COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1151

SECOND ORDER APPROXIMATION IN A LINEAR REGRESSION WITH HETEROSKEDASTICITY FOR UNKNOWN FORM

Oliver Linton

May 1997

SECOND ORDER APPROXIMATION IN A LINEAR REGRESSION WITH HETEROSKEDASTICITY OF UNKNOWN FORM

OLIVER B. LINTON

Cowles Foundation for Research in Economics, Yale University, CT 06520.

Key words and phrases: Semiparametric estimation, GLS, heteroskedasticity, local linear regression, asymptotic expansions.

JEL Classification: 210.

ABSTRACT

We develop stochastic expansions with remainder $o_P(n^{-2\mu})$, where $0 < \mu < 1/2$, for a standardised semiparametric GLS estimator, a standard error, and a studentized statistic, in the linear regression model with heteroskedasticity of unknown form. We calculate the second moments of the truncated expansion, and use these approximations to compare two competing estimators and to define a method of bandwidth choice.

1. INTRODUCTION

Heteroskedasticity is frequently found in residuals from estimated econometric models, in both cross-sectional and time series datasets. There are two predominant strategies for dealing with this problem. In the first approach, one specifies a parametric model for the heteroskedasticity, estimates all parameters by maximum likelihood, and conducts testing under this assumption. In the second, one estimates under the presumption of homoskedasticity, but employs standard errors that consistently estimate the relevant sampling variability that pertains when heteroskedasticity is in fact present. The latter methodology originates with Eicker (1968) and White (1980). If the parametric model is correctly specified, the first method provides optimal estimation and testing. However, it is not robust: in particular, test statistics with asymptotically incorrect size will result from misspecification of the second moments. The second method provides valid inference regardless of the form of the heteroskedasticity, but at the cost of a loss of efficiency and local power relative to an approach based on a correctly specified parametric model.

An alternative that appears to offer the advantages of both these procedures is to use semiparametric methods to account for the heteroskedasticity. In some cases, this approach involves no asymptotic efficiency loss relative to the MLE (and hence test statistics with maximum local asymptotic power result) from a correctly specified parametric model – whatever that may be. In the context of a linear regression model with heteroskedasticity of unknown form, Carroll (1982) and Robinson (1987) use a feasible GLS procedure in which the weights are nonparametric estimators of the conditional variance $\sigma^2(\bullet)$ to estimate the mean parameters β. These estimators of β are asymptotically equivalent to the infeasible GLS estimator that uses the unknown variances to weight the observations. Therefore, when the errors are normal, they are efficient. This would suggest that the semiparametric approach dominates the other two. However, this judgement is based exclusively on first order asymptotic approximations, whose validity is not entirely supported by the monte carlo evidence presented in Hsieh and Manski (1987), Stock (1989) and Stoker (1993) for related situations. Rothenberg (1984b) shows that the magnitude of the second order corrections for parametric GLS estimators generally increases with the number of nuisance parameters one has estimated to construct the GLS weights. The semiparametric procedure implicitly estimates an infinite number of nuisance parameters and therefore may be expected to have rather poor small sample properties. A second problem is that the first order theory does not reflect the bandwidth $h(n)$ that determines the amount of smoothing employed (i.e. the number of nuisance parameters being fitted) in the nonparametric procedure. This quantity can materially affect the magnitude of estimators and test statistics.

We propose using higher order asymptotic expansions to address these problems. This methodology has a long tradition of successful application in the econometric literature, starting with Nagar (1959), see *inter alia* Sargan (1975,1976), Phillips (1977ab), and Anderson and Sawa (1979), see Rothenberg (1984a) for a review. We derive an $o_P(n^{-2\mu})$ stochastic expansion, where $0 < \mu < 1/2$, for two standardised competing semiparametric GLS estimators, a standard error and a Wald statistic. We calculate approximations to the first two moments of the truncated expansions; these depend on the bandwidth $h(n)$ used in the kernel estimation of $\sigma^2(\bullet)$. We use our approximations to calibrate the likely small sample cost of this estimation strategy in a number of examples. We also use this formula to calculate an optimal bandwidth, which can be used as a method of bandwidth choice. Our work is related to that of Carroll and Härdle (1989), Cavanagh (1989), Härdle, Hart, Marron and Tsybakov (1992) and Linton (1993) for related semiparametric situations.

In section 2 we describe the sampling scheme we examine, while in section 3 we describe the estimators and test statistics. In section 4 we develop the asymptotic expansion, and give formulae for the second order approximations to the MSE of the various quantities. In section 5 we discuss optimality and bandwidth choice, while section 6 contains the results of a small simulation experiment. Section 7 concludes, while the Appendix contains an outline proof.

A word on notation. We use $\Rightarrow$ to denote convergence in distribution, $\xrightarrow{P}$ means convergence in probability, while the symbol $\approx$ denotes asymptotic equivalence in probability, all holding as $n \to \infty$.

2. SAMPLING SCHEME

We examine the following linear regression model:

$$y_i = \beta^T x_i + u_i \; ; \; u_i = \epsilon_i \sigma_i, \; i = 1, 2, .., n, \tag{1}$$

where ϵ_i is iid zero mean, variance one, skewness κ_3 and kurtosis κ_4, while the fixed design regressors $\{x_i\}_{i=1}^n$ have support a bounded domain $\Upsilon \subseteq \mathbb{R}^P$. The conditional variance $\sigma_i^2 = \sigma^2(x_i) \geq \underline{\sigma} > 0$, where $\sigma^2(\bullet)$ is of unknown functional form, possesses at least three continuous partial derivatives in each direction. We further require of the design that there be a positive differentiable density f, such that for any bounded continuous function $v(\bullet)$,

$$n^{-1} \sum_{i=1}^{n} v(x_i) \to \int v(x) f(x) dx. \tag{2}$$

We do not exclude stochastic regressors from our treatment. If x_i were iid with density f, then (2) holds with probability 1. In this case, the approximations developed in this paper also hold with probability 1.

Carroll (1982) and Robinson (1987) both consider sampling schemes where $(x^T, y)^T$ are iid; they examine estimator performance unconditionally. In our case, however, the marginal distribution of the regressors contains no information about the parameter β. Therefore, conditioning on them is more in line with conventional statistical practice – see Cox and Hinkley (1974, p33).

3. ESTIMATION

3.1 Semiparametric GLS

We examine the behavior of the following feasible GLS estimators:

$$\widehat{\beta} = [\sum_{i=1}^{n} x_i x_i^T \widehat{\sigma}_i^{-2}]^{-1}[\sum_{i=1}^{n} x_i y_i \widehat{\sigma}_i^{-2}] \ ; \ \widetilde{\beta} = [\sum_{i=1}^{n} x_i x_i^T \widetilde{\sigma}_i^{-2}]^{-1}[\sum_{i=1}^{n} x_i y_i \widetilde{\sigma}_i^{-2}], \tag{3}$$

where $\widehat{\sigma}_i^2 = \sum_{j \neq i} w_{ij} \widehat{u}_j^2$ and $\widetilde{\sigma}_i^2 = \sum_{j \neq i} w_{ij} y_j^2 - [\sum_{j \neq i} w_{ij} y_j]^2$ are nonparametric estimators of σ_i^2 of the "leave one out" type, where $\{w_{ij}\}_{j=1}^n$ is a sequence of weights defined below, while $\{\widehat{u}_j\}_{j=1}^n$ are the least squares residuals. Carroll (1982) and Robinson (1987) both establish the first order asymptotic theory for $\widehat{\beta}$, although Robinson (1987) also indicates the feasibility of $\widetilde{\beta}$.

We now turn to the choice of nonparametric weights $\{w_{ij}\}$. Carroll (1982) employed a Nadaraya-Watson kernel estimator, while Robinson (1987) used nearest neighbors. See Härdle and Linton (1993) for a comparison of these and other nonparametric regression smoothers. We use the fixed window local linear regression method suggested in Stone (1977) and further examined in Fan (1992). This procedure has several advantages: there are no boundary effects, and the interior pointwise bias does not, asymptotically, depend on the design density (Fan (1992) calls this latter property design adaptation). This method is motivated by the following argument.

A smooth regression function $g(x)$ can be expanded in a Taylor series, so that for x_j in a neighborhood $\mathcal{N}(x_i)$ of $x_i = (x_{i1}, .., x_{iP})^T$, $g(x_j) \approx \tau_i^T z_{ij}$, where $z_{ij} = (1, x_{j1} - x_{i1}, .., x_{jP} - x_{iP})^T$ and $\tau_i = (g(x_i), \partial g/\partial x_{1i}, .., \partial g/\partial x_P)^T$. Therefore, in $\mathcal{N}(x_i)$ we have an approximate linear regression in which the explanatory variables are z_{ij}, and τ_i are 'hyper-parameters'. Therefore, we take w_{ij} to be the $(1, j)'th$ element of the $P + 1$ by n regression weighting matrix

$$(Z_i^T K_i Z_i)^{-1} Z_i^T K_i, \tag{4}$$

where $Z_i = (z_{i1}, .., z_{in})^T$, while K_i is a diagonal matrix with j'th element $k((x_j - x_i)/h)$ and i'th equal to zero. Here, $k(\bullet)$ is a $P-$dimensional probability density function with bounded support and one continuous partial derivative in each direction, while $h(n)$ is a scalar bandwidth satisfying $h \to 0$ and $nh^P \to \infty$. The rate at which h converges to zero is determined in the sequel.

3.2. Standardised Quantities

Both $\widehat{\beta}$ and $\widetilde{\beta}$ approximate, provided only $h \to 0$ and $nh^P \to 0$, the infeasible GLS estimator

$$\overline{\beta} = [\sum_{i=1}^{n} x_i x_i^T \sigma_i^{-2}]^{-1} [\sum_{i=1}^{n} x_i y_i \sigma_i^{-2}], \tag{5}$$

which satisfies $\sqrt{n}(\overline{\beta} - \beta) \Rightarrow N(0, M_n^{-1})$, where $M_n = n^{-1} \sum_{i=1}^{n} x_i x_i^T \sigma_i^{-2}$. Therefore, the common asymptotic distribution of the semiparametric estimators does not depend on the bandwidth. We examine the higher order properties of $\widehat{\beta}$ and $\widetilde{\beta}$, which do depend on h. For convenience, we work with the scalar standardised quantities $T = \sqrt{n} c^T (\widehat{\beta} - \beta)/s$ and $T' = \sqrt{n} c^T (\widetilde{\beta} - \beta)/s$, where $s^2 = c^T M_n^{-1} c$ and c is any P by 1 vector. We also consider a standardised standard error $S = \sqrt{n}(\widehat{s} - s)/s$, where $\widehat{s}^2 = c^T \widehat{M}_n^{-1} c$ with $\widehat{M}_n = n^{-1} \sum_{i=1}^{n} x_i x_i^T \widehat{\sigma}_i^{-2}$. Finally, we also consider a Wald statistic $W = \sqrt{n} c^T (\widehat{\beta} - \beta_0)/\widehat{s}$ that can be used to test the hypothesis H_0: $c^T \beta = c^T \beta_0$. An important special case is where $c = (0, .., 0, 1, 0, ., 0)^T$ upon which W is the standard t-test for the significance of the corresponding regressor.

4. SECOND ORDER APPROXIMATIONS

We give second order stochastic expansions for T, T', S, and W and further approximate the second moments of the truncated expansions. In section 4.1 we derive the stochastic expansion for T, while in section 4.2 we present formulae for its asymptotic moments and justify their interpretation. In section 4.3 we compare the approximations for T and T', while in section 4.4 we give the expansions and asymptotic moments for S and W.

4.1 Stochastic Expansion

By a geometric series expansion, T can be written as

$$T = s^{-1} c^T M_n^{-1} \{ X_N - \frac{X_D M_n^{-1} X_N}{\sqrt{n}} + \frac{X_D M_n^{-1} X_D M_n^{-1} X_N}{n} \} + \frac{R}{n\sqrt{n}} \equiv T^* + R_T^*, \quad (6)$$

where $X_N = n^{-1/2} \sum_{i=1}^n x_i u_i \widehat{\sigma}_i^{-2}$ and $X_D = n^{-1/2} \sum_{i=1}^n x_i x_i^T (\widehat{\sigma}_i^{-2} - \sigma_i^{-2})$, while

$$R = c^T \widehat{M}_n^{-1} X_D M_n^{-1} X_D M_n^{-1} X_D M_n^{-1} X_N.$$

We write

$$\widehat{\sigma}_i^{-2} - \sigma_i^{-2} = \widehat{\sigma}_i^{-2} - \widehat{\sigma}_i^{*-2} + \widehat{\sigma}_i^{*-2} - \overline{\sigma}_i^{-2} + \overline{\sigma}_i^{-2} - \sigma_i^{-2},$$

where $\widehat{\sigma}_i^{*2} = \sum_{j \neq i} w_{ij} u_j^2$ and $\overline{\sigma}_i^2 = \sum_{j \neq i} w_{ij} \sigma_j^2$. We first drop $\widehat{\sigma}_i^{-2} - \widehat{\sigma}_i^{*-2}$: it is of smaller order in probability. Then we expand $\widehat{\sigma}_i^{*-2}$ about $\overline{\sigma}_i^{-2}$ and $\overline{\sigma}_i^{-2}$ about σ_i^{-2} to the third term, so that

$$\widehat{\sigma}_i^{*-2} - \overline{\sigma}_i^{-2} = -\overline{\sigma}_i^{-2}\{(\frac{\widehat{\sigma}_i^{*2}-\overline{\sigma}_i^2}{\overline{\sigma}_i^2}) - (\frac{\widehat{\sigma}_i^{*2}-\overline{\sigma}_i^2}{\overline{\sigma}_i^2})^2 + \widehat{\sigma}_i^{*-2}(\frac{\widehat{\sigma}_i^{*2}-\overline{\sigma}_i^2}{\overline{\sigma}_i^2})^3\}$$

$$\overline{\sigma}_i^{-2} - \sigma_i^{-2} = -\sigma_i^{-2}\{(\frac{\overline{\sigma}_i^2-\sigma_i^2}{\sigma_i^2}) - (\frac{\overline{\sigma}_i^2-\sigma_i^2}{\sigma_i^2})^2 + \overline{\sigma}_i^{-2}(\frac{\overline{\sigma}_i^2-\sigma_i^2}{\sigma_i^2})^3\}.$$

Then let ω_i^* and $\overline{\omega}_i$ be the first two terms in the respective expansions, and

$$X_N^* = n^{-1/2}\sum_{i=1}^n x_i u_i \sigma_i^{-2} + n^{-1/2}\sum_{i=1}^n x_i u_i(\omega_i^* + \overline{\omega}_i) \ ; \ X_D^* = n^{-1/2}\sum_{i=1}^n x_i x_i^T(\omega_i^* + \overline{\omega}_i).$$

Then let T^{**} be the corresponding truncation of T^* with X_N^* and X_D^* replacing X_N and X_D, and let R_T^{**} be the grand remainder that includes R_T^* as well as the remainders from replacing X_N and X_D by X_N^* and X_D^*.

4.2 Asymptotic Moments of T

In this section we compute approximations to the first two moments[1] of T^{**}. Let $MSE(h)$ denote n times the asymptotic mean squared error of $\widehat{\beta}$, i.e. $MSE(h) = E[T^{**2}]$. Then

$$MSE(h) \approx 1 + \{O(h^4) + O(n^{-1}h^{-P})\},$$

where the term in curly brackets we call the second order effect. The order of magnitude of the second order effect is minimised by setting h so that $h^4 \approx n^{-1}h^{-P}$, which requires $h(n) = O(n^{-\pi})$, where $\pi = 1/(P+4)$, and results in $MSE = 1 + O(n^{-2\mu})$ (in fact, $E[T^{**}] = O(n^{-1/2})$ and $Var[T^{**}] = 1 + O(n^{-2\mu})$), where $\mu = 2/(P+4)$ (when $P = 1$, $\pi = 1/5$ and $\mu = 2/5$). The second order effect on MSE is then $O(n^{-2\mu})$, is dominated by variance, and is strictly larger than the $O(n^{-1})$ effect typically found in parametric models − see Rothenberg (1984b).

[1]These are uniformly bounded under moment conditions on ε_i, see Linton (1993).

PROPOSITION 1A. *When* $h = O(n^{-\pi})$, $R_T^{**} = o_P(n^{-2\mu})$. *Furthermore,* $E[T^{**}] = O(n^{-1/2})$ *and*

$$Var[T^{**}] = 1 + h^4\mathcal{B} + n^{-1}h^{-P}\mathcal{V} + o(n^{-2\mu}), \tag{7}$$

where

$$\mathcal{B} = \frac{c^T M_n^{-1}[\Gamma_2 - \Gamma_1 M_n^{-1}\Gamma_1]M_n^{-1}c}{c^T M_n^{-1}c} \;;\; \mathcal{V} = (\kappa_3^2 + 2 + \kappa_4)\frac{c^T M_n^{-1}M_n^* M_n^{-1}c}{c^T M_n^{-1}c},$$

where $M_n^* = n^{-1}\sum_{i=1}^n x_i x_i^T \sigma_i^{-2}\rho_i$, *with* $\rho_i = nh^P \sum_{j\neq i} w_{ij}^2 = O(1)$ *uniformly in* i, *while* $\Gamma_1 = \Gamma_1(\sigma^2)$ *and* $\Gamma_2 = \Gamma_2(\sigma^2)$, *where for any function* g, $\Gamma_1(g) = n^{-1}\sum_{i=1}^n x_i x_i^T \widetilde{B}_i(g)\sigma_i^{-4}$ *and* $\Gamma_2(g) = n^{-1}\sum_{i=1}^n x_i x_i^T \widetilde{B}_i^2(g)\sigma_i^{-6}$, *where*

$$\widetilde{B}_i(g) = [\sum_{j\neq i} w_{ij}g(x_j) - g(x_i)]/h^2 = O(1).$$

REMARK. The quantity $c^T M_n^{-1}M_n^* M_n^{-1}c/c^T M_n^{-1}c$ depends on both the kernel k and on the design. It is positive, and $\mathcal{V} \geq 0$. Furthermore, $\mathcal{B} \geq 0$ by the Cauchy-Schwarz inequality. Therefore, the asymptotic (second order) variance of $\widehat{\beta}$ is not less than that of $\overline{\beta}$ – regardless of the error distribution. This contrasts with the results obtained by Carroll, Wu, and Ruppert (1988) for parametric GLS estimators in non-normal error situations. Their expansions indicate situations where a feasible (parametric) GLS estimator can have a lower asymptotic variance than GLS.

REMARK. We interpret the moments of T^{**} as approximations to the moments of T. When $\varepsilon_i \sim N(0,1)$, then $Sup_n E[T^2] < \infty$, and $E[T^2] = E[T^{**2}] + o(n^{-2\mu})$, see Carroll and Härdle (1989) and Rothenberg (1984b). However, when the original statistic does not possess moments, it is desirable to establish slightly stronger regularity on the remainder terms than merely $R_T^{**} = o_P(n^{-2\mu})$. If, for some positive constant δ,

$$\Pr[n^{2\mu} \log n |R_T^{**}| > \delta] = o(n^{-2\mu}), \tag{8}$$

then, following Sargan and Mikhail (1971) – see also Robinson (1988a) – the distribution of the truncated quantity agrees with that of the original statistic to order $n^{-2\mu}$. In this case, the asymptotic moments can be interpreted as the moments of a random variable whose distribution is close to that of the original quantity – see the discussion in Rothenberg (1984a) and Robinson (1988a). Condition (8) can be established under smoothness and moment conditions, see Linton (1992,1993).

By further asymptotic approximation

$$\widetilde{B}_i(\sigma^2) \approx \sum_{\alpha=1}^{P} \sum_{\delta=1}^{P} \theta_{\alpha\delta} \frac{\partial^2 \sigma^2}{\partial x_{i\alpha} \partial x_{i\delta}}(x_i), \tag{9}$$

where $\theta_{\alpha\delta}$ are constants depending on the kernel – see the appendix for details. The special case where the design is equally spaced on the unit interval provides especially simple formulae. In this case, M_n^* is proportional to M_n. Furthermore, when the errors are normal (and $c = 1$),

$$Var[T^{**}] \approx 1 + \frac{h^4}{4} a_1(k)^2 M_n^{-1}[\gamma_2 - \frac{\gamma_1^2}{M_n}] + \frac{2}{nh} a_2(k), \tag{10}$$

where $a_1(k) = \int t^2 k(t) dt$ and $a_2(k) = \int k(t)^2 dt$, while $\gamma_1 = n^{-1} \sum_{i=1}^n x_i^2 \sigma_i^{-4} \frac{\partial^2 \sigma^2}{\partial x^2}(x_i)$ and $\gamma_2 = n^{-1} \sum_{i=1}^n x_i^2 \sigma_i^{-6} [\frac{\partial^2 \sigma^2}{\partial x^2}(x_i)]^2$. The kernel constants can be evaluated for standard choices of k. For example, when $k(t) = \frac{3}{4}(1-t^2)\mathrm{I}(|t| < 1)$, $a_1 = 0.2$ and $a_2 = 0.6$.

When the errors are homoskedastic[2], $Var[T^{**}] \approx 1 + 2n^{-1}h^{-1}a_2(k)$, which suggests a quite modest variance inflation in this case.

We now calculate the correction for some simple heteroskedastic regressions that Carroll (1982) used in simulations. He examined the following design: $y_i = \phi_i + \epsilon_i \sigma_i$, $i = 1, 2, .., n = 60$, where $\phi_i = \beta_0 + \beta_1 x_i$ with x_i equally spaced[3] on $(-1/2, 1/2)$, while ϵ_i were iid N(0,1). Three models for the variance were considered:

(M1) $\sigma_i^2 = \delta_1 + \delta_2 \phi_i^2$

(M2) $\sigma_i = \delta_1 exp[\delta_2 |\phi_i|]$

(M3) $\sigma_i = \delta_1 exp[\delta_2 \phi_i^2]$,

where $\beta_0 = 50$ and $\beta_1 = 60$ throughout, and: $(\delta_1 = 100, \delta_2 = 0.25)$, $(\delta_1 = 0.25, \delta_2 = 0.04)$, and $(\delta_1 = 0.25, \delta_2 = 1/3200)$ in (M1), (M2), and (M3) respectively, while $k(t) = \frac{3}{2}(1 - |t|)^2 \mathrm{I}(|t| < 1)$ and $h = 0.13$. We focus on $\widehat{\beta}_1$. Figures 1-3 show the relationship, predicted from (7), between $Var[T^{**}]$ and bandwidth for the Carroll designs (M1-M3). Our approximations predict a percentage variance inflation of 23% in each case[4] relative to the variance of $\overline{\beta}_1$ at the Carroll bandwidth $h = 0.13$ – i.e. $Var[\widehat{\beta}_1]$ should be 162.6, 0.878, and 0.10 in M1-3 respectively. Carroll (1982) gives (in his Table 1) 144.46, 0.8034, and 0.0888 respectively from 500 replications. We were puzzled by the discrepancy and carried out a simulation experiment on model 1. We found, from 10,000 replications[5], the monte carlo variance of $\widehat{\beta}_1$ to be 181.39 and of $\overline{\beta}_1$ to be[6] 133.69. These numbers are in closer agreement with our second order

[2]Or if the nonparametric estimates are undersmoothed – i.e. $hn^{\pi} \to 0$.

[3]Carroll actually chose a random design with x_i uniformly distributed on $(-1/2, 1/2)$.

[4]For the given parameters, $\mathcal{B}$ is very small.

[5]The different results obtained by Carroll may be largely explained by the small number of replications he used.

[6]We also investigated the OLS estimator and found its monte carlo variance to be 172.03 which agees quite closely with the 172.07 reported in Carroll's Table 1.

theory, although (10) apparently understates[7] the true variance of $\widehat{\beta}_1$ a bit.

4.3 Comparison with T'

Using the same method as in section 4.1, we obtain a stochastic expansion for T'. Let T'^{**} be the truncated version of T'; then T'^{**} is identical to T^{**} except that $\widehat{\sigma}_i^2 - \sigma_i^2$ is replaced by $\widetilde{\sigma}_i^2 - \sigma_i^2$. Therefore, $h = O(n^{-\pi})$ is optimal for $\widetilde{\beta}$ too, and

PROPOSITION 1B. *Let $h = O(n^{-\pi})$. Then $E[T'^{**}] = O(n^{-1/2})$ and*

$$Var[T'^{**}] = 1 + h^4\mathcal{B}' + n^{-1}h^{-P}\mathcal{V} + o(n^{-2\mu}), \tag{11}$$

where $\mathcal{B}'$ is the same as $\mathcal{B}$, except that $\widetilde{B}_i(g)$, where $g(x_i) \equiv E[y_i^2] = \sigma^2(x_i) + \beta^T x_i x_i^T \beta$, replaces $\widetilde{B}_i(\sigma^2)$.

It may appear that $\widehat{\beta}$ dominates $\widetilde{\beta}$, since $\widehat{\sigma}_i^2$ imposes the parametric restriction on the mean function that $\widetilde{\sigma}_i^2$ ignores. However, our second order approximations do not completely support this argument. In the appendix we obtain the asymptotic approximation

$$\widetilde{B}_i(g) \approx \widetilde{B}_i(\sigma^2) + \vartheta,$$

where the $O(1)$ quantity ϑ depends only on the kernel and on β. Therefore, $\Gamma_1(g) \approx \Gamma_1 + \Delta_1$ and $\Gamma_2(g) \approx \Gamma_2 + \Delta_2 + 2\Delta_3$, where $\Delta_1 = \vartheta n^{-1}\sum_{i=1}^n x_i x_i^T \sigma_i^{-4}$, $\Delta_2 = \vartheta^2 n^{-1}\sum_{i=1}^n x_i x_i^T \sigma_i^{-6}$, and $\Delta_3 = \vartheta n^{-1}\sum_{i=1}^n x_i x_i^T \widetilde{B}_i(\sigma^2)\sigma_i^{-6}$. Combining, we obtain

[7] This may be due to the fact that our approximations relate to a local linear estimator, while we, in replicating the Carroll study, chose a Nadaraya-Watson estimator. In this case, boundary effects - see Linton (1991) - may be causing additional variation.

$$\Gamma_2(g) - \Gamma_1(g) M_n^{-1} \Gamma_1(g) \approx \Gamma_2 - \Gamma_1 M_n^{-1} \Gamma_1 + \Delta_2 - \Delta_1 M_n^{-1} \Delta_1 + 2\Delta_3 - 2\Delta_1 M_n^{-1} \Gamma_1.$$

Now, by the Cauchy-Schwarz inequality, $\Delta_2 - \Delta_1 M_n^{-1} \Delta_1 \geq 0$. Therefore, if

$$\Delta_3 - \Delta_1 M_n^{-1} \Gamma_1 \geq 0, \tag{12}$$

then, asymptotically, $E[T'^{**2}] \geq E[T^{**2}]$. In the special case that $\sigma^2(x)$ is a convex quadratic function of x, and hence $\widetilde{B}_i(\sigma^2) \approx \delta > 0$, (12) holds by the Cauchy-Schwarz inequality. In general, however, it is possible that $\Delta_3 - \Delta_1 M_n^{-1} \Gamma_1 < 0$, and even $E[T'^{**2}] < E[T^{**2}]$.

Although we cannot uniformly rank the two estimators, the tendency would appear to be in favour of $\widehat{\beta}$.

4.4. Asymptotic moments of S and W

Robinson (1989), Andrews (1989b) and Stoker (1989) establish the consistency of standard errors and test statistics in semiparametric situations. However, little is known about their small sample properties. Chesher and Jewitt (1987) and Chesher (1989) have shown that robust standard errors and test statistics based on them can have finite sample properties quite different from their limiting behavior. We suspect the usual asymptotic approximations may be even worse for the semiparametric standard errors and test statistics. Furthermore, the numerical value of standard errors and test statistics can vary considerably with bandwidth. Therefore, it is important to take account of the second order effects when designing testing procedures for empirical work.

In the sequel, we restrict attention to standard errors and test statistics derived from $\widehat{\sigma}^2$. Assuming that $X_D = O_P(1)$, we have

$$\sqrt{n}(\widehat{s}-s) = \frac{[c^T\sqrt{n}(\widehat{M}_n^{-1}-M_n^{-1})c]}{2[c^TM_n^{-1}c]^{1/2}} - \frac{3[c^T\sqrt{n}(\widehat{M}_n^{-1}-M_n^{-1})c]^2}{4\sqrt{n}[c^TM_n^{-1}c]^{3/2}} + O_P(n^{-1})$$

$$\sqrt{n}(\widehat{M}_n^{-1}-M_n^{-1}) = -M_n^{-1}X_DM_n^{-1} + \frac{M_n^{-1}X_DM_n^{-1}X_DM_n^{-1}}{\sqrt{n}} + O_P(n^{-1}).$$

Let S^{**} be the $O_P(n^{-1})$ truncation

$$S^{**} = -\frac{c^TM_n^{-1}X_D^*M_n^{-1}c}{2c^TM_n^{-1}c} + \frac{c^TM_n^{-1}X_D^*M_n^{-1}X_D^*M_n^{-1}c}{2\sqrt{n}c^TM_n^{-1}c} + \frac{3[c^TM_n^{-1}X_D^*M_n^{-1}c]^2}{4\sqrt{n}[c^TM_n^{-1}c]^2}$$

with X_D^* replacing X_D. Then S^{**} has mean $O(\sqrt{n}h^2)+O(n^{-1/2}h^{-P})$ and variance $O(1)+O(h^2)+O(n^{-1}h^{-P})$. Therefore, the squared bias dominates the correction to $E[S^{**2}]$, and can be exceptionally large to the extent that $\widehat{s}$ is not $\sqrt{n}$ consistent for any bandwidth, unless[8] $P=1$, although it will be consistent for any dimension.

The bandwidth that minimises $E[S^{**2}]$ sets $h^2 \approx n^{-1}h^{-P}$; it is $h(n)=O(n^{-\pi^*})$, where $\pi^*=1/(P+2)$, which is narrower than $O(n^{-\pi})$. In this case, $E[S^{**}]=O(n^{-\mu^*})$, where $\mu^*=(2-P)/(4+2P)$ (when $P=1$, $\pi^*=1/3$ and $\mu^*=1/6$). Clearly, the second order effects are larger for $\widehat{s}$ than for $\widehat{\beta}$ and $\widetilde{\beta}$.

[8] Or unless a bias reduction technique, such as higher order kernels (Robinson (1988b)), higher order polynomial regression (Fan (1992)), or the multiplicative bias correction method of Linton and Nielsen (1994), is used.

PROPOSITION 2. *Let* $h = O(n^{-\pi^*})$. *Then*

$$E[S^{**}] = h^2\sqrt{n}\frac{c^T M_n^{-1}\Gamma_1 M_n^{-1}c}{2c^T M_n^{-1}c} - n^{-1/2}h^{-P}(\kappa_4+2)\frac{c^T M_n^{-1}M_n^* M_n^{-1}c}{2c^T M_n^{-1}c} + o(n^{-\mu^*}) \quad (13)$$

and $Var[S^{**}] = O(1) + o(n^{-2\mu^*})$.

REMARK: When $P > 1$, this proposition still applies, although it may be preferable to express the results for $\widehat{s}$ itself, which is $O_P(1)$, rather than S. The expansion argument has to be suitably modified – i.e. a different normalization for $n^{-1}\sum_{i=1}^n x_i x_i^T(\widehat{\sigma}_i^{-2} - \sigma_i^{-2})$ should be used, such as h^{-2}, instead of $\sqrt{n}$.

The first term in (13) has the same sign as $\widetilde{B}_i(\sigma^2)$ – under homoskedasticity it is zero. The second quantity is negative and depends on the error kurtosis, on a design effect, and on the number of nuisance parameters (nh^P) used up. Either term could dominate, and the standard error could be[9] an upward or downward biased estimate of s. We suspect that a downward bias is quite frequent in applications – except in cases of extreme curvature or when a very large bandwidth is used, the second term in (13) should dominate.

We now turn to the behavior of the test statistic W. The large biases in the standard error can adversely affect the test statistic, especially when the same bandwidth of order $n^{-\pi}$ is used to estimate both β and s. Since $n^{-1/2}S = O_P(h^2)$, we have

$$W = T\{1 + n^{-1/2}S\}^{-1} = T - \frac{TS}{\sqrt{n}} + O_P(n^{-2\mu}).$$

Substituting $T = T^{**} + O_P(n^{-2\mu})$ and $S = S^{**} + O_P(n^{-2\mu})$, we obtain

[9]Note that from (7), s itself provides an underestimate of the sampling variability of $\widehat{\beta}$.

$$W = s^{-1} c^T M_n^{-1} \{ X_N^* - \frac{X_D^* M_n^{-1} X_N^*}{\sqrt{n}} - \frac{X_N^* S^{**}}{\sqrt{n}} \} + O_P(n^{-2\mu}) \equiv W^{**} + O_P(n^{-2\mu}).$$

Then, using the fact that $Cov\{X_N^*, n^{-1/2} X_N^* E[S^{**}]\} = O(h^2)$, the second order effect on the variance of W is of order $n^{-\mu}$. This is considerably larger than in (7) and (11).

PROPOSITION 3A. *Let a single bandwidth $h = O(n^{-\pi})$ be used in estimating β and s. Then $E[W^{**}] = O(n^{-1/2})$ and*

$$Var[W^{**}] = 1 - h^2 \frac{c^T M_n^{-1} \Gamma_1 M_n^{-1} c}{c^T M_n^{-1} c} + o(n^{-\mu}), \tag{14}$$

Therefore, when the curvature of $\sigma^2(\bullet)$ is large and positive, the first order theory overestimates the variability of the Wald statistic because it neglects the upward bias in $\widehat{s}$. In this case, we get under-rejection when the null hypothesis is true.

These approximations suggest it may be advantageous to use different bandwidths for the estimator and its standard error. Let $h_1 = O(n^{-\pi})$ be used to construct $\widehat{\beta}$ and let the narrower bandwidth $h_2 = O(n^{-\pi^*})$ be used for $\widehat{s}$. The behavior of the standard error still dominates, although an improved rate is obtained. In this case, the second order effect on $E[W^{**}]$ is $O(n^{-2\mu^{**}})$, where $\mu^{**} = 1/(P+2)$ (when $P = 1$, $\mu^{**} = 1/3$).

PROPOSITION 3B. *Let $h_1 = O(n^{-\pi})$ be used to construct $\widehat{\beta}$ and let $h_2 = O(n^{-\pi^*})$ be used for $\widehat{s}$. Then*

$$Var[W^{**}] = 1 - h_2^2 \frac{c^T M_n^{-1} \Gamma_1 M_n^{-1} c}{c^T M_n^{-1} c} + n^{-1} h_2^{-P} (\kappa_4 + 2) \frac{c^T M_n^{-1} M_n^* M_n^{-1} c}{c^T M_n^{-1} c} + o(n^{-2\mu^{**}}). \tag{15}$$

Under homoskedasticity, the second term in (15) dominates and the test statistic tends to over-reject under the null. When the design is equally spaced on the unit interval and the errors are normally distributed, $Var[W^{**}] \approx 1 + 2a_2 n^{-1} h_2^{-1}$, under homoskedasticity.

5. SECOND ORDER OPTIMALITY AND BANDWIDTH CHOICE

In this section we consider second order optimal estimation of β. We restrict attention to the class of semiparametric GLS estimators based on a local linear estimator $\widehat{\sigma}^2(x)$ of the variance function with bandwidths of the form $\gamma n^{-\pi}$, for $\gamma > 0$.

The optimal value[10] of γ can be found by calculus to be

$$\gamma_0 = [P\mathcal{V}/4\mathcal{B}]^{1/(4+P)}, \tag{16}$$

and at this bandwidth,

$$E[T^{**}] = 1 + n^{-2\mu}\{[P/4]^{4/(4+P)} + [4/P]^{P/(4+P)}\}\mathcal{V}^{4/(4+P)}\mathcal{B}^{P/(4+P)}. \tag{17}$$

Since $\mathcal{B}$ and $\mathcal{V}$ depend on $\sigma^2(\bullet)$, we cannot calculate γ_0 when the regression functions are unknown. However, we can estimate these quantities by

$$\widehat{\mathcal{B}} = \frac{c^T \widehat{M}_n^{-1}[\widehat{\Gamma}_2 - \widehat{\Gamma}_1 \widehat{M}_n^{-1} \widehat{\Gamma}_1]\widehat{M}_n^{-1} c}{c^T \widehat{M}_n^{-1} c} \ ; \ \widehat{\mathcal{V}} = (\widehat{\kappa}_3^2 + 2 + \widehat{\kappa}_4)\frac{c^T \widehat{M}_n^{-1} \widehat{M}_n^* \widehat{M}_n^{-1} c}{c^T \widehat{M}_n^{-1} c},$$

[10] The optimal bandwidth for estimating β differs from the optimal bandwidth for estimating $\sigma^2(x)$, although the difference is at the level of constants. The optimal bandwidths and MSE corrections are of the same order of magnitude in both estimation problems. Also the quadratic kernel is optimal for both estimation problems – see Müller (1988).

where $\widehat{\kappa}_3$ and $\widehat{\kappa}_4$ are the sample third and fourth cumulants respectively, $\widehat{M}_n^* = n^{-1}\sum_{i=1}^n x_i x_i^T \widehat{\sigma}_i^{-2}\rho_i$, $\widehat{\Gamma}_1 = n^{-1}\sum_{i=1}^n x_i x_i^T \widehat{B}_i \widehat{\sigma}_i^{-4}$, and $\widehat{\Gamma}_2 = n^{-1}\sum_{i=1}^n x_i x_i^T \widehat{B}_i^2 \widehat{\sigma}_i^{-6}$, where $\widehat{B}_i$ is an estimate of $\widetilde{B}_i(\sigma^2)$. One then plugs $\widehat{\mathcal{V}}$ and $\widehat{\mathcal{B}}$ into (16). We suggest two different estimators of $\widetilde{B}_i(\sigma^2)$. Firstly,

$$\widetilde{B}_i(\widehat{\sigma}^2) = [\sum_{j\neq i} w_{ij}(h^*)\widehat{\sigma}_j^2 - \widehat{\sigma}_i^2]/h^{*2}, \tag{18}$$

in which $h^* \to 0$ is any bandwidth sequence and $\widehat{\sigma}_i^2$ is a preliminary estimate of $\sigma^2(x_i)$. Note that one may need to use bias reduction methods to estimate $\widehat{\sigma}_i^2$ in order to achieve consistent estimation of $\widetilde{B}_i(\sigma^2)$. An alternative approach is to replace $\widetilde{B}_i(\sigma^2)$ by its limit (9), and then to estimate the relevant derivatives either by higher order local polynomial regression as discussed in Linton (1993) or by the series method discussed in Andrews (1991a). Under smoothness and moment conditions, the resulting estimator $\widehat{\gamma}_0$ will consistently estimate γ_0, and $\widehat{\beta}(\widehat{h})$, where $\widehat{h} = \widehat{\gamma}_0 n^{-1/(4+P)}$, should achieve the optimal second moment bound (17) – as shown in Linton (1993) for a related example.

The plug-in method constitutes an alternative to the cross-validation scheme considered in Robinson (1991a). Although it is not fully automatic – estimating $\mathcal{V}$ and $\mathcal{B}$ requires selection of a preliminary bandwidth – evidence presented in Park and Marron (1990) and Sheather and Jones (1991) suggests that the final estimate may be little affected by the preliminary choice of bandwidth. The so-called rule of thumb approach, see Silverman (1986), offers an alternative plug-in implementation that does not require a preliminary bandwidth to be chosen. In this approach one specifies, for the purposes of bandwidth choice only, a parametric model for $\sigma^2(\bullet)$ such as making it a quadratic function of x. Parametric procedures are then used to get a preliminary fit $\widehat{\sigma}_i^{\infty 2}$ and derivatives thereof which are then plugged into (9). This method achieves the more modest objective of being second order optimal for the particular model chosen for $\sigma^2(\bullet)$, although the correct order of magnitude for $\widehat{h}$ is guaranteed for all $\sigma^2(\bullet)$.

6. SIMULATIONS

We generated 5000 samples of size n =100 from

(MC1) $y_i = 1.66 - 0.175x_i + \varepsilon_i\sigma_i,$

(MC2) $\sigma_i^2 = 0.0592 - 0.229x_i + 0.00254x_i^2,$

where ε_i were independent standard normals. The chosen parameters were those estimated (by least squares) from the household survey dataset analyzed in Anand, Harris and Linton (1993) in which y is food share and x is the log of total expenditure per capita. The full dataset contained 7465 observations; we retained the 100 percentiles of $x's$ distribution as our design. Figure 4 plots $\sigma^2(x)$, while Figure 5 shows the design density estimated by a standard kernel procedure. We implemented the local linear procedure in rank space, i.e. the dataset was ordered by x, and the nonparametric regressions were estimated using i/n in place of x_i in (4). A normal kernel $k(t) = (2\pi)^{-1/2}\exp(-0.5t^2)$ was used throughout. The approximations of Proposition 1A are still valid for our implementation.

Firstly, we investigate how the semiparametric GLS estimator performed at a grid of bandwidths: h =0.05, 0.1,.., 1.0. Figures 6 and 7 show the simulation variance of $\widehat{\beta}$ compared with the asymptotic approximations predicted from (7) and with the exact variance of the OLS estimators. Although poor for large or small bandwidths, the asymptotic approximations are excellent for a range of bandwidths located close to the optimum. Also note that up to a 25% efficiency gain over OLS is possible for these bandwidths. These simulations emphasize the importance of bandwidth selection, which we now turn to.

Taking as preliminary bandwidth $h^* = 0.3$, we first obtained estimates $\widehat{\sigma}_i^2(h^*)$ and hence (18). The variance constant in this case was estimated by $\widehat{\mathcal{V}} = 2 \sum\sum_{j \neq i} nh^* w_{ij}^2(h^*)$. Although different bandwidths were estimated for constant and intercept, their simulation distribution was quite similar; we show only our results for the constant. The estimated bandwidth is highly concentrated near the true optimum. We point out that $\widehat{h}$ is local ancillary for β, see Cox (1980). Therefore, it is recommended to conduct inference about β conditional on the estimated bandwidth. Thus, one can read off the performance of $\widehat{\beta}$ from Figure 6.

7. CONCLUSIONS

The semiparametric estimators we examined, while first order efficient, are infinitely deficient, not only when compared with GLS but even when compared with feasible GLS estimators based on a correct finite dimensional parametric model for $\sigma^2(\bullet)$, in the sense that the increase of asymptotic MSE (over GLS) that such a strategy entails is $O(n^{-4/(4+P)})$ when an approximately optimal bandwidth is used, while for parametric estimators the small sample cost is $O(n^{-1})$.

However, the preliminary calculations we have done for some common variance models suggests that the cost of using this method for one-dimensional problems − such as when the variance depends only on the mean of the dependent variable − may not be great unless extreme heteroskedasticity is present, and provided certain precautions are taken. In particular, a large enough bandwidth should be chosen so that the degrees of freedom term in (7) is not too large.

The standard error estimates are likely to be severely downward biased in small samples which can also adversely affect the t-ratios. We recommend using different bandwidths for estimating β and s.

Finally, we have proposed a method of bandwidth selection which should be second order optimal, and which appears to fare well in practice.

APPENDIX

The appendix is divided into four sub-sections. In part 1 we derive properties of $\widehat{\sigma}_i^2$ and $\widetilde{\sigma}_i^2$ which are useful in establishing the main results. In section 2 we obtain expansions for the standardised sums X_N and X_D that determine the standardised quantities T, S, and W. In section 3 we derive the main expansions, while in section 4 we discuss the remainder terms.

A1. PROPERTIES OF NONPARAMETRIC ESTIMATORS

(1) We first consider the properties of $\widehat{\sigma}_i^2$. We have

$$\widehat{\sigma}_i^2 - \sigma_i^2 = [\widehat{\sigma}_i^2 - \widehat{\sigma}_i^{*2}] + [\widehat{\sigma}_i^{*2} - \overline{\sigma}_i^2] + [\overline{\sigma}_i^2 - \sigma_i^2] \equiv P_i + V_i + B_i, \tag{19}$$

where $\widehat{\sigma}_i^{*2} = \sum_{j \neq i} w_{ij} u_j^2$ and $\overline{\sigma}_i^2 = \sum_{j \neq i} w_{ij} \sigma_j^2$. In particular, $V_i = \sum_{j \neq i} w_{ij}(u_j^2 - \sigma_j^2)$. Note that $B_i = h^2 \widetilde{B}_i$. The kernel weights w_{ij} satisfy:

(a) $Card\{j : w_{ij} \neq 0\} = O(nh^P)$

(b) $w_{ij} \leq \chi n^{-1} h^{-P}$

for some $\chi < \infty$, where the order of magnitude in (a) is uniform over i, see Linton (1993). Therefore, $V_i = O_P(n^{-1/2}h^{-P/2})$. In section 4 we verify that $P_i = o_P(n^{-\mu})$, when $h = O(n^{-\pi})$. Therefore,

$$\widehat{\sigma}_i^2 - \sigma_i^2 = V_i + B_i + o_P(n^{-\mu}),$$

where $Var[V_i] \approx n^{-1}h^{-P}(\kappa_4 + 2)\sigma_i^4(nh^P \sum_{j \neq i} w_{ij}^2)$, while $B_i = h^2 \widetilde{B}_i$.

We now further approximate $\widetilde{B}_i$. Let $\widehat{\tau}_i^* = (Z_i^T K_i Z_i)^{-1} Z_i^T K_i (u_1^2, .., u_n^2)^T$ be the (infeasible) local linear estimator of $\tau_i = (\sigma^2(x_i), \partial\sigma^2/\partial x_{i1}, .., \partial\sigma^2/\partial x_{iP})^T$. Then,

$$E[\widehat{\tau}_i^*] - \tau_i = (Z_i^T K_i Z_i)^{-1} Z_i^T K_i [\sigma^2 - Z_i^T \tau_i],$$

where $\sigma^2 = (\sigma^2(x_1), .., \sigma^2(x_n))^T$. Then we use a Taylor series approximation:

$$\sigma^2(x_j) - \tau_i^T z_{ij} \approx \frac{1}{2} \sum_{\delta=1}^{P} \sum_{\alpha=1}^{P} \frac{\partial^2 \sigma^2}{\partial x_{i\alpha} \partial x_{i\delta}}(x_i)(x_{j\alpha} - x_{i\alpha})(x_{j\delta} - x_{i\delta})$$

in a neighborhood of x_i. Let

$$\Omega_i = \underset{n\to\infty}{Lim}\, n^{-1}h^{-P} H^{-1} Z_i^T K_i Z_i H^{-1}$$

$$\omega_i^{\alpha\delta} = \underset{n\to\infty}{Lim}\, n^{-1}h^{-P} H^{-1} Z_i^T K_i [(\tfrac{x_{1\alpha} - x_{i\alpha}}{h})(\tfrac{x_{1\delta} - x_{i\delta}}{h}), .., (\tfrac{x_{n\alpha} - x_{i\alpha}}{h})(\tfrac{x_{n\delta} - x_{i\delta}}{h})]^T,$$

where $H = diag\{1, h, .., h\}$ is a $P+1$ dimensional diagonal matrix, and let $\theta_{\alpha\delta}$ be the first element of the P by 1 vector $\Omega_i^{-1}\omega_i^{\alpha\delta}/2$; asymptotically this quantity does not depend on i. Then

$$\widetilde{B}_i(\sigma^2) \approx \sum_{\delta=1}^{P} \sum_{\alpha=1}^{P} \theta_{\alpha\delta} \frac{\partial^2 \sigma^2}{\partial x_{i\alpha} \partial x_{i\delta}}(x_i),$$

see Fan and Gijbels (1992), Fan (1992, Theorem 1) and Ruppert and Wand (1992, Theorem 4.2). These approximations are valid for any fixed point x_i. For boundary sequences, i.e. for $x_{ni} \to \partial\Upsilon$, where $\partial\Upsilon$ is the topological boundary of Υ, we still get $\widetilde{B}_{ni}(\sigma^2) = O(1)$ – see Ruppert and Wand (1992). ∎

(2) We now consider $\widetilde{\sigma}_i^2$, where

$$\widetilde{\sigma}_i^2 - \sigma_i^2 = [V_{2i} + B_{2i}] - [V_{1i} + B_{1i}][2\beta^T x_i + B_{1i} + V_{1i}],$$

with $V_{1i} = \sum_{j\neq i} w_{ij}u_j$, $B_{1i} = \sum_{j\neq i} w_{ij}\beta^T x_j - \beta^T x_i$, $V_{2i} = \sum_{j\neq i} w_{ij}v_j$ and $B_{2i} = \sum_{j\neq i} w_{ij}g(x_j) - g(x_i)$, where $v_j = y_j^2 - g(x_j)$, with $g(x) = \sigma^2(x) + \beta^T x x^T \beta$. Note that $B_{2i} = O(h^2)$ and $V_{1i}, V_{2i}, = O_P(n^{-1/2}h^{-P/2})$, while $B_{1i} = o(h^2)$; therefore, B_{1i} can be dropped. Furthermore, $V_{2i} - 2\beta^T x_i V_{1i} = \sum_{j\neq i} w_{ij}(u_j^2 - \sigma_j^2) + o_P(n^{-\mu})$. Therefore,

$$\widetilde{\sigma}_i^2 - \sigma_i^2 = B_{2i} + V_i + o_P(n^{-\mu}),$$

where $Var[\widetilde{\sigma}_i^2] \approx Var[V_i]$ and $E[\widetilde{\sigma}_i^2] - \sigma_i^2 \approx B_{2i}$. Furthermore, $B_{2i} = h^2\widetilde{B}_{2i}$, where

$$\widetilde{B}_{2i} \approx \sum_{\delta=1}^{P}\sum_{\alpha=1}^{P} \theta_{\alpha\delta}\frac{\partial^2 g^2}{\partial x_{i\alpha}\partial x_{i\delta}}(x_i) = \widetilde{B}_i(\sigma^2) + \sum_{\delta=1}^{P}\sum_{\alpha=1}^{P} \theta_{\alpha\delta}\beta_\alpha\beta_\delta.$$

A2. PROPERTIES OF STANDARDIZED SUMS

$T, S,$ and W all depend only on the standardised quantities, $X_N = n^{-1/2}\sum_{i=1}^n x_i u_i \widehat{\sigma}_i^{-2}$ and $X_D = n^{-1/2}\sum_{i=1}^n x_i x_i^T(\widehat{\sigma}_i^{-2} - \sigma_i^{-2})$, while T' depends only on similar quantities with $\widetilde{\sigma}_i^2$ replacing $\widehat{\sigma}_i^2$. Firstly, write

$$X_N = X_N^* + P_N + R_N \ ; \ X_D = X_D^* + P_D + R_D,$$

where

$$X_N^* = X_{N0} - n^{-1/2}\sum_{i=1}^n x_i u_i \left\{\sigma_i^{-4}B_i - \sigma_i^{-6}B_i^2 + \overline{\sigma}_i^{-4}V_i - \overline{\sigma}_i^{-6}V_i^2\right\}$$

$$X_D^* = -n^{-1/2}\sum_{i=1}^n x_i x_i^T \left\{\sigma_i^{-4}B_i - \sigma_i^{-6}B_i^2 + \overline{\sigma}_i^{-4}V_i - \overline{\sigma}_i^{-6}V_i^2\right\}$$

with $X_{N0} = n^{-1/2}\sum_{i=1}^n x_i u_i \sigma_i^{-2}$, while: $P_N = n^{-1/2}\sum_{i=1}^n x_i u_i(\widehat{\sigma}_i^{-2} - \widehat{\sigma}_i^{*-2})$, $P_D = n^{-1/2}\sum_{i=1}^n x_i x_i^T(\widehat{\sigma}_i^{-2} - \widehat{\sigma}_i^{*-2})$, and

$$R_N = -n^{-1/2}\sum_{i=1}^n x_i u_i \left\{\sigma_i^{-6}\overline{\sigma}_i^{-2}B_i^3 + \overline{\sigma}_i^{-6}\widehat{\sigma}_i^{*-2}V_i^3\right\} \equiv R_{NB} + R_{NV}$$

$$R_D = -n^{-1/2}\sum_{i=1}^n x_i x_i^T \left\{\sigma_i^{-6}\overline{\sigma}_i^{-2}B_i^3 + \overline{\sigma}_i^{-6}\widehat{\sigma}_i^{*-2}V_i^3\right\} \equiv R_{DB} + R_{DV}.$$

We establish in section 4 below that

(R1) $R_N = o_P(n^{-2\mu})$

(R2) $n^{-1/2}R_D = o_P(n^{-2\mu})$

(R3) $P_N, n^{-1/2}P_D = o_P(n^{-2\mu})$,

provided $h(n) = O(n^{-\pi})$. Therefore, collecting terms,

$$X_N = X_{N0} - L_{N1} + L_{N2} + L_{N3} - Q_{N1} + Q_{N2} + C_{N1} + o_P(n^{-2\mu})$$

$$X_D n^{-1/2} = -n^{-1/2}X_{D0} - n^{-1/2}b_{D1} + n^{-1/2}b_{D2} + n^{-1/2}b_{D3} + o_P(n^{-2\mu}),$$

where, letting $\zeta_j = \varepsilon_j^2 - 1$,

$X_{D0} = n^{-1/2}\sum_{i=1}^n x_i x_i^T \overline{\sigma}_i^{-4}V_i = O_P(1)$
$b_{D1} = n^{-1/2}\sum_{i=1}^n x_i x_i^T \sigma_i^{-4}B_i = O(\sqrt{n}h^2)$
$b_{D2} = n^{-1/2}\sum_{i=1}^n x_i x_i^T \overline{\sigma}_i^{-6}E[V_i^2] = O(\frac{1}{\sqrt{n}h^P})$
$b_{D3} = n^{-1/2}\sum_{i=1}^n x_i x_i^T \sigma_i^{-6}B_i^2 = O(\sqrt{n}h^4)$

$L_{N1} = n^{-1/2} \sum_{i=1}^n x_i u_i \sigma_i^{-4} B_i = O_P(h^2)$
$L_{N2} = n^{-1/2} \sum_{i=1}^n x_i u_i \overline{\sigma}_i^{-6} E[V_i^2] = O_P(n^{-1} h^{-P})$
$L_{N3} = n^{-1/2} \sum_{i=1}^n x_i u_i \sigma_i^{-6} B_i^2 = O_P(h^4)$
$Q_{N1} = n^{-1/2} \sum_{i=1}^n x_i u_i^4 \overline{\sigma}_i^{-4} V_i = n^{-1/2} \sum_{i=1}^n \sum_{j \neq i} w_{ij} x_i \overline{\sigma}_i^{-4} \sigma_j^2 u_i \zeta_j = O_P(n^{-1/2} h^{-P/2})$
$Q_{N2} = n^{-1/2} \sum_{i=1}^n \sum_{j \neq i} x_i \overline{\sigma}_i^{-6} \sigma_j^4 w_{ij}^2 u_i (\zeta_j^2 - E[\zeta_j^2]) = O_P(n^{-1} h^{-P})$
$C_{N1} = n^{-1/2} \sum_{i=1}^n \sum_{k \neq j \neq i}\sum w_{ij} w_{ik} x_i \overline{\sigma}_i^{-6} \sigma_j^2 \sigma_k^2 u_i \zeta_j \zeta_k = O_P(n^{-1} h^{-P})$

because $n^{-1/2} \sum_{i=1}^n x_i x_i^T \overline{\sigma}_i^{-6} (V_i^2 - E[V_i^2]) = O_P(n^{-1} h^{-P})$. ∎

We now establish the properties of X_N^* and X_D^*.

LEMMA 1*: $X_{N0} = O_P(1)$, $X_{D0} = O_P(1)$.*

The properties of X_{N0} are obvious. By interchanging summations we obtain

$$X_{D0} = n^{-1/2} \sum_{i=1}^n \sum_{j \neq i} w_{ij} x_i x_i^T \overline{\sigma}_i^{-4} (u_j^2 - \sigma_j^2) = n^{-1/2} \sum_{j \neq i} a_j \zeta_j,$$

where $a_j = \sum_{i=1}^n w_{ij} x_i x_i^T \overline{\sigma}_i^{-4} \sigma_j^2 = O(1)$, since x_i lies in a bounded set. ∎

LEMMA 2*: As $n \rightarrow \infty$,*

$$b_{D1} = h^2 \sqrt{n} n^{-1} \sum_{i=1}^n x_i x_i^T \sigma_i^{-4} \widetilde{B}_i = O(h^2 \sqrt{n}).$$

Follows by substituting the approximation for $\widetilde{B}_i$ into the formula for b_{D1}. ∎

LEMMA 3*: As $n \rightarrow \infty$,*

$$Var[L_{N1}] = h^4 n^{-1} \sum_{i=1}^n x_i x_i^T \sigma_i^{-6} \widetilde{B}_i^2 = O(h^4).$$

∎

LEMMA 4*: As $n \rightarrow \infty$,*

$$Var[Q_{N1}] \approx [\kappa_4 + 2 + \kappa_3^2] n^{-1} \sum_{i=1}^{n} x_i x_i^T \sigma_i^{-2} (\sum_{j \neq i} w_{ij}^2) = O(n^{-1} h^{-P}).$$

By interchanging summations, we obtain $Q_{N1} = n^{-1/2} \sum_{i=1}^{n} x_i u_i \overline{\sigma}_i^{-4} V_i = \sum\sum_{j \neq i} \rho_{ij} \epsilon_i \zeta_j$, where $\rho_{ij} = n^{-1/2} w_{ij} x_i \overline{\sigma}_i^{-4} \sigma_i \sigma_j^2$. Therefore, $Var[Q_{N1}] = \sum\sum_{j \neq i, k \neq l}\sum\sum \rho_{ij} \rho_{lk}^T E[\epsilon_i \zeta_j \epsilon_l \zeta_k]$. But $E[\epsilon_i \zeta_j \zeta_k \epsilon_l] = 0$, unless either $i = l$ and $j = k$, in which case it is $\kappa_4 + 2$, or $i = k$ and $j = l$, in which case it is κ_3^2. Therefore,

$$Var[Q_{N1}] = (\kappa_4 + 2) \sum\sum_{j \neq i} \rho_{ij} \rho_{ij}^T + \kappa_3^2 \sum\sum_{j \neq i} \rho_{ij} \rho_{ji}^T,$$

where $\rho_{ij} \rho_{ij}^T = n^{-1} w_{ij}^2 x_i x_i^T \overline{\sigma}_i^{-8} \sigma_i^2 \sigma_j^4$. Then by Taylor expansion, $\rho_{ij} \rho_{ij}^T \approx n^{-1} w_{ij}^2 x_i x_i^T \sigma_i^{-2} \approx \rho_{ij} \rho_{ji}^T$, and the result follows. ■

LEMMA 5*: As* $n \rightarrow \infty$,

$$b_{D2} = \sqrt{n} (\kappa_4 + 2) n^{-1} \sum_{i=1}^{n} x_i x_i^T \sigma_i^{-2} (\sum_{j \neq i} w_{ij}^2) = O(n^{-1/2} h^{-P}).$$

Replace $\overline{\sigma}_i^{-6}$ by σ_i^{-6} and $E[V_i^2]$ by $(\kappa_4 + 2) \sigma_i^4 \sum_{j \neq i} w_{ij}^2$, and the result follows. ■

A3. MAIN EXPANSIONS

PROOF OF PROPOSITION 1A,B

The proof of this theorem relies on Lemmas 1-5 above and the proof of (R1)-(R3) which is given in section 4 below.

Using the calculations of section A2, we can drop a number of terms and find that

$$T^{***} = s^{-1} c^T M_n^{-1} \{ X_{N0} - [L_{N1} - \tfrac{b_{D1}^T M_n^{-1} X_{N0}}{\sqrt{n}}] + [L_{N2} - \tfrac{b_{D2}^T M_n^{-1} X_{N0}}{\sqrt{n}}] + [L_{N3} - \tfrac{b_{D3}^T M_n^{-1} X_{N0}}{\sqrt{n}}]$$
$$- Q_{N1} + Q_{N2} + C_{N1} - \tfrac{X_{D0}^T M_n^{-1} X_{N0}}{\sqrt{n}} \} + o_P(n^{-2\mu}). \qquad (20)$$

The correlation between $\frac{X_{D0}^T M_n^{-1} X_{N0}}{\sqrt{n}}$ and the X_{N0} is $O(n^{-1})$. The quadratic and cubic terms (Q_{N1}, Q_{N2}, C_{N1}) are uncorrelated with X_{N0}; the same goes for $L_{N1} - \frac{b_{D1}^T M_n^{-1} X_{N0}}{\sqrt{n}}$, $L_{N2} - \frac{b_{D2}^T M_n^{-1} X_{N0}}{\sqrt{n}}$, and $L_{N3} - \frac{b_{D3}^T M_n^{-1} X_{N0}}{\sqrt{n}}$. For example,

$$Cov[X_{N0}, L_{N1} - \frac{b_{D1}^T M_n^{-1} X_{N0}}{\sqrt{n}}] = E[X_{N0} L_{N1}^T] - \frac{1}{\sqrt{n}} b_{D1}^T M_n^{-1} E[X_{N0} X_{N0}^T],$$

where $E[X_{N0} L_{N1}^T] = n^{-1} \sum_{i=1}^n x_i x_i^T \sigma_i^{-4} B_i = \frac{1}{\sqrt{n}} b_{D1}$ and $E[X_{N0} X_{N0}^T] = M_n$ – the moments cancel as required. Therefore, we have to calculate

$$Var[X_{N0}] + Var[L_{N1} - \frac{b_{D1}^T M_n^{-1} X_{N0}}{\sqrt{n}}] + Var[Q_{N1}],$$

and the result follows from Lemmas 1-5.

The asymptotic variance of T' is the same as that of T, except that we must replace $\widetilde{B}_i$ by $\widetilde{B}_i(g)$. ∎

PROOF OF PROPOSITION 2

When $h = O(n^{-\pi^*})$, $X_D = -X_{D0} - b_{D1} + b_{D2} + o_P(n^{-\mu^*})$. Therefore,

$$S = \frac{c^T M_n^{-1} X_{D0} M_n^{-1} c}{2s^2} + \frac{c^T M_n^{-1} [b_{D1} - b_{D2}] M_n^{-1} c}{2s^2} + o_P(n^{-\mu^*}) \equiv S_0 + b_S + o_P(n^{-\mu^*}).$$

Since $X_{D0} \approx n^{-1/2} \sum_{i=1}^n x_i x_i^T \sigma_i^{-2} \zeta_j$, we have $c^T M_n^{-1} X_{D0} M_n^{-1} c \approx n^{-1/2} \sum_{i=1}^n m_i^2 \zeta_i$, where $m_i = c^T M_n^{-1} x_i \sigma_i^{-1}$. Therefore,

$$Var[S^{**}] = \frac{(\kappa_4 + 2)}{4s^4} \frac{n^{-1} \sum_{i=1}^n m_i^4}{(n^{-1} \sum_{i=1}^n m_i^2)} + o(n^{-2\mu^*}) \ ; \ E[S^{**}] = \frac{c^T M_n^{-1} [b_{D1} - b_{D2}] M_n^{-1} c}{2s^2} + o(n^{-\mu^*}),$$

and *(13)* follows on substituting from Lemmas 2 and 5. ∎

PROOF OF PROPOSITION 3A,B

Using $S = S_0 + b_S + O_P(n^{-\mu})$, from above, we have to calculate the second moments of

$$s^{-1}c^T M_n^{-1}\{X_{N0} + [L_{N1} - \frac{b_{D1}^T M_n^{-1} X_{N0}}{\sqrt{n}}] + Q_{N1} - \frac{X_{N0} b_S}{\sqrt{n}}\},$$

since $Cov[X_{N0}, \frac{X_{D0}^T M_n^{-1} X_{N0}}{\sqrt{n}} + \frac{X_{N0} S_0}{\sqrt{n}}] = O(n^{-1})$. However,

$$Cov[X_{N0}, \frac{X_{N0} b_S}{\sqrt{n}}] = Var[X_{N0}]\{h^2 \frac{c^T M_n^{-1} \Gamma_1 M_n^{-1} c}{2c^T M_n^{-1} c} - \frac{\kappa_4 + 2}{nh^P} \frac{c^T M_n^{-1} M_n^* M_n^{-1} c}{2c^T M_n^{-1} c}\}. \quad (21)$$

Therefore, when the same bandwidth is used throughout, the dominant term in (21) is the first one. When different bandwidths are used, we obtain variance terms of order h_1^4 and $n^{-1}h_1^{-P}$ from $\widehat{\beta}$, and variance terms of order h_2^2 and $n^{-1}h_2^{-P}$ from $\widehat{s}$. Therefore, when the stated bandwidths are used

$$Var[W^{**}] = 1 - h_2^2 \frac{c^T M_n^{-1} \Gamma_1 M_n^{-1} c}{c^T M_n^{-1} c} + \frac{\kappa_4 + 2}{nh_2^P} \frac{c^T M_n^{-1} M_n^* M_n^{-1} c}{c^T M_n^{-1} c} + o(n^{-2\mu^{**}}),$$

as required. ■

A4. REMAINDER TERMS

We now sketch the proof of (R_1), (R_2), and (R_3). The basic arguments are very similar to those contained in Robinson (1987), but are somewhat simplified by our fixed design set up. We omit much detail, and merely remark that a large number of moments may be required to formally verify many of the orders of probability statements.

We first examine P_i. We have $P_i = P_i^A - 2P_i^B$, where

$$P_i^A = n^{-1}\sum_{j\neq i} w_{ij}[x_j^T(\tfrac{X^TX}{n})^{-1}\tfrac{X^Tu}{\sqrt{n}}\tfrac{u^TX}{\sqrt{n}}(\tfrac{X^TX}{n})^{-1}x_j] = O_P(n^{-1}),$$

$$P_i^B = n^{-1/2}\sum_{j\neq i} w_{ij}u_j x_j^T(\tfrac{X^TX}{n})^{-1}\tfrac{X^Tu}{\sqrt{n}} = O_P(n^{-1}h^{-P/2}),$$

where $u = (u_1, u_2, .., u_n)^T$ and $X = (x_1, x_2, .., x_n)^T$. Therefore, $P_i = o_P(n^{-2\mu})$ uniformly in i, and by slight extension, $(R3)$ is satisfied $-$ see below and Robinson (1987) for details.

$(R1)$ and $(R2)$ follow by Taylor expansion. The results for R_{NB} and R_{DB} follow by straightforward calculation. We examine R_{DV} which it is convenient to expand by one more term

$$n^{-1/2}R_{VD} = -n^{-1}\sum_{i=1}^n x_i x_i^T \overline{\sigma}_i^{-8}V_i^3 + n^{-1}\sum_{i=1}^n x_i x_i^T \overline{\sigma}_i^{-8}\widehat{\sigma}_i^{*-2}V_i^4,$$

where

$$V_i^3 = \sum_{j\neq i} w_{ij}^3 E[u_j^3] + \sum_{j\neq i} w_{ij}^3(u_j^3 - E[u_j^3]) + ..+ \sum\sum\sum_{j,k,l\neq i} w_{ij}w_{ik}w_{il}u_j u_k u_l$$

is $O_P(n^{-3/2}h^{-3P/2})$ and $n^{-1}\sum_{i=1}^n x_i x_i^T \overline{\sigma}_i^{-8}V_i^3 = O_P(n^{-3/2}h^{-3P/2})$. Furthermore, since everything is positive

$$|n^{-1}\sum_{i=1}^n x_i x_i^T \overline{\sigma}_i^{-8}\widehat{\sigma}_i^{*-2}V_i^4| \leq \delta[\underset{i\leq n}{Min}\ \widehat{\sigma}_i^{*2}]^{-1}[\underset{i\leq n}{Max}\ |V_i|]^4,$$

where δ is a finite constant, and $[\underset{i\leq n}{Min}\ \widehat{\sigma}_i^{*2}]^{-1} = O_P(1)$ by Robinson (1987) Lemma 13, while $\underset{i\leq n}{Max}\ |V_i| = O_P(n^{-(\mu-\theta)})$ for any $\theta > 0$ by application of Müller and Stadtmüller (1987) Lemma 5.2.

Similar methods work for R_{NV}. ∎

ACKNOWLEDGEMENT

I would like to thank T.J.Rothenberg, C.Cavanagh, W.Härdle, and O.Bunke for helpful discussions, and a referee for some very careful criticisms.

REFERENCES

ANAND, S., C. HARRIS, AND O. LINTON, (1993): "On the concept of Ultra-poverty," Harvard Center for Population, working paper no 9302.

ANDERSON, T.W., AND T. SAWA, (1979): "Evaluation of the Distribution Function of the Two-Stage Least Squares Estimate." *Econometrica* 47, 163-182.

ANDREWS, D.W.K., (1989a): "Semiparametric Econometric Models: I Estimation." Cowles Foundation Discussion paper no. 908.

ANDREWS, D.W.K., (1989b): "Semiparametric Econometric Models: II Testing." Cowles Foundation Discussion paper no. 910.

ANDREWS, D.W.K., (1991a): "Asymptotic Normality of Series Estimators for Non-parametric and Semiparametric Regression Models." *Econometrica* 59, 307-346.

ANDREWS, D.W.K., (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59, 817-858.

CARROLL, R.J. (1982): "Adapting for Heteroscedasticity in Linear Models." *Annals of Statistics* 10, 1224-1233.

CARROLL, R.J., AND W.HÄRDLE, (1989): "Second Order Effects in Semiparametric Weighted Least Squares Regression." *Statistics* 2, 179-186.

CARROLL, R.J., C.F. WU, AND D.RUPPERT, (1988): "The Effect of Estimating Weights in Weighted Least Squares." *Journal of the American Statistical Association* 83, 1045-1054.

CAVANAGH, C.J. (1989): "The Cost of Adapting for Heteroskedasticity in Linear Regression Models." Unpublished Manuscript, Harvard University.

CHESHER, A., AND I. JEWITT, (1987): "The Bias of a Heteroskedasticity Consistent Covariance Matrix." *Econometrica* 55, 1217-1222.

CHESHER, A., (1989): "Hajek Inequalities, Measures of Leverage and the Size of Heteroskedasticity Robust Wald Tests." *Econometrica* 57, 971-977.

COX, D.R., AND D.V. HINKLEY, (1974): *Theoretical Statistics* Chapman and Hall.

COX, D.R., (1980): "Local ancillarity," *Biometrika* 67, 279-286.

EICKER, F. (1968): "Asymptotic Normality and Consistency of the Least Squares Regression Model for Families of Linear Regressions." *Annals of Mathematical Statistics* 34, 447-456.

FAN, J. (1992): "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association*, 87, 998-1004.

FAN, J. and I.GIJBELS (1992): "Spatial and Design Adaptation: Adaptive Order Polynomial Approximation in Function Estimation." Manuscript, CORE.

HÄRDLE, W., J. HART, J. S. MARRON, AND A. B. TSYBAKOV (1992): "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association* 87, 218-226.

HÄRDLE, W., AND O. LINTON (1993): "Applied nonparametric methods," *The Handbook of Econometrics*, Vol. IV, eds D.F. McFadden and R.F. Engle III. North Holland.

HSIEH, D.A., AND C.F. MANSKI (1987): "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression." *Annals of Statistics* 15, 541-551.

LEE, A.J., (1990): *U-Statistics: Theory and Practice.* Marcel Dekker.

LINTON, O.B. (1991): *Edgeworth Approximation in Semiparametric Regression Models* PhD thesis, Department of Economics UC Berkeley.

LINTON, O.B. (1992): "Edgeworth Approximation for Generalised Minimum Contrast Estimators in Semiparametric Regression Models." Nuffield College Discussion Paper no. 73.

LINTON, O.B. (1993): "Second Order Approximation in the Partially Linear Model." Cowles Foundation Discussion Paper no 1065.

LINTON, O.B., AND J.P.NIELSEN, (1994): "A Multiplicative Bias Reduction Method for Nonparametric Regression." *Statistics and Probability Letters* 19, 181-7.

MÜLLER, H. G. (1988): *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics, Vol. 46. Heidelberg/New York: Springer-Verlag.

MÜLLER, H. G. AND U. STADTMÜLLER (1987): "Estimation of Heteroskedasticity in Regression Analysis," *Annals of Statistics* 15, 610-625.

NAGAR, A.L., (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica* 27, 575-595.

NEWEY, W.K. (1990a): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics* 5, 99-135.

NEWEY, W.K. (1990b): "Efficient Instrumental Variables Estimation of Nonlinear Models." *Econometrica* 58, 809-837.

PARK, B.U., AND J.S. MARRON, (1990): "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association* 85, 66-72.

PHILLIPS, P.C.B, (1977a): "A General Theorem in the Theory of Asymptotic Expansions as Approximations to the Finite Sample Distributions of Econometric Estimators," *Econometrica* 45, 1517-1534.

PHILLIPS, P.C.B, (1977b): "An Approximation to the Finite Sample Distribution of Zellner's Seemingly Unrelated Regression Estimator," *Econometrica* 6, 147-164.

ROBINSON, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown form." *Econometrica* 56, 875-891.

ROBINSON, P. M. (1988a): "The Stochastic Difference between Econometric Statistics," *Econometrica* 56, 531-548.

ROBINSON, P. M. (1988b): "Root-N-Consistent Semiparametric Regression," *Econometrica* 56, 931-954.

ROBINSON, P. M. (1991a): "Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models." *Econometrica* 59, 1329-1364.

ROBINSON, P. M. (1991b): "Best Nonlinear Three-stage Least Squares Estimation of certain Econometric Models." *Econometrica* 59, 755-786.

ROTHENBERG, T., (1984a): "Approximating the Distributions of Econometric Estimators and Test Statistics." Ch.14 in: *Handbook of Econometrics*, Vol 2, ed. by Z. Grilliches and M.Intriligator. North Holland.

ROTHENBERG, T., (1984b): "Approximate Normality of Generalized Least Squares Estimates." *Econometrica* 52, 811-825.

ROTHENBERG, T., (1988): "Approximate Power Functions for some Robust Tests of Regression Coefficients." *Econometrica* 56, 997-1019.

RUPPERT, D. AND M.P. WAND (1992): "Multivariate Weighted Least Squares Regression." Rice University, Department of Statistics Technical Report no 92-4.

SARGAN, J.D., AND W.M.MIKHAIL, (1971): "A General Approximation to the Distribution of Instrumental Variables Estimates." *Econometrica* 39, 131-169.

SARGAN, J.D. (1975): "Gram-Charlier Approximations Applied to t Ratios of k-Class Estimators," *Econometrica* 43, 327-346.

SARGAN, J.D. (1976): "Econometric Estimators and the Edgeworth Approximation," *Econometrica* 44, 421-448.

SHEATHER, S.J., AND M.C. JONES, (1991): "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society, Series B* 53, 683-690.

SILVERMAN, B. (1986): *Density estimation for statistics and data analysis.* London, Chapman and Hall.

STOCK, J. H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association* 84, 567-576.

STOCK, J. H. (1991): "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

STOKER, T.M, (1989): "Tests of Average Derivative Constraints," *Review of Economic Studies* 56, 535-552.

STOKER, T.M, (1993): "Smoothing Bias in Derivative Estimation." Forthcoming in *Journal of the American Statistical Association.*

STONE, C.J. (1977): "Consistent Nonparametric Regression," *Annals of Statistics* 5, 595-620.

WHITE, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48, 817-838.