho(z_i, \theta) = 0$) gives

$$\tilde{\theta} = \hat{\theta} - \left[ \sum_{i=1}^n \hat{B}(x_i)\, \partial\rho(z_i, \hat{\theta})/\partial\theta \right]^{-1} \sum_{i=1}^n \hat{B}(x_i)\rho(z_i, \hat{\theta}) . \tag{2.10}$$

One approach to estimation of the optimal instruments is to assume that $D(x) = D(x, \eta_0)$ and $\Omega(x) = \Omega(x, \eta_0)$ for some known functions $D(x, \eta)$ and $\Omega(x, \gamma)$ and real vector $\eta$, to construct an estimator $\hat{\eta}$ of $\eta_0$, and form $\hat{B}(x) = D(x, \hat{\eta})'\Omega(x, \hat{\eta})^{-1}$. For example, since $D(x)$ and $\Omega(x)$ are conditional expectations, the parameters $\eta$ could be estimated by, say, least squares regression with the elements of $\partial\rho(z_i, \hat{\theta})/\partial\theta$ and $\rho(z_i, \hat{\theta})\rho(z_i, \hat{\theta})'$ as dependent variables. This approach will lead to an efficient IV estimator if the specification of $D(x, \eta)$ and $\Omega(x, \eta)$ is correct. The IV estimator will not be efficient otherwise, although it will remain consistent. This approach will be discussed in the context of the examples of the next section.

Nonparametric estimates of the optimal instruments are also useful. Although the structure of the model may result in some elements of $D(x)$ and/or $\Omega(x)$ having known functional form, knowledge of the functional form of all these functions will generally require auxiliary distribution assumptions and/or difficult calculations. For example, in the linear model knowing the form of $\Omega(x) = \mathrm{E}[\varepsilon^2 \mid x]$ amounts to knowing the form of heteroskedasticity, and in models where $D(x) = \mathrm{E}[\partial\rho(z, \theta_0)/\partial\theta \mid x]$ it may be difficult to specify $D(x, \eta)$ so that it is consistent with the model, or to even calculate it when the conditional distribution of $z$ given $x$ is specified as an auxiliary assumption. Nonparametric estimation of the optimal instruments provides a means of constructing efficient estimators that do not rely on auxiliary assumptions or the outcome of difficult calculations.

## 3. Examples

This section describes a number of examples, in order to illustrate the broad applicability of conditional moment restriction models. These applications include standard ones, such as nonlinear regression, and some that are less familiar, but that are important in econometrics, such as models with endogenous dummy variables.

### *3.1. Nonlinear regression*

To start on common ground, the first example is one that is very familiar. Suppose that $z = (y, x')$ where $y$ is a dependent variable, that $\beta$ is a vector of parameters with true value $\beta_0$, and that

$$y = f(x, \beta_0) + \varepsilon\,, \qquad \mathrm{E}[\varepsilon \mid x] = 0\,. \tag{3.1.1}$$

This model is the standard nonlinear regression model, where the only restriction imposed on the disturbance distribution is that it has conditional mean zero. It is similar to the linear model example of equation (2.2), in that it is a special case of the general conditional moment restriction model of equation (2.1) with $\theta = \beta$ and $\rho(z, \theta) = y - f(x, \beta)$ being the regression residual.

In this example an optimal instrument $B(x) = D(x)'\Omega(x)^{-1}$ and its components $D(x)$ and $\Omega(x)$ have simple formulas,

$$\begin{aligned} D(x) &= -\frac{\partial f(x, \beta_0)'}{\partial\beta}\,, \qquad \Omega(x) = \mathrm{E}[\varepsilon^2 \mid x]\,, \\ B(x) &= (\mathrm{E}[\varepsilon^2 \mid x])^{-1}\frac{\partial f(x, \beta_0)}{\partial\beta}\,. \end{aligned} \tag{3.1.2}$$

Here the functional form of $D(x)$ is known, so that to construct an estimator $\hat{D}(x)$ it suffices to replace $\beta_0$ with an estimator $\hat{\beta}$. This feature of $D(x)$ having known functional form results from the derivative of the residual depending only on the conditioning variables. The more difficult components to deal with is the conditional variance $\mathrm{E}[\varepsilon^2 \mid x]$. An estimator that is efficient over a parametric family of conditional variances can be constructed by specifying a functional form $\Omega(x, \eta)$ for the conditional variance, and using an estimator $\hat{\eta}$ of the true value $\eta_0$ to construct $\hat{\Omega}(x) = \Omega(x, \hat{\eta})$, and then form estimated optimal instruments as

$$\hat{B}(x) = -\frac{\partial f(x, \hat{\beta})}{\partial \beta} \cdot \Omega(x, \hat{\eta})^{-1} . \tag{3.1.3}$$

For example, $\hat{\eta}$ might be formed from a weighted nonlinear regression of squared least squares residuals on $\Omega(x, \eta)$, or might be the result of several iterations of such an estimator; an extensive discussion of specification of $\Omega(x, \eta)$ and construction of $\hat{\eta}$ is given in Carroll and Ruppert (1988). With estimator $\hat{B}(x)$ of an optimal instrument in hand one could construct a one-step efficient estimator as in equation (2.10). The resulting estimator will be efficient in the class of all estimators that use only the conditional moments restriction of equation (3.1.1), as long as $\Omega(x, \eta)$ and $\hat{\eta}$ are correctly specified, under appropriate regularity conditions. This efficiency will result from the instrument being optimal, and the well-known property that the estimation of the conditional variance will have no effect on the asymptotic distribution of $\hat{\beta}$.

The linearized estimator in equation (2.10) amounts to one iteration of the Gauss–Newton algorithm for minimizing the weighted nonlinear least squares criteria $\Sigma_{i=1}^{n} \Omega(x_i, \hat{\eta})^{-1}\{y_i - f(x_i, \beta)\}^2$. It is well-known that such weighted least squares criteria use only the moment restriction of equation (3.1.1), so that this estimator will be consistent and asymptoticially normal even if $\Omega(x, \eta)$ is misspecified. Other estimators have this property, in particular estimators that are quasi-maximum likelihood estimator for exponential families with conditional mean function $f(x, \beta)$, see McCullagh and Nelder (1983). As is well known, these estimators are asymptotically equivalent to weighted nonlinear least squares estimators, and so are contained in the class of instrumental variables estimators (with instruments equal to the product of $\partial f(x, \beta)/\partial\beta$ and the weight). For asymptotically efficient estimation there is no reason to prefer such estimators to weighted nonlinear least squares (or its linearized counterpart of equation (2.10)), because nonlinear weighted least squares is efficient if $\Omega(x, \eta)$ and $\eta$ are correctly specified.

Nonparametric estimation of the conditional variance provides a way to guard against efficiency loss from misspecifying the form of the conditional variance, as suggested in Carroll (1982) and Robinson (1987). In the next two sections two different approaches to nonparmetric estimation of the optimal instruments will be discussed, and applied to the regression model as an example. Section 6 gives a Monte Carlo comparison of some of the different estimators.

### *3.2. Using second moment information*

An example that illustrates how additional conditional moment restrictions may improve efficiency, at the cost of additional specification sensitivity, is the addition of a moment restriction based on knowing the form of the heteroskedasticity in regression models. Suppose that there is a known function $h(x, \beta, \pi)$ of $\beta$ and some additional parameters $\pi$ such that

$$y = f(x, \beta_0) + \varepsilon\,, \qquad \mathrm{E}[\varepsilon \,|\, x] = 0\,, \qquad \mathrm{E}[\varepsilon^2 \,|\, x] = h(x, \beta_0, \pi_0)\,. \quad (3.2.1)$$

The specified functional form $h(x, \beta, \pi)$ of the heteroskedasticity could be used in construction of a weighted least squares estimator as discussed above. If one had great confidence in the heteroskedasticity specification, then one might want to incorporate this information in estimation. The way to do that in the context of equation (3.2.1), where only the conditional first and second moments are restricted, is to add the conditional variance residual as an additional conditional moment restriction, specifying

$$\rho(z, \theta) = (y - f(x, \beta), \{y - f(x, \beta)\}^2 - h(x, \beta, \pi))'\,. \quad (3.2.2)$$

The additional conditional moment restriction exploited by instrumental variables estimators using this residual will result in estimators that are at least as asymptotically efficient as the heteroskedasticity corrected least squares estimator, and more efficient in some cases. As usual, this efficiency gain comes at a price, that the resulting estimator may be inconsistent if the form of heteroskedasticity is not correctly specified.

The optimal instruments are straightforward to derive for the case, taking the form

$$D(x) = D(x, \theta_0)\,, \qquad D(x, \theta) = \begin{bmatrix} \dfrac{\partial f(x, \beta)}{\partial \beta'} & 0 \\ \dfrac{\partial h(x, \beta, \pi)}{\partial \beta'} & \dfrac{\partial h(x, \beta, \pi)}{\partial \pi'} \end{bmatrix}, \quad (3.2.3)$$

$$\Omega(x) = \mathrm{Var}((\varepsilon, \varepsilon^2)' \,|\, x)\,, \qquad B(x) = D(x)'\Omega(x)^{-1}\,.$$

One question that can be addressed using this formula is whether the additional moment restriction can give more efficient estimators. This question can be answered by comparing the asymptotic variance $(\mathrm{E}[\{\mathrm{E}[\varepsilon^2 \,|\, x]\}^{-1}\{\partial f(x, \beta_0)/\partial\beta\}\{\partial f(x, \beta_0)/\partial\beta\}'])^{-1}$ of the heteroskedasticity corrected least squares estimator with the block of the bound $(\mathrm{E}[D(x)'\Omega(x)^{-1}D(x)])^{-1}$ corresponding to $\beta$. It is easy to see that the two are equal if $\mathrm{E}[\varepsilon^3|x] = 0$ or $\beta$ does not enter the variance function. If either $\mathrm{E}[\varepsilon^3 \,|\, x] \neq 0$ or $\beta$ does enter the variance function it will generally be true that the conditional moment bound is less than the asymptotic variance of optimally weighted least squares. Surprisingly, this potential efficiency gain is present even when $h(x, \pi, \beta)$ and $\Omega(x)$ do not depend on $x$ or $\beta$, but $\mathrm{E}[\varepsilon^3] \neq 0$, as previously noted by MaCurdy (1982).

An estimator that is efficient over a parametric family of conditional third and second moments could be obtained by specifying $\Omega(x, \eta)$. Since the conditional variance is already specified in the model, one only needs to specify the conditional third and fourth moments. If $y$ is continuous then one reasonable, parsimonious specification is to assume that $\mathrm{E}[\varepsilon^3 \mid x] = 0$ and that $\mathrm{Var}(\varepsilon^2 \mid x) = \eta_0 h(x, \beta_0, \pi_0)^2$. This specification is more general than assuming normality, because it allows for a free parameter $\eta_0$, that essentially quantifies the degree of kurtosis. This parameter can be estimated by the sample variance of $\{y_i - f(x_i, \hat{\beta})\}^2 / h(x_i, \hat{\beta}, \hat{\pi})$, where $\hat{\beta}$ and $\hat{\pi}$ are some initial estimators. Then estimated optimal instruments can be constructed as

$$\hat{D}(x) = D(x, \hat{\theta}), \qquad \hat{\Omega}(x) = \begin{bmatrix} h(x, \hat{\beta}, \hat{\pi}) & 0 \\ 0 & \hat{\eta} \cdot h(x, \hat{\beta}, \hat{\pi})^2 \end{bmatrix}, \tag{3.2.4}$$
$$\hat{B}(x) = \hat{D}(x)' \hat{\Omega}(x)^{-1}.$$

The resulting linearized estimator in equation (2.10) is similar to that of Jobson and Fuller (1980), except that it does not depend on normality for it to be efficient relative to weighted least squares. It will be efficient in the class of all estimators that use only the conditional moment restrictions of equation (3.2.1), and hence efficient relative to least squares, as long as $\mathrm{E}[\varepsilon^3 \mid x] = 0$ and $\mathrm{Var}(\varepsilon^2 \mid x) = \eta_0 h(x, \beta_0, \pi_0)^2$ for some $\eta_0$.

One could use the results to follow to construct an estimator that is asymptotically efficient over all (suitably regular) conditional third and fourth moments by replacing $\Omega(x, \hat{\eta})$ by a nonparametric estimator in equation (3.2.4). As a practical matter, though, it may be difficult to estimate higher-order conditional moments, so that large sample sizes may be needed for the asymptotic theory to provide a good approximation. Alternatively, depending on the context, it should be possible to specify other parametric families of conditional third and fourth moments that give good efficiency.

### 3.3. *Box–Cox model*

An interesting and important example is models where the dependent variable has been transformed. As pointed out in Amemiya and Powell (1981), it is possible to estimate these models when the disturbance term is only restricted to have conditional mean zero. The essential idea is that nonlinear functions of the regressors provide information that can be used to identify the transformation parameters. Surprisingly, as shown in Newey (1989b), the resulting estimators can have good efficiency relative to other procedures that use stronger restrictions on the disturbance distribution, such as independence or symmetry, and even do tolerably well relative to maximum likelihood estimators (that are inconsistent if the disturbance distribution is incorrectly specified).

The Box–Cox transformation model with conditional mean restriction is

$$T(y, \lambda_0) = x'\beta_0 + \varepsilon, \qquad \mathrm{E}[\varepsilon \mid x] = 0, \qquad T(y, \lambda) = (y^\lambda - 1)/\lambda. \tag{3.3.1}$$

The conditional moment restriction $\mathrm{E}[\varepsilon \mid x] = 0$ will allow for $\beta$ to be estimated by IV where $\rho(z, \theta) = T(y, \lambda) - x'\beta$, $\theta = (\beta', \lambda)'$, and the instruments consist of suitably chosen functions of $x$, e.g., including $x$ and nonlinear functions of $x$. In this example, unlike the previous two, IV estimation rather than least squares is essential to finding a consistent estimator (the nonlinear least squares estimates are inconsistent).

The conditional moment restriction is not strong enough to allow interpretation of $\beta$ and $\lambda$ as parameters of the conditional distribution of $y$ given $x$. One way to deal with this problem is to add the restriction that $\text{med}(\varepsilon \mid x) = 0$, in which case $\text{med}(y \mid y) = T^{-1}(x'\beta_0, \lambda_0)$. It also might be more natural to just impose $\text{med}(\varepsilon \mid x) = 0$ in estimation, as in Powell (1990), although the asymptotic theory for this case is more complicated, and so for estimation attention here is restricted to the conditional mean case in equation (3.3.1).

The optimal instruments for this model are, for $T_\lambda(y, \lambda) = \partial T(y, \lambda)/\partial\lambda$,

$$\begin{aligned} &D(x) = (-x', \mathrm{E}[T_\lambda(y, \lambda_0) \mid x]) , \qquad \Omega(x) = \mathrm{E}[\varepsilon^2 \mid x] , \\ &B(x) = D(x)'\Omega(x)^{-1} . \end{aligned} \tag{3.3.2}$$

Thus, in this example, unlike the previous two, the functional form of $D(x)$ is unknown, and difficult to specify a priori. One might try and specify a functional form for $D(x)$ by calculating $\mathrm{E}[T_\lambda(y, \lambda_0) \mid x]$ for some distribution and/or values of $\lambda$, and then use this value in the instruments. For example, if $\mathrm{E}[\varepsilon^2 \mid x]$ is constant so there is no heteroskedasticity, and $x$ includes a constant, then as shown in Newey (1989b), at $\lambda_0 = 0$,

$$D(x) = C \cdot (x', (x'\beta_0)^2)$$

for a nonsingular matrix $C$. Thus, a specification of the optimal instruments that will be efficient at $\lambda_0 = 0$, corresponding to a log transformation, under homoskedasticity, is

$$\hat{B}(x) = (x', (x'\hat{\beta})^2)' , \tag{3.3.3}$$

where $\hat{\beta}$ is some initial estimator and $\hat{\Omega}(x)$ is not needed because it is assumed constant and $\rho$ is a scalar. By continuity the resulting estimator should have good efficiency when $\lambda$ is close to zero, as is often found in practice, and when there is little heteroskedasticity, as shown by Newey (1989b) for an example.

An estimator that is efficient over all true $\lambda$ when there is heteroskedasticity could be formed by using as an instrument

$$\hat{B}(x) = (x', \hat{\mathrm{E}}[T_\lambda(y, \hat{\lambda}) \mid x])(\hat{\mathrm{E}}[\{T(y, \hat{\lambda}) - x'\hat{\beta}\}^2 \mid x])^{-1} , \tag{3.3.4}$$

where $\hat{\mathrm{E}}[T_\lambda(y, \hat{\lambda}) \mid x]$ and $\hat{\mathrm{E}}[\{T(y, \hat{\lambda}) - x'\hat{\beta}\}^2 \mid x]$ are nonparametric regression estimators based on the estimated derivative $T_\lambda(y, \hat{\lambda})$ and squared residual $\hat{\mathrm{E}}[\{T(y, \hat{\lambda}) - x'\hat{\beta}\}^2 \mid x]$ respectively.

### *3.4. An endogenous dummy variable model*

For an example of the form of the optimal instruments, consider the model

$$y = \lambda_0 s + f(x, \beta_0) + \varepsilon\,, \quad s \in \{0, 1\}\,, \qquad \mathrm{E}[\varepsilon \mid x] = 0\,. \tag{3.4.1}$$

This model has many important applications in economics to the effect of some event $s$ on an economic variable, such as the effect of college education on income. In many of these applications it is important to allow $s$ to be correlated with $\varepsilon$, because $s$ represents a choice variable of the economic agent that may be affected by variables in $\varepsilon$ that are observed by the empirical worker, such as 'ability'. Correlation between $s$ and $\varepsilon$ is allowed here, because $\mathrm{E}[\varepsilon \mid x] = 0$ does not restrict the joint distribution of $\varepsilon$ and $s$. See Heckman and Robb (1985) for further discussion and motivation.

Optimal instruments for this model take the form, for $\pi(x) = \mathrm{Prob}(s = 1 \mid x) = \mathrm{E}[s \mid x]$,

$$\begin{aligned} D(x) &= \left(\frac{\partial f(x, \beta_0)}{\partial \beta'}, \pi(x)\right), \qquad \Omega(x) = \mathrm{E}[\varepsilon^2 \mid x]\,, \\ B(x) &= D(x)'\Omega(x)^{-1}\,. \end{aligned} \tag{3.4.2}$$

As in the last example, there is a component of $D(x)$ that does not have known functional form. Here this component is $\pi(x) = \mathrm{Prob}(d = 1 \mid x)$. An estimator that is efficient over a parametric family of possible $\pi(x)$ functions can be obtained by specifying a functional form for $\pi(x)$, estimating its parameters by binary choice maximum likelihood with $s$ as the dependent variable, and then substituting the predicted probability for $\pi(x)$. For example, assuming that $\mathrm{Prob}(s = 1 \mid x) = \Phi(x'\eta_0)$ for some $\eta_0$ and the standard normal CDF $\Phi(\cdot)$, one could form an estimate $\hat{\eta}$ from probit with $s$ as the dependent variable, and then construct $\hat{\pi}(x) = \Phi(x'\hat{\eta})$. Also one might assume homoskedasticity in constructing an instrument estimator. Then substituting into the formula gives

$$\hat{B}(x) = \left(\frac{\partial f(x, \beta_0)}{\partial \beta'}, \Phi(x'\hat{\eta})\right)'\,, \tag{3.4.3}$$

where $\hat{\beta}$ is some initial estimator and $\hat{\Omega}(x)$ is not needed because it is assumed constant and $\rho$ is a scalar. Estimators that are efficient over all (suitably regular) conditional probabilities for $s$, for unknown heteroskedasticity, could be constructed from nonparametric estimators of the optimal instruments.

## 4. Nearest neighbor estimation of the optimal instruments

The first approach to construction of nonparametric estimates of the optimal instruments is to nonparametrically estimate the conditional expectations that make up the optimal instruments and plug the estimates into the formula for the optimal instruments. Because of its technical convenience, the nearest neighbor estimation method is considered here.

A nearest neighbor estimator of a conditional expectation is formed by averaging over the values of the dependent variable for observations where the conditioning variable is closest to its evaluation value. To describe how this estimator is applied to estimation of the optimal instruments, let $\hat{\sigma}_{xl}$ be some estimate of the scale of the $l$th component $x_i$ of $x$, satisfying the conditions given in Stone (1977), such as the sample standard deviation of $x_{il}$, and define $\|x_i - x_j\|_n = \{\Sigma_{l=1}^{r} (x_{il} - x_{jl})^2/(\hat{\sigma}_l^2)\}^{1/2}$, where $r$ is the dimension of $x$, $(i, j = 1, \ldots, n)$. For a given integer $K \leqslant n$ consider constants $\omega_{kK}$, satisfying

$$\omega_{kK} \geqslant 0\,, \quad 1 \leqslant k \leqslant K\,; \qquad \omega_{kK} = 0\,, \quad k > K\,; \qquad \sum_{k=1}^{K} \omega_{kK} = 1\,. \tag{4.1}$$

For given $i$ let $W_{ii} = 0$, and rank the observations $j \neq i$ according to the distance $\|x_i - x_j\|_n$. If there are no ties among the distances assign to observation with $j$-th smallest value $\|x_i - x_j\|_n$ the weight $W_{ij} = \omega_{jK}$. If there are ties, follow the same procedure, but with equal weight given to observations with the same value of $\|x_i - x_j\|_n$. A nearest neighbor estimator of the conditional covariance $\Omega(x_i)$ at $x_i$ can then be formed as

$$\hat{\Omega}(x_i) = \sum_{j=1}^{n} W_{ij}\rho(z_j, \hat{\theta})\rho(z_j, \hat{\theta})'\,. \tag{4.2}$$

This is a nearest neighbor estimator that excludes the own observation that is analogous to Robinson's (1987) conditional variance estimator. Examples of weights include uniform ones where $\omega_{kK} = 1/K$, $k \leqslant K$, and other smoother versions, e.g., as discussed in Robinson (1987). The theory here will utilize a uniform boundedness restriction on the weights, that there is a constant $C$ such that

$$\omega_{kK} \leqslant \frac{C}{K}\,, \quad 1 \leqslant k \leqslant K\,. \tag{4.3}$$

This restriction is satisfied by the uniform weights, as well as the other weights discussed in Robinson (1987).

For many models at least some components of $D(x)$ will have known functional form, and depend only on $x$, and it seems wise to take advantage of this knowledge in construction of the estimator. One general formulation that allows for that knowledge is to specify an estimator that is a sum of parametric and nearest neighbor components. Let $D(x, \eta)$ be some prespecified function of $x$ and nuisance parameters $\eta$ with the same dimensions as $D(x)$, e.g., with components equal to those of $D(x)$ that are known and zero otherwise. For an estimator $\hat{\eta}$ of $\eta_0$, let

$$\hat{D}(x_i) = D(x_i, \hat{\eta}) + \sum_{j=1}^{n} W_{ij}\left[\frac{\partial\rho(z_j, \hat{\theta})}{\partial\theta} - D(x_j, \hat{\eta})\right]. \tag{4.4}$$

This estimator is fully nonparametric, in that it will be consistent for all $D(x, \eta)$ and (suitably regular) estimators $\hat{\eta}$, and components of the nearest neighbor term will be zero when the corresponding components of $D(x, \hat{\eta})$ and $\partial\rho(z, \theta)/\partial\theta$ are equal.

Another possible use of the $D(x, \eta)$ function is for 'detrending' in the sense of Stone (1977), where $D(x, \eta)$ is some simple function that is included to 'wash out' some of the dependence of $D(x)$ on $x$. For example, $D(x, \eta)$ might be linear in $x$, in which case the nearest neighbor term would be estimating the deviation from a linear function. This set-up does not allow for linearity, or other known functional form, to be imposed on the conditional expectation, e.g., as is often done in the linear simultaneous equations models (e.g., see Hausman, 1984). To get such an estimator one would have to make sure that the components of $D(x, \eta)$ had the right functional form and delete corresponding components from the nearest neighbor terms. Although this possibility is not explicitly allowed for in the paper, the conclusion of Theorem 1 below would still hold in this case.

With estimators of the conditional expectations in hand, one can combine them to form an estimator of the optimal instruments in the natural way, as $\hat{B}(x_i) = \hat{D}(x_i)'\hat{\Omega}(x_i)^{-1}$. An efficient estimator of $\theta_0$ can then be constructed in the way described in Section 2. An important purpose for such an estimator is to make asymptotically efficient inferences about the population parameter value $\theta_0$, such as to construct confidence intervals or tests statistics. For this purpose it is useful to have a consistent asymptotic variance estimator. One natural estimator can be obtained by replacing the conditional expectations in the variance bound by corresponding nearest neighbor estimates and the expectation by a sample average. The result is

$$\hat{\Lambda} = \left(\sum_{i=1}^{n} \hat{D}(x_i)'\hat{\Omega}(x_i)\hat{D}(x_i)/n\right)^{-1}. \tag{4.5}$$

Conditions for consistency of this estimator are given in Theorem 1 below.

An important problem of nearest neighbor estimation is the choice of number of nearest neighbors $K$. Because the focus here is on efficient estimation of $\theta$, it seems desirable to base that choice on a criteria that focuses on the properties of $\hat{\theta}$. Although a theoretical investigation of how to formulate such a criteria is beyond the scope of this paper, some heuristic ideas can be used to suggest a choice of $K$ that may lead to good properties for the estimator of $\hat{\theta}$. Suppose that one wants to choose $K$ so as to minimize 'remainder terms' in the asymptotic theory that arise form estimation of the optimal instruments. From an expansion similar to that in equation (2.4) it is easy to see that there will be remainder terms of both the Jacobian matrix and the cross product of instruments with the residuals. Suppose that we can ignore the Jacobian term, which may be possible under some circumstances (e.g., if it is root-$n$ consistently estimated and the size of the other remainder terms are greater than $1/\sqrt{n}$). This reasoning suggests a criteria based on an estimate of the magnitude of $\Sigma_{i=1}^{n}\{\hat{B}(x_i)-B(x_i)\}\rho(z_i,\theta_0)/\sqrt{n}$, a difference of estimated and true moment functions (or 'scores'). For the nearest neighbor estimator, there are two terms in this difference, one each for the estimation of $D(x)$ and $\Omega(x)$. Assuming higher-order terms can be ignored, this remainder can be linearized to give

$$\begin{aligned}&\sum_{i=1}^{n}\hat{R}(x_i)\rho(z_i,\theta_0)/\sqrt{n}\,,\\&\hat{R}(x_i)=\{\hat{D}(x_i)-D(x_i)+B(x_i)[\hat{\Omega}(x_i)-\Omega(x_i)]\}\Omega(x_i)^{-1}\,.\end{aligned}$$

This is a vector remainder term, so to quantify its magnitude for minimization one has to choose a distance metric. Let $Q$ be a positive definite matrix. Assuming that the 'hat' can be ignored in taking the conditional expectation over $z_i$, given the $x_i$ observations (e.g., as would be the case if $\hat{R}(x)$ were constructed from another sample), leads to the criteria $\operatorname{tr}[Q\,\Sigma_{i=1}^{n}\hat{R}(x_i)\Omega(x_i)\hat{R}(x_i)']$. This criteria cannot be computed, but because $\hat{R}(x_i)$ depends on 'cross-validation' nearest neighbor estimates, where the $i$-th observation is not used in estimation of $\hat{D}(x_i)$ or $\hat{\Omega}(x_i)$, the usual crossvalidation reasoning suggests that one can estimate this criteria up to a term that does not depend on $K$ by replacing the conditional expectations by their estimators and their estimators by the dependent variables, as in

$$\begin{aligned}&C\hat{V}(K)=\operatorname{tr}\left[Q\sum_{i=1}^{n}\tilde{R}(x_i)\hat{\Omega}(x_i)\tilde{R}(x_i)'\right],\\&\tilde{R}(x_i)=\left\{\frac{\partial\rho(z_i,\hat{\theta})}{\partial\theta}-\hat{D}(x_i)+\hat{B}(x_i)[\rho(z_i,\hat{\theta})\rho(z_i,\hat{\theta})'-\hat{\Omega}(x_i)]\right\}\hat{\Omega}(x_i)^{-1}\,.\end{aligned}\tag{4.6}$$

Thus, a cross-validation criteria for the choice of $K$, that takes some account for how the estimator of parameter of interest depends on $K$, is to choose $k$ to minimize $C\hat{V}(K)$.

To illustrate the estimator and the cross-validation criteria it is helpful to consider an example. A simple, important example is the linear regression model. In this case $\hat{\Omega}(x_i)=\hat{\sigma}_i^2=\Sigma_{j=1}^{n}W_{ij}\hat{\varepsilon}_j^2$, where $\hat{\varepsilon}_j=y_j-x_j'\hat{\beta}$ are residuals, and the linearized estimator of equation (2.10) and the asymptotic variance estimator of equation (4.5) are

$$\begin{aligned}&\tilde{\beta}=\left(\sum_{i=1}^{n}\hat{\sigma}(x_i)^{-2}x_ix_i'\right)^{-1}\sum_{i=1}^{n}\hat{\sigma}(x_i)^{-2}x_iy_i\,,\\&\hat{\Lambda}=\left(n^{-1}\sum_{i=1}^{n}\hat{\sigma}(x_i)^{-2}x_1x_1'\right)^{-1}\,.\end{aligned}\tag{4.7}$$

These estimators were suggested and analyzed in Robinson (1987). The cross-validation criteria obtained by setting $\partial\rho(z_i,\hat{\theta})/\partial\theta-\hat{D}(x_i)$ to zero, as appropriate here where the form of $D(x)$ is known, and by choosing $Q = (\Sigma_{j=1}^{n} x_i x_i')^{-1}$, is

$$C\hat{V}(K) = \sum_{i=1}^{n} (x_i' Q x_i)[(\hat{\varepsilon}_i/\hat{\sigma}_i)^2 - 1] , \qquad Q = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} . \tag{4.8}$$

The choice of $Q$ here makes this criteria invariant to nonsingular linear transformations of the regressors. In the Monte Carlo example of Section 6 this cross-validation criteria for choice of $K$, choosing $K$ to minimize $C\hat{V}(K)$ and then using this $K$ to form the estimator in equation (4.7), leads to $\hat{\beta}$ with good properties.

To prove asymptotic efficiency it is helpful to impose some regularity conditions. The first condition is essentially a standard one involving compactness, existence of certain moments, and nonsingularity conditions.

Assumption 4.1. $\theta_0$ is an element of the interior of $\Theta$, which is compact, there is a neighborhood $\mathcal{N}$ of $\theta_0$ and $d(z)$ such that with probability one $\rho(z, \theta)$ is continuous on $\Theta$, continuously differentiable on $\mathcal{N}$, $\sup_{\theta \in \Theta} \|\rho(z, \theta)\| \leq d(z)$, $\sup_{\theta \in \mathcal{N}} \|\partial \rho(z, \theta)/\partial \theta\| \leq d(z)$, $E[d(z)^2] < \infty$. Also, $\mathrm{E}[B(x)\Omega(x)B(x)']$ exists and is nonsingular.

The next Assumption is important for consistency of the fully iterated estimator given in equation (2.9).

Assumption 4.2. $\mathrm{E}[B(x)B(x')']$ exists and is nonsingular, and there is a unique solution to $\mathrm{E}[B(x)\rho(z, \theta)] = 0$ on $\Theta$ at $\theta = \theta_0$.

For the linearized estimator of equation (2.10) is important to have root-$n$ consistency of the initial estimator. The following condition helps guarantee this.

Assumption 4.3. $\hat{\theta}$ satisfies equation (2.3) for $\hat{P} = (\Sigma_{i=1}^{n} A(x_i)A(x_i)')^{-1}$, there is a unique solution to $\mathrm{E}[A(x)\rho(z, \theta)] = 0$ on $\Theta$ at $\theta = \theta_0$, $\mathrm{E}[\|A(x)\|^2] < \infty$, $\mathrm{E}[A(x)D(x)]$ and $\mathrm{E}[A(x)A(x)']$ are nonsingular, and $\mathrm{E}[A(x)\Omega(x)A(x)']$ is finite.

Finally, some additional smoothness and moment existence conditions are useful.

Assumption 4.4. There is a neighborhood $\mathcal{N}$ of $\theta_0$ and $d(z)$ such that for all $\theta \in \mathcal{N}$,

$$\|\rho(z, \theta)\|^4 < d(z) , \qquad \|\partial \rho(z, \theta)/\partial \theta\|^4 < d(z) ,$$
$$\|\partial^2 \rho(z, \theta)/\partial \theta \, \partial \theta\|^2 < d(z) ,$$

$$\|\partial^2\rho(z,\bar{\theta})/\partial\theta\,\partial\theta - \partial^2\rho(z,\theta)/\partial\theta\,\partial\theta\| \leq d(z)\|\bar{\theta}-\theta\|\,,$$

and $\mathrm{E}[d(z)^2] < \infty$.

For the special case of the heteroskedastic linear model, this condition is stronger than that imposed in Robinson (1987), in that it requires existence of eighth moments of the disturbance and of the regressors. This more restrictive condition allows for the presence of $\hat{D}(x)$ and leads to a simpler proof than given by Robinson (1987).

Assumption 4.5. $\mathrm{E}[\|D(x,\eta_0)\|^8] < \infty$, and there is $d(z)$ and a neighborhood $\mathcal{N}$ of $\eta_0$ such that for all $\eta \in \mathcal{N}$,

$$\|\partial D(x,\eta)/\partial\eta\|^2 \leq d(z)\,,$$
$$\|\partial D(x,\eta)/\partial\eta - \partial D(x,\eta_0)/\partial\eta\| \leq d(z)\|\eta-\eta_0\|\,,$$

$\mathrm{E}[d(z)^2] < \infty$. Also $\sqrt{n}(\hat{\eta}-\eta_0) = \mathrm{O}_\mathrm{p}(1)$.

Theorem 1. *If Assumptions 4.1–4.5 are satisfied and $K = K(n)$ such that $K(n)/n \to 0$ and $K(n)^2/n \to \infty$, then for $\bar{\theta} = \hat{\theta}$ or $\bar{\theta} = \tilde{\theta}$,*

$$\sqrt{n}(\bar{\theta}-\theta_0) \xrightarrow{d} \mathrm{N}(0,\Lambda)\,, \qquad \left[\sum_{i=1}^{n} \hat{D}(x_i)'\hat{\Omega}(x_i)^{-1}\hat{D}(x_i)/n\right]^{-1} \xrightarrow{p} \Lambda\,.$$

## 5. Series approximation of the optimal instruments

Another approach to estimation of the optimal instruments is by series approximation, where the estimator is formed as a linear combination of known functions. A potential advantage of this approach is that linear combinations of smooth functions can approximate a smooth function very well with only a few terms, while nearest neighbor estimators do not seem to exploit such smoothness.

As previously noted, it is useful to allow some components of the optimal instruments to have known functional form. A way to form a series approximation that allows for some known components is to let $D(x,\eta)'$ be a matrix with $q$ rows, $\{a_{kK}(x)\}_{k=1}^{K}$ be matrices of approximating functions with number of rows equal to the number of columns of $D(x,\eta)'$, and estimate the optimal instruments by

$$\hat{B}(x) = D(x,\hat{\eta})'\left[\sum_{k=1}^{K} \hat{\gamma}_k a_{kK}(x)\right], \tag{5.1}$$

where $\hat{\gamma}_{kK}$ are estimated scalars described below, and $\hat{\eta}$ is an estimator of $\eta$. For example in the linear model $D(x,\hat{\eta})'$ could be specified as $x$, so that $[\Sigma_{k=1}^{K} \hat{\gamma}_k a_{kK}(x)]$ corresponds to an estimator of $1/\sigma^2(x)$.

The two keys to asymptotic efficiency are the choice of $a_{kK}(x)$ and of

$\hat{\gamma}_1, \dots, \hat{\gamma}_K$. The functions $a_{kK}(x)$ should be specified in such a way that $D(x, \eta_0)'[\Sigma_{k=1}^K \gamma_k a_{kK}(x)]$ can approximate $B(x)$ for some choice of linear combination coefficient, $\gamma_1, \dots, \gamma_K$. The form of this approximation will be specific to the form of $B(x)$. Given functions $a_{kK}(x)$ with the property, it is possible to form $\hat{\gamma}_1, \dots, \hat{\gamma}_K$ as estimators of a minimum mean-square error approximation that leads to asymptotic efficiency. For any $\gamma = (\gamma_1, \dots, \gamma_K)$ let $A_k(x) = D(x, \eta_0)' a_{kK}(x)$, $\rho = \rho(z, \theta_0)$, $u_k = A_k(x)\rho$, and $U = [A_1(x)\rho, \dots, A_K(x)\rho]$. Also, let

$$\begin{aligned} M_k &= \mathrm{E}\left[A_k(x) \frac{\partial \rho(z, \theta_0)}{\partial \theta}\right] = \mathrm{E}[A_k(x) D(x)] = \mathrm{E}[A_k(x) \Omega(x) B(x)'] \\ &= \mathrm{E}[u_k \rho' b(x)'] , \end{aligned} \tag{5.2}$$

where the second and fourth equalities follow by iterated expectations. For any positive definite matrix $Q$ consider

$$\begin{aligned} \bar{\gamma} &= \underset{\gamma \in \mathbb{R}^K}{\operatorname{argmin}} \operatorname{tr}\left(Q \cdot \mathrm{E}\left[\left\{B(x) - \sum_{k=1}^K \gamma_k A_k(x)\right\} \Omega(x) \right.\right. \\ &\qquad \left.\left. \times \left\{B(x) - \sum_{k=1}^K \gamma_k A_k(x)\right\}'\right]\right) \\ &= \underset{\gamma \in \mathbb{R}^K}{\operatorname{argmin}} \mathrm{E}[\{B(x)\rho - U\gamma\}' Q \{B(x)\rho - U\gamma\}] \\ &= (\mathrm{E}[U'QU])^{-1} \mathrm{E}[U'QB(x)\rho] \\ &= (\mathrm{E}[U'QU])^{-1} [\operatorname{tr}(Q\mathrm{E}[B(x)\rho u_1']), \dots, \operatorname{tr}(Q\mathrm{E}[B(x)\rho u_K'])]' \\ &= (\mathrm{E}[U'QU])^{-1} [\operatorname{tr}(QM_1), \dots, \operatorname{tr}(QM_K)]' , \end{aligned} \tag{5.3}$$

where tr(·) denotes the trace of a square matrix, the first equality holds by iterated expectations, and the last equality follows by equation (5.2).

From the first two equalities it is apparent that these coefficients have two interpretations as mean-square approximations, one as a weighted (by $\Omega(x)$) mean-square approximation of the optimal instruments by $\Sigma_{k=1}^K \gamma_k A_k(x)$ and the other as a mean-square approximation of $B(x)\rho$ by $\Sigma_{k=1}^K \gamma_k A_k(x)\rho$. The matrix $Q$ is present to account for the fact that these approximations are scalar linear combinations rather than matrix linear combinations. For some choices of $a_{kK}(x)$ (where $\{a_{kK}(x)\}$ is zero except for one element), the approximation may actually be a matrix linear combination. In this case the coefficients $\gamma$ can also be interpreted as minimizing the asymptotic variance; see Newey (1989a) for details.

It follows by equations (5.2) and (5.3) that if a linear combination of $A_k(x) = (D(x, \eta_0)' a_{kK}(x)$ can approximate $B(x)$ arbitrarily closely as $K$ grows, in the weighted mean-square norm following the first equality of equation (5.3), then an IV estimator with $A(x) = \Sigma_{k=1}^K \bar{\gamma}_k A_k(x)$ will be approximately efficient for large $K$. Equation (5.2) implies equation (2.8), which in turn implies that the asymptotic variance of an IV estimator will be close to the bound if $A(x)\rho$

is close in mean-square to $B(x)\rho$. Then by the second inequality of equation (5.3) and $Q$ positive definite, $A(x)\rho$ will be close in mean-square to $B(x)\rho$ when the weighted norm approximation error for the optimal instruments is small. Thus, under the spanning condition that linear combination of $D(x,\eta_0)'a_{kK}(x)$ can approximate $B(x)$ arbitrarily closely as $k$ grows, the IV estimator with instruments $\Sigma_{k=1}^{K} \bar{\gamma}_k A_k(x)$ will be approximately efficient. This spanning condition is explicitly imposed in Assumptions 5.4 and 5.5 below.

The IV estimator based on $\bar{\gamma}_k$ is not feasible because $\bar{\gamma}_k$ are unknown, but the last equality in equation (5.3) can be used to construct an estimator. Let $\hat{Q}$ denote an estimator of $Q$, $\hat{u}_{ki} = D(x_i, \hat{\eta})'a_{kK}(x_i)\rho(z_i, \hat{\theta})$, $\hat{U}_i = [\hat{u}_{1i}, \ldots, \hat{u}_{Ki}]$, and

$$\hat{M}_K = n^{-1} \sum_{i=1}^{n} D(x_i, \hat{\eta})'a_{kK}(x_i) \frac{\partial \rho(z_i, \hat{\theta})}{\partial \theta} .$$

Plugging in estimates and sample averages to the last equality gives

$$\hat{\gamma} = \left( \sum_{i=1}^{n} \hat{U}_i' Q \hat{U}_i / n \right)^{-1} [\mathrm{tr}(\hat{Q}\hat{M}_1'), \ldots, \mathrm{tr}(\hat{Q}\hat{M}_K')]' . \tag{5.4}$$

A feasible IV estimator can then be constructed by using this $\hat{\gamma}$ in the instruments of equation (5.1). Under appropriate regularity conditions, $\hat{\gamma}$ will be consistent for $\bar{\gamma}$, so by estimation of instruments having no effect on the asymptotic variance of IV, the estimator will be approximately efficient for large $K$. The asymptotic efficiency result below is even stronger, giving a growth rate for $K$ as a function of sample size to achieve asymptotic efficiency.

An efficient estimator can be constructed as in equation (2.10) using the estimator of the optimal instruments in equation (5.1). Also, an estimator of the asymptotic variance of $\tilde{\theta}$ is

$$\hat{\Lambda} = \left( \sum_{k=1}^{K} \hat{\gamma}_k \hat{M}_k \right)^{-1} \left[ \sum_{i=1}^{n} (\hat{U}_i' \hat{\gamma})(\hat{U}_i' \hat{\gamma})' \Big/ n \right] \left( \sum_{k=1}^{K} \hat{\gamma}_k \hat{M}_k \right)^{-1\prime} .$$

This estimator is a consistent estimator for the asymptotic variance of $\hat{\theta}$, even for $K$ fixed. An alternative estimator that would be consistent as $K$ goes to infinity is $[\Sigma_{i=1}^{n} (\hat{U}_i' \hat{\gamma})(\hat{U}_1' \hat{\gamma})'/n]^{-1}$.

An important problem for this estimator is the choice of $K$. One way to choose $K$, that is suggested a Newey (1989a), is to minimize a cross-validation estimator of the mean square error of product of the differences of the true and estimated instruments with the residual. This choice can be motivated by similar reasoning gives for the choice of nearest neighbors described in the last section. In particular, by the second equality of equation (5.3), this criterion is based on the difference of instruments multiplied by the residual vector, and so may be related to the magnitude of remainder terms. To describe this cross-validation method, let $\hat{\gamma}_{-i}$ be the estimator in equation (5.4) except that the $i$-th observation has been deleted when calculating the sample averages. For

$\hat{\rho}_i = \rho(z_i, \hat{\theta})$ and $\hat{A}_{ki} = D(x_i, \hat{\eta})'a_{kK}(x_i)$ consider

$$\sum_{i=1}^{n} [B(x_i)\hat{\rho}_i - \hat{U}_i'\hat{\gamma}_{-1}]'\hat{Q}[B(x_i)\hat{\rho}_i - \hat{U}_i'\hat{\gamma}_{-i}]$$

$$= \sum_{i=1}^{n} [\hat{\rho}_i'B(x_i)'\hat{Q}B(x_i)\hat{\rho}_i - 2\hat{\rho}_i'B(x_i)'\hat{Q}\hat{U}_i\hat{\gamma}_{-i} + \hat{\gamma}'_{-i}\hat{U}_i'\hat{Q}\hat{U}_i\hat{\gamma}_{-i}] .$$

This criterion cannot be computed, because $B(x_i)$ is unknown. To obtain one that can be computed, the first term can be dropped because it does not depend on $K$ and so is not needed to minimize over $K$. Also, equation (5.3) implies $\mathrm{E}[\rho(z_i, \theta_0)'B(x_i)'QU_i] = [\mathrm{tr}(QM_1), \ldots, \mathrm{tr}(QM_K)]$ allowing replacement of the second term by $\mathrm{tr}(\hat{Q}\, \Sigma_{k=1}^{K} \hat{\gamma}_{-i,k}\hat{M}_{ki})$, for $\hat{M}_{ki} = D(x_i, \hat{\eta})'a_{kK}(x_i)\, \partial\rho(z_i, \hat{\theta})/\partial\theta$. These changes give

$$C\hat{V}(K) = -2\mathrm{tr}\left( \hat{Q} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\gamma}_{-i,k}\hat{M}_{ki} \right) + \mathrm{tr}\left( \hat{Q} \sum_{i=1}^{n} \hat{U}_i\hat{\gamma}_{-i}(\hat{U}_i\hat{\gamma}_{-i})' \right) . \quad (5.5)$$

To illustrate the estimator and the cross-validation criteria it is helpful to again consider the linear regression model as an example. In this case $B(x) = \sigma(x)^{-2}x$, and there are several ways that one might choose $D(x, \eta)$ and $a_{kK}(x)$ to approximate this function. One way is based on a multivariate approximation of the entire vector $\sigma(x)^{-2}x$. Let $D(x, \eta)$ be an identity matrix and each $a_{kK}(x)$ be a vector with the same dimension as $x$, such that there is an integer $J$ and a vector $p_J(x)' = (p_{1J}(x), \ldots, p_{JJ}(x))'$ with

$$[a_{1K}(x), \ldots, a_{KK}(x)] = p_J(x)' \otimes I ,$$

for an identity matrix $I$ with the same row dimension as $x$. From example, $p_{jJ}(x)$ might be a power series. In this case the matrix $\hat{Q}$ factors out and the estimator of the optimal instruments is given by

$$\hat{B}(x) = \hat{\Gamma}p_J(x) , \qquad \hat{\Gamma} = \left(\sum_{i=1}^{n} p_J(x_i)x_i'\right)\left(\sum_{i=1}^{n} p_J(x_i)p_J(x_i)'\hat{\varepsilon}_i^2\right)^{-1} . \quad (5.6)$$

The resulting IV estimator $\hat{\beta} = (\Sigma_{i=1}^{n} \hat{B}(x_i)x_i')^{-1} \Sigma_{i=1}^{n} \hat{B}(x_i)y_i$ is equal to the estimator of Cragg (1983). A cross-validation criterion for the choice of $J$ for this estimator can be constructed as described above. Let $\hat{\Gamma}_{-i}$ be the coefficients in equation (5.6) with the $i$-th observation deleted from the sums, and let $p_i = p_J(x_i)$. Then specializing the formula in equation (5.5) to this example gives

$$C\hat{V}(J) = -2\mathrm{tr}\left(\sum_{i=1}^{n} \hat{\Gamma}_{-i}p_ix_i'\right) + \mathrm{tr}\left[\sum_{i=1}^{n} \hat{\varepsilon}_i^2(\hat{\Gamma}_{-i}p_i)(\hat{\Gamma}_{-i}p_i)'\right] . \quad (5.7)$$

In the Monte Carlo example of Section 6 this cross-validation criteria for choice of $J$ leads to $\hat{\beta}$ with good properties.

It is possible to construct alternative estimators that are more parsimonious, in that they use fewer functions to achieve the same degree of approximation. The idea here is that the optimal instruments consist of a product of a known vector with a single unknown function $1/\sigma^2(x)$, so that it should be possible to construct an estimator where the only function being approximated is $1/\sigma^2(x)$. The way to do this is to let $D(x, \eta)' = x$ and $a_{kK}(x)$ be scalars. In particular, consider letting $k = j$, $K = J$, and $a_{kK}(x) = p_{jJ}(x)$ for the $p_{jJ}(x)$ in the previous estimator. Also, let $\hat{Q} = (\Sigma_{i=1}^n x_i x_1'/n)^{-1}$, which will make the estimator equivariant with respect to nonsingular linear transformations of $x$. Then for $s_i = x_i'\hat{Q}x_i$, the estimator of the optimal instruments is

$$\hat{B}(x) = x \cdot (p_J(x)'\hat{\gamma}) , \qquad \hat{\gamma} = \left(\sum_{i=1}^n s_i p_i p_i' \hat{\varepsilon}_i^2\right)^{-1} \sum_{i=1}^n s_i p_i . \tag{5.8}$$

In this example, it follows form equation (5.3) that $\hat{\gamma}$ will be consistent for

$$\bar{\gamma} = \arg\min \mathrm{E}[\sigma^2(x) x'Qx\{1/\sigma^2(x) - p_J(x)'\gamma\}^2] \quad \text{for } Q = (\mathrm{E}[x_i x_i'])^{-1} ,$$

so that $p_J(x)'\hat{\gamma}$ can be interpreted as an estimator of $1/\sigma^2(x)$. In comparison with the previous estimator, the approximation is more parsimonious, because only a linear combination of $J$ functions is required for a $j$-th order approximation, rather than the linear combination of $J \cdot \text{dimension}(x)$ functions. To choose $J$ by cross-validation, let $\hat{\gamma}_{-1}$ be as given in equation (5.8), except that the $i$-th observation is deleted from the sum. The criterion of equation (5.5) is

$$C\hat{V}(J) = -2 \sum_{i=1}^n s_i(p_i'\hat{\gamma}_{-i}) + \sum_{i=1}^n s_i(p_i'\hat{\gamma}_{-i})^2 \hat{\varepsilon}_i^2 . \tag{5.9}$$

The Monte Carlo example of Section 6 compares the performance of this estimator with the previous one.

To prove asymptotic efficiency of the estimators described in the section, it is useful to impose some regularity conditions. The first condition requires certain smoothness conditions and existence of certain moments.

Assumption 5.1. $Q$ is positive definite, there is $\nu > 2$, $\delta > 0$, such that $\mathrm{E}[\|\rho(z, \theta_0)\|^\nu] < \infty$, $\mathrm{E}[\|D(x, \eta_0)\|^{[2\nu/(\nu-2)]+\delta}] < \infty$. $\sqrt{n}(\hat{Q} - Q) = \mathrm{O}_p(1)$, $\sqrt{n}(\hat{\eta} - \eta_0) = \mathrm{O}_p(1)$, $D(x, \eta)$ is continuously differentiable in $\eta$, there is $d_1(z), d_2(z), d_1'(z), d_2'(z)$ and neighborhoods of $\eta_0$ and $\theta_0$ respectively such that

$$\|\partial\rho(z, \theta)/\partial\theta\| \leqslant d_1(z) ,$$

$$\|\partial\rho(z, \theta)/\partial\theta - \partial\rho(z, \theta_0)/\partial\theta\| \leqslant d_\rho(z)\|\theta - \theta_0\| ,$$

$$\sum_{j=1}^J \left\|\frac{\partial D(x, \eta)}{\partial\eta_j}\right\| \leqslant d_1'(z) ,$$

and

$$\sum_{j=1}^{J}\left\|\frac{\partial D(x,\eta)}{\partial\eta_j}-\frac{\partial D(x,\eta_0)}{\partial\eta_j}\right\|\leq d_2'(z)\|\theta-\theta_0\|\,,$$
$$\mathrm{E}[d_1(z)^{\nu}]<\infty\,,\qquad \mathrm{E}[d_2(z)^{2\nu/(\nu+2)}]<\infty\,,$$
$$\mathrm{E}[d_1'(z)^{2\nu/(\nu-2)}]<\infty\,,\qquad \mathrm{E}[d_2'(z)^{\nu/(\nu-1)}]<\infty\,.$$

The moment conditions required here are weaker than in Section 4. Only slightly higher than two moments of the residual are required to exist, unlike the eight moments condition of Section 4. The next assumption imposes some conditions on the approximating functions.

Assumption 5.2. $a_{kK}(x)$ is bounded, uniformly in $k, K$, and the smallest eigenvalue of $\mathrm{E}[U'QU]$ is bounded below by $\Delta^{-1}K^{-\Delta K^{1/r}}$ for some $\Delta>0$.

The boundedness of the approximating functions is not restrictive, because the choice of functions is controlled, and boundedness can be relaxed without affecting any of the following results. The eigenvalue condition is not primitive, and is essential for the results to follow. Primitive conditions for this hypothesis can be obtained for power series. Let

$$p_j(x)=\prod_{l=1}^{r}\tau_l(x_l)^{\lambda_l(j)}\,,\qquad p^J(x)=(p_1(x),\ldots,p_J(x))'\,,$$

where $\lambda_l(j)$ are nonnegative integers.

Assumption 5.3. $a_{kK}(x)=p_{j(k)}(x)C_k$ for a constant matrix $C_k$, $\tau_j(x)$ are bounded, the distribution of $(\tau_1(x),\ldots,\tau_r(x))'$ has a continuously distributed component with density bounded away from zero on an open set $\mathscr{X}$, the smallest eigenvalue of $\{D(x,\eta_0)QD(x,\eta_0)'\}\otimes\Omega(x)$ is bounded away from zero on $\mathscr{X}$, $(\lambda_1(j),\ldots,\lambda_r(j))_{j=1}^{\infty}$ consists of all distinct vectors of nonnegative integers, $\Sigma_{l=1}^{r}\lambda_l(j)$ is increasing in $j$, for every $K$ there is $J(K)$ and a constant matrix $L_K$ such that

$$L_K[\{p^{J(K)}(x)p^{J(K)}(x)\}\otimes\{D(x,\eta_0)QD(x,\eta_0)'\}$$
$$\otimes\{\rho(z,\theta_0)\rho(z,\theta_0)\}]L_K'=U'QU\,,$$

the smallest eigenvalue of $L_KL_K'$ is bounded away from zero uniformly in $K$, and $J(K)\leq CK$ for a constant $C$.

This condition, in particular the assumption that there is some continuity in the distribution of $\tau(x)$, will imply Assumption 5.2. It is also possible to allow for some discrete regressors with finite number of support points by including a full set of interactions of these variables with the power series, but for brevity this possibility is excluded here.

The next condition is the spanning condition.

ASSUMPTION 5.4. There exist $\tilde{\gamma}_{1K}, \ldots, \tilde{\gamma}_{KK}$ such that $\mathrm{E}[\|D(x,\eta_0)\,\Sigma_{k=1}^{K}\tilde{\gamma}_{kK}a_{kK}(x) - B(x)\|^2\|\Omega(x)\|] \to 0$, and either $\hat{K} \in \mathscr{K}_n$ with probability approaching one and the number of elements of $\mathscr{K}_n$ is bounded, or

$$\sum_{K=1}^{\infty}\left\{\mathrm{E}\left[\left\|D(x,\eta_0)\sum_{k=1}^{K}\tilde{\gamma}_{kK}a_{kK}(x) - B(x)\right\|^2\|\Omega(x)\|\right]\right\}^{1/2} < \infty\,.$$

This condition allows for an approximation rate that is useful in proving efficiency when $k$ is data-based. Primitive conditions for this hypothesis with power series are given in the following condition.

ASSUMPTION 5.5. $\hat{K} \in \hat{\mathscr{K}}_n$ with probability approaching one and the number of elements of $\mathscr{K}_n$ is bounded, $\tau(x)$ is one-to-one and there is a conformable matrix $R(x)$ such that $B(x) = D(x,\eta_0)'R(x)$, $\mathrm{E}[\|D(x,\eta_0)\|^2\|\Omega(x)\|\,\|R(x)\|^2] < \infty$. Also, $\mathrm{vec}(R(x)) = (r_1(x), \ldots, r_v(x))'$ such that for any $J$ and $\gamma_{11}, \ldots, \gamma_{1J}$, $\gamma_{21}, \ldots, \gamma_{vJ}$ there is $K$ and $\gamma_1, \ldots, \gamma_K$ such that

$$\left\|R(x) - \sum_{k=1}^{K}\gamma_k a_k(x)\right\| \leqslant \sum_{l=1}^{v}\left\|r_l(x) - \sum_{j=1}^{J}\gamma_{lj}p_j(x)\right\|$$

THEOREM 2. *If Assumptions* 4.1, 4.3, 5.1, *either* 5.2 *and* 5.4, *or* 5.3 *and* 5.5. *are satisfied, and* $K = \hat{K}$, $\hat{K} \xrightarrow{p} \infty$, *and there is* $\bar{K}(n)$ *such* $\hat{K} \leqslant \bar{K}(n)$ *with probability approaching one and* $\bar{K}(n)^{1/r}\ln[\bar{K}(n)]/\ln(n) \to 0$, *then*

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, \Lambda)\,, \qquad \hat{\Lambda} \xrightarrow{p} \Lambda\,.$$

## 6. Sampling experiment for the heteroskedastic linear model

A small Monte Carlo study can suggest how the estimators might perform in practice. The model and design considered was the heteroskedastic linear model with normally distributed disturbance, lognormal regressor, and quadratic variance function considered by Cragg (1983), taking the form

$$\begin{aligned} &y_i = \beta_{10} + \beta_{20}x_i + \varepsilon_i\,, \qquad \varepsilon_i/\sigma_i \sim \mathrm{N}(0,1)\,, \qquad \sigma_i^2 = 0.1 + 0.2x_i + 0.3x_i^2\,, \\ &\ln(x_i) \sim \mathrm{N}(0,1)\,, \qquad x_i \text{ and } \varepsilon_i \text{ independent}\,. \end{aligned} \tag{6.1}$$

The reported results are invariant to the values of $\beta_0$ and to multiplication of the disturbance by a fixed constant. Two sample sizes were considered, 50, and 200. Computations were carried out on a microcomputer using GAUSS, with 1000 replications of both the $x_i$ and $y_i$ samples.

Table 1 reports results for sample size 50 and Table 2 for sample size 200. Each table gives the ratios of standard deviation (STDEV), median absolute error (MAE), and coverage probabilities for the asymptotic 95% confidence

Table 1
Ratio of variance, median absolute error, and nominal 95 percent coverage probability to that of Aitken estimator, and distribution of cross-validated 'Bandwidth', $n = 50$

| | STDEV | MAE | COV PROB | Distribution of $K$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| OLS | 3.189 | 2.778 | 0.782 | | | | | | | |
| | 2.110 | 2.055 | 0.751 | | | | | | | |
| FGLS | 1.311 | 1.178 | 0.715 | | | | | | | |
| | 1.460 | 1.312 | 0.646 | | | | | | | |
| | | | | $k =$ | 6 | 9 | 12 | 15 | 18 | 24 |
| NN | 1.515 | 1.344 | 0.878 | | | 1 | | | | |
| | 1.436 | 1.294 | 0.733 | | | | | | | |
| | 1.523 | 1.333 | 0.875 | | 0.32 | 0.28 | 0.18 | 0.10 | 0.05 | 0.07 |
| | 1.442 | 1.321 | 0.742 | | | | | | | |
| | | | | $J =$ | 4 | 6 | 8 | 10 | 12 | |
| MM-CRAGG | 1.500 | 1.367 | 0.904 | | | | 1 | | | |
| | 1.491 | 1.376 | 0.752 | | | | | | | |
| | 1.795 | 1.622 | 0.894 | | 0.50 | 0.26 | 0.14 | 0.05 | 0.05 | |
| | 1.595 | 1.550 | 0.781 | | | | | | | |
| | | | | $J =$ | 2 | 3 | 4 | 5 | 6 | |
| MM-COMBINE | 1.515 | 1.278 | 0.866 | | | | 1 | | | |
| | 1.509 | 1.440 | 0.700 | | | | | | | |
| | 1.583 | 1.356 | 0.870 | | 0.28 | 0.44 | 0.12 | 0.11 | 0.05 | |
| | 1.528 | 1.404 | 0.748 | | | | | | | |

interval (COV PROB) of several estimators to the corresponding results for the generalized least squares (Aitken) estimator. The estimators for which results are reported are ordinary least squares (OLS), feasible generalized least squares (FGLS), nearest neighbor estimators (NN), and two varieties of series estimators (MM). The FGLS estimator was calculated by taking the predicted values from a regression of the squared residual on linear and quadratic terms in $x_i$, dividing by the estimated variance of the disturbance, censoring the result below at 0.04, and then using the inverse of the resulting quantity as a weight in weighted least squares results. Different truncation points were tried, in results not reported here, but they did not seem to make much difference.

Several nearest neighbor estimators were calculated, one each for the grid of $K$ values given in the table, and one where for each replication $K$ was chosen to minimize the cross-validation criterion described in Section 4. The tables only report results for the $K$ where the estimator has smallest variance and for the cross-validated $K$, with the distribution of $K$ across replications reported on the right.

Two varieties of series estimators were considered. Both used approximating functions $p_j(x) = \tau(x)^j$, $\tau(x) = x/(1 + |x|)$, where $x$ was normalized in each replication to have sample mean zero and variance one. The first type was the Cragg (1983) estimator that used these approximating functions (MM-CRAGG), and the second was the more parsimonious estimator of equation (5.8). For each type several estimators were calculated, one each for the grid

Table 2
Ratio of variance, median absolute error, and nominal 95 percent coverage probability to that of Aitken estimator, and distribution of cross-validated 'Bandwidth', $n = 200$

| | VAR | MAE | COV PROB | Distribution of $K$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| OLS | 4.892 | 3.933 | 0.856 | | | | | | | |
| | 2.949 | 2.731 | 0.836 | | | | | | | |
| FGLS | 1.369 | 1.156 | 0.717 | | | | | | | |
| | 1.321 | 1.231 | 0.733 | | | | | | | |
| | | | | $k =$ | 8 | 12 | 16 | 20 | 24 | 28 |
| NN | 1.462 | 1.200 | 0.871 | | 1 | | | | | |
| | 1.449 | 1.346 | 0.809 | | | | | | | |
| | 1.462 | 1.178 | 0.904 | | 0.28 | 0.27 | 0.18 | 0.13 | 0.08 | 0.06 |
| | 1.436 | 1.308 | 0.851 | | | | | | | |
| | | | | $J =$ | 6 | 8 | 10 | 12 | 14 | |
| MM-CRAGG | 1.154 | 1.089 | 0.001 | | 1 | | | | | |
| | 1.115 | 1.096 | 0.977 | | | | | | | |
| | 1.231 | 1.089 | 0.081 | | 0.62 | 0.19 | 0.10 | 0.04 | 0.05 | |
| | 1.205 | 1.173 | 0.938 | | | | | | | |
| | | | | $J =$ | 3 | 4 | 5 | 6 | 7 | |
| MM-COMBINE | 1.154 | 1.067 | 0.974 | | 1 | | | | | |
| | 1.192 | 1.250 | 0.948 | | | | | | | |
| | 1.246 | 1.089 | 0.953 | | 0.64 | 0.18 | 0.10 | 0.04 | 0.05 | |
| | 1.269 | 1.212 | 0.921 | | | | | | | |

of $k$ values listed on the right of the table, and the one where for each replication $K$ was chosen to minimize the cross-validation criterion. The tables report results for the estimator with fixed $K$ that had the smallest variance and for the cross-validation estimator.

Before discussing the tables it is useful to note that all the estimators here are symmetrically distributed around the true parameter values, so that there is no need to be concerned about central tendency. Also, unlike Cragg (1983), the results are not conditional on a particular set of regressors, which here were simulated for each replication.

The results in the tables can be summarized by saying that for sample size 50, the FGLS estimator does somewhat better than the NN estimator (cross-validated), which in turn does slightly better than the MM estimators (cross-validated), but for sample size 200 the MM estimator does best, the FGLS estimator next best, and the NN estimator least best. Also, the estimators that use the cross-validation criteria suggested earlier do about as well as the ones with best choice of nearest neighbor or number of series terms, both in terms of dispersion and accuracy of asymptotic confidence intervals, particularly for the larger sample sizes. Indeed, in a number of cases the cross-validated estimators perform better in some respects than the fixed $K$ versions. Also, the Cragg type series estimator does worse than the more parsimonious estimator for the smaller sample size, but better for the larger sample size. One would expect to find a bigger advantage for the more parsimonious estimator suggested here in more realistic cases where there are more regressors, and hence more linear combination coefficients in the Cragg estimator relative to the more parsimonious one.

## Appendix

In order to state and prove some useful lemmas it is necessary to introduce some additional notation. Throughout $C$ will denote a generic positive constant that can take on different values in different uses and $\Sigma_i = \Sigma_{i=1}^n$. The first lemma gives a general set of sufficient conditions for asymptotic efficiency of nonlinear instrumental variables estimators with estimated optimal instruments.

Lemma A.1. *Suppose that Assumption* 3.1 *is satisfied,*

$$\sum_i \|\hat{B}(x_i) - B(x_i)\|^2/n \xrightarrow{p} 0 ,$$

*and*

$$\sum_i \{\hat{B}(x_i) - B(x_i)\}\rho(z_i, \theta_0)\sqrt{n} \xrightarrow{p} 0 .$$

*Then if Assumption* 3.2 *is satisfied,* $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, \Lambda)$ *for* $\hat{\theta}$ *from equation* (2.9), *and if Assumption* 3.3 *is satisfied, and* $\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, \Lambda)$.

Proof. Let $\hat{B}_i = \hat{B}(x_i)$, $B_i = B(x_i)$, $\hat{g}(\theta) = \sum_i \hat{B}_i\rho(z_i, \theta)/n$ and $\tilde{g}(\theta) = \Sigma_i B_i\rho(z_i, \theta)/n$. By CS and M,

$$\sup_{\theta\in\Theta} \|\hat{g}(\theta) - \tilde{g}(\theta)\| \leq \Big(\sum_i \|\hat{B}_i - B_i\|^2/n\Big)^{1/2}\Big(\sum_i d(z_i)^2\Big/n\Big)^{1/2}$$
$$= \mathrm{o_p}(1)\mathrm{O_p}(1) = \mathrm{o_p}(1) ,$$

while by the usual uniform law of large numbers, $\sup_{\theta\in\Theta}\|\tilde{g}(\theta) - \mathrm{E}[\tilde{g}(\theta)]\| \xrightarrow{p} 0$, so by T, $\sup_{\theta\in\Theta}\|\hat{g}(\theta) - \mathrm{E}[\tilde{g}(\theta)]\| \xrightarrow{p} 0$. It follows similarly that for $\hat{g}_\theta(\theta) = \partial\hat{g}(\theta)/\partial\theta$ and $\tilde{g}_\theta(\theta) = \partial\tilde{g}(\theta)/\partial\theta$, $\sup_{\theta\in\mathcal{N}}\|\hat{g}_\theta(\theta) - \mathrm{E}[\tilde{g}_\theta(\theta)]\| \xrightarrow{p} 0$. Also, by the Sluztky and Lindbergh–Levy theorems, for $\rho_i = \rho(z_i, \theta)$,

$$\sum_i \hat{B}_i\rho_i/\sqrt{n} = \sum_i (\hat{B}_i - B_i)\rho_i/\sqrt{n} + \sum_i B_i\rho_i/\sqrt{n}$$
$$= \mathrm{o_p}(1) + \sum_i B_i\rho_i/\sqrt{n} \xrightarrow{d} \mathrm{N}(0, \Lambda^{-1}) .$$

The estimator of equation (2.9) is an IV estimator with $\hat{P}=(\Sigma_i \hat{B}_i\hat{B}_i/n)^{-1}\xrightarrow{p}P=(\mathrm{E}[B_iB_i'])^{-1}$ under Assumption 3.2, so $\hat{g}(\theta)'\hat{P}\hat{g}(\theta)$ converges uniformly in probability to $g(\theta)'Pg(\theta)$, $g(\theta)'Pg(\theta)$ is continuous, and has a unique minimum (of zero) at $\theta=\theta_0$, and hence is consistent by the standard Wald argument. The first conclusion then follows by a mean-value expansion like that of equation (2.4). For the second conclusion, asymptotic normality of the initial estimator follows by a similar argument to that just given. Then by a mean-value expansion

$$\begin{aligned}\sqrt{n}(\tilde{\theta}-\theta_0)&=\{I-[\hat{g}_\theta(\hat{\theta})]^{-1}\hat{g}_\theta(\bar{\theta})\}\sqrt{n}(\hat{\theta}-\theta_0)-[\hat{g}_\theta(\hat{\theta})]^{-1}\sqrt{n}\hat{g}(\theta_0)\\&=\mathrm{o_p}(1)\mathrm{O_p}(1)-[\hat{g}_\theta(\hat{\theta})]^{-1}\sqrt{n}\hat{g}(\theta_0)\xrightarrow{d}\mathrm{N}(0,\Lambda).\qquad\square\end{aligned}\tag{A.1}$$

Some other Lemmas are useful for proving Theorem 1. Let $h(z,\theta)$ be a function of $z$ and a parameter vector $\theta$ and let $\hat{\theta}$ be a consistent estimate of some value $\theta_0$. Let

$$\begin{aligned}g_i&=\mathrm{E}[h(z_i,\theta_0)\,|\,x_i]\,,\qquad &\bar{g}_i&=\sum_{j=1}^{n}W_{ij}\mathrm{E}[h(z_j,\theta_0)\,|\,x_j]\,,\\\tilde{g}_i&=\sum_{j=1}^{n}W_{ij}h(z_j,\theta_0)\,,\qquad &\tilde{g}_i&=\sum_{j=1}^{n}W_{ij}h(z_j,\hat{\theta})\,.\end{aligned}\tag{A.2}$$

The following assumption concerning $h(z,\theta)$ and $\theta$ will be maintained.

Assumption A.1. (i) $\hat{\theta}-\theta_0=\mathrm{O}_p(1\sqrt{n})$. (ii) For $p>2$, $\mathrm{E}[|h(z_i,\theta_0)|^p]$ is finite. (iii) there is $M(z)$ with $\mathrm{E}[M(z_i)^2]$ finite such that for $\theta$ close enough to $\theta_0$, $|h(z_i,\theta)-h(z_i,\theta_0)|\leqslant M(z_i)\|\theta-\theta_0\|$. (iv) $k/\sqrt{n}\rightarrow$, $k/n\rightarrow 0$.

Lemma A.2 (Stone, 1977, Proposition 1). $\lim_{n\to\infty}\mathrm{E}[|\bar{g}_i-g_i|^p]=0$.

The proofs of the following three lemmas are nearly identical to the proofs of Lemmas 8, 9, and 5, respectively, of Robinson (1987), and so will be omitted.

Lemma A.3. $\{\mathrm{E}[|\tilde{g}_i-\bar{g}_i|^p]\}^{1/p}=\mathrm{O}(k^{-1/2})$.

Lemma A.4. $\max_{i\leqslant n}|\tilde{g}_i-\bar{g}_i|=\mathrm{O_p}(n^{1/p}k^{-1/2})$

Lemma A.5. $\max_{i\leqslant n}|\hat{g}_i-\tilde{g}_i|=\mathrm{O_p}(k^{-1/2})$.

Let $Z_n=(z_1,\ldots,z_n)$. The following two lemmas are proven in Newey (1990).

LEMMA A.6. *Suppose that* (i) $\rho_n(z, X_n)$ *is a function such that* $\mathrm{E}[\rho_n(z_i, X_n) \mid x_i, Z_{-i}] = 0$ *and* $\mathrm{E}[|\rho_n(z_i, X_n)|^{2p/(p-2)}] = \mathrm{O}(1)$; (ii) $h(z_i, \theta)$ *is continuously differentiable with derivative* $h_\theta(z_i, \theta)$ *in a convex neighborhood* $N$ *of* $\theta_0$; (iii) *there are random variables* $M_\theta(z_i)$ *and* $M_{\theta\theta}(z_i)$ *satisfying* $\|h_\theta(z_i, \theta)\| \leqslant M_\theta(z_i) - h_\theta$ *and*

$$\|h_\theta(z_i, \theta) - h_\theta(z_i, \theta_0)\| \leqslant M_{\theta\theta}(z_i)\|\theta - \theta_0\| \quad \textit{for } \theta \textit{ in } N ,$$

*and* $E[M_\theta(z_i)^p]$ *and* $E[M_{\theta\theta}(z_i)^{2p/(p+2)}]$ *finite. Then*

$$\sum_i (\hat{g}_i - g_i)\rho_n(z_i, X_n)/\sqrt{n} = \mathrm{o}_\mathrm{p}(1) . \tag{A.3}$$

The following lemma is an 'asymptotic trimming' result, that allows us to ignore the 'denominator problem' associated with the nearest neighbor estimator $\hat{\Omega}(x_i)$. Let $\hat{\Omega}_i$, $\tilde{\Omega}_i$, $\bar{\Omega}_i$, $\Omega_i$, $\hat{D}_i$, $\tilde{D}_i$, $\bar{D}_i$, $D_i$ be the estimators corresponding to equation (A.1), where $h(z, \theta)$ is an element of $\rho(z, \theta)\rho(z, \theta)'$ and $\partial\rho(z, \theta)/\partial\theta$, respectively.

LEMMA A.7. *If the hypotheses of Theorem* 3.1 *are satisfied then there is a constant* $C$ *such that* $\|\Omega_i^{-1}\| < C$, $\|\bar{\Omega}_i^{-1}\| < C$ *and the indicator* $\hat{1}_n = 1(\max_{i \leqslant n} \max\{\|\hat{\Omega}_i^{-1}\|, \|\tilde{\Omega}_i^{-1}\|\} < C)$ *is equal to one with probability approaching one. Furthermore, for any sequence of random variables* $Y_n$ *and constants* $b_n$, *if* $Y_n\hat{1}_n = \mathrm{O}_\mathrm{p}(b_n)$ *then* $Y_n = \mathrm{O}_\mathrm{p}(b_n)$.

PROOF. Let $\lambda(A)$ denote the smallest eigenvalue of a symmetric matrix A. Then $\lambda(\Omega_i) > C$ for all $i$ by $\lambda(\Omega(x))$ bounded away from zero. Also, by $W_{ij}$ nonnegative and $\Sigma_{j=1}^n W_{ij} = 1$ the extremal characterization of $\lambda(\cdot)$ gives $\lambda(\bar{\Omega}_i) = \lambda(\Sigma_{j=1}^n W_{ij}\Omega_j) \geqslant \Sigma_{j=1}^n W_{ij}\lambda(\Omega_j) > C$, giving the first conclusion. As is well known, $|\lambda(\tilde{A}) - \lambda(A)| \leqslant C\|\tilde{A} - A\|$ where the constant $C$ here depends only on the dimension of $A$. By Lemma A.4, with $h(z, \theta)$ and element of $\rho(z, \theta)\rho(z, \theta)'$ and $p = \nu/2 > 4$,

$$\begin{aligned} \max_{i \leqslant n} |\lambda(\tilde{\Omega}_i) - \lambda(\bar{\Omega}_i)| &\leqslant C \max_{i \leqslant n} \|\tilde{\Omega}_i - \bar{\Omega}_i\| = \mathrm{O}_\mathrm{p}(n^{1/p}k^{-1/2}) \\ &= \mathrm{O}_\mathrm{p}(n^{1/4}k^{-1/2}) = \mathrm{o}_\mathrm{p}(1) . \end{aligned}$$

Similarly, by Lemma A.5, $\max_{i \leqslant n}|\lambda(\hat{\Omega}_i) - \lambda(\tilde{\Omega}_i)| \xrightarrow{p} 0$, so that $\mathrm{Prob}(\hat{1}_n = 1) \to 1$ by T and specification of small enough $C$. The final conclusion follows by noting that $(\hat{1}_n - 1)Y_n$ is equal to zero with probability approaching one, and hence $(1 - \hat{1}_n)Y_n = \mathrm{O}_\mathrm{p}(b_n)$ for *any* $b_n$, giving $Y_n = \hat{1}_nY_n + (1 - \hat{1}_n)Y_n = \mathrm{O}_\mathrm{p}(b_n)$. □

PROOF OF THEOREM 1. The result is first proved for the case where $D(x, \eta) = 0$, by verifying the hypotheses of Lemma A.6. For $\hat{1}_n$ in Lemma A.7, by CS,

$$\begin{aligned}\hat{1}_n\|\hat{\Omega}_i^{-1} - \bar{\Omega}_i^{-1}\| &= \hat{1}_n\|\hat{\Omega}_i^{-1}(\bar{\Omega}_i - \hat{\Omega}_i)\bar{\Omega}_i^{-1}\| \\ &\leqslant \hat{1}_n\|\hat{\Omega}_i^{-1}\|\ \|\hat{\Omega}_i - \bar{\Omega}_i\|\ \|\bar{\Omega}_i^{-1}\| \\ &\leqslant C\hat{1}_n\|\hat{\Omega}_i - \bar{\Omega}_i\| \leqslant C\|\hat{\Omega}_i - \bar{\Omega}_i\| .\end{aligned}$$

Therefore, by Lemmas A.4 and A.5,

$$\begin{aligned}&\hat{1}_n\Big(\sum_i \|\hat{\Omega}_i^{-1} - \bar{\Omega}_i^{-1}\|^4/n\Big)^{1/4} \\ &\quad\leqslant C\Big(\sum_i \|\hat{\Omega}_i - \bar{\Omega}_i\|^4/n\Big)^{1/4} \\ &\quad\leqslant C\max_{i\leqslant n}\|\hat{\Omega}_i - \bar{\Omega}_i\| + C\Big(\sum_i \|\tilde{\Omega}_i - \bar{\Omega}_i\|^4/n\Big)^{1/4} \\ &\quad= \mathrm{o_p}(k^{-1/2}) + \mathrm{O_p}(\{\mathrm{E}[\|\tilde{\Omega}_i - \bar{\Omega}_i\|^{\nu/2}]\}^{2/\nu}) = \mathrm{o_p}(k^{-1/2}) . \qquad \text{(A.4)}\end{aligned}$$

Similarly, $(\Sigma_i \|\hat{D}_i - \bar{D}_i\|^2/n)^{1/2} = \mathrm{o_p}(k^{-1/2})$. Then by CS, H, T, and $\hat{1}_n = 1$ with probability approaching one,

$$\begin{aligned}&\Big\|\sum_i (\hat{D}_i - \bar{D}_i)'(\hat{\Omega}_i^{-1} - \bar{\Omega}_i^{-1})\rho_i/\sqrt{n}\Big\| \\ &\quad= \hat{1}_n\Big\|\sum_i (\hat{D}_i - \bar{D}_i)'(\hat{\Omega}_i^{-1} - \bar{\Omega}_i^{-1})\rho_i/\sqrt{n}\Big\| + \mathrm{o_p}(1) \\ &\quad\leqslant \sqrt{n}\Big(\sum_i \|\hat{D}_i - \bar{D}_i\|^2/n\Big)^{1/2}\cdot\hat{1}_n\Big(\sum_i \|\hat{\Omega}_i^{-1} - \bar{\Omega}_i^{-1}\|^4/n\Big)^{1/4} \\ &\qquad\times\Big(\sum_i \|\rho_i\|^8/n\Big)^{1/8} + \mathrm{o_p}(1) = \mathrm{O_p}(\sqrt{n}/k) = \mathrm{o_p}(1) . \qquad \text{(A.5)}\end{aligned}$$

Similarly, by $\mathrm{E}[\|D_i\|^4] < \infty$, implying $\mathrm{E}[\|\bar{D}_i\|^4] = \mathrm{o}(1)$,

$$\begin{aligned}&\Big\|\sum_i \bar{D}_i'(\hat{\Omega}_i^{-1} - \bar{\Omega}_i^{-1})\rho_i/\sqrt{n} - \sum_i \bar{D}_i'\bar{\Omega}_i^{-1}(\bar{\Omega}_i - \hat{\Omega}_i)\bar{\Omega}_i^{-1}\rho_i/\sqrt{n}\Big\| \\ &\quad\leqslant \hat{1}_n\sum_i \|\bar{D}_i\|\ \|\bar{\Omega}_i^{-1}\|^2\|\hat{\Omega}_i - \bar{\Omega}_i\|^2\|\hat{\Omega}_i\|\ \|\rho_i/\sqrt{n} + \mathrm{o_p}(1) \\ &\quad\leqslant C\sum_i \|\bar{D}_i\|\ \|\hat{\Omega}_i - \bar{\Omega}_i\|^2\|\rho_i\|/\sqrt{n} \\ &\quad\leqslant C\Big(\sum_i \|\bar{D}_i\|^4/n\Big)^{1/4}\Big(\sum_i \|\hat{\Omega}_i - \bar{\Omega}_i\|^4/n\Big)^{1/2}\Big(\sum_i \|\rho_i\|^8/n\Big)^{1/8} \\ &\quad= \mathrm{O_p}(\sqrt{n}/k) = \mathrm{o_p}(1) .\end{aligned}$$

By Lemma A.6, with $\rho_n(z_i, X_n)$ equal to an element of $\mathrm{vec}([\bar{D}_i'\bar{\Omega}_i^{-1}]\otimes[\bar{\Omega}_i^{-1}])$,

$p = 4$ and $h(z, \theta)$ equal to an element of $\rho(z, \theta)\rho(z, \theta)'$,

$$\sum_i \bar{D}_i' \bar{\Omega}_i^{-1}(\hat{\Omega}_i - \Omega_i)\bar{\Omega}_i^{-1}\rho_i/\sqrt{n} \xrightarrow{p} 0 ,$$

and for $h(z, \theta)$ equal to an element of $\mathrm{E}[\rho(z, \theta)\rho(z, \theta)' \mid x]$,

$$\sum_i \bar{D}_i' \bar{\Omega}_i^{-1}(\bar{\Omega}_i - \Omega_i)\bar{\Omega}_i^{-1}\rho_i/\sqrt{n} \xrightarrow{p} 0 .$$

Then by the triangle inequality,

$$\sum_i \bar{D}_i'(\hat{\Omega}_i^{-1} - \Omega_i^{-1})\rho_i/\sqrt{n} \xrightarrow{p} 0 . \tag{A.6}$$

Also, by Lemma A.6 with $\rho_n(z_i, X_n)$ equal to an element of $\Omega_i^{-1}\rho_i$, $p = 4$ and $h(z, \theta)$ equal to an element of $\partial\rho(z, \theta)/\partial\theta$,

$$\sum_i (\hat{D}_i - D_i)' \bar{\Omega}_i^{-1}\rho_i/\sqrt{n} \xrightarrow{p} 0 . \tag{A.7}$$

Also, by Lemma A.2 and $\mathrm{E}[(\bar{D}_i - D_i)'(\bar{\Omega}_i^{-1} - \Omega_i^{-1})\rho_i] = 0$ and $\mathrm{E}[\rho_i\rho_j \mid X_n] = 0$ for $i \neq j$,

$$\begin{aligned}
&\mathrm{E}\left[\sum_i (\bar{D}_i - D_i)'(\bar{\Omega}_i^{-1} - \Omega_i^{-1})\rho/\sqrt{n}\right\|^2\Bigg] \\
&\quad \leqslant \mathrm{E}[\|\bar{D}_i - D_i\|^2 \|(\bar{\Omega}_i^{-1} - \Omega_i^{-1})\Omega_i(\bar{\Omega}_i^{-1} - \Omega_i^{-1})\|] \leqslant C , \\
&\mathrm{E}[\|\bar{D}_i - D_i\|^2 \|\bar{\Omega}_i - \Omega_i\|^2] \leqslant (\mathrm{E}[\|\bar{D}_i - D_i\|^4])^{1/2}(\mathrm{E}[\|\bar{\Omega}_i \\
&\qquad\qquad - \Omega_i\|^4])^{1/2} \to 0 ,
\end{aligned}$$

$$\sum_i (\bar{D}_i - D_i)'(\bar{\Omega}_i^{-1} - \Omega_i^{-1})\rho_i/\sqrt{n} \xrightarrow{p} 0 . \tag{A.8}$$

It then follows by equations (A.5)–(A.8) and the triangle inequality that $\Sigma_{i=1}^n (\hat{B}_i - B_i)\rho_i\sqrt{n} \xrightarrow{p} 0$. Furthermore, by T, CS, and Lemma A.2, it follows similarly to equation (A.4) that

$$\begin{aligned}
\sum_{i=1}^n \|\hat{B}_i - B_i\|^2/n &\leqslant \Big(\sum_i \|\hat{D}_i - D_i\|^4/n\Big)^{1/4}\Big(\sum_i \|\hat{\Omega}_i^{-1}\|^4/n\Big)^{1/4} \\
&\quad + \Big(\sum_i \|D_i\|^4/n\Big)^{1/4}\Big(\sum_i \|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|^4/n\Big)^{1/4} \xrightarrow{p} 0 .
\end{aligned}$$

Similarly, $\Sigma_{i=1}^n \|\hat{B}_i\hat{D}_i - B_iD_i\|/n \xrightarrow{p} 0$, and by the law of large numbers, $\Sigma_{i=1}^n B_iD_i/n \xrightarrow{p} \Lambda^{-1}$, so the final conclusion follows by T.

For the case where $D(x, \eta)$ is not equal zero, by Assumption 4.5, the conclusions previously given hold with $\partial\rho(z, \theta)/\partial\theta - D(x, \eta)$ replacing $\partial\rho(z, \theta)/\partial\theta$ throughout. For simplicity, assume $\eta$ is a scalar. Therefore, by

Assumption 4.5, a mean-value expansion, and H,

$$\begin{aligned}
&\left\|\sum_i \{D(x_i, \hat\eta) - D(x_i, \eta_0)\}(\hat\Omega_i^{-1} - \Omega_i^{-1}\}\rho_i/\sqrt{n}\right\| \\
&\quad \leqslant \sqrt{n}\|\hat\eta - \eta_0\| \sum_i \sup_{\eta \in \mathcal{N}} \left\|\frac{\partial D(x_i, \eta)}{\partial \eta}\right\| \|\hat\Omega_i^{-1} - \Omega_i^{-1}\| \, \|\rho_i\|/n \\
&\quad \leqslant \mathrm{O_p}(1)\Big(\sum_i d(z_i)^2/n\Big)^{1/4}\Big(\sum_i \|\hat\Omega_i^{-1} - \Omega_i^{-1}\|^4/n\Big)^{1/4}\Big(\sum_i \|\rho_i\|^8/n\Big)^{1/8} \\
&\quad = \mathrm{O_p}(1)\mathrm{o_p}(1)\mathrm{O_p}(1) = \mathrm{o_p}(1) .
\end{aligned}$$

Also, it follows similarly to equation (A.6) that

$$\sum_i D(x_i, \eta_0)(\hat\Omega_i^{-1} - \Omega_i^{-1})\rho_i/\sqrt{n} \xrightarrow{p} 0 .$$

Also by the usual uniform law of large numbers,

$$\sum_i \left\{\frac{\partial D(x_i, \bar\eta)}{\partial \eta}\right\}\rho_i/n \xrightarrow{p} \mathrm{E}[\{\partial D(x_i, \eta_0)/\partial\eta\}\rho_i] = 0 ,$$

so that by a mean value expansion,

$$\begin{aligned}
&\left\|\sum_i \{D(x_i, \hat\eta) - D(x_i, \eta_0)\}\Omega_i^{-1}\rho_i/\sqrt{n}\right\| \\
&\quad \leqslant \sqrt{n}\|\hat\eta - \eta_0\| \left\|\sum_i \left\{\frac{\partial D(x_i, \bar\eta)}{\partial \eta}\right\}\rho_i/n\right\| \xrightarrow{p} 0 .
\end{aligned}$$

Also, the previous proof gives

$$\sum_i \{(\hat D_i - D(x_i, \hat\eta))\hat\Omega_i^{-1} - (D_i - D(x_i, \eta_0))\Omega_i^{-1}\}\rho_i/\sqrt{n} \xrightarrow{p} 0 .$$

Therefore, by T,

$$\begin{aligned}
&\left\|\sum_i (\hat B_i - B_i)\rho_i/\sqrt{n}\right\| = \left\|\sum_i \{D(x_i, \hat\eta)\hat\Omega_i^{-1} - D(x_i, \eta_0)\Omega_i^{-1}\}\rho_i/\sqrt{n}\right\| + \mathrm{o_p}(1) \\
&\quad = \mathrm{o_p}(1) . \qquad \text{(A.9)}
\end{aligned}$$

Similar reasoning also gives $\Sigma_i \|\hat B_i - B_i\|^2/n \xrightarrow{p} 0$, so the first conclusion follows by Lemma A.1, and also gives, $\Sigma_i \|\hat B_i \hat D_i - D_i D_i\|/n \xrightarrow{p} 0$, so the second conclusion follows as above. □

Two lemmas are useful for the proof of Theorem 2. Let $m_k(z, \theta, \eta) =$

$D(x, \eta)' a_{kK}(x) \rho(z, \theta)$, so that

$$U = [m_1(z, \theta_0, \eta_0), \ldots, m_K(z, \theta_0, \eta_0)]\,,$$
$$\hat{U}_i = [m_1(z_i, \hat{\theta}, \hat{\eta}), \ldots, m_K(z_i, \hat{\theta}, \hat{\eta})]\,,$$
$$\hat{M}_k = \sum_{i=1}^{n} \partial m_k(z_i, \hat{\theta}, \hat{\eta}) / \partial \theta\,.$$

Lemma A.8. *If Assumption* 5.3 *is satisfied then Assumption* 5.2 *is satisfied.*

Proof. It follows by the extremal characterization of the smallest eigenvalue and Lemma A1 of Newey (1988) that

$$\begin{aligned} &\lambda(\mathrm{E}[U'QU]) \\ &\quad \geqslant \lambda(L_K L_K')\lambda(\mathrm{E}[\{p^{J(K)}(x)\rho^{J(K)}(x)\} \otimes \{D(x, \eta_0)QD(x, \eta_0)'\} \otimes \Omega(x)]) \\ &\quad \geqslant CJ(K)^{-CJ(K)^{1/r}} \end{aligned} \tag{A.10}$$

The conclusion then follows by $J(K) \leqslant CK$. □

Lemma A.9. *If Assumptions* 5.3 *and* 5.5 *are satisfied then Assumption* 5.4 *is satisfied.*

Proof. By Gallant (1980), for each $l$ their exist $\gamma_{l1}^{J}, \ldots, \gamma_{lJ}^{j}$ such that

$$\mathrm{E}\left[\|D(x, \eta_0)\|^2 \|\Omega(x)\| \left\{r_l(x) - \sum_{j=1}^{J} \gamma_{lj}^{j} p_j(x)\right\}^2\right] \to 0\,,$$

so the conclusion follows from Assumption 5.5. □

Proof of theorem 2. The proof proceeds by verifying the hypotheses of Lemma 5.1 of Newey (1989) (L5 henceforth). It follows by standard arguments that $\sqrt{n}\|\hat{\theta} - \theta_0\| = \mathrm{O_p}(1)$ for the initial estimator $\hat{\theta}$. Consider nonrandom $K = K(n)$ such that $K \to \infty$, and let the $\nu(K)$ of L5 be $K^{1/r}$. By hypothesis $Q$ is positive semidefinite. Also, by H and Assumption 5.1,

$$\begin{aligned} &\|\mathrm{E}[\partial m(z, \theta_0, \eta_0)/\partial\theta]\| \\ &\qquad \leqslant K \sup_{k,x} \|a_{kK}(x)\| \mathrm{E}\left[\|D(x, \eta_0)\| \left\|\frac{\partial \rho(z, \theta_0)}{\partial \theta}\right\|\right] \leqslant CK \leqslant C\nu(K)^{C\nu(K)} \end{aligned}$$

and there is $\varepsilon > 0$ small enough that

$$\mathrm{E}[\|\mathrm{Vec}(U)\|^{2+\varepsilon}] \leqslant K^C \mathrm{E}\left[\|D(x, \eta_0)\|^{2+\varepsilon} \left\|\frac{\partial \rho(z, \theta_0)}{\partial \theta}\right\|^{2+\varepsilon}\right] \leqslant C\nu(K)^{C\nu(K)}\,.$$

In addition, it follows by Assumption 5.2 or Assumption 5.3 and Lemma A.8 that $\lambda(\mathrm{E}[U'QU]) \geqslant C\nu(K)^{-C\nu(K)}$, so that hypothesis (i) of L5 is satisfied.

Next, by Assumption 5.4, $\|\hat{Q} - Q\| = \mathrm{O_p}(n^{-1/2})$. Also, by Assumption 5.1

$\mathrm{E}[\|\partial m_k(z,\theta_0,\eta_0)/\partial\theta\|^{1+\varepsilon}] < C$ for some $\varepsilon > 0$, so by Newey (1988),

$$\left\|n^{-1}\sum_i \frac{\partial m(z_i,\theta_0,\eta_0)}{\partial\theta} - \mathrm{E}\left[\frac{\partial m(z,\theta_0,\eta_0)}{\partial\theta}\right]\right\| = \mathrm{O}_{\mathrm{p}}(n^{-C}\nu(K)^{C\nu(K)}) . \tag{A.11}$$

By H, each of the following are finite:

$$\mathrm{E}[d_2(z_i)d_1'(z_i)] , \qquad \mathrm{E}[\|\rho_{\theta i}\|d_1'(z_i)] , \qquad \mathrm{E}[\|d_2(z_i)\|\ \|D_{0i}\|] ,$$
$$\mathrm{E}[d_1(z_i)^2 d_1'(z_i)^2] , \qquad \mathrm{E}[\|\rho_i\|^2 d_1'(z_i)^2] , \qquad \mathrm{E}[\|d_2(z_i)\|^2\|D_{0i}\|^2] .$$

Then, for $m(z,\theta,\eta) = (m_1(z,\theta,\eta)', \ldots, m_K(z,\theta,\eta)')'$, $\hat{D}_i = D(x_i, \hat{\eta})$, $D_{0i} = D(x_i,\eta_0)$, $\hat{\rho}_{\theta i} = \partial\rho(z_i, \hat{\theta})/\partial\theta$, $\rho_{\theta i} = \partial\rho(z_i,\theta_0)/\partial\theta$, $\Sigma_k = \Sigma_{k=1}^K$,

$$\begin{aligned}
&\left\|n^{-1}\sum_i \frac{\partial m(z_i,\hat{\theta},\hat{\eta})}{\partial\theta} - \mathrm{E}\left[\frac{\partial m(z,\theta_0,\eta_0)}{\partial\theta}\right]\right\| \\
&\quad \leqslant n^{-1}K\sum_i [\|\hat{\rho}_{\theta i} - \rho_{\theta i}\|\ \|\hat{D}_i - D_{0i}\| + \|\rho_{\theta i}\|\ \|\hat{D}_i - D_{0i}\| \\
&\qquad + \|\hat{\rho}_{\theta i} - \rho_{\theta i}\|\ \|D_{0i}\|] + \left\|n^{-1}\sum_i \frac{\partial m(z_i,\theta_0,\eta_0)}{\partial\theta} - \mathrm{E}\left[\frac{\partial m(z,\theta_0,\eta_0)}{\partial\theta}\right]\right\| \\
&\quad \leqslant n^{-1}K\sum_i \{d_2(z_i)d_1'(z_i) + \|\rho_{\theta i}\|d_1'(z_i) + \|d_2(z_i)\|\ \|D_{0i}\|\}\{\|\hat{\theta} - \theta_0\| \\
&\qquad + \|\hat{\eta} - \eta_0\|\} + \mathrm{O}_{\mathrm{p}}(n^{-C}\nu(K)^{C\nu(K)}) = \mathrm{O}_{\mathrm{p}}(n^{-C}\nu(K)^{C\nu(K)}) .
\end{aligned} \tag{A.12}$$

$$\begin{aligned}
&n^{-1}\sum_i \|m(z_i,\hat{\theta},\hat{\eta}) - m(z_i,\theta_0,\eta_0)\|^2 \\
&\quad \leqslant n^{-1}K\sum_i \{d_1(z_i)^2 d_1'(z_i)^2 + \|\rho_i\|^2 d_1'(z_i)^2 \\
&\qquad\qquad + \|d_2(z_i)\|^2\|D_{0i}\|^2\}\mathrm{O}_{\mathrm{p}}(1/\sqrt{n}) \qquad\qquad \text{(A.13)} \\
&\quad = \mathrm{O}_{\mathrm{p}}(n^{-C}\nu(K)^{C\nu(K)}) .
\end{aligned}$$

Furthermore, by a mean-value expansions and the same argument as for equation (A.12),

$$\begin{aligned}
&\sqrt{n}\left\|n^{-1}\sum_i m(z_i,\hat{\theta},\hat{\eta}) - n^{-1}\sum_i m(z_i,\theta_0,\hat{\eta}) \right. \\
&\qquad \left. - \mathrm{E}\left[\frac{\partial m(z,\theta_0,\eta_0)}{\partial\theta}\right](\hat{\theta} - \theta_0)\right\| \\
&\quad \leqslant \left\|n^{-1}\sum_i \frac{\partial m(z_i,\bar{\theta},\hat{\eta})}{\partial\theta} - \mathrm{E}\left[\frac{\partial m(z,\theta_0,\eta_0)}{\partial\theta}\right]\right\|\sqrt{n}\|\hat{\theta} - \theta_0\| \\
&\quad = \mathrm{O}_{\mathrm{p}}(n^{-C}\nu(K)^{C\nu(K)}) .
\end{aligned} \tag{A.14}$$

where $\bar{\theta}$ denotes the mean value, satisfying $\sqrt{n}\|\bar{\theta}-\theta_0\| \leqslant \sqrt{n}\|\hat{\theta}-\theta_0\| = \mathrm{O_p}(1)$. Also, $\mathrm{E}[\partial m(z_i, \theta_0, \eta_0)/\partial\eta] = 0$ and for some $\varepsilon > 0$, by H,

$$\mathrm{E}\left[\left\|\frac{\partial m_k(z_i, \theta_0, \eta_0)}{\partial\eta}\right\|^{1+\varepsilon}\right] \leqslant C\mathrm{E}\left[\left\|\frac{\partial D(x, \eta_0)}{\partial\eta}\right\|^{1+\varepsilon} \|\rho_i\|^{1+\varepsilon}\right] < C,$$

where it is assumed that $\eta$ is a scalar, for simplicity. Then by a lemma of Newey (1988),

$$\left\|n^{-1}\sum_i \frac{\partial m(z_i, \theta_0, \eta_0)}{\partial\eta}\right\| = \mathrm{O_p}(\mathrm{n}^{-\mathrm{C}}\nu(\mathrm{K})^{\mathrm{C}\nu(\mathrm{K})}),$$

and expanding around $\eta_0$ gives,

$$\begin{aligned}
&\sqrt{n}\left\|n^{-1}\sum_i m(z_i, \theta_0, \hat{\eta}) - n^{-1}\sum_i m(z_i, \theta_0, \eta_0)\right\| \\
&\quad \leqslant \left\|n^{-1}\sum_i \frac{\partial m(z_i, \theta_0, \eta_0)}{\partial\eta}\|\sqrt{n}\|\hat{\eta}-\eta_0\right\| \\
&\qquad + K\left[\sum_{i=1}^{n} d_2'(z_i)\|\rho_i\|/n\right]\sqrt{n}\|\hat{\eta}-\eta_0\|^2 \\
&\quad = \mathrm{O_p}(n^{-C}\nu(K)^{c\nu(K)}) .
\end{aligned} \tag{A.15}$$

It then follows by equations (A.12)–(A.15) that part (ii) of the hypotheses of L5 is satisfied, with $u = m(z, \beta_0, \eta_0)$. Part (iii) follows by equation (2.6). Also,

$$\begin{aligned}
&\mathrm{E}\left[\left\|B_i\rho_i - \sum_{k=1}^{K} \gamma_k m_k(z_i, \theta_0, \eta_0)\right\|^2\right] \\
&\quad \leqslant \mathrm{E}\left[\left\|B_i - D_{0i}\left\{\sum_{k=1}^{K} \gamma_k a_{kK}(x_i)\right\}\right\|^2 \|\rho_i\|^2\right] \\
&\quad \leqslant \mathrm{E}\left[\left\|B_i - D_{0i}\left\{\sum_{k=1}^{K} \gamma_k a_{kK}(x_i)\right\}\right\|^2 \|\Omega_i\|^2\right],
\end{aligned}$$

so the remainder of the hypotheses of L5 follow by Assumption 5.4 or Assumption 5.5 and Lemma A.9. The conclusion then follows by the conclusion of L5. □

## Acknowledgment

This research was partially supported by the NSF and the Sloan Foundation. An earlier version of this paper was presented at the 1986 European Meeting of the Econometric Society. Helpful comments were provided by L. Hansen and J. Robins.

## References

Amemiya, T. (1974). The non-linear two-stage least-squares estimator. *J. Econometrics* **2**, 105–110.

Amemiya, T. and J. L. Powell (1981). A comparison of the Box–Cox maximum likelihood estimator and the non-linear two-stage least squares estimator. *J. Econometrics* **17**, 351–381.

Berndt, E. R., B. H. Hall, R. E. Hall and J. A. Hausman (1974). Estimation and inference in nonlinear structural models. *Ann. Econom. Social Measurement* **3**, 653–666.

Box, G. E. P. and D. R. Cox (1967). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26**, 211–252.

Burguete, J. F., A. R. Gallant and G. Souza (1982). On the unification of the asymptotic theory of nonlinear econometric models. *Econometric Rev.* **1**, 151–190.

Carroll, R. J. (1982). Adapting for heteroskedasticity in linear models. *Ann. Statist.* **10**, 1224–1233.

Carroll, R. J. and D. Ruppert (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34**, 305–334.

Cragg, J. G. (1983). More efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* **49**, 751–764.

Gallant, A. R. (1980). Explicit estimators of parametric functions in nonlinear regression. *J. Amer. Statist. Assoc.* **75**, 182–193.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

Hansen, L. P. (1985). A method of calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *J. Econometrics* **30**, 203–238.

Hansen, L. P., J. Heaton and M. Ogaki (1988). Efficiency bounds implied by multi-period conditional moment restrictions. *J. Amer. Statist. Assoc.* **83**, 863–871.

Hausman, J. A. (1983). Simultaneous equations models. In: Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*. North-Holland, Amsterdam, Chapter 7.

Jobson, J. D. and W. A. Fuller (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related, *J. Amer. Statist. Assoc.* **75**, 176–181.

Jorgenson, D. W. and J. Laffont (1974). Efficient estimation of nonlinear simultaneous equations with additive disturbances. *Ann. Econom. Social Measurement* **3**, 615–640.

MaCurdy, T. E. (1982). Using information on the moments of the disturbance of increase the efficiency of estimation. Preprint, Stanford University.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models*. Chapman and Hall, New York.

Newey, W. K. (1988). Adaptive estimation of regression models via moment restrictions. *J. Econometrics* **38**, 301–339.

Newey, W. K. (1989a). Efficient estimation of semiparametric models via moment restrictions. MIT, Preprint.

Newey, W. K. (1989b). Locally efficient residual based estimation of nonlinear simultaneous equations models. MIT, Preprint.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica* **58**, 809–837.

Powell, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. In: W. A. Barnett, J. L. Powell, G. E. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge Univ. Press, Cambridge.

Reiersol, O. (1945). Confluence analyses by means of instrumental sets of variables. *Ark. Math. Astronom. Fys.* **32**, 1–119.

Robinson, P. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* **55**, 875–891.

Sargan, J. D. (1959). The estimation of relationships with autocorrelated results by the use of instrumental variables. *J. Roy. Statist. Soc. Ser. B* **21**, 91–105.

Stein, C. (1956). Efficient nonparametaric testing and estimation. In: *Proc. 3rd Berkeley Sympos. on Mathematical Statistics and Probability*, Vol. 1. Univ. California Press, Berkeley, CA.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5**, 595–645.

Theil, H. (1971). *Principles of Econometrics*. Wiley, New York.

White, H. (1982). Instrumental variables regression with independent observations. *Econometrica* **50**, 483–499.