ISSN 1755-5361

University of Essex

Department of Economics

Discussion Paper Series

No. 686 February 2010

Regression towards the mode

Gordon C.R. Kemp
J.M.C. Santos Silva

Note : The Discussion Papers in this series are prepared by members of the Department of Economics, University of Essex, for private circulation to interested readers. They often represent preliminary reports on work in progress and should therefore be neither quoted nor referred to in published work without the written consent of the author.

Regression towards the mode[*]

Gordon C.R. Kemp
Department of Economics, University of Essex

J.M.C. Santos Silva
Department of Economics, University of Essex and CEMAPRE

February 9, 2010

Abstract

We propose a semi-parametric mode regression estimator for the case in which the variate of interest is continuous and observable over its entire unbounded support. The estimator is semi-parametric in that the conditional mode is specified as a parametric function, but only mild assumptions are made about the nature of the conditional density of interest. We show that the proposed estimator is consistent and has a tractable asymptotic distribution. Simulation results and an empirical illustration are provided to highlight the practicality and usefulness of the estimator.

JEL classification code: C13, C14.

Key words: Bandwidth choice, Body mass index, Density estimation, Normal kernel.

[*] We are grateful to Marcus Chambers, Emma Iglesias, Myoung-jae Lee, Oliver Linton, Myung Hwan Seo, Paulo Parente, and to participants in seminars at Birmingham, Cambridge, GREQAM, ISER, LSE, Manchester, Reading, and University College Dublin, for their helpful comments and suggestions. The usual disclaimer applies. Santos Silva gratefully acknowledges the partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER.

Address for correspondence: Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. E-mail: kempgcr@essex.ac.uk (Kemp), jmcss@essex.ac.uk (Santos Silva).

1. INTRODUCTION

The mode is a characterizing feature of any statistical distribution or data set. Consequently, it is not surprising to find that the estimation of the mode has received considerable attention in the statistics literature (early references include Parzen, 1962, Chernoff, 1964, and Dalenius, 1965). Likewise, non-parametric estimation of the conditional mode is the subject of a large number of papers in statistical journals (see, among many others, Collomb, Härdle and Hassani, 1987, Samanta, and Thavaneswarn, 1990, Quintela-Del-Rio and Vieu, 1997, and Ziegler, 2003). However, very little attention has been devoted to the case that is most likely to be useful in econometric applications, that is, the semi-parametric case in which the conditional mode is specified as a parametric function, but only mild assumptions are made about the conditional distribution of interest.

In a pair of pathbreaking papers, Lee (1989, 1993) introduced semi-parametric mode regression estimators motivating them by noting that, under certain conditions, the conditional mode from the truncated data provides consistent estimates of the conditional mean for the original non-truncated data. To be able to handle truncation, the maximands considered by Lee (1989 and 1993) are based on kernels with bounded support. As a consequence, these estimators are difficult to implement and unattractive to practitioners, having seen little, if any, use in practice. However, applications with truncated continuous dependent variables are relatively rare and, therefore, tailoring mode regression to this kind of data unduly restricts its usefulness.

Although mode regression is appealing in the case of truncated data, its interest is much broader. Indeed, for the positively skewed data found in many applications (e.g., wages, prices, energy intake, expenditures on certain types of goods and services), the mode is generally located below the median and the mean. That is, the routinely used measures of central tendency convey little or no information on the location of the mode and on how it is affected by the regressors. Moreover, although in principle quantile regression (Koenker and Bassett, 1978) can completely characterize the shape of the conditional distribution, the way it is used in practice generally fails to reveal any information about the conditional mode. For example, it is easy to find examples where the mean and all quantiles are increasing functions of a regressor, while the mode decreases with the same regressor. Therefore, mode regression is a potentially very useful but much neglected tool that can be used to complement the standard mean and quantile regressions in the study of the features of conditional distributions.

In this paper we study the semi-parametric estimation of the conditional mode (mode regression) for the case in which the variate of interest is unbounded, continuous, and observable over its entire support.[1] In doing this, we depart from Lee (1989, 1993) by using smooth unbounded kernels and by letting the smoothing parameter δ pass to zero as the sample size increases. We show that in this case it is possible to obtain a consistent mode regression estimator that does not depend on the restrictive symmetry or independence assumptions required by Lee (1989, 1993). In addition, the estimator has a tractable asymptotic distribution and it is simple to implement using standard software. Furthermore, by using a Gaussian kernel with unbounded support, we obtain a family of estimators which includes both the conditional mode (when $\delta \to 0^+$) and the conditional mean (when $\delta \to \infty$) as limiting cases.

The reminder of the paper is organized as follows. Section 2 briefly reviews the rectangular and quadratic mode regression estimators proposed by Lee (1989, 1993). Section 3 details our approach to mode regression and presents the main asymptotic results. Section 4 provides simulation evidence on the finite sample performance of the proposed estimator, and Section 5 illustrates its application with a study of the recent evolution of the body mass index in England. Section 6 contains concluding remarks and discusses directions for further research. Finally, the proofs of the main results are collected in a technical appendix.

[1] For the polar opposite case, the function npconmode() in the np package (Hayfield and Racine, 2008) implements nonparametric mode regression for a categorical dependent variable based on the results of Hall, Racine and Li (2004).

2. RECTANGULAR AND QUADRATIC MODE REGRESSION

Let $\text{Mode}(y|x)$ denote the mode of the conditional distribution of y given x and assume that $\text{Mode}(y|x) = x'\beta_0$. Lee (1989) introduced the (rectangular) mode regression estimator based on a well-known loss function that can be written as

$$\mathrm{L}_R(y,x) = 1 - 2K_R\left(\frac{y - x'\beta}{\delta}\right), \tag{1}$$

where $K_R(u) = \mathbf{1}[|u| < 1]/2$ denotes the rectangular or uniform kernel often used in density estimation (Silverman, 1986), with $\mathbf{1}[A]$ being the indicator function for event A, and $\delta > 0$ the bandwidth parameter.

The expectation of $\mathrm{L}_R(y,x)$ is minimized when $x'\beta$ is the midpoint of the interval of length 2δ that has the highest probability of containing y (see Manski, 1991). If the conditional density of y, $f_{Y|X}(y|x)$, is strictly unimodal, the minimizer of this function approaches $\text{Mode}(y|x)$ as δ approaches zero. Moreover, for fixed δ, the minimizer of $\mathrm{L}_R(y,x)$ is $\text{Mode}(y|x)$ if $f_{Y|X}(y|x)$ is strictly unimodal and symmetric about $\text{Mode}(y|x)$ up to $\pm\delta$.[2] For fixed δ, Manski (1991) terms the minimizer of (1) the δ-mode.

The estimator proposed by Lee (1989) can be obtained by minimizing the sample analog of the expectation of (1). In particular, for a sample of size n, this is equivalent to maximizing

$$Q_n^R(\beta) = n^{-1}\sum_{i=1}^{n}\delta^{-1}K_R\left(\frac{y_i - x_i'\beta}{\delta}\right),$$

which can be recognized as a kernel estimation of the density of y_i at $x_i'\beta$.

Despite its elegance, Lee's (1989) rectangular mode regression estimator is of little practical use because, due to the nature of the objective function, its distribution is intractable (Kim and Pollard, 1990). In order to overcome this unappealing feature, Lee (1993) introduced the quadratic mode regression estimator $\hat{\beta}_Q$, which can be

[2] It is also interesting to notice that the minimizer of $\mathrm{L}_R(y,x)$ is parallel to $\text{mode}(y|x)$ if x only affects the location of $f_{Y|X}(y|x)$ (see Lee, 1989, 1993). However, this situation is not particularly interesting and consequently it will not be emphasized in what follows.

obtained by replacing the uniform kernel in (1) with the quadratic or Epanechnikov kernel (Silverman, 1986), defined as $K_Q(u) = \mathbf{1}[|u| < 1]\frac{3}{4}\{1 - u^2\}$.

Lee (1993) shows that, under the assumed regularity conditions, for fixed δ, $\hat{\beta}_Q$ is $\sqrt{n}$-consistent and asymptotically normal. As in the case of the rectangular kernel, consistency of $\hat{\beta}_Q$ requires $f_{Y|X}(y|x)$ to be unimodal and symmetric about the mode up to $\pm\delta$. Besides the enormous advantage of having a tractable asymptotic distribution, the quadratic mode estimator is also more appealing than the estimator based on the rectangular kernel in that maximizing its objective function is easier than maximizing $Q_n^R(\beta)$. Still, because the objective function based on $K_Q(u)$ is non-differentiable, relatively non-standard algorithms, like the two-step procedure proposed by Lee and Kim (1998), are needed to find $\hat{\beta}_Q$.

3. MODE REGRESSION FOR UNBOUNDED DATA

3.1. Motivation

In this section, we consider mode regression for a fully observed unbounded continuous variate, with a strictly unimodal conditional density. Given the nature of the data being considered, smooth unbounded kernels can be used in the construction of the objective function defining the mode regression estimator. This greatly facilitates the practical implementation of the estimator and the derivation of its asymptotic properties.

As noted before, for a fixed bandwidth, consistent estimation of the mode is only possible when the conditional distribution has some degree of symmetry, or when the regressors only affect the location of $f_{Y|X}(y|x)$. However, not only are these assumptions unlikely to hold in many interesting situations, but also, when they do, mode regression is likely to be less attractive. In particular, with a fixed bandwidth and an unbounded kernel, consistent estimation of the conditional mode is only possible when it coincides with, or is parallel to, the conditional mean and median. In these cases, the same slope parameters can be estimated, possibly more efficiently, by mean or quantile regression.

To widen the range of situations where mode regression is interesting and useful, we let the bandwidth parameter δ go to zero as the sample size passes to infinity. In this case it is possible to prove consistency of the proposed mode regression estimator, even for asymmetric conditional distributions with higher order moments that depend on the regressors. Of course, the fact that consistency is possible under much more general conditions has a cost. In particular, as in other cases where the objective function depends on a vanishing bandwidth (see, e.g., Parzen, 1962, Horowitz, 1992, and Seo and Linton, 2007), the estimator will not converge at the usual $\sqrt{n}$ rate. Nevertheless, as we will illustrate in Sections 4 and 5, the proposed mode regression estimator can still be useful in many empirical applications.

3.2. Model Framework

We consider a regression model of the form

$$y_i = x_i'\beta_0 + \varepsilon_i \quad (i = 1, 2, \ldots, n), \tag{2}$$

where x_i takes values in R^p for some finite p, β_0 is an unknown element of the parameter space B, which is a known subset of R^p, and the conditional density of ε_i given x_i has a strict global maximum at $\varepsilon_i = 0$ so that the conditional mode of y_i given x_i is equal $x_i'\beta_0$ (and is unique).[3] As in Lee (1989, 1993), our starting point is a loss function which can be written as one minus a (scaled) kernel. In particular, we consider a loss function of the form

$$\mathrm{L}_n(y, x) = 1 - \gamma K\left(\frac{y - x'\beta}{\delta_n}\right), \tag{3}$$

where $\gamma = K(0)^{-1} > 0$ is a scaling constant such that $\mathrm{L}_n(y, x) = 0$ when $y = x'\beta$, δ_n is a non-stochastic strictly positive bandwidth that vanishes with n, and $K(u)$ denotes

[3] Strictly speaking, the conditional density is not uniquely defined: we just require that there is a version of the conditional density with this property.

a smooth kernel function with finite third derivatives and unbounded support, such that $\int_{-\infty}^{\infty} K\left(u\right) du = 1$. Many smooth kernels are available, but throughout we focus on the popular choice $K\left(u\right) = \phi\left(u\right)$, where $\phi\left(u\right)$ denotes the standard normal density. This choice has the advantage of generating a loss function which has both the mode and the mean as minimizers in limiting cases.

Minimizing the sample analog of the expectation of (3) is equivalent to maximizing

$$Q_n(\beta) = n^{-1} \sum_{i=1}^{n} \delta_n^{-1} K\left(\frac{y_i - x_i'\beta}{\delta_n}\right), \tag{4}$$

which, for a given value of δ_n, can be done, for example, using a Newton-type algorithm (further discussion of estimation algorithms is provided in Subsection 3.4).

The maximizer of $Q_n(\beta)$, denoted $\hat{\beta}_n$, is a regression version of Parzen's (1962) mode estimator and it is possible to show that, under a set of mild regularity conditions to be detailed below, this estimator is consistent for β_0 and has a tractable asymptotic distribution.

For a fixed δ_n, the asymptotic distribution of $\hat{\beta}_n$ can be obtained using standard techniques (see, e.g., Amemiya, 1985). However, it was already noted that for a fixed δ_n this mode regression estimator is not particularly interesting. Therefore, in the next subsection, we drive the asymptotic distribution of $\hat{\beta}_n$ when δ_n is allowed to vanish as n passes to infinity.

3.3. Asymptotic Results

The basic model we consider is given by (2) and the estimator of interest is:

$$\hat{\beta}_n \equiv \arg\max_{\beta} Q_n(\beta), \tag{5}$$

where $Q_n(\beta)$ is defined as in (4).

For any given value $\beta \in B$, $Q_n(\beta)$ is a kernel-based estimator of the density function of the residuals, $\eta_i(\beta) \equiv y_i - x_i'\beta$, at 0. This identifies the parameters of interest because $f_{\eta_i(\beta)}(0) = \mathrm{E}[f_{Y|X}(x_i'\beta|x_i)]$, which is clearly maximized at $\beta = \beta_0$ provided that $f_{Y|X}(x_i'\beta|x_i) \leq f_{Y|X}(x_i'\beta_0|x_i)$ for all x and $\beta \in B$, with a strict inequality when $\beta \neq \beta_0$, on a set of x with positive probability.

Provided that the kernel is continuous and the bandwidth is finite and strictly positive, then the objective function is continuous in β for any realized data. If, in addition, the data have a well-defined joint distribution, then the value of the objective function at any fixed value of β is clearly a random variable. Then, if the parameter space B is a compact subset of a finite dimensional Euclidean space, it follows that our estimator is well-defined in that there exists a random variable $\hat{\beta}_n$ which satisfies Equation (5), except possibly on a set of probability zero.

Below we present the main results on the asymptotic properties of $\hat{\beta}_n$ as a set of three theorems whose proofs are provided in the Appendix. Before the theorems are presented, we give details on the assumptions under which they are valid.

3.3.1. Consistency

In order to prove consistency, we make the following assumptions.

A1 *Data Generation Process*

$\{(\varepsilon_i, x_i)\}_{i=1}^{\infty}$ is an iid sequence, where ε_i takes values in R and x_i takes values in R^p for some finite p.

A2 *Parameter Space and Parameter Value: I*

B is a compact subset of R^p and $\beta_0 \in B$.

A3 *Distribution of x: I*

(i) $\mathrm{E}\{|x_i|\} < \infty$, where $|a|$ denotes the Euclidean norm of a for any scalar or finite-dimensional vector a.

(ii) $\Pr\{x_i'\lambda = 0\} < 1$ for all fixed $\lambda \neq 0$.

A4 *Conditional Density of ε Given x: I*

There exists a version of the conditional density of ε given x, denoted $f_{\varepsilon|X}(\cdot|\cdot) : R \times R^p \to R$, such that:

(i) $\sup_{\varepsilon \in R,\ x \in R^p} f_{\varepsilon|X}(\varepsilon|x) = L_0 < \infty$.

(ii) $f_{\varepsilon|X}(\varepsilon|x)$ is continuous in ε for all ε and x.

(iii) $f_{\varepsilon|X}(\varepsilon|x) \leq f_{\varepsilon|X}(0|x)$ for all ε and x. In addition, there exists a set $A \subseteq R^p$ such that $\Pr\{x_i \in A\} = 1$ and $f_{\varepsilon|X}(\varepsilon|x) < f_{\varepsilon|X}(0|x)$ for all $\varepsilon \neq 0$ and $x \in A$.

A5 *Kernel Function: I*

$K(\cdot) : R \rightarrow R$ is a differentiable kernel function such that:

(i) $\int_{-\infty}^{\infty} K(u)du = 1$.

(ii) $\sup_{u \in R} |K(u)| = c_0 < \infty$.

(iii) $\sup_{u \in R} |K'(u)| = c_1 < \infty$, where $K'(u) = dK(u)/du$.

A6 *Bandwidth Sequence: I*

$\{\delta_n\}_{n=1}^{\infty}$ is a strictly positive bandwidth sequence such that:

(i) $\delta_n \rightarrow 0$.

(ii) $n\delta_n^{1+\sigma} \rightarrow \infty$ for some $\sigma > 0$.

We make Assumption A1 for convenience: the assumptions in the paper could be modified to allow the $\{(\varepsilon_i, x_i)\}_{i=1}^{\infty}$ process to exhibit some dependence but this would complicate the proofs quite substantially and there would be a trade-off between allowing some dependence in the $\{(\varepsilon_i, x_i)\}_{i=1}^{\infty}$ process (captured, for example, by mixing rates) and strengthening other assumptions (mostly on the moments of x_i). Assumptions A2 and A3 are standard. Parts (i) and (ii) of A4 are standard. Part (iii) of A4 is specific to the context of mode regression. Assumption A5 is fairly standard and is satisfied by many commonly used kernel functions though the required continuity does rule out the use of the rectangular kernel which was adopted in Lee's (1989) original analysis of mode regression. Assumption A6 is a fairly standard condition on the bandwidth sequence and specifies that the bandwidth goes to 0 at a suitably rapid rate. It is required for the proof of consistency since, unlike Lee (1989, 1993), we do not assume the conditional density of the errors given the regressors is symmetric on an interval around the mode.

Under these assumptions we can establish consistency.

Theorem 1 *Under Assumptions A1–A6, $\hat{\beta}_n \xrightarrow{p} \beta_0$.*

3.3.2. Asymptotic Normality

The proof of asymptotic normality requires the following additional assumptions.

B1 *Distribution of x: II*

$\mathrm{E}\{|x_i|^{5+\xi}\} < \infty$ for some $\xi > 0$.

B2 *Parameter Space and Parameter Value: II*

β_0 belongs to the interior of B.

B3 *Conditional Density of ε Given x: II*

$f_{\varepsilon|X}(\varepsilon|x)$ is three times differentiable with respect to ε for all x such that:

(i) $f^{(j)}_{\varepsilon|X}(\varepsilon|x) = \partial^j f_{\varepsilon|X}(\varepsilon|x)/\partial\varepsilon^j$ is uniformly bounded for $j = 1, 2, 3$.

(ii) $E\left[f^{(2)}_{\varepsilon|X}(0|x_i)x_i x_i'\right]$ is symmetric negative definite.

B4 *Kernel Function: II*

(i) $K(\cdot)$ is three times differentiable.

(ii) $\int_{-\infty}^{\infty} uK(u)du = 0$.

(iii) $\lim_{u\to\pm\infty} K(u) = 0$.

(iv) $\int_{-\infty}^{\infty} u^2|K(u)|du = M_0 < \infty$.

(v) $\int_{-\infty}^{\infty} |K'(u)|^2\, du = M_1 < \infty$.

(vi) $\sup_{u\in R} |K''(u)| = M_2 < \infty$.

(vii) $\sup_{u\in R} |K'''(u)| = M_3 < \infty$.

(viii) $\int_{-\infty}^{\infty} |K''(u)|^2 du = M_4 < \infty$.

B5 *Bandwidth Sequence: II*

(i) $n\delta_n^7 = o(1)$.

(ii) $n\delta_n^{5+\sigma} \to \infty$, for some $0 < \sigma < 2$.

Unsurprisingly, each of these additional assumptions involves strengthening a corresponding earlier assumption used in establishing consistency. Of these, the most interesting is Assumption B5 which pins down further the convergence rate used for the bandwidth sequence, and is closely related to the bandwidth assumption made by Parzen (1962) to establish consistency of the kernel mode estimator. In terms of our notation, Parzen assumed that $n\delta_n^6 \to \infty$ while $n\delta_n^{5+2\tau} = o(1)$ for some $0 < \tau < 1$. Since, like us, Parzen assumes that $\delta_n = o(1)$ then his assumptions only make sense if $1/2 < \tau < 1$. Parzen's bandwidth assumptions then imply that our bandwidth assumptions hold with $\sigma = 1$, and thus our bandwidth assumptions are more general than those used by Parzen. It should, however, be noted that this τ parameter is also involved in Parzen's assumptions on the smoothness of the density of the errors and on the smoothness of the kernel.

We are now in position to obtain the asymptotic distribution of $\hat{\beta}_n$.

Theorem 2 *Under Assumptions A1–A6 and B1–B5:*

$$(n\delta_n^3)^{1/2}\left[\hat{\beta}_n - \beta_0\right] \xrightarrow{D} \mathcal{N}[0, \Omega_0], \tag{6}$$

where:

$$\Omega_0 = B_0^{-1} A_0 B_0^{-1}, \tag{7}$$

$$A_0 = \lim_{n\to\infty} \mathrm{Var}\left[(n\delta_n^3)^{1/2}\left(\left.\frac{\partial Q_n(\beta)}{\partial \beta}\right|_{\beta_0}\right)\right] = M_1 \mathrm{E}\left[f_{\varepsilon|X}(0|x_i)x_i x_i'\right], \tag{8}$$

$$B_0 = \lim_{n\to\infty} \mathrm{E}\left(\left.\frac{\partial^2 Q_n(\beta)}{\partial \beta \partial \beta'}\right|_{\beta_0}\right) = \mathrm{E}\left[f_{\varepsilon|X}^{(2)}(0|x_i)x_i x_i'\right]. \tag{9}$$

This theorem reveals that, given our bandwidth assumptions, the proposed mode regression estimator converges to a normal distribution at a rate that can be made arbitrarily close to $n^{2/7}$. Moreover, we see that the variance of the asymptotic distribution, Ω_0, depends on the choice of kernel, through M_1, and on the interplay between characteristics of the distributions of the regressors and error term. In particular, Ω_0 depends both on how high and on how concave the conditional density of ε is at the mode.

The following theorem provides a way of obtaining a consistent estimator of Ω_0.

Theorem 3 *Under Assumptions A1–A6 and B1–B5:*

$$\widehat{\Omega}_n \xrightarrow{p} \Omega_0 \tag{10}$$

where:

$$\widehat{\Omega}_n = \widehat{B}_n^{-1} \widehat{A}_n \widehat{B}_n^{-1}, \tag{11}$$

$$\widehat{A}_n = n^{-1} \sum_{i=1}^{n} \delta_n^{-1} \left[K' \left(\frac{y_i - x_i' \hat{\beta}_n}{\delta_n} \right) \right]^2 (x_i x_i'), \tag{12}$$

$$\widehat{B}_n = n^{-1} \sum_{i=1}^{n} \delta_n^{-3} K'' \left(\frac{y_i - x_i' \hat{\beta}_n}{\delta_n} \right) (x_i x_i'). \tag{13}$$

Here, $\widehat{B}_n$ is the conventional observed Hessian estimator, while $\widehat{A}_n$ is an outer-product of the gradient variance estimator rescaled by the factor δ_n^3. This rescaling arises because the gradient needs to be multiplied by $(n\delta_n^3)^{1/2}$ rather than by $n^{1/2}$ to have a non-degenerate limiting distribution.

3.4. Implementation issues

Two issues are of paramount importance in the implementation of the proposed mode regression estimator. One, of course, is the choice of the bandwidth parameter to use in any particular application. The other is the choice of algorithm to use in the maximization because the objective function may have multiple maxima, especially for small values of δ_n, and therefore it is important to ensure that a global maximum is found.

Our approach to both of these problems is based on the observation that, for $K(u) = \phi(u)$,[4] maximization of (4) can be seen as solving the following set of moment conditions

$$\mathrm{E} \left[\exp \left(-\frac{(y_i - x_i'\beta)^2}{2\delta_n^2} \right) (y_i - x_i'\beta)\, x_i' \right] = 0. \tag{14}$$

[4] More generally, a similar result holds whenever the kernel used is a function of $(y_i - x_i\beta)^2$.

Equation (14) makes clear that maximization of (4) is essentially a weighted least squares problem that has as special cases mode regression, when δ_n passes to zero as $n \to \infty$, and mean regression, when $\delta_n \to \infty$. This equation also reveals the close link between the mode regression estimator proposed here and the family of robust M-estimators (Huber, 1973) that aim to "give a good fit to the bulk of the data without being perturbed by a small proportion of outliers" (Maronna, Martin and Yohai, 2006, p. 88). In particular, under certain conditions, M-estimators like the one based on biweights (Beaton and Tukey, 1974), can also be interpreted as mode regression estimators.[5] The link between mode regression and robust M-estimators was noted by Lee (1989, 1993) and is explored in Baldauf and Santos Silva (2009).

The most important feature of (14), however, is that it shows that (4) defines a continuum of conditional measures of central tendency, of which the two polar cases are of particular interest. Therefore, rather than just estimate the conditional mean and mode, for a chosen value of δ_n, we can estimate the parameters of interest for a wide range of values of δ_n and obtain a more detailed picture of how these parameters, say $\beta_{(\delta_n)}$, vary within this class of conditional measures of central tendency.

Of course, it is still necessary to define the limits for the sequence of values of δ_n to be used in the estimation. However, because inference will not be based on a single value of the smoothing parameter, this choice is less critical than the choice of an optimal value of δ_n to estimate the mode. In the application in Section 5, we estimate $\beta_{(\delta_n)}$ for 100 values of δ_n between $50\textsc{mad}$ and $0.5\textsc{mad}n^{-0.143}$, where $\textsc{mad}$ denotes the median of the absolute deviation from the median OLS residual, i.e., denoting by b the OLS estimates of β, $\textsc{mad} = \underset{i}{\text{med}}\left\{\left|(y_i - x_i'b) - \underset{j}{\text{med}}\left(y_j - x_j'b\right)\right|\right\}$.[6]

From a computational point of view, this strategy is attractive because OLS provides a natural set of starting values for the estimation of $\beta_{(\delta_n)}$ when δ_n is large

[5] These estimators are impelemented in popular software packages such as Stata (StataCorp., 2007), SAS (SAS Institute Inc., 2008), Matlab (Mathworks, 2008), and R and S-PLUS (Venables and Ripley, 2002, and Heiberger and Becker, 1992).

[6] For comparison, we note that *rreg* in Stata uses a smoothing parameter (for a triweight kernel) equal to $7\textsc{mad}$. This would correspond to a bandwidth of $2.33\textsc{mad}$ with a Gaussian kernel.

enough. Subsequently, the new estimation results can be used as starting values for the estimation with a smaller value of the smoothing parameter. Of course, there is no guarantee that the estimates obtained in this way will correspond to the global maxima of the objective function for each value of δ_n. Therefore, it is recommended that, at least for an interesting value of the smoothing parameter, additional checks are performed to try to ensure that a global maximum is indeed found.

To better interpret the estimates obtained with the different values of the smoothing parameter, it is interesting to compute an auxiliary estimation result. For large values of δ_n, the weights in (14) are approximately equal to one for every observation. That is, the weights sum to n. As δ_n passes to zero, the value of the weights will vary from observation to observation, being often much smaller than one. We suggest that, for each value of δ_n, the sum of suitably normalized weights (SNW)[7] is saved and used as an heuristic indication of the number of observations "effectively" used in the estimation.

4. SIMULATION EVIDENCE

This section presents the results of a small simulation study illustrating the finite sample performance of the proposed mode regression estimator. In these experiments data are generated by the simple linear model

$$y_i = \beta_0 + \beta_1 x_i + (1 + v x_i)\varepsilon_i \quad (i = 1, 2, \ldots, n),$$

where x_i is a random regressor, ε_i is a random disturbance that is statistically independent of the regressor, and v is the parameter that controls the degree of heteroskedasticity. Throughout, we set $\beta_0 = 0$ and $\beta_1 = 1$.

To avoid overly optimistic results, x_i is generated from a skewed distribution.[8] In particular, for each replication of the simulations, the regressor is newly generated as

[7] Because the estimation results are obviously invariant to a rescaling of the weights, these should be normalized so that their maximum is equal to one.

[8] See Chesher and Peters (1994) and Chesher (1995).

independent draws from the $\chi^2_{(3)}$ distribution, and scaled to have variance equal to one.

To complete the design of the experiments it is necessary to define how ε_i is generated. In the present context, it is important to generate ε_i using a distribution that meets the following criteria: 1) is unimodal, 2) has unbounded support, 3) is capable of exhibiting varying degrees of skewness, 4) is such that the mode and the first three moments are easy to parametrize, and 5) is easy to simulate. To satisfy all these requirements, we generate ε_i as independent draws from a re-scaled log-gamma random variable

$$\varepsilon_i = -\lambda \ln(Z_i), \qquad \lambda > 0,$$

where Z_i has a gamma distribution with mean α/κ and variance α/κ^2, for $\alpha, \kappa > 0$.

It is possible to show that the mode of ε_i is given by $\lambda \ln(\kappa/\alpha)$, and therefore we set $\kappa = \alpha$ to ensure that ε_i has zero mode. For this choice of parameters, ε_i will have positive expectation defined by $\mu_\varepsilon = \lambda [\ln(\kappa) - \psi_0(\alpha)]$, where $\psi_0(\cdot)$ denotes the digamma function. The variance of ε_i is given by $\lambda^2 \psi_1(\alpha)$, where $\psi_1(\cdot)$ is the trigamma function, and in our experiments the value of λ is set so that the unconditional variance of the error $(1 + vx_i)\varepsilon_i$ is equal to one.[9] Finally, ε_i is positively skewed, with coefficient of skewness $-\psi_2(\alpha)\psi_1(\alpha)^{-3/2}$, where $\psi_2(\cdot)$ is the quadrigamma function. Having fixed κ and λ, α can be used to control the degree of skewness of the distribution.

We perform experiments with $\alpha \in \{0.05, 5.00\}$,[10] $v \in \{0, 1, 2\}$ and $n \in \{250, 1000, 4000, 16000\}$. For each replication of the experiments, we estimate the conditional mean and the conditional mode of y_i, which are both linear functions of x_i. Specifically, with this design, $\text{Mode}(y_i|x_i) = x_i$ and $\text{E}(y_i|x_i) = \mu_\varepsilon + (1 + v\mu_\varepsilon)x_i$, which show that for the homoskedastic cases ($v = 0$) the conditional mean and the conditional mode have the same slope parameter. The mode regression estimator was

[9] Specifically, $\lambda = \left[\left(1 + 2\text{E}(x_i)v + \text{E}(x_i^2)v^2\right)\psi_1(\alpha)\right]^{-0.5}$.

[10] For $\alpha = 5.00$ the coefficient of skewness of ε_i is approximately 0.5, being approximately 2.0 for $\alpha = 0.05$.

implemented using the iterative weighted least squares estimator described in Subsection 3.4, for smoothing parameters defined as $\delta_n = k\text{MAD}n^{-0.143}$, with $k \in \{0.6, 1.2\}$.

Table 1 summarises the main simulation results obtained with 10000 replications of the simulation procedure. Specifically, for the 24 cases considered, the table displays the mean and standard error of the estimated intercepts and slopes for the three estimators included in these experiments: OLS, mode regression with $k = 1.2$, labelled Mode 1.2, and mode regression with $k = 0.6$, labelled Mode 0.6.

The OLS results are not surprising in any way and illustrate the well-known properties of this estimator. In particular, because OLS is unbiased and converges at the usual $\sqrt{n}$ rate, the mean of the OLS estimates is almost invariant to the sample size, but its standard errors are roughly halved each time the sample size increases by a factor of 4. These results, therefore, provide an interesting benchmark against which the performance of the mode regression estimators can be evaluated.

As for the results obtained with Mode 1.2 and Mode 0.6, perhaps the most remarkable finding is the fact that the intercept picks-up most of the bias, with the mean of the estimates of the slope being always close to one. Not surprisingly, the biases shrink with the sample size, but the rate at which the biases vanish depends on the degree of both skewness and heteroskedasticity of the errors. Like the biases, the standard errors of the mode estimators also shrink with the sample size and again the rate at which this happens depends on the characteristics of the conditional distribution. Generally speaking, as expected, Mode 1.2 has smaller standard errors but larger biases than Mode 0.6. The efficiency penalty of Mode 0.6 is especially severe for less skewed and less heteroskedastic errors.

As noted above, the OLS and the mode regression identify the same slope when $v = 0$. Therefore, for these cases, it is meaningful to compare the results of the mode estimators with those of the OLS. In particular, it is interesting to notice that for $\alpha = 5.00$ the slopes are estimated with much better precision by OLS, but for $\alpha = 0.05$ the mode estimators are strong competitors, with Mode 1.2 outperforming

Table 1: Simulations results

			Intercept				Slope			
α	v	$n =$	250	1000	4000	16000	250	1000	4000	16000
5.00	0	OLS	0.220 (0.101)	0.220 (0.050)	0.220 (0.025)	0.220 (0.013)	0.999 (0.064)	1.000 (0.032)	1.000 (0.016)	1.000 (0.008)
		Mode 1.2	0.035 (0.252)	0.021 (0.164)	0.014 (0.106)	0.010 (0.068)	1.005 (0.158)	1.003 (0.102)	1.002 (0.066)	1.000 (0.043)
		Mode 0.6	0.028 (0.396)	0.012 (0.277)	0.005 (0.199)	0.005 (0.142)	1.007 (0.246)	1.005 (0.166)	1.004 (0.114)	1.001 (0.081)
	1	OLS	0.090 (0.113)	0.090 (0.057)	0.090 (0.028)	0.090 (0.014)	1.089 (0.112)	1.090 (0.056)	1.090 (0.029)	1.090 (0.014)
		Mode 1.2	0.037 (0.134)	0.029 (0.083)	0.020 (0.053)	0.015 (0.034)	1.000 (0.215)	0.995 (0.145)	0.996 (0.099)	0.996 (0.067)
		Mode 0.6	0.022 (0.213)	0.014 (0.144)	0.007 (0.098)	0.005 (0.066)	1.005 (0.292)	1.000 (0.208)	1.002 (0.149)	1.000 (0.106)
	2	OLS	0.055 (0.119)	0.055 (0.061)	0.055 (0.031)	0.055 (0.015)	1.109 (0.123)	1.110 (0.062)	1.110 (0.031)	1.110 (0.016)
		Mode 1.2	0.036 (0.099)	0.029 (0.059)	0.022 (0.037)	0.017 (0.023)	0.999 (0.207)	0.992 (0.138)	0.993 (0.094)	0.993 (0.063)
		Mode 0.6	0.017 (0.142)	0.013 (0.090)	0.007 (0.059)	0.005 (0.039)	1.009 (0.274)	1.000 (0.190)	1.000 (0.137)	0.999 (0.096)
0.05	0	OLS	0.872 (0.101)	0.874 (0.050)	0.874 (0.025)	0.873 (0.012)	1.000 (0.064)	1.000 (0.032)	1.000 (0.016)	1.000 (0.008)
		Mode 1.2	0.313 (0.072)	0.261 (0.035)	0.222 (0.017)	0.183 (0.009)	1.005 (0.049)	1.001 (0.023)	1.001 (0.011)	1.000 (0.006)
		Mode 0.6	0.168 (0.094)	0.131 (0.042)	0.103 (0.022)	0.079 (0.012)	1.013 (0.068)	1.005 (0.023)	1.001 (0.015)	1.001 (0.008)
	1	OLS	0.358 (0.111)	0.359 (0.057)	0.358 (0.029)	0.358 (0.014)	1.357 (0.110)	1.358 (0.057)	1.358 (0.029)	1.358 (0.014)
		Mode 1.2	0.220 (0.076)	0.195 (0.028)	0.170 (0.014)	0.146 (0.007)	1.035 (0.071)	1.015 (0.034)	1.005 (0.017)	0.998 (0.009)
		Mode 0.6	0.130 (0.061)	0.110 (0.031)	0.092 (0.016)	0.075 (0.009)	1.022 (0.088)	1.002 (0.046)	0.993 (0.025)	0.990 (0.014)
	2	OLS	0.219 (0.117)	0.220 (0.061)	0.219 (0.031)	0.219 (0.015)	1.437 (0.120)	1.438 (0.063)	1.438 (0.031)	1.438 (0.016)
		Mode 1.2	0.181 (0.050)	0.164 (0.024)	0.146 (0.012)	0.128 (0.006)	1.048 (0.072)	1.025 (0.036)	1.011 (0.018)	1.002 (0.009)
		Mode 0.6	0.113 (0.049)	0.098 (0.025)	0.084 (0.013)	0.070 (0.007)	1.026 (0.086)	1.004 (0.046)	0.994 (0.025)	0.989 (0.014)

OLS for all sample sizes considered in these exercises. The competitiveness of the mode regression in this case is, of course, a reflex of the well-known fact that OLS can be outperformed by "robust" estimators when the distribution of the errors has high skewness and/or kurtosis (see, e.g., Maronna et al., 2006).

Overall, the results of these experiments are quite encouraging in that they show that the proposed mode estimator is likely to have a reasonable performance in samples of a realistic size. Naturally, the conditional mode is often estimated with much less precision than the conditional mean, but in most cases this comparison has little meaning as the two location functions generally provide very different information about the conditional distribution of interest.

5. AN EMPIRICAL ILLUSTRATION - THE RECENT EVOLUTION OF BMI IN ENGLAND

The economic effects of obesity have attracted substantial interest in recent years (see, for example, Averett and Korenman, 1996, Cawley, 2004, and Morris, 2006 and 2007), and are at the centre of attention for many policy makers in western countries (e.g., U.S. Department of Health and Human Services, 2001, and Department of Health, 2004). Therefore, the study of the trends in obesity is likely to be of interest to a wide audience (see Mills, 2009, for a recent example of a study of this kind for England).

In this section we illustrate the use of mode regression by studying the recent evolution of the body mass index (BMI)[11] in England. In particular, we use individual data from the Health Survey for England[12] to study the evolution over the period 1997-2006 of different location measures for the conditional distribution of the BMI,

[11] The body mass index of an individual is defined as his body weight, measured in kilograms, divided by the square of his height, measured in meters.

[12] The Health Survey for England is a set of cross-sectional surveys commissioned by the Department of Health and annually carried out since 1991.

for males and non-pregnant females, aged between 18 and 65 at the time of the interview, for whom a valid BMI measurement could be obtained.

It must be emphasized that the purpose of this study is not to attempt to explain the causes of the observed trends (as done for example by Cutler, Glaeser and Shapiro, 2003, and Chou, Grossman and Saffer, 2004), but simply to describe how the conditional distribution of the BMI has changed over time. Therefore, although the Health Survey for England contains detailed information on many behavioural risk factors, like eating and drinking habits, here we condition only on covariates characterizing the composition of the population. Specifically, besides gender and the year of the survey (Year), we condition only on the age of the respondent (Age) and on an indicator of whether or not the individual is white (Non-white).

Table 2 presents the estimation results obtained with the traditional mean and quantile regressions. Separate models are estimated for males and females and, in both cases, the regressors Year and Age are transformed so that the intercept corresponds to the BMI for a forty years old white individual in the year 2000.

The results for males indicate that Year has a positive and statistically significant effect, both on the mean and on the estimated quantiles. Moreover, the impact of Year is much stronger on the upper-tail of the distribution, indicating that over time the distribution is becoming more spread-out and positively skewed.

The results for females are not much different, although the effect of Year is less pronounced, and in this case it is not statistically significant for the lower estimated quantile ($\theta = 0.1$).

Although our interest is focused on the effect of Year, it is nonetheless noteworthy that the effect of the dummy Non-white on the mean and median regressions has different signs for males and females, but for the extreme quantiles the sign of this effect is the same for the two samples.

Turning now to the mode regression, Figure 1 displays the estimated coefficient on Year against the SNW (effective sample size) for a range of values of δ_n, for the samples of males and females. This picture was obtained by maximizing (4) for 100 values of δ_n between 50MAD and 0.5MAD$n^{-0.143}$. For each value of δ_n, several sets of starting values were used to try to ensure that a global maximum of the objective function was found. For values of δ_n smaller than 0.5MAD$n^{-0.143}$, the objective functions have multiple, almost identical, maxima and consequently the estimates become unstable.

Table 2: Estimation results for mean and quantile regressions

	Mean	Quantile Regression		
Regressors	Regression	$\theta = 0.1$	$\theta = 0.5$	$\theta = 0.9$
	Males, $n = 38125$			
Year	0.097 (0.008)	0.028 (0.010)	0.083 (0.009)	0.181 (0.018)
Non-white	−0.729 (0.080)	−0.565 (0.097)	−0.702 (0.102)	−0.962 (0.201)
ln (Age)	2.392 (0.130)	2.529 (0.153)	2.458 (0.136)	1.953 (0.321)
$[\ln(\text{Age})]^2$	−3.199 (0.271)	−2.547 (0.410)	−3.025 (0.317)	−3.676 (0.551)
$[\ln(\text{Age})]^3$	0.733 (0.520)	1.127 (0.700)	1.525 (0.567)	1.261 (1.247)
Intercept	27.344 (0.035)	22.519 (0.047)	26.875 (0.039)	32.649 (0.076)
	Females, $n = 44651$			
Year	0.064 (0.009)	0.007 (0.008)	0.051 (0.009)	0.154 (0.026)
Non-white	0.074 (0.094)	−0.158 (0.081)	0.428 (0.097)	−0.238 (0.217)
ln (Age)	3.051 (0.154)	2.522 (0.140)	3.554 (0.145)	2.342 (0.435)
$[\ln(\text{Age})]^2$	−0.342 (0.323)	0.166 (0.284)	0.566 (0.348)	−1.661 (0.871)
$[\ln(\text{Age})]^3$	0.733 (0.630)	0.038 (0.566)	0.828 (0.589)	3.087 (1.640)
Intercept	26.610 (0.041)	21.008 (0.035)	25.380 (0.042)	33.940 (0.123)

For both samples, it is clear that for large values of SNW (large δ_n) the estimated coefficient on Year is identical to the one obtained by OLS. For males, as δ_n passes to zero, the estimates of the parameter of interest smoothly decline to about 0.06, becoming reasonably stable except for the smaller values of SNW. These results suggest that the location of the mode of the conditional distribution is also drifting up with time, but not as quickly as the location of the conditional mean and median. For females, the results are much more striking. Indeed, the estimates of the coefficient of interest decline almost monotonically and, in sharp contrast to what was found with mean and quantile regression, become negative for values of the smoothing parameter smaller than about $\delta_n = 1.75\textsc{mad}n^{-0.143}$.

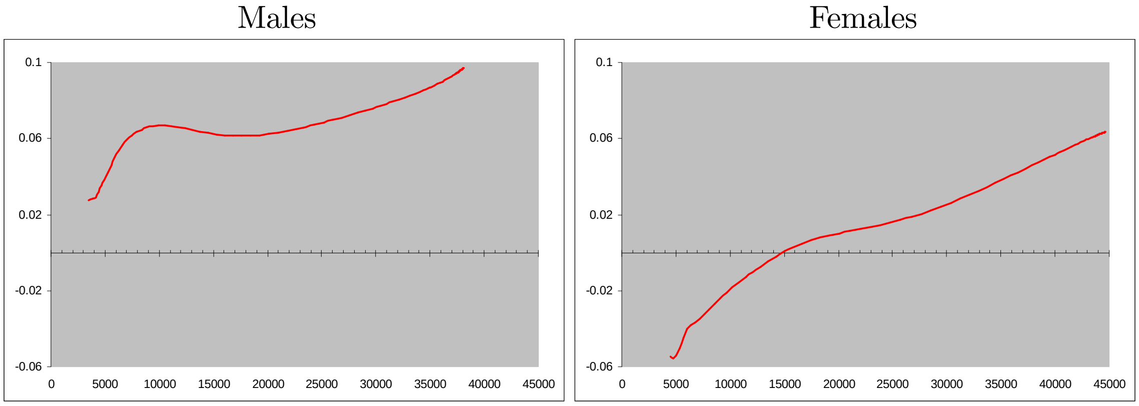

Fig. 1 – Mode regression results: estimated coefficient on Year versus the sum of normalized weights for different values of δ_n, for the samples of males and females

In view of the results in Figure 1, and consistently with the simulation results presented before, we focus on the mode regression estimator obtained with $\delta_n = k\textsc{mad}n^{-0.143}$, $k \in \{0.6, 1.2\}$, whose results are displayed in Table 3.

For males, as expected, the mode regression results lead to an estimated intercept that is smaller than those obtained by mean or median regression. More interestingly, we find that the coefficient of Year is also somewhat smaller than those obtained with other conditional measures of central tendency, and it is only significant for $k = 1.2$.

For females we again find that the estimated intercept in the mode regression is smaller than those obtained by mean or median regression. However, the most notable feature of the mode regression results for females is that, as Figure 1 revealed, we find that Year has a negative effect on the conditional mode, albeit not statistically significant at the usual levels.

Therefore, for females, we find that although the mean and most (if not all) quantiles of the distribution of interest are increasing functions of Year, the same does not seem to happen for the conditional mode. Moreover, in contradistinction with what was found for the mean and median regressions, in the mode regressions the dummy NON-WHITE has a negative effect both for males and females, and it is only statistically significant for males. Interestingly, the mode regression results for the coefficients on the powers of ln (AGE) are not much different, both in size and in statistical significance at conventional levels, from those obtained by mean regression.

It is also worth noting that, as could be expected for this kind of data and for samples of this size, mode regression parameters are estimated with less precision than the corresponding ones obtained by mean or median regressions. However, this just reflects the fact that we have more information about some features of the conditional distribution than about others. In spite of its larger variance, the mode regression estimator proved to be useful in this particular application by revealing that, in contrast with what is happening with the mean and median, the mode of the conditional distribution of the BMI for females does not seem to be increasing over time, and may actually be decreasing. Overall, these results illustrate that the mode regression can provide information on how the regressors affect the location and shape of the conditional distribution that cannot be easily elicited using the more standard mean and quantile regressions.

Table 3: Estimation results for mode regressions

	Males, $n = 38125$		Females, $n = 44651$	
Regressors	$k = 1.2$	$k = 0.6$	$k = 1.2$	$k = 0.6$
YEAR	0.064 (0.027)	0.029 (0.037)	−0.018 (0.028)	−0.051 (0.061)
NON-WHITE	−1.162 (0.196)	−1.588 (0.447)	−0.079 (0.444)	−0.536 (0.351)
ln (AGE)	2.690 (0.490)	2.448 (0.601)	3.904 (0.474)	3.828 (0.750)
$[\ln(\text{AGE})]^2$	−2.416 (1.019)	−3.730 (1.089)	0.968 (1.066)	−1.219 (1.482)
$[\ln(\text{AGE})]^3$	3.037 (1.617)	2.650 (1.655)	−0.828 (1.798)	−3.090 (2.441)
INTERCEPT	26.179 (0.169)	26.477 (0.176)	23.753 (0.153)	23.942 (0.304)
SNW	7987.334	4125.910	10282.773	5321.951
δ_n	0.76991	0.38496	0.97590	0.48795

6. CONCLUDING REMARKS

In this paper we provide the asymptotic results needed for valid inference about the conditional mode when estimation is based on unbounded smooth kernel and the bandwidth parameter is allowed to pass to zero as the sample size increases. The estimator is very easy to implement and it is valid under mild conditions. In particular, its asymptotic properties do not depend on the symmetry and homoskedasticity of the conditional distribution of interest. The main drawback of this estimator is that it converges at a rate much smaller that the usual $\sqrt{n}$. In spite of this, the simulation results presented in Section 4 and illustrative application in Section 5 suggest that the mode regression estimator can be a useful tool in many applications.

There are, of course, many aspects of mode regression that deserve further investigation. In particular, it would be interesting to define a goodness-of-fit criterion for mode regression and to use it to develop a cross-validation procedure to optimally select the bandwidth parameter.

APPENDIX

Throughout, $|\cdot|$ denotes the Euclidean norm so that $|a| = \text{abs}(a)$ for any scalar a, $|a| = (a'a)^{1/2}$ for any finite-dimensional vector, and $|A| = [\text{tr}(A'A)]^{1/2}$ for any finite-dimensional matrix A. Also, integrals are taken over their entire range unless explicitly indicated otherwise, so $\int a(u)\ du = \int_R a(u)\ du$, when $a(\cdot)$ is a scalar valued function, $\int a(x)\ dF_X(x) = \int_{R^s} a(x)\ dF_X(x)$, when $a(\cdot)$ is an s-dimensional vector valued function and $F_X(\cdot)$ is a cdf on R^s (for finite s), and so on. In addition, we use $a_n \sim b_n$ to denote that both a_n/b_n and b_n/a_n are $O(1)$.

A.1 Proof of Theorem 1

There are two main parts to the proof of this theorem. First, in Lemma 1, below, we establish that $Q_0(\beta)$ exists and that it is continuous in $\beta \in B$ with a unique global maximum at $\beta = \beta_0$. Second, in Lemma 2 below, we establish that

$$\sup_{\beta \in B} |Q_n(\beta) - Q_0(\beta)| = o_p(1), \tag{15}$$

i.e., $Q_n(\beta)$ satisfies a uniform law of large numbers. Since B is compact, then the result of the theorem follows by application of Theorem 2.1 of Newey and McFadden (1994).

Lemma 1 *Under Assumptions A1–A6,* $Q_0 = \lim_{n\to\infty} \mathrm{E}[Q_n(\beta)]$ *exists, is continuous in* $\beta \in B$*, and has a unique global maximum at* $\beta = \beta_0$*.*

Proof. First, observe that, since $\{(\varepsilon_i, x_i)\}_{i=1}^{\infty}$ are iid by Assumption A1, then:

$$\begin{aligned}
\mathrm{E}[Q_n(\beta)] &= \mathrm{E}\left[\delta_n^{-1} K\left(\frac{y_i - x_i'\beta}{\delta_n}\right)\right] \\
&= \int \delta_n^{-1} K\left(\frac{\varepsilon - x'(\beta - \beta_0)}{\delta_n}\right) f_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x) \\
&= \int K(u) f_{Y|X}(x'(\beta - \beta_0) + \delta_n u|x)\ du\ dF_X(x),
\end{aligned}$$

where $F_X(\cdot)$ is the distribution function of x. By Assumptions A4(i) and A4(ii) we have that $f_{\varepsilon|X}(\varepsilon|x)$ is continuous in ε for all ε and x, and is uniformly bounded from above. By Assumptions A5(i) and A5(ii) we have that $\int |K(u)|\ du\ dF_X(x) = \int |K(u)|\ du < \infty$ and that $\int K(u)\ du = 1$. Combining these with Assumption A6(i), it follows by dominated convergence that:

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{E}[Q_n(\beta)] &= \int K(u) f_{\varepsilon|X}(x'(\beta - \beta_0)|x)\ du\ dF_X(x) \\
&= \int K(u)\ du \cdot \mathrm{E}[f_{\varepsilon|X}(x_i'(\beta - \beta_0)|x_i)] \\
&= \mathrm{E}[f_{\varepsilon|X}(x_i'(\beta - \beta_0)|x_i)] = Q_0(\beta),
\end{aligned}$$

which establishes the existence of $Q_0(\beta)$.

Second, since $f_{\varepsilon|X}(\varepsilon|x)$ is continuous in ε for all ε and x, and is uniformly bounded from above by Assumptions A4(i) and A4(ii), it then follows by dominated convergence that $\mathrm{E}[f_{\varepsilon|X}(x_i'(\beta - \beta_0)|x_i)]$ is continuous in $\beta \in B$.

Third, since $\Pr(x_i'\lambda = 0) < 1$ for all fixed $\lambda \neq 0$ by Assumptions A3(ii), and $f_{\varepsilon|X}(\varepsilon|x)$ achieves a strict global maximum at $\varepsilon = 0$ for every x in a set of probability 1 by Assumption A4(iii), it follows that $\mathrm{E}[f_{\varepsilon|X}(x_i'(\beta-\beta_0)|x_i)]$ achieves a strict global maximum at $\beta = \beta_0$. ■

Lemma 2 *Under Assumptions A1–A6:*

$$\sup_{\beta\in B}|Q_n(\beta) - Q_0(\beta)| = o_p(1). \tag{16}$$

Proof. The proof follows lines somewhat similar to those of the proof of Theorem 1 from Hansen (1996). First, let $N(k) = 2^{pk}$ for $k = 1, 2, \ldots$. Then, since B is a compact subset of R^p by Assumption A2, there exists a constant G_1 and a set $B_\infty = \{\beta^s\}_{s=1}^{\infty} \subset B$ such that:

$$\sup_{\beta\in B}\left\{\min_{1\leq s\leq N(k)}|\beta - \beta^s|\right\} \leq \frac{G_1}{2^k} \qquad (k = 1, 2, \ldots). \tag{17}$$

Next, for each $k = 1, 2, \ldots$, define $B_k = \{\beta^s\}_{s=1}^{N(k)}$ and let $\bar\beta_k(\cdot) : B \to B_k$ be a function which satisfies $|\beta - \bar\beta_k| \leq G_1/2^k$ for all $\beta \in B$ (clearly such a function exists for each $k = 1, 2, \ldots$). Then, select $\{k_n\}_{n=1}^{\infty}$ to be a sequence of positive integers such that $2^{k_n} \sim \delta_n^{-(2+\tau)}$ as $n \to \infty$ for some $0 < \tau < \infty$. We can now express:

$$\begin{aligned}[Q_n(\beta) - Q_0(\beta)] &= [Q_n(\beta) - Q_n(\bar\beta_{k_n}(\beta))] + [Q_n(\bar\beta_{k_n}(\beta)) - Q_n^e(\bar\beta_{k_n}(\beta))] \\ &\quad + [Q_n^e(\bar\beta_{k_n}(\beta)) - Q_n^e(\beta)] + [Q_n^e(\beta) - Q_0(\beta)] \\ &= A_{1n}(\beta) + A_{2n}(\bar\beta_{k_n}(\beta)) + A_{3n}(\beta) + A_{4n}(\beta),\end{aligned}$$

where $Q_n^e(\beta) = \mathrm{E}[Q_n(\beta)]$.

Second, we have that:

$$\begin{aligned}|A_{1n}(\beta)| &\leq n^{-1}\sum_{i=1}^{n}\delta_n^{-1}\left|K\left(\frac{y_i - x_i'\beta}{\delta_n}\right) - K\left(\frac{y_i - x_i'\bar\beta_{k_n}}{\delta_n}\right)\right| \\ &\leq c_1 n^{-1}\sum_{i=1}^{n}\delta_n^{-2}|x_i'(\beta - \bar\beta_{k_n})| \leq c_1\left[n^{-1}\sum_{i=1}^{n}|x_i|\right]\delta_n^{-2}|\beta - \bar\beta_{k_n}|,\end{aligned}$$

and hence:

$$\sup_{\beta\in B}|A_{1n}(\beta)| \leq \left[n^{-1}\sum_{i=1}^{n}|x_i|\right]\left(\frac{c_1 G_1}{\delta_n^2 2^{k_n}}\right) = A_{1n}^*. \tag{18}$$

Now $n^{-1}\sum_{i=1}^n |x_i| = O_p(1)$ since x_i are iid and $\mathrm{E}\{|x_i|\} < \infty$ by Assumptions A1 and A3(i). But since $2^{k_n} \sim \delta_n^{-(2+\tau)}$ as $n \to \infty$ for some $0 < \tau < \infty$, then $\delta_n^2 2^{k_n} \sim \delta_n^{-\tau}$ as $n \to \infty$. Since $\delta_n \to 0$ as $n \to \infty$ by Assumption A6(i), then $\delta_n^{-\tau} \to \infty$ as $n \to \infty$. Hence $A_{1n}^* = o_p(1)$ and so $\sup_{\beta \in B} |A_{1n}(\beta)| = o_p(1)$.

Third, define:

$$m(\beta, \delta) = \int K(u) f_{\varepsilon|x}(x_i'(\beta - \beta_0) + \delta u | x_i)\ d\varepsilon\ dF_X(x),$$

so for any $\delta \neq 0$:

$$\begin{aligned} m(\beta, \delta) &= \int \delta^{-1} K\left(\frac{\varepsilon - x_i'(\beta - \beta_0)}{\delta}\right) f_{\varepsilon|x}(\varepsilon | x_i)\ d\varepsilon\ dF_X(x) \\ &= \mathrm{E}\left[\delta^{-1} K\left(\frac{y_i - x_i'\beta}{\delta}\right)\right], \end{aligned}$$

and thus:

$$A_{2n}(\beta) = n^{-1}\sum_{i=1}^n \left[\delta_n^{-1} K\left(\frac{y_i - x_i'\beta}{\delta_n}\right) - m(\beta, \delta_n)\right]. \tag{19}$$

From Assumption A5(ii) it follows that:

$$\left|\delta_n^{-1} K\left(\frac{y_i - x_i'\beta}{\delta_n}\right) - m(\beta, \delta_n)\right| \leq 2\delta_n^{-1} c_0,$$

while:

$$\begin{aligned} \mathrm{Var}\left[\delta_n^{-1} K\left(\frac{y_i - x_i'\beta}{\delta_n}\right) - m(\beta, \delta_n)\right] &\leq \mathrm{E}\left[\delta_n^{-2} K\left(\frac{y_i - x_i'\beta}{\delta_n}\right)^2\right] \\ &= \int \delta_n^{-2} K\left(\frac{\varepsilon - x'(\beta - \beta_0)}{\delta_n}\right)^2 f_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x) \\ &= \int \delta_n^{-1} K(u)^2 f_{\varepsilon|X}(x'(\beta - \beta_0) + \delta_n u | x)\ du\ dF_X(x) \\ &\leq \delta_n^{-1} \cdot L_0 \cdot c_2, \end{aligned}$$

where $c_2 = \int K(u)^2 du$, which is clearly finite and strictly positive since $|K(\cdot)|$ is uniformly bounded from above and $\int K(u)du$ exists and equals 1 by Assumptions A5(i) and A5(ii).

Thus $A_{2n}(\beta) = n^{-1}\sum_{i=1}^n w_{in}(\beta)$ where for any fixed $\beta \in B$, the $w_{in}(\beta)$ are independently distributed mean zero random variables, which are uniformly bounded from above in absolute value by $b_n = 2\delta_n^{-1}c_0$, and whose variances are uniformly bounded from above by $\nu_n^2 = \delta_n^{-1}L_0c_2$. Hence, by Bernstein's inequality (see Hoeffding, 1963), it follows that for all $\eta > 0$:

$$\begin{aligned}\Pr\{|A_{21n}(\beta)| \geq \eta\} &\leq 2\exp\left\{-\left(\frac{n\eta}{b_n}\right)h\left(\frac{b_n\eta}{\nu_n^2}\right)\right\} \\ &= 2\exp\left\{-(n\delta_n)\left(\frac{3\eta^2}{6L_0c_2 + 4c_0\eta}\right)\right\},\end{aligned} \tag{20}$$

where $h(s) = 3s/(6+2s)$ for all $s > 0$. Since:

$$\sup_{\beta\in B}|A_{2n}(\bar{\beta}_{k_n}(\beta))| = \sup_{\beta\in B_{k_n}}|A_{2n}(\beta)|,$$

then it follows that:

$$\begin{aligned}\Pr\left\{\sup_{\beta\in B}\left|A_{2n}(\bar{\beta}_{k_n}(\beta))\right| \geq \eta\right\} &= \Pr\left\{\sup_{\beta\in B_{k_n}}|A_{2n}(\beta)|\right\} \\ \leq \sum_{s=1}^{N(k_n)}\Pr\{|A_{2n}(\beta^s)| \geq \eta\} &\leq 2^{pk_n+1}\exp\left\{-(n\delta_n)\left(\frac{3\eta^2}{6L_0c_2 + 4c_0\eta}\right)\right\}.\end{aligned}$$

Assumption A6(ii) specifies that $n\delta_n^{1+\sigma} \to \infty$ as $n \to \infty$ for some $0 < \sigma < \infty$, and thus it follows that $n\delta_n$ tends to infinity at a faster rate than $\delta_n^{-\sigma}$. By choice, $2^{k_n} \sim \delta_n^{-(2+\tau)}$ as $n \to \infty$ for some $0 < \tau < \infty$, so 2^{pk_n+1} tends to infinity at the same rate as some positive power of $\delta_n^{-\sigma}$. Together these imply that:

$$\lim_{n\to\infty} 2^{pk_n+1}\exp\left\{-(n\delta_n)\left(\frac{3\eta^2}{6L_0c_2 + 4c_0\eta}\right)\right\} = 0,$$

for all $\eta > 0$, which thus implies that $\sup_{\beta\in B}|A_{2n}(\bar{\beta}_{k_n}(\beta))| = o_p(1)$.

Fourth, we have that:

$$\begin{aligned}\sup_{\beta\in B}|A_{3n}(\beta)| &= \sup_{\beta\in B}\left|\mathrm{E}[Q_n(\bar{\beta}_{k_n})] - \mathrm{E}[Q_n(\beta)]\right| \\ &\leq \mathrm{E}\left\{\sup_{\beta\in B}\left|Q_n(\bar{\beta}_{k_n}) - Q_n(\beta)\right|\right\} = \mathrm{E}\left\{\sup_{\beta\in B}|A_{1n}(\beta)|\right\}.\end{aligned}$$

But from Equation (18) we have that $\sup_{\beta\in B}|A_{1n}(\beta)| \leq A_{1n}^*$ and $\mathrm{E}[A_{1n}^*] = \mathrm{E}\{|x_i|\}\left(\frac{c_1G_1}{\delta_n^2 2^{k_n}}\right)$. But, as argued above, $\delta_n^2 2^{k_n} \to \infty$ as $n \to \infty$, and hence it follows that $\sup_{\beta\in B}|A_{3n}(\beta)| = o(1)$.

Fifth, observe that:

$$A_{4n}(\beta) = m(\beta, \delta_n) - Q_0(\beta) = m(\beta, \delta_n) - m(\beta, 0).$$

But by the line of argument used in the proof of Lemma 1, it follows that $m(\beta, \delta)$ is continuous in (β, δ). Since $\delta_n \to 0$ as $n \to \infty$ and B is compact, it then follows that $\sup_{\beta \in B} |A_{4n}(\beta)| = o(1)$.

Putting all of these properties together, it follows that $\sup_{\beta \in B} |Q_n(\beta) - Q_0(\beta)| = o_p(1)$ as desired. ■

A.2 Proof of Theorem 2

The proof of asymptotic normality involves more stages than the proof of consistency. First, since $\hat{\beta}_n$ is consistent by Theorem 1 and $K(\cdot)$ is twice continuously differentiable, then with probability tending to 1:

$$0 = \left(\left. \frac{\partial Q_n}{\partial \beta} \right|_{\hat{\beta}_n} \right) = \left(\left. \frac{\partial Q_n}{\partial \beta} \right|_{\beta_0} \right) + \left(\left. \frac{\partial^2 Q_n}{\partial \beta \partial \beta'} \right|_{\hat{\beta}_n^*} \right) \left(\hat{\beta}_n - \beta_0 \right),$$

where $\hat{\beta}_n^*$ lies on the line segment joining $\hat{\beta}_n$ and β_0 (as usual we may need to evaluate each row of the second-derivative matrix at different values of $\hat{\beta}_n^*$). Hence, with probability tending to 1, it follows that:

$$\left(\hat{\beta}_n - \beta_0 \right) = - \left(\left. \frac{\partial^2 Q_n}{\partial \beta \partial \beta'} \right|_{\hat{\beta}_n^*} \right)^{-1} \left(\left. \frac{\partial Q_n}{\partial \beta} \right|_{\beta_0} \right).$$

Second, we show in Lemma 3 below that $(n\delta_n^3)^{1/2} \left(\left. \frac{\partial Q_n}{\partial \beta} \right|_{\beta_0} \right)$ converges in distribution to a normal with mean 0 and variance A_0. Note that under the assumptions required for this theorem $(n\delta_n^7) = o(1)$. Third, we show in Lemma 4 below that $\left(\left. \frac{\partial^2 Q_n}{\partial \beta \partial \beta'} \right|_{\hat{\beta}_n^*} \right)$ converges in probability to B_0. Putting these properties together gives us the desired result.

Lemma 3 *Under Assumptions A1–A6 and B1–B5:*

$$(n\delta_n^3)^{1/2} \left(\left. \frac{\partial Q_n}{\partial \beta} \right|_{\beta_0} \right) \xrightarrow{D} \mathcal{N}[0, A_0],$$

where $A_0 = M_1 \cdot \mathrm{E}\left[f_{\varepsilon|X}(0|x_i)(x_i x_i') \right]$ *is positive definite.*

Proof. Observe that:

$$(n\delta_n^3)^{1/2}\left(\frac{\partial Q_n}{\partial\beta}\bigg|_{\beta_0}\right) = -\sum_{i=1}^{n} g_{in} = -ng_n^e - \sum_{i=1}^{n}[g_{in} - g_n^e]$$

where $g_{in} = n^{-1/2}\delta_n^{-1/2}K'\left(\frac{\varepsilon_i}{\delta_n}\right)x_i$ and $g_n^e = \mathrm{E}[g_{in}]$. The proof is based on establishing that $\lim_{n\to\infty} ng_n^e = 0$ and that $\sum_{i=1}^{n}[g_{in} - g_n^e]$ satisfies the conditions of the Liapunov Central Limit Theorem (CLT); see, for example, Theorem 2.4.2 from Bierens (1994).

First, observe that:

$$ng_n^e = \int n^{1/2}\delta_n^{-1/2}K'\left(\frac{\varepsilon}{\delta_n}\right)xf_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x).$$

So, applying integration by parts, we obtain:

$$\begin{aligned} ng_n^e = &\int\left[n^{1/2}\delta_n^{1/2}K\left(\frac{\varepsilon}{\delta_n}\right)f_{\varepsilon|X}(\varepsilon|x)\right]_{-\infty}^{\infty} x\ dF_X(x) \\ &-\int n^{1/2}\delta_n^{1/2}K\left(\frac{\varepsilon}{\delta_n}\right)xf_{\varepsilon|X}^{(1)}(\varepsilon|x)\ d\varepsilon dF_X(x). \end{aligned}$$

Now, Assumption B4(iii) states that $\lim_{u\to\pm\infty}K(u) = 0$, and Assumption A4(i) states that $f_{\varepsilon|X}(\varepsilon|x)$ is uniformly bounded, so these imply that $\left[n^{1/2}\delta_n^{1/2}K\left(\frac{\varepsilon}{\delta_n}\right)f_{\varepsilon|X}(\varepsilon|x)\right]_{-\infty}^{\infty} = 0$. Then, by defining $u = \varepsilon/\delta_n$, we obtain:

$$ng_n^e = -\int n^{1/2}\delta_n^{3/2}K(u)xf_{\varepsilon|X}^{(1)}(\delta_n u|x)\ du\ dF_X(x).$$

Since $f_{\varepsilon|X}(\varepsilon|x)$ is three times continuously differentiable in ε for all x by Assumption B3, then we can take a second-order Taylor series expansion of $f_{\varepsilon|X}^{(1)}(\delta_n u|x)$ around $u = 0$ for given x:

$$f_{\varepsilon|X}^{(1)}(\delta_n u|x) = f_{\varepsilon|X}^{(1)}(0|x) + (\delta_n u)f_{\varepsilon|X}^{(2)}(0|x) + \frac{1}{2}(\delta_n u)^2 f_{\varepsilon|X}^{(3)}(\lambda\delta_n u|x),$$

for some $0 \le \lambda \le 1$ (which may vary with δ_n, u, and x). The continuous differentiability of $f_{\varepsilon|X}^{(1)}(\varepsilon|x) = 0$ with respect to ε given x, combined with the property that $f_{\varepsilon|X}(\varepsilon|x)$ has a maximum at $\varepsilon = 0$ by Assumption A4(iii), implies that $f_{\varepsilon|X}^{(1)}(0|x) = 0$ since $f_{\varepsilon|X}(\varepsilon|x)$ has a maximum at $\varepsilon = 0$. By substituting this result into the Taylor series expansion we obtain:

$$
\begin{aligned}
ng_n^e = &-\int n^{1/2}\delta_n^{5/2}uK(u)xf_{\varepsilon|X}^{(2)}(0|x)\ du\ dF_X(x) \\
&-\left(\frac{1}{2}\right)\int n^{1/2}\delta_n^{7/2}u^2K(u)xf_{\varepsilon|X}^{(3)}(\lambda\delta_n u|x)\ du\ dF_X(x).
\end{aligned} \tag{21}
$$

Moreover:

$$
\begin{aligned}
&\int n^{1/2}\delta_n^{5/2}uK(u)xf_{\varepsilon|X}^{(2)}(0|x)\ du\ dF_X(x) \\
&\quad = (n\delta_n^5)^{1/2}\int uK(u)du\cdot\int xf_{\varepsilon|X}^{(2)}(0|x)\ dF_X(x) = 0,
\end{aligned}
$$

because $\int uK(u)du = 0$, by Assumption B4(ii), and because $\int xf_{\varepsilon|X}^{(2)}(0|x)\ dF_X(x)$ is finite since $f_{\varepsilon|X}^{(2)}(\varepsilon|x)$ is uniformly bounded, by Assumption B3(i), and $\mathrm{E}\{|x|\} < \infty$, by Assumption A3(i). In addition:

$$
\begin{aligned}
&\left|\int n^{1/2}\delta_n^{7/2}u^2K(u)xf_{\varepsilon|X}^{(3)}(\lambda\delta_n u|x)\ du\ dF_X(x)\right| \\
&\qquad\qquad \le (n\delta_n^7)^{1/2}\int u^2|K(u)|\cdot|x|\ \cdot\left|f_{\varepsilon|X}^{(3)}(\lambda\delta_n u|x)\right|\ du\ dF_X(x) \\
&\qquad\qquad \le (n\delta_n^7)^{1/2}\int u^2|K(u)|\ du\cdot\mathrm{E}\{|x_i|\}\ \cdot\sup_{\varepsilon,x}\left|f_{\varepsilon|X}^{(3)}(\varepsilon|x)\right| = o(1),
\end{aligned}
$$

since $\int u^2|K(u)|\ du = M_0 < \infty$, by Assumption B4(iv), $\mathrm{E}\{|x_i|\} < \infty$, by Assumption A3(i), $\sup_{\varepsilon,x}\left|f_{\varepsilon|X}^{(3)}(\varepsilon|x)\right| < \infty$, by Assumption B3(i), and $n\delta_n^7 = o(1)$, by Assumption B5(i). This establishes that $ng_n^e = O(n\delta_n^7) = o(1)$.

Second, fix any $\lambda \in R^p$ and set $z_{in} = [g_{in} - g_n^e]'\lambda$. Clearly, by construction $\mathrm{E}(z_{in}) = 0$. This implies that:

$$
\sum_{i=1}^n \mathrm{E}\{|z_{in}|^2\} = n\mathrm{E}\{|g_{in}'\lambda|^2\} - n^{-1}[(ng_n^e)'\lambda]^2,
$$

and clearly $n^{-1}[(ng_n^e)'\lambda]^2 = o(1)$, since $ng_n^e = o(1)$ as established immediately above. Now:

$$
\begin{aligned}
n\mathrm{E}\{|g_{in}'\lambda|^2\} &= \int \delta_n^{-1}\left[K'\left(\frac{\varepsilon}{\delta_n}\right)\right]^2(x'\lambda)^2 f_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x) \\
&= \int |K'(u)|^2(x'\lambda)^2 f_{\varepsilon|X}(\delta_n u|x)\ du\ dF_X(x),
\end{aligned}
$$

and since $\int |K'(u)|^2 = M_1 < \infty$, by Assumption B4(v), $\delta_n \to 0$, by Assumption A6(i), and $f_{\varepsilon|X}(\varepsilon|x)$ is continuous and uniformly bounded, by Assumptions A4(i) and A4(ii), it follows that:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}\{|z_{in}|^2\} = \int |K'(u)|^2 du \cdot \int (x'\lambda)^2 f_{\varepsilon|X}(0|x)\ du\ dF_X(x)$$

$$= M_1 \cdot \lambda' \mathrm{E}\left[f_{\varepsilon|X}(0|x_i)x_i x_i'\right] \lambda = \omega^2 < \infty. \tag{22}$$

But Assumptions A3(ii) and B1 imply that $\mathrm{E}(x_i x_i') > 0$ for all fixed $\lambda \neq 0$. This combined with the properties that $f_{\varepsilon|X}(0|x_i)$ is uniformly bounded, by Assumption A4(i), and is strictly positive on a set of x_i with probability one, by Assumption A4(iii), implies that $\mathrm{E}\left[f_{\varepsilon|X}(0|x_i)\,(x_i'\lambda)^2\right] = \lambda'\mathrm{E}\left[f_{\varepsilon|X}(0|x_i)x_i x_i'\right]\lambda > 0$ for all $\lambda \neq 0$. Since $\int K(u)du$ exists and is equal to 1, by Assumption A5(i), and $K(\cdot)$ is three times differentiable with $\int |K'(u)|^2 du = M_1 < \infty$, by Assumptions B4(i) and B4(v), it follows that M_1 must be strictly positive and hence that ω^2 is finite and strictly positive. Note that since the data is iid, by Assumption A1:

$$\mathrm{Var}\left[(n\delta_n^3)^{1/2}\left(\frac{\partial Q_n(\beta)}{\partial\beta}\bigg|_{\beta_0}\right)'\lambda\right] = \mathrm{Var}\left[\sum_{i=1}^{n} g_{in}'\lambda\right] = \sum_{i=1}^{n}\mathrm{E}|z_{in}|^2,$$

so:

$$\lim_{n\to\infty}\mathrm{Var}\left[(n\delta_n^3)^{1/2}\left(\frac{\partial Q_n(\beta)}{\partial\beta}\bigg|_{\beta_0}\right)'\lambda\right] = \omega^2,$$

and hence:

$$\lim_{n\to\infty}\mathrm{Var}\left[(n\delta_n^3)^{1/2}\left(\frac{\partial Q_n(\beta)}{\partial\beta}\bigg|_{\beta_0}\right)'\right] = M_1 \cdot \mathrm{E}\left[f_{\varepsilon|X}(0|x_i)x_i x_i\right].$$

Third, observe that for any $\rho > 0$ such that $\mathrm{E}\{|z_{in}|^{2+\rho}\} < \infty$:

$$\sum_{i=1}^{n}\mathrm{E}\{|z_{in}|^{2+\rho}\} \leq 2^{1+\rho} n \left[\mathrm{E}\{|g_{in}'\lambda|^{2+\rho}\} + |(g_n^e)'\lambda|^{2+\rho}\right]. \tag{23}$$

We obtain this by observing that for any value of $r \geq 1$, $|x|^r$ is a convex function of x and hence for any real x_1 and x_2 it follows that $|(x_1+x_2)/2|^r \leq (1/2)|x_1|^r + (1/2)|x_2|^r$ which implies that $|x_1+x_2|^r \leq 2^{r-1}(|x_1|^r + |x_2|^r)$. Equation (23) then follows by setting

$r = 2+\rho$, $x_1 = g'_{in}\lambda$, $x_2 = -(g^e_n)'\lambda$, and noting that the data is iid by Assumption A1. But $n|g^e_n|^{2+\rho} = n^{-(1+\rho)}|ng^e_n|^{2+\rho} = o(1)$, since $ng^e_n = O(n\delta^7_n) = o(1)$ as established above, and in addition:

$$\begin{aligned} n\mathrm{E}\{|g'_{in}\lambda|^{2+\rho}\} &= \int (n\delta_n)^{-\rho/2}\delta_n^{-1}\left|K\left(\frac{\varepsilon}{\delta_n}\right)\right|^{2+\rho}|x'\lambda|^{2+\rho}f_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x) \\ &= (n\delta_n)^{-\rho/2}\int |K(u)|^{2+\rho}|x'\lambda|^{2+\rho}f_{\varepsilon|X}(\delta_n u|x)\ du\ dF_X(x) \\ &\le (n\delta_n)^{-\rho/2}L_0\int |K(u)|^{2+\rho}du\cdot \mathrm{E}\{|x_i|^{2+\rho}\}\to 0, \end{aligned} \tag{24}$$

since $n\delta_n \to \infty$ as a consequence of Assumption A6(ii).

Together, Equations (22) and (24) imply that the conditions of the Liapunov CLT are satisfied, see Theorem 2.4.2 from Bierens (1994), and thus $\sum_{i=1}^n z_{in} \xrightarrow{D} \mathcal{N}[0,\omega^2]$, where $0<\omega^2<\infty$. Since $\lambda \neq 0$ was arbitrary, this implies that $(n\delta^3_n)^{1/2}\left(\left.\frac{\partial Q_n}{\partial\beta}\right|_{\beta_0}\right) \xrightarrow{D} \mathcal{N}[0, A_0]$, where:

$$A_0 = \lim_{n\to\infty}\mathrm{Var}\left[(n\delta^3_n)^{1/2}\left(\left.\frac{\partial Q_n(\beta)}{\partial\beta}\right|_{\beta_0}\right)'\right] = M_1\cdot \mathrm{E}\left[f_{\varepsilon|X}(0|x_i)(x_ix'_i)\right],$$

and is positive definite. ■

Lemma 4 *Under Assumptions A1–A6 and B1–B5:*

$$\left(\left.\frac{\partial^2 Q_n}{\partial\beta\partial\beta'}\right|_{\hat\beta_n}\right)\xrightarrow{p} B_0,$$

where $B_0 = \left(\left.\frac{\partial^2 Q_0}{\partial\beta\partial\beta'}\right|_{\beta_0}\right)$ *is negative definite.*

Proof. The proof of this Lemma follows a similar approach to that of Lemma 2 but in addition makes use of a trimming argument. Fix any $\lambda\in R^p$ and define:

$$H_n(\beta) = \lambda'\left(\left.\frac{\partial^2 Q_n}{\partial\beta\partial\beta'}\right|_{\beta}\right)\lambda = n^{-1}\sum_{i=1}^n \delta_n^{-3}K''\left(\frac{y_i - x'_i\beta}{\delta_n}\right)(x'_i\lambda)^2,$$

noting that this exists by Assumption B4(i). In addition, provisionally define:

$$H_0(\beta) = \lambda'\left(\left.\frac{\partial^2 Q_0}{\partial\beta\partial\beta'}\right|_{\beta}\right)\lambda,$$

(we will establish that $H_0(\beta)$ is well-defined in the course of the proof). Also define $N(k)$, B_k, and $\bar{\beta}_k(\cdot)$ as in the proof of Lemma 2, but now let $\{k_n\}$ be a sequence of monotonically increasing positive integers such that $2^{k_n} \sim \delta_n^{-(4+\tau)}$ for some $0 < \tau < \infty$. Then:

$$\begin{aligned}[H_n(\beta) - H_0(\beta)] &= [H_n(\beta) - H_n(\bar{\beta}_{k_n}(\beta))] + [H_n(\bar{\beta}_{k_n}(\beta)) - H_n^e(\bar{\beta}_{k_n}(\beta))] \\ &\quad + [H_n^e(\bar{\beta}_{k_n}(\beta)) - H_n^e(\beta)] + [H_n^e(\beta) - H_0(\beta)] \\ &= C_{1n}(\beta) + C_{2n}(\bar{\beta}_{k_n}(\beta)) + C_{3n}(\beta) + C_{4n}(\beta).\end{aligned}$$

First, observe that by Assumption B4(vii) and the mean value theorem:

$$\begin{aligned}|C_{1n}(\beta)| &\leq n^{-1}\sum_{i=1}^{n}\delta_n^{-3}\left|K''\left(\frac{y_i - x_i'\beta}{\delta_n}\right) - K''\left(\frac{y_i - x_i'\bar{\beta}_{k_n}(\beta)}{\delta_n}\right)\right|(x_i'\lambda)^2 \\ &\leq \delta_n^{-4}\left|\beta - \bar{\beta}_{k_n}(\beta)\right| \cdot |\lambda|^2 M_3 \cdot \left[n^{-1}\sum_{i=1}^{n}|x_i|^3\right],\end{aligned}$$

which thus implies that:

$$\sup_{\beta\in B}|C_{1n}(\beta)| \leq \left(\frac{G_1 M_3|\lambda|^2}{\delta_n^4 2^{k_n}}\right)\left[n^{-1}\sum_{i=1}^{n}|x_i|^3\right]. \tag{25}$$

Clearly, $n^{-1}\sum_{i=1}^{n}|x_i|^3 = O_p(1)$, by Assumptions A1 and B1, and $\delta_n^4 2^{k_n} \to \infty$ as $n \to \infty$ since $\delta_n^4 2^{k_n} \sim \delta_n^{-\tau}$, with $\tau > 0$ and $\delta_n \to 0$ as $n \to \infty$, by Assumption A6(i). Together these then imply that $\sup_{\beta\in B}|C_{1n}(\beta)| = o_p(1)$.

Second, define:

$$\begin{aligned}h_{in,1}(\beta) &= \delta_n^{-3}K''\left(\frac{y_i - x_i'\beta}{\delta_n}\right)(x_i'\lambda)^2\ \mathbf{1}[(x_i'\lambda)^2 \leq \delta_n^{-2}], \\ h_{in,2}(\beta) &= \delta_n^{-3}K''\left(\frac{y_i - x_i'\beta}{\delta_n}\right)(x_i'\lambda)^2\ \mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}],\end{aligned}$$

so $H_n(\beta) = H_{n,1}(\beta) + H_{n,2}(\beta)$, where $H_{n,j}(\beta) = n^{-1}\sum_{i=1}^{n}h_{in,j}(\beta)$ for $j = 1, 2$. Also define $h_{n,j}^e(\beta) = \mathrm{E}[h_{in,j}(\beta)]$ and $C_{2n,j}(\beta) = H_{n,j}(\beta) - h_{n,j}^e(\beta)$ for $j = 1, 2$. Now, by construction, $|h_{in,1}(\beta)| \leq \delta_n^{-5}M_2$, where $M_2 = \sup_{u\in R}|K''(u)|$ from Assumption B4(vi). Hence $|h_{n,1}^e(\beta)| \leq \delta_n^{-5}M_2$ and so $|h_{in,1}(\beta) - h_{n,1}^e(\beta)| \leq \bar{b}_n$, where $\bar{b}_n = 2\delta_n^{-5}M_2$.

In addition:

$$\begin{aligned}\mathrm{Var}[h_{in,1}(\beta)] &\leq \mathrm{E}[h_{in,1}(\beta)^2] \\ &\leq \int \delta_n^{-6}\left[K''\left(\frac{\varepsilon - x'(\beta-\beta_0)}{\delta_n}\right)\right]^2 (x_i'\lambda)^4 f_{\varepsilon|x}(\varepsilon|x)\ d\varepsilon\ dF_X(x) \\ &= \int \delta_n^{-5}[K''(u)]^2(x_i'\lambda)^4 f_{\varepsilon|x}(x'(\beta-\beta_0)+\delta_n u|x)\ du\ dF_X(x) \\ &\leq \delta_n^{-5}\cdot L_0 \cdot \int [K''(u)]^2 du\ \cdot \mathrm{E}[(x_i'\lambda)^4] = \delta_n^{-5}d_0 = \bar{\nu}_n^2, \qquad (26)\end{aligned}$$

where d_0 is a finite positive constant as a consequence of Assumptions A4(i), B1, and B4(viii). Then, by Bernstein's inequality, it follows that:

$$\Pr\left\{|C_{2n,1}(\beta)| \geq \eta\right\} \leq 2\exp\left\{-\left(\frac{n\eta}{\bar{b}_n}\right) h\left(\frac{\bar{b}_n\eta}{\bar{\nu}_n^2}\right)\right\} = 2\exp\left\{-\left(\frac{3n\delta_n^5\eta^2}{6d_0+4M_2\eta}\right)\right\},$$

where $h(s) = 3s/(6+2s)$ for all $s > 0$, as in the proof of Lemma 2 above. But $\sup_{\beta\in B}|C_{2n,1}(\bar{\beta}_{k_n}(\beta))| = \sup_{\beta\in B_{k_n}}|C_{1n}(\beta)|$, so it follows that:

$$\begin{aligned}\Pr\left\{\sup_{\beta\in B}|C_{2n,1}(\bar{\beta}_{k_n}(\beta))| \geq \eta\right\} &\leq \sum_{s=1}^{N(k_n)} \Pr\left\{|C_{2n,1}(\beta^s)| \geq \eta\right\} \\ &\leq 2^{pk_n+1}\exp\left\{-\left(\frac{3n\delta_n^5\eta^2}{6d_0+4M_2\eta}\right)\right\}.\end{aligned}$$

Now, $n\delta_n^{5+\sigma} \to \infty$ for some $\sigma > 0$, by Assumption B5(ii), and thus $n\delta_n^5$ tends to infinity more rapidly than $\delta_n^{-\sigma}$. Since $2^{k_n} \sim \delta_n^{-(4+\tau)}$ and $\delta_n \to 0$ as $n \to \infty$, by Assumption A6(i), it follows that $\Pr\left\{\sup_{\beta\in B}|C_{2n,1}(\bar{\beta}_{k_n}(\beta))| \geq \eta\right\}$ tends to zero as $n \to \infty$ for any fixed value of η, and thus $\sup_{\beta\in B}|C_{2n,1}(\bar{\beta}_{k_n}(\beta))| = o_p(1)$.

Next, observe that by Assumption B4(vi):

$$\sup_{\beta\in B_{k_n}}|h_{in,2}(\beta)| \leq \delta_n^{-3}M_2(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}]$$

for all β. So:

$$\sup_{\beta\in B_{k_n}}|C_{2n,2}(\beta)| \leq \sup_{\beta\in B_{k_n}}\left|h_{n,2}^e(\beta)\right| + \delta_n^{-3}M_2\left(n^{-1}\sum_{i=1}^n (x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}]\right),$$

and hence:

$$\mathrm{E}\{\sup_{\beta\in B_{k_n}}|C_{2n,2}(\beta)|\} \leq 2\delta_n^{-3}M_2\mathrm{E}\left\{(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}]\right\},$$

noting that $\sup_{\beta\in B_{k_n}}\left|h^e_{n,2}(\beta)\right| = \sup_{\beta\in B_{k_n}}\left|\mathrm{E}[h_{in,2}(\beta)]\right| \leq \mathrm{E}\{\sup_{\beta\in B_{k_n}}\left|h_{in,2}(\beta)\right|\}$.

Now, $r>1$ such that $E\{|X_i|^{2r}\}<\infty$ and then the Hölder inequality implies that:

$$\mathrm{E}\left\{(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2>\delta_n^{-2}]\right\} \leq \left[\mathrm{E}\left\{|x_i'\lambda|^{2r}\right\}\right]^{1/r}\left[\mathrm{E}\left\{\mathbf{1}\left[(x_i'\lambda)^2>\delta_n^{-2}\right]^s\right\}\right]^{1/s},$$

where $s=r/(r-1)$. But, since $s>0$, then:

$$\begin{aligned}\mathrm{E}\left\{\mathbf{1}\left[(x_i'\lambda)^2>\delta_n^{-2}\right]^s\right\} &= \mathrm{E}\left\{\mathbf{1}\left[(x_i'\lambda)^2>\delta_n^{-2}\right]\right\} = \Pr\left\{(x_i'\lambda)^2>\delta_n^{-2}\right\}\\ &= \Pr\left\{|x_i'\lambda|^{2r}>\delta_n^{-2r}\right\} \leq \frac{\mathrm{E}\left\{|x_i'\lambda|^{2r}\right\}}{\delta_n^{-2r}},\end{aligned}$$

by the Markov inequality. Hence:

$$\mathrm{E}\left\{(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2>\delta_n^{-2}]\right\} \leq \left[\mathrm{E}\left\{|x_i'\lambda|^{2r}\right\}\right]^{1/r}\left[\frac{\mathrm{E}\left\{|x_i'\lambda|^{2r}\right\}}{\delta_n^{-2r}}\right]^{1/s} = \mathrm{E}\left\{|x_i'\lambda|^{2r}\right\}\delta_n^{2r/s},$$

since $(1/s)+(1/r)=1$. But then $2r/s = 2r\times(r-1)/r = 2(r-1)$, so:

$$\mathrm{E}\left\{(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2>\delta_n^{-2}]\right\} \leq \mathrm{E}\{|x_i'\lambda|^{2r}\}\delta_n^{2(r-1)},$$

and thus $\mathrm{E}\{\sup_{\beta\in B_{k_n}}|C_{in,2}(\beta)|\} = O(\delta_n^{2r-5})$. Now, setting $r=(5+\xi)/2$, we have that $r>1$ and $E\{|x_i|^{2r}\}<\infty$ by Assumption B1, and hence $\mathrm{E}\{\sup_{\beta\in B_{k_n}}|C_{in,2}(\beta)|\} = O(\delta_n^{\xi}) = o(1)$ since $\xi>0$ and $\delta_n=o(1)$, by Assumption A6(i). Hence $\sup_{\beta\in B_{k_n}}|C_{2n,2}(\beta)| = o_p(1)$. Since $C_{2n}(\beta) = C_{2n,1}(\beta)+C_{2n,2}(\beta)$, this implies that $\sup_{\beta\in B_{k_n}}|C_{2n}(\bar{\beta}_{k_n}(\beta))| = o_p(1)$.

Third, observe that for any fixed $\beta\in B$, $C_{3n}(\beta) = -\mathrm{E}[C_{1n}(\beta)]$, and hence from Equation (25) it follows that:

$$\begin{aligned}\sup_{\beta\in B}|C_{3n}(\beta)| \leq \mathrm{E}\{\sup_{\beta\in B}|C_{1n}(\beta)|\} &\leq \left(\frac{G_1M_3|\lambda|^2}{\delta_n^4 2^{k_n}}\right)\mathrm{E}\left[n^{-1}\sum_{i=1}^n|x_i|^3\right]\\ &= O(\delta_n^{-4}2^{-k_n}) = o(1).\end{aligned}$$

Finally, observe that by repeated application of integration by parts, it follows that:

$$\begin{aligned}H^e_n(\beta) &= \int \delta_n^{-3}K''\left(\frac{\varepsilon - x'(\beta-\beta_0)}{\delta_n}\right)(x_i'\lambda)^2\ f_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x)\\ &= \int \delta_n^{-1}K\left(\frac{\varepsilon - x'(\beta-\beta_0)}{\delta_n}\right)(x_i'\lambda)^2\ f^{(2)}_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x)\\ &= \int K(u)(x_i'\lambda)^2\ f^{(2)}_{\varepsilon|X}(x'(\beta-\beta_0)+\delta_n u|x)\ du\ dF_X(x).\end{aligned}$$

Since $f^{(2)}_{\varepsilon|X}(\varepsilon|x)$ is uniformly bounded from above and continuous in ε for all x, by Assumption B3(i), then $H^e_n(\beta)$ converges to $H_0(\beta)$ uniformly over $\beta \in B$, where:

$$H_0(\beta) = \mathrm{E}\left[f^{(2)}_{\varepsilon|X}(x_i'(\beta-\beta_0)|x_i)(x_i'\lambda)^2\right] = \lambda'\mathrm{E}\left[f^{(2)}_{\varepsilon|X}(x_i'(\beta-\beta_0)|x_i)x_ix_i'\right]\lambda.$$

But since $\mathrm{E}(x_ix_i')$ is finite, by Assumption B1, and $f^{(j)}_{\varepsilon|X}(\varepsilon|x)$ is continuous in ε and uniformly bounded from above for $j = 0, 1, 2, 3$, by Assumptions A4(i) and B3(i), then we can interchange the order of taking derivatives with respect to β and taking expectations with respect to x_i to establish that:

$$\mathrm{E}\left[f^{(2)}_{\varepsilon|X}(x_i'(\beta-\beta_0)|x_i)x_ix_i'\right] = \frac{\partial^2}{\partial\beta\partial\beta'}\mathrm{E}\left[f_{\varepsilon|X}(x_i'(\beta-\beta_0)|x_i)\right],$$

and hence that $H_0(\beta) = \lambda'\left(\frac{\partial^2 Q_0(\beta)}{\partial\beta\partial\beta'}\right)\lambda$ as stated at the start of the proof.

Putting all of these results together, we have that:

$$\sup_{\beta\in B}|H_n(\beta) - H_0(\beta)| = o_p(1),$$

and since $\lambda \neq 0$ was set at an arbitrary value it follows that:

$$\sup_{\beta\in B}\left|\left(\frac{\partial^2 Q_n(\beta)}{\partial\beta\partial\beta'}\right) - \left(\frac{\partial^2 Q_0(\beta)}{\partial\beta\partial\beta'}\right)\right| = o_p(1),$$

as desired. ■

A.3 Proof of Theorem 3

It is sufficient to establish that $\widehat{A}_n$ converges in probability to A_0 and that $\widehat{B}_n$ converges in probability to B_0. Define:

$$\widetilde{A}_n(\beta) = n^{-1}\sum_{i=1}^n \delta_n^{-1}\left[K'\left(\frac{y_i - x_i'\beta}{\delta_n}\right)\right]^2(x_ix_i'),$$

$$\widetilde{B}_n(\beta) = n^{-1}\sum_{i=1}^n \delta_n^{-3}K''\left(\frac{y_i - x_i'\beta}{\delta_n}\right)(x_ix_i') = \frac{\partial^2 Q_n(\beta)}{\partial\beta\partial\beta'},$$

and note that $\widehat{A}_n = \widetilde{A}_n(\hat{\beta}_n)$ and $\widehat{B}_n = \widetilde{B}_n(\hat{\beta}_n)$. From Lemma 4 it follows that $\widehat{B}_n$ converges in probability to B_0.

To establish that $\widehat{A}_n$ converges in probability to A_0 we use an approach similar to that used in the proofs of Lemmas 2 and 4 above. Fix any p-vector $\lambda \neq 0$ and define $S_n(\beta) = \lambda'\widetilde{A}_n(\beta)\lambda$, $S_n^e(\beta) = \mathrm{E}[S_n(\beta)]$, and $S_0(\beta) = \lim_{n\to\infty} S_n^e(\beta)$; note that we will establish the existence of the relevant expectations and limits in the course of this proof. In addition, define $N(k)$, B_k, $\bar{\beta}_k(\cdot)$, and $\{k_n\}$ as in the proof of Lemma 4. Then we have that:

$$\begin{aligned} S_n(\beta) - S_0(\beta) &= \left[S_n(\beta) - S_n(\bar{\beta}_{k_n}(\beta))\right] + \left[S_n(\bar{\beta}_{k_n}(\beta)) - S_n^e(\bar{\beta}_{k_n}(\beta))\right] \\ &\quad + \left[S_n^e(\bar{\beta}_{k_n}(\beta)) - S_n^e(\beta)\right] + \left[S_n^e(\beta) - S_0(\beta)\right] \\ &= D_{1n}(\beta) + D_{2n}(\bar{\beta}_{k_n}(\beta)) + D_{3n}(\beta) + D_{4n}(\beta). \end{aligned}$$

First, observe that by Assumptions A5(iii) and B(vi) together with the mean value theorem:

$$\begin{aligned} |D_{1n}(\beta)| &\leq n^{-1}\sum_{i=1}^{n} \delta_n^{-1} \left| \left\{K'\left(\frac{y_i - x_i'\beta}{\delta_n}\right)\right\}^2 - \left\{K'\left(\frac{y_i - x_i'\bar{\beta}_{k_n}}{\delta_n}\right)\right\}^2 \right| (x_i'\lambda)^2 \\ &\leq 2c_1 M_2 n^{-1} \sum_{i=1}^{n} \delta_n^{-2} |x_i'(\beta - \bar{\beta}_{k_n})| (x_i'\lambda)^2 \\ &\leq 2c_1 M_2 |\beta - \bar{\beta}_{k_n}| \cdot |\lambda|^2 n^{-1} \sum_{i=1}^{n} \delta_n^{-2} |x_i|^3, \end{aligned} \tag{27}$$

and hence:

$$\sup_{\beta\in B} |D_{1n}(\beta)| \leq \left[n^{-1}\sum_{i=1}^{n} |x_i|^3\right] \left(\frac{c_1^3 G_1 |\lambda|^2}{\delta_n^2 2^{k_n}}\right) = o_p(1),$$

since $\mathrm{E}\{|x_i|^3\} < \infty$, by Assumption B1, and $\delta_n^2 2^{k_n} \sim \delta_n^{-(2+\tau)}$ for some $0 < \tau < \infty$ with $\delta_n = o(1)$, by Assumption A6(i).

Second, define:

$$s_{in,1}(\beta) = \delta_n^{-1}\left[K'\left(\frac{y_i - x_i'\beta}{\delta_n}\right)\right]^2 (x_i'\lambda)^2 \mathbf{1}[(x_i'\lambda)^2 \leq \delta_n^{-2}],$$

$$s_{in,2}(\beta) = \delta_n^{-1}\left[K'\left(\frac{y_i - x_i'\beta}{\delta_n}\right)\right]^2 (x_i'\lambda)^2 \mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}],$$

so $S_n(\beta) = S_{n,1}(\beta) + S_{n,2}(\beta)$, where $S_{n,j}(\beta) = n^{-1}\sum_{i=1}^{n} s_{in,j}(\beta)$ for $j = 1, 2$. Also define $s_{n,j}^e(\beta) = \mathrm{E}[s_{in,j}(\beta)]$ and $D_{2n,j}(\beta) = S_{n,j}(\beta) - s_{n,j}^e(\beta)$. Then $|s_{in,1}(\beta)| \leq \delta_n^{-3} c_1^2$

by Assumption A5(iii); hence $|s^e_{n,1}(\beta)| \leq \delta_n^{-3}c_1^2$ and so $|s_{in,1}(\beta) - s^e_{n,1}(\beta)| \leq \tilde{b}_n = 2\delta_n^{-3}c_1^2$. In addition, by Assumptions A4, A5(iii), and B1:

$$\begin{aligned}\mathrm{Var}[s_{in,1}(\beta)] &\leq \mathrm{E}[s_{in,1}(\beta)^2] \\ &\leq \int \delta_n^{-2}\left[K'\left(\frac{\varepsilon - x'(\beta-\beta_0)}{\delta_n}\right)\right]^4 (x_i'\lambda)^4 f_{\varepsilon|X}(\varepsilon|x)\ d\varepsilon\ dF_X(x) \\ &\leq \delta_n^{-1}\cdot L_0 \cdot \int [K'(u)]^4\ du \cdot \mathrm{E}[(x_i'\lambda)^4] = \delta_n^{-1}d_1 = \tilde{\nu}^2,\end{aligned}$$

where d_1 is a finite positive constant. By Bernstein's inequality it follows that:

$$\Pr\left\{|D_{2n,1}(\beta)| \geq \eta\right\} \leq 2\exp\left\{-\left(\frac{n\eta}{\tilde{b}_n}\right)h\left(\frac{\tilde{b}_n\eta}{\tilde{\nu}_n^2}\right)\right\} \leq 2\exp\left\{-\left(\frac{n\delta_n^3\eta h_0}{2c_1^2}\right)\right\},$$

where $h_0 = h\left(2c_1^2\eta\delta_0^{-2}d_1^{-1}\right)$ and $\delta_0 = \sup_{m\geq 1}\delta_n$. Since $\sup_{\beta\in B}|D_{2n,1}(\bar{\beta}_{k_n}(\beta))| = \sup_{\beta\in B_{k_n}}|D_{1n}(\beta)|$, it follows that:

$$\begin{aligned}\Pr\left\{\sup_{\beta\in B}|D_{2n,1}(\bar{\beta}_{k_n}(\beta))| \geq \eta\right\} &\leq \sum_{s=1}^{N(k_n)}\Pr\left\{|D_{2n,1}(\beta^s)| \geq \eta\right\} \\ &\leq 2^{pk_n+1}\exp\left\{-\left(\frac{n\delta_n^3\eta h_0}{2c_1^2}\right)\right\}.\end{aligned}$$

Now $n\delta_n^{5+\sigma} \to \infty$ for some $\sigma > 0$, by Assumption B5(ii), and thus $n\delta_n^3$ tends to infinity more rapidly than $\delta_n^{-(2+\sigma)}$. Since 2^{k_n} tends to infinity at the same rate as some negative power of δ_n, it then follows that $\Pr\left\{\sup_{\beta\in B}|D_{2n,1}(\bar{\beta}_{k_n}(\beta))| \geq \eta\right\}$ tends to zero as $n \to \infty$ for any fixed value of η and thus $\sup_{\beta\in B}|D_{2n,1}(\bar{\beta}_{k_n}(\beta))| = o_p(1)$.

Third, observe that by Assumption A5(iii):

$$\sup_{\beta\in B_{k_n}}|s_{in,2}(\beta)| \leq c_1^2\delta_n^{-1}(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}],$$

for all β, so:

$$\sup_{\beta\in B_{k_n}}|D_{2n,2}(\beta)| \leq \sup_{\beta\in B_{k_n}}\left|s^e_{n,2}(\beta)\right| + c_1^2\delta_n^{-1}\left(n^{-1}\sum_{i=1}^n(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}]\right),$$

and hence:

$$\mathrm{E}\{\sup_{\beta\in B_{k_n}}|D_{2n,2}(\beta)|\} \leq 2c_1^2\delta_n^{-1}\mathrm{E}\left\{(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}]\right\},$$

noting that $\sup_{\beta\in B_{k_n}} \left|s^e_{n,2}(\beta)\right| = \sup_{\beta\in B_{k_n}} |\mathrm{E}[s_{in,2}(\beta)]| \leq \mathrm{E}\{\sup_{\beta\in B_{k_n}} |s_{in,2}(\beta)|\}$. But as established in the proof of Lemma 4, $\mathrm{E}\left\{(x_i'\lambda)^2\mathbf{1}[(x_i'\lambda)^2 > \delta_n^{-2}]\right\} = O(\delta_n^{2(r-1)})$ for any $r > 1$ such that $E\{|x_i|^{2r}\} < \infty$. Setting $r = (5+\xi)/2$, as in the proof of Lemma 4, it follows that $\mathrm{E}\{\sup_{\beta\in B_{k_n}} |D_{in,2}(\beta)|\} = O(\delta_n^{2r-3}) = O(\delta_n^{2+\xi}) = o(1)$ since $\xi > 0$ and $\delta_n = o(1)$ by Assumption A6(i). It then follows from the Markov inequality that $\sup_{\beta\in B_{k_n}} |D_{2n,2}(\beta)| = o_p(1)$ and thus $\sup_{\beta\in B_{k_n}} |D_{2n}(\bar{\beta}_{k_n}(\beta))| = o_p(1)$ since $D_{2n}(\beta) = D_{2n,1}(\beta) + D_{2n,2}(\beta)$ and $\sup_{\beta\in B_{k_n}} |D_{1n,2}(\beta)| = o_p(1)$, as established above.

Fourth, observe that, for any fixed $\beta \in B$, $D_{3n}(\beta) = -\mathrm{E}[D_{1n}(\beta)]$ and hence:

$$\sup_{\beta\in B} |D_{3n}(\beta)| \leq \mathrm{E}\{\sup_{\beta\in B} |D_{1n}(\beta)|\} \leq \left(\frac{c_1^3 G_1 |\lambda|^2}{\delta_n^2 2^{k_n}}\right) \mathrm{E}\left[n^{-1}\sum_{i=1}^{n} |x_i|^3\right]$$
$$= O(\delta_n^{-2} 2^{-k_n}) = o(1),$$

by Assumptions A5(iii), A6(i), and B1.

Fifth, observe that:

$$S_n^e(\beta) = \int \delta_n^{-1}\left[K'\left(\frac{\varepsilon - x'(\beta-\beta_0)}{\delta_n}\right)\right]^2 (x_i'\lambda)^2 f_{\varepsilon|X}(\varepsilon|x)\; d\varepsilon\; dF_X(x)$$
$$= \int [K'(u)]^2 (x_i'\lambda)^2 f_{\varepsilon|X}(x'(\beta-\beta_0) + \delta_n u|x)\; du\; dF_X(x).$$

But B is compact, by Assumption A2, $\int [K'(u)]^2\; du < \infty$, by Assumption B4(v), $\delta_n \to 0$, by Assumption A6(i), $\mathrm{E}\{|x|^2\} < \infty$, by Assumption B1, and $f_{\varepsilon|X}(\varepsilon|x)$ is uniformly bounded from above and continuous in ε for all x, by Assumptions A4(i) and A4(ii). Hence it follows by dominated convergence that $S_n^e(\beta)$ converges uniformly over B to:

$$S_0(\beta) = \int [K'(u)]^2\; du \cdot \lambda' \mathrm{E}\left[f_{\varepsilon|X}(x'(\beta-\beta_0)|x)(xx')\right]\lambda,$$

as $n \to \infty$ so $\sup_{\beta\in B} |S_n(\beta) - S_0(\beta)| = o(1)$. But $\hat{\beta}_n$ converges in probability to β_0, by Theorem 1, so it follows that $S_n(\hat{\beta}_n)$ converges in probability to $S_0(\beta_0)$, where:

$$S_0(\beta_0) = \int [K'(u)]^2\; du \cdot \lambda' \mathrm{E}\left[f_{\varepsilon|X}(x'(0|x)(xx')\right]\lambda = \lambda' A_0 \lambda.$$

Since $\lambda \neq 0$ was fixed at an arbitrary value this implies that $\widehat{A}_n$ converges in probability to A_0.

REFERENCES

Amemiya, T. (1985). *Advanced Econometrics*, Cambridge (MA): Harvard University Press.

Averett, S. and Korenman, S. (1996). "The Economic Reality of the Beauty Myth," *The Journal of Human Resources*, 31, 304-330.

Baldauf, M. and Santos Silva, J.M.C. (2009), On the Use of Robust Regression in Econometrics, Department of Economics, University of Essex, Discussion Paper No 664.

Beaton, A.E. and Tukey, J.W. (1974). "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, 16, 146-185.

Bierens, H.J. (1994). *Topics in Advanced Econometrics*, Cambridge: Cambridge University Press.

Cawley, J. (2004). "The Impact of Obesity on Wages," *Journal of Human Resources*, 39, 451-474.

Chernoff, H. (1964)."Estimation of the Mode," *Annals of the Institute of Statistical Mathematics*, 16, 31-41.

Chesher, A. (1995). "A Mirror Image Invariance for M-Estimators," *Econometrica*, 63, 207-11.

Chesher, A. and Peters, S. (1994). "Symmetry, Regression Design, and Sampling Distributions," *Econometric Theory*, 10, 116-129.

Chou, S.-Y., Grossman, M. and Saffer, H. (2004). "An Economic Analysis of Adult Obesity: Results from the Behavioral Risk Factor Surveillance System," *Journal of Health Economics*, 23, 565-587.

Collomb, G., Härdle, W. and Hassani, S. (1987). "A note on prediction via estimation of the conditional mode function," *Journal of Statistical Planning and Inference*, 15, 227-236.

Cutler, D., Glaeser, E. and Shapiro, J. (2003). "Why have Americans Become more Obese?", *Journal of Economic Perspectives*, 17, 93-118.

Dalenius, T. (1965)."The Mode–A Neglected Statistical Parameter," *Journal of the Royal Statistical Society, Series A*, 128, 110-117.

Department of Health (2004). *Choosing Health: Making Healthy Choices Easier*, London: The Stationery Office.

Hall, P., Racine, J.S. and Li, Q. (2004), "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99, 1015-1026.

Hansen, B.E. (1996), "Stochastic Equicontinuity for Unbounded Dependent Heterogeneous Arrays," *Econometric Thoery*, 12, 347-349.

Hayfield, T. and Racine, J.S. (2008). "Nonparametric Econometrics: The np Package," *Journal of Statistical Software* 27(5). Available at: `http://www.jstatsoft.org/v27/i05/`.

Heiberger, R.M. and Becker, R.A. (1992). "Design of an S Function for Robust Regression Using Iteratively Reweighted Least Squares," *Journal of Computational and Graphical Statistics*, 1, 181-196.

Hoeffding, W. (1963). "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30.

Horowitz, J.L. (1992). "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

Huber, P.J. (1973). "Robust Regression: Asymptotics, Conjectures and Monte Carlo." *Annals of Statistics*, 1, 799-821.

Kim, J.K. and Pollard, D. (1990)."Cube-Root Asymptotics," *Annals of Statistics*, 18, 191-219.

Koenker, R. and Bassett Jr., G.S. (1978). "Regression Quantiles," *Econometrica*, 46, 33-50.

Lee, M.J. (1989). "Mode Regression," *Journal of Econometrics*, 42, 337-349.

Lee, M.J. (1993). "Quadratic Mode Regression," *Journal of Econometrics*, 57, 1-19.

Lee, M.J. and Kim, H.J. (1998). "Semiparametric Econometric Estimators for a Truncated Regression Model: A review with an extension," *Statistica Neerlandica*, 52, 200-225.

Manski, C.F. (1991). "Regression," *Journal of Economic Literature*, 29, 34-50.

Maronna, R.A., Martin R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*, Chichester (UK): John Wiley & Sons.

Mathworks. (2008). *Statistics Toolbox User's Guide, Version 7.* Natick (MA): The Mathworks Inc.

McFadden, D.L. and Newy, W.K. (1994). "Large Sample Estimation and Hypothesis Testing ," Ch 36 in R.F. Engle and D.L. McFadden, (eds.) *Handbook of Econometrics, Vol. 4*, 2111-2245, North Holland: Amsterdam.

Mills, T.C. (2009). "Forecasting Obesity Trends in England," *Journal of The Royal Statistical Society Series A*, 172, 107-117.

Morris, S. (2006). "Body Mass Index and Occupational Attainment," *Journal of Health Economics*, 25, 347-364.

Morris, S. (2007). "The Impact of Obesity on Employment," *Labour Economics*, 14, 413-433.

Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33, 1065-1076.

Quintela-Del-Rio, A. and Vieu, Ph. (1997). "A nonparametric conditional mode estimate," *Journal of Nonparametric Statistics*, 8, 253-266.

Samanta, M. and Thavaneswarn, A. (1990). "Non-parametric estimation of the conditional mode." Communications in Statistics – Theory and Methods, 19, 4515-4524.

SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide*, Cary (NC): SAS Institute Inc.

Seo, M.H. and Linton, O. (2007). "A Smoothed Least Squares Estimator for Threshold Regression Models." *Journal of Econometrics*, 141, 704-735.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

StataCorp. (2007). *Statistical Software: Release 10.0.* College Station (TX): Stata Corporation

U.S. Department of Health and Human Services (2001). *The Surgeon General's Call to Action to Prevent and Decrease Overweight and Obesity*, Rockville (MD): U.S. Department of Health and Human Services, Public Health Service, Office of the Surgeon General.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S, 4th ed.*, New York (NY): Springer.

Ziegler, K. (2003). "On the asymptotic normality of kernel regression estimators of the mode in the random design model." *Journal of Statistical Planning and Inference*, 115, 123-144.