

**A PARADOX OF INCONSISTENT PARAMETRIC
AND CONSISTENT NONPARAMETRIC REGRESSION**

By

Peter C. B. Phillips and Liangjun Su

June 2009

COWLES FOUNDATION DISCUSSION PAPER NO. 1704



**COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281**

<http://cowles.econ.yale.edu/>

A Paradox of Inconsistent Parametric and Consistent Nonparametric Regression*

Peter C. B. Phillips
Yale University, University of Auckland
University of York & Singapore Management University

Liangjun Su
School of Economics
Singapore Management University

May 31, 2009

Abstract

This paper explores a paradox discovered in recent work by Phillips and Su (2009). That paper gave an example in which nonparametric regression is consistent whereas parametric regression is inconsistent even when the true regression functional form is known and used in regression. This appears to be a paradox, as knowing the true functional form should not in general be detrimental in regression. In the present case, local regression methods turn out to have a distinct advantage because of endogeneity in the regressor. The paradox arises because additional correct information is not necessarily advantageous when information is incomplete. In the present case, endogeneity in the regressor introduces bias when the true functional form is known, but interestingly does not do so in local nonparametric regression. We examine this example in detail and propose two new consistent estimators for the parametric regression, which address the endogeneity in the regressor by means of spatial bounding and bias correction using nonparametric estimation. Some simulations are reported illustrating the paradox and the new procedures.

JEL classification: C13, C14.

Keywords: Bias-correction, Endogeneity, Kernel regression, \mathcal{L}_2 regression, Location shift, Nonparametric IV, Nonstationarity, Paradox, Spatial regression, Structural estimation.

*Phillips acknowledges partial support from NSF Grant #SES 06-47086. Address correspondence to Peter C. B. Phillips, Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, Connecticut USA 06520-8281; e-mail: peter.phillips@yale.edu. Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; e-mail: ljsu@smu.edu.sg.

1 Introduction and Motivation

Recent work (Phillips and Su, 2009, hereafter PS) drew attention to an interesting example in which parametric regression is inconsistent when the true functional form is known and yet, surprisingly, nonparametric regression is consistent. This note explores that example and the resulting paradox in more detail and suggests modifications to parametric regression which remove the inconsistency.

The example given in PS involves cross section regression with some systematic location shifts in an endogenous regressor. The location shifts may be regarded as a form of instrumental variable effect that influences observations of the regressors and assists in identifying the regression curve. In particular, in the linear regression model

$$y_i = \beta_0 + \beta_1 X_i + u_i, \quad E(u_i|X_i) \neq 0, \quad E(u_i) = 0, \quad (1.1)$$

the endogenous regressor X_i is assumed to satisfy

$$X_i = \mu_\alpha 1\{i \in A_\alpha\} + u_{xi}, \quad (1.2)$$

and, by virtue of its moving intercept

$$\mu_\alpha = \frac{\alpha L_n}{2m}, \quad \alpha = -m, -m+1, \dots, m, \quad (1.3)$$

X_i is subject to $(2m+1)$ location shifts that are equally spaced over the interval $(-L_n/2, L_n/2)$. The u_{xi} are random quantities that may have compact $[\underline{u}, \bar{u}]$ or infinite $(-\infty, \infty)$ support. The quantity L_n may be fixed or may increase slowly as $n \rightarrow \infty$ (e.g., $L_n = \log n$). The parameter m passes to infinity so that the distance $(\frac{L_n}{2m})$ between two contiguous location shifts shrinks to zero as $m \rightarrow \infty$. Data also accumulates at each location and we define

$$M = \# \{i \in A_\alpha\},$$

although it is not necessary that M be fixed across different locations. The total observation count is then $n = (2m+1)M$. As in PS, we require $m \rightarrow \infty$ but M can either be finite or pass to ∞ and this dual possibility is denoted by writing $(M, m) \rightarrow \infty$.

Model (1.1) is a simple example of a structural equation with an endogenous regressor. The endogeneity is obviously a source of potential bias in regression, particularly in the absence of an observed instrumental variable. Curiously, not knowing the functional form of (1.1) and performing nonparametric kernel regression produces a consistent estimator in spite of the endogeneity, whereas conventional least squares regression which uses functional form information is inconsistent.

In addition to the linear model (1.1), PS also considered the nonlinear structural equation

$$y_i = g(X_i) + u_i, \quad E(u_i|X_i) \neq 0, \quad E(u_i) = 0, \quad (1.4)$$

where the endogenous regressor X_i is generated via (1.2) and (1.3). PS studied the local level (Nadaraya-Watson) kernel estimator of $g(x)$

$$\widehat{g}(x) = \frac{\frac{1}{n} \sum_{i=1}^n y_i K_h(X_i - x)}{\frac{1}{n} \sum_{i=1}^n K_h(X_i - x)}, \quad (1.5)$$

where $K(\cdot)$ is a kernel function, $K_h(\cdot) = h^{-1}K(\cdot/h)$ and $h \equiv h(M, m)$ is a bandwidth parameter. Let $g'(x)$ and $g''(x)$ denote the first and second derivatives of $g(x)$, respectively. Let $f'_{u_x}(\cdot)$ denote the first derivative of the marginal probability density function (p.d.f.) of u_{xi} , and $f''_2(u, u_x)$ the second order partial derivative with respect to u_x of the p.d.f. $f(\cdot, \cdot)$ of (u_i, u_{xi}) . The central finding of PS is that the nonparametric estimator $\widehat{g}(x)$ is consistent for $g(x)$ under some standard regularity conditions, as detailed in the following theorem.

Theorem 1.1 *Under Assumptions A1-A6 of PS with m^λ replaced by $d_m = 2m/L_n$, or Assumptions A1-A2, A3(i), A4, A5 and A7-A9 of PS with L replaced by L_n , we have*

$$\begin{aligned} \sqrt{Md_m h} \left(\widehat{g}(x) - g(x) - h^2 \mu_2(K) \left\{ g'(x) \int_{\underline{u}}^{\bar{u}} f'_{u_x}(p) dp + \frac{1}{2} g''(x) \right\} \right) \\ \rightarrow_d N(0, \sigma^2 \nu_2(K)), \end{aligned} \quad (1.6)$$

where $\mu_2(K) = \int x^2 K(x) dx$, $\nu_2(K) = \int K(x)^2 dx$, and $\sigma^2 = E(u_i^2)$.

If the distribution of u_{xi} has infinite support, so that $f_{u_x}(p)$ vanishes at infinity, then $\int_{-\infty}^{\infty} f'_{u_x}(p) dp = 0$ and the linear bias term disappears in (1.6). Despite endogeneity in the regressor, the local level estimator is consistent and asymptotically normally distributed. The price for the presence of endogeneity in the regression is the potentially slower rate of convergence. The convergence rate of the local level estimator relies on $Md_m = 2Mm/L_n \sim n/L_n$. So the convergence rate, in the case of increasing L_n , is slower than the usual \sqrt{nh} -rate of convergence for local level estimate with a stationary regressor, which is related to the findings of Wang and Phillips (2009) for nonparametric structural cointegrating regression, where again nonparametric regression is consistent. As PS argue, the location shifts act in a manner analogous to the random wandering feature of unit root regressors in a cointegrating regression equation and add variation to the regressor, and thereby explaining the consistency of simple nonparametric regression in both cases. Nevertheless, if the support of the distribution of u_{xi} is compact, as is conventionally assumed in the nonparametric IV literature (e.g., Hall and Horowitz, 2005), the locational range L_n can be a large fixed constant, and in this event simple local level regression is consistent and the usual \sqrt{nh} nonparametric rate of convergence is attained. This outcome contrasts with the attainable convergence rates for nonparametric IV estimation, which can be slow.

Applying the above result to the linear regression (1.1), we see that the nonparametric local level estimate $\widehat{g}(x)$ of $g(x) = \beta_0 + \beta_1 x$ is consistent. Alternatively, if the functional form of

the regression is known, ordinary least squares regression can be applied to (1.1), giving

$$\hat{\beta}_1 = \sum_{i=1}^n (X_i - \bar{X}) y_i / \sum_{i=1}^n (X_i - \bar{X})^2, \quad \hat{\beta}_0 = \bar{y} - \bar{X} \hat{\beta}_1,$$

where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$, and $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. The parametric estimate of $g(x) = \beta_0 + \beta_1 x$ is then given by $\hat{g}^p(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. Somewhat surprisingly, as remarked by PS, $(\hat{\beta}_0, \hat{\beta}_1)$ is inconsistent for (β_0, β_1) when L_n is fixed and, correspondingly, $\hat{g}^p(x)$ is inconsistent for $g(x)$ at all points except $x = E(u_{xi})$. When $L_n \rightarrow \infty$ as $n \rightarrow \infty$, consistency of the parametric estimate can be achieved but at the cost of the presence of a substantial bias term which seems difficult to eliminate. This outcome is paradoxical, as local level regression is consistent whereas linear parametric regression that uses the true functional form and global information is inconsistent. Clearly, using all information is costly here. The reason, of course, is the endogeneity in the regressor (1.1). Intriguingly, however, partial information usage of the type employed in kernel regression avoids most of the problems associated with parametric regression.

To illustrate the phenomenon, we generate a sample of $n = 500$ observations from the following system

$$\begin{aligned} y_i &= \beta_0 + \beta_1 X_i + u_i, \quad \beta_0 = 10, \quad \beta_1 = -1, \\ X_i &= \mu_\alpha 1\{i \in A_\alpha\} + u_{xi}, \\ u_i &= 2(\epsilon_{yi} + \gamma u_{xi}) / (1 + \gamma^2)^{1/2}, \end{aligned}$$

where $\gamma = 2.07$, (ϵ_{yi}, u_{xi}) are *iid* $N(0, I_2)$, and $\mu_\alpha \in \{-4, -2, 0, 2, 4\}$. Using the above notation, this corresponds to the case where $m = 2$, and $L_n = 8$. When $\gamma = 2.07$, the error correlation coefficient is $\text{corr}(u_i, u_{xi}) = 0.9$, implying strong endogeneity in the structural equation. Fig 1 provides a typical sample plot of data generated from this system. Also displayed are the true linear regression line, the local level nonparametric estimate (using a Gaussian kernel and Silverman's rule of thumb for the bandwidth choice) and the fitted OLS estimate obtained for these data. Clearly, the local level estimate considerably outperforms the parametric estimate of the regression line over a wide range of the support of the regressor. But at the tails of the support the endogeneity bias becomes manifest and the location shifts lose power in identifying the regression line. Since it uses local information and does so increasingly as the sample size rises, the local level estimator at interior points of the domain of the regressor very effectively attenuates distortion from tail observations. On the other hand, parametric linear regression, which treats all observations as equally important and applies global information in fitting the regression line, is inevitably subject to potential distortions from tail observations.

This example makes clear that local nonparametric regression has more robustness advantages beyond robustness to specific functional form, for which it is commonly celebrated. As shown here, nonparametric regression may also display a robustness to endogeneity in a regression by concentrating attention on local information and attenuating tail information that may be more heavily subjected to endogeneity effects.

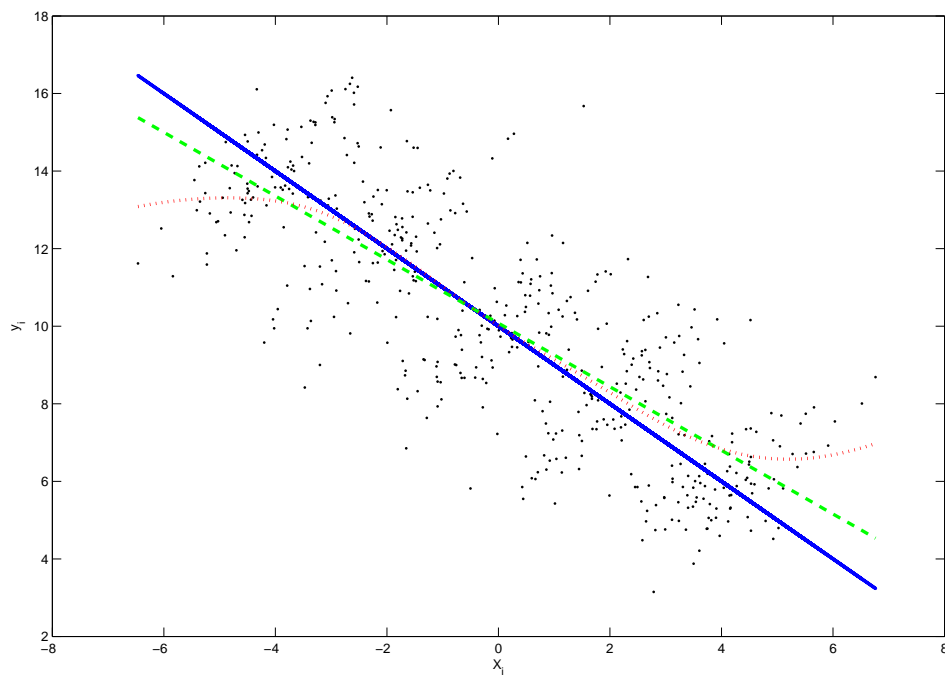


Figure 1: Local level nonparametric (dotted) and linear regression (dashed) estimates of the true regression line (solid) using the full sample (scatter plot) over all locations.

Intuitively, any regression approach like conventional parametric regression that uses global information can be subject to distortionary effects from outlying observations. Such behavior is very well known in statistics. Anscombe (1960) coined the term ‘outlier’, provided a brief history of the subject, and suggested trimming techniques to attenuate the effects of outliers in regression based on the insurance analogy of ‘protection and premium’ to guard against unwanted effects. In the present context, the reason for the outlier effect is that at the limits of the domain of definition, the observations are more affected by the endogeneity of X_i , so bias arising from the ends of the domain can dominate a global regression and result in inconsistency. By contrast, in nonparametric regression mainly local information is used in estimation so the endogeneity effects of X_i in the tail can be well controlled and first order bias, at least, can be eliminated in local regression. In effect, by concentrating attention on the cluster of observations around each point, nonparametric regression localizes attention and removes outlier effects. This heuristic suggests that to recover the true regression line by a parametric method, a natural approach is to modify the regression by removing the effects of tail observations. The idea is comparable to that of trimming or Winsorizing the data, on which there is a large literature in statistics stemming largely from Anscombe’s (1960) study (e.g. Welsh, 1987; Chen, Welsh and Chan, 2001). In the present context, to use Anscombe’s analogy, the idea is to provide protection (against the possible effects of endogeneity) by paying a premium in terms of losing some observations. Kernel regression accomplishes this task by using data that is effectively in the locality of each individual regression point, thereby sacrificing an (asymptotically larger) infinity of observations to achieve a local regression fit and, incidentally in the process, protection from the effects of endogeneity.

The remainder of the paper is organized as follows. Section 2 studies the asymptotic properties of the OLS estimator $(\hat{\beta}_0, \hat{\beta}_1)$, demonstrates its inconsistency for the fixed L_n case, and derives the asymptotic distribution of $(\hat{\beta}_0, \hat{\beta}_1)$. Section 3 proposes two new consistent estimators of (β_0, β_1) when L_n is either fixed or increasing slowly as $n \rightarrow \infty$. The first is a spatial \mathcal{L}_2 estimator which is obtained by regressing the nonparametric local level estimate $\hat{g}(x)$ on $(1, x)$, using a continuous number of pseudo-observations on $(x, \hat{g}(x))$ where x is spatially restricted to be bounded away from the two tails. The second is a bias-corrected OLS estimator of (β_0, β_1) that is based on the spatial \mathcal{L}_2 regression residuals. For both cases, we show that the resulting estimators are consistent and asymptotically normally distributed. The consistency rate is parametric in the case where $L_n = L$ is fixed, and $\sqrt{n/L_n}$ in the case where $L_n \rightarrow \infty$ as $n \rightarrow \infty$. Section 4 reports some simulation evidence and Section 5 concludes. Proofs of the main results are given in Section 6.

2 Limit Theory for Parametric Regression

To study the asymptotic properties of the OLS estimator $(\widehat{\beta}_0, \widehat{\beta}_1)$, we make the following assumptions.

Assumption 1.

- (i) (u_i, u_{xi}) , $i = 1, \dots, n$, are independent and identically distributed (iid).
- (ii) $E(u_i) = 0$, and $E(u_i^2) = \sigma^2$.
- (iii) $E(u_{xi}) = \mu_x$, and $\text{Var}(u_{xi}) = \sigma_x^2$.

The above assumption is fairly standard in cross-sectional regression, although we do not make allowance for unconditional heterogeneity in (u_i, u_{xi}) . Note that it is not assumed that the mean of u_{xi} is zero, and conditional homoskedasticity is not assumed for the error term u_i in the structural equation.

2.1 Inconsistency of $(\widehat{\beta}_0, \widehat{\beta}_1)$

Under Assumption 1, we first derive the probability limit of the OLS estimator $\widehat{\beta}_1$ and show that it is inconsistent for β_1 when L_n is fixed. The probability limit of $\widehat{\beta}_0$ follows straightforwardly.

Write

$$\widehat{\beta}_1 = \beta_1 + \frac{n^{-1} \sum_{i=1}^n (X_i - \bar{X}) u_i}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad (2.1)$$

First, by the definition of $\{\mu_\alpha\}$ and the WLLN, we have

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{2m+1} \sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \left[\frac{L_n \alpha}{2m} + (u_{xi} - \bar{u}_x) \right]^2 \\ &= \frac{L_n^2}{4m^2(2m+1)} \sum_{\alpha=-m}^m \alpha^2 + \frac{1}{n} \sum_{i=1}^n (u_{xi} - \bar{u}_x)^2 + \frac{L_n}{m(2m+1)} \sum_{\alpha=-m}^m \frac{\alpha}{M} \sum_{i \in A_\alpha} (u_{xi} - \bar{u}_x) \\ &= \frac{L_n^2}{2m^2(2m+1)} \sum_{\alpha=1}^m \alpha^2 + \sigma_x^2 + o_P(1) \\ &= \frac{L_n^2}{12} \{1 + o(1)\} + \sigma_x^2 + o_P(1), \end{aligned} \quad (2.2)$$

where $\bar{u}_x = n^{-1} \sum_{i=1}^n u_{xi}$, and the third line follows from the WLLN and Chebyshev inequality provided $Mm/L_n^2 \rightarrow \infty$ as $n \rightarrow \infty$.¹

¹Letting $T_n \equiv \frac{L_n}{m(2m+1)} \sum_{\alpha=-m}^m \frac{\alpha}{M} \sum_{i \in A_\alpha} (u_{xi} - \mu_x) = \frac{L_n}{m(2m+1)} \sum_{\alpha=-m}^m \frac{\alpha}{M} \sum_{i \in A_\alpha} (u_{xi} - \bar{u}_x)$, then $E(T_n) = 0$, and $\text{Var}(T_n) = \frac{L_n^2}{Mm^2(2m+1)^2} \sum_{\alpha=-m}^m \alpha^2 \sigma_x^2 = O\left(\frac{L_n^2}{Mm}\right) = o(1)$.

Next

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) u_i \\
&= \frac{1}{2m+1} \sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \left(\frac{L_n \alpha}{2m} + u_{xi} - \bar{u}_x \right) u_i \\
&= \frac{L_n}{2m(2m+1)} \sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \alpha u_i + \frac{1}{n} \sum_{i=1}^n u_{xi} u_i - \frac{\bar{u}_x}{n} \sum_{i=1}^n u_i \\
&= \frac{L_n}{2m(2m+1)} \sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \alpha u_i + E(u_{xi} u_i) + o_P(1).
\end{aligned}$$

Consider the first term in the last expression. Let $T_{1n} = \frac{L_n}{2m(2m+1)} \sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \alpha u_i$. Then $E(T_{1n}) = 0$ as $E(u_i) = 0$, and

$$\begin{aligned}
\text{Var}(T_{1n}) &= \frac{L_n^2}{4m^2(2m+1)^2} \text{Var} \left(\sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \alpha u_i \right) \\
&= \frac{L_n^2 \sigma^2}{4Mm^2(2m+1)^2} \sum_{\alpha=-m}^m \alpha^2 = O \left(\frac{L_n^2}{Mm} \right) = o(1).
\end{aligned}$$

Hence $T_{1n} = o_P(1)$ and

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) u_i = E(u_{xi} u_i) + o_P(1). \tag{2.3}$$

Combining (2.1), (2.2) and (2.3) yields

$$\hat{\beta}_1 = \beta_1 + \frac{E(u_{xi} u_i) + o_P(1)}{\frac{L_n^2}{12} \{1 + o(1)\} + \sigma_x^2 + o_P(1)} = \beta_1 + \frac{E(u_{xi} u_i)}{\frac{L_n^2}{12} + \sigma_x^2} \{1 + o_P(1)\}, \tag{2.4}$$

thereby giving the following result.

Lemma 2.1 *Suppose Assumption 1 holds and $Mm/L_n^2 \rightarrow \infty$. Then*

$$\hat{\beta}_1 = \beta_1 + \frac{E(u_{xi} u_i)}{\frac{L_n^2}{12} + \sigma_x^2} \{1 + o_P(1)\}.$$

The following remarks explore the implications of the above lemma, considering the two cases where $L_n = L$ is fixed and $L_n \rightarrow \infty$ as $n \rightarrow \infty$.

Remark 1. In the first case ($L_n = L$ fixed), $\hat{\beta}_1$ has the probability limit

$$\text{plim}_{n \rightarrow \infty} \hat{\beta}_1 = \beta_1 + \frac{E(u_{xi} u_i)}{\frac{L^2}{12} + \sigma_x^2},$$

and is inconsistent unless $E(u_{xi}u_i) = 0$, viz., X_i is exogenous. For $\widehat{\beta}_0$, we have

$$\begin{aligned}\widehat{\beta}_0 - \beta_0 &= -\overline{X} \left(\widehat{\beta}_1 - \beta_1 \right) + n^{-1} \sum_{i=1}^n u_i \\ &\rightarrow_p -\mu_x \frac{E(u_{xi}u_i)}{\frac{L^2}{12} + \sigma_x^2}.\end{aligned}\tag{2.5}$$

so $\widehat{\beta}_0$ is inconsistent for β_0 unless either $\mu_x = 0$ or $E(u_{xi}u_i) = 0$. Hence, the parametric estimator $\widehat{g}^p(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is inconsistent for $g(x) = \beta_0 + \beta_1 x$ at all points except $x = \mu_x$. By contrast, according to Theorem 1.1, the nonparametric estimator is consistent for all x satisfying certain domain restrictions.

Remark 2. In the second case ($L_n \rightarrow \infty$), the OLS estimator $\widehat{\beta}_1$ is consistent for β_1 as

$$\widehat{\beta}_1 = \beta_1 + \frac{12E(u_{xi}u_i)}{L_n^2} \{1 + o_P(1)\} \rightarrow_p \beta_1,$$

due to the strengthening signal in the regressor as $L_n \rightarrow \infty$. This result, together with (2.5), implies that $\widehat{\beta}_0$ is consistent for β_0 , and so the parametric regression estimator of $g(x) = \beta_0 + \beta_1 x$ is also consistent. However, if L_n diverges to infinity slowly like $L_n = \log n$, the estimation bias may disappear at a very slow rate. For inferential purposes, we now derive the limit distribution of $(\widehat{\beta}_0, \widehat{\beta}_1)$.

2.2 Limit distribution

To find the limit distribution of $(\widehat{\beta}_0, \widehat{\beta}_1)$, we add the following assumption.

Assumption 2. $E[|u_i|^{2+\delta}] < \infty$ and $E[|u_i u_{xi}|^{2+\delta}] < \infty$ for some $\delta > 0$.

Let $\theta = (\beta_0, \beta_1)'$ and $\widehat{\theta} = (\widehat{\beta}_0, \widehat{\beta}_1)'$. Let $\underline{X}_i = (1, X_i)'$, $\mathbf{X} = (\underline{X}_1, \dots, \underline{X}_n)'$, $\mathbf{u} = (u_1, \dots, u_n)'$, and $\mathbf{y} = (y_1, \dots, y_n)'$. Define $D_n = \text{diag}(1, L_n)$,

$$\Gamma = \lim_{n \rightarrow \infty} \begin{pmatrix} 1 & \frac{\mu_x}{L_n} \\ \frac{\mu_x}{L_n} & \frac{1}{12} + \frac{E(u_{xi}^2)}{L_n^2} \end{pmatrix}, \text{ and } \Omega = \lim_{n \rightarrow \infty} \begin{pmatrix} \sigma^2 & \frac{E(u_{xi}u_i^2)}{L_n} \\ \frac{E(u_{xi}u_i^2)}{L_n} & \frac{\sigma^2}{12} + \frac{\text{Var}(u_{xi}u_i)}{L_n^2} \end{pmatrix}.$$

After centering, the limiting distribution of $\widehat{\theta}$ is given in the following theorem.

Theorem 2.2 *Suppose Assumptions 1-2 hold and $Mm/L_n^2 \rightarrow \infty$. Then*

$$\sqrt{n}D_n \left(\widehat{\theta} - \theta - (\mathbf{X}'\mathbf{X})^{-1} E(\mathbf{X}'\mathbf{u}) \right) \rightarrow_d N(0, \Gamma^{-1}\Omega\Gamma'^{-1}).\tag{2.6}$$

Remark 3. Straightforward calculations show that

$$(\mathbf{X}'\mathbf{X})^{-1} E(\mathbf{X}\mathbf{u}) = \frac{E(u_{xi}u_i)}{n^{-1} \sum_{i=1}^n (X_i - \overline{X})^2} \begin{pmatrix} -\overline{X} \\ 1 \end{pmatrix},$$

and

$$\Gamma^{-1}\Omega\Gamma'^{-1} = \lim_{n \rightarrow \infty} c_n^{-2} \begin{pmatrix} \omega_{11n} & \omega_{12n} \\ \omega_{12n} & \omega_{22n} \end{pmatrix},$$

where $c_n = \frac{1}{12} + \frac{\sigma_x^2}{L_n^2}$,

$$\begin{aligned} \omega_{11n} &= \frac{\sigma^2}{144} + \frac{2E(u_{x1}^2)\sigma^2 + \mu_x^2\sigma^2 - 2\mu_x E(u_{x1}u_1^2)}{12L_n^2} \\ &\quad + \frac{[E(u_{x1}^2)]^2\sigma^2 + \mu_x^2\text{Var}(u_{x1}u_1) - 2\mu_x E(u_{x1}^2)E(u_{x1}u_1^2)}{L_n^4}, \\ \omega_{12n} &= \frac{\sigma^2}{12} - \frac{\mu_x\sigma^2}{12} + \frac{\text{Var}(u_{x1}u_1) - 2\mu_x E(u_{x1}u_1^2)}{L_n^2} - \frac{E(u_{x1}^2)\mu_x\sigma^2}{L_n^3}, \text{ and} \\ \omega_{22n} &= \frac{\sigma^2}{12} + \frac{\text{Var}(u_{x1}u_1) + \mu_x^2\sigma^2 - 2\mu_x E(u_{x1}u_1^2)}{L_n^2} = \frac{\sigma^2}{12} + \frac{\text{Var}((u_{x1} - \mu_x)u_1)}{L_n^2}. \end{aligned}$$

An immediate implication of Theorem 2.2 is therefore that

$$\sqrt{n} \left(\hat{\beta}_0 - \beta_0 + \frac{\bar{X} E(u_{x1}u_1)}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2} \right) \rightarrow_d N \left(0, \lim_{n \rightarrow \infty} c_n^{-2} \omega_{11n} \right), \quad (2.7)$$

$$L_n \sqrt{n} \left(\hat{\beta}_1 - \beta_1 - \frac{E(u_{x1}u_1)}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2} \right) \rightarrow_d N \left(0, \lim_{n \rightarrow \infty} c_n^{-2} \omega_{22n} \right), \quad (2.8)$$

If $L_n \rightarrow \infty$ as $n \rightarrow \infty$, then the above calculation simplifies and

$$\Gamma^{-1}\Omega\Gamma'^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{12} \end{pmatrix}^{-1} \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{12} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{12} \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 12 \end{pmatrix}.$$

In this case, the intercept estimator, after centering, is asymptotically independent of the slope estimator, and the slope coefficient estimator has a faster convergence rate, due to the stronger signal in the regressor. In contrast, in the fixed L_n case, the two estimators are asymptotically dependent, have the same rate of convergence, and the range of location shifts contributes to the asymptotic variance formula in (2.6) in a complicated way.

Remark 4. In spite of the $O(\sqrt{n})$ convergence rate for the intercept estimator and the $O(L_n\sqrt{n})$ convergence rate for the slope estimator, the result in (2.6) does not seem useful for inferential purposes because the bias term in (2.6) does not appear to be estimable at the required \sqrt{n}/L_n^2 -rate (for the intercept parameter) or \sqrt{n}/L_n -rate (for the slope parameter) to be eliminated (recall from (2.2) that $n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2 = O_P(L_n^2)$). It is worth mentioning that the residuals $\{\hat{u}_i\}$ from the OLS regression are useless in constructing a consistent estimate of $E(X_i u_i)$ because of the orthogonality of \hat{u}_i and X_i . One may consider consistent estimation of $E(X_i u_i)$ by estimating the model first via the local level (or local linear) nonparametric method at interior points and then obtaining the nonparametric residuals from the structural equation. Unfortunately, this approach usually requires uniform consistency of the local level (or local

linear) estimate over the whole support of the regressor, which seems extremely difficult here given the fact that the variance of the regressor is expanding as the sample size increases or that the nonparametric estimates are only consistent at points within a subset of the interior of the support of the regressor for the fixed L_n case.

Instead, the next section proposes two alternative methods to achieve $\sqrt{n/L_n}$ -consistent estimation of (β_0, β_1) by direct use of the nonparametric level estimate in a parametric regression.

3 Spatial \mathcal{L}_2 and Bias-Corrected OLS Estimation

In this section we propose two methods for consistent estimation of (β_0, β_1) . The first method is a spatial \mathcal{L}_2 regression and the second method involves bias-corrected OLS estimation. The \mathcal{L}_2 method treats the linear regression function as unknown, estimates it nonparametrically by $\hat{g}(x)$, and then regresses $\hat{g}(x)$ on $(1, x)$ to estimate the unknown parameter (β_0, β_1) by minimizing a spatial \mathcal{L}_2 criterion function, using a continuum of pseudo-observations on $(x, \hat{g}(x))$ where x is restricted to be bounded away from the two tails. We prove that the resulting \mathcal{L}_2 estimator is $\sqrt{n/L_n}$ -consistent, where L_n can be either fixed or pass to infinity as $n \rightarrow \infty$. In either case, we show that the OLS bias terms are corrected by using the residuals from the \mathcal{L}_2 regression, and the bias-corrected OLS estimators can only attain an $O(\sqrt{n/L_n})$ consistency rate, which is inherited from that of the \mathcal{L}_2 estimates.

3.1 Spatial \mathcal{L}_2 regression

Noting that the local level estimate $\hat{g}(x)$ is consistent for $g(x) = \beta_0 + \beta_1 x$ in the interior of the regressor support, we propose to estimate the unknown parameter $\theta \equiv (\beta_0, \beta_1)'$ by minimizing the following (spatial) \mathcal{L}_2 criterion

$$S_n(\beta_0, \beta_1) = \int_a^b (\hat{g}(x) - \beta_0 - \beta_1 x)^2 \hat{f}(x) dx \quad (3.1)$$

where a and b are finite integration limits that serve to truncate observations in the two tails, $\hat{f}(x) = N^{-1} \sum_{i=1}^n K_h(x - X_i)$ is a pseudo-estimate of the “density” of X_i , and $N = 2mM/L_n$ signifies the effective number of observations used in the nonparametric estimation, which does not need to be observed in practice for implementation. Note that $\hat{f}(x)$ is a weight function that serves to avoid division by zero and to perform trimming in areas of sparse support, and $[a, b]$ defines a compact set on which the nonparametric estimates $\hat{g}(x)$ are used in the estimation of (β_0, β_1) . Clearly, (3.1) provides a trimming operation implicitly via the local nature of the estimate $\hat{g}(x)$ and explicitly via the use of the (truncated) domain $[a, b]$.

The minimizer of (3.1) $\tilde{\theta} \equiv (\tilde{\beta}_0, \tilde{\beta}_1)'$ is given by

$$\tilde{\theta} = \begin{pmatrix} \int_a^b \hat{f}(x) dx & \int_a^b x \hat{f}(x) dx \\ \int_a^b x \hat{f}(x) dx & \int_a^b x^2 \hat{f}(x) dx \end{pmatrix}^{-1} \begin{pmatrix} \int_a^b \hat{g}(x) \hat{f}(x) dx \\ \int_a^b x \hat{g}(x) \hat{f}(x) dx \end{pmatrix}.$$

To develop the limit theory, we make the following assumptions.

Assumption 3. *The probability density function (p.d.f.) $f(\cdot, \cdot)$ of (u_i, u_{xi}) exists. $f(\cdot, \cdot)$ has second order partial derivative $f_2''(u, u_x)$ with respect to u_x such that $f_2''(u, u_x)$ is continuous in u_x and $\int \int |u f_2''(u, u_x)| du du_x < \infty$. The marginal p.d.f. of u_{xi} , $f_{u_x}(\cdot)$, has second order continuous derivatives such that $\int_{-\infty}^{\infty} |f_{u_x}'(p)| dp < \infty$, and $\int_{-\infty}^{\infty} |f_{u_x}''(p)| dp < \infty$.*

Assumption 4. *Either one of the following conditions holds:*

(i) *The error term u_{xi} has infinite support such that there exists a majorizing function $C_f(\cdot)$ and a diverging sequence $c_n = c_n(L_n)$ such that $|\int_{-\infty}^{-L_n} f(u, u_x) du_x + \int_{L_n}^{\infty} f(u, u_x) du_x| \leq c_n^{-1} C_f(u)$ with $c_n^{-1} = O(h^2)$ and $\int_{-\infty}^{\infty} |u| C_f(u) du < \infty$. $L_n \rightarrow \infty$ as $n \rightarrow \infty$, and $f_{u_x}(L_n) = O(h^2)$ as $L_n \rightarrow \infty$;*

(ii) *The error term u_{xi} has compact support, i.e., $u_{xi} \in [\underline{u}, \bar{u}]$ a.s. for some finite numbers \underline{u} and \bar{u} . L_n is either fixed or tends to ∞ as $n \rightarrow \infty$. If $L_n = L$ is fixed, L is sufficiently large that $x \in (-L/2 + \bar{u}, L/2 + \underline{u})$ for all $x \in [a, b]$.*

Assumption 5. *The kernel function $K(x)$ is a uniformly bounded, symmetric p.d.f. such that $\int x^4 K(x) dx < \infty$.*

Assumption 6. *As $(M, m) \rightarrow \infty$, $Mm/L_n^2 \rightarrow \infty$, $Mmh^4/L_n \rightarrow 0$, $Mm^{-3}L_n^3 \rightarrow 0$, and $N^{-\delta/2}L_n \rightarrow 0$.*

Assumptions 3-4 are comparable to Assumptions A3 and A7-A9 in PS. Assumption 5 is standard and the symmetry assumption greatly simplifies derivations. Note that we impose undersmoothing ($Mmh^4/L_n \rightarrow 0$) on the bandwidth in Assumption 6. The last requirement in Assumption 6 is needed to verify the Liapounov condition.

Theorem 3.1 *Suppose Assumptions 1-6 hold. Then*

$$\sqrt{N}(\tilde{\theta} - \theta) \rightarrow_d N(0, \sigma^2 Q^{-1}), \quad (3.2)$$

$$\text{where } Q = \begin{pmatrix} b-a & \frac{b^2-a^2}{2} \\ \frac{b^2-a^2}{2} & \frac{b^3-a^3}{3} \end{pmatrix}.$$

Remark 5. *Despite the nonparametric convergence rate of the regressand $\hat{g}(x)$, Theorem 3.1 indicates that the \mathcal{L}_2 estimate $\tilde{\theta}$ is \sqrt{N} -consistent. It achieves the parametric \sqrt{n} -rate of consistency for the case of fixed L_n and its rate is only slightly worse than the parametric rate in the case where L_n is an increasing slowly varying function at infinity. In addition, $\tilde{\theta}$ is not subject to any non-negligible asymptotic bias term. To obtain these results in Theorem 3.1, we*

have applied two tricks implicitly. First, undersmoothing is required to eliminate the $O(h^2)$ bias terms from the first stage nonparametric regression estimate of $g(x)$. This is standard in the nonparametric or semiparametric literature when the first stage nonparametric estimates are used in a second stage parametric or nonparametric estimation. Second, to reduce the variation of the nonparametric estimates, we have used integration in the \mathcal{L}_2 regression. The smoothing operation of integration helps to produce the (nearly) parametric convergence rate of $\tilde{\theta}$ despite the slow nonparametric convergence rate of $\hat{g}(x)$. The mechanism is analogous to that of average marginal effect or derivative estimation.

Remark 6. Even though the structural error terms may be conditionally heteroskedastic, the asymptotic variance of $\tilde{\theta}$ only depends on the unconditional variance σ^2 . For inferential purposes, we need to choose (a, b) in the \mathcal{L}_2 regression and estimate σ^2 . Let $\tilde{u}_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_i$. We can estimate σ^2 by $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{u}_i^2$. It is straightforward to show that $\tilde{\sigma}^2$ is consistent for σ^2 . With this estimate, in principle one can obtain a consistent estimate of the asymptotic covariance matrix of $\sqrt{N}(\tilde{\theta} - \theta)$ using $\tilde{\sigma}^2 Q^{-1}$. But since N is not observed, this last estimate is not directly applicable in statistical inference. One approach is to replace Q by its consistent estimate

$$Q_n = \begin{pmatrix} \int_a^b \hat{f}(x) dx & \int_a^b x \hat{f}(x) dx \\ \int_a^b x \hat{f}(x) dx & \int_a^b x^2 \hat{f}(x) dx \end{pmatrix}$$

because the quantity N that appears in the definition of $\hat{f}(x) = N^{-1} \sum_{i=1}^n K_h(x - X_i)$ and also in the scaling of the estimate, $\sqrt{N}(\tilde{\theta} - \theta)$, will cancel in inference. More explicitly, consider testing the null hypothesis $H_0 : R\theta = r$, where R is a full rank $k \times 2$ matrix and r is $k \times 1$ vector. Using Q_n we can construct the following Wald statistic as usual

$$W_n = \tilde{\sigma}^{-2} (R\tilde{\theta} - r)' (NQ_n) (R\tilde{\theta} - r),$$

and without knowledge of N .

Remark 7. For the integration limits, we recommend choosing $a = F_X^{-1}(\lambda)$ and $b = F_X^{-1}(1 - \lambda)$ where $F_X^{-1}(\lambda)$ denotes the sample λ -th quantile of $\{X_i\}$ and $\lambda \in (0, 0.5)$, although there is no reason for symmetric truncations in the two tails. Since in the extreme tails the nonparametric estimates $\hat{g}(x)$ are highly distorted, so we recommend $\lambda \geq 0.05$ depending on the number of observations n . The larger λ , the greater the proportion of observations that are trimmed and the greater the efficiency loss that results. We therefore do not want to trim too many observations and recommend $\lambda \leq 0.30$. In the simulations below, we study the effect of the truncation parameter (λ) on the \mathcal{L}_2 and bias-corrected OLS estimators. We find that for sample sizes $n = 500 \sim 2000$, the choice $\lambda = 0.15$ works fairly well. It is worth mentioning that in the above theory, we only establish the asymptotic result for fixed (a, b) . We conjecture the theory also works when one allows the integration range to expand slowly as $n \rightarrow \infty$, under suitable controls on the rate of expansion.

3.2 Bias-corrected OLS estimation

We now propose a bias-correction procedure for the OLS estimator of $\theta = (\beta_0, \beta_1)'$ in the linear structural equation (1.1). Let $\tilde{c} = n^{-1} \sum_{i=1}^n \tilde{u}_i X_i$. We define bias-corrected OLS estimators of β_0 and β_1 , respectively, as

$$\hat{\beta}_{0c} = \hat{\beta}_0 + \frac{\bar{X} \tilde{c}}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2} \text{ and } \hat{\beta}_{1c} = \hat{\beta}_1 - \frac{\tilde{c}}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

Due to the $\sqrt{n/L_n}$ -rate of convergence of $\tilde{\theta} = (\tilde{\beta}_0, \tilde{\beta}_1)'$, the \mathcal{L}_2 residuals \tilde{u}_i only converge to the true structural errors u_i at a $\sqrt{n/L_n}$ -rate (viz. $\tilde{u}_i - u_i = O_P(\sqrt{L_n/n})$). As a result, $n^{-1} \sum_{i=1}^n \tilde{u}_i X_i - E(u_i u_{xi}) = O_P(\sqrt{L_n^5/n})$ due to the fact that $n^{-1} \sum_{i=1}^n X_i^2 = O_P(L_n^2)$, which is not the $o_P(L_n/\sqrt{n})$ rate required for the complete removal of the bias of the OLS estimator $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1)'$ in (2.7) and (2.8). This indicates that $\hat{\theta}_c \equiv (\hat{\beta}_{0c}, \hat{\beta}_{1c})'$ does not have the same standardization and asymptotic variance as $\hat{\theta}$.

Nevertheless, after re-scaling, we can show that $\hat{\theta}_c$ is $\sqrt{n/L_n}$ -consistent for θ and it is asymptotically normally distributed with variance that can be easily estimated. The following theorem establishes the consistency and asymptotic normality of $\hat{\theta}_c$.

Theorem 3.2 *Suppose Assumptions 1-3, and 5-6 hold. (i) If Assumption 4(i) holds or 4(ii) holds with $L_n \rightarrow \infty$ as $n \rightarrow \infty$, then $\sqrt{N}(\hat{\theta}_c - \theta) \rightarrow_d N(0, \Psi)$, where*

$$\Psi = \sigma^2 q^{22} \begin{pmatrix} \mu_x^2 & -\mu_x \\ -\mu_x & 1 \end{pmatrix}$$

and q^{22} is the (2, 2) element of Q^{-1} .

(ii) *If Assumption 4(ii) holds with $L_n = L$ fixed, then $\sqrt{N}(\hat{\theta}_c - \theta) \rightarrow_d N(0, \Psi)$, where*

$$\Psi = B^{-1} \Upsilon B^{-1}, \quad \Upsilon = \sigma^2 \begin{pmatrix} L^{-1} & L^{-1} \\ L^{-1} & c' Q^{-1} c \end{pmatrix}, \quad B = \begin{pmatrix} 1 & \mu_x \\ \mu_x & \frac{L^2}{12} + E(u_{xi}^2) \end{pmatrix},$$

and $c = (\mu_x, c_x)'$ with $c_x = \frac{L^2}{12} + E(u_{xi}^2)$.

Remark 8. Despite its potential slower convergence than the OLS estimator $\hat{\theta}$, Theorem 3.2 indicates that the estimator $\hat{\theta}_c$ is $\sqrt{n/L_n}$ -consistent for θ . For inferential purposes we need to estimate the asymptotic covariance matrix Ψ . Since L_n is not observed in practice and the researcher may not know whether L_n is fixed or tends to ∞ as $n \rightarrow \infty$, we propose an estimator of Ψ that is consistent under either scenario. In the Appendix, we show that

$$\sqrt{N}(\hat{\theta}_c - \theta) = \sqrt{N} B_n^{-1} n^{-1} \sum_{i=1}^n \begin{pmatrix} u_i \\ (X_i, X_i^2) \end{pmatrix} (\tilde{\theta} - \theta) = B_n^{-1} R_n$$

where $B_n = n^{-1}\mathbf{X}'\mathbf{X}$,

$$R_n = \sqrt{N}n^{-1} \sum_{i=1}^n \begin{pmatrix} u_i \\ (X_i, X_i^2) (\tilde{\theta} - \theta) \end{pmatrix} = \begin{pmatrix} \sqrt{N}n^{-1} \sum_{i=1}^n u_i \\ \tilde{\mathcal{C}}' \sqrt{N}(\tilde{\theta} - \theta) \end{pmatrix},$$

and $\hat{c} = n^{-1} \sum_{i=1}^n (X_i, X_i^2)'$. Let $\hat{u}_{ic} = y_i - \hat{\beta}_{0c} - \hat{\beta}_{1c}X_i$. Theorems 3.1-3.2 suggest that we can consistently estimate the asymptotic variance of $\sqrt{N}n^{-1} \sum_{i=1}^n u_i$ and $\sqrt{N}(\tilde{\theta} - \theta)$ by $Nn^{-2} \sum_{i=1}^n \hat{u}_{ic}^2$ and $n^{-1} \sum_{i=1}^n \hat{u}_{ic}^2 \tilde{\mathcal{C}}' Q_n^{-1} \hat{c}$, respectively.² To estimate the asymptotic covariance between $\sqrt{N}n^{-1} \sum_{i=1}^n u_i$ and $\tilde{\mathcal{C}}' \sqrt{N}(\tilde{\theta} - \theta)$, we need to use the Bahadur representation of $\sqrt{N}(\tilde{\theta} - \theta)$ given in the Appendix:

$$\sqrt{N}(\tilde{\theta} - \theta) = \sum_{i=1}^n \begin{pmatrix} \xi_{i1} \\ \xi_{i2} \end{pmatrix} + o_P(1),$$

where ξ_{i1} and ξ_{i2} are defined in (6.11)-(6.12). Due to the automatic correction of endogeneity bias in nonparametric estimation and the use of undersmoothing, centering in the definition of ξ_{i1} and ξ_{i2} is not necessary. This representation motivates us to estimate the asymptotic covariance between $\sqrt{N}n^{-1} \sum_{i=1}^n u_i$ and $\tilde{\mathcal{C}}' \sqrt{N}(\tilde{\theta} - \theta)$ by

$$\hat{\Upsilon}_{n,12} = n^{-1} \sum_{i=1}^n \hat{u}_{ic} \tilde{\mathcal{C}}' (\hat{\xi}_{i1}, \hat{\xi}_{i2})',$$

where $\hat{\xi}_{i1} = q^{11} \int_a^b K_{ix} \hat{u}_{ic} dx + q^{12} \int_a^b x K_{ix} \hat{u}_{ic} dx$, $\hat{\xi}_{i2} = q^{21} \int_a^b K_{ix} \hat{u}_{ic} dx + q^{22} \int_a^b x K_{ix} \hat{u}_{ic} dx$, $K_{ix} = K_h(X_i - x)$, and q^{st} is the (s, t) element of Q^{-1} . Plugging these expressions for $\hat{\xi}_{i1}$ and $\hat{\xi}_{i2}$ directly into $\hat{\Upsilon}_{n,12}$ yields the simplification

$$\hat{\Upsilon}_{n,12} = n^{-1} \sum_{i=1}^n \hat{u}_{ic}^2 \tilde{\mathcal{C}}' Q^{-1} \left(\int_a^b K_{ix} dx, \int_a^b x K_{ix} dx \right)'$$

Again, for the same reason as that used to obtain the asymptotic variance of $\sqrt{N}(\tilde{\theta} - \theta)$ in Remark 6, we need to replace Q by Q_n in practice. It follows that we can estimate the asymptotic variance-covariance matrix Ψ by

$$\hat{\Psi}_n = B_n^{-1} \hat{\Upsilon}_{nc} B_n^{-1}$$

where

$$\hat{\Upsilon}_{nc} = \begin{pmatrix} Nn^{-2} \sum_{i=1}^n \hat{u}_{ic}^2 & \hat{\Upsilon}_{nc,12} \\ \hat{\Upsilon}_{nc,12} & n^{-1} \sum_{i=1}^n \hat{u}_{ic}^2 \tilde{\mathcal{C}}' Q_n^{-1} \hat{c} \end{pmatrix}.$$

and $\hat{\Upsilon}_{nc,12} = n^{-1} \sum_{i=1}^n \hat{u}_{ic}^2 \tilde{\mathcal{C}}' Q_n^{-1} \left(\int_a^b K_{ix} dx, \int_a^b x K_{ix} dx \right)'$. It is straightforward to show that $\hat{\Psi}_n \rightarrow_p \Psi$, and statistical inference can be conducted as usual without observing N . Thus we have the following corollary.

²Here we use Q_n in place of Q for the same reason stated in Remark 6.

Corollary 3.3 Under the conditions of Theorem 3.2, $\widehat{\Psi}_n \rightarrow_p \Psi$.

Remark 9. Even though we only focus on the linear structural equation model as specified in (1.1), it is straightforward to extend our theory to the general nonlinear structural equation model. For simplicity and in order to apply the result of PS directly, we focus on the case of one endogenous regressor. Suppose $\{y_i\}$ is generated according to

$$y_i = g(X_i, \theta) + u_i, \quad E(u_i|X_i) \neq 0, \quad (3.3)$$

where $g(\cdot, \theta)$ is known up to the finite dimensional parameter θ , and the endogenous regressor X_i satisfies (1.2) and (1.3). It is straightforward to show that the nonlinear least squares (NLS) estimator $\widehat{\theta}$ of θ is inconsistent. As before, the nonparametric local level estimate $\widehat{g}(x)$ of $g(x) = g(x, \theta)$ is still consistent for a large portion of the domain of the regressor. Then, the unknown parameter θ can be estimated by minimizing the following (spatial) \mathcal{L}_2 criterion

$$S_n(\theta) = \int_a^b (\widehat{g}(x) - g(x, \theta))^2 \widehat{f}(x) dx, \quad (3.4)$$

just as before. Let $\widetilde{\theta}$ denote the solution to the above minimization problem. Following the proof of Theorem 3.1, we can show that $\widetilde{\theta}$ is $\sqrt{n/L_n}$ -consistent for θ under some regularity conditions. In addition, it can be established in an analogous way to that of Theorem 3.2 that the NLS estimator, after bias correction, is also $\sqrt{n/L_n}$ -consistent for θ . The details are straightforward and are thus omitted.

4 Simulations

This section reports a small set Monte Carlo experiment to evaluate the finite sample performance of the \mathcal{L}_2 and bias-corrected OLS estimators. Data is generated according to the following data generating process (DGP):

$$\begin{aligned} y_i &= \beta_0 + \beta_1 X_i + u_i, \quad \beta_0 = 10, \quad \beta_1 = -1, \\ X_i &= \mu_\alpha \mathbf{1}\{i \in A_\alpha\} + u_{xi}, \quad \mu_\alpha = \frac{\alpha L_n}{2m}, \quad \alpha = -m, -m+1, \dots, m, \\ u_i &= \sigma(\epsilon_{yi} + \gamma(u_{xi} - 10)) / (1 + \gamma^2)^{1/2}, \quad \sigma = S_X, \end{aligned}$$

where ϵ_{yi} are *iid* $N(0, 1)$, u_{xi} are *iid* $N(10, 1)$ and independent of ϵ_{yi} , and S_X is the sample standard deviation of X_i . By construction, the signal to noise ratio is maintained at unity throughout the simulations in order to enhance comparability across experiments. Simulations are performed for $\gamma = 0.32$ (weak endogeneity, $\text{corr}(u_i, u_{xi}) = 0.3$) and $\gamma = 2.07$ (strong endogeneity, $\text{corr}(u_i, u_{xi}) = 0.9$), and for the sample sizes $n = 500, 1000, \text{ and } 2000$. We generate the location shift points μ_α as $2m + 1$ evenly spaced points between $[-\log n, \log n]$, where $m = \lceil n^{1/3} \rceil$ and $\lceil \cdot \rceil$ denotes the integer part of the argument.

We consider the three estimators discussed in previous sections, namely, the OLS estimator $(\widehat{\beta}_0, \widehat{\beta}_1)$, the \mathcal{L}_2 estimator $(\widetilde{\beta}_0, \widetilde{\beta}_1)$, and the bias-corrected OLS estimator $(\widehat{\beta}_{0c}, \widehat{\beta}_{1c})$. For the latter two estimators, we need to choose both a kernel and bandwidth. We use the normalized Epanechnikov kernel (with variance 1),

$$K(u) = \frac{3}{4} \left(1 - \frac{1}{5}u^2\right) \mathbf{1}(|u| \leq \sqrt{5}). \quad (4.1)$$

Two methods of bandwidth selection were considered: (i) rule of thumb (ROT) setting $h = S_X n^{-1/3}$ where S_X is the sample standard deviation of $\{X_i\}$; and (ii) least squares cross-validation (LSCV) to find a preliminary bandwidth h_0 (which converges to 0 at $n^{-1/5}$ in the stationary regressor case) for calculating $\widehat{g}(x)$ and then re-normalize this bandwidth as $h = h_0 n^{1/5} n^{-1/3}$ to achieve the undersmoothing required for the \mathcal{L}_2 and bias-corrected OLS estimators. Simulations show that results based on the LSCV are similar to those based on ROT. The LSCV is much more costly in terms of computation time, so only the ROT results are reported in what follows.

Figs 2 and 3 report bias (Bias), standard deviation (Std dev) and root mean squared error (Rmse) for the three estimates of the intercept and slope parameters, respectively. On the horizontal axis is the truncation parameter (λ) that indicates the integration lower and upper limits of the \mathcal{L}_2 estimators given by $F_X^{-1}(\lambda)$ and $F_X^{-1}(1 - \lambda)$, respectively. The number of replications is 10,000. The top panel of Figs 2 and 3 reports the bias for each estimator. Under both weak and strong endogeneity, the OLS estimator has non-negligible bias, and the \mathcal{L}_2 and bias-corrected estimators have much smaller bias than the OLS estimators for all values of λ . As expected, when λ is small, the latter two estimators are still subject to the distortion of the endogeneity effect of X_i in the tails. But as λ increases, endogeneity bias dies off quickly. The middle panel of Figs 2 and 3 reports the standard deviation of each estimator. Unsurprisingly, the OLS estimator has the least Std dev and the bias-corrected OLS estimator has the largest Std dev. Also as expected, the larger is λ , the less the number of effective observations used in obtaining the \mathcal{L}_2 and bias-corrected estimators, and the larger the variance of the \mathcal{L}_2 and bias-corrected estimators in consequence. The bottom panel of Figs 2 and 3 reports the Rmse of each estimator. Interestingly, in the case of weak endogeneity, the \mathcal{L}_2 and bias-corrected estimators outperform OLS in terms of Rmse only for a small degree of truncation. As λ increases, the increase of the Std dev of the \mathcal{L}_2 and bias-corrected estimators can dominate the decrease of bias. In sharp contrast, in the case of strong endogeneity, the \mathcal{L}_2 and bias-corrected estimators dominate OLS in terms of Rmse for all values of λ under investigation.

To see the effect of sample size on the various estimators, Table 1 reports the Bias, Std dev, and Rmse for different numbers of observations ($n = 500, 1000, \text{ and } 2000$), where the truncation parameter λ takes the value 0.15. The number of replications is 100,000. From Table 1, we see that the endogeneity bias of OLS remains largely unchanged when the sample size is doubled or quadrupled. This is true despite the fact that the support of the location

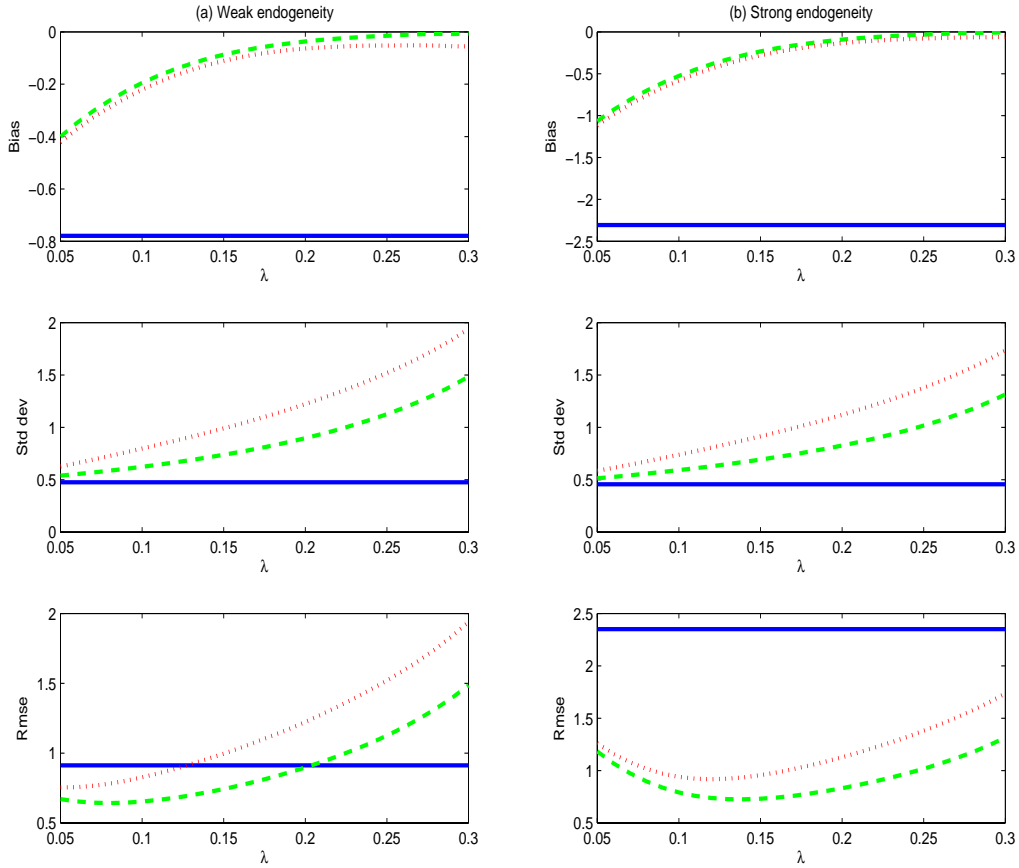


Figure 2: Bias, Std dev, and Rmse of various intercept estimators with $n = 500$ observations. OLS estimator (solid line), \mathcal{L}_2 estimator (dashed line), bias-corrected OLS estimator (dotted line). $\text{Corr}(u_i, u_{xi}) = 0.3$ and 0.9 for weak and strong endogeneity cases, respectively.

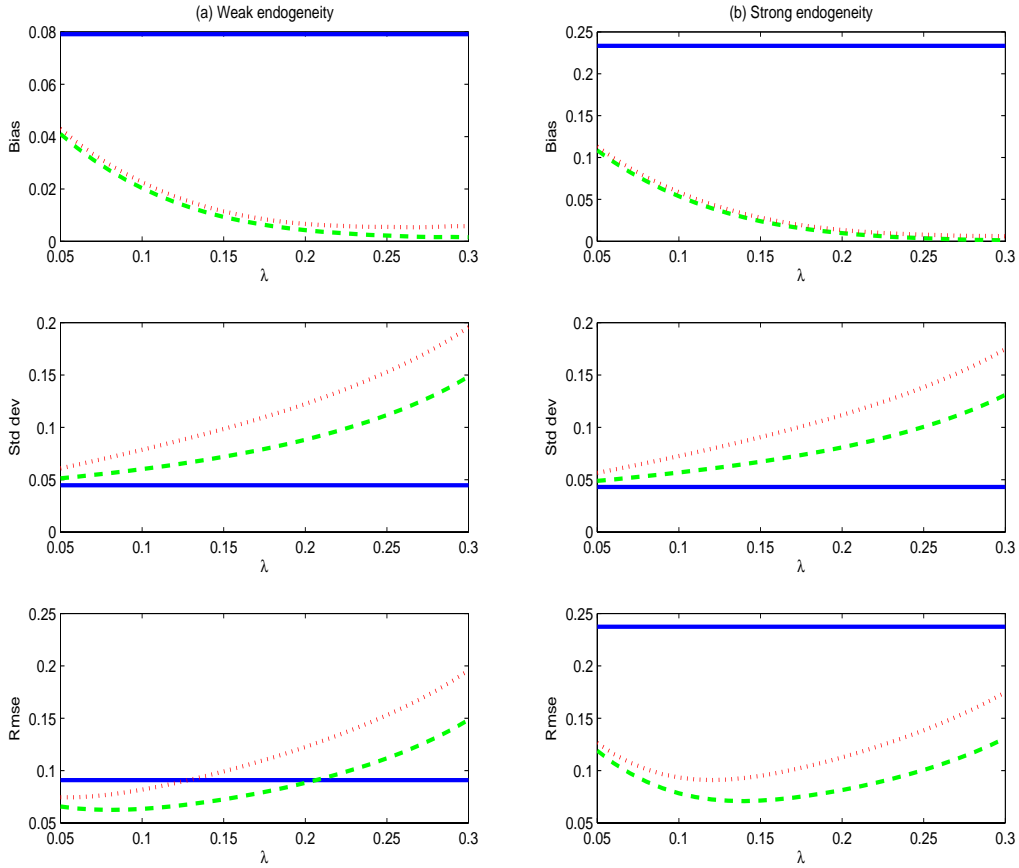


Figure 3: Bias, Std dev, and Rmse of various slope estimators with $n = 500$ observations. OLS estimator (solid line), \mathcal{L}_2 estimator (dashed line), bias-corrected OLS estimator (dotted line). $\text{Corr}(u_i, u_{xi}) = 0.3$ and 0.9 for weak and strong endogeneity cases, respectively.

Table 1: Comparison of finite sample performance of various estimates

Estimators	$n = 500$			$n = 1000$			$n = 2000$		
	Bias	Std dev	Rmse	Bias	Std dev	Rmse	Bias	Std dev	Rmse
Weak endogeneity									
$\widehat{\beta}_0$	-0.774	0.473	0.908	-0.709	0.341	0.787	-0.654	0.244	0.698
$\widetilde{\beta}_0$	-0.076	0.731	0.735	-0.037	0.532	0.533	-0.020	0.382	0.383
$\widehat{\beta}_{0c}$	-0.086	0.975	0.979	-0.041	0.683	0.684	-0.023	0.473	0.474
$\widetilde{\beta}_1$	0.078	0.045	0.090	0.071	0.032	0.078	0.066	0.022	0.070
$\widehat{\beta}_1$	0.008	0.071	0.072	0.004	0.051	0.051	0.002	0.037	0.037
$\widehat{\beta}_{1c}$	0.009	0.097	0.098	0.004	0.067	0.067	0.002	0.047	0.047
Strong endogeneity									
$\widehat{\beta}_0$	-2.290	0.461	2.336	-2.093	0.334	2.120	-1.932	0.239	1.947
$\widetilde{\beta}_0$	-0.210	0.681	0.712	-0.106	0.501	0.512	-0.058	0.361	0.366
$\widehat{\beta}_{0c}$	-0.244	0.902	0.934	-0.117	0.638	0.648	-0.066	0.446	0.451
$\widetilde{\beta}_1$	0.232	0.044	0.236	0.211	0.031	0.213	0.195	0.022	0.196
$\widehat{\beta}_1$	0.022	0.066	0.070	0.011	0.048	0.049	0.006	0.034	0.035
$\widehat{\beta}_{1c}$	0.025	0.090	0.093	0.012	0.063	0.064	0.007	0.044	0.044

Note. $\text{Corr}(u_i, u_{xi}) = 0.3$ and 0.9 for the case of weak and strong endogeneity, respectively.

shifts is expanding. But since the support expands slowly as n increases, so too is the reduction of the bias of the OLS estimator. In contrast, both the \mathcal{L}_2 and bias-corrected OLS estimators have substantially smaller bias than OLS. As the sample size increases, the bias of the latter two estimators continues to decrease. As expected, the variance of the \mathcal{L}_2 and bias-corrected OLS estimators are larger than that of OLS. In terms of Rmse, the \mathcal{L}_2 estimator generally dominates the bias-corrected OLS estimator which in turn outperforms OLS.

5 Concluding remarks

The present paper explores a paradox where the greater use of correct prior information on a model can be detrimental in regression. The explanation for this paradox is that even when we use additional correct information about the specification of a model, that information may still not be complete and, in consequence, may distort regression results. In the example studied here, the correct additional information used is considerable and is the full functional form specification of the model. Nevertheless, the omitted information (endogeneity) that makes the model specification incomplete is very important and leads to inconsistency in parametric regression.

In such situations, it is very interesting that partial information can be successful where more complete information fails. In applied statistics, it has long been known that controlling for outliers in regression can help to achieve robustness. What the results of the present pa-

per show, is that nonparametric kernel regression naturally utilizes this mechanism to great advantage in structural regression. More specifically, kernel regression has a considerable additional advantage beyond its usually touted advantage of robustness to (unknown) functional form. Local nonparametric regression also provides robustness to endogeneity in the regressor when there are systematic influences that assure identification, such as location shifts or nonstationarity in the data.

It is also possible to obtain consistent estimation by parametric methods in such cases. In particular, spatial \mathcal{L}_2 regression is shown to successfully remove endogeneity bias and inconsistency by bounding the domain of the regression. This approach is analogous to the treatment of outliers – it provides protection against possible effects of endogeneity in parametric regression by paying a premium through the loss of tail information in the data.

6 Appendix: Proofs and supplementary technical results

6.1 Proof of Theorem 2.2

Noting that $\sqrt{n}(\hat{\theta} - \theta) = (n^{-1}\mathbf{X}'\mathbf{X})^{-1}n^{-1/2}\mathbf{X}'\mathbf{u}$, we have

$$\begin{aligned} & \sqrt{n}D_n \left(\hat{\theta} - \theta - (\mathbf{X}'\mathbf{X})^{-1} E(\mathbf{X}'\mathbf{u}) \right) \\ &= (D_n^{-1}n^{-1}\mathbf{X}'\mathbf{X}D_n^{-1})^{-1} D_n^{-1}n^{-1/2} (\mathbf{X}'\mathbf{u} - E(\mathbf{X}'\mathbf{u})). \end{aligned}$$

We prove the theorem by showing that

$$\Gamma_n \equiv D_n^{-1}n^{-1}\mathbf{X}'\mathbf{X}D_n^{-1} \rightarrow_p \lim_{n \rightarrow \infty} \begin{pmatrix} 1 & L_n^{-1}\mu_x \\ L_n^{-1}\mu_x & \frac{1}{12} + \frac{E(u_{x1}^2)}{L_n^2} \end{pmatrix} \equiv \Gamma, \quad (6.1)$$

and

$$A_n \equiv D_n^{-1}n^{-1/2}(\mathbf{X}\mathbf{u} - E(\mathbf{X}\mathbf{u})) \rightarrow_d N(0, \Omega), \quad (6.2)$$

where

$$\Omega = \lim_{n \rightarrow \infty} \begin{pmatrix} \sigma^2 & \frac{E(u_{x1}u_1^2)}{L_n} \\ \frac{E(u_{x1}u_1^2)}{L_n} & \frac{\sigma^2}{12} + \frac{\text{Var}(u_{x1}u_1)}{L_n^2} \end{pmatrix}.$$

First, by the fact that $\bar{X} = \bar{u}_x \rightarrow_p \mu_x$, and (2.2), we have

$$\begin{aligned} L_n^{-2}n^{-1} \sum_{i=1}^n X_i^2 &= L_n^{-2}n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2 + L_n^{-2}\bar{X}^2 \\ &= \frac{1}{12} + \frac{\sigma_x^2}{L_n^2} + \frac{\mu_x^2}{L_n^2} + o_P(1) = \frac{1}{12} + \frac{E(u_{x1}^2)}{L_n^2} + o_P(1). \end{aligned}$$

Thus (6.1) follows. To show (6.2), by the Cramér-Wold device it suffices to show that for any $\omega = (\omega_1, \omega_2)'$ with $\|\omega\| = 1$, we have

$$\omega' A_n = n^{-1/2} \sum_{i=1}^n \{ \omega_1 u_i + \omega_2 L_n^{-1} [X_i u_i - E(X_i u_i)] \} \rightarrow_d N(0, \omega' \Omega \omega). \quad (6.3)$$

By construction, $E(\omega' A_n) = 0$. We now calculate the asymptotic variance of $\omega' A_n$:

$$\begin{aligned} \text{Var}(\omega' A_n) &= n^{-1} \sum_{i=1}^n \text{Var}(\omega_1 u_i + \omega_2 L_n^{-1} [X_i u_i - E(X_i u_i)]) \\ &= \omega_1^2 \sigma^2 + 2\omega_1 \omega_2 L_n^{-1} n^{-1} \sum_{i=1}^n E(X_i u_i^2) + \omega_2^2 L_n^{-2} n^{-1} \sum_{i=1}^n \text{Var}(X_i u_i) \\ &\rightarrow \omega' \Omega \omega, \end{aligned}$$

because $L_n^{-1} n^{-1} \sum_{i=1}^n E(X_i u_i^2) = L_n^{-1} E(u_{xi} u_i^2)$, and

$$\begin{aligned} &\frac{1}{n L_n^2} \sum_{i=1}^n \text{Var}(X_i u_i) \\ &= \frac{1}{(2m+1) L_n^2} \sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \text{Var}\left(\left(\frac{L_n \alpha}{2m} + u_{xi}\right) u_i\right) \\ &= \frac{1}{(2m+1) L_n^2} \sum_{\alpha=-m}^m \frac{1}{M} \sum_{i \in A_\alpha} \left\{ \frac{L_n^2 \alpha^2}{4m^2} \sigma^2 + \text{Var}(u_{xi} u_i) + \frac{L_n \alpha}{m} E[u_{xi} u_i^2] \right\} \\ &= \frac{\sigma^2}{(2m+1) 2m^2} \frac{m(m+1)(2m+1)}{6} + \frac{\text{Var}(u_{xi} u_i)}{L_n^2} \\ &\rightarrow \frac{\sigma^2}{12} + \lim_{n \rightarrow \infty} \frac{\text{Var}(u_{xi} u_i)}{L_n^2}. \end{aligned}$$

Let $\xi_{in} = n^{-1/2} \{\omega_1 u_i + \omega_2 L_n^{-1} [X_i u_i - E(X_i u_i)]\}$. By the C_r inequality,

$$\begin{aligned} &\sum_{i=1}^n E |\xi_{in}|^{2+\delta} \\ &= \frac{1}{n^{1+\delta/2}} \sum_{i=1}^n E |\omega_1 u_i + \omega_2 L_n^{-1} [X_i u_i - E(X_i u_i)]|^{2+\delta} \\ &\leq \frac{2^{1+\delta}}{n^{1+\delta/2}} \sum_{i=1}^n \left\{ E |\omega_1 u_i|^{2+\delta} + |\omega_2 L_n^{-1}|^{2+\delta} E \|X_i u_i - E(X_i u_i)\|^{2+\delta} \right\} \\ &\leq \frac{2^{1+\delta} |\omega_1|^{2+\delta}}{n^{1+\delta/2}} \sum_{i=1}^n E |u_i|^{2+\delta} + |\omega_2 L_n^{-1}|^{2+\delta} \frac{2^{3+2\delta}}{n^{1+\delta/2}} \sum_{i=1}^n E \|X_i u_i\|^{2+\delta} \\ &= O(n^{-\delta/2}) + O((mM)^{-\delta/2}) = o(1). \end{aligned}$$

Then (6.3) follows by the Liapounov CLT.

6.2 Proof of Theorem 3.1

For notational simplicity, let $K_{ix} = K_h(X_i - x)$. Define

$$Q_n = \begin{pmatrix} \int_a^b \hat{f}(x) dx & \int_a^b x \hat{f}(x) dx \\ \int_a^b x \hat{f}(x) dx & \int_a^b x^2 \hat{f}(x) dx \end{pmatrix}.$$

We first prove some lemmas that are used in the proof of Theorem 3.1.

Lemma 6.1 Let $\Theta_{jn} = \int_a^b x^j \widehat{f}(x) dx$ for $j = 0, 1, 2$. Then

$$(i) E\Theta_{jn} = \frac{b^{j+1} - a^{j+1}}{j+1} + o(1),$$

$$(ii) \text{Var}(\Theta_{jn}) = O\left(\frac{L_n}{mM} + \frac{hL_n^2}{mM}\right) = o(1) \text{ for } j = 0, 1, 2.$$

Proof. Recall $N = 2mM/L_n$ and $\widehat{f}(x) = N^{-1} \sum_{i=1}^n K_h(X_i - x)$. By the Fubini theorem and Taylor expansion,

$$\begin{aligned} & E\Theta_{jn} \\ &= \frac{L_n}{2m} \sum_{a=-m}^m \int_a^b \int x^j K(z) f_{u_x}(x - \mu_\alpha + hz) dz dx \\ &= \frac{L_n}{2m} \sum_{a=-m}^m \int_a^b x^j f_{u_x}(x - \mu_\alpha) dx + \frac{h^2 \mu_2(K) L_n}{2m} \sum_{a=-m}^m \int_a^b x^j f''_{u_x}(x - \mu_\alpha) dx + R_{jn} \end{aligned} \quad (6.4)$$

where $R_{jn} = \frac{h^2 L_n}{2m} \sum_{a=-m}^m \int_a^b \int \int x^j z^2 K(z) [f''_{u_x}(x - \mu_\alpha + whz) - f''_{u_x}(x - \mu_\alpha)] (1-w) dw dz dx$ is the remainder term. Noting

$$\begin{aligned} \frac{L_n}{2m} \sum_{a=-m}^m f_{u_x}(x - \mu_\alpha) &\approx \int_{L_n/2}^{L_n/2} f_{u_x}(x - p) dp \rightarrow \int_{-\infty}^{\infty} f_{u_x}(x - p) dp = 1, \\ \frac{L_n}{2m} \sum_{a=-m+1}^m f''_{u_x}(x - \mu_\alpha) &\approx \int_{L_n/2}^{L_n/2} f''_{u_x}(x - p) dp \rightarrow \int_{-\infty}^{\infty} f''_{u_x}(x - p) dp < \infty, \end{aligned}$$

it follows that the first term in (6.4) tends to

$$\int_a^b x^j dx = \frac{b^{j+1} - a^{j+1}}{j+1}, \quad (6.5)$$

the second term in (6.4) is approximately

$$h^2 \mu_2(K) \int_{-\infty}^{\infty} f''_{u_x}(p) dp \int_a^b x^j dx = O(h^2),$$

and $R_{jn} = o(h^2)$ by dominated convergence.

Now, by Jensen's inequality, the Fubini theorem, a change of variables, and since (u_i, u_{xi}) is *iid*, we have

$$\begin{aligned}
& \text{Var}(\Theta_{jn}) \\
&= \frac{L_n^2}{4m^2M^2} \sum_{i=1}^n \text{Var} \left(\int_a^b x^j K_h(X_i - x) dx \right) \\
&= \frac{L_n^2}{4m^2M} \sum_{a=-m}^m \text{Var} \left(\int_a^b x^j K_h(\mu_\alpha + u_{x1} - x) dx \right) \\
&\leq \frac{L_n^2}{4m^2M} \sum_{a=-m}^m \int_a^b \int_a^b x^j \tilde{x}^j E[K_h(\mu_\alpha + u_{x1} - x) K_h(\mu_\alpha + u_{x1} - \tilde{x})] dx d\tilde{x} \\
&= \frac{L_n^2}{4m^2M} \sum_{a=-m}^m \int_a^b \int_a^b x^j \tilde{x}^j \int K_h(\mu_\alpha + u_{x1} - x) K_h(\mu_\alpha + u_{x1} - \tilde{x}) f_{u_x}(u_{x1}) du_{x1} dx d\tilde{x} \\
&= \frac{L_n^2}{4hm^2M} \sum_{a=-m}^m \int_a^b \int_a^b x^j \tilde{x}^j \int K(z) K\left(z + \frac{x - \tilde{x}}{h}\right) f_{u_x}(x - \mu_\alpha + hz) dz dx d\tilde{x} \\
&= \frac{L_n}{2hmM} \int_a^b \int_a^b x^j \tilde{x}^j \int K(z) K\left(z + \frac{x - \tilde{x}}{h}\right) \frac{L_n}{2m} \sum_{a=-m}^m f_{u_x}(x - \mu_\alpha) dz dx d\tilde{x} + O\left(\frac{hL_n^2}{mM}\right) \\
&= \frac{L_n}{2hmM} \int_a^b \int_a^b x^j \tilde{x}^j \int K(z) K\left(z + \frac{x - \tilde{x}}{h}\right) dz dx d\tilde{x} \{1 + o(1)\} + O\left(\frac{hL_n^2}{mM}\right) \\
&= \frac{L_n}{2mM} \int_a^b \int_{(a-x)/h}^{(b-x)/h} x^j (x + hv)^j \int K(z) K(z - v) dz dv dx + O\left(\frac{hL_n^2}{mM}\right) \\
&= O\left(\frac{L_n}{mM} + \frac{hL_n^2}{mM}\right).
\end{aligned}$$

■

Lemma 6.2 *Let $\Xi_{jn} = \int_a^b N^{-1/2} \sum_{i=1}^n x^j (X_i - x) K_{ix} dx$ for $j = 0, 1$. Then $\Xi_{jn} = o_P(1)$.*

Proof. By the Fubini theorem and Taylor expansion,

$$\begin{aligned}
E(\Xi_{jn}) &= \sqrt{\frac{ML_n}{2m}} \sum_{a=-m}^m \int_a^b E[x^j (\mu_\alpha + u_{x1} - x) K_h(\mu_\alpha + u_{x1} - x)] dx \\
&= \sqrt{\frac{ML_n}{2m}} \sum_{a=-m}^m \int_a^b \int x^j z K(z) f_{u_x}(x - \mu_\alpha + hz) dz dx \\
&= h^2 \sqrt{\frac{2mM}{L_n}} \mu_2(K) \left\{ \int_a^b x^j \frac{L_n}{2m} \sum_{a=-m}^m f_{u_x}''(x - \mu_\alpha) dx \{1 + o(1)\} \right\} \\
&= h^2 \sqrt{N} \int_a^b x^j dx \int f_{u_x}''(p) dp \{1 + o(1)\} \\
&= o(1).
\end{aligned}$$

Analogous to the proof of Lemma 6.1(ii), we can show that $\text{Var}(\Xi_{jn}) = O(h^2) = o(1)$. The result follows from the Chebyshev inequality. ■

Lemma 6.3 Let $f(u)$ denote the p.d.f. of u_i . Suppose Assumptions 3 and 4 hold. Then

$$\int u \frac{L_n}{2m} \sum_{a=-m}^m [f(u, x - \mu_\alpha) - f(u)] du = O((m/L_n)^{-2} + h^2).$$

Proof. The proof follows from that of Lemma 6.1 in Phillips and Su (2009). The main difference is that $2m/L_n$ here plays the role of m^λ in Phillips and Su (2009). ■

To prove Theorem 3.1, noticing that if $g(x) = \beta_0 + \beta_1 x$, then

$$\begin{aligned} & \int_a^b \widehat{g}(x) \widehat{f}(x) dx \\ = & \int_a^b N^{-1} \sum_{i=1}^n K_h(X_i - x) y_i dx \\ = & \int_a^b N^{-1} \sum_{i=1}^n K_{ix} [\beta_0 + \beta_1 x + \beta_1 (X_i - x) + u_i] dx \\ = & \beta_0 \int_a^b \widehat{f}(x) dx + \beta_1 \int_a^b x \widehat{f}(x) dx + \beta_1 \int_a^b N^{-1} \sum_{i=1}^n (X_i - x) K_{ix} dx + \int_a^b N^{-1} \sum_{i=1}^n K_{ix} u_i dx, \end{aligned}$$

and similarly

$$\begin{aligned} & \int_a^b x \widehat{g}(x) \widehat{f}(x) dx \\ = & \beta_0 \int_a^b x \widehat{f}(x) dx + \beta_1 \int_a^b x^2 \widehat{f}(x) dx + \beta_1 \int_a^b N^{-1} \sum_{i=1}^n x (X_i - x) K_{ix} dx + \int_a^b N^{-1} \sum_{i=1}^n x K_{ix} u_i dx. \end{aligned}$$

It follows that

$$\begin{aligned} & \sqrt{N} (\widetilde{\theta} - \theta) \\ = & \beta_1 Q_n^{-1} \left(\begin{array}{c} \int_a^b N^{-1/2} \sum_{i=1}^n (X_i - x) K_{ix} dx \\ \int_a^b N^{-1/2} \sum_{i=1}^n x (X_i - x) K_{ix} dx \end{array} \right) + Q_n^{-1} \left(\begin{array}{c} \int_a^b N^{-1/2} \sum_{i=1}^n K_{ix} u_i dx \\ \int_a^b N^{-1/2} \sum_{i=1}^n x K_{ix} u_i dx \end{array} \right) \end{aligned} \quad (6.6)$$

By Lemma 6.1,

$$Q_n \rightarrow_p \left(\begin{array}{cc} b-a & \frac{b^2-a^2}{2} \\ \frac{b^2-a^2}{2} & \frac{b^3-a^3}{3} \end{array} \right) = Q, \quad (6.7)$$

where $\det(Q) = (b-a)^4/12 > 0$ as $b \neq a$. This, together with Lemma 6.2, implies that the first term in (6.6) is $o_P(1)$. We are left to show that the second term in (6.6) is asymptotically $N(0, \sigma^2 Q^{-1})$.

Let $\omega = (\omega_1, \omega_2)'$ be such that $\|\omega\| = 1$, and define

$$\Theta_n = \omega' \left(\begin{array}{c} \int_a^b N^{-1/2} \sum_{i=1}^n K_{ix} u_i dx \\ \int_a^b N^{-1/2} \sum_{i=1}^n x K_{ix} u_i dx \end{array} \right) = N^{-1/2} \sum_{i=1}^n \int_a^b (\omega_1 + \omega_2 x) K_{ix} u_i dx.$$

We complete the proof by showing that $E(\Theta_n) = O(N^{1/2}h^2) = o(1)$ by Assumption 6, and

$$\Theta_n - E(\Theta_n) \rightarrow_d N(0, \sigma^2 \omega' Q \omega). \quad (6.8)$$

By a change of variables, the Fubini theorem, Lemma 6.3, and Assumptions 1(ii), 3 and 6, we obtain

$$\begin{aligned} E\Theta_n &= \sqrt{\frac{ML_n}{2m}} \sum_{a=-m}^m \int_a^b E[(\omega_1 + \omega_2 x) K_h(\mu_\alpha + u_{x1} - x) u_1] dx \\ &= \sqrt{\frac{ML_n}{2m}} \sum_{a=-m}^m \int_a^b \int \int (\omega_1 + \omega_2 x) u K(z) f(u, x - \mu_\alpha + hz) dz du dx \\ &= \sqrt{\frac{2mM}{L_n}} \int_a^b (\omega_1 + \omega_2 x) \int K(z) \left\{ \int u \frac{L_n}{2m} \sum_{a=-m}^m f(u, x - \mu_\alpha + hz) du \right\} dz dx \\ &= \sqrt{N} \int_a^b (\omega_1 + \omega_2 x) \int K(z) \left\{ O\left((m/L_n)^{-2} + h^2\right) + \int u f(u) du \right\} dz dx \\ &= \sqrt{N} O\left((m/L_n)^{-2} + h^2\right) = o(1). \end{aligned}$$

Next, letting $w(x) = \omega_1 + \omega_2 x$, we have

$$\begin{aligned} &\text{Var}(\Theta_n) \\ &= N^{-1} \sum_{i=1}^n \text{Var} \left(u_i \int_a^b (\omega_1 + \omega_2 x) K_{ix} dx \right) \\ &= \frac{L_n}{2m} \sum_{a=-m}^m \int_a^b \int_a^b E[w(x) w(\tilde{x}) u_1^2 K_h(\mu_\alpha + u_{x1} - x) K_h(\mu_\alpha + u_{x1} - \tilde{x})] dx d\tilde{x} + o(1) \\ &= \frac{L_n}{2m} \sum_{a=-m}^m \int_a^b \int_a^b w(x) w(\tilde{x}) \int \int u^2 K_h(\mu_\alpha + u_x - x) K_h(\mu_\alpha + u_x - \tilde{x}) f(u, u_x) du_x du dx d\tilde{x} + o(1) \\ &= \frac{L_n}{2m} \sum_{a=-m}^m \int_a^b \int_a^b w(x) w(\tilde{x}) \int \int u^2 K(z) K\left(z + \frac{x - \tilde{x}}{h}\right) f(u, x - \mu_\alpha + hz) dz du dx d\tilde{x} + o(1) \\ &= \int_a^b \int_a^b w(x) w(\tilde{x}) \int \int u^2 K(z) K\left(z + \frac{x - \tilde{x}}{h}\right) \frac{L_n}{2m} \sum_{a=-m}^m f(u, x - \mu_\alpha) dz du dx d\tilde{x} + o(1) \\ &= \sigma^2 \int_a^b \int_a^b w(x) w(\tilde{x}) \int K(z) K\left(z + \frac{x - \tilde{x}}{h}\right) dz dx d\tilde{x} + o(1) \\ &= \sigma^2 \int_a^b \int_{(a-x)/h}^{(b-x)/h} w(x) w(x + hv) \int K(z) K(z - v) dz dv dx + o(1) \\ &= \sigma^2 \int_a^b (\omega_1 + \omega_2 x)^2 dx + o(1) \rightarrow \sigma^2 \omega' Q \omega, \end{aligned}$$

where we have used the fact that $\frac{L_n}{2m} \sum_{a=-m}^m f(u, x - \mu_\alpha) \rightarrow \int_{-\infty}^{\infty} f(u, x - p) dp = f(u)$. To show the asymptotic normality of $\Theta_n - E(\Theta_n)$, by the above variance calculation and the

independence of (u_i, u_{xi}) across i , it suffices to check the Liapounov condition. Let $\bar{Z}_i = Z_i - E(Z_i)$, where $Z_i = N^{-1/2} \int_a^b (\omega_1 + \omega_2 x) K_{ix} u_i dx$. Then by the C_r inequality,

$$\begin{aligned} \sum_{i=1}^n E |\bar{Z}_i|^{2+\delta} &\leq 2^{1+\delta} N^{-(1+\delta/2)} \sum_{i=1}^n E \left| u_i \int_a^b (\omega_1 + \omega_2 x) K_{ix} dx \right|^{2+\delta} \\ &\quad + 2^{1+\delta} N^{-(1+\delta/2)} \sum_{i=1}^n \left| E \left[u_i \int_a^b (\omega_1 + \omega_2 x) K_{ix} dx \right] \right|^{2+\delta} \\ &\equiv L_{n1} + L_{n2}. \end{aligned}$$

First,

$$\begin{aligned} L_{n1} &= 2^{1+\delta} N^{-\delta/2} \frac{L_n}{2m} \sum_{a=-m}^m \int \int \left| u \int_a^b w(x) K_h(\mu_\alpha + u_x - x) dx \right|^{2+\delta} f(u, u_x) du_x du \\ &\leq c_\delta N^{-\delta/2} \frac{L_n}{2m} \sum_{a=-m}^m \int \int |u|^{2+\delta} \left| \int_a^b K_h(\mu_\alpha + u_x - x) dx \right|^{2+\delta} f(u, u_x) du_x du \\ &\leq c_\delta N^{-\delta/2} \frac{L_n}{2m} \sum_{a=-m}^m \int \int |u|^{2+\delta} f(u, u_x) du_x du \\ &= c_\delta N^{-\delta/2} E |u|^{2+\delta} \frac{L_n(2m+1)}{2m} \\ &= O(N^{-\delta/2} L_n) = o(1), \end{aligned}$$

where $c_\delta = 2^{1+\delta} \sup_{a \leq |x| \leq b} |w(x)|^{2+\delta} < \infty$ as $\|\omega\|$, a and b are finite. By the Jensen inequality, $L_{n2} \leq L_{n1} = o(1)$. Then by the Liapounov CLT, (6.8) follows, and the proof is complete.

6.3 Proof of Theorem 3.2

Recall $N = 2mM/L_n$, and $B_n = n^{-1} \mathbf{X}' \mathbf{X}$. Noting that

$$\frac{1}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2} \begin{pmatrix} \bar{X} \tilde{c} \\ -\tilde{c} \end{pmatrix} = B_n^{-1} \sum_{i=1}^n \begin{pmatrix} 0 \\ -X_i \tilde{u}_i \end{pmatrix},$$

we have

$$\begin{aligned} \sqrt{N} (\hat{\theta}_c - \theta) &= \sqrt{n} C_n (\hat{\theta} - \theta) + \frac{1}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{N} C_n \begin{pmatrix} \bar{X} \tilde{c} \\ -\tilde{c} \end{pmatrix} \\ &= \sqrt{N} B_n^{-1} n^{-1} \sum_{i=1}^n \begin{pmatrix} u_i \\ X_i u_i - X_i \tilde{u}_i \end{pmatrix} \\ &= \sqrt{N} B_n^{-1} n^{-1} \sum_{i=1}^n \begin{pmatrix} u_i \\ (X_i, X_i^2) (\tilde{\theta} - \theta) \end{pmatrix}. \end{aligned}$$

Let $\omega = (\omega_1, \omega_2)'$ with $\|\omega\| = 1$. Define

$$T_n = \sqrt{N}\omega' B_n^{-1} n^{-1} \sum_{i=1}^n \begin{pmatrix} u_i \\ (X_i, X_i^2)' (\tilde{\theta} - \theta) \end{pmatrix}.$$

By the Cramér-Wold device, it suffices to show that

$$T_n \rightarrow_d N(0, \omega' \Psi \omega). \quad (6.9)$$

We show (6.9) by distinguishing whether L_n is allowed to approach ∞ as $n \rightarrow \infty$.

Case 1. $L_n \rightarrow \infty$ as $n \rightarrow \infty$. Noting that $\sqrt{N}n^{-1} = o(n^{-1/2})$, $\bar{X} = n^{-1} \sum_{i=1}^n u_{xi} \rightarrow_p \mu_x$, $L_n^{-2} n^{-1} \sum_{i=1}^n X_i^2 = L_n^{-2} n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2 + L_n^{-2} \bar{X}^2 \rightarrow_p \frac{1}{12}$, we have $\sqrt{N}n^{-1} \sum_{i=1}^n u_i \rightarrow_p 0$, $S_{xnc}^2/S_x^2 \rightarrow_p 1$ and $\bar{X}/S_x^2 \rightarrow_p 0$, where $S_x^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ and $S_{xnc}^2 = n^{-1} \sum_{i=1}^n X_i^2$. Also note that

$$B_n^{-1} = \frac{1}{S_x^2} \begin{pmatrix} S_{xnc}^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}. \quad (6.10)$$

It follows that

$$\begin{aligned} T_n &= \sqrt{N}n^{-1}\omega' \sum_{i=1}^n \frac{1}{S_x^2} \begin{pmatrix} S_{xnc}^2 u_i - \bar{X} (X_i, X_i^2)' (\tilde{\theta} - \theta) \\ -\bar{X} u_i + (X_i, X_i^2)' (\tilde{\theta} - \theta) \end{pmatrix} \\ &= \sqrt{N}n^{-1} \sum_{i=1}^n \frac{1}{S_x^2} \left\{ (\omega_1 S_{xnc}^2 - \omega_2 \bar{X}) u_i + (\omega_2 - \omega_1 \bar{X}) (X_i, X_i^2)' (\tilde{\theta} - \theta) \right\} \\ &= \frac{\omega_2 - \omega_1 \bar{X}}{S_x^2} \sqrt{N} (\tilde{\theta} - \theta)' n^{-1} \sum_{i=1}^n \begin{pmatrix} X_i \\ X_i^2 \end{pmatrix} + o_P(1) \\ &= \frac{(\omega_2 - \omega_1 \bar{X}) \sqrt{N} (\tilde{\beta}_1 - \beta_1)}{S_x^2} S_{xnc}^2 + \frac{(\omega_2 - \omega_1 \bar{X}) \sqrt{N} (\tilde{\beta}_0 - \beta_0)}{S_x^2} \bar{X} + o_P(1) \\ &= (\omega_2 - \omega_1 \mu_x) \sqrt{N} (\tilde{\beta}_1 - \beta_1) + o_P(1) \\ &\rightarrow_d N\left(0, \sigma^2 (\omega_2 - \omega_1 \mu_x)^2 q^{22}\right) = N\left(0, \sigma^2 q^{22} \omega' \begin{pmatrix} \mu_x^2 & -\mu_x \\ -\mu_x & 1 \end{pmatrix} \omega\right), \end{aligned}$$

where q^{22} is the (2, 2) element of Q^{-1} :

$$Q^{-1} = \frac{12}{(b-a)^4} \begin{pmatrix} \frac{b^3-a^3}{3} & -\frac{b^2-a^2}{2} \\ -\frac{b^2-a^2}{2} & b-a \end{pmatrix} \equiv \begin{pmatrix} q^{11} & q^{12} \\ q^{21} & q^{22} \end{pmatrix}.$$

Case 2. $L_n = L$ is fixed as $n \rightarrow \infty$. By the proof of Theorem 3.1 (see (6.6) and arguments thereafter),

$$\begin{aligned} \sqrt{N} (\tilde{\theta} - \theta) &= Q^{-1} N^{-1/2} \sum_{i=1}^n \begin{pmatrix} \int_a^b [K_{ix} u_i - E(K_{ix} u_i)] dx \\ \int_a^b x [K_{ix} u_i - E(K_{ix} u_i)] dx \end{pmatrix} + o_P(1) \\ &\equiv \sum_{i=1}^n \begin{pmatrix} \xi_{i1} \\ \xi_{i2} \end{pmatrix} + o_P(1), \end{aligned}$$

where

$$\xi_{i1} = N^{-1/2} \left\{ q^{11} \int_a^b [K_{ix} u_i - E(K_{ix} u_i)] dx + q^{12} \int_a^b x [K_{ix} u_i - E(K_{ix} u_i)] dx \right\}, \quad (6.11)$$

and

$$\xi_{i2} = N^{-1/2} \left\{ q^{21} \int_a^b [K_{ix} u_i - E(K_{ix} u_i)] dx + q^{22} \int_a^b x [K_{ix} u_i - E(K_{ix} u_i)] dx \right\}. \quad (6.12)$$

Let

$$R_n = \sqrt{N} n^{-1} \sum_{i=1}^n \begin{pmatrix} u_i \\ (X_i, X_i^2) (\tilde{\theta} - \theta) \end{pmatrix}.$$

Then

$$\begin{aligned} \omega' R_n &= \omega_2 \sqrt{N} (\tilde{\beta}_1 - \beta_1) \left\{ n^{-1} \sum_{i=1}^n X_i^2 \right\} + \omega_2 \sqrt{N} (\tilde{\beta}_0 - \beta_0) \left\{ n^{-1} \sum_{i=1}^n X_i \right\} + \omega_1 \sqrt{N} n^{-1} \sum_{i=1}^n u_i \\ &= \omega_2 c_x \sqrt{N} (\tilde{\beta}_1 - \beta_1) + \omega_2 \mu_x \sqrt{N} (\tilde{\beta}_0 - \beta_0) + \omega_1 \sqrt{N} n^{-1} \sum_{i=1}^n u_i + o_P(1) \\ &= \sum_{i=1}^n \left(\omega_2 c_x \xi_{i2} + \omega_2 \mu_x \xi_{i1} + \omega_1 \sqrt{N} n^{-1} u_i \right) + o_P(1) \\ &\equiv \bar{R}_n + o_P(1), \end{aligned}$$

where recall $c_x = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n E(X_i^2) = \frac{L^2}{12} + E(u_{xi}^2)$. Let $c = (\mu_x, c_x)'$. Note that $E(\bar{R}_n) = 0$, and

$$\begin{aligned} &\text{Var}(\bar{R}_n) \\ &= \sum_{i=1}^n \text{Var} \left(\omega_2 c_x \xi_{i2} + \omega_2 \mu_x \xi_{i1} + \omega_1 \sqrt{N} n^{-1} u_i \right) \\ &= \omega_2^2 \sum_{i=1}^n \text{Var} (c_x \xi_{i2} + \mu_x \xi_{i1}) + 2\omega_2 \omega_1 \sum_{i=1}^n \text{Cov} (c_x \xi_{i2} + \mu_x \xi_{i1}, \sqrt{N} n^{-1} u_i) \\ &\quad + \omega_1^2 \sum_{i=1}^n \text{Var} (\sqrt{N} n^{-1} u_i) \\ &= \sigma^2 \omega_2^2 c' Q^{-1} c + 2\omega_2 \omega_1 \sigma^2 L^{-1} \left\{ q^{11} \int_a^b dx + q^{12} \int_a^b x dx + q^{21} \int_a^b dx + q^{22} \int_a^b x dx \right\} \\ &\quad + \omega_1^2 \sigma^2 L^{-1} + o(1) \\ &\rightarrow \sigma^2 \omega_2^2 c' Q^{-1} c + 2\omega_2 \omega_1 \sigma^2 L^{-1} + \omega_1^2 \sigma^2 L^{-1} = \sigma^2 \omega' \begin{pmatrix} L^{-1} & L^{-1} \\ L^{-1} & c' Q^{-1} c \end{pmatrix} \omega, \end{aligned}$$

where the third line follows from the definitions of ξ_{i1} and ξ_{i2} , Fubini, a change of variables,

and the following two limits:

$$\begin{aligned}
& \sum_{i=1}^n E \left(\xi_{i1} \sqrt{N} n^{-1} u_i \right) \\
&= n^{-1} \sum_{i=1}^n E \left\{ \left[q^{11} \int_a^b [K_{ix} u_i - E(K_{ix} u_i)] dx + q^{12} \int_a^b x [K_{ix} u_i - E(K_{ix} u_i)] dx \right] u_i \right\} \\
&= q^{11} n^{-1} \sum_{i=1}^n E \left\{ u_i^2 \int_a^b K_{ix} dx \right\} + q^{12} n^{-1} \sum_{i=1}^n E \left\{ u_i^2 \int_a^b x K_{ix} dx \right\} \\
&= q^{11} (2m+1)^{-1} \sum_{\alpha=-m}^m \int_a^b \int \int u^2 K(z) f(u, x - u_\alpha + hz) dz du dx \\
&\quad + q^{12} (2m+1)^{-1} \sum_{\alpha=-m}^m \int_a^b x \int \int u^2 K(z) f(u, x - u_\alpha + hz) dz du dx \\
&= q^{11} L^{-1} \int_a^b \int u^2 \left\{ \frac{L}{2m} \sum_{\alpha=-m}^m f(u, x - u_\alpha) \right\} du dx \{1 + o(1)\} \\
&\quad + q^{12} L^{-1} \int_a^b x \int u^2 \left\{ \frac{L}{2m} \sum_{\alpha=-m}^m f(u, x - u_\alpha) \right\} dx \{1 + o(1)\} \\
&= \sigma^2 q^{11} L^{-1} \int_a^b dx \{1 + o(1)\} + \sigma^2 q^{12} L^{-1} \int_a^b x dx \{1 + o(1)\} \\
&\rightarrow \sigma^2 L^{-1} \left\{ q^{11} \int_a^b dx + q^{12} \int_a^b x dx \right\};
\end{aligned}$$

and, in a similar way,

$$\begin{aligned}
\sum_{i=1}^n E \left(\xi_{i2} \sqrt{N} n^{-1} u_i \right) &= n^{-1} \sum_{i=1}^n E \left[q^{21} u_i^2 \int_a^b K_{ix} u_i dx + q^{22} u_i^2 \int_a^b x K_{ix} dx \right] \\
&\rightarrow \sigma^2 L^{-1} \left\{ q^{21} \int_a^b dx + q^{22} \int_a^b x dx \right\}.
\end{aligned}$$

The Liapounov condition follows from the verification in the proof of Theorem 3.1, the fact that $\sqrt{N} n^{-1} \sum_{i=1}^n u_i$ also satisfies the Liapounov condition, and the C_r inequality. It follows that

$$R_n \rightarrow_d N \left(0, \sigma^2 \begin{pmatrix} L^{-1} & L^{-1} \\ L^{-1} & c' Q^{-1} c \end{pmatrix} \right)$$

and

$$T_n = \omega' B_n^{-1} R_n \rightarrow_d N \left(0, \sigma^2 \omega' B^{-1} \begin{pmatrix} L^{-1} & L^{-1} \\ L^{-1} & c' Q^{-1} c \end{pmatrix} B^{-1} \omega \right).$$

References

Anscombe, F. J. (1960) Rejection of outliers. *Technometrics*, 2, 123-147.

- Chen, L-A., A. H. Welsh, and W. Chan (2001) Estimators for the linear regression model based on winsorized observations. *Statistica Sinica* 11, 147-172.
- Hall, P. and J. L., Horowitz (2005) Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics* 33, 2904-2929.
- Phillips, P. C. B. and L. Su (2009) Nonparametric structural estimation via continuous location shifts in an endogenous regressor. *Mimeo*, Dept. of Economics, Yale University.
- Wang, Q. and P. C. B., Phillips (2009) Structural nonparametric cointegrating regression. Forthcoming in *Econometrica*.
- Welsh, A. H. (1987) The trimmed mean in the linear model. *Annals of Statistics* 15, 20-36.