SEMIPARAMETRIC WEAK-INSTRUMENT REGRESSIONS WITH AN APPLICATION TO THE RISK-RETURN TRADEOFF

Benoit Perron*

Abstract—We extend the local-to-zero analysis of models with weak instruments to models with estimated instruments and regressors and with higher-order dependence between instruments and disturbances. This framework is applicable to linear models with expectation variables that are estimated nonparametrically, such as the risk-return tradeoff in finance and the effect of inflation uncertainty on real economic activity. Our simulation evidence suggests that Lagrange multiplier confidence intervals have better coverage in these models. We apply these methods to excess returns on the S&P 500 index, yen-dollar spot returns, and excess holding yields between 6-month and 3-month Treasury bills.

I. Introduction

RECENTLY, the problem of weak correlation between instruments and regressors in instrumental variable (IV) regressions has become a focal point of much research. Staiger and Stock (1997) developed an asymptotic theory for this type of problem using a local-to-zero framework. They show that standard asymptotics for IV estimators can be highly misleading when this correlation is low. Following this methodology, Zivot, Startz, and Nelson (1998), Wang and Zivot (1998), and Startz, Nelson, and Zivot (2001) show that usual testing procedures are unreliable in such situations, and Chao and Swanson (2001) provide expressions for the bias and MSE of the IV estimator based on higher-order asymptotic approximations. Extensions of this approach to nonlinear models have been developed in Stock and Wright (2000). Earlier analyses of models under partial identification conditions are given in Phillips (1989), Choi and Phillips (1992), and Dufour (1997).

This paper extends the weak-instrument literature using Staiger and Stock’s framework in two ways: first, we analyze a restricted class of semiparametric models in which both regressors and instruments are estimated, and second, we allow for higher-order dependence between the instruments and the disturbances. These extensions are meant to make the analysis applicable to the many theoretical models in finance and macroeconomics that suggest a linear relationship between a random variable and an expectation term of the general form

\[ y_t = \gamma' x_t + \delta Z_t + \epsilon_t \]  

where \( y_t \) is a scalar, \( x_t \) is a vector of exogenous and predetermined variables, and \( Z_t \) is a vector of unobservable expectation variables.

The estimation of these models has proven difficult because a proxy has to be constructed for the unobservable expectation term. A complete parametric approach would assume functional forms for the expectation processes of agents, which could then be estimated along with equation (1) by, for example, maximum likelihood. A semiparametric approach, which is of interest in this paper, leaves the functional form of the expectation terms unspecified, but uses the linear structure in equation (1) to estimate the parameters of interest once estimates of the expectation terms are obtained.

Of particular interest is the case where \( Z_t \) is a conditional variance term, and in this framework, interest centers on the parameter \( \delta \), as it measures the response of \( y_t \) to increased risk. One such example includes the risk-return tradeoff in finance, where agents have to be compensated with higher expected returns for holding riskier assets. This tradeoff has been examined by several authors, including French, Schwert, and Stambaugh (1987), Glosten, Jaganathan, and Runkle (1993), and Braun, Nelson, and Sunier (1995); a good survey can be found in Lettau and Ludvigson (2001).

In this case, \( Z_t \) is the conditional variance of the asset, and \( x_t \) will generally include variables measuring the fundamental value of the asset. A second example is the analysis of the effect of inflation uncertainty on real economic activity, where \( Z_t \) is the variance of the inflation rate conditional on past information, and \( y_t \) is some real aggregate variable such as real GDP or industrial production.

In the case where \( Z_t \) is a variance term, Engle, Lilien, and Robins (1987) have introduced the parametric autoregressive conditional heteroskedasticity-in-mean (ARCH-M) model, which postulates that \( Z_t = \sigma_t^2 \), the variance of returns, follows an ARCH(p) model. A popular generalization is the generalized ARCH-M (GARCH-M) model with \( \sigma_t^2 \) of the form

\[ \sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_p \epsilon_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2 \]  

with equations (1) and (2) estimated jointly by maximum likelihood. Two problems surface when using such models. First, global maximization of the likelihood function can be difficult unless \( p \) and \( q \) are kept small. Second, estimates in the mean equation will be inconsistent if the variance equation is misspecified, because the information matrix is not block-diagonal. Given the lack of restrictions on the behavior of the conditional variance provided by economic theory, this seems quite problematic.

* Université de Montréal, CIREQ and CIRANO.

Thanks go to two referees, the Editor, Peter Phillips, Oliver Linton, Donald Andrews, Hyung-jick Roger Moon, John Galbraith, and seminar participants at Yale, Toronto, Montréal, Concordia, Rutgers, Windsor, Florida International University, and the 2000 World Congress of the Econometric Society for helpful comments and suggestions. The usual disclaimer necessarily applies. Financial assistance for the Alfred P. Sloan Foundation, the Social Science Research Council of Canada, the Fonds pour la Formation des chercheurs et l’Aide à la recherche (FCAR), and the Mathematics of Information Technology and Complex Systems (MITACS) network is gratefully acknowledged.

© 2003 by the President and Fellows of Harvard College and the Massachusetts Institute of Technology
An alternative approach that is robust to specification was suggested by Pagan and Ullah (1988) and Pagan and Hong (1991). Their suggestion is to first replace $Z_t$ by its realized values, say $Y_t$, estimating this quantity nonparametrically and using a nonparametric estimate of $Z_t$ as an instrument. This approach is itself problematic in that it does not avoid the necessity of keeping the number of conditioning variables low, due to the curse of dimensionality. Moreover, a common problem when using such a semiparametric approach is that the estimated conditional variance is poorly covered than more standard Wald confidence intervals. This approach is itself problematic in that it does not avoid the real effects of inflation uncertainty.

The first step in tackling this problem is to replace the conditional expectation $Z_t$ by its realized value $Y_t$. In the following, we assume that $Y_t$ is not observable as it is in the variance example, because $Y_t$ is itself a function of an expectation. Thus, an extra step is required in replacing $Y_t$ by an estimate, $\hat{Y}_t$. The model to be estimated is then

$$y_t = \gamma' x_t + \delta' \hat{Y}_t + e_t + \delta'(Y_t - \hat{Y}_t) + \delta'(Z_t - Y_t)$$

$$= \gamma' x_t + \delta' \hat{Y}_t + u_t.$$ 

In general, an ordinary least squares regression of $y_t$ on $x_t$ and $\hat{Y}_t$ will lead to inconsistent estimates of $\gamma$ and $\delta$ due to the correlation between $\hat{Y}_t$ and $Z_t - Y_t$. The solution suggested by Pagan (1984) and by Pagan and Ullah (1988) is to use an instrumental variables estimator with $\hat{Z}_t$ used as instruments for $\hat{Y}_t$. In fact, to obtain consistent estimates, any variable in $\mathcal{F}_t$ could be used as instrument. We could consider finding an optimal instrument as $E[\hat{Y}_t|\mathcal{F}_t]$, which in general will be different from $\hat{Z}_t$ because of the bias arising from the estimation of $Y_t$. The steps used to construct the estimator are displayed as follows:

$$Z_t \rightarrow Y_t \rightarrow \hat{Y}_t \rightarrow \hat{Z}_t.$$ 

This problem will be semiparametric when $Y_t$ and $Z_t$ are estimated nonparametrically. As in many semiparametric models, despite the lower rate of convergence of the nonparametric estimators, the estimates of $\gamma$ and $\delta$ will converge at the usual $\sqrt{n}$ rate under certain conditions, where $n$ is the sample size.

Define $\hat{Z}_t = (x_t, Z_t)$, $\hat{Y}_t = (x_t, Y_t)$, $\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_n)'$, $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)'$ with $\hat{Z}_t$ and $\hat{Y}_t$ replacing $Z_t$ and $Y_t$. Further let $\bar{u}_t = e_t + \delta'(Z_t - Y_t)$ and $\hat{\theta} = (\gamma, \delta)'$. Consider the IV estimator for this model:

$$\hat{\theta} = (\hat{Z}' \hat{Y})^{-1} \hat{Z}' \bar{u}_t.$$ 

Andrews (1994) proved the asymptotic normality of this estimator. There are two conditions of interest here.

The first one is that $\hat{Y}$ is $\sqrt{n}$-consistent. This ensures that the asymptotic distribution of the IV estimator of $\hat{\theta}$ is not affected by replacing $Y_t$ and $Z_t$ with $\hat{Y}_t$ and $\hat{Z}_t$, respectively. This will generally not be the case when $\hat{Y}$ is estimated nonparametrically. However, it will hold in the special case where $Z_t$ is a variance term as long as the mean of $E[\psi_t|\mathcal{F}_t]$
multivariate standard normal and are plotted in versus 3-month Treasury bills between 1959:1 and 2000:2. The bandwidth of the conditioning variables. The bandwidth takes the form

\[ b_j = c_j s n^{-(1/4 + p_j)}, \]

where \( s \) is the standard deviation of \( y_t \) and \( c_j \) is a constant to be selected. We then define \( \hat{e}_t^2 = (y_t - \hat{\tau}_{1t})^2 \) and obtain an estimate of \( \sigma_t^2 \) as

\[ \hat{\sigma}_t^2 = \hat{\tau}_{2t} - (\hat{\tau}_{1t})^2. \]

A theoretical analysis of this nonparametric estimator of the conditional variance can be found in Masry and Tjostheim (1995). In order to avoid unbelievably small bandwidth choices for all three series, we left out outliers in the bandwidth selection process. The extreme 25% of the data was not used in the computation of the information criteria. The second estimator was first proposed by Engle and Ng (1993). It provides more structure to the conditional variance and will approximate the conditional variance function much better than the kernel when the variance is persistent [see Perron (1999) for simulation evidence]. The estimator is implemented by first estimating the mean by a kernel estimate as above and then fitting an additive function for \( \sigma_t^2 \) as follows:

\[ \sigma_t^2 = \omega + f_1(\hat{e}_{t-1}) + \cdots + f_p(\hat{e}_{t-p}) + \beta \sigma_{t-1}^2, \]

where the \( f(\cdot) \) are estimated as splines with knots using a Gaussian likelihood function. This allows for a flexible effect of recent information on the conditional variance while allowing for persistence. This framework includes most parametric models suggested in the literature, such as the GARCH class. The number of segments in the spline functions acts as a smoothing parameter and is selected using BIC. The knots in the spline were selected using the order statistics so that each bin has roughly the same number of observation subject to the constraint of equal numbers of bins in the positive and negative regions.

A quick look at table 1 reveals that only the excess holding yield data have \( R^2 \) greater than 5.5%. The reason for this low correlation is that \( e_t^2 \) and \( \sigma_t^2 \) have very different volatilities. Even if \( E[e_t^2|\mathcal{F}_t] = \sigma_t^2 \), financial returns are extremely volatile and therefore the difference between \( e_t^2 \)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Period</th>
<th>Kernel (%)</th>
<th>Engle-Ng (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S&amp;P 500 returns</td>
<td>1965:1–1997:12</td>
<td>5.26</td>
<td>2.02</td>
</tr>
<tr>
<td>Yen-dollar returns</td>
<td>1978:10–1998:12</td>
<td>5.07</td>
<td>0.38</td>
</tr>
</tbody>
</table>

is estimated at rate \( n^{1/4} \). Conditions under which this holds can be found in Andrews (1995).

The second key assumption is that the matrix \( n^{-1/2} \hat{Y} \) converges to a nonsingular limit. It is a key assumption because the quality of the instrument \( \hat{Z} \) will determine the quality of the asymptotic approximation obtained by Andrews (1994). This assumption is nearly violated in many practical situations, and that is the motivation for the development of the weak-instrument literature. The next section will document this phenomenon for financial data.

### III. Evidence of Weak Instruments

In the case of interest in which \( Y_t = e_t^2 \) and \( Z_t = \sigma_t^2 \), it will generally be the case that the correlation between the two estimates, \( \hat{e}_t^2 \) and \( \hat{\sigma}_t^2 \), is very low, suggesting a weak-instrument problem. Table 1 shows the value of \( R^2 \) for the regression of \( \hat{e}_t^2 \) on a constant and \( \hat{\sigma}_t^2 \) for three financial data sets using two different nonparametric estimators.

The first data set analyzed represents monthly excess returns on the Standard and Poor’s 500 between January 1965 and December 1997 measured at the end of each month. The data are taken from CRSP, and the risk-free rate is the return on 3-month Treasury bills. The second data set is made of monthly returns on the yen-dollar spot rate obtained from International Financial Statistics between September 1978 and June 1998. Finally, the last data series consists of quarterly excess holding yields on 6-month versus 3-month Treasury bills between 1959:1 and 2000:2.

A similar, but shorter, data set has already been analyzed by Engle, Lilien, and Robins (1987) using their ARCH-M methodology, and by Pagan and Hong (1991) using the above instrumental variables estimator. The three data sets are plotted in figure 1.

The first nonparametric estimator is based on a multivariate leave-one-out kernel. First, we estimate the means of \( y_t \) and \( \hat{y}_t^2 \), denoted \( \tau_{1t} \) and \( \tau_{2t} \), respectively, as

\[ \hat{\tau}_{jt} = \frac{\sum_{i \neq t} y_i K(w_i - w_t)}{\sum_{i \neq t} K(w_i - w_t)} \]

for \( j = 1, 2 \) with the kernel function \( K(w) \) taken to be the multivariate standard normal and \( w_t \) to be a vector of conditioning variables. The bandwidth \( b_j \) and the number of lags of \( y_t \) in the conditioning set \( p_j \) are selected using a modified version of the criterion suggested in Tjostheim and Auestad (1994) that penalizes small bandwidths and large lag lengths. Accordingly, we choose the bandwidth \( (b_j) \) and lag length \( (p_j) \) so as to minimize

\[ \ln \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\tau}_j)^2 \right) + \ln \left( \frac{K(0)}{b_j} \right) \sum_{i=1}^{n} (y_i - \hat{\tau}_j)^2, \]

where \( K(0) \) is the kernel evaluated at 0 and \( f(w_j) \) is the density of the conditioning variables. The bandwidth takes the form

\[ b_j = c_j s n^{-(1/4 + p_j)}, \]

where \( s \) is the standard deviation of \( y_t \) and \( c_j \) is a constant to be selected. We then define \( \hat{e}_t^2 = (y_t - \hat{\tau}_{1t})^2 \) and obtain an estimate of \( \sigma_t^2 \) as

\[ \hat{\sigma}_t^2 = \hat{\tau}_{2t} - (\hat{\tau}_{1t})^2. \]

The second estimator is first proposed by Engle and Ng (1993). It provides more structure to the conditional variance and will approximate the conditional variance function much better than the kernel when the variance is persistent [see Perron (1999) for simulation evidence]. The estimator is implemented by first estimating the mean by a kernel estimate as above and then fitting an additive function for \( \sigma_t^2 \) as follows:

\[ \sigma_t^2 = \omega + f_1(\hat{e}_{t-1}) + \cdots + f_p(\hat{e}_{t-p}) + \beta \sigma_{t-1}^2, \]

where the \( f(\cdot) \) are estimated as splines with knots using a Gaussian likelihood function. This allows for a flexible effect of recent information on the conditional variance while allowing for persistence. This framework includes most parametric models suggested in the literature, such as the GARCH class. The number of segments in the spline functions acts as a smoothing parameter and is selected using BIC. The knots in the spline were selected using the order statistics so that each bin has roughly the same number of observation subject to the constraint of equal numbers of bins in the positive and negative regions.

A quick look at table 1 reveals that only the excess holding yield data have \( R^2 \) greater than 5.5%. The reason for this low correlation is that \( e_t^2 \) and \( \sigma_t^2 \) have very different volatilities. Even if \( E[e_t^2|\mathcal{F}_t] = \sigma_t^2 \), financial returns are extremely volatile and therefore the difference between \( e_t^2 \)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Period</th>
<th>Kernel (%)</th>
<th>Engle-Ng (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S&amp;P 500 returns</td>
<td>1965:1–1997:12</td>
<td>5.26</td>
<td>2.02</td>
</tr>
<tr>
<td>Yen-dollar returns</td>
<td>1978:10–1998:12</td>
<td>5.07</td>
<td>0.38</td>
</tr>
</tbody>
</table>
and $\sigma^2_t$ can be quite large. This would be true even if we did not have to estimate these two quantities; having to estimate them complicates matters further. We can illustrate by looking at the GARCH(1, 1) model:

$$y_t = \mu + \sigma_t \varepsilon_t = \mu + e_t,$$

$$\sigma_t^2 = \omega + \alpha e_{t-1}^2 + \beta \sigma_{t-1}^2.$$ 

Andersen and Bollerslev (1998) show that the population $R^2$ in the regression

$$(y_t - \mu)^2 = a_0 + a_1 \hat{\sigma}_t^2 + v_t,$$

where $\hat{\sigma}_t^2$ is the one-period-ahead forecast obtained from the GARCH model, is

$$R^2 = \frac{\alpha^2}{1 - \beta^2 - 2\alpha \beta},$$

which will in general be very small even though $E[(y_t - \mu)^2|\hat{\sigma}_t^2] = \sigma^2_t$. Figure 2 plots the value of $R^2$ for different values of $\alpha$ and $\beta$ for this GARCH(1, 1) example. The value of $R^2$ is highly sensitive to the value of $\alpha$, and this reflects the fact that $\alpha = 0$ makes the model unidentified. It is usual in the literature to find point estimates of GARCH(1, 1) models in the neighborhood of $\alpha = 0.05$ and $\beta = 0.9$. The figure clearly shows that for such values, the correlation between $e_t^2$ and $\sigma_t^2$ will typically be quite low. The problem in this case is that $\sigma_t^2$ has very low variance relative to that of $y_t^2$; a low value of $\alpha$ means that $\sigma_t^2$ is nearly constant locally.

We can expect that Table 1 will not even provide an accurate picture of the problem of weak instruments. Using data sampled at higher frequency (daily or even intra-day) would result in even lower correlation. The lower frequency allows some averaging, which reduces the variance of $e_t^2$. Potentially a better solution is to use model-free measures of volatility such as those proposed by Andersen et al. (2001), which are obtained

---

**Figure 1.**—Data

**Returns on S&P 500 index - 1865-1997**

**Yen-dollar spot returns - 1979-1999**

**Excess holding yield - 1959:1-2003:2**
by summing squared returns from higher-frequency data. We do not pursue this possibility here, but note that its variance-reducing property could be helpful in this context.

IV. Asymptotics with Weak Instruments

Staiger and Stock (1997) have recently shown, in the framework of a linear simultaneous equation system, that having instruments that are weakly correlated with the explanatory variables makes the usual asymptotic theory work poorly. Their assumed model is

\[ y = Y\delta + X\gamma + u, \]
\[ Y = Z\Pi + X\Gamma + V, \]

where \( Y \) is the matrix of included endogenous variables that are to be replaced by at least \( k_2 \) instruments. Because in our case it will always be true that the model is exactly identified (that is, there will be as many regressors as instruments, because the instruments are estimates of the expected value of the regressors), we will concentrate on the case where \( Z \) is a \( n \times k_2 \) matrix. The weak-instrument assumption is imposed by assuming that

\[ \Pi = \frac{G}{\sqrt{n}} \]

for some fixed \( k_2 \times k_2 \) matrix \( G \neq 0 \). This assumption implies that in the limit, \( Y \) and \( Z \) are uncorrelated.

We extend the analysis of weak instruments in Staiger and Stock (1997) to our case of interest by allowing \( Y \) and \( Z \) to be unobserved and estimated by \( \hat{Y} \) and \( \hat{Z} \) respectively. Moreover, we allow for the possibility of higher-order dependence between the instruments and the disturbances. Simple algebra leads to

\[ \hat{Y} = \hat{Z}\Pi + (Z - \hat{Z})\Pi + (\hat{Y} - Y) + X\Gamma + V \]
\[ = \hat{Z}\Pi + X\Gamma + \zeta, \]

so that the correlation between \( \hat{Y} \) and \( \hat{Z} \) is also low.

There are two possible reasons for a low correlation between the estimated instrument and explanatory variable in a given data sample. The first is that the estimators used in constructing \( \hat{Z} \) and \( \hat{Y} \) may be poor and not approach their true value in small samples. On the other hand, the estimators may not be poor in any sense, but \( Y \) and its expected value may be weakly correlated as in the GARCH(1, 1) example above.

Recall that the IV estimator of \( \delta \) is

\[ \hat{\delta} = (\hat{Z}'M_{\delta}\hat{Y})^{-1}\hat{Z}M_{\delta}Y \]
\[ = \delta + (\hat{Z}'M_{\delta}\hat{Y})^{-1}\hat{Z}'M_{\delta}u, \]
where $M_Y = I - X(X'X)^{-1}X'$. In order to derive the distribution of $\hat{\gamma}$, we need to make an extra assumption on the reduced-form coefficients of $X$. We will also assume that they are local to zero:

$$
\Gamma = \frac{H}{\sqrt{n}}
$$

for some $k_1 \times k_2$ matrix $H \neq 0$. This assumption is made because if $\Gamma$ were fixed, $X$ and $Y$ would be collinear in the limit and the moment matrices would be singular. The assumption plays no role in the analysis of the behavior of $\hat{\delta}$.

The distribution of the estimators is given in the following theorem. All proofs are relegated to the appendix.

**Theorem 4.1.** In the model (4)–(6), assume the following:

1. $\sqrt{n}(\hat{\gamma} - \gamma) \to^d 0$;
2. $Z = Z + o_p(1), Z < \infty$ a.s.;
3. $\theta_0$ is the interior of $\Theta \subset R^{k_1+1}$;
4. $(n^{-1/2}X', n^{-1}XZ, n^{-1/2}Z'M_XZ) \to^d (\Sigma_{XX}, \Sigma_{XZ}, \Sigma_{ZZ})$;
5. $(n^{-1/2}X'u, n^{-1/2}Z'M_X'u, n^{-1/2}Z'V't, \alpha^{1/2}Z'XV't, \alpha^{1/2}Z'XZV't) \to^d (\Psi_{Xu}, \Psi_{Zu}, \Psi_{XV}, \Psi_{ZV})$.

Define

$$
\sigma_{ZZ} = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} Z_i^2 \epsilon_i^2 u_i Z_i'^2,
$$

$$
\sigma_{ZV} = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} Z_i V'_i V_i Z_i'^2,
$$

$$
\sigma_{Xu} = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i u_i u_i,'
$$

$$
\sigma_{XV} = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i V'_i V_i X_i',
$$

$$
\rho_Z = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} Z_i V_i \sigma_{ZV}^{-1/2} \sigma_{ZZ}^{-1/2} u_i Z_i'^2,
$$

$$
\rho_X = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i V_i \sigma_{ZV}^{-1/2} \sigma_{ZZ}^{-1/2} u_i X_i,
$$

where $Z_i^t$ is the residuals from the projection of $Z_i$ onto $X$, that is, it is the transpose of the

$$(t^{th} \text{ row of } Z^t) = M_XZ.$$
**Proposition 4.2.** Under the null hypothesis $H_0 : R\delta = r$, the $W$ statistic:

$$W = (R\hat{\delta} - r)'[R(\hat{Z}'M\hat{Y})^{-1} \text{var}(\hat{Z}'M\hat{u})]^{-1}(R\hat{\delta} - r).$$

The following proposition gives the asymptotic theory of Wald statistics in the above model:

**Proposition 4.3.** Let $g = n^{-1}\hat{Z}'M_y(y - \hat{Y}_0)$. Then under the null hypothesis $H_0 : \delta = \delta_0$ we have $LM = ng'(\hat{Z}_{\delta} \delta_0^{\ast} g) \rightarrow X^2(k_2)$, where $\hat{Z}_{\delta,0} = n^{-1} \sum_t Z_t' u_t g_0 u_t g_0 Z_t'$ is an estimator of $\sigma_{\delta,0}$ computed under the null hypothesis.

Unfortunately, in this case there is no easy way to write the inequality that defines the confidence intervals as a quadratic equation in $\delta$. Confidence intervals must be computed numerically by defining a grid of $\delta$ and verifying for each point on the grid whether the LM statistic defined in the above proposition is less than the appropriate critical value from the $X^2(k_2)$ distribution. This method is easily implemented in the scalar case, but could hardly be carried out in high dimensions.

Another approach to obtaining confidence intervals, suggested by Staiger and Stock (1997), is to use the Anderson-Rubin (1949) statistic. It is usually defined as the $F$-statistic for the significance of $\delta^{\ast}$ in the regression

$$y - \hat{Y}_0 = X\gamma^{\ast} + \hat{Z}\delta^{\ast} + u^{\ast},$$

where $\gamma^{\ast} = \gamma + \Gamma(\delta - \delta_0)$, $\delta^{\ast} = \Pi(\delta - \delta_0)$, and $u^{\ast} = u + \nu(\delta - \delta_0)$. Since we have a case with heteroskedasticity, we need to use robust standard errors to compute the test statistic. It turns out that in the just-identified case, this statistic is identical to the above LM statistic. This fact is stated in the following proposition:

**Proposition 4.4.** Let $AR = n\hat{\delta}^{\ast}V^{-1}\hat{\delta}^{\ast}$, where $V = (\hat{Z}'M_yZ)^{-1}\hat{Z}_{\delta,0}^{\ast} \hat{Z}_{\delta,0}^{\ast'} (\hat{Z}'M_yZ)^{-1}$. Then, under the null hypothesis $H_0 : \delta = \delta_0$, we have $AR = LM$.

The above propositions thus give us two equivalent ways to construct asymptotically valid confidence intervals for the entire vector $\delta$. The two methods are exactly the same as long as the same estimate of $\sigma_{\delta,0}$ is used to construct either LM or AR. The performance of these intervals in a small-sample situation will be analyzed in the simulation experiment in the next section. In the case where a confidence interval on a linear combination of a subvector of $\delta$ is desired, one can proceed by the projection method discussed in Dufour and Jasiak (2001) and further analyzed in Dufour and Taamouti (2001). Such an approach would be valid but conservative.

In a related paper, Dufour and Jasiak (2001) have obtained exact tests based on AR-type statistics in models with generated regressors and weak instruments. However, their results only apply to parametrically estimated regressors that will converge at rate $\sqrt{n}$, and not to the nonparametric estimators analyzed here.
Stutz, Nelson, and Zivot (2001) have developed an alternative set of statistics, which they call S-statistics, that take into account the degree of identification. They show in the case of a single regressor and instrument \((k_1 = k_2 = 1)\) that these are equivalent to the AR statistic. We suspect that this correspondence is more general and carries over to the exactly identified case that we treat here, but we have no proof for this conjecture.

### V. Simulation Results

In this section, the behavior of the procedures described above will be analyzed through a small simulation experiment. Important issues to be analyzed include the choice of smoothing parameters, the appropriateness of the various confidence intervals, and the distribution of the resulting estimators.

Consider the GARCH-M(1, 1) DGP:

\[
y_t = \gamma + \delta \sigma_t^2 + \epsilon_t = \gamma + \delta \sigma_t^2 + \sigma \varepsilon_t,
\]

\[
\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2,
\]

\[
\varepsilon_t \sim \text{i.i.d.}(0, 1).
\]

In the above notation, we have \(y_t = \epsilon_t^2 - \sigma_t^2, \ u_t = \epsilon_t - \delta \varepsilon_t, \ Y_t = \epsilon_t^2, \) and \(Z_t = \sigma_t^2. \) The distribution of \(\varepsilon_t\) is either normal or Student \(t. \) This allows us to check the robustness of the procedures to the restrictive moment assumptions required by the asymptotic theory developed above. We use six sets of parameters, all estimated from data, which are presented in table 2.

The point estimates for the stock data are similar to those usually obtained in this context, for example by Glosten, Jagannathan, and Runkle (1993), and will lead to a rather persistent \(\sigma_t^2\) and to a weak instrument. Sample sizes of 450, 300, and 150 are used for the experiments, with the first 50 observations deleted to remove the effect of the initial condition (taken as the mean of the unconditional distribution). The lengths of the samples nearly match those of the S&P, exchange-rate, and excess-holding-yield data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>I (S&amp;P 500)</th>
<th>II (yen-dollar)</th>
<th>III (holding yields)</th>
<th>IV (S&amp;P 500—t)</th>
<th>V (yen-dollar—t)</th>
<th>VI (holding yields—t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\gamma)</td>
<td>-0.099</td>
<td>0.059</td>
<td>-0.001</td>
<td>-0.012</td>
<td>0.109</td>
<td>0.0005</td>
</tr>
<tr>
<td>(\omega)</td>
<td>(1.44 \times 10^{-4})</td>
<td>(8.42 \times 10^{-4})</td>
<td>0.17 \times 10^{-7}</td>
<td>2.03 \times 10^{-4}</td>
<td>8.91 \times 10^{-4}</td>
<td>2.03 \times 10^{-4}</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>0.066</td>
<td>0.061</td>
<td>0.312</td>
<td>0.064</td>
<td>0.043</td>
<td>0.330</td>
</tr>
<tr>
<td>(\beta)</td>
<td>0.855</td>
<td>0</td>
<td>0.680</td>
<td>0.821</td>
<td>0</td>
<td>0.651</td>
</tr>
<tr>
<td>(\delta)</td>
<td>6.676</td>
<td>-65.661</td>
<td>48.722</td>
<td>8.444</td>
<td>-115.349</td>
<td>36.265</td>
</tr>
<tr>
<td>(\nu)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(n)</td>
<td>400</td>
<td>250</td>
<td>150</td>
<td>400</td>
<td>250</td>
<td>150</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>2.145</td>
<td>0.757</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(\rho)</td>
<td>-0.472</td>
<td>0.953</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(\sigma^{Zu}_2)</td>
<td>(6.819 \times 10^{-10})</td>
<td>(7.304 \times 10^{-11})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(\sigma^{ZV}_2)</td>
<td>(3.404 \times 10^{-12})</td>
<td>(1.540 \times 10^{-14})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(R^2) (%)</td>
<td>2.77</td>
<td>0.37</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

DGP: \(y_t = \gamma + \delta \sigma_t^2 + \sigma \varepsilon_t, \ \sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \ \varepsilon_t \sim \text{i.i.d.}(0, 1)\) for experiments I-III; \(y_t \sim \text{i.i.d.}(\varepsilon'\varepsilon)\) for experiments IV-VI.

One disadvantage of the current setup is that the correlation between \(\hat{\sigma}_t^2\) and \(\hat{\epsilon}_t^2\) cannot be controlled. We can control the correlation between the unobservable variables, but due to estimation, the correlation between observable variables will be different in general.

The values of the nuisance parameters in this setup can be obtained in terms of the moments of the conditional variance process as

\[
\sigma^{ZV} = (\kappa_4 - 1)[E(\sigma_t^4) - 2E(\sigma_t^2)E(\sigma_t^2)]
\]

\[
+ E(\sigma_t^2)^2E(\sigma_t^2),
\]

\[
\sigma^{Zu} = \delta^2 \sigma^{ZV}_2 + E(\sigma_t^4) - 2E(\sigma_t^2)E(\sigma_t^2) + E(\sigma_t^2)^3
\]

\[
\sigma^{Zu}_2 = \delta^2 \sigma^{ZV}_2 + E(\sigma_t^4) - 2E(\sigma_t^2)E(\sigma_t^2) + E(\sigma_t^2)^3
\]

\[
- 2\delta \kappa_3[E(\sigma_t^3) - 2E(\sigma_t^2)E(\sigma_t^2) + E(\sigma_t^2)^2],
\]

\[
\rho_Z = \frac{-\delta \sigma^{Zu}_2 + \kappa_3[E(\sigma_t^3) - 2E(\sigma_t^2)E(\sigma_t^2) + E(\sigma_t^2)^2]}{\sigma^{ZV}_2 \sigma^{ZV}_2},
\]

\[
\lambda = \frac{\sqrt{\delta}}{\sigma^{ZV}_2},
\]

\[
\sigma^{Zu}_2 = \sigma^{Zu}_2 - 2\sigma^{Zu}_2 \Xi + \sigma^{ZV}_2 \Xi^2,
\]

where \(\kappa_j = E(\varepsilon_j)\) is the \(j\)th moment of \(\varepsilon_t. \) The values of the first four even moments of \(\sigma_t^2\) are derived recursively in Bollerslev (1986) as a function of \(\omega, \alpha, \) and \(\beta\) and the moments of \(\varepsilon_t. \) This allows for the easy computation of the nuisance parameters, which are included in table 2. Note that the moment condition assumed in theorem 4.1 is only satisfied for the first two sets of parameters.

Figure 3 shows a plot of the asymptotic distribution of the usual \(t\)-statistic (from proposition 4.2) using the above estimates of the nuisance parameters and the standard normal distribution obtained under the usual asymptotic theory for the first two sets of parameters. The figure is based on 500,000 draws taken from each distribution. For the other experiments, because the higher-order moments necessary to obtain the limit distribution do not exist, we cannot use
the weak-instrument limiting distribution to describe the behavior of the estimator.

The top panel in figure 3 represents the distribution for the S&P 500 data. For those values of the parameters, the t-statistic has a highly skewed distribution. On the other hand, the bottom panel reveals that for the second experiment, the t-statistic is highly skewed and also has fat tails. In fact, a good part of the probability mass (about 7%) lies outside the \([-4, 4]\) interval. The shape of the distribution is controlled by two nuisance parameters, \(\lambda\) and \(\rho_Z\). These experiments show that low \(\lambda\) and high \(|\rho_Z|\) give distributions very far from normality. To measure the effect of these properties on coverage probabilities, note that only 77.7% of the mass is between \(-1.96\) and 1.96 in the bottom panel; the corresponding fraction in the top panel is 96.5%. We conclude that the first experiment will have usual (Wald-based) 95% confidence intervals with coverage rates higher than their nominal level, whereas those in the second experiment will exhibit low coverage.

To demonstrate convergence to normality, figure 4 shows the same picture for \(n = 50,000\) for both experiments. Because the weak-instrument approximation approaches the standard normal as \(n \rightarrow \infty\) in this case (because \(\lambda \rightarrow \infty\) at rate \(\sqrt{n}\), we see that both skewness and excess kurtosis are much reduced. In this case, 95.2% and 95.1% of the mass lies between \(-1.96\) and 1.96 in the two experiments, respectively. Because of the parameter values, the distribution for the second experiment requires a much larger sample size than the first one in order to have a reasonably normal distribution and accurate 95% Wald-based confidence intervals.

The simulation results are presented in figures 5 and 6 and tables 3 and 4. Figure 5 provides a plot of the density of the weak IV approximation and of the infeasible IV estimator that uses the actual values of \(\sigma^2_t\) and \(\epsilon^2_t\) generated; this estimator is infeasible because these values are unobservable in practice. Figure 6 provides the same information for the two nonparametric estimators. In tables 3 and 4, the first
column shows the median of the IV estimator (rather than the mean, because of the heavy tails of the distributions). The next two columns indicate the coverage rate of the appropriate 95% confidence intervals. The fifth column contains the mean $R^2$ of a regression of $\hat{e}_t^2$ on a constant and $\hat{\sigma}_t^2$. The next two columns provide the Kolmogorov-Smirnov (KS) statistic as a measure of fit of the small-sample distribution to the two alternative asymptotic approximations (if applicable). Finally, the last two columns compare the fits of the nonparametric estimates of both the regressor and instrument by reporting the $R^2$ from a regression of the true values on a constant and the nonparametric estimates. The first line of each panel reports results of the infeasible estimator discussed above.

We first discuss the results for the infeasible estimator. All experiments with the infeasible estimator were repeated 10,000 times. The asymptotic approximation captures the finite-sample distribution of the $t$-statistic well. It matches the skewness and kurtosis well and thus provides a much better description than the normal approximation.

In all six experiments, the infeasible IV estimator is biased upward. The Wald confidence intervals have a coverage rate that is higher than its nominal level for the S&P data and lower (and sometimes much lower) for the other two data sets, while the LM interval has coverage rate that is only slightly too low in all cases. Not surprisingly, the weak-instrument approximation is more accurate according to the KS statistic in both cases where it can be computed. The improvement is much more dramatic in the very non-normal case of experiment 2. Note also that the overall results are not sensitive to conditional normality or the existence of moments.

We now turn our attention to the semiparametric estimators. Estimates of $e_t^2$ and $\sigma_t^2$ are obtained using the same two nonparametric methods as above, either a kernel or the semiparametric Engle-Ng estimator, using data-based selec-
**TABLE 3.** SIMULATION RESULTS: GARCH-M (1, 1) PARAMETERS—CONDITIONAL NORMALITY

<table>
<thead>
<tr>
<th>Method</th>
<th>Median</th>
<th>Coverage of 95% CI</th>
<th>First-stage $R^2$ (%)</th>
<th>KS statistic</th>
<th>Fit of instrument (%)</th>
<th>Fit of regressor (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>LM</td>
<td></td>
<td>Weak IV</td>
<td>Normal</td>
</tr>
<tr>
<td>Infeasible</td>
<td>7.90</td>
<td>97.4</td>
<td>93.7</td>
<td>2.19</td>
<td>0.087</td>
<td>0.107</td>
</tr>
<tr>
<td>Kernel</td>
<td>2.79</td>
<td>92.3</td>
<td>94.1</td>
<td>2.33</td>
<td>0.204</td>
<td>0.216</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>6.31</td>
<td>98.3</td>
<td>95.6</td>
<td>1.99</td>
<td>0.048</td>
<td>0.094</td>
</tr>
</tbody>
</table>

Experiment I—S&P Data (True Slope Parameter = 6.676)

| Infeasible | −41.73 | 81.8 | 92.6 | 0.76 | 0.095 | 0.328 | 100.0 | 100.0 |
| Kernel     | −13.72 | 19.3 | 73.7 | 2.73 | 0.583 | 0.797 | 28.7  | 96.0  |
| Engle-Ng   | −28.87 | 66.6 | 97.8 | 2.28 | 0.111 | 0.349 | 22.8  | 88.9  |

Experiment II—Yen-Dollar Data (True Slope Parameter = −65.661)

| Infeasible | 76.68  | 89.8 | 92.8 | 14.82 | — | — | 100.0 | 100.0 |
| Kernel     | 72.75  | 88.2 | 93.2 | 14.75 | — | — | 30.7  | 85.3  |
| Engle-Ng   | 79.71  | 96.0 | 94.0 | 12.23 | — | — | 66.5  | 77.1  |

Experiment III—Excess-Holding-Yield Data (True Slope Parameter = 48.736)
TABLE 4.—SIMULATION RESULTS: GARCH-M(1, 1) PARAMETERS—CONDITIONAL $t$

<table>
<thead>
<tr>
<th>Method</th>
<th>Median</th>
<th>Coverage of 95% CI</th>
<th>First-stage $R^2$ (%)</th>
<th>Fit of instrument (%)</th>
<th>Fit of regressor (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>LM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment IV—S&amp;P Data (True Slope Parameter = 8.444)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infeasible</td>
<td>9.60</td>
<td>97.1</td>
<td>94.3</td>
<td>1.66</td>
<td>100.0</td>
</tr>
<tr>
<td>Kernel</td>
<td>3.00</td>
<td>68.6</td>
<td>90.3</td>
<td>3.61</td>
<td>10.7</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>7.26</td>
<td>92.1</td>
<td>95.7</td>
<td>2.27</td>
<td>49.3</td>
</tr>
<tr>
<td>Experiment V—Yen-Dollar Data (True Slope Parameter = −115.349)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infeasible</td>
<td>−46.16</td>
<td>72.2</td>
<td>92.3</td>
<td>0.62</td>
<td>100.0</td>
</tr>
<tr>
<td>Kernel</td>
<td>−10.81</td>
<td>7.8</td>
<td>71.4</td>
<td>3.70</td>
<td>23.9</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>−32.31</td>
<td>48.8</td>
<td>98.5</td>
<td>1.32</td>
<td>20.1</td>
</tr>
<tr>
<td>Experiment VI—Excess-Holding-Yield Data (True Slope Parameter = 36.265)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infeasible</td>
<td>54.95</td>
<td>95.3</td>
<td>91.2</td>
<td>17.11</td>
<td>100.0</td>
</tr>
<tr>
<td>Kernel</td>
<td>0.89</td>
<td>28.4</td>
<td>92.4</td>
<td>26.21</td>
<td>22.3</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>3.59</td>
<td>35.4</td>
<td>80.9</td>
<td>10.81</td>
<td>51.3</td>
</tr>
</tbody>
</table>

FIGURE 6.—DISTRIBUTION OF IV ESTIMATOR

Experiment 1—Estimates from S&P 500 data

Experiment 2—Estimates from yen-dollar data
tion for all smoothing parameters. Each experiment with the nonparametric estimators was repeated 5000 times.

The need to estimate $\sigma_t^2$ and $e_t^2$ changes the result quite dramatically relative to the infeasible estimator. The results using the kernel estimates are presented in the second row of each panel of tables 3 and 4 and as the dashed line in figure 6; those for the Engle-Ng are presented in the third row of each panel and as the dotted line in figure 6. Overall, the Engle-Ng procedure leads to an IV estimator that much more closely matches the infeasible one. In particular, its distribution has a similar shape to that of the infeasible IV (and that of the weak IV approximation), and the coverage rate of the confidence intervals based on it are much closer to those of the infeasible estimator. The reason for this is clear: it provides a better approximation to the instrument ($\sigma_t^2$) than does the kernel, as evidenced by the higher $R^2$ in the regression of true conditional variance on a constant and its estimate, which is consistent with the simulation evidence in Perron (1999). The regressor ($e_t^2$) is well approximated by any method. Note also that once we estimate the regressor and instrument, the IV estimator of $\sigma_t^2$ is strongly biased towards zero (with the exception of experiment 3).

An important practical result is that LM-based confidence intervals are more robust (in terms of having correct coverage) to both the presence of weak instruments and the estimation of regressors and instruments. In all cases, the coverage rate of LM-based confidence intervals is closer to 95% than that of Wald-based intervals. If, in addition, the Engle-Ng estimator is used, coverage is almost exact. These should therefore be preferred in empirical work.

Table 5 provides details on the nonparametric estimators used in the simulation. We report the mean bandwidth constant, the lag length selected, the sum of the first 10 squared autocorrelation coefficients of the variance residuals, and the median constant and slope coefficient from the regression of the true instrument and regressor on a constant and the nonparametric estimates. The $R^2$ from these regressions has already been reported in tables 3 and 4.

The BIC-type criterion seems to overpenalize the number of lags, as it always chooses a single lag for all kernel estimates. However, it does suggest that some oversmoothing relative to the i.i.d. normal case is typically warranted (since in that case, the optimal bandwidth constant is 1.06). This is not surprising and is usually the case for dependent data. The criterion also seems to penalize heavily the number of bins in the Engle-Ng estimator, as the mean number of bins is not much above 2. However, it frequently chooses more than one lag.

The main feature of table 5 however is the tight relation between the bias of the instrument estimates and the behavior of the resulting IV estimator relative to the infeasible estimator. In cases where the IV estimator with estimated regressor and instrument performs poorly (experiments 2, 5, and 6 for both estimators, and experiment 4 for the kernel only), the median slope parameter from the instrument

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Estimation</th>
<th>Variance Estimation</th>
<th>Sum of 10 Squared Autocorrelations</th>
<th>Instrument Regression</th>
<th>Regressor Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bandwidth/Lag Length</td>
<td>Bandwidth/Bins/Lag Length</td>
<td>Constant/Slope</td>
<td>Constant/Slope</td>
<td></td>
</tr>
<tr>
<td>Kernel</td>
<td>1.99/1.00</td>
<td>1.84/1.00</td>
<td>2.61</td>
<td>0.001/0.532</td>
<td>0.000/1.000</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>1.99/1.00</td>
<td>2.01/1.47</td>
<td>2.73</td>
<td>0.000/0.991</td>
<td>0.000/1.023</td>
</tr>
<tr>
<td>Kernel</td>
<td>1.97/1.00</td>
<td>1.80/1.00</td>
<td>0.05</td>
<td>0.001/0.252</td>
<td>0.000/0.990</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>1.97/1.00</td>
<td>2.00/1.61</td>
<td>0.14</td>
<td>0.001/0.447</td>
<td>0.000/1.040</td>
</tr>
<tr>
<td>Kernel</td>
<td>1.82/1.00</td>
<td>1.32/1.00</td>
<td>1.81</td>
<td>0.000/1.141</td>
<td>0.000/1.013</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>1.82/1.00</td>
<td>2.40/1.09</td>
<td>2.03</td>
<td>0.000/1.899</td>
<td>0.000/1.041</td>
</tr>
<tr>
<td>Kernel</td>
<td>1.95/1.00</td>
<td>1.58/1.00</td>
<td>1.28</td>
<td>0.001/0.300</td>
<td>0.000/0.997</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>1.95/1.00</td>
<td>2.06/1.42</td>
<td>0.85</td>
<td>-0.000/0.879</td>
<td>0.000/1.027</td>
</tr>
<tr>
<td>Kernel</td>
<td>1.66/1.00</td>
<td>1.40/1.00</td>
<td>0.06</td>
<td>0.001/0.065</td>
<td>0.000/0.986</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>1.66/1.00</td>
<td>2.04/1.66</td>
<td>0.15</td>
<td>0.001/0.248</td>
<td>0.000/1.057</td>
</tr>
<tr>
<td>Kernel</td>
<td>1.05/1.86</td>
<td>0.69/1.82</td>
<td>0.51</td>
<td>0.005/0.015</td>
<td>0.007/0.015</td>
</tr>
<tr>
<td>Engle-Ng</td>
<td>1.05/1.86</td>
<td>3.58/1.39</td>
<td>1.53</td>
<td>0.001/0.078</td>
<td>0.007/0.018</td>
</tr>
</tbody>
</table>
regression is always less than 0.5, suggesting a severe bias of the nonparametric estimator. This result is akin to the typical result in semiparametric estimation that it is preferable to undersmooth the nonparametric component so as to reduce bias. The averaging in the second step mitigates the higher variance that this undersmoothing typically entails, though it does not eliminate bias.

VI. Empirical Results

In this section, we analyze our three financial data sets to seek evidence of a risk-return tradeoff. To reiterate, the series are monthly returns on the S&P 500 index, monthly returns on the yen-dollar spot rate, and quarterly excess holding yield between 6-month and 3-month Treasury bills. For each series, we postulate a model of the form

\[ y_t = \gamma + \delta \sigma^2_t + e_t \]

with \( \sigma^2_t = E[(y_t - E[y_t|\mathcal{F}_{t-1}])^2|\mathcal{F}_{t-1}] \) where \( \mathcal{F}_{t-1} \) are lagged values of \( y_t \). For all three series, the conditional variance was estimated using either the kernel or Engle-Ng estimator described above with the data-based selection of the tuning parameters. For comparison, we also report the results from a GARCH-M(1, 1) model estimated using Gaussian quasi-maximum likelihood.

The convergence to normality shown in the simulation might suggest that the use of higher-frequency data is greatly desirable, as it would increase the sample size, but higher frequency would also lead to a more persistent conditional variance and hence a weaker instrument. The impact of this choice on the behavior of the IV estimator and its related statistics is therefore ambiguous. As discussed already, another potential use of high-frequency data (not pursued here) is to get better estimates of low-frequency volatility.

The estimation results are presented in table 6. In addition to the point estimates and their robust (White) standard errors, we present Wald-based and LM-based 95% confidence intervals for the coefficient on the risk variable (\( \delta \)), the \( R^2 \) in a regression of \( \hat{\sigma}^2_t \) on \( \hat{\sigma}^2_t \) and a constant, and the values of the tuning parameters used to construct the nonparametric estimates. The LM confidence intervals were computed by numerically inverting the LM statistic using a grid of 20,000 equispaced points between −1000 and 1000. For this reason, the infinite or very large confidence intervals are truncated at these two endpoints.

The tradeoff between risk and return has been extensively studied for stocks, with conflicting results. For example, French, Schwert, and Stambaugh (1987) find a positive relation between returns and the conditional variance,
whereas Glosten, Jagannathan, and Runkle (1993) find a negative relationship using a modified GARCH-M methodology. This conflicting evidence is not surprising in view of the results obtained by Backus, Gregory, and Zin (1989) and Backus and Gregory (1993). Using a general-equilibrium setting, they provide simulation evidence that the relationship between expected returns and the variance of returns can go in either direction, depending on specification. Further doubt on the validity of the linearity assumption is provided in Linton and Perron (2003) using nonparametric methods.

Our results suggest that no significant risk premium exists in stock returns, using any of the three methods. However, the main feature of the results is the wider confidence intervals obtained using the LM principle. Wald confidence intervals understate the uncertainty of the estimated parameters; the differences are not dramatic, however. The results are also similar to those obtained from the GARCH-M(1, 1) model.

The results for the yen-dollar returns are presented next, with all point estimates negative. In the case of the kernel estimator, this finding is actually significantly different from...
0. The relationship for this series appears to be the least identified, as all estimators have large standard errors (and the first-stage $R^2$ is very low). Both the Wald and LM confidence intervals are quite wide, reflecting poor identification of the model. The LM interval with the Engle-Ng estimator is even unbounded in this case.

Finally, the results of the estimation for excess holding yields present a similar picture. All point estimates are positive, with the GARCH-M result being significantly different from 0. This conclusion is the same as that of Engle, Lilien, and Robbins [who used a restricted ARCH-M(4) structure]. For the kernel estimator, the effect is almost significant at the 5% level. Once again, the LM intervals are much wider than their Wald counterparts.

Figure 7 presents a time plot of the estimated conditional variance for all three series. Except for the excess holding yield, the Engle-Ng and GARCH-M models offer very similar pictures. On the other hand, the kernel estimates are much more volatile (not surprisingly, given that they do not have an autoregressive structure) over time. The results for the excess holding yield might seem strange at first sight, in that the GARCH-M gives such a different picture (especially around the Volker experiment of 1979–1982). The reason lies in the bandwidth choice for the estimation of the conditional mean of this series. The mean is estimated with a very small bandwidth (constant is 0.28) thus implying little smoothing of neighboring observations, and as a result, the residuals are much smaller than with GARCH-M (and hence have smaller variance).

VII. Conclusion

This paper follows several others in showing that inference using instrumental variables is greatly affected by a low correlation between the instruments and the explanatory variables. It extends the current literature to linear semiparametric models with nonparametrically estimated regressors and instruments and to cases with higher-order dependence. The analysis shows that the limit theory is similar to that currently available in the literature.

Simulation evidence reveals that the additional step of estimating both the regressors and the instruments may lead to a loss in the quality of asymptotic approximations. Using a semiparametric estimator proposed by Engle and Ng (1993) and carrying out inference using Lagrange multiplier procedures allows for inference that is more robust than the alternatives considered here.

Empirical application to three financial series suggests that conclusions may hinge on the use of appropriate confidence intervals. Using the appropriate LM confidence intervals and the semiparametric estimator of the conditional variance leads us to conclude that none of the series considered includes a statistically significant risk premium. This differs in some cases from inference based on the usual Wald confidence intervals and on a parametric GARCH-M model. However, because of the wide confidence intervals, the results are also consistent with the presence of large risk premia. The data is simply not informative enough to precisely estimate the relationship between risk and returns.

Further work on this problem is clearly warranted. In particular, other more commonly used estimators such as maximum likelihood are likely to face similar problems to the IV estimator analyzed here. This analysis could follow the methodology developed in Stock and Wright (2000) for GMM estimators. Finally, a critical avenue for future research is the development of techniques to diagnose cases where weak identification hinders inference using usual methods. Recent testing procedures along these lines have been suggested by Arellano, Hansen, and Sentana (1999), Wright (2000), and Hahn and Hausman (2002).

REFERENCES


Arellano, Manuel, Lars P. Hansen, and Enrique Sentana, “Underidentification?” CEMFI manuscript (July 1999).


Chao, John, and Norman R. Swanson, “Bias and MSE of the IV Estimator under Weak Identification,” University of Maryland and Purdue University manuscript (August 2001).


APPENDIX
Proofs

1. Preliminary results

Before proving the results in the paper, we will collect the required preliminaries in the following lemma.

Lemma A.1. Suppose the conditions of theorem 4.1 are satisfied. Then the following hold:

1. \( \frac{1}{\sqrt{n}} \left( \hat{Z}'M_x \hat{Y} - \frac{1}{\sqrt{n}} (Z'M_x Y) \right) \rightarrow o_p(1) \).

2. \( \frac{1}{\sqrt{n}} \left( \hat{Z}'M_x (Z - Y)\hat{\delta} \right) = \frac{1}{\sqrt{n}} \left( Z'M_x (Z - Y)\hat{\delta} \right) + o_p(1) \).

3. \( \frac{1}{\sqrt{n}} \left( \hat{Z}'M_x e \right) = \frac{1}{\sqrt{n}} (Z'M_x e) + o_p(1) \).

4. \( \frac{1}{n} \left( \hat{Z}'M_x \hat{Z} \right) \rightarrow \frac{1}{n} \left( Z'M_x Z \right) + o_p(1) \).

5. \( \frac{1}{\sqrt{n}} X' \hat{Y} \rightarrow \frac{1}{\sqrt{n}} X'Y + o_p(1) \).

6. \( \frac{1}{\sqrt{n}} \left( \hat{Z}'M_x (Y - \hat{Y} \hat{\delta}) \right) \rightarrow 0 \).

7. \( \frac{1}{\sqrt{n}} \left( \hat{Z}'M_x a \right) = \Psi_{a'} + o_p(1) \).

Proof. To prove the first result, note that

\[
\frac{1}{\sqrt{n}} \left( \hat{Z}'M_x \hat{Y} \right) = \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x (\hat{Y} - Y)
\]

\[+ \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x Y + \frac{1}{\sqrt{n}} Z'M_x (\hat{Y} - Y)
\]

\[+ \frac{1}{\sqrt{n}} Z'M_x Y \]

\[= \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x (\hat{Y} - Y) + \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x Y + \frac{1}{\sqrt{n}} Z'M_x (\hat{Y} - Y)
\]

\[+ \frac{1}{\sqrt{n}} Z'M_x Y \]

\[= \frac{1}{\sqrt{n}} Z'M_x Y + \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x (\hat{Y} - Y)
\]

\[+ \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x Y + \frac{1}{\sqrt{n}} Z'M_x (\hat{Y} - Y)
\]

\[+ \frac{1}{\sqrt{n}} Z'M_x Y \]

\[= \frac{1}{\sqrt{n}} Z'M_x Y + \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x (\hat{Y} - Y)
\]

\[+ \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x Y + \frac{1}{\sqrt{n}} Z'M_x (\hat{Y} - Y)
\]

\[+ \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x E(Z)
\]

\[= \frac{1}{\sqrt{n}} Z'M_x Y + A_1 + A_2 + A_3 + A_4 \]

\[+ \frac{1}{\sqrt{n}} \left( \hat{Z}'Z \right) M_x E(Z)
\]
We will next bound each of the $A_i$, $i = 1, \ldots, 4$. Let $|A|$ be the matrix norm of $A$. First,

$$|A| = \frac{1}{\sqrt{n}} (\hat{Z} - Z)'M_i(\hat{Y} - Y)$$

$$\leq |\hat{Z} - Z| \frac{1}{\sqrt{n}} M_i(\hat{Y} - Y)$$

$$= o_p(1)$$

by assumptions 1 and 2. Next,

$$|A| = \frac{1}{\sqrt{n}} (\hat{Z} - Z)'M_i(Y - Z)$$

$$\leq |\hat{Z} - Z| \frac{1}{\sqrt{n}} M_i(Y - Z)$$

$$= o_p(1)$$

by assumption 2 and because the quantity inside the second norm will be $O_p(1)$. The third term is

$$|A| = \frac{1}{\sqrt{n}} (\hat{Z} - Z)'M_i[Z - E(Z)]$$

$$\leq |\hat{Z} - Z| \frac{1}{\sqrt{n}} M_i[1 - E(Z)]$$

$$= o_p(1),$$

again by assumption 2 and because the quantity inside the second norm is $O_p(1)$. Finally, the fourth term can be bounded as

$$|A| = \frac{1}{\sqrt{n}} (\hat{Z} - Z)'M_iE(Z)$$

$$\leq \frac{1}{\sqrt{n}} (\hat{Z} - Z)'M_i|E(Z)|$$

$$= o_p(1)$$

as $\hat{Z} \xrightarrow{p} Z$, and $|E(Z)| < \infty$ with probability 1, because $Z_i < \infty$ for all $i$.

Thus,

$$\frac{1}{\sqrt{n}} \hat{Z}'M_2\hat{Y} = \frac{1}{\sqrt{n}} Z'M_2Y + o_p(1),$$

as required.

The second result is obtained as

$$\frac{1}{\sqrt{n}} [\hat{Z}'M_4(Z - Y)\hat{\delta}] = \frac{1}{\sqrt{n}} [(\hat{Z} - Z)'M_4(Z - Y)\delta] + \frac{1}{\sqrt{n}} [Z'M_4(Z - Y)\delta]$$

$$= \frac{1}{\sqrt{n}} [Z'M_4(Z - Y)\delta] + o_p(1),$$

where the last line follows from

$$\left\{ \frac{1}{\sqrt{n}} (\hat{Z} - Z)'M_4(Z - Y)\delta \right\} \leq |\hat{Z} - Z| \frac{1}{\sqrt{n}} M_4(Z - Y)\delta$$

$$= o_p(1)$$

The third result follows from

$$\frac{1}{\sqrt{n}} \hat{Z}'M_4e = \frac{1}{\sqrt{n}} [(\hat{Z} - Z)'M_4e + \frac{1}{\sqrt{n}} Z'M_4e]$$

and noting that the first term can be bounded by

$$\left\{ \frac{1}{\sqrt{n}} [(\hat{Z} - Z)'M_4e] \right\} \leq |\hat{Z} - Z| \frac{1}{\sqrt{n}} M_4e$$

$$= o_p(1) \cdot O_p(1)$$

$$= o_p(1)$$

by assumptions 2 and 4.

The fourth result is proven by rewriting the left-hand side as

$$\frac{1}{\sqrt{n}} \hat{Z}'M_2\hat{Z} = \frac{1}{n} (\hat{Z} - Z)'M_2\hat{Z} + \frac{1}{n} Z'M_2\hat{Z}$$

$$= \frac{1}{n} Z'M_2Z + \frac{1}{n} (\hat{Z} - Z)'M_2(\hat{Z} - Z)$$

$$+ \frac{1}{n} (\hat{Z} - Z)'M_2Z + \frac{1}{n} Z'M_2(\hat{Z} - Z)$$

$$= \frac{1}{n} Z'M_2Z + B_1 + B_2 + B_3,$$

where $B_j, j = 1, 2$, is bounded in turn by an $o_p(1)$ term. For $B_1$, we do so as

$$|B_1| = \left\{ \frac{1}{n} (\hat{Z} - Z)'M_2\hat{Z} \right\}$$

$$\leq \frac{1}{n} \left( \hat{Z} - Z \right)'M_2 \left\{ \frac{1}{n} \hat{Z}' - \hat{Z} \right\}$$

$$= o_p(1)$$

by assumption 2, where $\hat{\iota}$ is a vector of ones. The second term is bounded as

$$|B_2| = \left\{ \frac{1}{n} (\hat{Z} - Z)'M_2\hat{Z} \right\}$$

$$\leq \frac{1}{n} \left( \hat{Z} - Z \right)'M_2 \left\{ \frac{1}{n} \hat{Z}' - \hat{Z} \right\}$$

$$= o_p(1) \cdot O_p(1)$$

$$= o_p(1)$$

by assumption 2. The fourth result follows.

The fifth result is obtained as

$$\frac{1}{n} X'\hat{Y} = \frac{1}{n} X'Y + \frac{1}{\sqrt{n}} X'(\hat{Y} - Y)$$

$$= \frac{1}{\sqrt{n}} X'Y + o_p(1)$$

by assumption 1.

The sixth result is obtained from the decomposition

$$\frac{1}{\sqrt{n}} [\hat{Z}'M_4(\hat{Y} - \hat{\delta})]$$

$$= \frac{1}{\sqrt{n}} [(\hat{Z} - Z)'M_4(Y - \hat{\delta}) + \frac{1}{\sqrt{n}} Z'M_4(Y - \hat{\delta})]$$
where the last line follows from assumption 1 and \( E(Z) < \infty \).

Finally, the last result is obtained by rewriting the left-hand side as

\[
\frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} [\hat{Z} - Z] (Y - \hat{Y} - \delta) + \frac{1}{\sqrt{n}} [Z - E(Z)] (Y - \hat{Y} - \delta) + \frac{1}{\sqrt{n}} E(Z) M_2 (Y - \hat{Y} - \delta)
\]

and using results 2, 3, and 6 of the lemma.

2. Proof of Theorem 4.1

The instrumental variable estimator of \( \delta \) is

\[
\hat{\delta} - \delta = (\hat{Z}' M_2 \hat{Y})^{-1} \hat{Z}' M_2 u.
\]

To derive the asymptotic distribution, we use the first result of the lemma to obtain

\[
\frac{1}{\sqrt{n}} Z' M_2 \hat{Y} = \frac{1}{\sqrt{n}} Z' M_2 Y + o_p(1)
\]

where the factor in parentheses is derived from

\[
\frac{1}{\sqrt{n}} X' Y = \frac{1}{\sqrt{n}} X' (Z \Pi + X \Gamma + V)
\]

\[
= \frac{1}{\sqrt{n}} X' Z G + \frac{1}{n} X' X H + \frac{1}{\sqrt{n}} X' V \Sigma_G + \Sigma_H + \Psi_H^V
\]

by assumption.

3. Proof of Proposition 4.2

From the proof of theorem 4.1, \( n^{-1/2} \hat{Z} M_2 \hat{Y} \xrightarrow{a.s.} \sigma^2 [\lambda + z] \). The only part that remains to derive is the limiting behavior of \( \lim_{n \to \infty} \text{var} \left( \frac{1}{\sqrt{n}} Z' M_2 \hat{Y} \right) \). The residual orthogonal to \( X \), \( M_2 u \), can be written as \( M_2 (y - \hat{Y}) = M_2 u - M_2 \hat{Y} (\hat{\delta} - \delta) \), and the term of interest is therefore

\[
\lim_{n \to \infty} \text{var} \left( \frac{1}{\sqrt{n}} Z' M_2 \hat{Y} \right)
\]

\[
= \text{var} \left( \frac{1}{\sqrt{n}} Z' M_2 [M_2 u - M_2 \hat{Y} (\hat{\delta} - \delta)] \right)
\]

\[
= \text{var} \left( \frac{1}{\sqrt{n}} [Z' M_2 u - Z' M_2 \hat{Y} (\hat{\delta} - \delta)] \right)
\]

\[
= \text{var} \left( \frac{1}{\sqrt{n}} \left( Z' M_2 u - Z' M_2 \hat{Y} (\hat{\delta} - \delta) + o_p(1) \right) \right)
\]

\[
= \text{var} \left( \frac{1}{\sqrt{n}} \left( Z' u - Z' Z' \Pi (\hat{\delta} - \delta) - Z' V (\hat{\delta} - \delta) \right) \right)
\]

\[
+ o_p(1)
\]

\[
= \lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{i} Z_i u_i Z_i^\top
\]

\[
+ \frac{1}{\sqrt{n}} \sum_{i} Z_i Z_i^\top \Pi (\hat{\delta} - \delta) (\hat{\delta} - \delta) \Pi Z_i^\top
\]

\[
+ \frac{1}{\sqrt{n}} \sum_{i} Z_i V (\hat{\delta} - \delta) (\hat{\delta} - \delta) V Z_i^\top
\]

\[
- \frac{1}{\sqrt{n}} \sum_{i} Z_i u_i (\hat{\delta} - \delta) \Pi Z_i^\top
\]

\[
- \frac{1}{\sqrt{n}} \sum_{i} Z_i u_i (\hat{\delta} - \delta) V Z_i^\top
\]

\[
= \sigma_{zz} + C_1 + C_2 - C_3 - C_4 - C_5 + C_6 + C_7.
\]
The second term is

$$C_1 = \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' \Pi (\hat{\delta} - \delta)(\hat{\delta} - \delta)' \Pi' Z_i$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' G \Xi' G' Z_i$$

$$= o_p(1),$$

and the third one is

$$C_2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' V (\hat{\delta} - \delta)(\hat{\delta} - \delta)' V Z_i$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' V \Xi' V Z_i$$

$$= o_p(1).$$

The next term is

$$C_3 = \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' u(\hat{\delta} - \delta)' \Pi' Z_i$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' u \Xi' G' Z_i$$

$$= o_p(1),$$

the fifth term in the sum is

$$C_5 = \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' u(\hat{\delta} - \delta)' V Z_i$$

$$= \lim_{n \to \infty} \frac{1}{n^2} \sum_{i} Z_i' u \Xi' V Z_i$$

$$= o_p(1).$$

and finally

$$C_1 = \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' \Pi (\hat{\delta} - \delta)(\hat{\delta} - \delta)' V Z_i$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i} Z_i' G \Xi' V Z_i$$

$$= o_p(1).$$

4. Proof of Proposition 4.3

By result 7 of the lemma, \( \sqrt{n} \mathbf{d} \to \Psi_{Zu} \approx N(0, \sigma_{Zu}) \) under the null hypothesis, and \( \mathbf{d}_{Zu} \to \sigma_{Zu} \). Standard arguments show the desired result, \( n \mathbf{g} \mathbf{d}_{Zu} \to \chi^2(k_2) \).

5. Proof of Proposition 4.4

The estimator of \( \delta^* \) is defined as

$$\delta^* = (\hat{Z}' M \hat{Z})^{-1} \hat{Z}' M \delta + \hat{Z} \hat{\beta}^* + u + v(\hat{\delta} - \delta_0)$$

so that

$$\sqrt{n} (\hat{\delta} - \delta^*) = \left( \frac{\hat{Z}' M \hat{Z}}{n} \right)^{-1} \hat{Z}' M u + \left( \frac{\hat{Z}' M \hat{Z}}{n} \right)^{-1} \hat{Z}' M \delta \hat{Z} (\hat{\delta} - \delta_0)$$

under the null hypothesis. By results 4 and 7 of the lemma, \( \sqrt{n} (\hat{\delta} - \delta^*) \to N(0, \Sigma_{Zu}(\sigma_{Zu})^2) \).

Define \( \Omega = [ (\hat{Z}' M \hat{Z})^{-1} \mathbf{d}_{Zu} (\hat{Z}' M \hat{Z})^{-1}] \). The robust AR statistic is

$$AR = n(y - \hat{Y}_{\delta_0})' M_j Z_j (\hat{Z}' M_j)\Omega (\hat{Z}' M_j)' M_j (y - \hat{Y}_{\delta_0})$$

$$= n(y - \hat{Y}_{\delta_0})' M_j \hat{Z} \mathbf{d}_{Zu}(\hat{Z}' M_j)' M_j (y - \hat{Y}_{\delta_0})$$

$$= LM$$

after simplification.